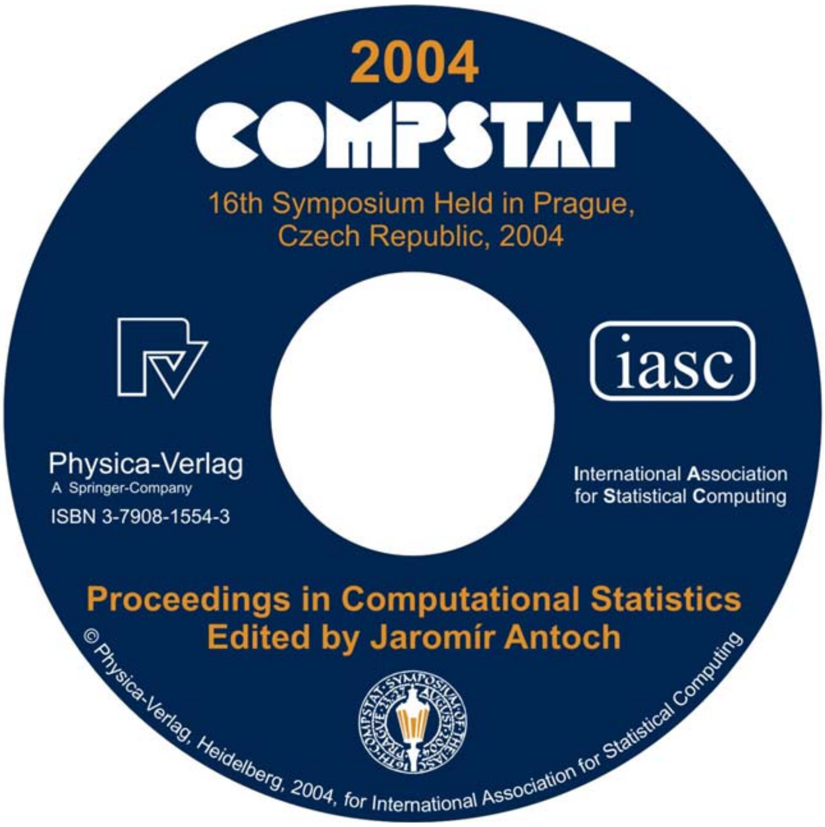


2004 COMPSTAT

16th Symposium Held in Prague,
Czech Republic, 2004

Physica-Verlag
A Springer-Company
ISBN 3-7908-1554-3

iasc

International Association
for Statistical Computing

Proceedings in Computational Statistics
Edited by Jaromír Antoch

COMPSTAT 2004

COMPSTAT

Proceedings
in Computational Statistics

16th Symposium Held in Prague,
Czech Republic, 2004

Edited by
Jaromir Antoch

With 151 Figures
and 38 Tables

Physica-Verlag
A Springer Company

Prof. Dr. Jaromir Antoch
Charles University Prague
Faculty of Mathematics and Physics
Department of Statistics and Probability
Sokolovska 83
18675 Prague 8 – Karlin
Czech Republic
antoch@karlin.mff.cuni.cz

ISBN 3-7908-1554-3 Physica-Verlag Heidelberg

Cataloging-in-Publication Data
Library of Congress Control Number: 2004108446

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Softcover-Design: Erich Kirchner, Heidelberg

SPIN 11015154 88/3130-5 4 3 2 1 0 – Printed on acid-free paper

Foreword

Statistical computing provides the link between the statistical theory and applied statistics. As at previous COMPSTATs, the scientific programme covered all aspects of this link, from the development and implementation of new statistical ideas through to user experiences and software evaluation. Following extensive discussions, a number of changes have been introduced by giving more focus to the individual sessions, involve more people in the planning of sessions, and make links with other societies as Interface or International Federation of Classification Societies (IFCS) involved in statistical computing. The proceedings should appeal to anyone working in statistics and using computers, whether in universities, industrial companies, government agencies, research institutes or as software developers.

This proceedings would not exist without the help of many people. Among them I would like to thank especially to the SPC members D. Banks (USA), H. Ekblom (S), P. Filzmoser (A), W. Härdle (D), J. Hinde (IRE), F. Murtagh (UK), J. Nakano (JAP), A. Prat (E), A. Rizzi (I), G. Sawitzki (D) and E. Wegman (USA); the session organizers D. Cook (USA), D. Banks (IFCS, USA) C. Croux (B), L. Edler (D), V. Esposito Vinzi (I), F. Ferraty (F), V. Kůrková (CZ), M. Müller (D), J. Nakano (ARS IASC, JAP), H. Nyquist (S), D. Peña (E), M. Schimek (A), G. Tunnicliffe-Wilson (GB) and E. Wegman (Interface, USA); as well as to all who contributed and/or refereed the papers.

Last but not least, I must sincerely thank my colleagues from Department of Statistics of the Charles University, Institute of Computer Science of the Czech Academy of Sciences, Czech Technical University, Technical University of Liberec and to Mme Anna Kotěšovcová from Conforg Ltd. Without their substantial help neither this book nor the COMPSTAT 2004 would exist.

My final thanks go to Mme Bilkova and Mme Pickova, who retyped most of the contributions and prepared the final volume, and Mme G. Keidel from the Springer Verlag, Heidelberg, who extremely carefully checked the final printing.

Prague May 15, 2004
Jaromír Antoch

Contents

Invited papers

Contributed papers

THE HISTORY OF COMPSTAT AND KEY-STEPS OF STATISTICAL COMPUTING DURING THE LAST 30 YEARS

Wilfried Grossmann, Michael G. Schimek and Peter Paul Sint

Key words: COMPSTAT symposium, computational statistics, history of statistics, statistical computing, statistical languages, statistical software.

COMPSTAT 2004 section: Historical keynote.

1 Introduction

First of all we try to trace the situation and the ideas that culminated in the first COMPSTAT symposium in the year 1974 held at the University of Vienna, Austria. Special emphasis is given to the memories of our founding member P. P. Sint who had been the driving force behind early COMPSTAT and had served it for twenty years.

At the time COMPSTAT was established computing technology was in its infancy. Yet it was well understood that computing would play a vital role in the future progress of statistics. The impact of the first digital computer in the Department of Statistics at the University of Vienna on the local statistics community is described. After the first computational statistics event in 1974 it was anything but clear that the COMPSTAT symposia would go on for decades as an international undertaking to be incorporated as early as 1978 into the International Association for Statistical Computing (IASC, `http://www.iasc-isi.org/`), a Section of the International Statistical Institute (ISI).

After the description of the background against which the COMPSTAT idea emerged, the subject area of computational statistics is critically discussed from a historical perspective. Key steps of development are pointed out. Special consideration is given to the impact of statistical theory, computing (algorithms), computer science, and applications. Further we provide an overview of the symposia and trace the topics across 30 years, the period of historic interest. Finally we draw conclusions, also with respect to recent developments.

2 The early history of electronic computing

To start off we describe the situation of computing technology in post-war Vienna and the prominent role of the Department of Statistics (later on Statistics and Informatics) at the University of Vienna. Also the Mathematics Department of this university is of historic interest. There, a well-attended seminar was held in summer 1962 by the Viennese mathematician N. Hofreiter “Zur Programmiernug von elektronischen Rechenmaschinen” (“On the programming of electronic calculators"). Topics were one and two address machines, connectors, and programming of simple loops. The treatment was purely theoretical and no specific machine was envisioned. Highlight was an excursion to the first-ever electronic computer at the university, a Burroughs Datatron 205, installed 1960 at the Department of Statistics. The same professor held classes in computing which started from slide rulers and did not go beyond mechanical calculators (Brunsviga type) because of lacking electro-mechanical machines for teaching.

While finishing his studies in physics Sint became a scholar of the Institute of Advanced Studies (Institut für Höhere Studien, IHS) in Vienna and ended up rather by chance in the Sociology Department. (A planned formal science department had not been realized.) There he learned, besides the basics of sociology, to handle card counting machines, especially the IBM Electronic Statistical Machine Type 101. With such machines one could not only count but also perform simple calculations. Only much later he learned from a historical article [72] that and how it could have been used even to invert matrices.

The punching machines did not produce printouts on the cards. Hence users had to learn the encoding. This was simple for numeric codes but more demanding for alphabetic characters. Sorting of the cards by alphabet or numerically was done sequentially starting from the last digit or character in the field. Sorted output card stacks were stapled and resorted for the next digit respectively character. This procedure made it also possible to perform multiplications of punched multi-digit numbers by "progressive digiting". Statistical machines had been popular in Vienna since the late 19th century. Programming was carried out on plugboard tablets – an invention of the Austrian O. Schäffler [75] – based on telephone switchboard technology [54]. This technology was adopted in the census of the Austro-Hungarian Monarchy in 1890 (at the same time also in the USA). Schäffler later sold his patents [88] to Hollerith's Tabulating Machine Company (which ended up in International Business Machines, IBM). By some tricky programming of the boards sorting was possible by two columns (i.e. characters) of the card in one run. Not knowing this, the famous economist F. Machlup destroyed one of Sint's nearly finished sortings while explaining to him the "proper" way of doing the job (with appropriate excuses afterwards).

3 The institutional environment in which COMPSTAT was born

During these early years so-called statistical machines were also used in sociology. A front runner of the use of formal mathematical methods in social sciences was the Institute for Advanced Studies. It was founded with essential financial help from the US Ford Foundation (hence locally known as "Ford Institute"). The famous sociologist P. Lazarsfeld, founder of the Institute for Applied Social Psychology at the University of Vienna in 1929, and its director until his emigration to the USA in 1933, later professor at Columbia University and O. Morgenstern (together with J. von Neumann), the father of game theory and a former director of the Austrian Institute of Trade Cycle Research, were the driving forces behind the foundation of the Ford Institute [39]. At that time formal-mathematical as well as empirical methods were practically absent from the syllabus of economics and sociology in most academic institutions in Austria.

S. Sagoroff was a key person during the foundation of the Ford Institute and also its first director. He had already an interesting personal history: After receiving his doctor degree from the University of Leipzig (Germany) and studying in the USA under the supervision of J. A. Schumpeter 1933/34 on a Rockefeller grant, he became professor of statistics, president of the statistical office, and director of the Rockefeller Institute for Economic Research in Bulgaria before World War II. Later he was Bulgarian Royal Ambassador to Germany in Berlin until 1942 (when Bulgaria joined the Allies). In that function he was involved in the delay of the delivery of Bulgarian Jews. While in Berlin and with a broad interest in science he had befriended with some of Germany's intellectual elite, including a number of Nobel laureates who cherished his dinner parties. After liberation from his internment in Bavaria he had worked for the US Ambassador R. D. Murphy and had spent some time at Stanford University, before becoming professor of statistics at the University of Vienna.

Sagoroff was certainly an able organizer for the start-up of IHS but might not have been the best choice for running the institution in a way ensuring high scientific standards. Still, the Ford Institute was a tremendous place to learn and to get acquainted with current thoughts in social and economic sciences, offering contacts to researchers of high reputation. In the following decade IHS played an important role in the reversal of the former situation at Viennese academic institutions, advocating mathematical and statistical approaches.

Sagoroff's USA experience had also been crucial to the fact that he was successful in receiving a Rockefeller grant for the University of Vienna to buy a digital computer. The foundation paid for half of the price (83.500 US$) and the computer company gave an educational grant covering the other half. The university had to pay just for transportation and installation. That Sagoroff was interested in computers and on the lookout for one was most likely fueled by the fact that at the very time H. Zemanek was constructing the first transistorized computer in Europe at the Technische Hochschule (now University of Technology) in Vienna. At that same time Sagoroff's assistant at the Statistics Department, A. Adam, also tried to build a simple electronic statistical calculator and obtained also a patent on this device. But he definitely had not the technical expertise of Zemanek and his machine was never used in practice. Nevertheless his historical findings on the early history of computing remain a landmark in the historiography of the area ([18] widely distributed during the 1973 ISI Session in Vienna).

The arrival of the first "electronic brain" in Vienna in 1960 was not only of interest for the scientific community but meant also a major event for the Austrian media. The electronic tube-based machine needed a special powerful electricity generator to convert the 50 Hz alternating current in Austria to the 60 Hz used in the USA. It was installed in the cellar of the new university annex building. The windows of the computer room had to be equipped with specially coated glass to ensure constant temperature.

This Datatron 205 was a one address machine with one command (or address) and two calculating registers. The machine owned a drum storage with 4000 cells. Each cell held 10 binary-decimal digits (each digit was represented by 4 bits and the uppermost values beyond 0-9 were not used). The 11th digit was used for signs and as a modifier in some commands. The 4000 cells were divided in 40 cylinders on the drum each containing 100 words with an average access time (half turn of the drum) in the millisecond domain. It possessed a feature later reinvented by IBM and marketed in a more elaborate form under the name virtual memory: two cylinders could accept repeated identical runs of 20 words (commands) which reduced access time to one fifth. The critical parts of the program code were shifted to this 'fast storage' with one block command and the program execution shifted (often simultaneously) to the first command in this storage which meant it was transferred into command register A.

The implementation was in digital code: Each command was a two digit number acting on one address, for instance the command "64" imported a number into register A:

0000641234 Import the content of cell 1234 (on the drum)
into calculating register A
While 74
0000741235 Add the content of cell 1235
to the content of the register A

60 stood for multiplication, 61 for division. Other arithmetic operations, floating point operations, shift operations, logical operations, conditional jumps, printing of registers were performed similarly. An additional register could be used independently or to enlarge the number of digits in register A. 02 stored results back to the drum. 08 stopped the run.

In principle there existed an assembler with mnemonic alphabetic codes, however, there was no tape punching device to enter alphabetic characters. Because one had to know the digit codes for operating the machine (entering and changing commands bit by bit only guided by a display of the registers on the console) the direct way was definitely faster. As one could actually see each bit stored in the registers during programming and debugging one could also spot a malfunctioning hardware unit if one of the bits did not show up properly. In this case one had to open the machine and take out the concerned unit (a flip flop with four tubes). Usually it was easy to spot the culprit by visual inspection or alternatively exchanging the tubes one by one. Only the (preliminary) finished program was printed or punched out on a paper tape. As space was scarce and each letter had to be encoded by two decimal digits, comments accompanying the results were kept to a minimum.

The arrival of this computer was essential to the fact that the Statistics Department became the hub of computing inside the University of Vienna.

Sint's first experiences with real computing in the early nineteen sixties are connected to a programming course for digital computers held by the mathematician J. Roppert, assistant in the Department of Statistics. As one of the few who took an exam in computer programming and as a scholar of IHS, Sint was offered an assistantship at this department. His statistical qualifications were elementary probability theory (not based on measure theory) and some statistics for sociologists. (The type of statistics used in quantum physics were not of much help in a statistics department.) At the IHS he also obtained a first training in game theory from O. Morgenstern. Later, while spending a year in Oxford, he learned more statistics and got interested in cluster analysis. This contact with English statistics helped him doing "real" statistics in the following.

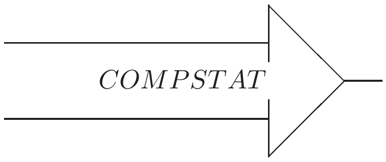
Before Sint could use the new generation of computers (an IBM /360-44 was installed at the University of Vienna in 1968) he had to learn his first programming language, Fortran. For W. Winkler, a professor emeritus of statistics, he wrote his first Fortran program for the calculation of a Lexis-type population distribution on an off-site computer. When he had finished Winkler remarked that it would have been much faster to do the job on a mechanical calculator. At that time correcting card decks and working on a remote machine was extremely time consuming.

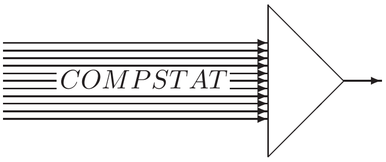
About that time IBM had started developing and distributing statistical software. Most developments were open source Fortran code. Naturally Fortran was a large step forward going along with third generation digital computers. Programme codes for algorithms were published by the US Association of Computing Machinery (ACM). About that time also the first commercial packages arrived. In statistics one could choose between OSIRIS, BMD, P-STAT, and SPSS. The Department of Statistics at the University of Vienna decided for SPSS in December 1973. SPSS, like BMD and P-STAT was implemented in Fortran, offering high portability. All the implementations of statistical methods at the department were programmed in Fortran, not a user-friendly environment from a today's perspective. This included the first administrative program for the enrollment of students and production of corresponding statistics.

4 The first symposium and the Compstat Society

Access to elaborate algorithms on computers increased the awareness of more recent methodological developments in statistics, primarily in the Anglo-American world. In the Statistics Department at the University of Vienna, with its tradition in conventional economic and demographic statistics, the younger members tried hard to establish contacts with the international statistical community. Not having had access to sufficient travel funds, Sint and his colleague J. Gordesch, a trained mathematician, encouraged by A. Liebing, the publisher of the journal Metrika, envisioned a conference on an up-to-date statistical topic in Vienna. Sint was interested in cluster analysis and Gordesch rather in computational probability and model building. These and other topics were ventilated until one settled on a conference on computers and statistics. As for the name in English they took the Journal of the Royal Statistical Society as a model: it comprised series A for *Theoretical Statistics* and series B for *Applied Statistics*. Thus they assumed *Symposium on Computational Statistics* would be a proper name. Sint came up with the acronym *COMPSTAT* arguing that one needs a short name which would still be near to an understandable expression to be easily remembered (this is what is called a logo now).

For the first call for papers the word *COMPSTAT* was embedded in an arrow like graph derived from the symbols used in analog computing: several input lines ending in a triangle (the statistical engines or algorithms). The condensed final result we are still using is displayed in the left figure. Sint and his colleagues were thinking about statistical methods (they were the hub of our ideas about the conference) as means of compressing a large number of inputs in a few meaningful results and *COMPSTAT* as an input to improve the algorithms (being quite aware of the recursivity of these processes).





The original design idea was rather something like the right figure. A sketched drawing similar to this one (without the small arrows and with a smaller number of input lines) had been dropped by the graphics designer of the publisher.

As we know now this was the first freely accessible international conference with an open call for papers in this area. The first COMPSTAT meeting was announced in the American Statistician (attracting some participants from the USA) which helped later to defend the right of name in that country. The only preceding international conference of that kind was organized and financed by IBM. Preceding were also the at first rather local North American Interface symposia starting in Southern California in 1967, sponsored by the local chapters of both the American Statistical Association and the ACM, obtaining an international flavor as late as 1979 (twelfth Interface symposium held at the University of Waterloo, Ontario, Canada). For the Interface Foundation of North America, Inc., and its history see `http://www.galaxy.gmu.edu/stats/IFNA.html`.

Any organizer of a new kind of conference is uncertain about its success and the number of participants he/she might attract. According to the preface of the proceedings [1], Sint and Gordesch were not sure whether "mathematicians specialized in probability theory or statistics, or experts in electronic data processing would look at computational statistics as a serious subject". As the deadline of the call for papers came nearer the organizers became increasingly anxious and started to muster locals for participation. Fortunately, in the first few days after the deadline had expired, a reasonable number of additional abstracts appeared, all together enough to give them peace of mind.

In 1972 Sint had attended a conference where the proceedings papers had to be retyped by clerical staff which turned out to be a disaster. This experience in mind it was decided to ask for camera-ready copies. For the COMPSTAT proceedings it worked out smoothly and the copies could be distributed during the symposium, a practice that has survived till now. The formal invitation to the conference was signed by G. Bruckmann and L. Schmetterer, both professors of statistics at the department, because the young colleagues hoped that the appearance of internationally known personalities would be more acceptable to participants and to the potential buyers of the proceedings (Sint and Gordesch just signed the preface; F. Ferschl was added as an editor by the publisher).

Gordesch had at the time of the conference already left Vienna, and Sint had moved to the Austrian Academy of Sciences. Thus, although the latter was still around (his new boss was Schmetterer, the successor of Sagoroff as professor of statistics), a lot of the preparatory work had to be done by the young colleagues W. Grossmann, G. Pflug, and W. Schimanovich. M.G. Schimek, a first-year student of statistics and informatics in 1974, learning Fortran and SPSS at that time, was a keen observer of all these activities going on in the Department of Statistics and Informatics at the University of Vienna.

The interest of Gordesch in COMPSTAT had remained awake and so the next conference was naturally held in Berlin. From that time onwards it has never been a problem to find places to go. Someone has always been willing to organize the symposium.

To have a permanent platform a *Compstat Society* was created in 1976. Membership was by invitation only. Mainly organizers and chair persons of the first conferences were approached. Sint recalls that only selected members were asked (no formal board decision) when COMPSTAT was transferred to the International Association for Statistical Computing (IASC) in 1978. It was an initiative of N. Victor (1991–1993 IASC President). Readers interested in the history of the IASC are referred to the *Statistical Software Newsletter*, edited for almost three decades by A. Hörmann, and since 1990 integrated as special section into the official journal of the IASC *Computational Statistics and Data Analysis*. Furthermore we want to mention P. Dirschedl and R. Ostermann, (1994 [32]) as a valuable reference for developments in computational statistics (including IASC activities in Germany, the history of the legendary *Reisensburg Meetings* and of the Statistical Software Newsletter).

Formally the Compstat Society was dissolved by the Austrian Registration Office due to inactivity. Numerous members reappeared in the newly founded European Regional Chapter (now European Section) of the IASC. The main stumbling block in the transfer was Physica-Verlag and its owner A. Liebing. He had contributed a lot to the planning of the first symposium to make it a success and was then afraid that, if the conference is taken over by a large organization, other publishers would get interested and grab the proceedings and the then started COMPSTAT Lectures (a series of books apart from the proceedings). The result of the heated discussions during COMPSTAT 1978 in Leiden was a most favourable treatment clause which gave Liebing an advantage over competitors. This worked out satisfactorily until he sold Physica-Verlag to the Springer company because of his retirement as a publisher.

Sint's continued active involvement ceased after 20 years at the second COMPSTAT symposium that took place in Vienna, organized by R. Dutter (University of Technology, Vienna) and W. Grossmann. The 1994 anniversary was also marked by a *COMPSTAT Satellite Meeting on Smoothing* – smoothing having been a hot topic at that time – held in the famous alpine spa Semmering (on the border between Lower Austria and Styria), bringing additional audiences mainly from outside Europe to COMPSTAT. It was organized by M. G. Schimek (Karl-Franzens-University, Graz; currently IASC Vice President). The COMPSTAT baby had become off age and a new generation was following the tradition of P. P. Sint.

5 Some remarks on the development of computational statistics

The idea of COMPSTAT was borne at the University of Vienna in an environment typical for statistics departments in continental Europe at that time against the background of new computer technology, rather specific with respect to statistical methodology. In order to obtain a more detailed picture of the role of COMPSTAT we need to sketch some important issues in the development of computational statistics in connection with other topics. Starting point for our considerations is the following working definition of the term *Computational Statistics*, which is according to a statement of N. Victor in 1986 (cf. Antoni et al., 1986 [7], p. vi) ".....not an independent science but rather an important area of statistics and indispensable tool for the statistician". This statement is made more precise in a definition proposed by A. Westlake (cf. Lauro, 1996 [61]): "Computational statistics is related to the advance of statistical theory and methods through the use of computational methods. This includes both the use of computation to explore the impact of theories and methods, and development of algorithms to make these ideas available to users". This definition gives on the one hand a clarification of the term “area of statistics” in Victor’s statement, on the other hand it emphasizes also the instrumental aspect of statistical methods with repect to their application.

Starting from this definition it is quite clear that we have to consider the progress of computational statistics in connection with developments in statistical theory, developments in computation and algorithms, developments in computer science, and last but not least developments in the application of statistics. In many ways there has always been an exchange of ideas, important for the understanding of computational statistics, stemming from these four areas. In the following we sketch some of these ideas and discuss their interplay.

5.1 Computational statistics and statistical theory

According to B. Efron in 2002 [36] the development of statistics in general can be divided into a *theory area* and a *methodology area*. Efron illustrates the theory area as a journey from applications towards the mathematical formulation of statistical ideas. According to him it all starts around 1900 with the work of K. Pearson and goes on to the contributions of J. Neyman and Sir R. Fisher, finally approaching the decision-theoretic framework for statistical procedures due to A. Wald. A key feature in this development is the foundation of statistical theory on optimality principles. This decision-theoretic framework is capable of bolstering statistical methods by a sound mathematical theory, provided that the problems are stated in precise mathematical form by a number of assumptions. In that sense the theoretical background is a prerequisite for the application of statistics and for the computations in connection with the statistical models. Obviously computation meant in early times paper and pencil calculations or using rather simple (mechanical) computing devices.

To some extent the early investigations were oriented more towards the analysis of mathematical properties of procedures and less towards the analysis of data. A milestone in the shift from the theory area towards the methodology area was the paper of J. W. Tukey in 1962 [83] about the future of data analysis. It emphasizes a number of important aspects, in particular the distinction between confirmatory and explanatory analysis, the iterative and dynamic nature of data analysis, the importance of robustness, and the use of graphical techniques for data analysis. In this paper Tukey is not so enthusiastic about the computer with respect to data analysis. He states that the computer is in many instances “important but not vital”, in others “vital”. However due to the technological development the computer has definitely become more important for the methodology area than one could foresee 40 years ago.

In fact, the methodology area is in many aspects characterized by a strong interplay between statistics and computing, ranging from the implementation of procedures over the definition of new types of models up to the discovery of new aspects of statistical theory. A typical example is Bayesian data analysis, the progress of which has been driven to a considerable extent by new computational techniques (cf. Gelman et al., 1996 [44]). High computing power is needed for these methods, hence they are often summarized under the heading *computer intensive methods*. Another interesting feature of many of these developments is the fact that optimality principles are not necessarily applied in a closed form by defining one objective function in advance, but rather by outlining a number of optimization problems in an iterative and more dynamic way than in traditional statistics. This iterative process is rather statistical in nature compared to the iterative numerical solutions of nonlinear equations. Hence, from a statistical (data analytic) point of view one is sometimes not solely interested in the final solution but also in the behaviour of the algorithm.

In many instances theoretical insight into methods and the development of models go hand in hand with the implementation of these methods respectively models. In the following we list (in alphabetical order) a number of key developments that have resulted in standard approaches of applied statistics (together with early references): Bootstrap Methods (Efron 1979 [35]), EM-Algorithm (Dempster, Laird and Rubin, 1977 [30]), Exploratory Data Analysis (EDA; Tukey, 1970 [84]), Generalized Additive Models (GAM; Buja, Hastie and Tibshirani, 1989 [22], Hastie and Tibshirani, 1990 [50]), Generalized Linear Models (GLM; Nelder and Wedderburn, 1972 [70]), Graphical Models (Lauritzen and Wermuth, 1989 [60]), Markov Chain Monte Carlo (MCMC) – in particular Gibbs Sampling – (Hastings, 1970 [52], Geman, 1984 [45]), Nonparametric Regression (Stone, 1977 [77]), Projection Pursuit (Fisherkeller et al., 1974 [38], Friedman and Tukey, 1974 [43]), Proportional Hazard Models (Cox, 1972 [28]), Robust Statistics (Huber, 1964 [56]), and Tree Based Methods (Breiman et al., 1982 [21]). Besides these developments inside statistics we wish to point out that new aspects of statistical data analysis have in addition occurred in connection with Data Mining (Frawley et al., 1992 [41]), recently explored from a statistical learning perspective by T. Hastie, R. Tibshirani and J. Friedman (2001 [51]).

Apart from these examples that are all characterized by a strong interplay between statistical theory and computational statistics in the sense of Westlake's definition, it should be noted that there are also methods which had been formulated long before they were feasible to compute. An interesting example with respect to the interplay between theory and computation are rank procedures. According to R. A. Thisted (1988 [80]) the motivation of F. Wilcoxon for defining his rank test was the fact that for moderate sample sizes calculation of the rank sum by hand is easier than calculation of the sum and the variance. However, the situation is completely changed in case of large sample sizes and machine calculation. Other examples of theoretical models introduced long before it was feasible to numerically evaluate them are conditional inference for logistic regression as formulated by Sir D. J. Cox

(Mehta and Patel, 1992 [62]) or the empirical Bayes approach of H. Robbins (1956 [74]) that nowadays sees interesting applications in microarray analysis (Efron, 2003 [37]).

Besides these new developments in statistical theory, the advance of computers has also influenced other areas of statistical theory in the sense of providing tools for experimental checking of statistical models under various scenarios. Such types of computer experiments are of interest even in cases where the methods are well underpinned from a theoretical point of view. A well known early example is the Princeton study on robust statistics (Andrews et al., 1972 [20]). Today in theoretical investigations it is rather common to support the results by simulation and graphical displays. In this context one should know that according to H. H. Goldstine (1972 [48]) such computer experiments were already envisioned by J. von Neuman and S. Ulam in 1945 at the very beginning of digital computing. This led to the development of simulation languages, rather independently of conventional statistics, but with an important impact on computer science (see also [65]). Note that Simula was the first object-oriented language ever (Dahl and Nygaard, 1966 [29]). A good overview of simulation from a statistical perspective can be found in B. Ripley's book of 1987 [73].

5.2 Computational statistics and algorithms

Computation in statistics is based on algorithms which have their origin either in numerical mathematics or in computer science. Such methods are summarized under the topic *statistical computing*. Usually textbooks emphasize the numerical aspects (for instance Monahan, 2001 [67]). However in the following we want to review briefly some important developments in numerical mathematics as well as in computer science.

For mainstream statistics the most important area is numerical analysis. The core topics are numerical linear algebra and optimization techniques but practically all areas of modern numerical analysis may be useful. Approximation techniques applying specific classes of functions, for examples splines or wavelets, play an important role in smoothing. Numerical integration is essential for the calculation of probability distributions, and for time series analysis Fourier transforms are of utmost importance (note that the fast Fourier transform, which is one of the most important algorithms of numerical analysis, was invented by J. Tukey in connection with statistical problems (Tukey and Cooley, 1965 [85])). Recursive algorithms and filtering are traditionally linked to time series but recently these methods are also of interest in connection with data streams [86]. However it seems that statisticians apply these methods often more like a tool from the shelf. New innovative aspects occur on the one hand in the theoretical analysis of algorithms in the context of statistical models, on the other hand as adaptation of methods according to statistical needs, which is in fact one of the key issues in computational statistics. The organization of the recent textbook by J. E. Gentle (2002 [46]) is a good example.

Another core topic is generation of random numbers which is conceptually close to computational statistics and computational probability theory and is the basic technique for discrete event simulation. Most early applications concerned the generation of random variates of different distributions for sampling as well as for numerical integration. Nowadays this technique is fundamental to many new developments in statistics like bootstrap methods, Bayesian computation, or multiple imputation techniques. However, also in this field it seems that statisticians are mainly interested in using these technique for their own purposes, in particular the theory of uniform random number generation is traditionally rather linked to number theory and computer science. An important contribution inside statistics is the quite exhaustive monograph on the generation of non-uniform random variates by L. Devroye (1986 [31]).

Apart from numerical analysis, there are algorithms of statistical interest for sorting, searching and combinatorial problems sometimes summarized under the heading *semi-numerical algorithms*. They are of utmost importance for exact nonparametric test procedures and for exact logistic regression as implemented in StatXact and LogXact (see for example [63] and [64]). Combinatorial algorithms are also used in the context of experimental design.

There is another group of algorithms highly relevant for computational statistics. Their origin is mainly in computer science, in particular we are thinking of machine learning, artificial intelligence (AI), and knowledge discovery in data bases. Neural Networks, Genetic Algorithms, Decision Trees, Belief Networks or Boosting are important and actual examples. These developments have given rise to a new research area on the borderline between statistics and computer science. New challenges arise from the need to interpret these non-statistical approaches in a statistical framework. In addition to [51], papers by D. Hand (1996 [49]), and R. Coppi (2002 [27]) discuss some of these issues.

All the above mentioned computational topics cover methods that are also adopted in other areas of mathematical modelling. If one looks into a book of mathematical modelling one might find similar algorithms and techniques as in a textbook about computational statistics. For example the book of N. Gershenfeld (1999 [47]) distinguishes between Analytical Models, Numerical Models and Observational Models. Analytical Models (mainly difference and differential equations) occur also in statistical applications, in particular in finance and epidemiology, but, as Sir D. J. Cox had stated in the preface to COMPSTAT 1992 [10], these topics are not core topics in computational statistics. It seems that the situation has not changed since. Obviously, in the area of Observational Models there is large overlap with methods used in statistical modelling but the focus is a different one. This had already been noticed in the early days of computational statistics by Sir J. A. Nelder (1978 [69]) who identified the following peculiarities of computing in statistics compared to other areas: (i) *Complex data structures*:

Problems analyzed by statisticians have often a rather complex data structure and adaptation of this structure towards the requirements of an algorithmic procedure is many times a genuine statistical task; (ii) *Exploratory nature of statistical analysis*: Usually in a statistical analysis we have not only a pure algorithmic cycle (defined by: get data, do algorithm, put results, stop) but rather a cycle of different computations, which are to some extent defined according to the interpretation of the previous results; (iii) *Competence of users*: Users of statistical methods are not necessarily experts in the area of statistics or in the area of numerical mathematics, but experts in a domain and want to interpret their methods according to their domain knowledge.

With these specific points in mind it is not surprising that graphical computation plays a more prominent role in statistics than in other areas of modelling. J. Tukey is one of the statistical pioneers, in particular with respect to dynamic graphics (Friedman and Stuetzle, 2002 [42]). Statistics has contributed to the development of graphical computation complementary to computer science. L. Wilkinson et al. (2000 [87]) stress the following three key ideas in the progression of statistical graphics, which may be seen as main driving factors behind most genuine statistical innovations: (i) Graphics are not only a tool for displaying results but rather a tool for perceiving statistical relationships directly; (ii) Dynamic interactive graphics are an important tool for data analysis, and (iii) Graphics are a means of model formalization reflecting quantitative and qualitative traits of its variables.

5.3 Computational statistics and computer science

Due to the specific needs of statistical data analysis mentioned in the previous section it was quite natural that even in the early days of computers statisticians were interested in developing specific software tools tailored more towards their needs than mathematical subroutine libraries like NAGLIB or IMSL. As early as 1965 Sir J. A. Nelder started with the development of GENSTAT in Adelaide (South Australia) on a CDC 3000 computer (Nelder, 1974 [68]). The data structure was at that time the data matrix, but in the further developments at Rothamsted Experimental Station (UK) the design was changed towards increasingly statistics-oriented data structures like variates, vectors, matrices or tables with main emphasis on the variate as well as the development of a statistical language. Around the same time also other projects had been started that resulted in major packages: BMD (later BMDP) was developed by W. J. Dixon and M.B. Brown from 1964 onwards at the University of California at Los Angeles as a coherent combination of different analysis subroutines with a common control language (first manual in 1972 [33]). SAS was designed by J. Goodnight and A. J. Barr starting in 1966 (the commercial SAS Institute was founded in 1976 by J. Goodnight, J. Sall, A. Barr and J. Helwigand;

`http://www.sas.com/presscenter/bgndr_history.html`,
`http://www.theexaminer.biz/Software/Goodnight.htm`).

Finally in 1967 N. H. Nie, C. H. Hull and D. H. Bent commenced at the University of Stanford the SPSS project

(`http://www.spss.com/corpinfo/history.htm`).

The latter two packages still flourish as products of service companies.

Many other statistical packages were designed in the subsequent years with the aim of supporting data manipulation and statistical computing. The major developers tried to keep track of the progress made in computing infrastructure in order to improve their products with respect to data storage and data management and to offer numerically more reliable statistical analysis methods. The book of I. Francis (1981 [40]) provides an overview over this early period of statistical software. It describes more than 100 packages available at the beginning of the nineteen eighties. The scope of these programs ranged from data management systems and survey programs to general purpose statistical programs and programs for specific analysis tasks. With respect to programming Fortran was the dominant source language and most of the products were offered for different hardware configurations and operating systems. Today for a number of reasons most of these products are only of historical interest. For specific purpose packages at the forefront of statistical methodology it was difficult to keep their competitive advantage after its methods had become widespread. For other products it was nothing but easy to keep path in their program design with the fast progress of computer technology. Only the major producers were able to follow the developments which also meant a switch from Fortran to other languages like C or C++, an adaptation to new computer architectures, and integration of modern user interfaces as well as of graphic facilities into their packages. Their new orientation towards customized analysis procedures made these products increasingly attractive for statisticians as well as non-statisticians.

More important for computational statistics were other developments aiming at the design of statistical languages as basis for statistical programming environments. Based on the conceptual formulation of the Generalized Linear Model, GLIM seems to have been the first system that was oriented towards the definition of an interactive analytical language for a large class of statistical problems in a unified manner, taking advantage of the previous GENSTAT experiences. The most important step in this direction was the S language, a project starting in 1976. The goal was the definition of a programming language for the support of data analysis processes
(`http://cm.bell-labs.com/cm/ms/departments/sia/S/history.html`).

The computer science oriented concepts of the S language are best described in the so called “green” S book by J. Chambers (1998 [23]). For the statistical aspects we refer to the “white” S book of J. Chambers and T. Hastie (1992 [24]). The general approach, a clever combination of functional programming and object oriented programming, supports perfectly the iterative nature of the statistical data analysis process and forms a new paradigm for computing, which is independent of the statistical application. The ACM honored this contribution to computer science: In 1998 Chambers received the ACM Software System Award for his seminal work which “has forever altered the way people analyze, visualize and manipulate data” [17].

In 1992 based on the S language, R. Ihaka and R. Gentlemen started the R-project at the University of Auckland (New Zealand; cf. Gentleman and Ihaka, 1996, [59] for the early history of R). Due to free availability the R-community grew rather fast and in 1996 the Comprehensive R Archive Network (CRAN) was established at the University of Technology in Vienna (cf. Hornik and Leisch, 2002, [55] for recent developments). A further important step in the development of statistical environments, closely related to R, was the formation of the Omegahat-project (`http://www.omegahat.org/`) for statistical computing in 1998. It serves as an umbrella for a number of other recent open source projects. Its goal, as described in detail by D. Temple Lang [79], is to meet the challenges for statistical computing resulting from new developments in computer science like distributed computing or Web-based services. Examples are extensions of existing systems such as StatDataML (Meyer et al., 2002 [66]) offering a XML interface for data exchange or embedding R into a spreadsheet environment (Neuwirth and Baier, 2000 [71]).

Besides S and R there were a number of other important projects in the area of statistical software development. For instance we want to mention W. Härdle’s XploRe [53], an interactive statistical computing environment, realizing new concepts of nonparametric curve and density estimation as well as statistical graphics in the mid nineteen eighties. In connection with XploRe recent efforts to extend its scope to statistical teaching and to Web applications are worth mentioning. Another project of interest due to L. Tierney in the late nineteen eighties was XLISP-STAT ([81], [82]), a statistical environment based on the public X-LISP language freely available from the *statlib archive*.

A further line of development are efforts to use parallel architectures in statistical computing. Such computer architectures are typically used for the implementation of demanding numerical algorithms. In recent years computer science has widened the scope of parallel computing towards distributed computing. We expect this research area to grow quite rapidly in the future, with an impact on statistical computing.

Another statistically relevant area of computer science is data management. While data structures in statistical computing are usually closely related to formal specifications of data types (e.g. lists, vectors, or matrices), the interpretation of an analysis process makes often use of conceptual and relational structures. Traditionally this topic is treated in the theory of data bases. A major breakthrough in this area was the introduction of the relational data model by E. F. Codd (1970 [26]). It offeres the opportunity to describe complex real world problems from a conceptual point of view in a unified manner. The description of data by data models is nowadays captured under the heading metadata. In this context it is worth mentioning that the term metadata occurred for the first time in connection with official statistics in a book by B. Sundgren (1975 [78]). Modern data base systems offer not only tools for storage and retrieval, but also statistical functionalities, in particular for tabulation (core instruments for official statistics). Despite the fact that they are rather simple with respect to statistical methodology, there are numerous pitfalls from a conceptual point of view. The latter raise interesting operational questions which are treated in the context of data warehouses. An interesting reference which helps to understand the connections as well as the differences between the statistical approach and the computer science approach to multi-dimensional tables is [76].

5.4 Computational statistics and applications

With respect to the interplay between applications and computational statistics we want to discuss now the challenges that arise from application problems. Besides the difficulties resulting from new problems in various research areas, for example analysis of microarrays in biology, one can identify – rather independently from the field of research – the following three interwoven challenges for computational statistics: handling of problems stemming from new data capture techniques, from the complexity of data structures, and from the size of data.

Since the early times of computational statistics a major effort has been the development of tools for automatic data capture and of interfaces to data management systems. This has led to the development of computer-aided survey information collection (CASIC) tools, an area which seems to be nowadays more a topic in official statistics and management of statistical data base systems. Inside computational statistics we observe an increasing interest in the handling of efficient data generation systems. Many times such systems occur in connection with automated monitoring of networks, in particular the Internet. Such data streams are of interest from a computer science as well as a statistical point of view. The statistical perspective is treated in a number of papers in a recent issue of the *Journal of Computational and Graphical Statistics* (e.g. Wegman and Marchette, 2003 [86].

With respect to the data structures the traditional model was characterized by the relation between sample and universe or by a properly designed measurement process. Such data structures can be represented quite well in a relational scheme and appropriate statistical models can be formulated for the analysis, for instance hierarchical models. In connection with data mining applications statisticians are confronted with new data structures which do not fit into the standard model. One has to analyze data combined from different sources which are often rather inhomogeneous with respect to quality (e.g. problem of missing values) and have no immediate interpretation in a traditional statistical framework. Combination of such data sources is a statistical problem in its own right.

The last difficulty is the size of the data. P. J. Huber (1994 [57]) classified data sets from tiny (about 100 bytes) up to huge (about 10^{10} bytes). One can definitely argue that size is always an issue relative to computing power and storage capacity, and problems practically intractable 30 years ago are nowadays routine applications. Nevertheless, today's statisticians and computer scientists have to solve problems for huge datasets. Specific problems concerning the data structure, the data base management, and the computational complexity are discussed in Huber (1999 [58]).

A second important topic for computational statistics with respect to applications is the statistical analysis process itself. The ubiquitous availability of the computer and of statistical software packages has changed the context in many ways. On the one hand statistical software packages support statisticians in the phase of exploratory data analysis and allow them the evaluation of numerous tentative models for the data without careful planning in advance. On the other hand they enable non-statisticians to perform rather complex analyses for their data, in former times solely carried out by professional statisticians. This evolution has weakened in some sense the role of statisticians as custodians of the data and has caused many discussions inside the statistical profession. Here we only want to mention Y. Dodge and J. Whittaker (1992 [34]) who raised the point that this development might bring about a de-skilling of certain parts of the profession. However they also argued that the democratization of facilities does not automatically mean a threat to the profession in the long run. We claim that statistical analysis is definitely more than the application of certain algorithms because an analysis strategy is required too. For instance in the current scientific development of the bio-sciences we see an explosion of highly complex data problems that can only be managed in part with the resources at hand.

In the nineteen eighties the question of automated analysis strategies was intensively discussed in connection with the issue of statistical expert systems. This undertaking ended without substantial success making it clear that it is rather implausible to assume statisticians can be easily substituted by machines in the near future. To put it in a nutshell, not even standard data-analytic problems can be handled easily via routine applications and simple rule systems. Another area of interest in this context is certainly the role of computers in statistical education, in particular for non-professionals, taking advantage of the various opportunities offered in the field of computational statistics.

6 The COMPSTAT symposia

In this section we review the COMPSTAT symposia, giving a tabulated summary of the occurrence of topics and a verbal description of the meetings and proceedings.

As for the summary of topics covered in the COMPSTAT symposia we have produced two self-explaining tables, Table 1 for the period 1974–1988, and Table 2 for the period 1990–2002. The notation in these tables is the following: "p" denotes that a topic was present in the proceedings, "f" denotes that a topic was frequently present in the proceedings (i.e. more than 3 times), "K" represents a keynote paper, "I" represents one or two invited papers, and finally "T" signifies a tutorial. We suggest to read the respective table in parallel with the verbal description of the chronologically ordered COMPSTAT symposia.

The very first COMPSTAT symposium was held at the University of Vienna in 1974, initiated by P. P. Sint and J. Gordesch. Both were also in fact the editors of the proceedings [1]. There were about 50 presentations organized according to five subject areas, reflecting to some extent the interests of the organizers: Computational Probability, Automatic Classification, Numerical and Algorithmic Aspects of Statistical Computing, Simulation and Stochastic Processes, and last but not least Software Packages. In 1974 there were neither formal keynotes nor invited lectures. However, during the opening session a special lecture was delivered by the well-know mathematical statistician L. Schmetterer on stochastic approximation (not in the proceedings).

Naturally the topics within the subject areas were rather scattered, but some of them remained popular across the whole period of 30 years such as Robustness (note that P. J. Huber was present at the first symposium), Time Series Analysis, and Modelling (the latter in its beginning primarily meaning factor analysis and dimension reduction techniques). It is remarkable that a number of statistical packages popular at the time were already covered: R. Buhler's P-STAT and Sir J. A. Nelder's GENSTAT. The presentation of a SAS system, not to be confounded with the later much more successful namesake [25], should also be mentioned. Further, as in succeeding conferences, APL (for details see e.g. [19]) appeared as a popular statistical environment.

With all this in mind Gordesch and Sint speculated in the preface of [1] about a spectacular growth of the field, in writing "which as we hope will now result in techniques of model building being very different today from what it was in pre-computer days".

The second COMPSTAT symposium took place in Berlin 1976, organized by J. Gordesch and P. Naeve (also the editors of the volume [2]). Altogether 58 papers were presented. The subject areas were more or less the same as at the first meeting but the names had changed somewhat: Computational Probability, Automatic Classification and Multidimensional Scaling, Numerical and Algorithmic Aspects of Statistical Models (with subtopics Linear Models, Multivariate Analysis and Sampling), Simulation and Stochastic Processes, and finally Software. A new section "Applications" was introduced (mainly in economics and biology). This selection reflects the understanding of computational topics in the mid nineteen seventies: Multivariate Analysis comprised mainly ANOVA as well as Factor Analysis and Computational

Probability meant random number generators and the calculation of distributions in statistics. Apart from statistical computing Software also comprised recent developments in data bases. In addition there was a dedicated interest in the comparison of software packages with respect to certain technical as well as practical criteria.

The third COMPSTAT symposium in Leiden 1978 was organized by the Department of Medical Statistics in cooperation with the Computer Centre (both Leiden University) and headed by L. C. A. Corsten and J. Hermans. 68 papers were presented and published in the proceedings [3]. For the first time two keynotes were included, delivered by Sir J. A. Nelder and J. Tinbergen. The main topics consisted of Linear and Nonlinear Regression, Time Series, Discriminant Analysis, Contingency Tables, Cluster Analysis, Exploratory Techniques, Simulation and Optimization, Teaching Statistics, and Statistical Software. It is interesting to note that Exploratory Techniques was mainly an umbrella for problems in connection with multidimensional scaling. The topics Simulation and Optimization as well as Computational Probability also comprised contributions which would nowadays hardly find their way into a statistical meeting.

The fourth COMPSTAT symposium was organized by and held at the University of Edinburgh in 1980 with a record number of about 750 participants. Four invited and 82 (out of 250 submissions) contributed papers were presented and published in the proceedings volume [4], edited by M.M. Barrit and D. Wishart. This meeting clearly marks the beginning of the transition from batch to interactive computer processing, reflected in a special session. Invited lectures were given by J. Tukey on styles of data analysis, by E.M.L. Beale on branch and bound methods for optimization, by R. Tomassone on survey management of large data sets, and by I. Francis on a taxonomy of statistical software. Other topics were Sampling Methods, Data Base Management, Education, Analysis of Variance/Covariance, Interactive Computing, Linear and Nonlinear Regression, Multivariate Analysis, Optimization and Simulation, Cluster Analysis, Statistical Software, and Time Series Analysis.

The diffusion of interactive personal computing (marking the shift from mainframe to personal computers in the early nineteen eighties) can be clearly identified in COMPSTAT 1982 held in Toulouse (the fifth symposium) with about 500 participants. H. Caussinus, who also published the proceedings [5] together with P. Ettinger and R. Tomassone, chaired the program committee. One finds several new features at this COMPSTAT: the number of invited speakers was increased to 15 in order to cover new developments in computational statistics like Experimental Design, Computing Environments, Numerical Methods, EDA, Parallel Processing in Statistics, and Artificial Intelligence. In contrast to previous proceedings volumes comprising also papers at the border of statistics to other areas, the focus was now less theoretical and more computing oriented (60 papers out of 250 submissions).

COMPSTAT Symposium	74	76	78	80	82	84	86	88
Algorithms	f	f	f	p	p	p	fI	fI
Applications	f	f	p		p	p	fI	p
Bayes/MCMC/EM		p	p					
Categorical Data	p	p	f	p	p	f	pI	
Classification/Discrimination	p	p	f	p	p	fI	f	p
Cluster Analysis	f	f	f	f	p	f	fI	
Computational Probability	f	f	f			p	p	
Data Bases/Metadata	p				p	f	fI	p
Data Imput./Survey Design	p	f	p	fI		f		pI
Data Visualization/Graphics			p	p			pI	pT
Dimension Reduction	f	f	f	p	p	f	pI	pI
Experimental Design	p	p	p			pI	p	p
Expert Systems/AI					I	pI	fI	fIT
Exploratory Data Analysis			p	p	pI	p	p	p
Foundations/History		p	K	pI		I	I	
Graphical Models						p		pT
Handling of Huge Data			p					
Image Analysis			p					
Internet-based Methods								
MANOVA		p	p	p		p	f	p
Modelling/GLM/GAM	p	f	f	f	p	fI		
Neural Networks			p			p	f	
Numerics/Optimization	p	f	p	f	I	fI	f	
Parallel Computing					pI			K
Reliability and Survival		p	p					
Regression (linear/nonlinear)		p	f	p	p	f		pI
Resampling							p	fK
Robustness	p	p			p	fI	f	p
Simulations	f	p	f			f		
Smoothing/Curve Estimat.	p	f	p	f	p			pI
Spatial Statistics		p	p	p		p		
Statistical Software	f	f	fK	fI	pI	f	f	p
Stat. Learning/Data Mining			p					
Stochastic Systems	p	f	f	p	p			
Teaching Statistics			f	p		p	pI	
Time Series Analysis	f	p	p	f	pI	fI		p
Tree-based Methods				p				
Wavelets								

Table 1: Topics in the proceedings of the COMPSTAT symposia 1974–1988. p–present, f–frequent (p>3), K–Keynote, I–Invited, T–Tutorial.

COMPSTAT Symposium	90	92	94	96	98	00	02
Algorithms	p			f	fI	fI	f
Applications		fI	p	p	fI	K	f
Bayes/MCMC/EM		f	pK	f	fI	fI	f
Categorical Data	I						
Classification/Discrimination	f	fI	fI	fK	f	f	f
Cluster Analysis						p	p
Computational Probability		I	I	p			p
Data Bases/Metadata	pI	fI	pT				f
Data Imput./Survey Design		I			p	fKI	p
Data Visualization/Graphics	p			p	pI	p	pI
Dimension Reduction			I	p	p	p	
Experimental Design		f	f	f	p		p
Expert Systems/AI	fI	p					
Exploratory Data Analysis	p					p	
Foundations/History		pI		K			p
Graphical Models	p	fI	p	p	p		
Handling of Huge Data			K		p		p
Image Analysis		pI		pI			p
Internet-based Methods				I	p		fI
MANOVA	p				I	p	
Modelling/GLM/GAM	p		fI	p	pK	fI	p
Neural Networks			I	p			
Numerics/Optimization	pI	p	fI		p		
Parallel Computing			pI	p			
Reliability and Survival			p	f	p	pI	p
Regression (linear/nonlinear)	p	fI		fI	p	p	p
Resampling	f	f	p	p		p	f
Robustness	fI	fI	f	fI			p
Simulations		p		p	p	p	
Smoothing/Curve Estimat.		f	f	pI	fI	p	pI
Spatial Statistics	p		pI	p	p	fI	
Statistical Software	p	fI	f	p		f	fT
Stat. Learning/Data Mining		f	p	p	p	pI	fK
Stochastic Systems				pI	I		pI
Teaching Statistics			p	pK	pI	pI	fI
Time Series Analysis	fI	f	fI	fI	pI	fI	fI
Tree-based Methods				p	pI	p	p
Wavelets		p	pI	pI	K		p

Table 2: Topics in the proceedings of the COMPSTAT symposia 1990–2002. p–present, f–frequent (p>3), K–Keynote, I–Invited, T–Tutorial.

Many of them reflect the trends of the time, especially the penetration of personal computers and improved graphical displays into the world of statistics. The wish of statisticians to apply these new technologies, not yet covered by commercial software packages, can be clearly seen. Another novelty was the production of a complementary volume with short communications and posters.

The sixth symposium took place in Prague in 1984, extending the scope of COMPSTAT to the Eastern European countries. As a matter of fact IASC had planned for a meeting in Bratislava (a Slovakian town only 65 kilometers from Vienna) but the (communist) Czechoslovakian Academy of Sciences decided for the central location of Prague. Luckily there were several dedicated statisticians, among them T. Havranek, Z. Sidak and M. Novak, the organizers of the meeting. Many colleagues, who at that time did not have the chance to participate in Western meetings, could attend. Out of a record number of about 300 submissions 65 papers were selected. T. Havranek, Z. Sidak and M. Novak also edited the proceedings [6] and a companion volume of short communications and posters, following the example of 1982. Commemorating the tenth anniversary of the COMPSTAT symposia P.P. Sint was invited to deliver a lecture entitled “Roots in Computational Statistics”. The main topics covered in invited talks were Computational Statistics in Random Processes, Computational Aspects of Robustness, Discriminant Analysis, Statistical Expert Systems, Optimization Techniques, Linear Models, and Formal Computation in Statistics. Besides these topics also the traditional COMPSTAT themes like Cluster Analysis, Multivariate Analysis, Statistical Modelling and Software were present. It is worth mentioning that also a number of more computer science-oriented papers on data management and data preprocessing had found their way into the proceedings, reflecting some of the local interests.

COMPSTAT 1986 (the seventh symposium) was held in Rome and attracted an ever record of about 900 participants. From around 300 submissions for contributions about 60 contributed papers as well as 13 invited papers were published in the proceedings [7], edited by F. De Antoni, N. Lauro and A. Rizzi. A keynote lecture was given by E. B. Andersen about information, science and statistics, discussing the challenges for statistics resulting from the development of statistical software, graphics, interactive computing, and new methods and styles of data analysis. Apart from the invited program the proceedings volume presents itself well-balanced between statistically oriented themes, computer science oriented topics and novel applications. The main statistical themes comprised the traditional COMPSTAT topics like Probabilistic Models in Exploratory Data Analysis, Computational Approaches of Inference, Numerical Aspects of Statistical Computation, Cluster Analysis and Robustness, but also a rather specialized topic entitled Three Mode Data Matrices. The more computer science oriented topics reflect the trend towards Expert Systems and Artificial Intelligence, typical for the mid nineteen eighties. Altogether 9 papers on statistical expert systems were presented. Not so much in the mainstream of the time we identify sections on Computer Graphics, Data Representation, Statistical Software and Statistical Data Base Management. Main application areas were Clinical Trials and Econometric Computing. Additionally there was a section about Teaching Statistics.

Also for COMPSTAT 1988 (the eighth symposium), taking place in Copenhagen, the number of participants remained high with more than 800. It was organized by D. Edwards who also published the proceedings (coeditor N. E. Raun, [8]) and the additional volume of short communications and posters. There were two keynotes delivered by G. W. Stewart on parallel linear algebra in statistical computations and by B. Efron on computer-intensive statistical inference, and 7 invited papers. They were related to Non-Parametric Estimation, Projection Pursuit, Expert Systems, Algorithms, Statistical Methods, Statistical Data Bases, and Survey Processing. Out of approximately 300 submissions 51 contributed papers were selected. At that time computational statistics had become an integrated part of statistics research with new emerging areas, especially graphical techniques and models, Bayes methods, and smoothing techniques. Nonparametric curve estimation and dimension reduction techniques are discussed at COMPSTAT for the first time. At the same time the COMPSTAT evergreen Expert Systems is still quite present. A real innovation were tutorials in the programme. They covered the fields Dynamic Graphics (R. Becker), Artificial Intelligence (W. Gale), and Graphical Models (N. Wermuth). The new availability of modern computing also makes itself visible in the appearance of the proceedings volume with a relatively larger number of electronically produced papers.

The ninth meeting in Dubrovnik 1990 marks a dramatic change in the positive development of the COMPSTAT symposia seen so far. Submissions were down to 115 (43 contributed papers selected). After six years COMPSTAT was back in a communist country, however when this decision was taken, nobody could foresee the disintegration of Yugoslavia. During the conference around Dubrovnik first road barricades were erected and soon after the civil war broke out (the conference hotel on the sea shore was destroyed in the following years). Anticipating the unrest many participants and speakers did not show up (an audience of around 180 was present). Thus the proceedings volume [9], edited by the organizer K. Momirović does not really represent the conference (many more papers than presentations). The programme was dominated by the subject areas Expert Systems, Multivariate Data Analysis and Model Building, and Computing for Robust Statistics. Special topics were Optimization and Analysis of Spatial Data. All these comprised invited talks (6 invited papers in the proceedings). In addition some of the traditional COMPSTAT topics such as Algorithms, Time Series (with an invited paper) and Computational Inference were present. Despite all external problems it is noteworthy that aspects of modelling and appropriate software played an important part in this meeting, establishing a new COMPSTAT focus. T. Hastie (replacing J. Chambers) presented statistical models in S for the first time and new strategies for GLIM4 were outlined by B. Francis. As a matter of fact it was for the first time that the statistical and graphical environment S (the S-Plus package) was discussed in the COMPSTAT community.

In 1992 the tenth symposium was held in Neuchâtel. It was the general hope that COMPSTAT would recover from the Dubrovnik adventure, however the problems went on. Submissions remained low with about 115. Despite the fact that participation was only around 200, some participants had to stay in remote accommodations, forced to use the cable car to get from Chaumont (great views!) down to the conference site and back, reducing the audience even further.

Y. Dodge, the organizer and proceedings editor (co-editor J. Whittaker), decided to reshape the symposium and the volume. In response to the unexpected low number of submissions and the fact that Physica-Verlag had been sold to the Springer company, he changed the format of the proceedings, giving up the established layout and format as well as the tradition of a complementary volume for short communications and posters, in accepting almost all submitted papers as full contributions (Computational Statistics Volume 1 and 2 [10] of a new Springer-Verlag series).

There is an interesting foreword by Sir D. J. Cox with the title "The Role of Computers in Statistics". In a prologue Y. Dodge and J. Whittaker feared that a de-skilling of the profession due to the dissemination of commercial software packages could take place. When studying the two volumes of COPMPSTAT 1992, one dedicated to statistics and modelling and the other to computation, we were astonished by the broad range of topics. The main subject areas in Volume 1 are Statistical Modelling, Multivariate Analysis, Classification and Discrimination, Symbolic and Relational Data, Graphical Models, Time Series, Nonlinear Regression, Robustness and Smoothing Techniques, Industrial Applications and Bayesian Statistics. Volume 2 comprises Programming Environments, Computational Inference, Package Developments, Experimental Design, Image Processing and Neural Networks, Meta Data, Survey Design and Data Bases. Almost all these topics included an invited lecture. There were neither official keynotes nor tutorials.

The new proceedings format had not been approved by the European Regional Section of the IASC and was changed back to the previous COMPSTAT appearance for the year 1994, remaining in this style up to the present.

The twentieth anniversary of COMPSTAT was celebrated at the eleventh symposium held in Vienna 1994. The program committee, chaired by R. Dutter, tried to find a compromise between the traditional COMPSTAT topics and actual topics when selecting the keynote speaker and the invited speakers. The keynote was given by P. Huber and concerned the treatment of huge data sets. The themes of the invited papers were Multivariate Analysis, Classification and Discrimination, Dynamic Graphics, Numerical Analysis, Nonparametric Regression, MCMC, Selection Procedures, Neural Networks, Change Point Problems, Wavelet Analysis, and Time Series Forecasting. Besides these invited lectures two tutorials were organized: W. Schachermayer introduced statistical problems in finance and insurance and B. Sundgren gave an overview on metadata. Furthermore a discussion about the nature of computational statistics was organized. All together about 280 participants attended this meeting. The organizers returned to the traditional format of publishing the proceedings and an additional volume of short communications and posters. The proceedings [11] were edited by R. Dutter and W. Grossmann and contained the invited and 60 contributed papers, selected from approximately 200 submissions. With respect to statistical software the increasing dominance of S for the development of computational statistics was evident. Other more commercially oriented products were presented during the conference and documented in a separate booklet.

After the symposium in Vienna there was a COMPSTAT Satellite Meeting on Smoothing held at Semmering, attracting almost 50 participants. Because of the COMPSTAT anniversary a historic train was bringing COMPSTAT participants and accompanying persons on the oldest mountain railroad in the world (now a World Cultural Heritage) from Vienna to the spa of Semmering in the Austrian Alps.

The meeting was organized by M. G. Schimek and comprised 7 invited lectures (presenters were B. Cleveland, M. Delecroix, R. Eubank, Th. Gasser, R. Kohn, A. van der Linde, and W. Stuetzle) and two software presentations (S-Plus and for the first time XploRe). W. Härdle and the organizer edited a proceedings volume [12] consisting of 10 papers (not published elsewhere) out of 26 given at the meeting. It also includes an expository discussed paper by J. S. Marron ("A Personal View of Smoothing and Statistics") and two other discussed contributions by W. S. Cleveland and C. Loader ("Smoothing by Local Regression: Principles and Methods") and by B. Seifert and Th. Gasser ("Variance Properties of Local Polynomials and Ensuing Modifications"). It is worth mentioning that local regression smoothing is now a principal tool for normalization of microarray data in genetic research. Since the symposium in Copenhagen 1988 nonparametric smoothing techniques and relevant software had played a steadily increasing role in COMPSTAT.

The twelfth COMPSTAT symposium was organized under the auspices of A. Prat in Barcelona 1996, attracting an estimated number of 300 participants. An opening keynote was delivered by G. Box entitled "Statistics, Teaching, Learning and the Computer" and a closing keynote "Information Markets" was presented by A. G. Jordan. Eleven invited papers covered topics like Time Series, Functional Imaging Analysis, Applications of Statistics in Economics, Classification and Computers, Image Processing, Optimal Design, Wavelet Analysis, Profile Methods, Web-based Computing, and Multidimensional Nonparametric Regression. Apart from the invited lectures the proceedings [13] edited by A. Prat present also 56 contributed papers selected from about 250 submissions, arranged in alphabetical order and grouped according to subjects at the end of the proceedings. From the subject areas one gets the overall impression that the main emphasis was on statistical modelling, in particular Bayesian Methods, Classification, Experimental Design and Time Series from the classical areas, and Neural Networks, Genetic Algorithms, Wavelets and Classification Trees as more recent methodologies. Also of interest is a rather broad spectrum of applications presented at the conference. A novelty was the introduction of awards for the best papers of young researchers.

The thirteenth symposium held at the University of Bristol in 1998 had seen less participants than the previous COMPSTAT. Organizers were R. Payne and P. Green. There was a methodological keynote on wavelets delivered by B. W. Silverman and in addition an applied keynote on the analysis of clustered multivariate data in toxicity studies presented by G. Molenberghs. Three of the 10 invited lectures dealt with various statistical techniques in connection with applications like Mortality Pattern Prediction, Covariance Structures in Plant Improvement Data, and Markov Models in Modeling Bacterial Genoms. The other invited lectures consider rather methodological issues like Design Algorithms, Scaling for Graphical Display, MCMC for Latent Variable Models, Decision Trees, Semi- and Nonparametric Techniques in Time Series, and Time Series Forecasting. In addition there was an invited lecture on teaching in network environments. The 58 contributed papers contained in the proceedings volume (edited by R. Payne and P. Green, [14]) were selected from about 180 submissions. Taking R. Payne's affiliation (IACR Rothamsted) into account, it is not surprising that the proceedings show a strong orientation towards statistical modelling and applications. However, there are also papers dealing with more computer science oriented aspects of computational statistics, in particular computing environments and software packages for special problems. The proceedings were accompanied by a volume comprising the short communications and posters, edited by IACR Long Ashton.

The fourteenth COMPSTAT symposium was held in Utrecht 2000. It was organized by P. van der Heijden (Utrecht University) and J. G. Bethlehem (Statistics Netherlands). The number of participants was around 220. It had a substantial applied focus on the social sciences and official statistics. There were two keynotes (one on multiple imputation by D. B. Rubin and the other on official statistics in the IT-era by P. Kooiman) and 13 invited papers. The invited lectures concerned Algorithms, Bayesian Model Selection, GLMs, HGLMs (a further generalization of GLMs), Imputation, Data Mining, Spatio-Temporal Modelling, Survival Techniques, Time Series, and Teaching. Further there were 60 contributed papers (out of around 250 submissions) following mostly the conventional subject areas of COMP-

STAT. A proceedings volume [15] and a supplement comprising the short communications and posters were published (editors P. van der Heijden and J.G. Bethlehem).

The last (fifteenth) COMPSTAT symposium we can report on took place at Humboldt-Universität zu Berlin in 2002. It was organized by W. Härdle and attracted approximately 220 submissions. This time the primary focus was on business applications, especially in connection with the Internet (such as E-Commerce and Web-Mining) and on the handling of massive and complex data sets (e.g. in genetic research). The idea was to expand the traditional scope of COMPSTAT and to make it attractive for new audiences. However the number of about 260 participants made it clear that this endeavour was not sufficient to substantially enlarge the audience for such a meeting. However it is only fair mentioning that many young researchers showed up for the first time, also joining IASC because of a special promotion scheme.

There was a keynote delivered by T. Hastie entitled “Supervised Learning from Microarray Data”. The other 8 invited talks concerned the topics Bayes Methods, Graphical Methods, Internet Traffic, Smoothing, Teaching, and Time Series. Further there were 90 contributed papers connected to the above topics as well as to Algorithms, Classification, Computational Inference, Computing Environments, Data Mining, Meta Data, and Multivariate Methods. Two additional areas of interest have emerged because of submissions received, the statistical language R and functional data analysis. Innovations were that the printed proceedings volume (edited by W. Härdle and B. Rönz [16]) also appeared as a Springer-Verlag e-book and that the companion volume of short communications and posters was published on a CD. Moreover several prices were granted (among them a new one for software innovation).

7 Conclusions

The evolution of computational statistics has always been strongly influenced by developments in statistical theory, in algorithms, in computer science, and by the problems statisticians are confronted with. In statistical theory many actual topics are connected to concepts and methods of computational statistics requiring definitely more than the proper implementation of well-defined algorithms. With respect to computation we can observe a shift from pure numerical analysis to more graphically oriented techniques and algorithms developed in computer science. This brings about a new quality of cooperation between statistics and computer science with a high potential for future development. The traditional knowledge transfer from computer science to computational statistics was primarily in the areas of statistical packages, statistical languages, statistical graphics and statistical data management systems. Yet these conventional areas are still open to new developments, in particular with regard to statistical Web services and the seamless integration of various tools. Concerning applications the main challenges for computational statistics are complex data structures and very large or huge data, as well as the demand for new analysis strategies. Due to the penetration of all areas of our life by computers one can expect an ever increasing number of challenging tasks.

Although we can identify many inter-connections between computational statistics and computer science, symbolic computation has not received the attention it deserves. Mathematica has been used to implement a number of statistical approaches applying general mathematical notation, this way making it feasible to calculate the results with (at least in principle arbitrarily) high precision. One might envisage a development where similar approaches are introduced in environments like S and R or complement these environments. The ability to use abstractions, symbolic representations and/or general objects/classes in everyday work while having access to low level constructs to improve statistical methods based on experiments or to solve non-anticipated practical problems, could be a promising way for the future.

The review of the COMPSTAT symposia has shown that the meetings and proceedings reflect clearly the international developments of computational statistics of the last 30 years, although with some delay in certain subject areas (e.g. Bayesian Methods, Resampling, Statistical Environments, Smoothing Techniques, Statistical Learning, Tree Based Methods, Wavelets). On the other hand the anticipation of new ideas in connection with Dimension Reduction Methods, Expert Systems, and Robust Techniques was very fast. More recently, with respect to content, there seems to be a shift of focus towards topics related to statistical modelling and at the same time less interest in computer science contributions useful in statistics.

There was a continuous uptrend in conference participation during the first 16 years with symposia covering a rather broad spectrum of computational statistics topics. The nineteen eighties have certainly seen the high time of COMPSTAT with many innovations in statistical computing, a boost in algorithms and professional software (emphasizing personal computing in the second half of the decade), and the early adoption of expert systems. This is also reflected in the size of the meetings, going beyond 300 submissions and ranging between 800 and 900 participants in Rome 1986 and in Copenhagen 1988. After the problems with COMPSTAT 1990 in Dubrovnik and COMPSTAT 1992 in Neuchâtel, the symposia have stabilized since at a lower level of participation. In recent years we have typically seen around 200 submissions and about 250 participants (i.e. very little non-contributing participants). This has already led to new formats of presentation to keep the number of parallel sessions as low as possible.

In general, COMPSTAT was probably not the main forum for the presentation of state-of-the-art research results in computational statistics. It was rather an important forum for the exchange of relevant information in the

European statistical community about current developments from all over the world as well as on practical aspects such as new algorithms and statistical software. This was largely achieved by a dedicated invitation policy. One can say that the organizers of the symposia always have given their best to identify distinguished personalities for keynotes and invited lectures. This way the European research community has received a great deal of valuable impulses that often have proven influential for subsequent projects in Europe. Occasional tutorials were another means of this successful policy, not only attracting young researchers.

COMPSTAT has always been an international undertaking. However, most recently there have been discussions at IASC business meetings focusing on strategies for opening up the European COMPSTAT symposia even further to make them world-wide events in the future, integrating other regional sections such as the Asian Section and the planned African Section (an initiative of S. Azen as President). As far as the Interface Foundation of North America, Inc., is concerned, there was a formal proposal in 1987 to transform it into the North American Section of IASC, however it was voted down by the Interface Board. E. J. Wegman (1997–1999 IASC President) from George Mason University (USA) initiated an informal connection between IASC and Interface which has finally led to the establishment of an IASC–Interface Liaison Committee, chaired by M. G. Schimek as IASC Vice President, to foster mutual interests and to organize invited sessions at each others symposia.

We are all looking forward to this year's COMPSTAT symposium in Prague, celebrating the thirtieth anniversary, chaired by J. Antoch (Charles University of Prague). Maybe the first step towards the next generation of COMPSTAT meetings has already been taken as it is organized in compliance with new guidelines. According to its scientific programme on the Web we can already say that the Prague symposium is going to be truly international with contributions from all IASC regional sections, from Interface, and beyond.

References

[1] Bruckmann, G., Ferschl, F. and Schmetterer, L. (1974, eds.). *COMPSTAT 1974. Proceedings in Computational Statistics.* Physica-Verlag, Wien.

[2] Gordesch, J. and Naeve, P. (1976, eds.). *COMPSTAT 1976. Proceedings in Computational Statistics. 2nd Symposium Berlin/FRG.* Physica-Verlag, Wien.

[3] Corsten, L. C. A. and Hermans, J. (1978, eds.). *COMPSTAT 1978. Proceedings in Computational Statistics. 3rd Symposium Leiden/The Netherlands.* Physica-Verlag, Wien.

[4] Barritt, M. M. and Wishart, D. (1980, eds.). *COMPSTAT 1980. Proceedings in Computational Statistics. 4th Symposium Edinburgh/UK.* Physica-Verlag, Wien.

[5] Caussinus, H., Ettinger, P. and Tamassone, R. (1982, eds.). *COMPSTAT 1982. Proceedings in Computational Statistics. 5th Symposium Toulouse/France.* Physica-Verlag, Wien.

[6] Havránek, T., Šidák, Z. and Novák, M. (1984, eds.).*COMPSTAT 1984. Proceedings in Computational Statistics. 6th Symposium Prague/CSSR.* Physica-Verlag, Wien.

[7] De Antoni, F., Lauro, N. and Rizzi, A. (1986, eds.). *COMPSTAT 1986. Proceedings in Computational Statistics. 7th Symposium Rome/Italy.* Physica-Verlag, Wien.

[8] Edwards, D. and Raun, N. E. (1988, eds.). *COMPSTAT 1988. Proceedings in Computational Statistics. 8th Symposium Copenhagen/Denmark.* Physica-Verlag, Heidelberg.

[9] Momirović, K. and Mildner, V. (1990, eds.). *COMPSTAT 1990. Proceedings in Computational Statistics. 9th Symposium Dubrovnik/Yugoslavia.* Physica-Verlag, Heidelberg.

[10] Dodge,Y. and Whittaker, J. (1992, eds.). *Computational Statistics. Volume 1 and 2. Proceedings of the 10th Symposium, COMPSTAT, Neuchâtel/Switzerland.* Physica-Verlag, Heidelberg.

[11] Dutter, W. and Grossmann, W. (1994, eds.). *COMPSTAT 1994. Proceedings in Computational Statistics. 11th Symposium Vienna/Austria.* Physica-Verlag, Heidelberg.

[12] Härdle, W. and Schimek, M. G. (1996, eds.) *Statistical Theory and Computational Aspects of Smoothing. Proceedings of the COMPSTAT '94 Satellite Meeting held in Semmering, Austria, 27-28 August 1994*, Physica-Verlag, Heidelberg.

[13] Prat, A. (1996, ed.). *COMPSTAT 1996. Proceedings in Computational Statistics. 12th Symposium Barcelona/Spain.* Physica-Verlag, Heidelberg.

[14] Payne, R. and Green, P. (1998, eds.). *COMPSTAT 1998. Proceedings in Computational Statistics. 13th Symposium Bristol/UK.* Physica-Verlag, Heidelberg.

[15] Bethlehem, J. G. and van der Heijden, P. G. M. (2000, eds.). *COMPSTAT 2000. Proceedings in Computational Statistics. 14th Symposium Utrecht/The Netherlands.* Physica-Verlag, Heidelberg.

[16] Härdle, W. and Rönz, B. (2002, eds.). *COMPSTAT 2002. Proceedings in Computational Statistics. 15th Symposium Berlin/Germany.* Physica-Verlag, Heidelberg.

[17] ACM (1999). *Software Systems Award. Press Release.* New York, March 23, 1999. `http://www.acm.org/announcements/ss99.html`.

[18] Adam, A. (1973). *Von himmlischen Uhrwerk zur statistischen Fabrik. 600 Jahre Entdeckungsreise in das Neuland österreischer Statistik und Datenverarbeitung.* Munk, Wien.

[19] Anscombe, F. (1981). *Computing in Statistical Science through APL.* Springer-Verlag, New York.

[20] Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimation of Location: Survey and Advances.* Princeton University Press, Princeton/NJ.

[21] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Pacific Grove/CA.

[22] Buja, A., Hastie, T. and Tibshirani, R. (1989). *Linear smoothers and additive models* (with discussion), Ann. Statist., **17**, 453 − 555.

[23] Chambers, J. M. (1998). *Programming with Data - A Guide to the S Language.* Springer-Verlag, New York.

[24] Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Chapman & Hall, London.

[25] Christeller, S., Meystre, A., Ballmer, U. and Glutz, G. (1974). *SAS. A Software System for Statistical Data Analysis.* In Bruckmann, G., Ferschl, F. and Schmetterer, L. (eds.). COMPSTAT. Proceedings in Computational Statistics, 479 − 488.

[26] Codd, E. F. (1970). *A Relational Model for Large Shared Data Banks.* CACM, **13**, 377 − 387.

[27] Coppi, R. (2002). *A Theoretical Framework for Data Mining: the Informational Paradigm.* Computat. Statist. Data Anal., **38**, 501 − 515.

[28] Cox, D. J. (1972). *Regression models and life-tables* (with discussion). J. Royal Statist. Soc., **B 34**, 187 − 220.

[29] Dahl, O. R. and Nygaard, K. (1966). *Simula – an Algol-based Simulation Language.* CACM, **9**, 671 − 678.

[30] Dempster, A. P., Laird, N. and Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm* (with discussion). J. Royal Statist. Soc., **B 39**, 1 − 38.

[31] Devroye, L. (1986). *Non-Uniform Random Variate Generation.* Springer-Verlag, New York.

[32] Dirschedl, P. and Ostermann, R. (1994, eds.). *Computational Statistics.* Physica-Verlag, Heidelberg.

[33] Dixon, W. J. (1971, ed.). *BMD. Biomedical Computer Programs.* University of California Press, Los Angeles/CA.

[34] Dodge, Y. and Whittaker, J. (1992). In Dodge, Y. and Whittaker, J. (eds.) *Science, Data, Statistics and Computing.* In *Computational Statistics. Volume 1*, 3 − 7.

[35] Efron, B. (1979). *Bootstrap methods: another look at the jackknife.* Ann. Statist., **7**, 1 − 26.

[36] Efron, B. (2002). Statistics in the 20th Century and the 21th. In Dutter, R. (ed.) *Festschrift 50 Jahre Österreichische Statistische Gesellschaft 1951–2001.* Austrian Statistical Society, Vienna, 7 – 20.

[37] Efron, B. (2003). Robbins, empirical Bayes and microarrays. Ann. Statist., **31**, 366 – 378.

[38] Fisherkeller, M. A., Friedman, J. H. and Tukey, J. W. T. (1974). *PRIM-9. An Interactive Multidimensional Data Display System.* Stanford Linear Accelerator Publication No. 1408. Palo Alto/CA.

[39] Fleck, C. (2000). *Wie Neues nicht entsteht. Die Gründung des Instituts für Höhere Studien in Wien durch Ex-Österreicher und die Ford Foundation.* Österreichische Zeitschrift für Geschichtswissenschaften, **1**, 129 – 177.

[40] Francis, I. (1981). *Statistical Software. A Comparative Review.* North Holland, New York.

[41] Frawley, W., Piatetsky-Shapiro, G. and Matheus, C.(1992). *Knowledge Discovery in Databases: An Overview.* AI Magazine, Fall 1992, 213 – 228.

[42] Friedman, J. H. and Stuetzle, W. (2002). *John W. Tukey's work on interactive graphics.* Ann. Statist., **30**, 1629 – 1639.

[43] Friedman, J. H. and Tukey, J. W. (1974). *A projection pursuit algorithm for exploratory data analysis.* IEEE Trans. Comp., **C 23**, 881 – 890.

[44] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1996). *Bayesian Data Analysis.* Chapman &Hall, London.

[45] Geman, S. and Geman, D. (1984). *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.* IEEE Trans. Pattern Anal. Machine Intellig., **6**, 721 – 741.

[46] Gentle, J. E. (2002). *Elements of Computational Statistics.* Springer-Verlag, New York.

[47] Gershenfeld, N. (1999). *The Nature of Mathematical Modeling.* Cambridge University Press, Cambridge/UK.

[48] Goldstine, H. H. (1972). *The Computer from Pascal to von Neumann.* Princeton University Press, Princeton/NJ.

[49] Hand, D. (1996). Classification and Computers, Shifting the Focus. In Prat, A. (ed.) *COMPSTAT 1996. Proceedings in Computational Statistics.*, 77 – 88.

[50] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall, London.

[51] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer-Verlag, New York.

[52] Hastings, W. K. (1970). *Monte Carlo sampling methods using Markov chains and their applications.* Biometrika, **57**, 97 – 109.

[53] Härdle, W., Klinke, S. and Turlach, B. A. (1995). *XploRe: An Interactive Statistical Computing Environment.* Springer-Verlag, New York.

[54] Heide, L. (2003). *Diffusing the emerging punched card technology in Europe 1889-1914.* Information Systems and Technology in Organizations and Society. ISTOS-Workshop Universitat Pompeu Fabra, Barcelona. `http://cbs.dk/staff/lars.heide/ISTOS/paper-10.pdf`.

[55] Hornik, K. and Leisch, F. (2002). Vienna and R: Love, Marriage and the Future. In Dutter, R. (ed.) *Festschrift 50 Jahre Österreichische Statistische Gesellschaft 1951–2001*, Austrian Statistical Society, 61 – 70.

[56] Huber, P. J. (1964). *Robust estimation of a location parameter*, Ann. Math. Statist., **35**, 73 – 101.

[57] Huber, P. J. (1994). Huge Datasets. In Dutter, W. and Grossmann, W. (eds.) *COMPSTAT 1994. Proceedings in Computational Statistics*, 1 – 13.

[58] Huber, P. J. (1999). *Massive Dataset Workshop: Four Years After*, J. Computat. Graph. Statist., **8**, 635 – 652.

[59] Ihaka, R. and Gentleman, R. (1996). *R: A language for data analysis and graphics*. J. Computat. Graph. Statist., **5**, 299 – 314.

[60] Lauritzen, S. L. and Wermuth, N. (1989). *Graphical models for association between variables, some of which are qualitative and some quantitative*. J. Royal Statist. Soc., **B 50**, 157 – 224.

[61] Lauro, C. (1996). *Computational Statistics or Statistical Computing, is that the question?* Computat. Statist. Data Anal., **23**, 191 – 193.

[62] Mehta, C. R. and Patel, N. R. (1992). Exact Logistic Regression: Theory, Applications, Software. In Dodge, Y and Whittacker, J. (eds.) *Computational Statistics. Volume 2*, 63 – 78.

[63] Mehta, C. R. and Patel, N. R. (1997). *Exact Inference for Categorical Data*. Electronic Publication: Harvard University and Cytel Software Corporation, `http:`
`www.cytel/Library/articles.asp`.

[64] Mehta, C. R., Patel, N. R. and Senchaudhuri, P. (2000). *Efficient Monte Carlo Methods for Conditional Logistic Regression*. J. Amer. Statist. Assoc., **95**, 99 – 108.

[65] Metropolis, N. and Ulam, S. (1949). *The Monte Carlo Method*. J. Amer. Statist. Assoc., **44**, 335 – 342.

[66] Meyer, D. Leisch, F., Hothorn, T. and Hornik, K. (2002). StatDataML: An XML Format for Statistical Data. In Härdle, W. and Rönz, B. (eds.)*COMPSTAT 2002. Proceedings in Computational Statistics.*, 545 – 550.

[67] Monahan, J. F. (2001). *Numerical Methods of Statistics*. Cambridge University Press, Cambridge/UK.

[68] Nelder, J. A. (1974). Genstat - A Statistical System. In Bruckmann, G., Ferschl, F. and Schmetterer, L. (eds.) *COMPSTAT. Proceedings in Computational Statistics*, 499 – 506.

[69] Nelder, J. A. (1978). The Future of Statistical Software. In Corsten, L. C. A. and Hermans, J. (eds.) *COMPSTAT 1978. Proceedings in Computational Statistics*, 11 – 19.

[70] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Royal Statist. Soc.*, **A 135**, 370 – 84.

[71] Neuwirth, E. and Baier, T. (2002). Embedding R in standard software, and the other way round. In Hornik, K. and Leisch, F. (eds.) *DSC 2001 Proceedings. 2nd International Workshop on Distributed Statistical Computing*, `http://www.ci.tuwien.ac.at/Conferences/DSC-2001`.

[72] Owen, D. B. (1976). *On the history of statistics and probability. Proceedings of a symposium on the American mathematical heritage.* Dekker, New York.

[73] Ripley, B. D. (1987). *Stochastic Simulation.* Wiley, New York.

[74] Robbins, H. (1956). *An empirical Bayes Approach to Statistics.* Proc. Third Berkeley Symp. Statist. Probab., **1**, 157 – 163.

[75] Schäffler, O. (1895). *Neuerungen an statistischen Zählmaschinen.* Österreichisches Patentprivileg No.46/3182, Patentarchiv, Wien.

[76] Shoshani, A. (1997). *OLAP and Statistical Databases: Similarities and Differences.* Proceedings 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems 1997, 185 – 196.

[77] Stone, C. (1977). *Consistent nonparametric regression* (with discussion). Ann. Statist., **5**, 595 – 645.

[78] Sundgren, B. (1975). *Theory of Data Bases.* Petrocelli/Charter, New York.

[79] Temple Lang, D. (2000). *The Omegahat Environment: New Possibilities for Statistical Computing.* J. Computat. Graph. Statist., **9**, 423 – 451.

[80] Thisted, R. A. (1988). *Elements of Statistical Computing.* Chapman & Hall, New York.

[81] Tierney, L. (1989). *XLISP-STAT: A Statistical Environment Based on the XLISP Language*, Technical Report No. 528, School of Statistics, University of Minnesota, `http://www.stat.umn.edu/luke/xls/tutorial/techreport/techreport.html`.

[82] Tierney, L. (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics.* Wiley, New York.

[83] Tukey, J. W. (1962). *The future of data analysis.* Ann. Math. Statist., **33**, 1 – 67 and 812.

[84] Tukey, J. W. (1970). *Exploratory Data Analysis. Volume I and II (limited preliminary edition).* Addison-Wesley, Reading/MA.

[85] Tukey, J. W. and Cooley, J. W. (1965). *An algorithm for the machine calculation of complex Fourier series.* Math. Comput., **19**, 237 – 301.

[86] Wegman, E. J. and Marchette, D. J (2003). *On Some Techniques for Streaming Data: A Case Study of Internet Packet Headers.* J. Computat. Graph. Statist., **12**, 893 – 914.

[87] Wilkinson, L., Rope, D. J., Carr, D. B. and Rubin, M. A. (2000). *The Language of Graphics*. J. Computat. Graph. Statist., **9**, 558 − 581.

[88] Zemanek, H. (1975). *Otto Schäffler. Ein vergessener Österreicher. Die Biographie eines genialen Unternehmers und Erfinders*. Österreichischer Gewerbeverein. Jahrbuch, **92**, 71 − 92.

Acknowledgement: First of all the authors wish to thank Prof. Jaromir Antoch (Charles University of Prague) for giving them the opportunity to present a historical keynote. Further the authors appreciate valuable hints and comments from the following colleagues: Dr. Lutz Edler (German Cancer Research Center Heidelberg), Dr. Karl A. Fröschl (Electronic Commerce Competence Center Vienna), Dr. Walter Grafendorfer (Austrian Computer Society), Prof. Kurt Hornik (Wirtschaftsuniversität Wien), and Prof. Edward J. Wegman (George Mason University). However, all errors and omissions are in the responsibility of the authors.

Address: W. Grossmann, University of Vienna, Institute for Statistics and Decision Support Systems, Universitätsstraße 5, A-1010 Wien, Austria
M.G. Schimek, Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2, A-8036 Graz, Austria
P.P. Sint, Austrian Academy of Sciences, Institute for European Integration Research, Prinz Eugen Straße 8-10/2, A-1040 Wien, Austria

E-mail: `wilfried.grossmann@univie.ac.at`,
`michael.schimek@meduni-graz.at`, `sint@oeaw.ac.at`

HYBRID ALGORITHMS FOR CONSTRUCTION OF D-EFFICIENT DESIGNS

Abdul Aziz Ali and Magnus Jansson

Key words: Exact D-optimal designs, genetic algorithms, local search.

COMPSTAT 2004 section: Design of experiments.

Abstract: We construct exact D-efficient designs for linear regression models using a hybrid algorithm that consists of genetic and local search components. The genetic component is a genetic algorithm (GA) with a 100% mutation rate and ranking selection. The local search methods we use are based on the G-bit improvement and a combination of the Powel multidimensional and Brent line optimization techniques. Computational results show that the hybrid algorithm generates designs that are comparable in efficiency to those found using the modified Fedorov algorithm (MFA), but without being limited to using a given set of candidate points.

1 Introduction

An experimental design is said to be optimal if it meets predefined criteria that determine the precision with which the model parameters or response is estimated. The D-optimality criterion Keifer and Wolfowitz [12] puts emphasis on the precision with which the model parameters are estimated by maximizing the determinant of the model's information matrix. This criterion has the intuitively appealing interpretation of minimizing the volume of the joint confidence ellipsoid of the least squares regression parameter estimates.

Exact D-optimal designs are calculated using optimization algorithms such as those given by Cook and Nachtsheim [6] and Johnson and Nachtsheim [11] among others. These algorithms iteratively maximize the determinant of the information matrix by sequentially, or simultaneously, adding and deleting points to the design. Many of the most used algorithms require an explicit set of candidate points to work with, thus putting heavy demands on prior domain-specific knowledge of the optimization problem. Although not as common, evolutionary algorithms have also been used to calculate D-optimal designs. Govaerts and Sanchez [8] were the first to use genetic algorithms (GAs) to find exact D-optimal designs. However, their algorithm incorporated the use of a candidate set of design points, much like the more traditional algorithms. Poland et al. [17] used a GA to improve on the standard Monte Carlo algorithms by applying DETMAX and k-exchange as the mutation operator. Compared to the exchange algorithms, their algorithm was slower but yielded better results. Broudiscou et al. [5] successfully applied a purely genetic algorithm to the exact D-optimal design problem in a chemometrics setting. GAs have since then been used by Montepiedra et al. [15] who omitted the mutation operator in favor of faster convergence and Heredia-Langner et al. [10] who used real value encoding in place of the more traditional binary encoding. The latter named also give an excellent introduction to the use of GAs in calculating optimal designs.

This paper presents the use of hybrid algorithms in calculating D-efficient or near D-optimal designs. The hybrid algorithms considered here consist of a genetic component with 100% mutation rate and local search methods. The mutation operator is extensively used in order to escape from local optima. The hybrid algorithm is therefore implemented in two stages: The genetic component finds a neighborhood point of a local optimum and the local search finds the local optimum. The genetic component is then updated with the coordinates of the local optimum and the process is repeated until some termination condition is met.

1.1 Model and the exact D-optimal design problem

In many experimental situations, the experimenters usually approximate the relationship between the response variable and the input factors with the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{X}$ is the $(n \times p)$ matrix of factor levels (design matrix), $\boldsymbol{\beta}$ is the $p \times 1)$ vector of unknown regression parameters, $\boldsymbol{y}$ is the $(n \times 1)$ vector of observations and $\boldsymbol{\epsilon}$ is the $(n \times 1)$ vector of error terms that are assumed to be *iid* (possibly normally distributed) with $E(\boldsymbol{\epsilon}) = 0$ and $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}) = \boldsymbol{\sigma}^2\boldsymbol{I}$.

When the goal is to construct exact designs, the problem becomes one of how to determine $\boldsymbol{x}_i^{\mathrm{T}}$ $i = 1, 2, 3, \ldots, n$ from the region defined by all the level combinations of the factors called the design region χ, so that the resulting design will estimate some function of $\boldsymbol{\beta}$ with a precision that is at least as good as that provided by any other design in χ. The exact n-point design is denoted by

$$\xi_n = \left\{ \begin{matrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_k \\ r_1/n & r_2/n & \cdots & r_k/n \end{matrix} \right\},$$

where $\sum_{i=1}^{k} r_i = n$ and r_i is the number of trials at $\boldsymbol{x}_i$. The standardized predictor variance is given by,

$$d(\boldsymbol{x}, \xi) = n\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}^{-1}(\xi_n)\boldsymbol{x},$$

a function of the design ξ_n and the point at which the prediction is made.

The design ξ^* is an exact D-optimal design if $\boldsymbol{M}$ is a non-singular matrix and the following is satisfied:

$$\min_{\xi} \left|\boldsymbol{M}^{-1}(\xi)\right| = \left|\boldsymbol{M}^{-1}(\xi^*)\right|.$$

A measure of efficiency is the D-efficiency which is defined as follows: A design ξ_1 has a D-efficiency relative to ξ_2 given by

$$D - eff = 100 \cdot \left[\frac{|\boldsymbol{M}(\xi_1)|}{|\boldsymbol{M}(\xi_2)|} \right]^{\frac{1}{p}} .$$

This comparison is valid even when the designs being compared are of different sizes because the comparison is based on the information per point for each design. For the interested reader an excellent review of optimum design theory is given by Ash and Hedayat [3] and books by Atkinson & Donev [2] and Silvey [18].

1.2 Commonly used algorithms and genetic search

Exact D-optimal designs are calculated using optimization algorithms such as those given by Dykstra [7], Cook and Nachtsheim [6], Mitchell [14], Wynn [20] and Johnson and Nachtsheim [11] among others. These algorithms are search heuristics that iteratively maximize the determinant of the information matrix by sequentially adding and deleting points to the design or exchanging points between the existing design and a candidate set of points. The algorithms update the design matrix with rank-one matrices derived from the candidate points as shown by the following formula which is often used for computational efficiency. Upon the addition of a point to a n point design ξ_n, the change in the information matrix is,

$$\left| \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right| \rightarrow \left| \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{x} \right) \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{x}^{\mathrm{T}} \end{pmatrix} \right| = \left| \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right| \cdot \left(1 + \frac{d(\boldsymbol{x}, \xi_n)}{n} \right) .$$

As a consequence the point $\mathbf{x}$ whose addition to the design maximizes the determinant of the information matrix is the point whose standardized predicted response variance calculated from the current design is largest.

A major drawback of these algorithms is that for each iteration the sequential algorithms have to calculate the variance functions of the current designs. Exchange algorithms calculate the variance functions of all possible pairs of candidate and the current design's points, a process that puts heavy demands on memory and speed even for moderately large designs.

Although not as commonly used as the exchange algorithms, evolutionary algorithms have also been used to calculate D-optimal designs. Genetic algorithms (GAs) have been successfully used to search for optimal or near optimal solutions in large-scale optimization because of their versatility. GAs do not require convexity or even continuity of a function and have their strong points as a powerful computational tool for function optimization because they are less susceptible to being trapped in local optima as compared to many other numerical optimization techniques.

GAs usually, but not always, encode the possible solutions to an optimization problem using binary strings. For example, if the range of possible solutions lies in the interval $[-\alpha, \alpha]$ then the 8-bit binary string 00000000 will represent $-\alpha$ and 11111111 will represent $+\alpha$. A randomly generated set of strings forms the initial population from which the GA starts its search. Initial candidate solutions (strings) are usually uniformly sampled from the search space in order to introduce variability in the set of candidate solutions. This initialization process is a random search whereby a number of possible solutions are randomly generated and the best solutions (the fittest strings) are remembered.

2 GA implementation for finding D-efficient designs

2.1 Encoding the designs

The GA is implemented by encoding each complete design, including the number of experimental runs, as a one bit-string. Binary encoding is the most widely used form of representation because of its flexibility and also because its theoretical framework is well developed Goldberg [9]. Binary encoding also allows for a simple way to apply the mutation and recombination operators. Consider the m-bit representation of a single factor design at the high and low levels. The base 10 (decimal) representation of the coordinate points will be 0 and $2^m - 1$ respectively. This design region is transformed to the familiar $[-1, 1]$ by the function $f : x \mapsto \frac{2x}{2^m-1} - 1$, where x is the decimal representation.

The length of the bit-string is determined by the number of bits required to code the coordinates of the levels taken by each factor, the number of factors, and the number of trials. For example, an n-trial experimental design with k factors each requiring p bits to code its coordinates would require npk bits.

2.2 Initialization and selection

The initial population of strings (at iteration 0) consists of N strings. Because the D-optimality criterion pushes the design points to the edges or vertices of the design region, initial designs are generated by drawing random variates from a U-shaped distribution which puts more mass on the edges. The *Beta* distribution with $\alpha = 0.35$, $\beta = 0.35$, transformed from $[0, 1]$ to cover the design region for each factor is used. This is done in order to sample fitter strings than would have otherwise been found using the commonly used uniform distribution. The designs are then evaluated, ranked according to their fitness, and encoded as bit-strings. The first iteration produces the N fittest strings.

In the second and subsequent iterations, the N fittest strings from the earlier iteration are selected, N mutated copies of these are made, and M strings which result from their recombination are generated. These $2N + M$ strings are evaluated and ranked according to their fitness and the N fittest strings are kept. This type of selection leads to what is known as an elitist algorithm. It ensures that the fittest strings are preserved from one iteration to the next and removes the possibility that all strings found in iteration $i+1$ are poorer than the fittest string found in iteration i. Other methods of selection such as selection with probability proportional to fitness may result in the loss of the fittest strings as there is a positive probability that any one string could be lost.

2.3 Recombination

Recombination when applied to strings with binary coding is usually performed by single or multi-point crossover. Single point cross-over is used in this application because of its simplicity and ease of execution. This is done by sampling without replacement of a pair of strings with probability proportional to their fitness. A point is randomly chosen and each string is divided into two segments. The strings then swap their segments and a new pair of strings is created. In this way, strings with high fitness are paired with each other and exchange sub-strings. Those that inherit segments which result in high fitness (also called building blocks) are kept for the next iteration.

2.4 Mutation

Mutation relocates the candidate solutions to some other points in the search space. Although it is common to use mutation with low probability so as not to destroy highly fit strings and prolong the computation times, we always apply mutation with probability $p_m = 1$. The reason for mutating in this way is that copies of the strings are made prior to mutating them so that strings are not lost because of mutation. Also a ranking selection which results in the elitist algorithm is used. This algorithm implements mutation by switching one randomly selected bit per string. The inversion operator is a generalization of the mutation operator. Whereas the mutation operator switches one bit per string, the inversion operator flips a whole string segment. The start and end positions for the inversion are randomly decided. Inversion is used when there is no improvement in fitness in at least one iteration.

The GA search process is thus iterative: evaluation, selection and recombination using the basic operators: selection, cross-over and mutation, until some termination condition is met. The basic algorithm is given by the pseudo code below.

If $s(i)$ is the set of strings processed by the GA at iteration i and f is the objective function then,

```
$i = 0$; initialize $s(i)$;
 evaluate $f(i)$;
  do while (termination condition is not met);
   select $s(i+1)$ from $s(i)$;
```

```
        recombine $s(i+1)$;
        mutate $s(i+1)$;
        evaluate $s(i+1)$;
    $i = i + 1$;
    End;
```

3 Local search methods

Local search is a strategy of searching a neighborhood until a gradient is found, moving along the gradient, then updating the starting point and generating a new neighborhood. We will examine two local search methods. The first method is local improvement on the genetic algorithm using a modified version of the G-bit improvement, Goldberg [9]. The G-bit improvement is implemented in the following manner.

1. Select the fittest string which the genetic algorithm generates.
2. Sweep the string bit by bit, evaluating the fitness of every string that results from one-bit switches. If a bit change results in a violation of any of the constraints then discard the string.
3. When a string is found that has a better fitness than the first (starting) string then replace the starting string with the fitter string.
4. Repeat the process until no further improvement is made after sweeping through the fittest string.

An objective function is evaluated for every switch which makes the method somewhat slow. The method is therefore most useful when the genetic algorithm converges to a point on the search grid that is very close to the optimum and there is a steep gradient between the two points. This method is only used on the fittest string found after the termination condition has been met by the GA.

Because of the difficulty of computing the directional derivatives of poorly characterized functions, we use methods that do not require differentiability. Local search is traditionally done using greedy algorithms such as those of Lawler [13] and Syslo et al. [19]. We implement local search by a combination of Powell's method and Brent line optimization as given in Press et al. [16]. Powell's method is given below. Readers interested in the technical details are referred to *Numerical recipes in C* available on-line at `www.library.cornell.edu/nr`. The algorithm establishes the direction along which the optimization takes place and then the Brent line optimization is used iteratively. Because minimization and maximization are trivially related, we consider the optimization problem as the minimization of a function f without loss of generality.

The algorithm begins by initializing the direction set to the basis vectors of the n dimensional space i.e.

$\mathbf{u}_i = \mathbf{e}_i \; i = 1, \ldots, n$.

1. Save the starting position as $\mathbf{p}_0$.
2. For $i = 1, \ldots, n$, move $\mathbf{p}_{i-1}$ to a minimum of f along the direction $\boldsymbol{u}_i$ and call this point $\boldsymbol{p}_i$.
3. Set $\boldsymbol{u}_{n+1} \leftarrow \boldsymbol{p}_n - \boldsymbol{p}_0$.
4. Move $\boldsymbol{p}_n$ to a minimum along the direction $\boldsymbol{u}_{n+1}$ and call this point $\boldsymbol{p}_0$.
5. Set $\boldsymbol{u}_k \leftarrow \boldsymbol{u}_{n+1}$, $(1 \leq k \leq n)$ where k is the index where the objective function made its greatest decrease.

In addition to the design region itself which is a constrained space in $\mathbb{R}^n$, it is not unusual to encounter constraints in design problems. The main difficulty when using these methods in constrained spaces is that the direction set degenerates to vectors of null norm at the edges of the search space. Bracketing minima may be impossible because one of the points needed to bracket the minimum may not be within the limits of the constraints. To overcome these limitations we have modified the algorithm to re-initialize the direction set along the edges of the design region. When local search leads to a point that violates any constraint, a new search is initiated closer to the starting point.

4 Examples

4.1 Response surface design in two factors

Box and Draper [4] analytically determined D-optimum designs for a second order response surface model in two factors using 6 to 9 design points. Exact D-efficient designs for their model are found using the hybrid algorithm and the genetic component of the hybrid algorithm used alone for comparison as well as for validation and for testing the performance of the algorithms.

The second order response surface model in two factors is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon \,.$$

The design region is given as: $\chi = [-1, 1]^2$.

8 bits were used to encode each coordinate point and 6 strings were used to initialize the algorithm.

The hybrid and genetic algorithms were run $10,000$ times and the average efficiencies of the resulting designs computed. Details of the performance and the average D-efficiencies of the designs found using the algorithms are summarized in Table 1.

The hybrid algorithm required an initial population of only 6 strings and 12 iterations to calculate exact D-efficient designs for the response surface

		Analytical	*Hybrid Algorithm*		*Genetic Algorithm*	
N	i	$\|(\mathbf{X}^T\mathbf{X})^{-1}\|$	$\|(\mathbf{X}^T\mathbf{X})^{-1}\|$	*D-eff**	$\|(\mathbf{X}^T\mathbf{X})^{-1}\|$	*D-eff**
6	12	3.7350E-3	3.7698E-3	99.84	3.6233E-2	68.47
7	12	1.0196E-3	1.0790E-3	99.06	8.0309E-3	70.89
8	12	4.2340E-4	4.5345E-4	98.86	2.9834E-3	72.22
9	12	1.9290E-4	2.2701E-4	97.32	1.3027E-3	72.73

N=number of design points i=number of iterations *D-eff**=Average D-efficiency

Table 1: Comparison of the the Hybrid Algorithm and the GA with the analytically calculated values for the response surface model.

model with 6 to 9 points. This indicates that the local search component of the hybrid algorithm was used to a large extent to find the designs that minimize $|(\mathbf{X}^T\mathbf{X})^{-1}|$.

Exact D-efficient designs are rarely found using analytical function optimization as shown above. When the design region is poorly characterized and/or constrained, it is usual practice to generate efficient designs using computerized algorithms. The next two examples are mixture designs with both linear and non-linear as well as single and multi-component constraints imposed on their design regions.

4.2 Mixture experiment with quadratic constraints

This example is found in Atkinson & Donev [2, pp. 186–187]. Using a three component mixture experiment, models have been first fitted to two responses after which measurements are made on a third response, but only in the region where the other two responses have satisfactory values. The requirements that $\widehat{Y}_1 \geq c_1$ and $\widehat{Y}_2 \geq c_2$ for specified c_1 and c_2 lead to the following quadratic constraints:

$$-4.062x_1^2 + 2.962x_1 + x_2 \geq 0.6075$$

$$-1.174x_1^2 + 1.057x_1 + x_2 \geq 0.5019$$

The D-optimum continuous design for the second order canonical polynomial uses 6 support points with equal weight and is given in Atkinson & Donev [2]. We applied the hybrid and genetic algorithms to finding exact D-efficient designs for this problem using 12 design points and compared them to the designs found using 200 iterations of the MFA with a randomly generated set of 72 points that satisfy all the constraints. The candidate points were generated by one execution of the GA. The MFA used the value ϵ=1.0E-7 as the smallest value that is considered to be non-zero when the search no longer yields an improved design.

The hybrid and GA were initialized using 6 strings (designs) assembled from the same set of candidate points that were used by the MFA. Each coordinate point was encoded using 16 bits which gives a search grid step

Algorithm	i	$\lvert(\mathbf{X}^T\mathbf{X})^{-1}\rvert$	*D-eff**	*Time**(s)
Modified Fedorov	200	7.5698E5	100	0.5
Hybrid Algorithm	200	8.9377E5	97.29	22.33
Genetic Algorithm	200	1.6356E6	87.97	20.88

i=number of iterations *D-eff**=Average D-efficiency *Time**=Average time

Table 2: Results for the example in section 4.2.

of size $1/2^{16}$=1.52587E-5. This search grid is finer than that used for the previous example because the design region for this example is not as regular and symmetric. The termination condition was when 200 iterations had been completed regardless of when the last improvement was made. The genetic and hybrid algorithms were run $10,000$ times and the average efficiencies and times are shown in Table 2.

The results show that the combination of the GA and local search finds efficient designs in a relatively short time using few iterations as seen from the optimized objective function value. This, in the presence of non-linear constraints on the design region.

4.3 Resin vehicle characterization

Altekar and Scarlatti [1] designed an experiment to characterize gel vehicles for use in lithographic inks. A combination of a factorial and a mixture design was used to study the effects of varying the ratio of two resins and other formulation variables on the viscosity of the inks. Each formulation consisted of two resin solids, gelling agent, ink oil and alkyd varnish. The amount of alkyld varnish was fixed at 7% in each formulation and the ink oil was an inert variable used as a filler. The ratio of the two resins was varied as follows: Resin A/Resin B ratio 60/40, 50/50 and 40/60, and were coded as $[-1, 0, 1]$ for the low, mid and high levels respectively.

In order to test the hybrid and genetic algorithms, the same mixture proportions were used to find D-efficient designs. To make the problem more challenging, the ratio of the solids was allowed to vary continuously between 6/10 and 10/6. Table 3 shows the constraints on the design region for the mixture experiment.

Because the resin solids, gelling agent and ink oil had to add up to 93%, the amount of ink oil was automatically restricted to 37–47.67%. In addition to the constraint that all the mixture proportions sum to unity, the following multi-component constraints are also imposed:

$$\text{Resin solids: } 0.45 \leq x_1 + x_2 \leq 0.55.$$

$$\text{Ratio of solids: } \frac{6}{10} \leq \frac{x_1}{x_2} \leq \frac{10}{6}.$$

Component	*Minimum*	*Maximum*
x_1- Resin A	0.0000	0.5500
x_2- Resin B	> 0.0000	0.5500
x_3- Gelling agent	0.0033	0.0100
x_4- Ink oil	0.3700	0.4767
x_5- Alkyd varnish	0.0700	0.0700

Table 3: Restrictions on the design region.

Algorithm	i	$\lvert(\boldsymbol{X}^T\boldsymbol{X})^{-1}\rvert$	*D-eff**	*Time**(s)
Modified Fedorov	200	5.3805E26	79.02	1.32
Hybrid Algorithm	200	8.1824E25	100	31.84
Genetic Algorithm	200	1.0412E27	72.76	29.32

i=number of iterations *D-eff**=Average D-efficiency *Time**=Average time

Table 4: Results for the example in section 4.3.

Let the solids be given by $x_2^* = x_1 + x_2$. The following model was considered for the purposes of evaluating the algorithm:

$$y = \beta_1 x_1 + \beta_2 x_2^* + \beta_3 x_3 + \beta_4 x_4 + \beta_{23} x_2^* x_3 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{24} x_2^* x_4 + \epsilon\,.$$

A 24 point design was generated using the hybrid algorithm. For comparison purposes 200 iterations of the MFA with a candidate set of 144 points which satisfy all the constraints was used. The candidate set was again generated using the GA. The hybrid algorithm and the GA were later re-initialized using the same set of points assembled into 6 designs. Each coordinate point was coded using 16 bits and the termination condition was when 200 iterations had been completed. The GA and hybrid algorithm were run $10,000$ times. Details of the average efficiencies and times are shown in table 4.

Table 4 shows that the hybrid algorithm finds on average designs with higher relative efficiency than those found using the MFA for this problem. Whereas the MFA can only be as good as the quality of its candidate points, the hybrid algorithm generates new design points through local search, selection, and recombination. As a result, the hybrid algorithm arrives at efficient designs without the benefit of using a specific set of candidate points.

5 Conclusions

A hybrid algorithm used to find D-efficient designs for linear regression models is presented in this paper. The genetic component of the hybrid algorithm allows for a high mutation probability without necessarily prolonging the time to convergence. This is possible because mutated copies of the strings are re-injected into the population of strings during every iteration and only the fittest strings are selected for the succeeding iterations. This greatly increases the chances of escaping local optima when applied to poorly characterized functions with many local extrema. Genetic algorithms are very efficient and are designed to search large spaces. However, they require a large initial population of strings to work with and the resulting variation inevitably leads to long computing times if the search domain is to be thoroughly explored. Searching the neighborhood of each point and updating the population of strings at every iteration of the GA with fitter strings that result from local search leads to much faster convergence than using the GA alone. The hybrid algorithm presented here therefore uses a small population of strings to search for efficient designs. It also requires a relatively few number of iterations and as a consequence less computing time is required to find efficient designs. The computing times for the examples used in this paper are real times (not CPU times) when using a 2.0 GHz Pentium PC. It should be noted that although the hybrid algorithm provides designs that are as efficient as those obtained using the MFA, it usually is slower depending on the the number of the candidate points supplied to the MFA, but has a distinct advantage when the candidate set of points is not of high quality or even not available. This relieves the experimenter from having to start with some previous knowledge of the search domain.

The algorithm presented in this paper is coded in *Pascal* using *Borland Delphi version* 4 and is available as a *.exe* file upon contacting the authors. The application that runs the algorithm allows for customizing of all the GA and local search parameters and generates the design points, the design matrix, the information matrix and its eigenvalues, the variance function plots as well as the records and graphical history of the optimization process, among other things.

References

[1] Altekar M., Scarlatti. A. N (1997). *Resin vehicle characterization using statistically designed experiments.* Chemometrics and Intelligent Laboratory Systems **36** 207 − 211.

[2] Atkinson A.C., Donev A.N (1992). *Optimum experimental designs.* Oxford: Oxford University Press.

[3] Ash H., Hedayat A (1978). *An introduction to design optimality with an overview of the literature.* Comm. Statist. Theory Methods. **7**, 1259 − 1325.

[4] Box G.E.P., Draper N.R (1971). *Factorial designs, the* $\left|F^{'}F\right|$ *criterion and some related matters.* Technometrics **13**, 731 − 742.

[5] Broudiscou A., Leardi R., Phan-Tan-Luu R (1996). *Genetic algorithm as a tool for selection of D-optimal design.* Chemometrics and Intelligent Laboratory Systems **35**, 105 − 116.

[6] Cook R.D., Nachtsheim C.J. (1980). *A comparison of algorithms for constructing exact D-optimum designs.* Technometrics **22**, 315 – 324.

[7] Dykstra O. (1971). *The augmentation of experimental data to maximize* $\left|X'X\right|$. Technometrics **13**, 682 – 688.

[8] Govaerts B., Sanchez R.P. (1992). *Construction of exact D optimal designs for linear regression models using genetic algorithms.* Belgian Journal of Operations Research, Statistics and Computer Science **1-2**, 153 – 174.

[9] Goldberg D.E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison Wesley.

[10] Heredia-Langner A., Carlyle. W.M., Montgomery D.C., Borror C.M., Runger G.C. (2003). *Genetic algorithms for the construction of D-optimal designs.* Journal of Quality Technology **35** 28-46.

[11] Johnson M.E., Nachtsheim C.J. (1983). *Some guidelines for constructing exact D-optimal designs on convex design spaces.* Technometrics **25**, 271 – 277.

[12] Keifer J., Wolfowitz J. (1959). *Optimum designs in regression problems.* Ann. Math. Statist. **30**, 271 – 294.

[13] Lawler E.L. (1976). *Combinatorial optimization: networks and matroids.* New York: Holt, Reinhart and Winston.

[14] Mitchell T.J. (1974). *An algorithm for the construction of D-optimal experimental designs.* Technometrics **20** 211 – 220.

[15] Montepiedra G., Myers D., Yeh A.B. (1998). *Application of genetic algorithms to the construction of exact D-optimal designs.* Journal of Applied Statistics **6**, 817 – 826.

[16] Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. (1992). *Numerical recipes in C, second edition: the art of scientific computing*, Cambridge: Cambridge University Press.

[17] Poland, J.A., Mitterer K., Knödler A., Zell (2001). *Genetic algorithms can improve the construction of D-optimal experimental designs.* In: Mastorakis N. (Ed.), Advances In Fuzzy Systems and Evolutionary Computation, WSES 2001, 227 – 231.

[18] Silvey S.D (1980). *Optimum design*. London: Chapman & Hall.

[19] Syslo M.M., Deo N., Kowalik J.S (1983). *Discrete optimization with pascal programs.* Engelwood Cliffs, NJ: Prentice-Hall.

[20] Wynn, H.P. (1970). *The sequential generation of D-optimum experimental designs.* Ann. Math. Statist. **41** 1655- 1664.

Acknowledgement: The authors would like to thank Professor Hans Nyquist of Stockholm University for his review and useful suggestions on this paper.

Address: A. Aziz Ali, Clinical Information Management, AstraZeneca R&D Södertälje, S-151 85 Södertälje, Sweden

E-mail: `Abdul.Aziz.Ali@AstraZeneca.com`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

GEOMETRY OF LEARNING IN MULTILAYER PERCEPTRONS

Shun-ichi Amari, Hyeyoung Park and Tomoko Ozeki

Key words: Learning, information geometry, singular statistical model, neural networks.

COMPSTAT 2004 section: Neural networks and machine learning.

Abstract: Neural networks provide a good model of learning from statistical data. Multilayer perceptron is regarded as a statistical model in which a nonlinear input-output relation is realized. The set of multilayer perceptrons forms a statistical manifold in which learning and estimation takes place. This is a Riemannian manifold with Fisher information metric. However, such a hierarchical model includes algebraic singularities at which the Fisher information matrix degenerates. This causes various difficulties in learning and statistical estimation. The present paper elucidates the structure of singularities, and how they influence the behavior of learning. The paper describes a new learning algorithm, named the natural gradient method, to overcome such difficulties. Various statistical problems in singular models are discussed, and the models selection criteria (AIC and MDL) are studied in this framework.

1 Introduction

The multilayer perceptron is a simple feedforward model of neural networks, which transforms input signals to output signals nonlinearly. It is a universal approximator in the sense that any nonlinear transformation is approximated sufficiently well by an adequate perceptron, if the number of hidden units is large.

In order to realize a good approximator, examples of input-output pairs are used. On-line learning receives a series of training examples one by one, and modifies the parameters of a perceptron each time when one example is given. Usually old examples are then discarded. Batch learning keeps all the examples and modifies the parameters in a batch mode.

A multilayer perceptron is an old model of learning machines, and the error-correcting learning algorithm was established for simple perceptrons in the sixties. Amari proposed a gradient descent learning method for multilayer perceptrons [2], which was rediscovered later independently and became popular under the name of backpropagation [23].

We study the set of multilayer perceptrons of a fixed architecture, which include a number of modifiable parameters called connection weights and biases. The set forms a multi-dimensional manifold, where all these parameters play a role of admissible coordinate systems. Learning takes place in the manifold, drawing a trajectory.

It is important to study the geometrical structure of the manifold which we call a neuromanifold. We will show by statistical considerations that the neuromanifold is Riemannian whose metric is specified by the Fisher information matrix [3]. Moreover, it has a pair of affine connections [4], but we do not state them in the present paper. The neuromanifold has singularities where the Fisher information (or the Riemannian metric) degenerates [5]. This is an interesting statistical model, because the conventional Cramér-Rao paradigm excludes such a model, assuming the existence and non-degeneracy of the Fisher information matrix as regularity conditions.

It is known that the convergence speed of a multilayer perceptron is usually very slow. This is caused by the Riemannian character in particular by its degeneracy, because the conventional backprop learning method does not take the Riemannian nature into account. The state of a network is often attracted by singularities by the conventional algorithm and takes long time before getting rid of them. The natural gradient learning algorithm was proposed to overcome the flaw, which takes the Riemannian gradient instead of the conventional gradient [3]. We show in the present paper the reasons why it works so well. We also explain an adaptive method of implementing the natural gradient [8]. In the case of the squared error criterion under Gaussian noises, the natural gradient algorithm coincides with the adaptive version of the Gauss-Newton method, but they differ in more general models (see [17]).

We finally study the dynamics of learning and the nature of singularities and explain the reason why learning trajectories are attracted to and stay longer in a neighborhood of singularities. The statistical analysis of behaviors of estimators in a neighborhood of singularities is another important problem to be studied. We show the conventional criteria of model selection such as AIC and MDL fail in this case.

2 Neuromanifold of multilayer perceptrons

Let us consider a multilayer perceptron of h hidden units and one output unit, which receives n-dimensional input signals $\boldsymbol{x} = (x_1, \cdots, x_n)$. A hidden unit, say the i-th unit receives $\boldsymbol{x}$, and takes its weighted sum, resulting in the potential

$$u_i = \boldsymbol{w}_i \cdot \boldsymbol{x}. \tag{1}$$

Here $\boldsymbol{w}_i = (w_{i1}, \cdots, w_{in})$ is the weight vector of the i-th unit, and we neglect the bias term for the sake of simplifying the notation. The unit calculates the nonlinear transform of the potential, $\varphi(u)$, where the nonlinear function is

$$\varphi(u) = \tanh u. \tag{2}$$

The final units collects all the outputs of the hidden units, and its final output is their weighted sum, if no noise intervenes. We put

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum v_i \varphi\left(u_i\right) = \sum v_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right), \tag{3}$$

where we summarized all the modifiable parameters in a large vector $\boldsymbol{\theta} = (v_1, \cdots, v_m \,;\, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_m)$. The final output of the perceptron is disturbed by noise, so that

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \varepsilon, \tag{4}$$

where we assume that ε is a Gaussian noise with mean 0 and variance 1. Therefore, its behavior is represented by the conditional probability of y given $\boldsymbol{x}$,

$$p\left(y|\boldsymbol{x}, \boldsymbol{\theta}\right) = c \exp\left\{-\frac{1}{2}\left(y - f(\boldsymbol{x}, \boldsymbol{\theta})\right)^2\right\} \tag{5}$$

or the joint probability distribution of $(\boldsymbol{x}, y)$,

$$p(y, \boldsymbol{x}; \boldsymbol{\theta}) = q(\boldsymbol{x}) p\left(y|\boldsymbol{x}, \boldsymbol{\theta}\right) \tag{6}$$

where $q(\boldsymbol{x})$ is the probability distribution of inputs $\boldsymbol{x}$. The set of all the perceptrons is a manifold called a neuromanifold M where $\boldsymbol{\theta}$ plays the role of the coordinate system. Each point of the neuromanifold corresponds to the probability distribution (5) or (6).

3 Fisher information matrix and the Riemannian metric

The Fisher information matrix G is given by

$$G(\boldsymbol{\theta}) = E\left[\nabla \log p(y, \boldsymbol{x}; \boldsymbol{\theta}) \nabla \log p(y, \boldsymbol{x}; \boldsymbol{\theta})^T\right] \tag{7}$$

which is further calculated as

$$G(\boldsymbol{\theta}) = E\left[\nabla f(\boldsymbol{x}, \boldsymbol{\theta}) \nabla f(\boldsymbol{x}, \boldsymbol{\theta})^T\right], \tag{8}$$

where E denotes expectation, $\nabla = (\partial/\partial\theta_i)$ is the gradient and T denotes transpose of a vector. Let us define the square of the distance between two nearby perceptrons whose parameters are $\boldsymbol{\theta}$ and $\boldsymbol{\theta}+d\boldsymbol{\theta}$. Information geometry gives the squared distance by the quadratic form

$$ds^2 = d\boldsymbol{\theta}^T G(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{9}$$

This is the Riemannian metric, where the Fisher information metric is used as the Riemannian metric tensor [19]. This is the only invariant metric to be introduced in the manifold of probability distributions.

Given a (large) number N of independently generated input-output pairs $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)$, the maximum likelihood estimator (or any other first order efficient estimator) satisfies the Cramér-Rao bound. Hence, the distance is large when two perceptrons are well separated in the sense that their estimation can be done precisely. However, different from the ordinary statistical model, the neuromanifold includes points at which the Fisher information degenerates and its inverse diverges. This is related to the unidentifiability of network parameters.

4 Identifiability of perceptrons and singularity

The behavior of a perceptron is invariant under the following two operations [10]:

1. Change of signs of v_i and $\boldsymbol{w}_i$ at the same time.

2. Permutation of the hidden units, which causes permutation of the weight vectors $\{\boldsymbol{w}_i\}$ and the output weight $\{v_i\}$ at the same time.

This causes the following unidentifiability:

1. When $v_i = 0$ or $\boldsymbol{w}_i = 0$, the behavior is the same whatever value $\boldsymbol{w}_i$ or v_i takes.

2. When $\boldsymbol{w}_i = \boldsymbol{w}_j$ (or $\boldsymbol{w}_i = -\boldsymbol{w}_j$), the behavior is the same when

$$v_i + v_j = v_i' + v_j' \qquad \left(v_i - v_j = v_i' - v_j'\right) \tag{10}$$

holds for two perceptrons $\{\boldsymbol{w}_i, v_i\}$ and $\{\boldsymbol{w}_i, v_i'\}$.

We call the set

$$C = \{\boldsymbol{\theta}|\ v_i \left|\boldsymbol{w}_i\right| = 0 \quad \text{or} \quad \boldsymbol{w}_i = \pm \boldsymbol{w}_j\} \tag{11}$$

the critical set on which unidentifiability takes place. The Fisher information degenerates on the critical set, because the unidentifiability implies that the estimation error does not converge to 0 even when N goes to infinity. Hence the statistical model is non-regular, and the Riemannian metric is singular. See also [7], [8], [9], [15]; [24], [27].

Let us introduce the equivalence relation $\approx$, by which two perceptrons with different parameters are equivalent when their input-output behaviors are the same. Then the set

$$\tilde{M} = M/\approx \tag{12}$$

includes algebraic singularities and dimensions are reduced on the critical set. The conventional theory of statistical estimation does not hold in a neighborhood of singularities.

5 Natural gradient learning algorithm

Let

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\} \tag{13}$$

be the set of input-output examples, which we call the training set. Here we assume that the examples are generated independently by using the true perceptron whose parameters are given by $\boldsymbol{\theta}_0$.

Given the training set, we want to obtain the estimated parameters $\hat{\boldsymbol{\theta}}$ which is closest to the true one. The performance of the estimator $\hat{\boldsymbol{\theta}}$ is measured by the generalization error, which is the expectation of the squared error for a new example $(\boldsymbol{x}, y)$,

$$L\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{2} E\left[\left\{y - f\left(\boldsymbol{x}, \hat{\boldsymbol{\theta}}\right)\right\}^2\right]. \tag{14}$$

The conventional on-line learning algorithm uses the gradient of the instantaneous error at time t,

$$\nabla e\left(\boldsymbol{x}_t, y_t, \hat{\boldsymbol{\theta}}_t\right) = \frac{1}{2}\nabla\left\{y_t - f\left(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_t\right)\right\}^2 \tag{15}$$

to update the current parameters $\hat{\boldsymbol{\theta}}_t$ to the new one,

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - c\nabla e. \tag{16}$$

The gradient of a function is believed to be the steepest direction of change. This is true only when the coordinate system $\boldsymbol{\theta}$ is orthonormal in a Euclidean space. The steepest direction of e is given by

$$\tilde{\nabla} e = G^{-1}\nabla e \tag{17}$$

in a Riemannian space, where G^{-1} is the inverse of the Riemannian metric matrix.

Amari [3] proposed to use the Riemannian gradient for learning,

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - c\,G^{-1}\nabla e, \tag{18}$$

which is called the natural gradient method. The natural gradient method is proved to give an Fisher efficient estimator, even though examples are used only once when they are observed, and then discarded.

The performance of the natural gradient method is largely different from the conventional method, when the Riemannian structure is very different from the Euclidean one. It will be seen that this is indeed the case with multilayer perceptrons, because they include singularities where the Riemannian metric degenerates.

It is known that the learning trajectory is often trapped in the so called plateaus, at which the parameters change so slowly, and it takes long time to get rid of. The statistical physical approach made it clear that the parameters are once attracted to the critical set of the neuromanifold, so that the set becomes plateaus of learning [25], [21], [18]. Rattray, Saad and Amari [20] analyzed the dynamics of the natural gradient learning method, and showed that it has an idealistic characteristic for avoiding plateaus. See also [14].

6 Implementation of natural gradient—Adaptive natural gradient method

In order to implement the natural gradient method, one needs to use the inverse G^{-1} of the Fisher information matrix. However, it is in general difficult to calculate the Fisher information matrix, because it uses the expectation with respect to the unknown distribution $q(\boldsymbol{x})$ of inputs. Moreover, it is computationally heavy to invert the matrix G when the number of parameters is large.

Amari, Park and Fukumizu [8] proposed an adaptive method to obtain an estimate of the inverse of the Fisher matrix. It is an iterative method, and the estimate $\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_t\right)$ is calculated by

$$\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_t\right) = (1+c')\,\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_{t-1}\right) - c'\tilde{\nabla} f_t \left(\tilde{\nabla} f_t\right)^T, \tag{19}$$

where $f_t = f(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$ and c' is another learning constant which may depend on t. One should choose c and c' carefully. By using this estimate $\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_t\right)$, we can obtain the update rule of the adaptive natural gradient method of the form,

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t - c\,\hat{G}^{-1}\left(\hat{\boldsymbol{\theta}}_t\right)\nabla e, \tag{20}$$

Park, Amari and Fukumizu [17] generalized the idea to be applicable to more general cost functions.

7 Dynamics of learning in the neighborhood of the critical set

In order to see the dynamics of learning, let us consider the special case of perceptrons consisting of two hidden units. Let us consider the set $Q(\boldsymbol{w}, v)$

$$Q(\boldsymbol{w}, v) = \{\boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w}, v_1 + v_2 = v\} \tag{21}$$

which is a part of the critical set. This corresponds to the set of all the perceptrons which have only one hidden unit, where the weight vector is $\boldsymbol{w}$ and the output weight is v. Let the true parameters be $\boldsymbol{\theta}_0 = \{\boldsymbol{w}_1, \boldsymbol{w}_2, v_1, v_2\}$, where $\boldsymbol{w}_1 \neq \boldsymbol{w}_2$ so that it needs two hidden units.

Let $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{w}}, \bar{v})$ be the best perceptron with one hidden unit that approximates the input-output function $f(\boldsymbol{x}, \boldsymbol{\theta}_0)$ of the true perceptron. Then, all the perceptrons of two hidden units on the line:

$$\boldsymbol{w}_1 = \boldsymbol{w}_2 = \bar{\boldsymbol{w}}, \quad v_1 + v_2 = \bar{v} \tag{22}$$

corresponds to the best approximation by one hidden unit perceptron. Let us transform the two weights as

$$\boldsymbol{w} = \frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2), \quad \boldsymbol{u} = \frac{1}{2}(\boldsymbol{w}_1 - \boldsymbol{w}_2). \tag{23}$$

Then, the derivative of $L(\boldsymbol{\theta})$ along the line is 0, because all the perceptrons are equivalent along the line. The derivative in the direction of changing $\bar{\boldsymbol{w}}$

and $\bar{v}$ are zero, because they are the best approximator. The derivative in the direction of $\boldsymbol{u}$ is again 0, because the perceptrons having $\boldsymbol{u}$ is equivalent to that having $-\boldsymbol{u}$ that is derived by changing the two hidden units. Hence the line forms critical points of the cost function. This implies that it is very difficult to get rid of it once the parameters are attracted to $Q(\bar{\boldsymbol{w}}, \bar{v})$.

Fukumizu and Amari [12] calculated the Hessian of L. When it is positive definite, the line is really attracting. When it includes the negative eigenvalues, the state is escaping in these directions eventually. They showed that, in some cases, a part of the line is really attracting in some region, while it is really a saddle having directions of escape (although the derivative is 0). In such a case, the perceptron is once truly attracted to the line, and stays inside the line fluctuating around it because of random noise until it finds the place from which it can escape from the line. This is clearly a plateau.

This explains the plateau phenomenon. In order to show why the natural gradient works well, we need to evaluate the natural gradient in the neighborhood of the critical points. We can then prove that the natural gradient has a large magnitude in the neighborhood of the critical set, so that the plateau phenomena will disappear. Computer simulations confirm this observation.

8 Estimation and testing in the neighborhood of the critical set

The Fisher information matrix degenerates on the critical set. Therefore, the Cramér-Rao paradigm cannot be valid in the neighborhood of the critical set. Let us consider the statistical test

$$H_0 \; : \; \boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{24}$$

against

$$H_1 \; : \; \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \tag{25}$$

The likelihood ratio statistics is given by

$$\lambda = \frac{1}{2} \log \frac{\sum p\left(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right)}{\sum p\left(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0\right)}, \tag{26}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator. When the true point $\boldsymbol{\theta}_0$ is a regular point, that is, it is not in the critical region, the mle (maximum likelihood estimator) is asymptotically subject to the Gaussian distribution with mean 0 and the variance-covariance matrix $G^{-1}(\boldsymbol{\theta}_0)/N$, where N is the number of observations. In such a case, the log likelihood-ratio statistics is expanded in the Taylor series, giving

$$\lambda = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^T G^{-1}(\boldsymbol{\theta}_0) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right). \tag{27}$$

Hence this is due to the χ^2-distribution of the degrees of freedom equal to the number k of parameters. Its expectation is

$$E[\lambda] = \frac{k}{N}. \tag{28}$$

However, when the true distribution $\boldsymbol{\theta}_0$ lies on the critical set, the situation changes. The Fisher information matrix degenerates, and G^{-1} diverges, so that the expansion is no more valid. The expectation of the log likelihood estimator is asymptotically written as

$$E[\lambda] = \frac{c(N)}{N} k \tag{29}$$

where the term $c(N)$ takes various forms depending on the nature of singularities. Fukumizu [11] showed that

$$c(N) = \log N \tag{30}$$

in the case of multilayer perceptrons under a certain condition. In the case of the Gaussian mixture,

$$c(N) = \log \log N \tag{31}$$

holds [13], [16].

Since the parameters are not identifiable, we cannot estimate the parameters when the true one is on the critical set. However, we can estimate its equivalence class, and the consistency holds with the order of $\sqrt{N}$. When the true one is close to the singular point, the estimation of the parameters suffers from similar difficulty. For fixed N, the variance of the estimators diverges in inversely proportional to the square of the distance from the critical set. We need a new framework to analyze such singular cases.

9 Bayesian estimator

The Bayesian estimator is used in many cases where an adequate prior distribution is assumed for the purpose of penalizing complex models based on data. It is empirically known that the Bayesian posterior distribution or its maximizer behaves well in the case of large scale neural networks. In such a case, one uses a non-zero smooth prior on the neuromanifold.

However, a smooth prior is not regular in the equivalence class $\tilde{M}$ of the neuromanifold, because a point in the equivalence class includes infinitely many equivalent parameters when it is in the critical point. This implies that the Bayesian smooth prior is in favor of singular points (perceptrons with a smaller number of hidden units) with an infinitely large factor. Hence the Bayesian method works well in such a case to avoid overfitting. One may use a very large perceptrons with a smooth Bayesian prior, and an adequate smaller model is selected.

The Bayesian estimator of singular models was studied by Watanabe [28], [29] by using the method of algebraic geometry, in particular Hironaka's theory of resolution of singularity and Sato's formula in the theory of algebraic analysis.

10 Model selection

In order to obtain an adequate model, one should select a good class of models based on data, that is, one should determine the number of hidden units. This is the problem of model selection. AIC, BIC and MDL have been widely used as criteria of model selection.

AIC [1] is the criterion to minimize the generalization error. The model that minimizes

$$\mathrm{AIC} = \text{training error} + \frac{k}{N} \tag{32}$$

is selected by this criterion. This is derived from the asymptotic statistical analysis, where the mle estimator $\hat{\boldsymbol{\theta}}$ is subject to the Gaussian distribution asymptotically.

MDL [22] is the criterion to minimize the length of encoding the observed data by using a family of parametric models. It is given asymptotically by the minimizer of

$$\mathrm{MDL} = \text{training error} + \frac{\log N}{2N} k \tag{33}$$

The Bayesian criterion BIC [26] gives the same criterion as MDL.

However, in the case of multilayer perceptrons, the neuromanifold of perceptrons with a smaller number of hidden units are included in that with a larger number, but the former is the critical set of the larger neuromanifold. Therefore, the maximum likelihood estimator (or any other efficient estimators) is no more subject to the Gaussian distribution even asymptotically. Model selection is required when the estimator is close to the critical set, and hence the validity of AIC and MDL fails to hold. One should evaluate the log likelihood-ratio statistics more carefully in such a case [6].

There have been reported many computer simulations of applications of AIC and MDL. Sometimes AIC works better, while MDL does better in other cases. Such confusing reports seem to be given rise to by the difference of regular and singular models and also the different nature of singularities.

11 Conclusions

Multilayer perceptrons are popular nonlinear models for nonlinear regression analysis of observed data. A class of perceptrons is specified by the number of hidden units, and a smaller class is included in a larger class. A class of multilayer perceptrons forms a manifold named the neuromanifold, where modifiable parameters play the role of the coordinate system.

The neuromanifold is a Riemannian space, where the Fisher information matrix plays the role of the Riemannian metric. A remarkable point is that it is singular in the sense that the Riemannian metric degenerates on a subset of the manifold, in which the neuromanifold of a smaller hidden units are embedded.

We proposed the natural gradient learning method, which takes the Riemannian nature into account. It works well because it avoids the plateaus existing in to the critical set corresponding to the neuromanifold of a smaller number of hidden units.

Conventional statistical analysis assumes the existence and non-degeneracy of the Fisher information matrix. However in the case of multilayer perceptrons, as well as other similar hierarchical models, the singularity is unavoidable in its nature. The criteria of model selection such as AIC and MDL fail their validity under such circumstances.

The present paper reviews such aspects of learning in neural networks, which require a new statistical analysis of singular models. Geometry will be useful for this purpose.

References

[1] Akaike H. (1974). *A new look at the statistical model identification.* IEEE Trans. Automatic Control AC-19, 716 – 723.

[2] Amari S. (1965). *Theory of adaptive pattern classifiers.* IEEE Trans. Elect. Comput. EC-16, 299 – 307.

[3] Amari S. (1998). *Natural gradient works efficiently in learning.* Neural Computation **10**, 251 – 276.

[4] Amari S., Nagaoka H. (2000). *Information geometry.* AMS and Oxford University Press, New York.

[5] Amari S., Ozeki T. (2001). *Differential and algebraic geometry of multilayer perceptrons.* IEICE Trans., E84-A, 31 – 38.

[6] Amari S. (2003). *New consideration on criteria of model selection.* Neural Networks and Soft Computing (Proceedings of the Sixth International Conference on Neural Networks and Soft Computing), L. Rutkowski and J. Kacprzyk (eds.), 25 – 30.

[7] Amari S., Ozeki T., Park H. (2003). *Learning and inference in hierarchical models with singularities.* Systems and Computers in Japan **34** (7), 701 – 708.

[8] Amari S., Park H., Fukumizu K. (2000). *Adaptive method of realizing natural gradient learning for multilayer perceptrons.* Neural Computation **12**, 1399 – 1409.

[9] Amari S., Park H., Ozeki T. (2002). *Geometrical singularities in the neuromanifold of multilayer perceptrons.* Advances in Neural Information Processing Systems,T.G. Dietterich, S. Becker, and Z. Ghahramani (eds.) **14**, 343 – 350.

[10] Chen A.M., Liu H., Hecht-Nielsen R. (1993). *On the geometry of feedforward neural network error surfaces.* Neural Computation **5**, 910 – 927.

[11] Fukumizu K. (2003). *Likelihood ratio of unidentifiable models and multilayer neural networks.* The Annals of Statistics **31** (3), 833 – 851.

[12] Fukumizu K., Amari S. (2000). *Local minima and plateaus in hierarchical structures of multilayer perceptrons.* Neural Networks **13**, 317 – 327.

[13] Hartigan J.A. (1985). *A failure of likelihood asymptotics for normal mixtures.* Proc. Barkeley Conf. in Honor of J. Neyman and J. Kiefer **2**, 807 – 810.

[14] Inoue M., Park H., Okada M. (2003). *On-line learning theory of soft committee machines with correlated hidden units - Steepest gradient descent and natural gradient descent -*. J. Phys. Soc. Jpn **72** (4), 805 – 810.

[15] Kůrková V., Kainen P.C. (1994). *Functionally equivalent feedforward neural networks.* Neural Computation **6**, 543 – 558.

[16] Lin X., Shao Y., (2003). *Asymptotics for likelihood ratio tests under loss of identifiability.* The Annals of Statistics **31** (3), 807 – 832.

[17] Park H., Amari S., Fukumizu K. (2000). *Adaptive natural gradient learning algorithms for various stochastic models.* Neural Networks **13**, 755 – 764.

[18] Park H., Inoue M., Okada M. (2003). *Learning dynamics of multilayer perceptrons with unidentifiable parameters.* J. Phys. A: Mathe. Gen. **36** (47), 11753 – 11764.

[19] Rao C.R. (1945). *Information and accuracy attainable in the estimation of statistical parameters.* Bulletin of the Calcutta Mathematical Society **37**, 81 – 91.

[20] Rattray M., Saad D., Amari S. (1998). *Natural gradient descent for online learning.* Physical Review Letters **81**, 5461 – 5464.

[21] Riegler P., Biehl M. (1995). *On-line backpropagation in two-layered neural networks.* J. Phys. A; Mathe. Gen. **28**, L507 – L513.

[22] Rissanen J. (1978). *Modelling by shortest data description.* Automata **14**, 465 – 471.

[23] Rumelhart D.E., Hinton G.E., Williams R.J. (1986). *Learning internal representations by error propagation.* In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group (eds.), Parallel distributed processing (Vol. **1**, 318 – 362), Cambridge, MA:MIT Press.

[24] Rüger S.M., Ossen A. (1995). *The metric of weight space.* Neural Processing Letters **5**, 63 – 72.

[25] Saad D., Solla A. (1995). *On-line learning in soft committee machines.* Phys. Rev. E **52**, 4225 – 4243.

[26] Schwarz G. (1978). *Estimating the dimension of a model.* The Annals of Statistics **6**, 461 – 464.

[27] Sussmann H.J. (1992). *Uniqueness of the weights for minimal feedforward nets with a given input-output map.* Neural Networks **5**, 589 – 593.

[28] Watanabe S. (2001a). *Algebraic analysis for non-identifiable learning machines.* Neural Computation **13**, 899 – 933.

[29] Watanabe S. (2001b). *Algebraic geometrical methods for hierarchical learning machines.* Neural Networks **14** (8), 1409 – 1060.

Address: S. Amari, T. Ozeki, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Satitma, 351-0198, Japan
H. Park, Dept. of Computer Science, College of Natural Science, Kyungpook National University, Sangyuk-dong, Buk-gu, Daegu, 702-701, Korea

E-mail: {amari,tomoko}@brain.riken.jp, {hypark}@knu.ac.kr

VISUAL DATA MINING FOR QUANTIZED SPATIAL DATA

Amy Braverman and Brian Kahn

Key words: Massive data sets, cluster analysis, multivariate visualization.

COMPSTAT 2004 section: Applications.

Abstract: In previous papers we've shown how a well known data compression algorithm called Entropy-constrained Vector Quantization (ECVQ; [3]) can be modified to reduce the size and complexity of very large, satellite data sets. In this paper, we discuss how to visualize and understand the content of such reduced data sets. We developed a Java tool to facilitate this using simple multivariate visualization, and interactively performing further data reduction on user selected spatial subsets. This enables analysts to compare reduced representations of the data for different regions and varying spatial resolutions. The ultimate aim is to explain physically observed differences, trends, patterns and anomolies in the data.

1 Introduction

This work came about because of challenges posed by NASA's Earth Observing System (EOS). EOS is a long-term data collection program for studying climate change, its consequences for life on Earth, and effects of human activities on it. The centerpieces of EOS are three satellites, Terra, Aqua and Aura. Terra and Aqua are already in orbit, and Aura is due for launch in 2004. Each carries a suite of instruments that collect massive amounts of observational data; so massive that it is difficult to take full advantage of them. Different instruments have different sampling strategies, resolutions, file naming conventions, and collect data about different physical processes. The information is provided to users in files corresponding to individual spacecraft orbits or parts of orbits, each of which can be very large, and must be stitched together properly to provide a global or even a regional picture. To make these data more accessible, NASA produces global summary data sets called Level 3 data products.

Traditionally, Level 3 products are simple maps of mean quantities and standard deviations at coarse spatial resolution, by month. In [2], we proposed methods for constructing nonparametric, multivariate distribution estimates to replace traditional maps. For instance, the Multi-angle Imaging SpectroRadiometer (MISR) aboard Terra collects data about clouds. A key goal is to better understand the spatial distribution of clouds since they have great influence on Earth's energy budget. The information MISR collects includes three variables seen at high resolution: scene albedo, height, and cloud presence indicator. Albedo is a measure of scene reflectivity measured roughly on a scale of zero to one. Scene height is measured in meters above the Earth's surface ellipsoid. The cloud indicator is a binary variable taking value one if the scene is cloudy, and zero otherwise. To summarize this information traditional Level 3 products are created by partitioning one month's data into spatial subsets corresponding to one degree latitude–longitude grid cells. Six maps are then produced: mean and standard deviation of albedo, mean and standard deviation of height, and mean and standard deviation of cloud indicator.

The Level 3 product we proposed regards each triplet of albedo, height and cloud indicator as a three-element vector, and uses ECVQ to cluster data each grid cell. We report a set of cluster representatives, the number of original data points belonging to each cluster, and within–cluster mean squared error, also called distortion. We call this a summary, or a compressed or quantized version of the grid cell's data. Figure 1 illustrates. For one grid cell it shows a three dimensional scatterplot of the original data in light gray. Positions of cluster representatives are shown by the embedded balls, and ball shading shows cluster population according to the color bar on the right. Two key features of the summary are that i) cluster representatives be centroids of cluster members, and ii) data vectors must be assigned to clusters with the nearest (euclidian distance) representatives. This ensures that mean squared error between grid cell data points and their representatives are at least locally minimized, and that representatives and mean squared errors resulting from aggregation to coarser resolutions will be properly preserved. Details of the algorithm like the one used to produce these summaries can be found in [1].

Starting with a monthly summary of MISR cloud data at one degree resolution, our challenge is to discover and understand how relationships among grid cell distributions change spatially, and over different resolutions. In other words, instead of examining spatial patterns of average behavior and variability only, we want to examine spatial patterns of other distributional characteristics such as the number of modes, presence of outliers, and nonlinear regressions. This requires interactively comparing summaries of different grid cells, and of aggregated spatial areas. Thus, we want to quickly visualize summaries, and construct summaries of summaries in hierarchical fashion. The main subject of this paper is the Java tool L3View, written to facilitate this.

2 L3View

The basic data structure underlying L3View is a 180 × 360 array of objects called L3Cell's. An L3Cell contains a variable–length vector of Cluster objects, with the number of objects depending on grid cell data complexity. A Cluster records a three-dimensional cluster representative, a cluster count, and a within–cluster mean squared error. L3View presents a map of the world, and when the user clicks on it the with the mouse, L3View translates the mouse position into geographic coordinates. L3View opens a separate

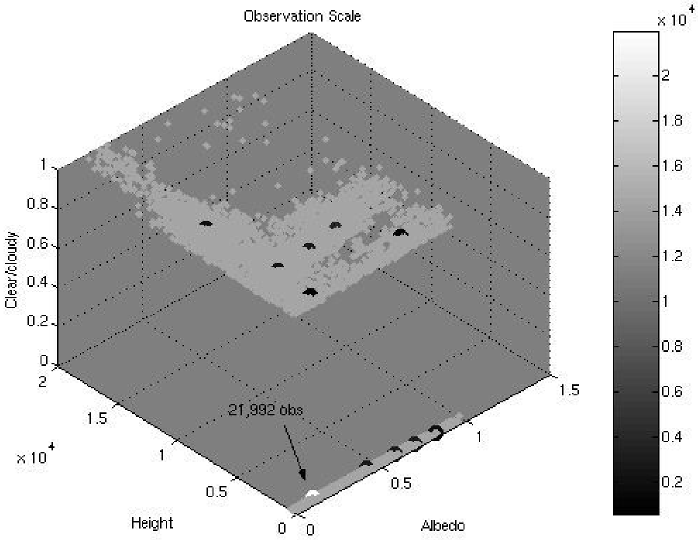


Figure 1: Three-dimensional scatterplot of MISR albedo, height and cloudiness data, in light gray, for a one degree grid cell in northern Oklahoma (southwest corner 38°N, 98°W) in March 2000. The embedded balls show the locations of cluster representatives. The ball colors show cluster populations using the gray-scale color bar on the right.

window, and displays a simple, multivariate visualization of the summary for the one degree grid cell at that location. Further, the user can select a subregion of the map with a rubberband box, and choose to summarize summaries of all grid cells within the box. This too is shown in a new window using the simple multivariate display.

2.1 Main map and control panel

The left panel of Figure 2 is a screenshot of the main L3View control panel. L3View uses Java Swing components to interact with users. The image displayed is constructed from information in the grid cells' clusters and combined with a GIF file containing continental outlines using Java image processing functions. L3View knows the position of the mouse in a graphic coordinate space native to the underlying Java object type, JPanel. L3View has methods to convert back and forth between this coordinate system, the 180×360 grid, and latitude and longitude. Latitude and longitude are displayed interactively as the mouse is moved, and the tool knows when the mouse is clicked, dragged, or leaves the map area. Clicking on a grid cell spawns a GraphView

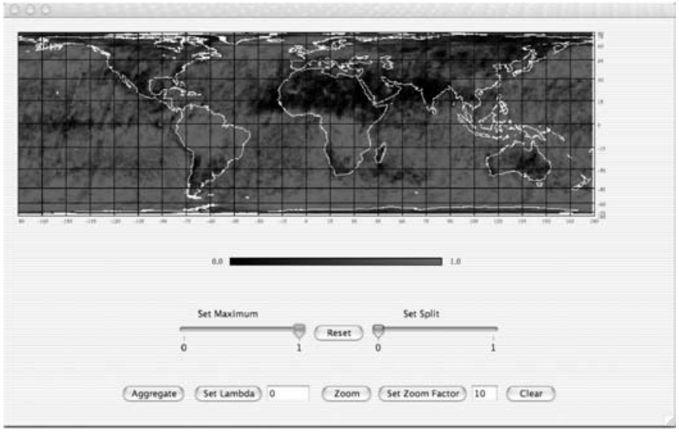


Figure 2: L3View main control panel showing MISR cloud fraction for March 2000.

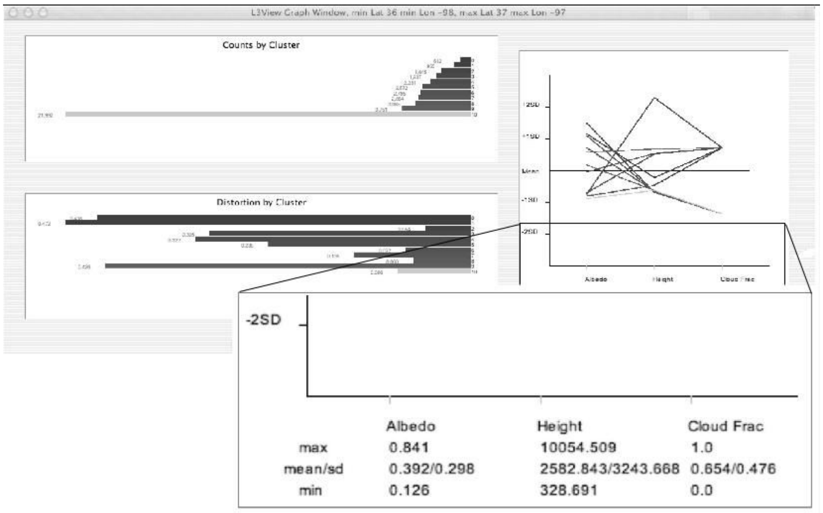
window, which contains three graphics for visualizing the clusters representing that grid cell's data.

If the mouse is used to isolate a rectangular geographic region with a rubberband box, L3View calculates the corresponding geographic and index limits. These are subsequently used in two cases. First, if the Zoom button is pushed, a new window containing a magnified image of the isolated area is spawned. Second, if the Aggregate button is pushed, all clusters from all grid cells inside the box are summarized, and the result is displayed in a new GraphView window. The lambda text box accepts user specified values for a parameter of the summarization algorithm that specifies how much data reduction is applied. This is discussed in Section 3.

Finally, the Set Maximum and Set Split sliders are used to study spatial patterns in the cumulative distribution function of the display variable. Set Maximum truncates the upper end of the color scale so that all grid cells with display values at or above the maximum display white. Set split is similar: all values above the split value are displayed white, while all values below the split value display in black.

2.2 The GraphView window

The GraphView window is a simple, three panel multivariate visualization of a set of clusters. A typical GraphView window is shown in Figure 3. It includes two bar plots and a parallel coordinate plot. The bar plots are two instances of the same class, instantiated to display cluster counts and mean squared errors (distortions). Each has one bar per cluster, and bars are sorted in order of increasing cluster count. Actual values of counts and distortions



	Albedo	Height	Cloud Frac
max	0.841	10054.509	1.0
mean/sd	0.392/0.298	2582.843/3243.668	0.654/0.476
min	0.126	328.691	0.0

Figure 3: A GraphView window showing the summary of MISR albedo, height and cloud indicator for the grid cell with southwest corner 36°N, 98°W over northern Oklahoma. A zoom-in view of the parallel coordinate plot legend is shown in superimposed box.

relative to the norms of corresponding cluster representatives, are shown at the bars' left edges. Though not apparent in these black and white figures, bars are colored using a scheme that transitions smoothly from blue to red with increasing count.

The parallel coordinate plot occupies the right side of the window. Each line plot shows the representative values of albedo, scene height, and cloudiness for a single cluster on scales normalized using the global means and standard deviations. These are shown at the bottom of the parallel coordinate plot area. Lines are color coded to match bars in the other two panels so users can see which representatives belong to which clusters. In addition, clicking on any bar or any line highlights the bars and line in all plot corresponding to that cluster.

GraphView windows are spawned to visualize a set of clusters, either for a single grid cell or when a set of grid cells are to be summarized collectively. In the latter case, one could simply display the entire collection of clusters, but that would become more unwieldy for large areas as more clusters are included. Complexity of the parallel coordinate plots could grow to the point where it is impossible to resolve individual lines. Therefore, distributions represented by cluster sets must be summarized before they are displayed. The next section describes the theoretical rationale for this.

3 Hierarchical aggregation and quantization

Braverman [2] described how entropy–constrained vector quantization (ECVQ; [3]) is modified to function as a data reduction tool for large, spatial data sets. The basic idea is to partition these data into one degree spatial subsets, and use ECVQ to cluster the subsets in a coordinated way. ECVQ is a randomized, iterative algorithm similar to K-means, except it minimizes the expected value of the penalized loss function,

$$L_\lambda(\boldsymbol{X}, \alpha(\boldsymbol{X})) = \|\boldsymbol{X} - q(\boldsymbol{X})\|^2 + \lambda \left[-\log \frac{N_{\alpha(\boldsymbol{X})}}{N} \right]. \tag{1}$$

$\boldsymbol{X}$ represents a randomly drawn observation from the empirical distribution of the grid cell's data. $\alpha(\boldsymbol{X})$ is an integer that specifies the id number of the cluster to which $\boldsymbol{X}$ is assigned, and $q(\boldsymbol{X})$ is the corresponding cluster centroid. N is the total number of data points in the grid cell, $N_{\alpha(\boldsymbol{X})}$ is the number of data points assigned to the same cluster as $\boldsymbol{X}$, and the logarithm is base two. λ is a fixed parameter that specifies how important the second term on the right in Equation (1) is. For K-means, one must specify K, the number of clusters a priori. For ECVQ, one must specify K, the maximum allowable number of clusters, and λ. The algorithm then determines the number of clusters and the assignment of data points to them. We added a final step in which each data point is subsequently reassigned to the cluster with the nearest euclidian distance representative, and the representatives updated again. This ensures cluster representatives are centroids of cluster members, and mean squared errors between data points and their representatives are minimized. In [1] we introduced a further modification of ECVQ in which $\boldsymbol{X}$ is a random variable having the distribution of $q(\boldsymbol{X})$ rather than the original empirical distribution of the data. In other words, we allow realizations to have unequal mass. That is precisely the situation in which we find ourselves when summarizing sets of clusters formed by combining multiple grid cells.

Consider Figure 4. It shows a schematic representation of a one degree spatial grid. Each grid cell contains a summary instantiated as an L3Cell object. Figure 4 also shows a two degree grid cell superimposed, and suppose we want to summarize the four, one degree L3Cell's inside. Let $\boldsymbol{X}_{1uv}$ be a random variable having the distribution of the summary for the one degree grid cell with southwest corner at row u and column v. Suppose this grid cell is the lower-left most grid cell in the light box in Figure 4, and denote the other three grid cells' random variables by $\boldsymbol{X}_{1(u+1)v}$, $\boldsymbol{X}_{1u(v+1)}$, and $\boldsymbol{X}_{1(u+1)(v+1)}$. At coarser, two degree resolution the light box is represented by $\boldsymbol{X}_{2uv}$,

$$\boldsymbol{X}_{2uv} = \sum_{i=0}^{1} \sum_{j=0}^{1} \boldsymbol{X}_{1(u+i)(v+j)} 1[V = v_{(u+i)(v+j)}],$$

with

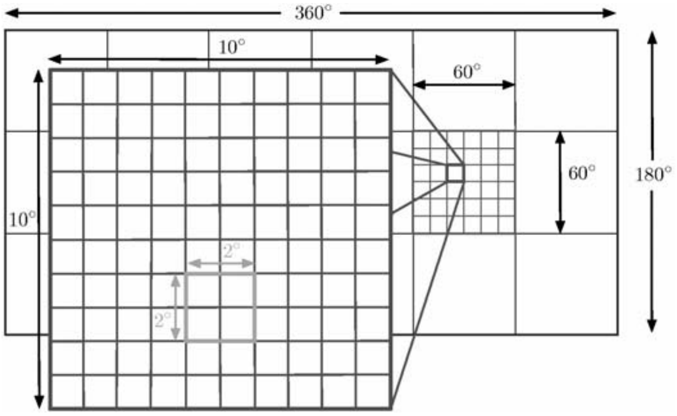


Figure 4: Schematic representation of a gridded map. The large rectangle represents a 180×360 array shown broken into 3×6 = 18, 60×60 arrays. Each of these is further subdivided into a 6 × 6 array. Each cell in the 6 × 6 array is a 10 × 10 arrangement of one degree grid cells. The lighter box illustrates how four one degree grid cells can make up a grid cell at coarser, two degree resolution.

$$P\left(V = v_{(u+i)(v+j)}\right) = \frac{N_{(u+i)(v+j)}}{\sum_{i=0}^{1}\sum_{j=0}^{1} N_{(u+i)(v+j)}},$$

and N_{ij} is the total number of data points represented by the summary of the corresponding grid cell. In other words, $\boldsymbol{X}_{2uv}$ is a mixture of $\boldsymbol{X}_{1uv}$, $\boldsymbol{X}_{1(u+1)v}$, $\boldsymbol{X}_{1u(v+1)}$, and $\boldsymbol{X}_{1(u+1)(v+1)}$ with weights equal to the proportions of the total count represented by $\boldsymbol{X}_{2uv}$ contributed by each one degree cell. The idea is illustrated on the left side of Figure 5, which shows the mixture distribution positioned directly above the four component distributions. Any nesting of fine-scale grid cells in a coarser grid can be represented in a similar way, and ensures mass, expectation, and mean squared error are all properly preserved between resolutions.

If data reduction were not a concern, we could proceed directly to visualizing mixture distributions like the middle layer in Figure 5. However, the greater the number of grid cells being aggregated, the greater the number of support points in the mixture, and the number of corresponding clusters. So, we compress the mixture distribution using a mass-weighted version of ECVQ described in [1], but implemented here in Java with the user specifying λ directly via the "Set Lambda" button and text box in the main control panel. K, the maximum number of clusters is nominally set to 10, and the default value of λ is zero, thus essentially implementing the K-means. If λ is changed to a positive value, the algorithm becomes ECVQ.

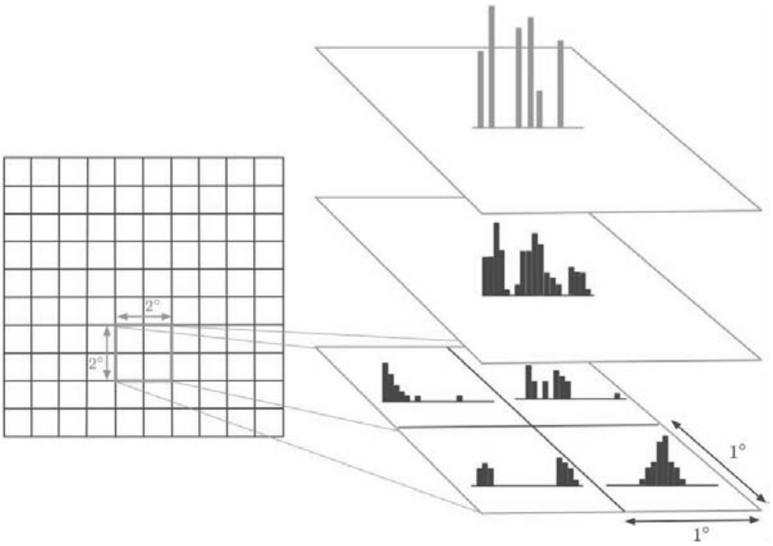
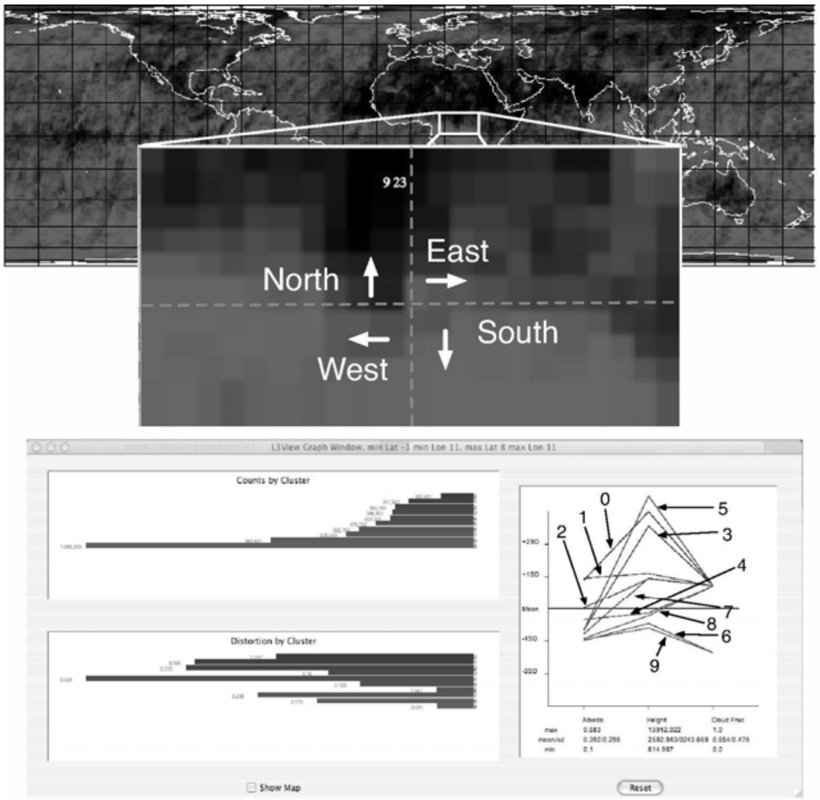


Figure 5: A hierarchy of distributions within a two degree spatial region. The bottom square on the right corresponds to the 2° × 2° area, and shows conceptual representations of cluster sets for constituent one degree grid cells as histograms. The middle layer on the right depicts the mixture distribution formed by the union of the cluster sets from the one degree cells. The top layer is the reduced distribution after summarization.

By first considering aggregated distributions for large areas, and then systematically summarizing subregions, we can begin to understand how the prevalence of various types of phenomena change spatially. The next section demonstrates how this can be done.

4 Visual data mining

As an example of how a scientist might use L3View for data exploration, we focus on an area in central Africa shown in Figure 6. The rectangular region extends from latitude 1°S to latitude 9°N, and from longitude 11°E to 31°E. The background L3View image shows cloud fraction. There is a clear difference between the northern and southern parts of this region, approximately demarcated by the horizontal dashed line in embedded, zoomed-in view. The southern area is very cloudy, and the northern area contains grid cells varying cloudiness. This is consistent with the climatological location of a persistent band of clouds called the Inter-Tropical Convergence Zone (ITCZ). The lower panel of Figure 6 shows the GraphView window of the summary of the entire 10 × 20 degree region.



	Albedo	Height	Cloud Frac
max	0.883	13912.022	1.0
mean/sd	0.392/0.256	2582.843/3243.868	0.654/0.476
min	0.1	614.597	0.0

Figure 6: Screenshots from L3View visualization of central Africa. Top: Main L3View map with embedded zoom of central Africa. Bottom: GraphView window for the entire, aggregated central area. All the parallel coordinate plots are annotated to show the cluster id numbers corresponding to the individual line plots.

The region contains 200 grid cells, with a total of 2,099 clusters. These represent 6,205,769 original MISR albedo–height–cloud indicator vectors. We begin by aggregating the whole region using the default value of $\lambda = 0$ and the number of clusters, K, set to 15. The resulting GraphView summary is shown in the lower panel of Figure 6. The figure is small making the graph labels difficult to see, but we can see from the bar chart of cluster counts that one cluster dominates in size. Using L3View interactively, we find that this is cluster 9, and it contains about 30 percent of the distribution's mass. Cluster 9 corresponds to one of two clear clusters, 6 being the other. Cluster 6 accounts for another eight percent of the distribution's mass. 9 and 6 have representatives with low albedo, low height, and are clear cloud indicators.

This is a dark, vegetated region of jungle. Areas to its north show significant numbers of low altitude, bright, clear scenes. This is the Sahara desert.

The remaining clusters have cloudy representatives, and form three subgroups. Clusters 8 and 4 constitute a subgroup with low albedo and below average height. Clusters 1, 2, and 7 form a second subgroup. Their heights are nearly one standard deviation above the mean, but their albedos range from nearly one standard deviation below to one standard deviation above average. The final subgroup is characterized by very large heights, two standard deviations above the mean at least. They too show a range of albedos similar to that of the second subgroup. These high clouds are likely the tops of thunderstorms prevalent in central Africa at this time of year, and the surrounding cloud formations. The first two subgroups are more mysterious. Clusters 1, 2, and 7 could be low and mid-level cumulus and stratus clouds. 8 and 4 are possibly dust, clear land surface misclassified as cloud, or simply dark, low clouds, as implied by the classification.

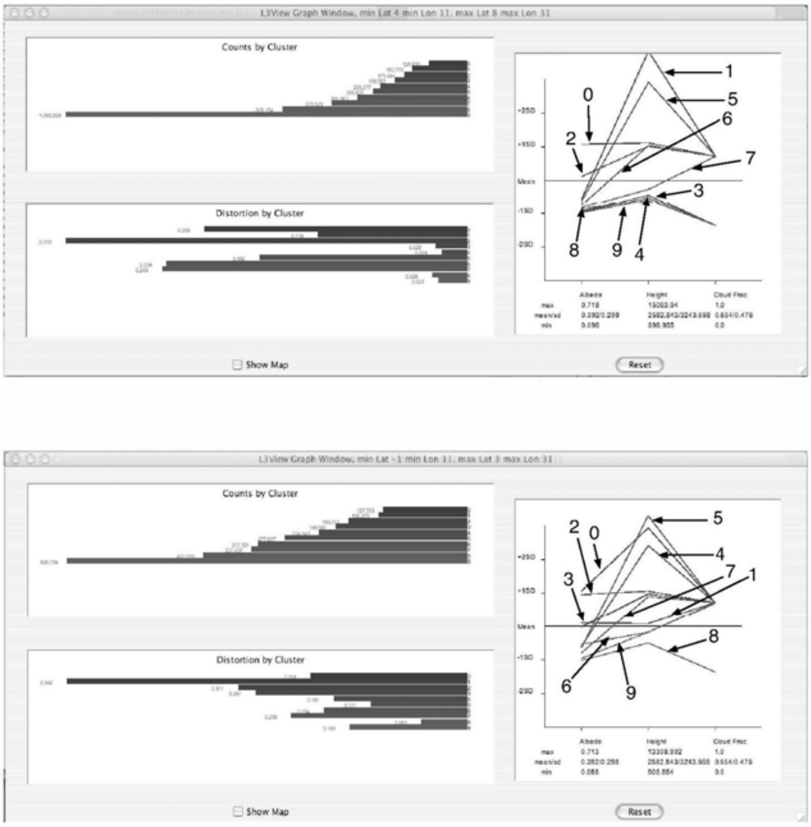
Noting the relatively sharp difference in cloud fractions between the northern and southern areas, we separately summarize them are shown in Figure 7. Signatures of southern region representatives look much like signatures for the region as a whole. The north's representatives also look roughly like those of the whole region except clusters similar to 0 and 4 are missing. The absence of clusters similar to cluster 0 in the lower panel is encouraging, since this cluster represents deep convective clouds. Corroborating sources indicate these are in fact absent in this region at this time.

These distributional differences are summarized in Table 1. Not surprisingly the joint distribution shows that the south is cloudier than the north. The fact that the south is dominated by low clouds while the north is dominated by mid-level clouds is less obvious but clear from the conditional distribution.

	Joint			Conditional	
Type	*North*	*South*	*Total*	*North*	*South*
Clear	0.330	0.064	0.394		
Low cloud	0.060	0.171	0.231	0.273	0.442
Mid-level cloud	0.094	0.114	0.207	0.426	0.295
High cloud	0.066	0.101	0.168	0.301	0.263
Total cloudy	0.220	0.386	0.606	1.000	1.000
Total	0.550	0.450	1.000		

Table 1: Joint and conditional distributions of cloud type/clear and location. Columns 2, 3, and 4 show the full, joint distribution. Columns 5 and 6 show the conditional distribution of cloud type given cloudy scene, and location.

The presence of low, dark clouds in both the north and the south at this time of year is something of a surprise. To see if these clouds can be


L3View Graph Window, min Lat 4 min Lon 11, max Lat 8 max Lon 31

	Albedo	Height	Cloud Frac
max	0.718	15063.94	1.0
mean/sd	0.392/0.298	2582.843/3243.688	0.654/0.479
min	0.096	596.955	0.0

L3View Graph Window, min Lat -1 min Lon 11, max Lat 3 max Lon 31

	Albedo	Height	Cloud Frac
max	0.713	13308.932	1.0
mean/sd	0.282/0.298	2582.843/3243.688	0.654/0.478
min	0.088	505.884	0.0

Figure 7: Top: GraphView window for the aggregated area in the northern half of the region (above the dashed line). Bottom: GraphView window for the aggregated area in the southern half of the region (below the dashed line).

attributed to specific areas, we subdivided the north and south regions into east and west. We found no distributional differences related to east-west division for either the north or south. We then investigated areas along the prominent clear/cloudy boundary in Figure 6, and contrasted them to areas away from the boundary. We did this separately for east and west, but none of these visualizations revealed definitive distributional differences. We are therefore reasonably confident that Table 1 tells a complete story.

5 Discussion

The example of the previous section is a small scale, simple example of one way we think L3View may be useful for exploring spatial summaries of satellite data. Guided by the background map in the main L3View window, we focused on an area of interest, and hierarchically examined the data distribution summaries. We discovered that, in addition to the cloud fraction differences apparent from the background image, there is also a difference in the types of cloud present in the northern and southern parts of the region. We will have to perform many more exercises like this one to gain confidence that summarized data have enough detail to be scientifically useful, and to gain experience interpreting physically what we see in L3View.

Two main computational issues are brought to light in this exercise. First, L3View's implementation of ECVQ/K-means is not fast enough to summarize large geographic regions in reasonable time. One would ideally like to summarize whole hemispheres in the same sort of hierarchical exploration performed here. Second, we have not yet made use of the tool's ability to summarize the same data for different values of λ's. We would like to look at data at different quantization resolutions as well as different spatial ones. We believe there is important information in how distributions collapse as greater data reduction is imposed. To achieve greater interactivity both these issues must be addressed. We look forward to working on these and other improvements to L3View as it matures. We also eagerly anticipate working with our geoscience colleagues to better understand the connection between global physical processes and their expression through rich, Earth Observing System data sets.

References

[1] Braverman Amy, Fetzer Eric, Eldering Annmarie, Nittel Silvia, Leung Kelvin (2003). *Semi-streaming quantization for remote sensing data.* Journal of Computational and Graphical Statistics **12**, 4, 759 – 780.

[2] Braverman Amy (2002). *Compressing massive geophysical datasets using vector quantization.* Journal of Computational and Graphical Statistics **11**, 1, 44 – 62.

[3] Chou P.A., Lookabaugh T., Gray R.M. (1989). *Entropy-constrained vector quantization.* IEEE Transactions on Acoustics, Speech, and Signal Processing, **37**, 31 – 42.

Acknowledgement: The authors would like to thank Eric Fetzer, Annmarie Eldering and Barbara Gaitley for their helpful comments. This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

Address: A. Braverman, Jet Propulsion Laboratory, California Institute of Technology, Mail Stop 169-237, 4800 Oak Grove Drive, Pasadena, CA 91109-8099

B. Kahn, Department of Atmospheric Science, UCLA, 405 Hilgard Avenue, Los Angeles, CA 90095

E-mail: Amy.Braverman@jpl.nasa.gov, kahn@atmos.ucla.edu.

GRAPHS FOR REPRESENTING STATISTICS INDEXED BY NUCLEOTIDE OR AMINO ACID SEQUENCES

Daniel B. Carr and Myong-Hee Sung

Key words: Letter sequences, graphics, layouts.

COMPSTAT 2004 section: Graphics.

Abstract: This paper develops coordinates and layouts for graphs that represent statistics indexed by repetitive letter sequences. The need for such graphics arises in a variety of applications. The examples in this paper concern sequences of nucleotides, such as AGTGGC, and sequences of amino acids.

1 Introduction

In contrast to maps that represent statistics indexed by geospatial coordinates, the development of graphics methodology for statistics indexed by repetitive letter sequences has been modest. One interesting exception is the sequence logo display [13] that can show a sequence of categorical frequencies. Statistical graphics methods for categorical data [7], [8] are relevant for relatively simple multivariate combinations but so far have seen little use in nucleotide and amino acid indexing examples.

Journal articles typically show short tables with one column giving the sequence of letters and one more column providing statistics. The rows are often sorted by one of the statistical columns. Both the one-dimensional linear ordering and the restriction to a modest number of rows reduce the opportunity to see patterns that may lead to new understanding. One-dimensional linear orderings produced by clustering, the first principal component, minimal spanning tree traversal, space filling curves or other methods do not exploit the human ability to see multivariate patterns based on 2-D and 3-D connectedness and proximity. Connectedness and proximity are among the most powerful of human perceptual grouping principles [18]. Thus this paper seeks to develop 2-D and 3-D coordinates for representing letter-indexed statistics.

The graphical design objectives include providing an overview along with interactive focusing and re-expression methods. For long sequences, combinatorics grow exponentially with sequence length and quickly lead to an overwhelming number of statistics. Overviews require substantial statistical summarization. The modest research here concerns developing representations for short sequences.

One approach not investigated here is the use of pixel oriented visualization [9]. It is possible to encode univariate statistics on all nucleotide sequences of length ten ($4^{10} = 1024 \times 1024$) in a pixel plot on a 1280×1024 monitor. Large high resolution prints will handle somewhat longer length sequences. Interactive pan and zoom methods can support layouts for longer sequences but showing all the values at once is problematic. The color of individual pixels is hard to identify with increased monitor resolution. The use of multiple monitors cannot keep up with the exponentially growing combinations.

Layout details are an issue. A convenient layout for a pixel plot may use lexicographic order for the first half (last half) of the sequence along the x-axis (y-axis). A following sectional on fractal coordinates provide another approach to layouts. In both cases indexing regularity help to keep the analyst oriented with interpreting the plots. However, indexing that is convenient for human memory may be poor at bringing out meaningful patterns. Maps often work well for showing geospatially-indexed statistics because geospatial attributes often have locally similar values. This applies to covariates as well as to the primary variables of interests. Proximity that reflects scientific relationships can be crucial to seeing meaningful patterns.

The layouts in this paper have limitations because they are primarily based on indexing regularity. However, the layouts provide some opportunities to rearrange letter order or axis placement either for perceptual simplification (such as reducing line crossings) or for incorporating physical/chemical properties (such as hydrophobicity) of the sequence constituents. Interestingly these two objectives can lead to the same display. Axes ordering problems are in general NP complete [1]. While many people prefer 2-D layouts, 3-D layouts not only allow better preservation of interpoint distances of higher dimensional points, they also provide more opportunities arranging axes. Thus the layout options are not as restrictive as might be assumed at first glance.

This paper develops three approaches to constructing coordinates while mentioning some alternatives along the way. Three different data sets motivate the development of the coordinates. Section 2 describes self-similar coordinates at different scales. Section 3 concerns self-similar coordinates at the same scale with focus on 3-D extension of parallel coordinates. The application shows cell statistics from a 4-D table. Section 4 illustrates the use of simple additive vector coordinates for showing all quadruples of amino acids (ignoring order). This approach can be useful despite some substantial overplotting problems. The section also hints at other 2-D layouts that avoid overplotting. Comments appear along the way about software and interactive tools used for rendering.

2 Self-similar coordinates, dimensionality, and fractals

The regularity of self-similar coordinates speeds learning and helps analysts devote short-term memory to other issues. A natural approach to developing multivariate coordinates encodes letters as integers and then uses Cartesian coordinate product sets with a coordinate for each position in the sequence. With A=1, C=2, G=3, and T=4 the sequence ATCG is located at (1, 4, 2, 3). The similar treatment of each position in the sequence and the same ordering of nucleotides for each axis motivates the description as a self-similar coordinate system. With coordinates in hand, multivariate glyphs can encode multivariate statistics associated with the sequence. The most immediate problem with this approach is that straightforward graphical representation of points is only available through three dimensions.

2.1 Rendering approaches and difficulties

There are many approaches to rendering multivariate data with more than three coordinates. As further mentioned in Section 4, nested coordinate plots provide one approach. Another approach encodes some coordinates as glyph features. For example with four coordinates the ray angle of a stereo-ray glyph can encode the value for the fourth coordinate [3]. Ray length and color can encode more coordinates. Good perceptual accuracy of extraction for the angle encoding and modest use of "ink" make glyph a good choice for revealing hyperplanes in 4-D data and other tasks. However in the current context all 3-D coordinate glyphs lose self similarity when rendering more than three coordinates.

Before developing 3-D coordinates to represent sequences longer than three, brief comments about limits and merits of 3-D graphics are appropriate. Many people prefer 2-D graphics to 3-D graphics. Common arguments for 2-D rendering are that people only see surfaces, occlusion is a problem in 3-D and that motion and/or binocular parallax depth cues are impossible or inconvenient to convey on a printed page. The position here is that many humans are endowed with the cognitive ability to see 3-D images based on motion and binocular parallax. They should be allowed to utilize this capability whenever it helps in dealing with difficult scientific challenges. Three dimensions provide a richer environment for conveying relationships and produce less distortion than 2D and 1D plots when scaling multivariate data into lower dimensions.

2.2 Weighted vector addition and fractals

Vector addition provides an enticing starting point for developing 3-D coordinates. For nucleotides, associate each letter with a vector from the origin to the vertices of a tetrahedron. Let A=(1,1,1), C=(1,-1,-1), G=(-1,1,-1), and T=(-1,-1,1). Then using of vector addition for each letter in a sequence produces a 3-D coordinate for representing the sequence. However, vector addition is commutative so all permutations of the same set of letters yield the same point. When the goal is to represent all hexamers (six letter sequences) the result should be $4^6 = 4096$ distinct plotting points but rather 84 points corresponding with the multinomial terms in multinomial $(\mathrm{A} + \mathrm{C} + \mathrm{G} + \mathrm{T})^6$.

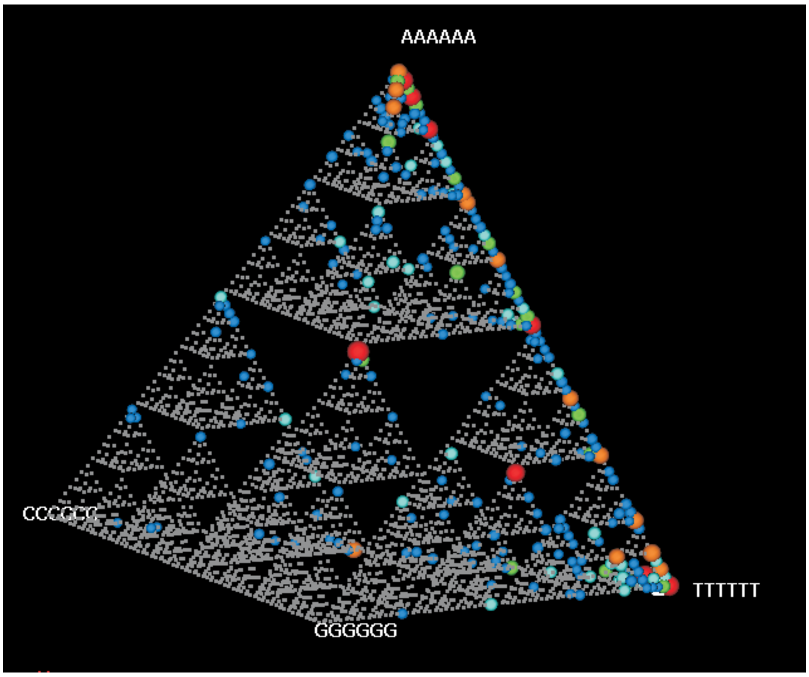


Figure 1: Fractal coordinates for nucleotide sequences. Sphere size (and color) show counts. Small rectangles show the plotting locations of low count spheres.

Weighted vector addition provides an approach that can produce unique points for plotting. Consider power of two weights $2^{(6-i)}/63$ where i is the position along the sequence and the weights sum to 1. The sequence ACGTTC then maps into the point (.555, .270, .206).

Rendering and rotating reveals the coordinates creating a Sierpinski Gasket similar to one shown in Mandelbrot [11]. The self-similarity provides means of decoding the indexing of a point based on its location. A point in the large tetrahedron toward the C attractor has C as the first letter. If a point within this C tetrahedron is as close as possible to the T attractor, then all the remaining letters are T. A tetrahedron zooming widget can reveal the sequence of conditioning letters. Similarly, zooming can be controlled by entering the first letters of the sequence.

A delight occurs when rotating the gasket. In some orthographic projections the appearance is a square lattice. Construction using a pair of coordinates indicated earlier makes this clear. However, this is not intuitive from looking at the gasket in other views. The 2-D layout is also self-similar and extends immediately to the 10-mer by 10-mer pixel display mentioned earlier.

The motivation for Figure 1 was an early effort to find transcription regulation docking sites for the Stanford yeast genes [5]. The study clustered genes based on their expression levels. This produced groups of seemingly co-regulated genes. For the genes in a group, the 300 (nucleotide)-letter regions upstream of the protein-coding regions of the genes were scanned with a sliding window of length six. This produced the basic statistics on the occurrence frequencies of the different hexamers encountered. The sphere glyphs in the plot encode counts using size and color. A glance reveals that most of the higher count hexamers appear along the AAAAAA to TTTTTT edge. Relatively little was known about transcription regulation when the Stanford yeast data was first made available and the statistics for Figure 1 was produced. Today many transcription regulation sites of various lengths have been identified and the regions as far as 800 nucleotides upstream are relevant for some genes. The plots could be improved by obtaining better data and by highlighting the hexamers known to be associated with transcription regulation.

Different kinds of software can produce figures similar to Figure 1. With a little work most standard statistical software can produce projected static views. Software that provides rotation, filtering and brushing, such as Xgobi and CrystalVision, provide better visualization environments. Efforts to produce multilayer 3-D visualization methodology similar to GIS software led to the development of software called GLISTEN (geometric letter-indexed statistical table encoding). GLISTEN supports point and path layers that are used in the graphs below.

Efforts to extend the fractal layout to amino acids were not very successful. One generalization used the 20 face centers of the icosahedron as attractors and adapted the weights so the clouds of points associated with each of the 20 attractors would be separated. Only three letter sequences were shown to restrict the view to $20^3 = 8000$ points. The high density of points for the smallest scale icosahedra and the occlusion, partly due to more points, made this layout less desirable.

2.3 Connecting coordinates for longer sequences

Paths that connect points can represent longer letter sequences. For example, a path through three points in Figure 1 can represent nucleotide sequences 18 letters long. The use of paths is advantageous since occlusion problems do not grow too quickly. Experimenting with translucent triangles and tetrahedra for showing triples and quadruples were not successful due to inability to see through much more than two layers.

Paths can encode statistics using path thickness and color. The path direction also needs to be encoded unless the sequence is explicit or intended to be reversible. There are limitations to this approach. If the data contains a large number of sequences, overplotting precludes providing an overview. Filtering widgets can then help to cope with very large databases. A second

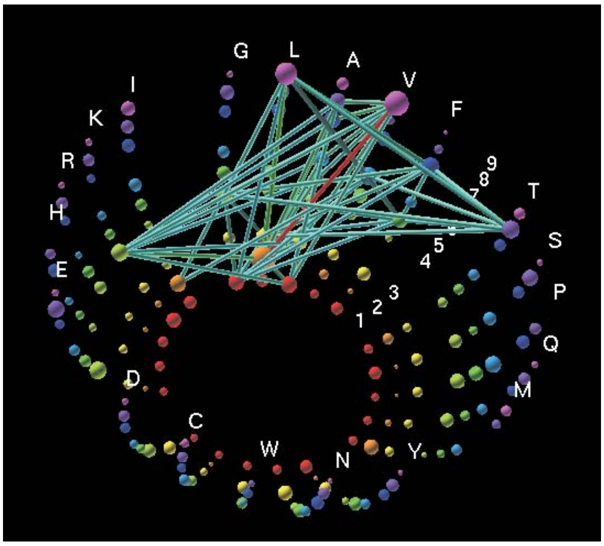


Figure 2: Capless hemisphere coordinates. Sphere size shows counts from 1-D table. Path thickness, color, and filtering enable focus on high counts from 2-D tables.

limitation is that a single coordinate system with each point representing a sequence of p letters does not accommodate sequence lengths that are not a multiple of p. A third issue that especially applies to fractal coordinates is that the apparent distance between paths is heavily influenced by the subset of coordinates receiving heavy weight. Fourth, there can be ambiguity when multiple paths go through the same point. Still, such displays can often turn up meaningful patterns that are otherwise missed.

3 Parallel coordinates escape the plane

Parallel coordinates (PC) plots also provide self-similar representations. Analysts are increasingly using these plots to show multivariate data and to provide interactive input in a multivariate context. A limitation of parallel coordinates is the lack of a natural way to connect non-adjacent axes. Figure 2 shows a capless hemisphere coordinate system that partially addresses the problem. The coordinate system encodes the 20 natural amino acids as longitude and nine positions along a sequence as latitude. The gap create perceptual groups that facilitate focusing on subsets. The path connecting L2 and L9 and the path between L2 and its neighbor L3 do not overlap due to the hemisphere curvature. The 3-D setting with curvature means the axes are not longer parallel, but there is a simple mapping to parallel coordinates.

The data motivating Figure 2 comes from a database of peptides [2], or in this case amino acid sequences of length 9 known to bind to important immune molecules called HLA. This binding reaction is crucial in initiating the recognition by the human body of peptides from 'foreign' sources such as viruses or cancer. When a T-cell finds the peptide-HLA combination on the target cell surface, it activates coordinated processes for the purpose of clearing the infected cells. The immune system is an incredibly complex 'search-and-destroy' system. Autoimmune diseases are examples of false positives, where the immunity is mistakenly directed toward normal tissue. An example of an immune system false negative is the inability to detect and clear certain infections. Bioinformatic prediction of peptides from pathogenic proteomes such as HIV has emerged as a valuable tool in vaccine development and cancer immunotherapy [16].

The data used in the example concerns peptides binding to the HLA A-2 molecule (a specific form among many different genetic versions of this HLA). Most of the binding peptides listed were 9-mers. Typical statistics would be just the counts of amino acid for each of the nine positions. Such can be represented by sequence logo displays or as sequence of bar charts. The sphere size and color (when not shown in gray level) in Figure 2 convey this information just as effectively. The paths in Figure 2 encode counts from the nine choose two (36) two-way (20×20) tables. A filtering widget removed all but the highest count cells. The counts are encoded by the line color (when not shown in gray level) based on a color ramp. The layout also requires some sorting considerations. Putting the hydrophobic amino acids adjacent to each other reduces line crossing.

Since paths can have more than one line segment, the capless hemisphere coordinate framework can also represent statistics from higher dimensional tables. The three-segment paths in Figure 3 show high count (frequency) cells from the nine choose four (126) four-way ($20 \times 20 \times 20 \times 20$) tables (see also [10]). The lowest count path shown goes through L2, A7, A8, and V9 as might be expected from the 1-D margin counts. There are large counts for L2 and V9. The rest of the high frequency paths shown go through G4. This is not apparent from the 1-D margins counts.

4 Additive vector coordinates and overplotting

When the ordering of letters in a sequence is not important, the additive vector coordinate approach mentioned in Section 2 is more appropriate. A 3-D tessellation application provides such data [15], [17]. The data arise from tessellating the space of proteins based on the location of their backbone Carbon alpha atoms. The Delauney tessellation yields tetrahedra indexed by the associated amino acid residues at the vertices. The ordering of the amino acids for a tetrahedron is not considered important. There are $(20 - 1 + 4)$ choose 4 or 8855 distinct tetrahedra (see [6] for a discussion of classical occupancy problems.)

Figure 4 provides an additive vector coordinate example with the vec-

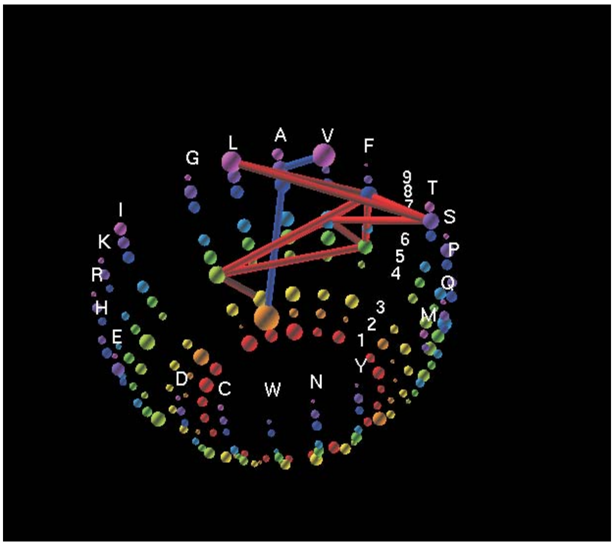
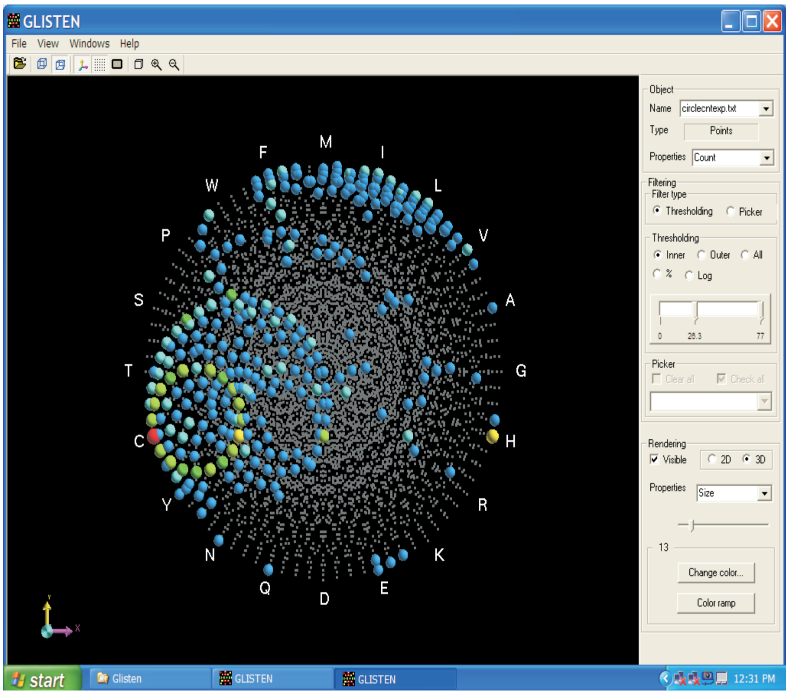


Figure 3: Paths with three segments show frequently occurring quadruples, i.e. high counts from 4-way tables.

tors point to twenty points evenly spaced around a circle. Again point size and color encode the counts, and dynamic filtering has removed low count tetrahedra. High count patterns jump out. One is a circle involving three Cysteins and one each of the amino acids.

While Figure 4 reveals a lot of structure, there are at least three problems worth noting. First, over some 2000 points are overplotted. This is partly related to symmetric construction with equal angles between vectors. Second, zooming reveals many points that to close together to see closer in a overview. Third, for over 4000 points involving four distinct amino acids the connection between plotting location and the indexing is almost impossible to untangle without mouseovers. Figure 4 is mostly useful for points in an outer annulus of the circle.

There are several possibilities for alternative views. It is possible to show a statistic encoded by color in a casement display of 204 = 160000 points [12]. In this example the casement display is a 20 × 20 layout of 20 × 20 matrices. However it remains desirable to study plots with a factor of 18 less points. The 8855 points can be placed in a 4-D simplex. (See also pentagonal numbers [4].) Space prohibits showing a layout composed of two-dimensional slices of the simplex. There is also a layout in the plane for all tetrahedra with 2 or more of the same amino acid. While this layout involves duplicates the regularity makes the layout easier to study. Such a layout can provide a starting point for drilling down to conditioned views of the 1-1-1-1 combinations.



Object
Name: circlecntexp.txt
Type: Points
Properties: Count

Filtering
Filter type: ● Thresholding ○ Picker

Thresholding
● Inner ○ Outer ○ All
○ % ○ Log
0 26.3 77

Picker
□ Clear all ☑ Check all

Rendering
☑ Visible ○ 2D ● 3D
Properties: Size
13

Figure 4: Vector addition coordinates. Sphere size (and color) encode statistics for protein tetrahedra. Small rectangles show plotting locations of low count spheres.

5 Closing remarks

Just as map projections have been devised to serve different purposes, coordinates systems for encoding statistics can be developed to serve different purposes. A worthy goal is to develop coordinates systems with a regularity that minimizes memory burdens and helps analysts keep oriented with respect to the coordinates. A tension arises when one desires to show complex relationships faithfully in some abstract sense while keeping the relationships cognitively accessible. In many cases there are no easy answers and the graphics are a compromise. Still, analysts can make discoveries from imperfect graphs. It is worthwhile to consider graphics that lean toward the cognitive accessibility and work toward incorporating as much scientific structure as possible options. Accessible graphics enable analysts to look and, if they look, they have a chance to see.

References

[1] Ankerst M., Berchtold S., Keim D.A. (1998). *Similarity clustering of dimensions for an enhanced visualization of multidimensional data.* Proceedings IEEE Symposium on Information Visualization, IEEE Computer Society, Washington, 51 – 60.

[2] Brusic V., Rudy G., Harrison L.C. (1998). *MHCPEP, a database of MHC-binding peptides: update 1997.* Nucleic Acids Research **26** (1), 368 – 371.

[3] Carr D.B., Nicholson W.L. (1988). *EXPLOR4: a program for exploring four?dimensional data.* Dynamic Graphics for Statistics, W.S. Cleveland and M.E. McGill (eds.), Wadsworth, Belmont, California, 309 – 329.

[4] Conway J.H., Guy R.K. (1996). *The book of numbers.* Copernicus Books, Inc. New York.

[5] DeRisi J.L., Iyer V.R., Brown P.O. (1997). *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science **278**, 680 – 686.

[6] Feller W. (1968). *An introduction to probability theory and its applications.* Third Edition. John Wiley and Sons. New York.

[7] Friendly M. (1999). *Extending mosaic displays: marginal, conditional, and partial views of categorical data.* Journal of Computational and Graphical Statistics **8** (3), 373 – 395.

[8] Hoffman H. (2000). *Exploring categorical data: interactive mosaic plots.* Metrika **51**, 11 – 26.

[9] Keim D.A. (1996). *Pixel-oriented visualization techniques for exploring very large databases.* Journal of Computational and Graphical Statistics, 58 – 77.

[10] Lee J.P., Carr D., Grinstein G., Kinney J., Saffer J. (2002). *The next frontier for bio- and cheminformatics visualization.* T-M Rhine, Ed. IEEE Computer Graphics and Applications, 6 – 11.

[11] Mandelbrot B.B. (1983). *The fractal geometry of nature.* W.H. Freeman and Company.

[12] Munson P.J, Singh R.K. (1997). *Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignments.* Protein Science **6**, 198 – 201.

[13] Schneider T.D., Stephens R.M. (1990). *Sequence logos: a new way to display consensus sequences.* Nucleic Acids Research **18**, 6097 – 6100.

[14] Segal M. Cummings R., Hubbard A. (2001). *Relating amino acid sequences to phenotype: analysis of peptide binding data.* Biometrics V57, 632-643

[15] Singh R.K., Tropsha A., Vaisman I.I. (1996). *Delaunay tessellation of proteins: four body nearest neighbor propensities of amino acid residues.* J. Computational. Biology. **3** (2), 213 – 221.

[16] Sung M.-H., Simon R. (2004). *Genome-wide conserved epitope profiles of HIV-1 predicted by biophysical properties of MHC binding peptides* J. Computational Biology **11** (1), 125 – 145.

[17] Vaisman I.I., Tropsha A., Zheng W. (1998). *Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis.* Proceedings of the IEEE Symposia on Intelligence and Systems, 163 – 168.

[18] Ware C. (2000). *Information visualization, perception for design.* Morgan Kaufman Publishers, New York.

Acknowledgement: The work was supported by NSF cooperative grant # 9983461. Duoduo Liao and Yanling Liu implemented versions of GLISTEN.

Address: D.B. Carr, George Mason University, Dept. AES MS 4A7, George Mason University, Fairfax, VA 22030, USA
M.-H. Sung, Biometric Research Branch, National Cancer Institute Room 8146 6130 Executive Plaza Rockville, MD 20852

E-mail: `dcarr@gmu.edu, sungm@mail.nih.gov`

COMPSTAT'2004 Symposium

MATRIX VISUALIZATION AND INFORMATION MINING

Chun-Houh Chen, Hai-Gwo Hwu, Wen-Jung Jang, Chiun-How Kao, Yin-Jing Tien, ShengLi Tzeng, Han-Ming Wu

Key words: Dimension free visualization, effect-ordered data display, generalized association plots, heat-map, matrix map, relativity of statistical graph, re-orderable matrix, sufficient statistical graph.

COMPSTAT 2004 section: Data visualization.

Abstract: Many statistical techniques, particularly multivariate methodologies, focus on extracting information from data and proximity matrices. Rather than rely solely on numerical characteristics, matrix visualization allows one to graphically reveal structure in a matrix. This article reviews the history of matrix visualization, then gives a more detailed description of its general framework, along with some extensions. Possible research directions in matrix visualization and information mining are sketched. Color versions of figures presented in this article, together with software packages, can be obtained from `http://gap.stat.sinica.edu.tw/`.

1 Introduction

The seminal work of Tukey [34] states a basic principle of Exploratory Data Analysis (EDA):

> *It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.*

In his concept of EDA, Tukey would allow the data to speak for themselves prior to adoption of any standard assumptions or formal modeling. Much of this preliminary work can be achieved with graphics-oriented tools - the box and whisker plot, the scatterplot, etc.

Many visualization techniques have now been developed to assist us in looking at data. Much of this literature has been devoted to dimension reduction: multidimensional scaling [11], projection pursuit [20], self-organizing maps [24], and sliced inverse regression [26]. These techniques are very useful for exploring data structure when the number of variables is of moderate size and when structure is not too complex. Yet, with striking advances in computing, communication, and high-throughput biomedical instruments, the number of variables can easily reach tens of thousands, and the need for practical data analysis remains. Dimension reduction tools generally lose effectiveness when it comes to visual exploration for information structure embedded in very high dimensional data sets. On the other hand, matrix visualization, integrated with computing, memory, and display, has great potential for visually exploring the structure that underlies massive and complex data sets.

A brief review of matrix visualization is provided in Section 2. Section 3 introduces the general framework of matrix visualization and some extensions of it appear in Section 4. An outline of possible research directions is sketched in Section 5 and there are concluding remarks in Section 6.

2 Matrix visualization (MV)

The technique of matrix visualization discussed in this article is deemed dimension-free [3]. The only limitation of size for a given data set is the resolution of a computer display or the size of the printing device used. Matrix visualization is not a new technique but CPU speed, memory size, and display capability of modern computers give it a brand-new platform.

Given a data matrix, $X = [x_{ij}]_{n \times p}$, the idea is to graphically present all numerical values, x_{ij}, in a matrix map. With small values of n and p, this map can be the numerical matrix itself. Bertin [1] considered re-orderable or permutation matrices, Hartigan [19] introduced the direct clustering of a data matrix, Lenstra [25] and Slagel et al. [33] linked the traveling-salesman problem and shortest spanning path to matrix reordering, Wegman [35] proposed the idea of a color histogram, while Minnotte and West [30] implemented The Data Image package. The Cluster and TreeView packages by Eisen et al. [12] are probably the most popular visualization packages because of the wide application to gene expression profiling for cDNA microarray experiments. Heatmaps with classification tree partitioning are commonly used to summarize stock market structure (`http://www.smartmoney.com/marketmap/`). On the other hand, Carmichael and Sneath (taxometric maps [2]) and Ling [27] dealt with visualization of a proximity matrix, $D = [\rho_{ij}]_{n \times n}$, whose elements are measures of degree of relationship between pairs of a set of n objects. Murdoch and Chow [31] used elliptical glyphs to represent correlation matrices while Friendly [15] called them Corrgrams. Church and Helfman [9] developed Dotplot for exploring self-similarity in millions of lines of text and code. Chen [4], [5], [6] and Chang et al. [3] integrated visualization for data and proximity matrices into the framework of generalized association plots (GAP). We use matrix visualization (MV) to refer to all the aforementioned terminologies.

The basic principle of MV, then, is to effectively present a complete data or proximity matrix on a computer display or printout. We use the psychosis disorder data set described in Hwu et al. [22] to illustrate the framework of an MV analysis. Thirty-three Positive and Negative Syndrome Scales, PANSS [23], for one hundred and sixty-three schizophrenic patients were used in the original analysis. We randomly sampled 40 patients with 17 symptoms for our illustration. PANSS symptoms are rated on an ordinal scale from 1 (normal) to 7 (severe), but we treat them on a continuous scale for simplicity. Psychiatrists [22] have addressed three fundamental issues: the grouping structure among the symptoms, the clustering structure of patients, and the general behavior of every patient-cluster in each symptom-group. These three issues are closely related to the three major pieces of information contained in any multivariate data set: the linkage amongst n subject points in the p-dimensional space; the linkage between p variable vectors in the n-dimensional space; and the interaction linkage between the sets of subjects and variables. Factor analysis and clustering related methods are commonly applied to answer the first two issues, but there is no general technique for studying the interaction effects for subjects and variables. With appropriate presentation (permutation and color/shape coding) and integration for the raw data and proximity matrices, MV can be used to effectively display all three pieces of information with many types of data formats and sampling schemes.

3 Four components of matrix visualization

Most existing MV methods deal with data and proximity matrices separately. Chen [6] integrated them in a framework of generalized association plots with four major components. We focus our introduction on these four components.

3.1 Presentation of raw data matrix and selection of proximity matrices

The first task of MV is to convert a numerical matrix into a matrix map with a color dot (symbol) representing each entry. Then each row vector, call it a patient's symptom profile, is converted into a horizontal color band and every column vector, a symptom's patient distribution, is replaced by a vertical color strip. The information in a numeric matrix is thus comprehensively displayed in a matrix map (Figure 1a).

3.1.1 Color spectrum and variable transformation The selection of a suitable color spectrum is crucial in an MV analysis [1], [13]. With the PANSS symptoms, we need only find a color spectrum capable of expressing its ordinal nature, a rainbow spectrum in Figure 1a for example. For measurements with bi-directional structure, such as the logarithmic gene expression profiles used for cDNA microarray experiments, an integration of two monotonic color spectrums is needed (Figure 2).

For the PANSS example, the rainbow spectrum is the same for each symptom since they share the same scale. When variable structure becomes more complicated, transformation of variables may be necessary before the MV can effectively present the data structure. In particular, in order to make simultaneous visualization of multiple variables in MV meaningful, it is essential to standardize variables (sometimes for subjects) with different scales. When outliers are present, the relationship between outlying observations and

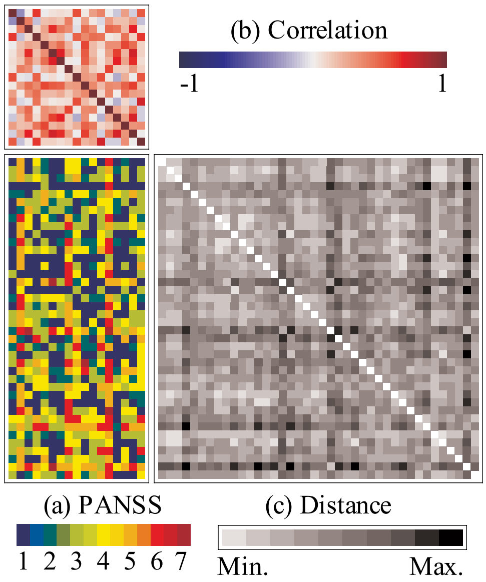


Figure 1: Presentation of the PANSS data matrix with proximity matrices.

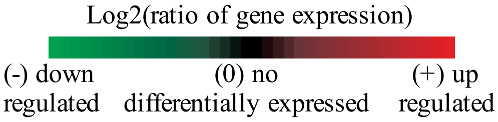

Log2(ratio of gene expression)
(-) down regulated
(0) no differentially expressed
(+) up regulated

Figure 2: Bi-directional color spectrum for gene expression profile.

the main body of the data set can exhaust the color spectrum and only the relative structure of outliers and the main body can be observed [3], [28]. A logarithm or similar transformation can be applied to variables or proximities to diminish the outlier effect. Transformation of variable (symptom), also termed the column conditioned transformation, is commonly practiced. Row (patient) and matrix conditioned transformations are used from time to time.

3.1.2 Selection of proximity measures The second important task is to identify appropriate measurements for representing between-variables and between-subjects association. The importance of this choice for proximity matrices is two-fold: it is used to directly assess the strength of variable-interaction with subject-relationship; it will be used to permute the raw data or proximity matrix. Suitable color spectrums are also needed to project numerical proximity matrices to matrix maps. Euclidean distance is used in the psychosis disorder example to measure the patient-to-patient dissimilarity so a uni-directional grey-scale spectrum (Figure 1c) is adopted, while a bi-directional blue-white-red spectrum is applied to illustrate the between symptoms correlation coefficients (Figure 1b). Appropriate transformations of variables before computing proximities, and of proximity measurements directly, are necessary for both numerical and visual considerations.

3.2 Seriation of proximity matrices and raw data matrix

Although Figure 1a has already converted three numerical matrices into MV format, not much information can be obtained from it since variables and subjects are randomly permuted in these matrices. In order for a statistical graph (including MV) to reveal structure embedded in the data being displayed it is necessary to place objects with similar (different) properties at closer (distant) positions in the graph. Chen [6] called this concept the relativity of a statistical graph. The corresponding mechanism in an MV display is to identify the best seriations (permutations) for the two proximity matrices. Friendly and Kwan [16] used a similar term, effect-ordered data display, for ordering information in general visual displays.

3.2.1 Robinson matrix Criteria are necessary to identify "good" seriations (permutations) for a given matrix. Seriation is a data analytic tool for finding a permutation or ordering of a set of objects using a data matrix (symmetric or asymmetric). Hubert [21] and Marcotorchino [29] discussed the seriation problem from the aspect of problem setting, methodology and algorithms. Two major considerations in permuting a matrix are global pattern identification and local cluster formation, see [6] and [16] for more detailed discussions. The global and local criteria usually conflict with each other unless the embedded structure has a simple uni-dimensional pattern. One familiar global criterion is the Robinson form [32], [8], and [6]. A matrix is said to be a Robinson (anti-Robinson) matrix if the elements in its rows and columns do not increase (decrease) when moving horizontally or vertically away from the main diagonal (Figure 3a). A permuted Robinson matrix is a pre-Robinson matrix (Figure 3b & c). A Robinson matrix satisfies both the relativity condition [6] and the effect-ordered requirement [16] and it is optimized for global and local criteria. A Robinson proximity matrix may be obtained from a raw data matrix with a monotonic structure for variables and/or for subjects. For our PANSS data, this monotonic structure can be a positive-neutral-negative pattern. Robinson form takes relative positions of any two columns and rows into consideration, so it focuses more on global consideration than on local structure. Visualization is usually more globally focused to reflect our ability to perceive the information in a given graph and organize it into a global pattern. There is no practical algorithm which optimizes the Robinson criterion because of its computing complexity.

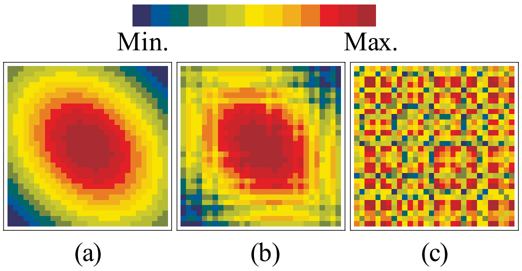


Figure 3: Robinson (a) and pre-Robinson (b & c) matrices.

The Iris data [14] is used to compare the performance of seriations with several commonly used sorting algorithms. The target proximity matrix is the Euclidean distance matrix of the 150 iris flowers on four variables. Using the convergence properties of a series of Pearson correlation matrices, Chen [6] proposed an elliptical seriation which identifies permutations with very good near-Robinson structure (Figure 4g & h).

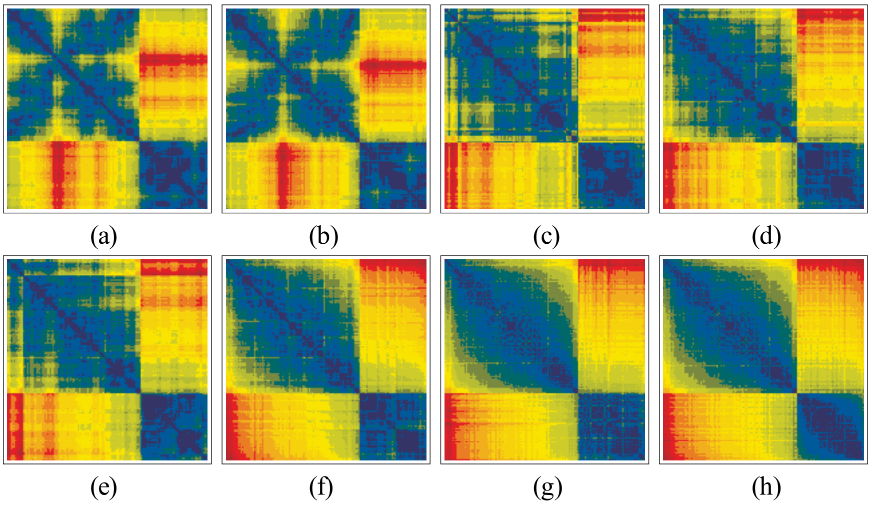


Figure 4: Permuted Euclidean distance maps for Iris data with eight seriation algorithms: (a) farthest insertion spanning; (b) nearest insertion spanning; (c) single linkage tree; (d) complete linkage tree; (e) average linkage tree; (f) GAP rank-one tree; (g) GAP rank-two ellipse; (h) GAP double ellipse.

3.2.2 Tree seriation Most of the seriation algorithms try to optimize some local properties. Travelling-salesman [25] and minimal spanning path [33] algorithms orient toward local optimization. Figure 4a & b show the distance matrix of the Iris data sorted by two minimal spanning algorithms. The hierarchical cluster analysis with a tree-architecture (dendrogram) is the most popular permutation fulfilling criteria for local optimization [27], [12] (Figure 4c, d, & e). Relative positions of the terminal nodes in the final tree grown are employed as the permutation to sort the input objects. The two dendrograms in Figure 5 are generated from the correlation matrix for symptoms and the distance matrix for patients in Figure 1. The two proximity matrix maps are permuted accordingly. The data matrix is two-way sorted using the corresponding permutations. After the permutations, relativity for both subjects and variables is satisfied. That is, patients with similar symptom profiles are placed in closer rows while symptoms with comparable patient distributions correspond to columns nearby each other. Patient-clusters and symptom-groups can now be easily identified using the sorted proximity matrix maps with the tree-architectures. These pieces of information are usually summarized through factor analysis for variables and clustering analysis for subjects. The two-way sorted data map in Figure 5a is actually a condensed version of two sets of scatter-plot matrix with $C(17, 2) = 136$ and $C(40, 2) = 780$ pair-wise plots each for studying the interaction of patient-clusters and symptom-groups.

One fundamental problem arises in applying the dendrogram for matrix permutation. There are $n - 1$ intermediate nodes for a dendrogram with n terminal nodes. Each of these $n - 1$ intermediate nodes can be flipped independently resulting in 2^{n-1} possible final permutations for a given dendrogram. Figure 6 shows the results of different flipping mechanisms for intermediate nodes applied to the same tree-architecture (dendrogram) for a given correlation coefficient matrix. As can be seen, different flipping mechanisms result in totally different visual perceptions and grouping effects. Figure 6a is the only permuted matrix map with perfect scores on both local and global criteria such as the minimum spanning and anti-Robinson scores [6]. It is also possible to use external and internal references [17], [36] for identifying desired flipping patterns.

3.3 Partitions of permuted matrix maps

The next step after the matrices have been permuted is to identify clusters in the resulting maps. This is a constrained clustering problem since the variables and subjects are sorted and listed in a one-dimensional manner. The goal becomes one of finding partitioning points on the two permutations. For a matrix map sorted with a dendrogram, the dendrogram structure can help in identifying suitable cutting points. Two purple lines are drawn in Figure 7b & c to partition the two dendrograms (and matrix maps) into three symptom-groups coded in (red, green, and blue) and four patient-clusters

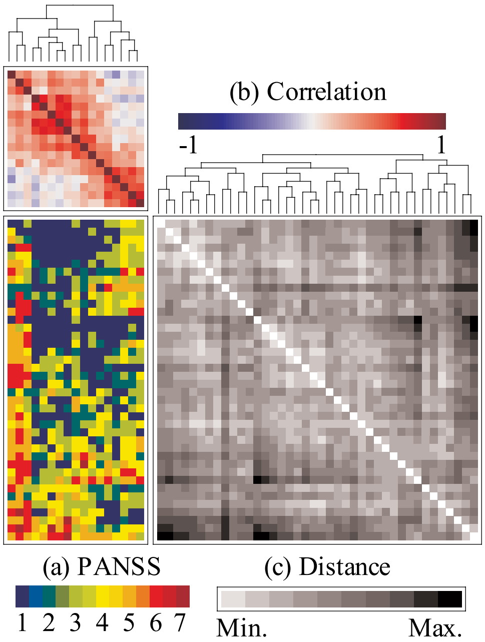


Figure 5: Proximity and raw data maps with dendrograms after permutation.

coded in (cyan, magenta, yellow, and grey). Without dendrograms, characteristics in a permuted data matrix (map) and in proximity matrices (maps) must be employed for identifying possible partitions. Chen [6] contains some discussion on matrix partition using a convergent sequence of Pearson correlation matrices.

3.4 Sufficient statistical graph

Symptoms and patients are partitioned into three groups and four clusters in Figure 7. These two partitions thus cut the correlation map, the distance map, and the raw data map into nine, sixteen, and twelve blocks accordingly. Blocks for proximity maps can be categorized as within-group blocks on the main diagonal and between-group blocks off the diagonal. Blocks for the raw data map have different combinations of symptom-groups and patient-clusters. Chen [6] proposed a concept of sufficient statistical graph for these partitioned blocks. The purpose is to comprehensively and effectively summarize information embedded in the raw data matrix and two proximity matrices with a simplified version of MV. Individual values within a block are replaced by a single summary statistic such as mean, median, or standard deviation to represent the information for that particular block. Figure 8

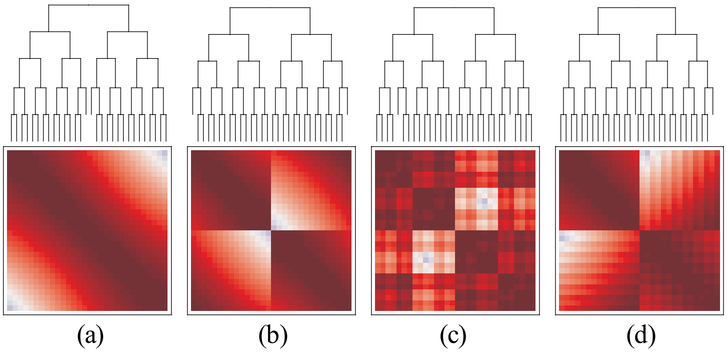


Figure 6: Results of different intermediate nodes flipping mechanisms applied to one tree-architecture (dendrogram) for a given proximity matrix.

displays the mean sufficient statistical graph for Figure 7. This presentation clearly illustrates the within-groups strength and the between-group relationship for symptoms and patients. More importantly, the sufficient graph for the raw data map effectively summarizes the interaction patterns of four patient-clusters on three symptom-groups. These three mosaic-displays of MV in Figure 8 can now easily reveal all three components of linkages for a given multivariate data set.

4 Generalization and flexibility of matrix visualization

Matrix visualization is very flexible and can be easily generalized for various purposes and situations. Two examples are illustrated in this section.

4.1 Sediment MV

Regular MV preserves the identity of each subject and variable, each dot in Figure 9a is the score of a specific symptom for a particular patient. It is possible to ignore symptom identity and sort the symptom profile for each patient according to severity. This results in the sediment MV for patients, as seen in Figure 9b, to express severity structure. One could also omit patients' identities and create the sediment MV for symptoms, as in Figure 9c. This is a side-by-side bar-chart and box-plot which displays the distribution structure for all symptoms simultaneously.

4.2 Sectional MV

The goal of a sectional MV is to display only those numerical values that satisfy certain conditions in the original MV display. Each sub-figure in Figure 10 exhibits correlation coefficients with p-values smaller than certain

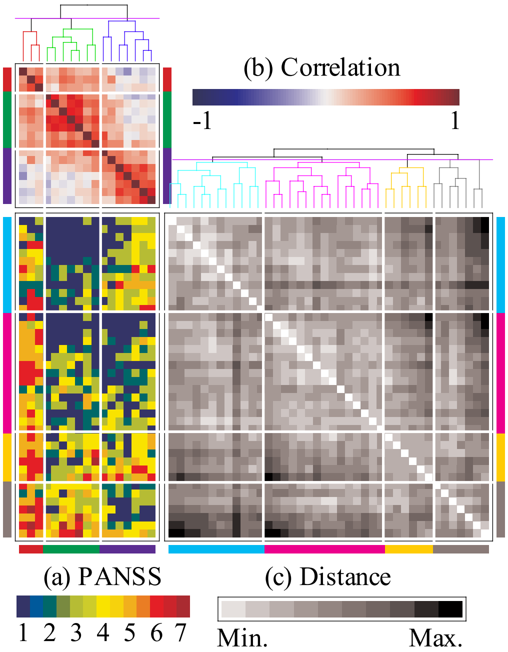


Figure 7: Partitions of permuted matrix maps with dendrograms.

significant levels for a student t-test. Figures with smaller p-values preserve more significant correlation coefficients along the main diagonal to reveal major (tight) symptom-groups, since the matrix maps have already been permuted.

5 Future directions

Matrix visualization is not a new research field but there are still many topics to be explored. All the available MV methods focus on seriation algorithms or coloring (shading) schemes for a data or proximity matrix with entries along a continuous scale. This is insufficient for exploring more complicated information structures in the statistical modelling of longitudinal, categorical, dependent or other complex data. We discuss several possible MV related issues in this section.

5.1 Categorical data

Two major difficulties arise in applying MV to categorical data (especially nominal): computing of proximities - given a categorical data matrix, it is necessary to define the proximity measure so that the numerical version of

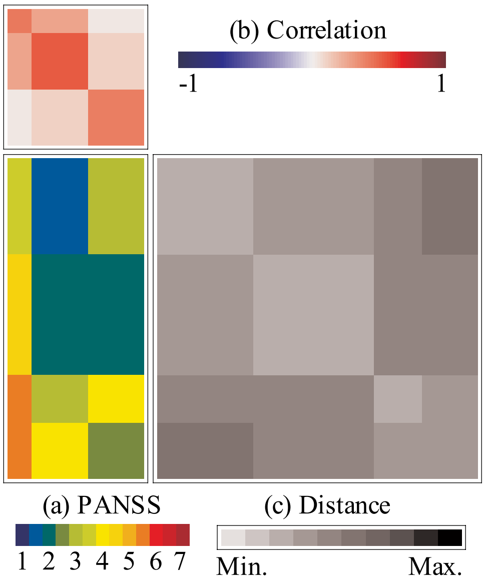


Figure 8: Sufficient statistical graph in GAP.

relativity is valid; selection of color spectrum - categories sharing similar (different) subject-distributions should be assigned with comparable (distinct) colors in order to satisfy the color version of the relativity concept. Chen [5] and Chang et al. [3] introduced a categorical version of GAP which can resolve these two difficulties.

5.2 Longitudinal multivariate data

The PANSS symptom profiles were collected when patients were admitted. There are also PANSS profiles collected at discharge and follow-ups. How to use a single or multiple MV displays to summarize and present the integration of structure for patients, symptoms, and time is a complicated and challenging task.

5.3 Multi-conditioned multivariate data

There are symptom tables other than PANSS in Hwu et al. [22] that can be analyzed simultaneously. The setup is similar to the longitudinal MV problem. Both data structures have identical sets of subjects across multiple time points or multiple tables. The longitudinal one has only one set of variables measured multiple times while the multi-conditioned one has various sets of variables. Statistical methods and theories from canonical correlation are discussed in Chen [5] and Chi [7] for multi-conditioned version of GAP.

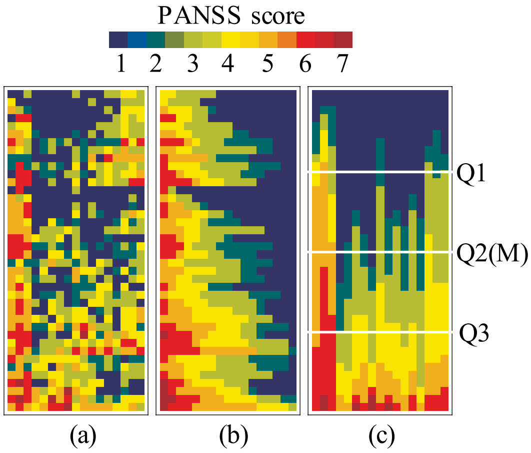


Figure 9: Sedimented MV for patients and symptoms.

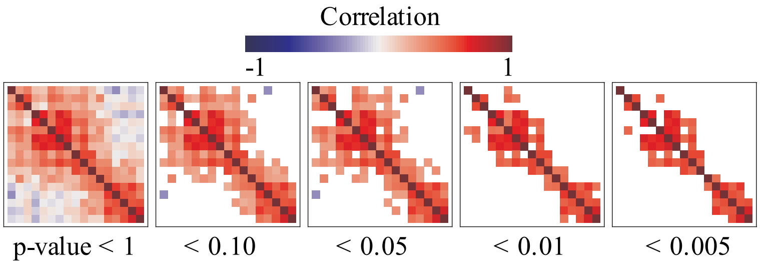


Figure 10: Sectional MV for the permuted correlation coefficient map.

5.4 MV with covariates adjustment

When effects of covariates such as gender or age are of concern in an MV analysis, covariate adjustment has to be taken into consideration. When gender acts as the covariate, it is not easy to create an MV display. One possibility is to decompose the correlation matrix into two component matrices, one for between-group (gender) structure, one for within-group pattern. The between-group correlation matrix can then be used to study the covariate effects on the original correlation matrix.

5.5 MV with dependent variables

The MV problems discussed so far do not include dependent variables. MV for a regression context with dependent variables is similar but not identical to MV with adjusting covariates. Sliced inverse regression by Li [26] is a natural staring point. The design matrix X can be row-wise sorted first by the magnitude of the dependent variable y with or without slicing. The proximity for variables can be the original one or the sliced version. Finally the reduced variables can be treated as the raw data and fed to the regular MV environment. There are certainly many different kinds of MV that can be developed for various regression problems.

5.6 Data with dependent (clustered) structure

When samples are collected with dependent structure, such as familial data for genetics studies, two difficulties emerge. First, it is numerically difficult to define the proximities with multi-levels of relationship and two issues must be addressed: the definition of between-cluster distance or similarity; whether or not within-cluster structure should be preserved while computing the between-cluster relationship. Second, it is graphically hard to display proximity and data maps simultaneously for the individuals' relationship and clusters' (families) structure. Generation of indexing variables for clusters may be necessary in forming MVs for data with dependent structure.

5.7 Mixed data

Categorical variables introduce difficulties in the computation of proximities and the selection of color spectrums. Problems get even more complicated when variables collected are mixed with quantitative, ordinal, binary, and nominal types. General similarity coefficients proposed by Gower [18] and general weighted two-way dissimilarity coefficients introduced by Cox and Cox [10] may aid the calculation of proximity matrices for variables and subjects in constructing the MV display with mixed data. Color coding for a data matrix with mixed data is a more difficult task.

5.8 MV for huge data

When data size (variable or subject) exceeds the limitations of computers used, such as thousands of variables and millions of subjects, burdens may come from the CPU speed, computer memory and display. Parallel and distributed computation with PC clusters may speed up computation time. When hardware support is not available, procedures from sampling techniques, sequential analysis, smoothing methods, and image processing all can be of help in creating MV for studying structures of huge data sets.

There are many interesting research areas not yet mentioned - MV for colorblind people, MV for spatial data, and MV with missing observations are good examples.

6 Conclusion

MV tools are not created to replace existing mathematical or statistical procedures. Instead, they can be applied in advance to obtain a general picture of the information structure and build up confidence for choosing and using more rigorous and appropriate mathematical and statistical operations. Of course it is possible that a good MV display alone can answer all the questions a user has in mind and reveal more comprehensive understanding about a data set than formal mathematical operations and statistical modellings.

References

[1] Bertin J. (1967). *Semiologie graphique*, Paris: Editions Gauthier-Villars. English translation by William J. Berg. as Semiology of Graphics: : Diagrams, Networks, Maps. TheUniversity of Wisconsin Press, Madison, WI, 1983.

[2] Carmichael J., Sneath P. (1969). *Taxometric maps*. Systematic Zoology **18**, 402 – 415.

[3] Chang S.C., Chen C.H., Chi Y.Y., Ouyoung C.W. (2002). *Relativity and resolution for high dimensional information visualization with generalized association plots (GAP)*. Proceedings in Computational Statistics 2002 (Compstat 2002), Berlin, Germany, 55 – 66.

[4] Chen C. H. (1996). *The properties and applications of the convergence of correlation matrices*. In: 1996 Proceedings of the Section on Statistical computing, 49 – 54, American Statistical Association.

[5] Chen C. H. (1999). *Extensions of generalized association plots (GAP)*. In: 1999 Proceedings of the Section on Statistical Graphics, 111 – 116, American Statistical Association.

[6] Chen C. H. (2002). *Generalized association plots: information visualization via iteratively generated correlation matrices*. Statistica Sinica **12**, 7 – 29.

[7] Chi Y. Y. (1999). *Information visualization for comparing two sets of variables*. Master Thesis. Division of Biomedical Statistics, Graduate Institute of Epidemiology, College of Public Health, National Taiwan University.

[8] Chepoi V., Fichet B. (1997). *Recognition of Robinsonian dissimilarities*, Journal of Classification **14**, 311 – 325.

[9] Church K.W., Helfman J.I. (1993). *Dotplot: a program for exploring self-similarity in millions of lines of text and code*. Journal of Computational and Graphical Statistics **2**, 153 – 174.

[10] Cox T.F., Cox M. A.A. (2000). *A general weighted two-way dissimilarity coefficient*. Journal of Classification **17**, 101 – 121.

[11] Cox T.F., Cox M.A.A. (2001). Multidimensional scaling. 2nd ed. Chapman & Hall/CRC.

[12] Eisen M.B., Spellman P.T., Brown P.O., Botstein B. (1998). *Cluster analysis and display of genome-wide expression patterns.* Proc. Nat'l. Acad. Sci. U. S. A. **95**, 14863 – 14868.

[13] Encarnacao J., Fruhauf M. (1994). *Global information visualization: the visualization challenge for the 21st Century*, in Scientific Visualization Advances and Changes L. Rosenblum et al (eds), Academic Press.

[14] Fisher R.A. (1936). *The use of multiple measurements in axonomic problems.* Annals of Eugenics **7**, 179 – 188.

[15] Friendly M. (2002). *Corrgrams: exploratory displays for correlation matrices.* Amer. Statist **56**, 316 – 324.

[16] Friendly M., Kwan E. (2003). *Effect ordering for data displays.* Computational Statistics & Data Analysis **43**, 509 – 539.

[17] Gale N., Halperin C.W., Costanzo C.M. (1984). *Unclassed matrix shading and optimal ordering in hierarchical cluster analysis.* J. Classification **1**, 75 – 92.

[18] Gower J.C. (1971). *A general coefficient of similarity and some of its properties.* Biometrics **27**, 857 – 874.

[19] Hartigan J.A. (1972). *Direct clustering of a data matrix.* Journal of the American Statistical Association **67**, 123 – 129.

[20] Huber P.J. (1985). *Projection pursuit.* The Annals of Statistics **13**, 435 – 475.

[21] Hubert L. (1976). *Seriation using asymmetric proximity measures.* British J. Math. Statist. Psych. **29**, 32 – 52.

[22] Hwu H.G., Chen C.H., Hwang T.J., Liu C.M., Cheng J.J., Lin S.K., Liu S.K., Chen C.H., Chi Y.Y., Ouyoung C.W., Lin H.N., Chen W. J. 2002). *Symptom patterns and subgrouping of schizophrenic patients: significance of negative symptoms assessed on admission.* Schizophrenia Research **56**, 105 – 119.

[23] Kay S.R., Fiszbein A., Opler L.A. (1987). *The positive and negative syndrome scale (PANSS) for schizophrenia.* Schizophr. Bull. **13**, 261 – 276.

[24] Kohonen T. (1995). *Self-organizing maps.* Berlin, Heidelberg: Springer.

[25] Lenstra J.K. (1974). *Clustering a data array and the traveling salesman problem.* Operations Research **22**, 413 – 414.

[26] Li K.C. (1991). *Sliced inverse regression for dimensional reduction (with discussion).* Journal of the American Statistical Association **86**, 316 – 342.

[27] Ling R.F. (1973). *A computer generated aid for cluster analysis.* Communications of the ACM **16**, 355 – 361.

[28] Marchette D.J., Solka J.L. (2003). *Using data images for outlier detection.* Computational Statistics and Data Analysis **43**, 541 – 552.

[29] Marcotorchino F. (1991). *Seriation problems: an overview.* Applied Stochastic Models and Data Analysis **7**, 139 – 151.

[30] Minnotte M., West W. (1998). *The data image: a tool for exploring high dimensional data sets*. In: 1998 Proceedings of the ASA Section on Statistical Graphics, Dallas, Texas, 25 – 33.

[31] Murdoch D.J., Chow E.D. (1996). *A graphical display of large correlation matrices*. Statistical Computing **50**, 178 – 180.

[32] Robinson W. S. (1951). *A method for chronologically ordering archaeological deposits*. American Antiquity **16**, 293 – 301.

[33] Slagel J.R., Chang C.L., Heller S.R. (1975). *A clustering and data reorganizing algorithm*. IEEE Transactions on Systems, Man, and Cybernetics **5**, 125 – 128.

[34] Tukey J.W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

[35] Wegman E. (1990). *Hyperdimensional data analysis using parallel coordinates*. Journal of the American Statistical Association **85**, 664 – 675.

[36] Ziv B.J., David K.G., Tommi S.J. (2001). *Fast optimal leaf ordering for hierarchical clustering*. Bioinformatics **17**, S22 – S29.

Acknowledgement: This work is partially supported by the National Science Council, Taiwan, R.O.C. (NSC-90-2118-M-001-030), and the National Health Research Institutes, Taiwan R.O.C. (NHRI-GT-EX908825PP and NHRI-EX92-9113PP). The authors are grateful to Chen-Hsin Chen and Donald Ylvisaker for valuable suggestions. The first author also wants to express his gratitude to the Institute for Pure and Applied Mathematics (IPAM), UCLA for part of the current work was carried out during his visiting IPAM.

Address: C.-H. Chen, W.-J. Jang, C.-H. Kao, S. Tzeng, H.-M. Wu, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
H.-G. Hwu, Department of Psychiatry, National Taiwan University Hospital and College of Medicine, National Taiwan University, Taipei, Taiwan
Y.-J. Tien, Institute of Statistics, National Central University, Chung-Li, Taiwan

E-mail: `cchen@stat.sinica.edu.tw`

st-apps AND EMILEA-STAT: INTERACTIVE VISUALIZATIONS IN DESCRIPTIVE STATISTICS

Katharina Cramer, Udo Kamps, and Christian Zuckschwerdt

Key words: Applied statistics, descriptive statistics, interactive visualizations, multimedia, teaching and learning environment, web-based.

COMPSTAT 2004 section: Teaching statistics.

Abstract: Within the "New Media in Education Funding Programme" the German Federal Ministry of Education and Research (bmb+f) has supported the project e-stat to develop and to provide a multimedia, web-based, and interactive learning and teaching environment in applied statistics called EMILeA-stat. After sketching the structure of EMILeA-stat, its scope and objectives briefly we focus on interactive visualizations in descriptive statistics as a specific and typical aspect of the system. Alternatively, the visualizations are available off-line as a graphical package called **st·apps**.

1 Introduction

Within the "New Media in Education Funding Programme" the German Federal Ministry of Education and Research (bmb+f) supports the project e-stat (project period April 2001 – June 2004) to develop and to provide a multimedia, web-based, and interactive learning and teaching environment in applied statistics called EMILeA-stat, which is a registered brand name. It is accessible via internet (`emilea-stat.uni-oldenburg.de`).

The project was set up by 13 partners at that time working at seven German universities: Bonn, Berlin (Humboldt-University), Dortmund, Karlsruhe, Münster, Oldenburg (leading university), and Potsdam. In test and evaluation phases of EMILeA-stat other universities are involved, too. The project is also supported by further partners in advice and it cooperates with economic partners such as SPSS Software, Springer-Verlag, MD*Tech Method & Data Technologies (XploRe-Software), and AON Re. Including the group of associated partners who are providing additional content, about 70 people are co-working in developing und realizing EMILeA-stat at the present time. For more detail about the project we refer to its web page `www.emilea.de`.

2 The system EMILeA-stat

Statistical and quantitative thinking and acting have become fundamental skills in several branches of natural sciences, life sciences, social sciences, economics, and engineering. Models, tools, and methods, which have been developed in statistics, are applied in modelling and data analysis, e.g., in business and industry, in order to obtain decision criteria and to gain more insight into structural correlations. Owing to these various applications and the necessity of using statistical methodology in so many fields, there have to be consequences for the processes of learning and teaching: Pupils, for example, should get to know elementary and application-oriented statistics. Therefore, statistics and data analysis, theoretically and practically, have to become part of teachers' studies at university and at in-service training courses. Moreover, students of many different disciplines with a statistics impact should be familiar with basic and advanced statistics. These goals gave the main impact to develop EMILeA-stat

- as one system suitable for teaching statistics at schools, universities, and in further vocational training,
- as one system which supports supervised and self-directed learning, and
- as one system which is accessible anywhere, anytime, and for anyone.

The basic concept offers on the one hand the opportunity to tailor individual courses covering specific learning needs. On the other hand, EMILeA-stat serves as an intelligent statistical encyclopaedia.

Basic statistical contents are presented on three levels of abstraction in order to take into account that different types of users have – owing to their individual mathematical and theoretical backgrounds – different needs. If sensible the contents are written on level

A (elementary level): presentation in a popular scientific way by assuming no or only a low previous (mathematical) knowledge,

B (basic level): like undergraduate courses in applied statistics for students, e.g., of economics, psychology, and social sciences, and

C (advanced level): containing deeper material and special topics within the broad field of statistics and applied probability.

Furthermore, user-oriented views and scenarios, which are near to real world applications, are integrated.

The following fields and subjects of quantitative methodology are or will be contained in EMILeA-stat: Descriptive and inductive statistics, exploratory data analysis, interactive statistics, graphical representations and methods, basic mathematics needed in statistics, probability theory, statistical methods in finance and insurance mathematics, modelling and prediction of data in financial markets, statistical methods in marketing, virtual productions and virtual company, experimental design, statistical quality management, and business games.

3 Interactive visualizations

The theoretical statistical content in EMILeA-stat is supplemented by interactive visualizations which are programmed as Java-Applets. By offering a variety of interactive options (for a detailed description see below) they support the learner in her/his learning process by offering the possibility to explore the explained method, to experiment with data, and to make own experiences with the discussed topic. Due to the fact that many places, e.g., at universities or schools, where teaching takes place still do not have access to the internet, these visualizations are not only part of the system but also realized as an off-line graphical package called **st·apps** . A German version of this tool – including an additional textbook with explanations, instructions, proposals for the use in teaching, etc. – is available via the publishing company Springer. An English edition is planned.

In the following we give an overview about the visualizations included in **st·apps** and present this tool by giving some examples. Finally the differences between the on-line version and the graphical package are briefly sketched.

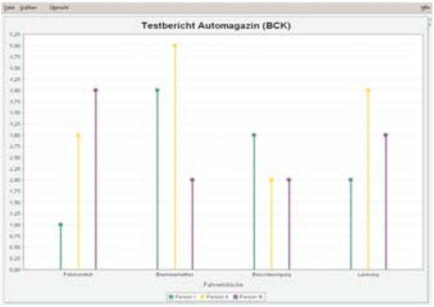
3.1 Types of visualizations

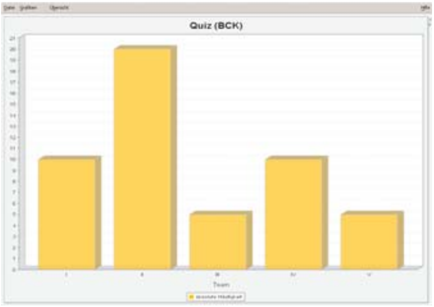
Due to their use the interactive visualizations are divided into two groups: The first one works like a simple statistical engine. It is used for analyzing own data and preparing presentations. The others are designed in the first place regarding didactical aspects in order to support the exploratory learning process for becoming familiar with the explained content or method. By offering many interactive elements the user is invited to experiment and explore the presented content by her/his own activity. Examples are given below.

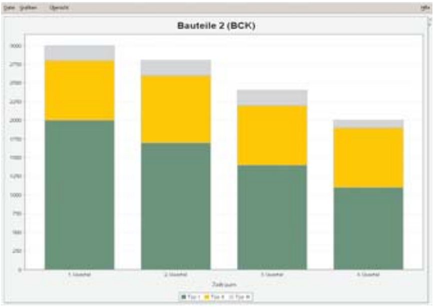
3.2 The range concerning descriptive statistics

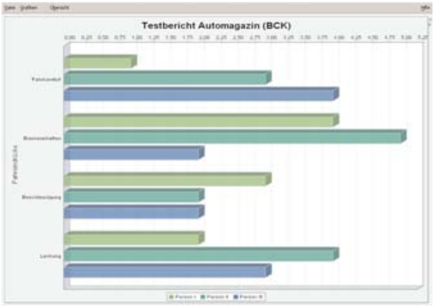
Starting with elementary visualizations like bar charts (vertical and horizontal), pie and ring charts, line charts, box plots, stem-and-leaf plots, histograms, and plots of the empirical and the approximate empirical distribution function, traditional measures like mean, standard deviation, quantiles, etc. are considered as well as measures of relationships between measurement or categorical variables. Methods for the description of economic data like price index numbers and measures of concentration, e.g., Lorenz curves and Gini coefficients, are also illustrated by this kind of interactive illustration. The field of univariate approaches is completed by regression and time series analysis. Furthermore, applets with scatterplots and scatter matrices are realized. All the mentioned visualizations are available on-line as part of EMILeA-stat, while in **st·apps** the following subjects focusing on analyzing and presenting data are included:

Bar charts

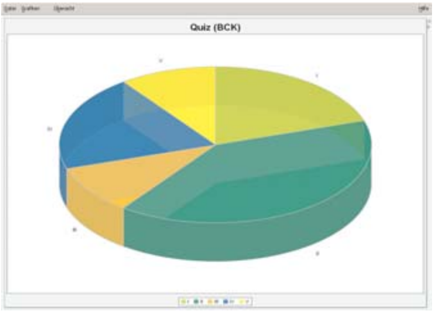


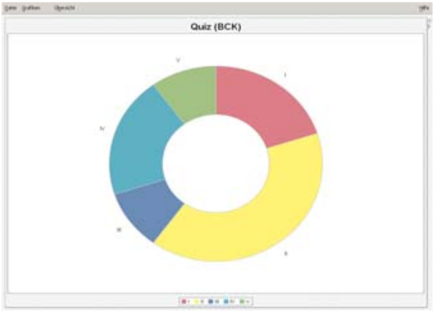




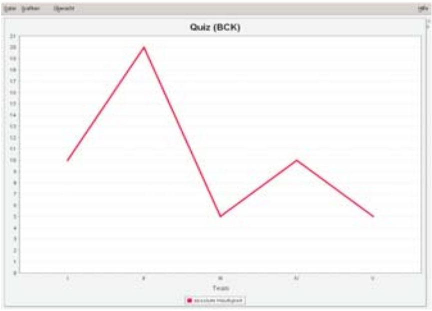


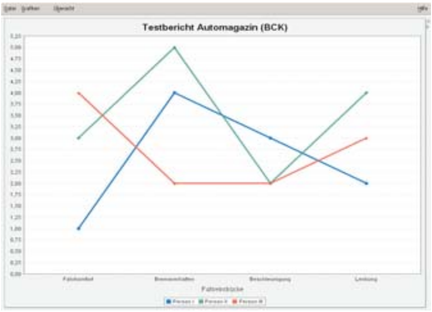
Pie and ring charts



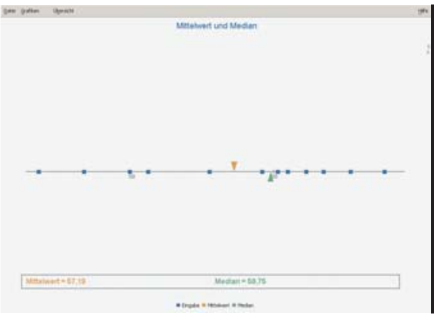


Line charts



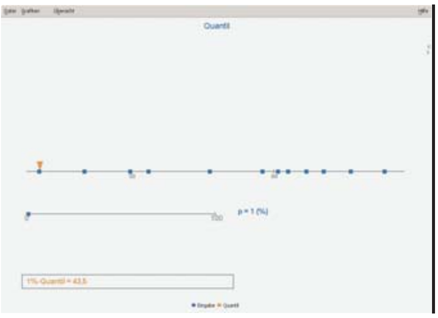


Location parameters



Mittelwert = 57,18 Median = 58,75

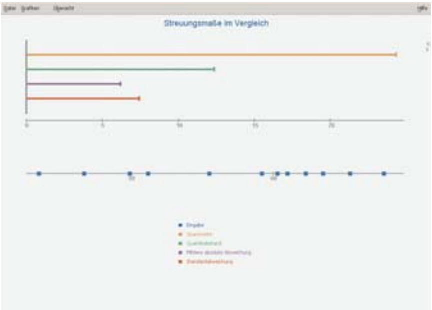
Mean and median



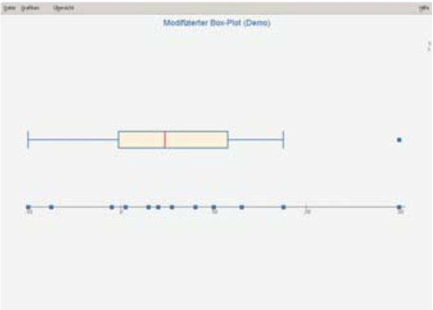
1%-Quantil = 43,5

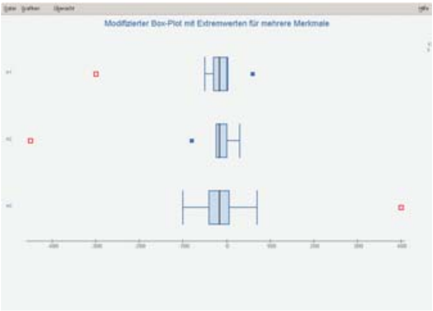
Quantiles

Scale parameters

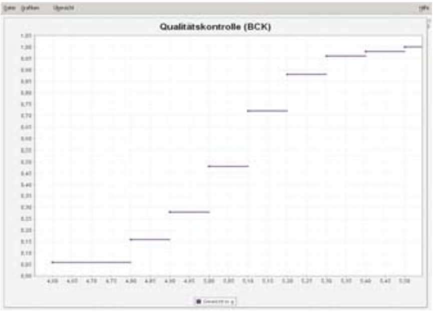


Box plots

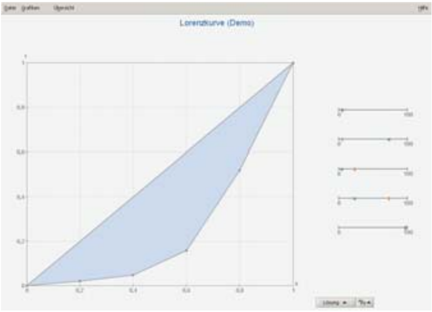


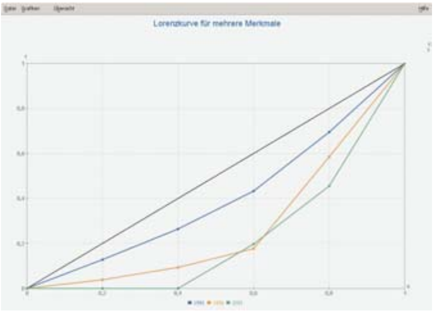

Modifizierter Box-Plot mit Extremwerten für mehrere Merkmale

Empirical distribution function

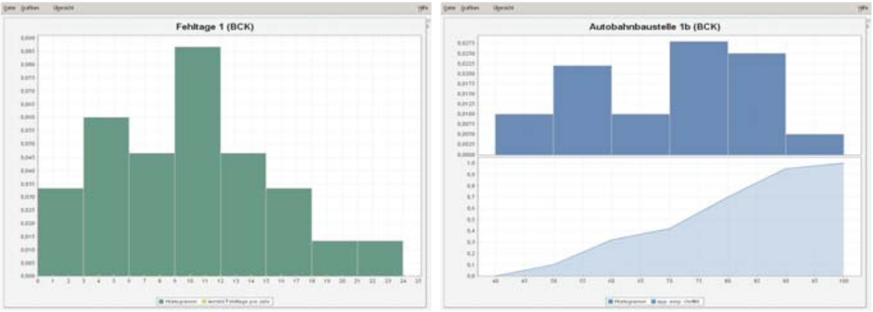


Lorenz curves

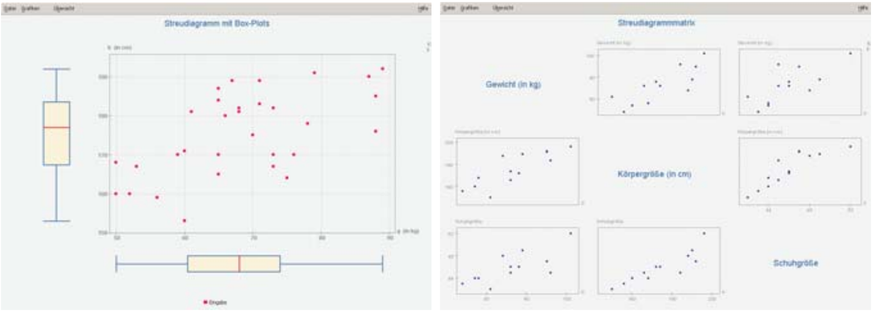




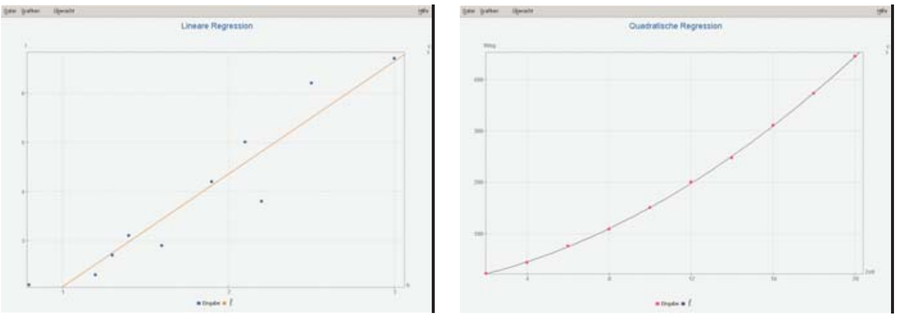
Histograms and approximate empirical distribution function



Scatter plot and scatter matrix



Regressions

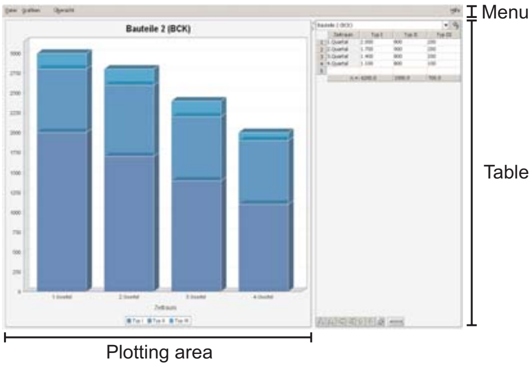


Linear regression

Quadratic regression

3.3 The structure

Each interactive visualization consists of three parts: the plotting area, a table for the data, and the menu.



When opening a visualization in **st·apps** a given data set is loaded automatically into the table and presented in the plotting area.

The size of the diagram and the table can be changed manually. Furthermore, it is possible to close one of these two components, a facility which is, e.g., useful for presentations.

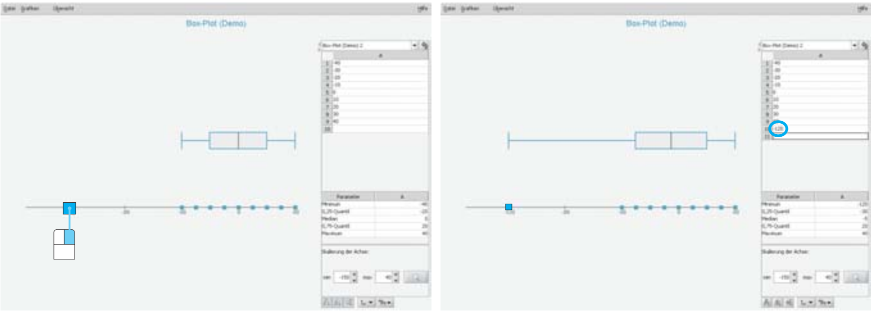
The "user interface" – the aspects of interactivity in the plot and the table or the menu – is standardized such that the frequent user should be able to work with a new visualization easily. Instructions accompanying the visualizations indicate the concept of learning by discovery enabled via the available interactivity.

Owing to the fact that the menu is organized similar to the menus of known standard software we will not go into details about this part while the plotting area and the table are described in more detail in the next paragraphs.

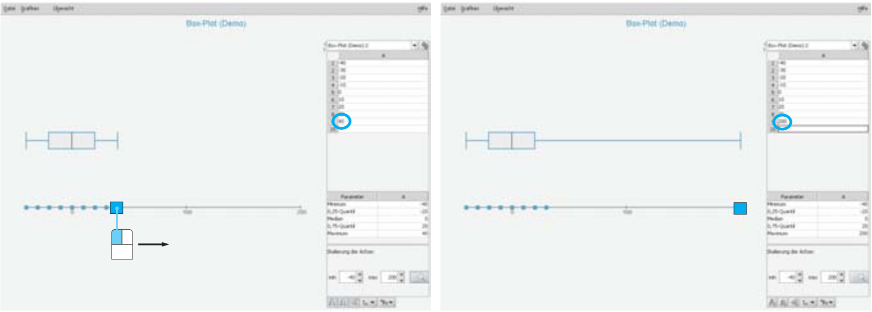
3.4 The diagram

Depending on the type of visualization different aspects of interactivity are implemented. Some of them will be explained in the following.

The automatically loaded data set can be modified by adding new data points. They can be given numerically by adding them in the table (see below) or by clicking the axis with the right mouse button.



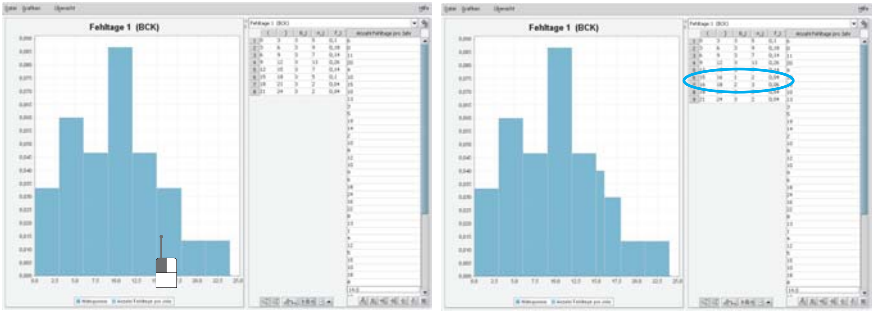
Moreover, existing data (points on the axis) can be moved to the right or left with the left mouse button. The axes are automatically rescaled.



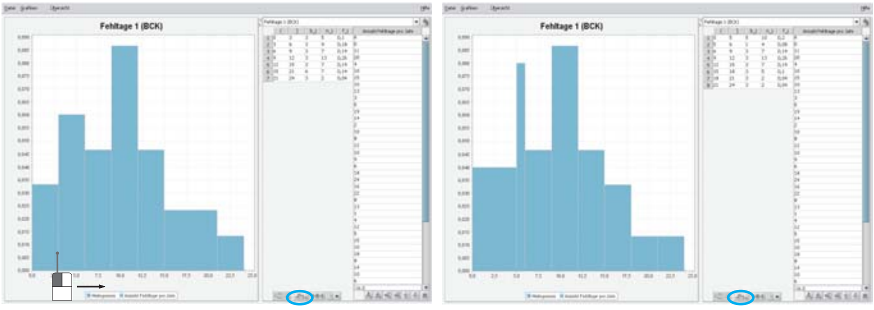
These options are included in many visualization such as those concerning location and scale parameters, box plots, scatter plots, regressions, histograms and the approximate empirical distribution function.

Furthermore, there are interactive aspects which are matched only with specific visualizations. Three examples are given in the following:

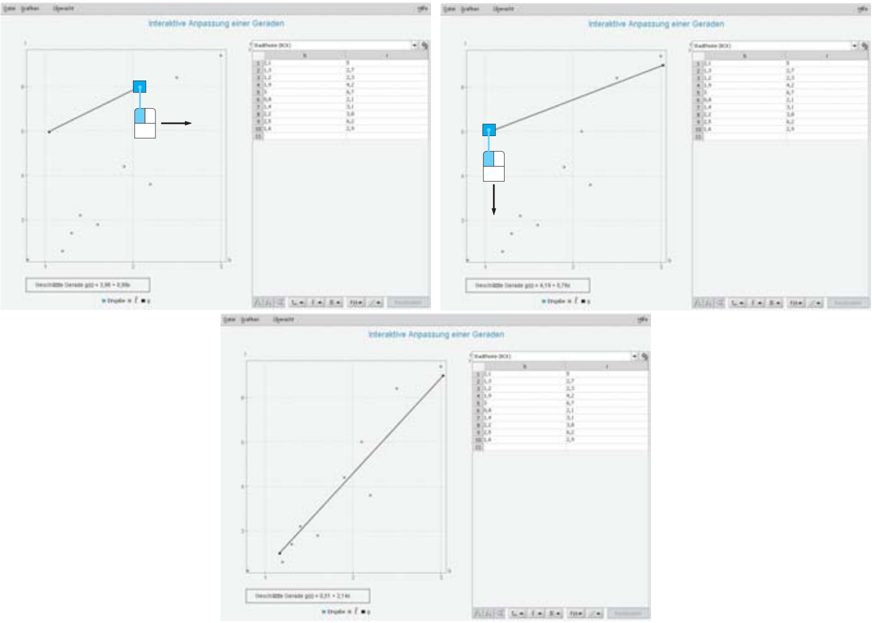
Histogram The histogram applet offers the most interactivity. Each bar can, for example, be split into two bars by clicking with the mouse into the respective bar.



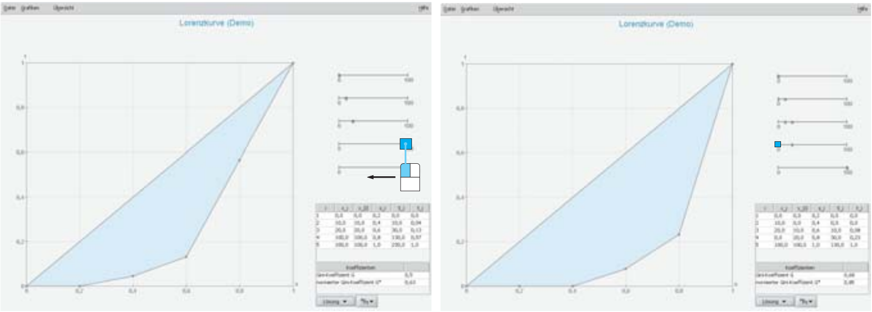
By shifting the endpoints of the bars the width of the classes and eventually the number of classes change.



Fitting a straight line Concerning linear regression there is, e.g., one visualization available where a straight line has to be fitted manually to the data. The correct linear regression function obtained by least squares can also be added for checking the manually fitted line.

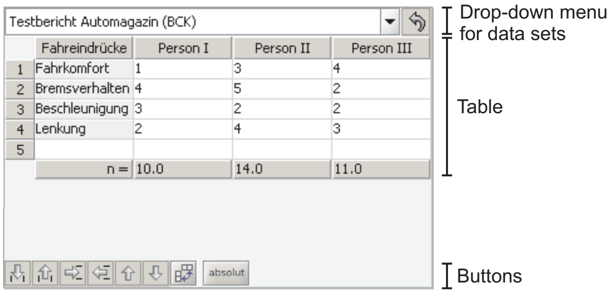

Interaktive Anpassung einer Geraden
Interaktive Anpassung einer Geraden
Interaktive Anpassung einer Geraden

Lorenz curve The Lorenz curve is offered in two interactive versions. In the applet shown in the illustration different market situations are modelled by shifting the marked buttons.



3.5 The table

The part which includes the table is composed of three parts:



Testbericht Automagazin (BCK)

	Fahreindrücke	Person I	Person II	Person III
1	Fahrkomfort	1	3	4
2	Bremsverhalten	4	5	2
3	Beschleunigung	3	2	2
4	Lenkung	2	4	3
5				
	n =	10.0	14.0	11.0

In some cases a fourth component with further information such as parameters or coefficients is realized.

In the drop-down-menu a selection of data sets suitable for visualization is offered. If a data set is loaded, the accompanying table can, e.g., be modified in the following ways:

Symbol	Action
⇩ ⇧	Add a column Delete the marked column(s)
⇨ ⇦	Add a row Delete the marked row(s)
⇧ ⇩	Shift the marked row(s) up Shift the marked row(s) down

An error in the table is indicated by (!) and ↶ restores the original data set. If a functionality makes no sense for the actual applet the corresponding button does not appear.

Depending on the specific visualization further buttons are offered. The button Korrelation ▲, for example, calculates the correlation coefficient in the regression applets while in the histogram applet ▅ generates a histogram with equidistant classes. Also the already mentioned optional fourth part with information about used parameters or coefficients is inserted by pushing a button, namely αβγ▲.

Some visualizations consist of two tables: one for the original data and one for the frequency table. The latter one depends on the first one. Therefore, it can only be modified as explained if the original data has been deleted.

3.6 Interactive visualizations available online

As already mentioned a wider range of applications, such as price index numbers or visualizations, e.g., concerning time series analysis, are offered online in EMILeA-stat. Moreover, in contrast to the off-line tool **st·apps** each interactive visualization is – similar to the theoretical content – available on three levels of abstraction. The elementary level A offers at least interactivity whereas on level C (advanced) the full range of functionality as described is accessible. Concerning the data sets loaded, this level dependent design means that the systems offers to a user working on level A only one data set (given by the teacher), while on level B she/he can choose between a wide range of data sets. On level C analyzing own data is possible. In other words the described "user interface" of the off-line tool is available only on level C to its full extent. On the other hand **st**·apps offers – because of these facilities – a variety of helpful and powerful tools for analyzing and presenting data which are also useable without an access to the internet.

References

[1] Burkschat, M., Cramer, E., Kamps, U. (2003). *Beschreibende Statistik: Grundlegende Methoden.* Springer, Heidelberg (in German).

[2] Cramer, E., Cramer, K., Kamps, U., Zuckschwerdt, Ch. (2004). *Beschreibende Statistik: Interaktive Grafiken.* Springer, Heidelberg (in German).

[3] Cramer, E., Cramer, K., Kamps, U. (2002). *e-stat: A web-based learning environment in applied statistics.* Proceedings in Computational Statistics, W. Härdle, B. Rönz (Eds.), Physica-Verlag, Heidelberg, 309 – 314.

[4] Cramer, E., Härdle, W., Kamps, U., Witzel R. (2003). *E-stat: Views, methods, applications.* Bulletin of the International Statistical Institute 54th Session, Contributed Papers, Volume LX, Book 2, 82 – 85.

[5] Cramer, K., Kamps, U. (2003). *Interactive graphics for elementary statistical education.* Bulletin of the International Statistical Institute 54th Session, Contributed Papers, Volume LX, Book 1, 222 – 223.

Address: K. Cramer, U. Kamps, C. Zuckschwerdt, University of Oldenburg, Institute of Mathematics, D-26111 Oldenburg, Germany

E-mail: `e-stat@uni-oldenburg.de`

COMPSTAT'2004 Symposium

THE CASE SENSITIVITY FUNCTION APPROACH TO DIAGNOSTIC AND ROBUST COMPUTATION: A RELAXATION STRATEGY

Frank Critchley, Michael Schyns, Gentiane Haesbroeck, David Kinns, Richard A. Atkinson and Guobing Lu

Key words: Combinatorial optimisation, convexity, diagnostics, Euclidean geometry, masking, multiple case effects, relaxation, robustness.

COMPSTAT 2004 section: Robustness.

Abstract: The present paper focuses on the case sensitivity function approach to diagnostics and robustness that are combinatorial by definition and hard to solve exactly. Attention is also given to the visual displays.

1 Overview and organisation

Central to both diagnostics and robustness are a range of optimisation problems that are combinatorial by definition and correspondingly hard to solve exactly. A variety of multiple case effects – such as masking – may be present, further complicating appropriate inference.

The present paper offers a computation-focused progress report on the case sensitivity function approach to diagnostics and robustness introduced in [4], on which we draw. A key idea here is that *relaxation brings benefits*. Specifically, the strategy outlined below shows how such high-dimensional ($O({}^nC_m)$) discrete optimisation problems can be embedded in low-dimensional ($O(n)$) smooth reformulations, in which both the insights of geometry and the power of analysis are available. In particular, informative plots become possible, while additional convexity and derivative information can be exploited.

Overall motivation for the case sensitivity function approach derives from considerations of (A) *unity*, (B) *insight* and (C) *innovation,* examples including – in order of appearance:

- (A1): an emphasis (throughout) on the connectivity of diagnostics and robustness,
- (A2): a single setting for a range of optimisation problems (Sections 3 and 4),
- (B1): visual displays affording insight into the nature and variety of multiple case effects (Section 5),
- (C1): new diagnostic methodologies (Section 6),

- (B2): insight into the performance of existing algorithms (Section 7), and:
- (C2): enhanced (potentially, encompassing) sets of algorithms for a class of robustness problems (Section 7).

A few preliminaries are established in Section 2.

2 Preliminaries

To gain focus, attention is restricted to one-sample contexts, with $\{z_i : i \in N\}$, $N := \{1, ..., n\}$ denoting a random sample of $n > 1$ cases from an unknown distribution F in $\dim(z)$ dimensions. The associated empirical distribution is $\widehat{F} := \sum_{i \in N} n^{-1} \widehat{F}_i$, where $\widehat{F}_i$ denotes the distribution degenerate at z_i. Throughout, analysis is conducted conditional on the observed $\{z_i\}$.

Assuming, as we do, that no further information is available about the observed cases, it is desirable that any analysis of these data should be invariant under permutation of the arbitrary labels attached to them. Given n, this invariance is achieved – without loss of information – by replacing $\{z_i : i \in N\}$ by $\widehat{F}$. In particular, every statistic of interest here is of the form $T[\widehat{F}]$, for some functional $T[\cdot]$. This may, for example, be (the observed significance level of) a test statistic, a parameter estimate, a prediction of future values of an observable, or a nonparametric density or regression function estimate. In particular, $T[\cdot]$ may be scalar, vector or function valued.

Let $Z := (z_i^T)$. In multivariate contexts where all the random variables in $\widetilde{z} \sim F$ are on the same footing, we put $\dim(z) = k$, $\widetilde{z} = \widetilde{x}$, $z_i = x_i$ and $Z = X$. In the usual linear model $y = X\beta + \epsilon$, we put $\dim(z) = 1 + k$, $\widetilde{z}^T = (\widetilde{y}, \widetilde{x}^T)$ and $z_i^T = (y_i, x_i^T)$, so that $Z = (y|X)$, (a constant term being assumed and accommodated by supposing that the distribution of the first element of $\widetilde{x}$ is degenerate at the value 1).

3 A combinatorial optimisation problem

Two integers $h > 0$ and $m > 0$ are called *n-complementary* if $h + m = n$, in which case:

$$A \in \mathbb{N}_h \Leftrightarrow A^c \in \mathbb{N}_m \tag{1}$$

where, for any integer $0 < a < n$, $\mathbb{N}_a := \{\varnothing \subset A \subset N : |A| = a\}$. In particular, $|\mathbb{N}_h| = |\mathbb{N}_m|$ or, in the familiar combinatorial identity, ${}^nC_h = {}^nC_m$.

Throughout, $\{H, M\}$ denotes a bipartition of N. That is, H and M are nonempty, complementary subsets of N. In particular, $|H|$ and $|M|$ are n-complementary. Of course, *holding onto* the cases labelled by H is the same thing as *missing* out those labelled by M. That is,

$$\widehat{F}_H = \widehat{F}_{-M} \tag{2}$$

where, for any $\varnothing \subset A \subset N$, $\widehat{F}_A := \sum_{i \in A} |A|^{-1} \widehat{F}_i$ and $\widehat{F}_{-A} := \widehat{F}_{A^C}$.

As is well-known, diagnostics and robustness meet at the influence function. The simple but general relations (1) and (2) provide a second, global connection between these two areas of statistics, as we now discuss. For brevity, each scalar *target* functional $t[\cdot]$ below is implicitly assumed to be defined wherever it is evaluated, and its possible dependence on $\widehat{F}$ or T suppressed notationally.

A general problem arising in diagnostics is to identify subsets M of given size m whose omission causes maximal changes $T[\widehat{F}] \to T[\widehat{F}_{-M}]$ in a statistic of interest, as measured by $t[\widehat{F}_{-M}]$ for some appropriate target functional $t[\cdot]$. A lead example is Cook's (squared) distance in the linear model. With $T[F] = \beta[F] := (\mathbb{E}_F(\widetilde{x}\widetilde{x}^T))^{-1}\mathbb{E}_F(\widetilde{x}\widetilde{y})$, $\widehat{\beta} := \beta[\widehat{F}]$ and $\widehat{\beta}_{-M} := \beta[\widehat{F}_{-M}]$, we have:

$$t_{Cook}[\widehat{F}_{-M}] := (ks^2)^{-1}(\widehat{\beta}_{-M} - \widehat{\beta})^T X^T X(\widehat{\beta}_{-M} - \widehat{\beta})$$

where s^2 is the usual estimate of error variance. Again, a range of robust estimates are defined in terms of subsets H of given size h which optimise a specified target functional $t[\widehat{F}_H]$. A lead example is minimum covariance determinant (MCD) estimation in multivariate analysis based on minimisation of $t_{MCD}[\widehat{F}_H] := \log(\det(\mathrm{cov}[\widehat{F}_H]))$. These two lead examples are developed below.

Summarising, a range of optimisation problems arising naturally in both diagnostics $(\mathcal{D})$ and robustness $(\mathcal{R})$ have *combinatorial* complexity and *entirely equivalent* $(\mathcal{D}) \leftrightarrow (\mathcal{R})$ forms expressed in Problem 3.1, in which h and m are given n-complementary integers:

Problem 3.1. *(Combinatorial optimisation problem)*
$(\mathcal{D})$ Optimise $t[\widehat{F}_{-M}]$ *over* $M \in \mathbb{N}_m$.
$(\mathcal{R})$ Optimise $t[\widehat{F}_H]$ *over* $H \in \mathbb{N}_h$.

We note in passing that a variety of other combinatorial problems, not necessarily linked to diagnostics and robustness, can also be formulated in this way.

This high-dimensional discrete problem can be embedded in a low-dimensional smooth one, as follows. It suffices to express such a relaxation strategy in, say, the $(\mathcal{D})$ form, that in the $(\mathcal{R})$ form following at once via (1) and (2).

4 A relaxation strategy

Throughout this section, h and m denote given n-complementary integers. Again, M denotes a general member of $\mathbb{N}_m$, and H its complement in N.

4.1 Probability vectors as labels for weighted empirical distributions

The first step in the relaxation strategy adopted here is to use probability vectors as labels for weighted empirical distributions.

For any $p \equiv (p_i) \subset \mathbb{P}^n :=$ {all probability n-vectors}, let $\widehat{F}(p) := \sum_{i \in N} p_i \widehat{F}_i$ denote the distribution attaching probability p_i to z_i, and $\widehat{\mathbb{F}} := \{\widehat{F}(p) : p \in \mathbb{P}^n\}$. For brevity, the $\{z_i\}$ are assumed distinct (this avoids an elaboration required in the general case). Accordingly, (indeed, equivalently),

$$p \leftrightarrow \widehat{F}(p) \text{ is a bijection between } \mathbb{P}^n \text{ and } \widehat{\mathbb{F}}. \tag{3}$$

In particular, every weighted empirical distribution corresponds to one and only one probability vector, which provides a convenient label for it. For example, $p_0 := (n^{-1})$ labels $\widehat{F}$.

Moreover, p_{-M} labels $\widehat{F}_{-M}$, where the i^{th} element of p_{-M} is zero if $i \in M$ and h^{-1} otherwise. That is, (3) specialises to:

$$p_{-M} \leftrightarrow \widehat{F}_{-M} \text{ is a bijection between } \mathbb{V}^n_{-m} \text{ and } \widehat{\mathbb{F}}_{-m},$$

where $\mathbb{V}^n_{-m}$ comprises the ${}^nC_m$ distinct probability vectors arising from permutation of $h^{-1}(0^T_m, 1^T_h)^T$ and $\widehat{\mathbb{F}}_{-m} := \{\widehat{F}_{-M} : M \in \mathbb{N}_m\}$ is the set of distributions optimised over in Problem 3.1.

The $(\mathcal{R})$ form is immediate, writing p_{-M}, $\mathbb{V}^n_{-m}$ and $\widehat{\mathbb{F}}_{-m}$ as p_H, $\mathbb{V}^n_h$ and $\widehat{\mathbb{F}}_h$ respectively. Of course, in the limit when $m = (n-1)$ (equivalently, $h = 1$), $\mathbb{V}^n_{-m}$ comprises the n unit vectors in $\mathbb{P}^n$ which label the degenerate distributions $\{\widehat{F}_i\}$ in the obvious way.

Again, with $0 < \lambda_a := a/n < 1$ denoting the proportion of cases in $\varnothing \subset A \subset N$, the identity:

$$\widehat{F} = (1 - \lambda_a)\widehat{F}_{-A} + \lambda_a \widehat{F}_A \tag{4}$$

has an exactly analogous probability vector form:

$$p_0 = (1 - \lambda_a)p_{-A} + \lambda_a p_A. \tag{5}$$

Finally, let $T[\cdot]$ denote any statistic of interest. Following [4], perturbation is defined here as movement $p \to p^*$ between probability n-vectors, with primary effect (corresponding to the identity functional T) the induced change $\widehat{F}(p) \to \widehat{F}(p^*)$ in distribution, and general effect $T[\widehat{F}(p)] \to T[\widehat{F}(p^*)]$.

4.2 Size and direction of perturbations

Again following arguments set out in [4], the second relaxation step embeds $\mathbb{P}^n$ in n-dimensional Euclidean space $\mathbb{E}^n$, this choice of geometry assigning both *size* and *direction* to perturbations.

In particular, the size $r^{(n)}_{-m} \equiv r^{(n)}_h = \sqrt{m/(nh)}$ of the perturbation $p_0 \to p_{-M}$ (*not*, note, of its primary effect $\widehat{F} \to \widehat{F}_{-M}$):

(i) does not depend on which m cases are deleted,
(ii) increases with m for fixed n, and
(iii) decreases with n for fixed m,

each of which is intuitive.

Again, for any nonzero vector v in $\mathbb{E}^n$, let $d(v) := v/\|v\|$ denote its direction. Then, for any $\varnothing \subset A \subset N$, $d_A := d(p_A - p_0)$ and $d_{-A} := d_{A^c}$ are the directions of the perturbations (from p_0) which *hold onto* and *miss out* A, respectively. In particular, (4) and (5) can be tellingly re-expressed as $d_A = -d_{-A}$. In words, *for any nonempty proper subset of cases, the perturbation which holds onto it is in the* ***opposite*** *direction to that which misses it out.*

Finally, let $\{M_r : r = 1, 2, 3\}$ denote a tripartition of N. Then it is easy to see that the perturbations $\pm d_{M_1}$ (from p_0) holding onto and missing out M_1 are *orthogonal* to those, $\pm d(p_{M_2} - p_{M_3})$, which trade probability weight between the cases labelled M_2 and M_3, exactly similar relations holding under cyclic permutation of subscripts.

4.3 Convexification of the feasible region

Recalling that $\mathbb{V}^n_{-m}$ labels the distributions over which an optimum is sought, the third relaxation step is to embed $\mathbb{V}^n_{-m}$ in its convex hull, $\mathbb{P}^n_{-m}$ say, this larger set serving (below) as the feasible region for the smooth embedding of Problem 3.1.

It follows that $\mathbb{P}^n_{-m} = \{p \in \mathbb{P}^n : p_i \leq h^{-1} \ (i \in N)\}$, a closed convex polyhedron of maximal dimension $(n-1)$ in $\mathbb{E}^n$. And, dually, that $\mathbb{V}^n_{-m}$ is the set of all *vertices* (extreme points) of $\mathbb{P}^n_{-m}$. That is, all those members of $\mathbb{P}^n_{-m}$ which cannot be written as a strict convex combination of two other members. Geometrically, all those points in $\mathbb{P}^n_{-m}$ which do not lie in the interior of a line segment joining two others. Again, we have:

$$\{p_0\} = \{p \in \mathbb{P}^n : p_i \leq n^{-1} \ (i \in N)\} \subset \mathbb{P}^n_{-1} \subset \mathbb{P}^n_{-2} \subset \ldots \subset \mathbb{P}^n_{-(n-1)} = \mathbb{P}^n \quad (6)$$

while, writing $\mathbb{P}^n_{-m}$ as $\mathbb{P}^n_h$, the $(\mathcal{R})$ form is immediate.

4.4 Examples

Figure 1 illustrates the $n = 3$ case. $\mathbb{P}^3 = \mathbb{P}^3_{-2}$ is the outer equilateral triangle, whose vertices $\mathbb{V}^3_{-2}$ are the unit vectors. $\mathbb{P}^3_{-1}$ is the inverted, inner equilateral triangle, whose vertices $\mathbb{V}^3_{-1}$ are the midpoints of the sides of $\mathbb{P}^3$. Both triangles are centred on p_0. All perturbations (from p_0) which miss out a single case are the same size, and smaller than all which miss out two. Again, each perturbation (from p_0) that holds onto a given case is in the opposite direction to that which misses it out, and orthogonal to that which trades weight between the other two.

The $n = 4$ case is illustrated in the 3-D polyhedra of Figure 2. The leftmost of these is the regular triangular pyramid $\mathbb{P}^4 = \mathbb{P}^4_{-3}$, whose vertices $\mathbb{V}^4_{-3}$ (again, the unit vectors) are shown as solid circles. The four square symbols shown there are the vertices $\mathbb{V}^4_{-1}$, each $p_{-\{i\}}$ being the centroid of the face of $\mathbb{P}^n$ opposite to $p_{\{i\}}$, (a result that holds for any $n > 1$). Again, the six oval symbols at the mid-points of the edges of $\mathbb{P}^4$ are the vertices $\mathbb{V}^4_{-2}$.

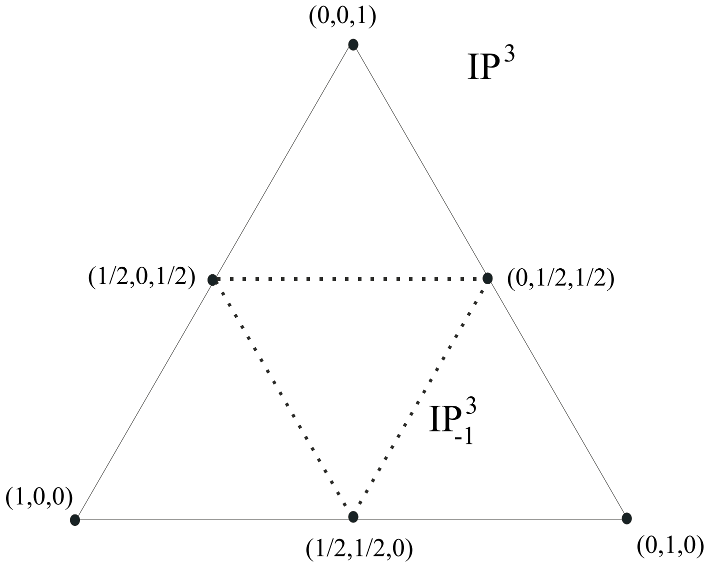


Figure 1: $\mathbb{P}^3$ and some of its key features.

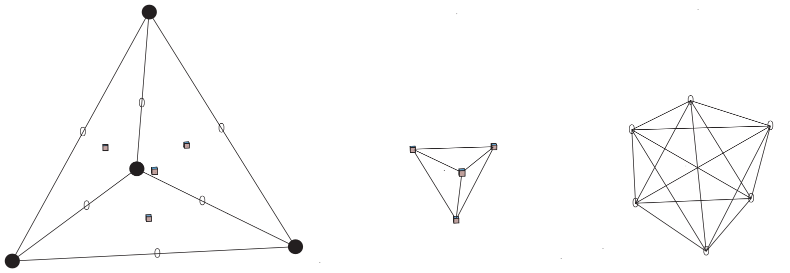


Figure 2: $\mathbb{P}^4$ and some of its key features.

The convex hulls $\mathbb{P}^4_{-1}$ and $\mathbb{P}^4_{-2}$ of these two vertex sets comprise the other two polyhedra shown, all three being centred on p_0. The inclusions (6) are clear.

Overall, the three sides of $\mathbb{P}^3$ are scaled copies of $\mathbb{P}^2$, each being the region where zero weight is attached to a given case. For the same reason, the four faces of $\mathbb{P}^4$ are scaled copies of $\mathbb{P}^3$, similar results holding in general.

4.5 A smooth reformulation

Now, exploiting (3), we define the *case sensitivity function* $T(\cdot)$ for the statistic $T[\cdot]$ via $T(p) := T[\widehat{F}(p)]$. Similarly, we define the *smooth* target function $t(\cdot)$ via $t(p) := t[\widehat{F}(p)]$. In particular, $t_{MCD}(p) = \log(\det(\text{cov}[\widehat{F}(p)]))$, while

$t_{Cook}(p) = (ks^2)^{-1}(\widehat{\beta}(p) - \widehat{\beta})^T X^T X(\widehat{\beta}(p) - \widehat{\beta})$, where $\widehat{\beta}(p) := \beta[\widehat{F}(p)]$.

The final relaxation step is to embed Problem 1 in:

Problem 4.1. *($O(n)$ smooth reformulation of Problem 3.1)*
Optimise $t(p)$ over $p \in \mathbb{P}^n_{-m} \equiv \mathbb{P}^n_h$.

It follows at once that any concave (respectively, convex) smooth target function $t(\cdot)$ attains its minimum (respectively, maximum) over the feasible region $\mathbb{P}^n_{-m} \equiv \mathbb{P}^n_h$ of Problem 4.1 at a member of the feasible region $\mathbb{V}^n_{-m} \equiv \mathbb{V}^n_h$ of Problem 3.1 and, in the strict case, only at such a vertex.

In particular, [7] show that $t_{MCD}(\cdot)$ is concave, exploiting this in their *smooth*-MCD algorithms. Although its convexity in a neighbourhood of p_0 need not extend to the whole feasible region of Problem 4.1, [4] present numerical results which support the conjecture that p-generalised Cook's distance $t_{Cook}(\cdot)$ enjoys similar extremal properties (as they note, it would be helpful to have either a proof of – or counterexample to – such a conjecture). We note, in passing, that further positive evidence for it turns up in Figure 3 of the following section.

5 Visual displays of multiple case effects

One outcome of the above relaxation strategy is the availability of visual displays offering insight into the nature and variety of multiple case effects that can occur in different contexts. We focus here on graphs of $t_{Cook}(\cdot)$ in the linear model, following [10] from which Figure 3 is taken.

For all but the smallest values of n, direct visualisation of the graph of any smooth target function $t(\cdot)$ over $\mathbb{P}^n$ – or one of its subsets $\mathbb{P}^n_{-m}$ – is prevented by the fact that each has dimension $(n-1)$. Instead, the approach adopted here uses tripartitions of N as devices providing informative triangular subsets of $\mathbb{P}^n$, over which the graph of $t(\cdot)$ can then be displayed. The key idea is to attach *equal* probability weight to cases in the same member of a tripartition. This turns out to be a rich enough structure to provide insight into a range of multiple case effects – allowing us, in effect, to *see* the nature of each, and their variety.

5.1 Tripartitions

Suppose then that $\mathcal{M} := \{M_r : r = 1,2,3\}$ is a given partition of N into three disjoint subsets, with $m_r := |M_r| > 0$ and $\sum_r m_r = n$, and let

$$\mathbb{T} = \mathbb{T}(\mathcal{M}) := \{p \in \mathbb{P}^n : [i \in M_r,\, j \in M_r] \Rightarrow p_i = p_j\}.$$

It follows that $\mathbb{T}$ is the convex hull of $\{p_{M_r} : r = 1,2,3\}$. That is, $\mathbb{T}$ is the triangle which has these three points as vertices which, when convenient, we abbreviate to $\{M_r\}$. Otherwise said, $p \in \mathbb{P}^n$ belongs to $\mathbb{T}$ if and only if, for some $\pi \equiv (\pi_r) \in \mathbb{P}^3$, $p = \sum_r \pi_r \, p_{M_r}$. In this case, $\pi = \pi(p)$ is *unique*, $\pi_r(p)$ being the total probability assigned (equally) by p to the m_r cases in M_r.

Accordingly, we may identify $\mathbb{T}$ with $\mathbb{P}^3$ via the bijection $p \leftrightarrow \pi(p)$. For example, $p_0 \leftrightarrow (\kappa_r)$, where $\kappa_r := m_r/n$ is the proportion of cases in M_r. However, whereas $\mathbb{P}^3$ is a fixed equilateral, the shape and size of $\mathbb{T}$ vary with the $\{m_r\}$. Nevertheless, important inclusions, collinearities and orthogonalities in $\mathbb{P}^3$ survive in $\mathbb{T}$ for every $\mathcal{M}$.

Two obvious cyclic permutations applying, the identity:

$$p_{-M_1} = (1 - \kappa_1)^{-1}(\kappa_2 p_{M_2} + \kappa_3 p_{M_3})$$

shows that p_{-M_1} lies on the M_2M_3 side of $\mathbb{T}$, being closer to whichever vertex labels the larger number of cases. In particular, writing $p_r(\lambda) := (1 - \lambda)p_{-M_r} + \lambda p_{M_r}$, the line segment $\mathbb{L}_r := \{p_r(\lambda) : \lambda \in [0, 1]\}$ lies in $\mathbb{T}$, all three such meeting at p_0 by (5). Again, using Section 4.2, each $\mathbb{L}_r$ is orthogonal to the side of $\mathbb{T}$ containing p_{-M_r}, $\mathbb{S}_{-r}$ say, along which probability weight is traded between the other two subsets. Thus, the probability attached to M_r increases linearly along $\mathbb{L}_r$ from zero at the p_{-M_r} end to unity at the other. Indeed, for each $\lambda \in [0, 1]$, this probability is constant at the value λ for all points in $\mathbb{T}$ on the line through $p_r(\lambda)$ parallel to $\mathbb{S}_{-r}$. In particular, it vanishes on $\mathbb{S}_{-r}$.

5.2 Four multiple case effects in the linear model

[3] and [11] discuss a variety of possible effects that a pair of cases may have on Cook's distance. Here, with M_3 representing a convenient 'null' data set, and restricting ourselves to the special case $m_1 = m_2 = 1$ (for a fuller account, see [10]), we consider four effects defined in the table below, and illustrated in the corresponding rows of Figure 3:

	Effect	**Joint presence of M_1 and M_2 ...**
(a)	Masking	. . . conceals presence of either
(b)	Cancellation	. . . has no effect on fitted line
(c)	Swing	. . . swings fitted line, (intercept ∼ unchanged)
(d)	Raise & Lower	. . . translates fitted line, (slope ∼ unchanged)

For clarity, stylised simple linear regression data sets are used, shown in the middle column of Figure 3. In each case, M_3 contains $m_3 = 20$ points, comprising five replicates at each corner of the square $\{\pm 1\}^2$, whose fitted line is the horizontal axis. Both M_1 and M_2 consist of a single point at the corner of $\{\pm 4\}^2$ indicated.

The righthand column of Figure 3 gives the corresponding graph of $t_{Cook}(\cdot)$ over $\mathbb{T}$, limits being used where needed (since, of course, a line cannot be fitted to a single case). Some linear rescaling between plots has been applied, both vertically and horizontally, to enhance their visual clarity (a minor cost being some loss of visual perception that the angle at M_3 exceeds 87°). Note that p_0 (corresponding to $\widehat{F}$) is close to M_3, being just one-eleventh of the way along the line $\mathbb{L}_3$ joining M_3 to the midpoint of the opposite side. The

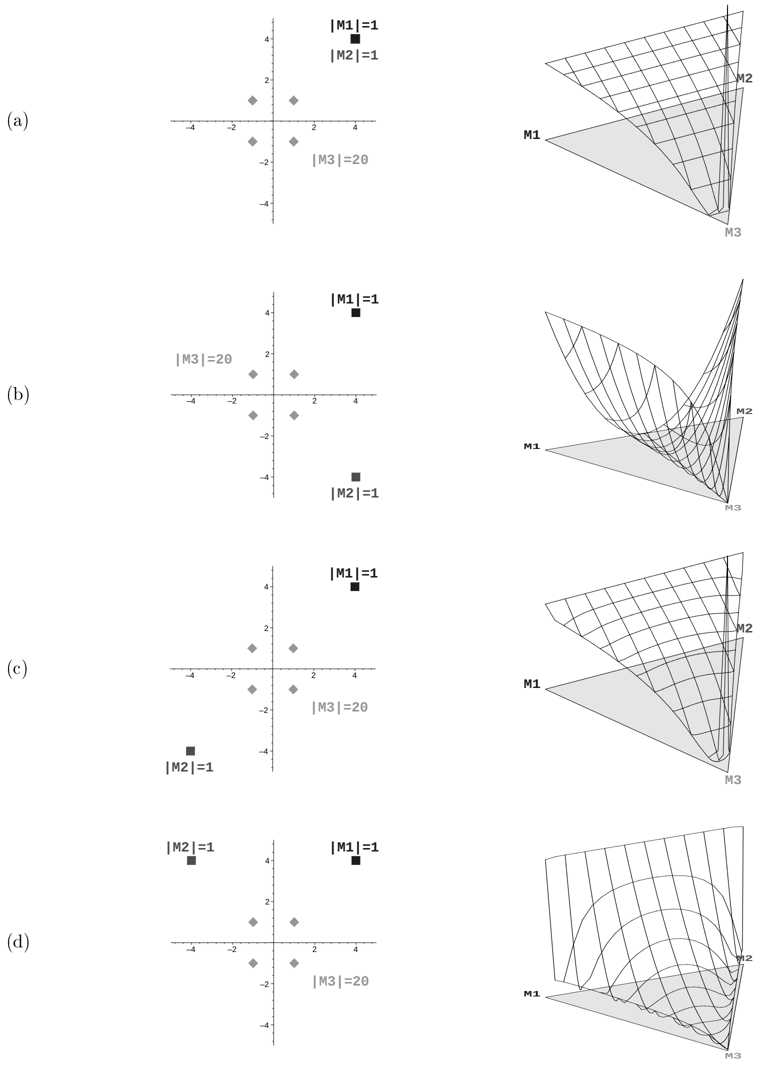


Figure 3: Four multiple case effects in the linear model: (a) masking, (b) cancellation, (c) swing and (d) raise & lower.

inbuilt $M_1 - M_2$ symmetry is evident throughout. Overall, the four graphs have visibly different *shapes*, discussed next:

(a) Masking. The ‘spike’ at M_3 reflects the dominant effect of removing both M_1 and M_2, while the parallelism of the contours to $\mathbb{S}_{-3}$ corresponds to the fact that there is, of course, no effect here in trading weight between these sets.

(b) Cancellation. The contours of $t_{Cook}(\cdot)$ here are straight lines fanning out from M_3. In particular, $\mathbb{L}_3$ is the zero height contour, since varying π_3 while keeping $\pi_1 = \pi_2$ has no effect on the fitted line. Trading weight between M_1 and M_2 now has a quadratic, globally dominant, effect.

(c) Swing. The overall shape of the surface here is very similar, but not identical, to that in the masking case. The 'spike' at M_3 remains dominant, but the surface contours are no longer parallel to $\mathbb{S}_{-3}$.

(d) Raise & Lower. This is perhaps the most interesting graph. As is intuitive from the data, the dominant global effect occurs along $\mathbb{S}_{-3}$. Looking at the surface, we see two 'troughs'. These run along $\mathbb{L}_1$ and $\mathbb{L}_2$, showing that varying the weight on one of these subsets alone has little effect. The contours of $t_{Cook}(\cdot)$ are parallel to $\mathbb{S}_{-3}$ when there is little weight on M_3, but become more curved as π_3 increases. Locally to p_0, trading weight between M_1 and M_2 produces the largest effects.

6 A relaxed diagnostic approach to detecting heavy mutual masking

Multiple case effects can be strong and yet intrinsically hard to detect with standard diagnostic procedures, while the burden of full enumeration increases combinatorially with m. Heavy mutual masking is a well-known example, challenge data sets comprising 60% of cases from one distribution and 40% from a second, suitably remote from the first. [4] present a widely applicable, relaxed, two-stage approach to detecting such effects (cf. [2]), briefly reviewed here.

Adopting the standard assumption in the literature that at most half the cases are discordant from a common pattern followed by the rest, Stage I consists on maximising (say) a suitable target function $t(\cdot)$ over $\mathbb{P}^n_{-m}$, with m the integer part of $n/2$, the optimum being known or assumed to occur at a vertex. This corresponds precisely to missing out a specified subset $\widehat{M}$ of m cases. The (in)equality constraints defining $\mathbb{P}^n_{-m}$ being linear, this relaxed optimisation can be carried out with standard software (or some alternative, as indicated in Section 7). The assumed internal consistency of the cases in $\widehat{H} := \widehat{M}^c$ may also be checked.

Stage II back-checks for swamping. That is, for cases in $\widehat{M}$ which are *not* inconsistent with the pattern followed by the majority. [4] envisage doing this *separately* for each case in $\widehat{M}$, although a *sequential* approach is possible. Having augmented $\widehat{H}$ with any such cases, a final check on their internal consistency can be made while, if required, the possibility of further structure within the cases in $\widehat{M}$ may be made.

[4] report encouraging results for this general strategy, using regression as a test problem and several forms of challenge data set. Specifically, they maximise $t_{Cook}(\cdot)$ in Stage I, using the mean shift outlier test in Stage II.

Finally, a remark on local maxima. On those occasions when the final check for a common pattern fails, the possibility that this is because omission of $\widehat{M}$ is a particular form of non-trivial local maximum can be easily explored as follows. The value of $t(\cdot)$ there can be compared to that where $\widehat{M}$ is held onto. If this is greater, replacing $\widehat{M}$ by its complement, and then continuing as before, is indicated. On the relatively few occasions where it was needed in their regression study, [4] report that this simple strategy was successful. The original $\widehat{M}$ containing no mutually masked cases, moving to its complement produced a large increase in t_{Cook} and led again to correct identification of the structure in the data.

7 Developments in relaxed robust computation

Consider now minimisation over $\mathbb{P}_h^n$ of the particular function $t(\cdot) = t_{MCD}(\cdot)$ as an exemplar of the class of robust estimation procedures that can be defined in this way. Algorithms for the MCD problem include those reported in [1], [8], [9] and [12]. These are all *discrete* in the sense that they address Problem 3.1, iteratively 'jumping' between members of $\mathbb{V}_h^n$.

We briefly sketch here some of the work reported in [6] and, more fully, in [7], recalling that these papers show that $t_{MCD}(\cdot)$ is, indeed, concave. Collectively, the new approaches reported therein are referred to as *smooth*-MCD algorithms.

Figure 4 shows two views of the same t_{MCD} surface over $\mathbb{P}_2^3$ for univariate data. This simple example offers some general geometric insight: the graph of t_{MCD} contains multiple local minima, separated by hills, with corresponding limitations for any purely descent algorithm. In particular, it motivates the use of *swapping* strategies aimed at 'getting you over a hill to a lower valley'. At the same time, the swapping strategy employed by the feasible subsets algorithm – while optimal in its own terms – is relatively expensive to perform and may not always be needed, in the sense that not every vertex is a local minimum.

Again, [4] note the benefits of using explicit gradient information, when this is available. [5] develop local *projected* (here, *centred*) Taylor expansions in generality. They show that such expansions are possible even when, as here, one or more constraints (here, $p^T 1_n = 1$) imply that there are *no* open sets in a function's domain (here, a subset of $\mathbb{P}^n$). Indeed, they exist uniquely under mild conditions and can be used to guide algorithms downhill, in the usual way. They also provide also a useful necessary and sufficient condition for a vertex in $\mathbb{V}_h^n$ to be a local minimum, for any t. In the t_{MCD} case, it is shown that these are precisely the points where the C-steps of FAST-MCD converge.

Now, conditional on robustness, there are two key performance criteria in any problem such as this: speed and optimality. Perfection (*i.e.* instant, global optimality!) being unachievable, different algorithms aim for it, while striking different trade-offs between these criteria. Accordingly, the state-of-

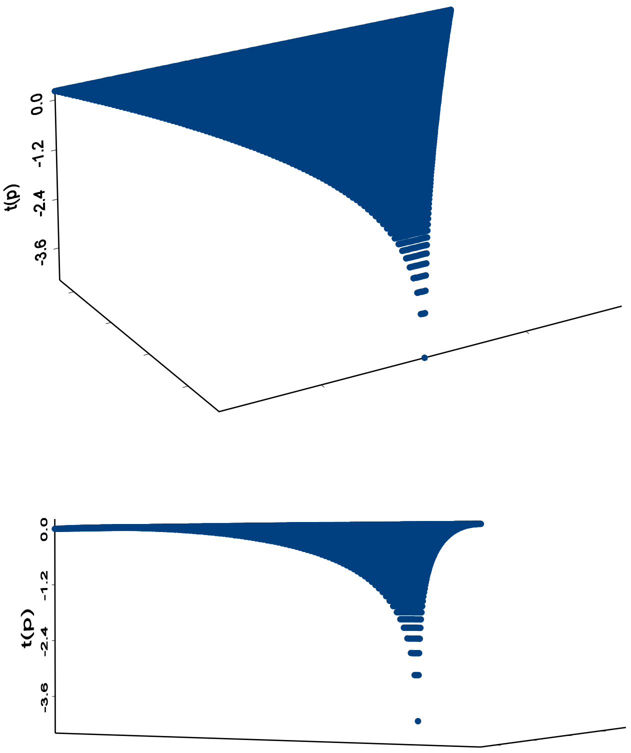


Figure 4: Two views of a t_{MCD} surface ($n = 3$; $k = 1$).

the-art can be thought of as a boundary of limiting speed/optimality trade-offs that are currently feasible, the different algorithms appearing at different points along it.

[6] and [7], to which the reader is referred for further details, exploit features of the case sensitivity function approach – in particular, insights from (convex) geometry, the power of analysis, and a unifying structure – both to understand better *why* current algorithms occur where they do along this boundary, and to add new algorithms that fill it out and/or nudge it nearer to perfection.

References

[1] Agulló J. (1998). *Computing the minimum covariance determinant estimator.* Technical report, Universidad de Alicante.

[2] Atkinson A.C. (1986). *Masking unmasked.* Biometrika **73**, 533 - 541.

[3] Barrett B.E. and Gray J.B. (1997). *Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression.* Computational Statistics and Data Analysis **26**, 39 - 52.

[4] Critchley F., Atkinson R.A., Lu G. and Biazi E. (2001). *Influence analysis based on the case sensitivity function.* J. Royal Statistical Society, **B 63**, 307 - 323.

[5] Critchley F., Lu G., Atkinson R.A. and Wang D.Q. (2003). *Projected Taylor expansions for use in Statistics.* Under consideration.

[6] Critchley F., Schyns M. and Haesbroeck G. (2003). *Smooth optimization for the MCD estimator.* International Conference on Robust Statistics, Antwerp, 29 - 30.

[7] Critchley F., Schyns M., Haesbroeck G., Lu G., Atkinson R.A. and Wang D.Q. (2004). *A convex geometry approach to algorithms for the MCD method of robust statistics.* Under consideration.

[8] Hawkins D.M. (1994). *A feasible solution algorithm for the minimum covariance determinant estimator in multivariate data.* Computational Statistics and Data Analysis ***17**, 197 - 210.*

[9] Hawkins D.M. and Olive D.J. (1999). *Improved feasible solution algorithms for high breakdown estimation.* Computational Statistics and Data Analysis ***30**, 1 - 11.*

[10] Kinns D.J. (2001). *Multiple case influence analysis with particular reference to the linear model.* PhD thesis, University of Birmingham.

[11] Lawrance A.J. (1995). *Deletion influence and masking in regression.* J. Royal Statistical Society, **B 57**, 181 - 189.

[12] Rousseeuw P.J. and Van Driessen K. (1999). *A fast algorithm for the minimum covariance determinant estimator.* Technometrics **41**, 212 - 223.

Acknowledgement: The UK authors are grateful for EPSRC support under research grant GR/K08246 and to D.Q. Wang for helpful discussions.

Address: F. Critchley, M. Schyns, G. Haesbroeck, D. Kinns, R.A. Atkinson, G. Lu, The Open University, Milton Keynes; University of Namur; University of Liège; (formerly) University of Birmingham; University of Birmingham and University of Bristol

E-mail: `F.Critchley@open.ac.uk`

COMPSTAT'2004 Symposium

ON THE BOOTSTRAP METHODOLOGY FOR FUNCTIONAL DATA

Antonio Cuevas and Ricardo Fraiman

Key words: Bootstrap validity, bootstrap consistency, bounded Lipschitz metric, differentiable functionals, functional data analysis.

COMPSTAT 2004 section: Functional data analysis.

Abstract: The current theory of statistics with functional data provides only a few results [21] of asymptotic validity for the bootstrap methodology. Roughly speaking, these validity results guarantee that the bootstrap versions of the sampling distribution of a statistic tend (as the sample size increases) to the same limit as the true sampling distributions. From a computational and practical point of view, such results have an special interest when dealing with functional data, as the distributional properties of the statistics are usually difficult to handle in this setup. Of course, the point is that while the true sampling distributions are usually very difficult to handle, the corresponding bootstrap versions can be approximated with arbitrary precision.

In this work, a uniform inequality is obtained for the Bounded Lipschitz distance between the empirical distribution of a function-valued random variable and the corresponding underlying distribution that generates the sample. As a consequence, a result of bootstrap validity (consistency) is obtained for functional statistics defined from differentiable operators.

Our proof is based on the use of a differential methodology for operators, similar to that used by Parr [19], and relies also on a result of empirical processes theory proved by Yukich [29].

1 Introduction

We deal here with the statistical setups where the available sample information consists of (or can be considered as) a set of functions. Depending on the approach and on the assumed structure of the data (which come often in a discretized version) this statistical field is called "longitudinal data analysis" or "functional data analysis" (FDA). We will follow here a purely functional approach which entails to consider the available data as true functions and, as a consequence, to define and motivate the methods in a functional framework.

The books by Ramsay and Silverman [22], [23] have greatly contributed to popularize the FDA techniques among the users, offering a number of appealing case studies and practical methodologies. Simultaneously, this increasing popularity motivates the need of a solid theoretical foundation for the FDA methods, as many basic issues (concerning, e.g., the asymptotic behavior) are often rather involved in the FDA setup.

In general terms, the FDA theory is still incomplete as many topics remain unexplored from the mathematical point of view. Some theoretical developments with functional data have been made in fields as principal component analysis ([5], [11], [17], [20], [25]), linear regression ([6], [7], [8], [9], [13]), data depth [14], clustering [1] and anova models ([12], [18], [10]).

An important issue in this field has to do with the asymptotic validity (usually called consistency) of bootstrap procedures for functional data. This looks as an interesting research line since the exact calculation of sampling distributions in FDA problems presents an obvious difficulty so that the bootstrap methodology turns out to be often the only practical alternative. Of course, the point is that while the sampling distribution of a function-valued statistic can be formally defined in the same way as the analogous concept for a real-valued statistic, the effective calculation and handling of such "functional" sampling distributions is usually very difficult since they are in fact probability measures defined on function spaces. Thus the case for using bootstrap versions is quite strong as they are discrete measures which can be in turn approximated by resampling with arbitrary precision. An example of the use of resampling methods in a functional data framework can be found in [10].

The classical works by Bickel and Friedman [3], Singh [26] and Parr [19], among others, have established the validity of the bootstrap methodology, in the case of real variables, for a number of useful statistics, including the sample mean and those generated by differentiable statistical functionals. The functional counterpart of this theory is much less developed. However, Giné and Zinn [15] have proved, in a very general setup, a bootstrap version of Donsker theorem for the empirical processes. A partial extension of this result is given in [24]. Politis and Romano [21] have proved the consistency of the bootstrap for the sample mean in the case of uniformly bounded functional variables taking values in a separable Hilbert space imposing very general assumptions on the dependence structure which include the independent case to be considered here. The main purpose of this paper is to partially extend this consistency result to (function-valued) statistics defined from differentiable operators. So we are concerned here with a functional version of some classical validity theorems, as those in [19] or [2], where the methodology based on functional differentiation plays a relevant role.

More precisely, we want to get a bootstrap validity result for statistics of type $T(P_n)$ where T is a differentiable operator (taking values in a functional space) and P_n is the empirical distribution associated with a sample $X_1, \ldots, X_n$ of n functions drawn from a common distribution P. In practical terms, this result will establish that the distribution of $\sqrt{n}(T(P_n) - T(P))$ can be approximated by its corresponding bootstrap version in $\sqrt{n}(T(P_n^*) - T(P_n))$, where P_n^* is the empirical distribution based on an artificial (bootstrap) sample drawn from the original sample. Our approach is much in the spirit of Theorem 4 in [19] although the fact that we are dealing with functional data entails some additional technical complications.

Our main result establishes that $\sqrt{n}(T(P_n) - T(P))$ and $\sqrt{n}(T(P_n^*) - T(P_n))$ converge (weakly) to the same limit. It is proved in Section 3 below. An essential auxiliary step in the proof of this theorem is a uniform (universal) bound, similar to the classical Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (see, e.g., [27]), for the distance $d(P_n, P)$ between P_n and P. It will be established in Section 2. This bound is universal in the sense that it does not depend on the underlying distribution P; this is crucial in a bootstrap setup as the bound for $d(P_n, P)$ will also hold for its bootstrap counterpart $d(P_n^*, P_n)$. Let us recall that P stands here for a probability distribution in a function space so that in order to establish a DKW-type inequality we would need a distance $d(P_n, P)$ compatible with the weak convergence and making sense in a functional framework. We will use the so-called Bounded Lipschitz metric defined by

$$d(P_n, P) = \sup_{f \in \mathcal{F}} |\int f\, dP_n - \int f\, dP|, \tag{1}$$

P being a probability on a normed space $\mathcal{X}$, P_n a empirical drawn from P and

$$\mathcal{F} = \{f : \mathcal{X} \longrightarrow \mathbb{R} :\ f \text{ is Lipschitz with } \|f\|_\infty \leq 1 \text{ and Lipschitz constant } 1\}. \tag{2}$$

2 A uniform inequality for the Bounded Lipschitz metric

Let P be a probability on a Banach space $\mathcal{X}$ whose support is included in the ball $B(0, r) \subset \mathcal{X}$. Let P_n be the empirical distribution associated with a sample $X_1, \ldots, X_n$ drawn from P. Let d denote the Bounded Lipschitz metric defined in 1). We next show a version of the DKW inequality for $d(P_n, P)$.

Theorem 2.1. *For all $\epsilon > 0$ there exists $K = K(\epsilon)$ such that*

$$\mathbb{P}\left\{\sqrt{n}\, d(P_n, P) > K\right\} < \epsilon, \text{ for all } n, \text{ for all } P \text{ with support in } B(0, r). \tag{3}$$

Proof: This will result as a direct consequence of a exponencial bound obtained by Yukich ([29], Theorem 1) by a empirical process methodology.

Define the ϵ-entropy $N(\epsilon, \mathcal{F})$ by

$$N(\epsilon, \mathcal{F}) = \min\{m : \text{there exist } f_1, \ldots, f_m \in \mathcal{F} \text{ such that } \sup_Q \min_i \|f - f_i\|_Q^2 < \epsilon^2,\ \forall f \in \mathcal{F}\}, \tag{4}$$

where the supremum on Q is taken on the set of all the probability distributions with finite support and $\|f\|_Q^2 = \int f^2 dQ$ is the $L^2(Q)$-norm.

Yukich's Theorem establishes that if the envelope function $F := \sup\{|f(x)| : f \in \mathcal{F}\}$ fulfills $F \leq 1$ and there are constants $0 < \epsilon_0 \leq 1$, $0 < \delta < 1$, and $C \geq 1$ such that

$$N(\epsilon, \mathcal{F}) \leq \exp(C/\epsilon^{2-\delta}),\ \forall \epsilon,\ 0 < \epsilon \leq \epsilon_0, \tag{5}$$

then

$$\mathbb{P}\left\{\sqrt{n}\, d(P_n, P) > M\right\} \leq 8 \exp(-M^2/5)\ \forall n \geq 1, \tag{6}$$

for all M greater than, or equal to, some constant $M(\delta, C, \epsilon_0)$ whose explicit expression is given in the statement of Theorem 1 in [29].

In fact, the proof will be more simple and intuitive by replacing the distances $\|f - f_i\|_Q^2$ in (5) by the supremum distances $\|f - f_i\|_\infty$. As a consequence, we will prove a stronger version of condition (5), by taking the supremum in (5) over all the possible probability measures (instead of just considering those of finite support). The reason is that we in fact will provide a bound for $\|f - f_i\|_\infty$ and, as the Q's are probability measures and the f's are bounded, we will also get bounds for the $L^2(Q)$ norms.

Given $0 < \epsilon < 1$, divide the interval $[-1, 1]$ (where the functions $f \in \mathcal{F}$ take values) into $q = [2/\epsilon] + 1$ subintervals with extreme points in the set

$$R_\epsilon = \{0, \epsilon, -\epsilon, 2\epsilon, -2\epsilon, \ldots, -1, 1\}.$$

Let us also consider a finite sequence of q_1 balls defined by

$$B(0, \epsilon) \subset B(0, 2\epsilon) \subset \ldots \subset B(0, r)$$

Observe that q_1 is either r/ϵ or $r/\epsilon + 1$.

Let $\mathcal{F}_m = \{f_1, \ldots, f_m\}$ be a class of functions taking values in R_ϵ such that every f_i is constant on the domains $B(0, \epsilon)$, $B(0, 2\epsilon) \setminus B(0, \epsilon)$, $B(0, 3\epsilon) \setminus B(0, 2\epsilon)$,... and the differences between the values of f_i on two adjacent domains (for example on $B(0, \epsilon)$ and $B(0, 2\epsilon) \setminus B(0, \epsilon)$) is at most ϵ. Note that $\#(\mathcal{F}_m) = m \leq q3^{q_1}$.

We have that for every $f \in \mathcal{F}$ there exists $i \in \{1, \ldots, m\}$ such that $\|f - f_i\|_\infty \leq 2\epsilon$. Indeed, given $f \in \mathcal{F}$, there exists $y_0 \in R_\epsilon$ such that $|f(0) - y_0| < \epsilon$. Now let $\mathcal{G}_0$ be the set of all functions f_i in $\mathcal{F}_m$ such that $f_i(0) = y_0$. As f has Lipschitz constant 1, we have $\sup_{x \in B(0,\epsilon)} |f(x) - g(x)| \leq 2\epsilon$. On the other hand, as

$$\sup_{x \in B(0,\epsilon)} f(x) \leq f(0) + \epsilon \text{ and } \inf_{x \in B(0,\epsilon)} f(x) \geq f(0) - \epsilon,$$

the class $\mathcal{G}_1$ of all functions $g \in \mathcal{G}_0$ such that $\sup_{x \in B(0,2\epsilon)} |f(x) - g(x)| \leq 3\epsilon$ is not empty. In a similar way, by the Lipschitz property of f, we can choose a non-empty class $\mathcal{G}_2 \subset \mathcal{G}_1$ such that $\sup_{x \in B(0,3\epsilon)} |f(x) - g(x)| \leq 3\epsilon$, for all $g \in \mathcal{G}_2$. By recurrence, define the (non-empty) class $\mathcal{G}_{q_1 - 1}$ of functions such that $\sup_{x \in B(0,r)} |f(x) - g(x)| \leq 3\epsilon$ for all $g \in \mathcal{G}_{q_1 - 1}$.

Thus we have shown

$$N(3\epsilon, \mathcal{F}) \leq q3^{q_1} \leq \left(\frac{2}{\epsilon}+1\right)3^{r/\epsilon+1} \leq \exp\left(\frac{C}{\epsilon^{1+\eta}}\right), \tag{7}$$

for all $\eta \in (0,1)$, and $C=(2r)^{1+\eta}\log 3$. Finally, using Yukich's [29] Theorem 1, (observe that $2-\delta$ in (5) has been denoted $1+\eta$ in (7)), we conclude (6).

3 A bootstrap validity result for functional data

We establish now a validity result for function-valued statistics defined on functional data. The methodology will be based on differentiability arguments very much in the line of [19]. As pointed out in the introduction, the functional version of the DKW-inequality obtained in the previous section will be a crucial step in the proof.

Theorem 3.1. *Let $\mathcal{H}$ be a bounded set in a Banach space (endowed with the Borel σ-algebra). Let $\mathcal{P}(\mathcal{H})$ be the set of all probability measures whose support is included in $\mathcal{H}$. Let T be an operator defined in $\mathcal{P}(\mathcal{H})$ with values in another Banach space $\mathcal{C}$. Denote by P_n the empirical measure corresponding to i.i.d. $\mathcal{H}$-valued variables with distribution P. Let P_n^* be the corresponding empirical associated with a bootstrap sample $X_1^*,\ldots,X_n^*$.*

(a) Assume that T satisfies the following differentiability condition for some given $P \in \mathcal{P}(\mathcal{H})$,

$$T(Q)=T(P)+T'_P(Q-P)+o(d(Q,P)), \tag{8}$$

where the remainder term $o(d(Q,P))$ denotes, as usual, an operator such that

$$\lim_{Q\to P}\frac{o(d(Q,P))}{d(Q,P)}=0,$$

and $T'_P : \mathcal{P}(\mathcal{H}) \longrightarrow \mathcal{C}$ is a linear (not necessarily continuous) operator for which is valid the bootstrap for the sample mean in the sense that

$$\sqrt{n}T'_P(P_n^*-P_n) \text{ converges weakly a.s. to the same limit as } \sqrt{n}T'_P(P_n-P). \tag{9}$$

Then,

$$\sqrt{n}\,(T(P_n^*)-T(P_n)) \xrightarrow{w} Z, \tag{10}$$

Z being the weak limit of $\sqrt{n}\,(T(P_n)-T(P))$.

(b) Assume that the operator T takes values in a separable Hilbert space $\mathcal{C}$ and it is differentiable in the sense (8). If the function $\Psi(x)=T'_P(\delta_x-P)$ is bounded (δ_x being the degenerate distribution at x), then condition (9) is fulfilled and therefore (10) holds.

Proof: (a) The result is a simple consequence of Theorem 2.1. Indeed, using the differentiability assumption (8),

$$T(P_n) = T(P) + T'_P(P_n - P) + o(d(P_n, P))$$

and

$$T(P_n^*) = T(P) + T'_P(P_n^* - P) + o(d(P_n^*, P)).$$

Hence

$$\sqrt{n}\,(T(P_n^*) - T(P_n)) = \sqrt{n}T'_P(P_n^* - P_n) + \sqrt{n}o(d(P_n^*, P)) + \sqrt{n}o(d(P_n, P)). \tag{11}$$

The first term in the right-hand side tends, by assumption (9), to the same limit as $\sqrt{n}\,(T(P_n) - T(P))$. Also, from the triangle inequality, $\sqrt{n}d(P_n^*, P)$ is bounded in probability (uniformly on P), as both $\sqrt{n}d(P_n^*, P_n)$ and $\sqrt{n}d(P_n, P)$ are. Therefore the remainder terms in (11) tend to zero in probability almost surely, which concludes the proof of (a).

(b) Since the operator T'_P is linear,

$$\sqrt{n}T'_P(P_n^* - P_n) = \sqrt{n}\left(\sum_{i=1}^{n} \Psi(X_i^*) - \Psi(X_i)\right).$$

Then, we may apply Theorem 3.1 in [21] to conclude that (9), and therefore (10), holds in this case.

Some final remarks:

(i) The hypothesis of uniform boundedness is not very restrictive in practice. It is in some sense similar to the assumption of compact support in nonparametric estimation. If one is willing to renounce to the usual gaussian models (which is also the case in nonparametrics) the hypothesis of boundedness looks quite natural as every observable phenomenon provides in fact observations taking values in a bounded domain (whose limits are imposed by the measurement instruments). From a technical point of view, boundedness is required for Theorem 2.1 (in order to be able to apply the entropy argument involved in the proof) and also for the result by Politis and Romano ([21], Theorem 3.1) used in the proof of part (b). Note also that the boundedness condition must be fulfilled in the metric of the space where the random elements X_i take values. For example, if this space is $L^2[a,b]$ the assumption that $X_i \in \mathcal{H}$, where $\mathcal{H}$ is bounded in $L^2[a,b]$, does not entail that the realizations of X_i have to be bounded in the supremum sense.

(ii) The above theorem can be applied, for example, to show the validity of the bootstrap for statistics of type $g(\bar{X})$ which may arise in different problems theoretical and applied. In particular, this type of statistics could appear if we are looking for robust alternatives (similar to M-estimators) for the sample mean in a functional data setup. Such functional statistics are often called Z-estimators; see [28], ch. 3.3. Since they are usually defined in an implicit way (as the solution of a functional equation) the effective use of our validity theorem for them would require an additional result in order to ensure that the required differentiability conditions are fulfilled. A detailed study on the asymptotic behavior of Z-estimators can be found in [30].

(iii) As an example of a differentiable operator $T = T(P)$ let us consider the variance operator

$$T(P)(t) = \int X^2(t,\omega)dP(\omega) - \mu_P^2(t),$$

where $X(t) = X(t,\omega)$ is a process with distribution P and mean function $\mu_P(t)$. It can be easily seen that the differential T'_P is the linear operator given by

$$T'_P(Q)(t) = \int X^2(t,\omega)dQ(\omega) - \mu_P^2(t).$$

References

[1] Abraham C., Cornillon P.A., Matzner-Lober E., Molinari N. (2003). *Unsupervised Curve Clustering using B-Splines*. Scandinavian Journal of Statistics, **30**, 581 − 595.

[2] Arcones M. A., Giné, E. (1992). *On the bootstrap of M-estimators and other statistical functionals*. In *Exploring the limits of bootstrap* (Edited by Raoul Le Page and Lynne Billard), Wiley, New York, 13 − 47.

[3] Bickel, P. J., Freedman, D. A. (1981). *Some asymptotic theory for the bootstrap*. The Annals of Statistics **9**, 1196 − 1217.

[4] Billingsley P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[5] Boente G., Fraiman R. (2000). *Kernel-based functional principal components*. Statistics and Probability Letters **48**, 335 − 345.

[6] Cardot H. Ferraty F., Sarda P. (1999). *Functional linear model*. Statistics and Probability Letters **45**, 11 − 22.

[7] Cardot H, Ferraty F., Mas A., Sarda P. (2003). *Testing hypotheses in the functional linear model*. Scandinavian Journal of Statistics **30**, 241 − 255.

[8] Cardot H., Sarda P. (2003). *Estimation in generalized linear models for functional data via penalized likelihood*. Journal of Multivariate Analysis, to appear.

[9] Cuevas A., Febrero M., Fraiman R. (2002). *Linear functional regression: the case of fixed design and functional response*. Canadian Journal of Statistics **30**, 285 − 300.

[10] Cuevas A., Febrero M., Fraiman R. (2004). *An anova test for functional data.* Computational Statistics and data Analysis, to appear.

[11] Dauxois J., Pousse A., Romain Y. (1982). *Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference.* Journal of Multivariate Analysis **12**, 136 – 154.

[12] Fan J., Lin S.K. (1998). *Test of significance when the data are curves.* Journal of the American Statistical Association **93**, 1007 – 1021.

[13] Ferraty F., Vieu P. (2002). *The functional nonparametric model and application to spectrometric data.* Computational Statistics **17**, 545 – 564.

[14] Fraiman R., Muniz, G. (2001). *Trimmed means for functional data.* Test **10**, 419 – 440.

[15] Giné E., Zinn J. (1990). *Bootstrapping general empirical measures.* The Annals of Probability **18**, 851 – 869.

[16] Kneip A., Gasser T. (1992). *Statistical tools to analyze data representing a sample of curves.* The Annals of Statistics **20**, 1266 – 1305.

[17] Locantore N., Marron J.S., Simpson D.G., Tripoli N., Zhang J.T., Cohen K.L. (1999). *Robust principal component analysis for functional data (with discussion).* Test **8**, 1 – 74.

[18] Muñoz-Maldonado Y., Staniswalis J.G., Irwin L.N., Byers, D. (2002). *A similarity analysis of curves.* Canadian Journal of Statistics **30**, 373 – 381.

[19] Parr W. C. (1985). *The bootstrap: some large sample theory and connections with robustness.* Statistics and Probability Letters **3**, 97 – 100.

[20] Pezzulli S., Silverman, B.W. (1993). *Some properties of smoothed principal components analysis for functional data.* Computational Statistics **8**, 1 – 16.

[21] Politis D.N., Romano J.P. (1994). *Limit theorems for weakly dependent Hilbert space valued random variables with application to the stationary bootstrap.* Statistica Sinica **4**, 461 – 476.

[22] Ramsay J.O., Silverman B.W. (1997). *Functional data analysis.* Springer-Verlag, New York.

[23] Ramsay J.O., Silverman B.W. (2002). *Applied functional data analysis.* Springer-Verlag, New York.

[24] Sheehy A., Wellner J.A. (1992). *Uniform Donsker classes of functions.* The Annals of Probability **20**, 1983 – 2030.

[25] Silverman B.W. (1996). *Smoothed functional principal components analysis by choice of norm.* The Annals of Statistics **24**, 1 – 24.

[26] Singh K. (1981). *On the asymptotic accuracy of Efron's bootstrap.* The Annals of Statistics **9**, 1187 – 1195.

[27] van der Vaart A. (2000). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

[28] van der Vaart A., Wellner J. (1996). *Weak convergence and empirical processes.* Springer-Verlag, New York.

[29] Yukich J.E. (1986). *Uniform exponential bounds for the normalized empirical process.* Studia Mathematica **84**, 71 − 78.

[30] Zhan Y. (2002). *Central limit theorems for functional Z-estimators.* Statistica Sinica **12**, 609 − 634.

Acknowledgement: The first author has been partially supported by grant BFM2001-0169 from the Spanish Ministry of Science and Technology.

Address: A. Cuevas, Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049-Madrid (Spain).
R. Fraiman, Departamento de Matemática, Universidad de San Andrés, Vito Dumas 284, Victoria, Provincia de Buenos Aires (Argentina).

E-mail: `antonio.cuevas@uam.es, rfraiman@udesa.edu.ar`

A NOVEL APPROACH TO PARAMETRIZATION AND PARAMETER ESTIMATION IN LINEAR DYNAMIC SYSTEMS

Manfred Deistler, Thomas Ribarits and Bernard Hanzon

Key words: Identification, parametrization, multivariate state space systems.

COMPSTAT 2004 section: Time series analysis.

Abstract: We describe a novel approach, called data driven local coordinates (DDLC), for parametrizing linear systems in state space form, and we analyze some of its properties which are relevant for e.g. maximum likelihood estimation. In addition we describe how this idea can be used for a concentrated likelihood function, obtained by a least squares type concentration step, which gives the so called sls (separable least squares) DDLC approach. Both approaches give favourable results in numerically optimizing the likelihood function in simulation studies.

1 Introduction

Despite the fact that identification (in the sense of model selection and parameter estimation) of linear dynamic systems is a quite mature subject now, there still exist severe problems in applying identification procedures, in particular in the multivariable case.

As is well known, one of the major problems is the 'curse of dimensionality'; in the (linear) multivariable case the dimension of the parameter space is a quadratic function of the number of outputs, unless additional restrictions, e.g. of factor analysis – or reduced rank regression type or of 'structural' type are imposed.

In this contribution our main focus will be on another important issue. For simplicity of notation, we only consider linear systems with unobserved white noise inputs. Then the most common models are AR, ARMA and state space (StS) models. In applications AR models still dominate, mainly for two reasons:

(i) The structure of parameter spaces for AR models is much simpler than in the case of ARMA and StS models. In particular, in the most common parametrization of AR(p) models (where the coefficient matrix of the present output is the identity) the entries of all other coefficient matrices are free parameters (of course satisfying the stability condition) and identifiable, including the parameters corresponding to the lower dimensional systems.

(ii) The maximum likelihood method gives least squares-type estimators, which are asymptotically efficient and numerically robust and fast; in other words parameter estimation is simple.

On the other hand ARMA and StS systems are more flexible and thus in many cases less parameters may be required.

As is well known every causal (stable) rational transfer function (describing the input-output behaviour of a linear system) can be described by an ARMA or a StS system; in this sense ARMA and state space system are equivalent. However, when embedded in 'naive' parameter spaces, typically the classes of observational equivalence are larger in the state space case. For instance, in the univariate case, for ARMA (n, n) systems, the equivalence classes are singletons in $\mathbb{R}^{2n}$ for the ARMA case (unless common factors occur), whereas they are n^2 dimensional manifolds for (minimal) state space systems in the embedding $\mathbb{R}^{2n+n^2}$. Identifiability is obtained by selecting representatives from equivalence classes and the advantage of large equivalence classes lies in the possibility to select (in some sense) better representatives. This is the reason why we here restrict ourselves to StS systems.

Both, typical ARMA and StS model classes suffer from the fact that the parametrization problem is non-trivial and that in general no explicit formula for the maximum likelihood estimator exists. For instance, in general, the boundary of the identifiable parameter spaces contains lower dimensional systems, which are not identifiable and algorithmic problems occur if the true system is close to the boundary. Some of these problems cannot be fully understood in the framework of the usual asymptotic analysis or are even better reflected by numerical rather than by statistical analysis. In a certain sense, asymptotic properties are parametrization independent, to be more precise:

(i) Under general assumptions, consistency can be shown for transfer functions in a coordinate-free way (see e.g. [2]); if we have identifiable parameter spaces and the function attaching parameters to transfer functions is continuous, then the corresponding parameter estimates are consistent, independent of the choice of the particular parametrization.

(ii) Under certain conditions the asymptotic variances of the maximum likelihood estimators change in a well defined way.

On the other hand a number of numerical properties are parametrization dependent. Numerical problems may arise for instance if the grid is too coarse in relation to the curvature of the likelihood function or if the likelihood function has 'long valleys' in relevant parts of the parameter space. It can be shown (see e.g. [4], [8]) that the choice of the parametrization has a severe impact on e.g. success rates or the number of iterations in numerical optimization of the likelihood function.

In the following we present two 'data driven' parametrizations as a contribution to the aim of increasing the 'market penetration' for state space modelling in applications.

2 Parametrization by state space systems

A common approch is to commence from the model class U_A of all causal and rational $s \times s$ transfer functions

$$k(z) = \sum_{j=0}^{\infty} K_j z^j \tag{1}$$

For a number of reasons, e.g. in order to obtain finite dimensional parameter spaces, U_A has to be broken into bits, where each bit is parametrized separately. In many cases, in a first step, the subclasses $M(n)$ of all transfer functions of order n are considered. Here we deal with parametrizations of $M(n)$ via state space systems (in innovations form):

$$\begin{aligned} x_{t+1} &= Ax_t + B\varepsilon_t && (2)\\ y_t &= Cx_t + \varepsilon_t && (3) \end{aligned}$$

where y_t is the s-dimensional observed output, x_t is the n-dimensional state and ε_t is (unobserved) s-dimensional white noise with $E\varepsilon_t\varepsilon_t' = \Sigma > 0$.

Usually it is assumed that

$$|\lambda_{\max}(A)| < 1 \quad \text{(stability)} \tag{4}$$

and

$$|\lambda_{\max}(A - BC)| < 1 \quad \text{(strict minimum phase assumption)} \tag{5}$$

hold. Here $\lambda_{\max}(D)$ denotes an eigenvalue of D of maximal modulus. However, mainly for the sake of notational simplicity, we here do not impose (4) and (5). For the stable case, the steady state solution of (1) is given by

$$y_t = \sum_{j=1}^{\infty} K_j\varepsilon_{t-j} + \varepsilon_t, \quad K_j = CA^{j-1}B \tag{6}$$

Let $S(n)$ denote the set of all $(A, B, C) \in \mathbb{R}^{n^2+2ns}$ (we identify (A, B, C) with $(vecA', vecB', vecC')'$). Clearly, $S(n) = \mathbb{R}^{n^2+2ns}$ and it can be shown that the set $S_m(n) \subseteq S(n)$ of all minimal (A, B, C) is open and dense in $\mathbb{R}^{n^2+2ns}$. Let us endow U_A with the pointwise topology, i.e. the topology corresponding to the product topology in the space $(\mathbb{R}^{s\times s})^N$ of power series coefficients $(K_j|j \in N)$ of the transfer functions. As can be shown, the closure $\bar{M}(n)$ of $M(n)$ satisfies $\bar{M}(n) = \cup_{i=1}^{n} M(i)$.

Finally, we define the mapping

$$\pi : S(n) \rightarrow \bar{M}(n) \tag{7}$$

by

$$\pi(A, B, C) = C(z^{-1}I - A)^{-1}B = k(z) \tag{8}$$

For describing $M(n)$ by state space systems the following approach (see e.g. [1] and [5]) may be used:

(i) Full state space parametrizations, i.e. $M(n)$ is described by $S_m(n)$. The drawback of this approach is that $S_m(n)$ is non-identifiable. The classes of observational equivalence are given by

$$\mathcal{E}(A, B, C) = \{(TAT^{-1}, TB, CT^{-1} | T \in GL(n)\} \tag{9}$$

and are real analytic manifolds of dimension n^2. Thus there are n^2 unnecessary parameters.

(ii) $M(n)$ can be shown to be a real analytic manifold of dimension $2ns$, which in general cannot be described by one coordinate system. One approach is to use socalled overlapping parametrizations, an alternative approach is the use of canonical forms, such as echelon form. In both cases a model selection procedure has to be applied in order to select a subclass of $M(n)$ from a fixed finite number of subclasses.

(iii) The approach described here, namely data driven local coordinates DDLC, (see [3], [4]) is as follows: We commence from an initial (minimal) $(A, B, C) \in S_m(n)$ and the tangent space to the equivalence class $\mathcal{E}(A, B, C)$ at (A, B, C). (A, B, C) may be obtained by an initial estimate, using e.g. a subspace or an instrumental variable estimation method. Then we take the orthocomplement (in $S(n)$) to the tangent space as (preliminary) parameter space: Let $Q^{\perp}$ denote a $(n^2 + 2ns) \times 2ns$ matrix whose columns form a basis for this orthocomplement. Then we have the parametrization:

$$\begin{aligned} \varphi_D : \mathbb{R}^{2ns} &\rightarrow S(n) \\ \tau_D &\mapsto \begin{pmatrix} vecA(\tau_D) \\ vecB(\tau_D) \\ vecC(\tau_D) \end{pmatrix} = \begin{pmatrix} vecA \\ vecB \\ vecC \end{pmatrix} + Q^{\perp} \cdot \tau_D \end{aligned} \tag{10}$$

The corresponding parameter space $T_D \subseteq \mathbb{R}^{2ns}$ is defined by removing the non-minimal systems and the corresponding space for transfer functions is $V_D = \pi(\varphi_D(T_D))$.

The intuitive motivation behind the $DDLC$ approach is that, due to orthogonality to the tangent space, the numerical properties of optimization based estimators, such as the maximum likelihood estimator, are at least locally favourable. Comparisons with other parametrizations corroborate this notion (see e.g. [4] and [8]). In particular these comparisons show that echelon forms (whose parameters correspond to the usual ARMA parameters) are clearly outperformed. $DDLC$ is now the default option in the system identification toolbox in $MATLAB6.x$. The success of $DDLC$ was the motivation for a careful investigation of the topological and geometrical properties of $DDLC$ relevant for estimation, described in the next section.

3 Topological and geometrical properties of DDLC

Important properties of $DDLC$ are summarized in the following theorem: see [5], [9].

Theorem 3.1. *Let an initial minimal system* (A, B, C) *be given. Then the parametrization by* $DDLC$ *as given in (10) has the following properties:*

(i) T_D *is an open and dense subset of* $\mathbb{R}^{2ns}$*.*

(ii) There exist open neighborhoods $T_D^{loc} \subseteq T_D$ *of* $0 \in T_D$ *and* V_D^{loc} *of* $\pi(A, B, C)$ *in* $M(n)$ *such that* T_D^{loc} *is identifiable,* $V_D^{loc} = \pi(T_D^{loc})$ *and the mapping* $\psi_D^{loc} : V_D^{loc} \to T_D^{loc}$ *defined by* $\psi_D^{loc}(\pi(\tau_D)) = \tau_D$ *is a homeomorphism.*

(iii) For $n > 0$, $\pi(\bar{T}_D)$ *contains transfer functions of lower McMillan degree.*

(iv) There exists an open and dense subset V_D^{fin} *of* V_D *such that for every* $k \in V_D^{fin}$*, the corresponding equivalence class in* T_D *consists of a finite number of points.*

(v) V_D° *is dense in* V_D*, where* V_D° *denotes the interior of* V_D *in* $\mathbb{M}(n)$*. Additionally,* V_D *is open (and trivially dense) in* $\pi(\bar{T}_D)$*, but not necessarily open in* $M(n)$*.*

(vi) $\pi(\bar{T}_D) \subseteq \bar{V}_D$*, where equality can hold, but the inclusion may also be strict.*

In a certain sense this theorem is an analogue to the theorems given in [2] for the overlapping description of $M(n)$ and for echelon forms. We give a short discussion of the consequences of the results of Theorem 3.1:

(i) Openness means that the parameters are free and in particular not restricted to a thin subset of $\mathbb{R}^{2ns}$. This is an important requirement for gradient-type optimization procedures to work properly. Note that openness also holds if the stability assumption (4) and the miniphase assumption (5) are imposed. Clearly then denseness will not hold.

(ii) states that there exist neighborhoods T_D^{loc} and V_D^{loc} where the parametrization is well-posed in the sense of being injective (and thus identifiable) and the parameters are attached to transfer functions in a continuous way. In particular 'coordinate free' consistency of transfer function estimates in V_D^{loc} (see [2]) then implies consistency of the corresponding parameter estimates. However, we have no statements concerning the size of T_D^{loc} and V_D^{loc}, respectively.

(iii) For $n > 0$, the following holds: The closure of the parameter space T_D – note that $\bar{T}_D = \mathbb{R}^{2ns}$ – corresponds to transfer functions of equal *and* lower McMillan degrees. The equivalence classes in $\bar{T}_D \setminus T_D$ are generally given by nonlinear restrictions and are thus difficult to describe.

(iv) In general, T_D is not identifiable; as a 'second best' result, the equivalence classes for a generic subset V_D^{fin} are at least finite and thus consist of isolated points in T_D.

(v) deals with the structure of the set V_D; for a discussion of the relevance of (v) see [9].

(vi) The fact that $\bar{V}_D$ may contain more transfer functions than those described by the closure of the parameter space T_D can effect the actual estimation procedure. In that case the norm of the parameter vector may diverge to infinity whereas the corresponding sequence of transfer functions converges to a well defined transfer function estimate in $M(n)$. Problems of nonconvergence of algorithms due to this phenomenon have actually puzzled researchers in the past when using echelon canonical forms.

4 Separable least squares for ML-type estimation

One way to reduce the dimension of the parameter space over which a likelihood function has to be numerically optimized is to concentrate out parameters which enter linearly by an (ordinary or generalized) least squares step. For the concentrated likelihood again the $DDLC$-approach is used; see [5], [8] and [7].

Here, we commence from the inverse state space system

$$\begin{aligned} x_{t+1} &= \overbrace{(A - BC)}^{\bar{A}} x_t + \overbrace{B}^{\bar{B}} y_t \\ \varepsilon_t &= \underbrace{-C}_{\bar{C}} x_t + y_t \end{aligned} \tag{11}$$

and the corresponding parameters $(\bar{A}, \bar{B}, \bar{C})$ are in a one-to-one relation with

(A, B, C). The (Gaussian) conditional likelihood function is of the form

$$L_T(\bar{A}, \bar{B}, \bar{C}, \Sigma) = \log\det\Sigma + \frac{1}{T}\sum_{t=1}^{T} tr\left\{\varepsilon_t(\bar{A}, \bar{B}, \bar{C})\varepsilon_t(\bar{A}, \bar{B}, \bar{C})'\Sigma^{-1}\right\} \quad (12)$$

Substituting a consistent estimate $\hat{\Sigma}$ for Σ, we get an approximation of this criterion function, which we again denote by L_T. Because $\bar{B}$ enters linearly in L_T, we obtain

$$vec\hat{\bar{B}} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'Y_1^T \quad (13)$$

by minimizing L_T with respect to $\bar{B}$ for fixed $\bar{A}$ and $\bar{C}$. Here, $Y_1^T = (y_1', \ldots, y_T')'$ is the stacked vector of observations and $\tilde{X}$ depends on Y_1^T, $\bar{A}$, $\bar{C}$ and $\hat{\Sigma}$ and we assume that $\tilde{X}$ has full column rank. This leads to the following new criterion function depending on $\bar{A}$ and $\bar{C}$ only:

$$L_T^{cc}(\bar{A}, \bar{C}) = tr\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t(\bar{A}, \bar{C})\varepsilon_t(\bar{A}, \bar{C})'\hat{\Sigma}^{-1} \quad (14)$$

Given the (observable) pair $(\bar{A}, \bar{C})$, Y_1^T and $\hat{\Sigma}$, we obtain the original system by $\Delta_Y(\bar{A}, \bar{C}) = (A, B, C) = (\bar{A} - \hat{\bar{B}}\bar{C}, \hat{\bar{B}}, -\bar{C})$. The pairs $(\bar{A}_1, \bar{C}_1)$ and $(\bar{A}_2, \bar{C}_2)$ are called observationally equivalent if they correspond to the same transfer function, i.e. if $\pi(\Delta_Y(\bar{A}_1, \bar{C}_1)) = \pi(\Delta_Y(\bar{A}_2, \bar{C}_2))$. If $(\bar{A}, \bar{C})$ is observable, then, under certain additional assumptions, all observational equivalent pairs are given by $\mathcal{E}_{cc}(\bar{A}, \bar{C}) = (T\bar{A}T^{-1}, \bar{C}T^{-1})$, $T \in GL(n)$.

$\mathcal{E}_{cc}(\bar{A}, \bar{C})$ is a real analytic manifold of dimension n^2 and the *DDLC* construction is performed again by taking the orthocomplement in $\mathbb{R}^{n^2+ns}$ to the tangent space of $\mathcal{E}_{cc}(\bar{A}, \bar{C})$ at an initial point $(\bar{A}, \bar{C})$. Let us denote the new parameter space, where the non-minimal systems have been removed, again by $T_D \subseteq \mathbb{R}^{ns}$ and let us put $V_D = \pi(\Delta_Y(\varphi_D(\tau_D)))$. Here, φ_D is given by

$$\begin{array}{rcl} \varphi_D : T_D & \rightarrow & \mathbb{R}^{n^2+ns} \\ \tau_D & \mapsto & \begin{pmatrix} vec\bar{A}(\tau_D) \\ vec\bar{C}(\tau_D) \end{pmatrix} = \begin{pmatrix} vec\bar{A} \\ vec\bar{C} \end{pmatrix} + Q^{\perp} \cdot \tau_D \end{array} \quad (15)$$

where $Q^{\perp} \in \mathbb{R}^{n^2+ns\times ns}$ is now a matrix with orthonormal columns spanning the new orthocomplement to $\mathcal{E}_{cc}(\bar{A}, \bar{C})$ at the point $(\bar{A}, \bar{C})$; (15) is called the *slsDDLC* parametrization. For the following theorem see [7]:

Theorem 4.1. *Let Y_1^T and an initial (A, B, C) be given and let $(\bar{A}, \bar{B}, \bar{C})$ denote the corresponding inverse system in (11). The parametrization by slsDDLC as given in (15) has the following properties:*

(i) T_D *is an open and dense subset of* $\mathbb{R}^{ns}$.

(ii) *There exist open neighborhoods* $T_D^{loc} \subseteq T_D$ *of* $0 \in T_D$ *and* V_D^{loc} *of* $\pi(A,B,C)$ *in* V_D *such that* T_D^{loc} *is identifiable,* $V_D^{loc} = \pi(T_D^{loc})$ *and the mapping* $\psi_D^{loc} : V_D^{loc} \to T_D^{loc}$ *defined by* $\psi_D^{loc}(\pi(\tau_D)) = \tau_D$ *is a homeomorphism.*

(iii) $\pi(\bar{T}_D \cap T_X)$ *may (but need not necessarily) contain transfer functions of lower McMillan degree.*

(iv) *There exists an open and dense subset* V_D^{fin} *of* V_D *such that for every* $k \in V_D^{fin}$*, the corresponding equivalence class in* T_D *consists of a finite number of points.*

(v) $\pi(\bar{T}_D \cap T_X) \subseteq \bar{V}_D$*, where equality can hold, but the inclusion may also be strict.*

Here, T_X *denotes the (generic) set of* τ_D *such that* $\tilde{X}$ *has full column rank.*

Note that here, as opposed to ordinary $DDLC$, V_D^{loc} is not open in $M(n)$.

An alternative procedure is to concentrate out $\bar{C}$. In this case, ML estimation of Σ can also be incorporated; see [8].

5 A numerical comparison

Eight different minimal, stable and strictly minimum phase state space models (A,B,C) with two outputs are specified. The models are denoted by M_1, ..., M_8 and are of order 2, 4, ..., 16. The poles and zeros are quite close to each other and close to the unit circle, but they do not cancel.

Simulation data for models M_1, ..., M_8 comprising $T = 500$ output observations are created, where the white noise sequence (ε_t) is chosen to be Gaussian distributed with $\Sigma = I_2$.

In the next step, 50 random initial state space models are created by randomly perturbing the matrices (A,B,C) corresponding to the true systems. It is ensured that the perturbed models remain minimal, stable and minimum phase.

All computations are carried out using the system identification toolbox of the software package `MATLAB`, version `6.5.0.180913a (R13)`. The identification procedure itself is performed by using the built-in function `pem`. The option `'SearchDirection'` is set to `'Gn'` (a plain Gauss-Newton type algorithm is used for minimizing the criterion function). For a more detailed discussion of the simulation results presented below we refer to [6]. We confine ourselves to the following statements: $slsDDLC$ leads to

- better success rates. Note that an identification experiment is considered to have failed if the algorithm yields a final parameter estimate with a likelihood value larger than 1.2 times the value of the asymptotic likelihood at the true system; see Table 1 (A). Note e.g. that for *slsDDLC* only 1 out of 400 estimation runs failed, whereas usage of *DDLC* leads to 8 failed runs.

- fewer iterations until convergence due to the reduction of the dimension of the parameter space; see Table 1 (B).

- the lowest condition numbers of the Gauss-Newton approximation to the Hessian of the criterion function. Note that the echelon canonical form *Can* lead to the highest condition numbers; see Table 1 (D).

- better estimates, i.e. better or at least equally good values of the likelihood function at convergence; see Table 1 (F).

In total, the echelon canonical form performs worst, and *slsDDLC* is slightly better than *DDLC*. However, the actual computations turn out to be more time-consuming for *slsDDLC*, but still remain within a feasible range.

A	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Can	0	78	18	8	50	28	74	68
DDLC	0	0	4	0	0	0	4	8
slsDDLC	0	0	0	0	0	2	0	0

B	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Can	8	24	18	20	27	35	35	28
DDLC	6	10	13	9	21	18	12	12
slsDDLC	8	9	10	8	16	13	8	8

C	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Can	0	39	22	23	23	32	33	28
DDLC	0	0	12	0	0	0	6	8
slsDDLC	0	0	0	0	0	6	0	0

D	M_2	M_3	M_4	M_5
Can	1.4e + 12	1.3e + 10	4.9e + 16	1.1e + 17
DDLC	9.7e + 4	1.1e + 5	8.0e + 3	1.0e + 6
slsDDLC	3.5e + 1	5.4e + 2	2.8e + 2	7.8e + 2
D	M_6	M_7	M_8	
Can	1.3e + 16	4.2e + 16	6.1e + 20	
DDLC	7.9e + 5	1.1e + 6	3.8e + 6	
slsDDLC	3.3e + 2	6.3e + 2	3.2e + 3	

E	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Can	0.	1.3e + 17	3.9e + 19	1.4e + 19	1.3e + 21	3.3e + 18	8.1e + 21	1.6e + 23
DDLC	0.	0.	3.1e + 5	0.	0.	0.	1.1e + 6	3.4e + 5
slsDDLC	0.	0.	0.	0.	0.	5.8e + 2	0.	0.

F	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Can	8.02e − 1	1.13	1.16	8.88e − 1	1.17	1.14	1.03	1.1
DDLC	7.88e − 1	1.09	1.11	8.18e − 1	1.08	1.04	9.03e − 1	9.46e − 1
slsDDLC	7.75e − 1	1.09	1.11	8.18e − 1	1.08	1.02	8.67e − 1	9.35e − 1

G	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Can	0.	1.42	2.68	1.77	2.96	2.28	3.12	3.06
DDLC	0.	0.	1.34	0.	0.	0.	1.33	2.18
slsDDLC	0.	0.	0.	0.	0.	1.21	0.	0.

Table 1: Identification of ARMA-type models
(A) Percentage of failed runs out of 50 identification experiments.
(B) Average number of iterations for successful runs. Test cases with no successful runs are indicated by 0.
(C) Average number of iterations for failed runs. Test cases, where no run failed, are denoted by 0.
(D) Average maximum condition number of the Gauss-Newton approximations to the Hessians for succesful runs. Test cases with no successful runs are indicated by 0.
(E) Average maximum condition number of the Gauss-Newton approximations to the Hessians for failed runs. Test cases, where no run failed, are indicated by 0.
(F) Average criterion value for successful runs.
(G) Average criterion value for failed runs.

References

[1] Deistler M. (2000). *System identification – general aspects and structure.* In G. Goodwin (ed.), System Identification and Adaptive Control, Springer, London, 3 – 26. (Festschrift for B.D.O. Anderson).

[2] Hannan E.J., Deistler M. (1988). *The statistical theory of linear systems.* John Wiley & Sons, New York, 1988.

[3] McKelvey T., Helmersson A. (1997). *System identification using an over-parametrized model class – improving the optimization algorithm.* In Proc. 36th IEEE Conference on Decision and Control, San Diego, California, USA **3**, 2984 – 2989.

[4] McKelvey T., Helmersson A., Ribarits T. (2004). *Data driven local coordinates for multivariable linear systems and their application to system identification.* Forthcoming in Automatica.

[5] Ribarits T. *The role of parametrizations in identification of linear dynamic systems.* PhD thesis, TU Wien.

[6] Ribarits T., Deistler M. (2003). *A new parametrization method for the estimation of state-space models.*

[7] Ribarits T., Deistler M., Hanzon B. (2004). *An analysis of separable least squares data driven local coordinates for maximum likelihood estimation of linear systems.* Submitted to Automatica.

[8] Ribarits T., Deistler M., Hanzon B. (2004). *On new parametrization methods for the estimation of state-space models.* Forthcoming in Intern. Journal of Adaptive Control and Signal Processing.

[9] Ribarits T., Deistler M., McKelvey T. (2004). *An analysis of the parametrization by data driven local coordinates for multivariable linear systems.* Automatica **40** (5), 789 – 803.

Address: M. Deistler, T. Ribarits, Institute for Mathematical Methods in Economics, Research Unit Econometrics and System Theory (EOS), Vienna University of Technology, Argentinierstrasse 8, 1040 Vienna, Austria
B. Hanzon, Mathematical Institute, Leiden University, P.O. Pox 9512,
2300 RA Leiden, The Netherlands

E-mail: {`Manfred.Deistler,Thomas.Ribarits`}`@tuwien.ac.at,`
`bhanzon@math.leidenuniv.nl`

STATISTICAL ANALYSIS OF HANDWRITTEN ARABIC NUMERALS IN A CHINESE POPULATION

Wing K. Fung, C.T. Yang, C.K. Li and N.L. Poon

Key words: Writing habits, Arabic numerals, statistical study, classification system, test for independence, probability of occurrence.

COMPSTAT 2004 section: E-statistics.

Abstract: A sample of 187 subjects from the Chinese population in Hong Kong was selected to participate in a handwriting study of Arabic numerals. Characteristic features such as slant, direction of writing, angularity of turnings, directions of initial and/or ending strokes etc. were developed. A set of characteristic codes representing the profile of writing habits was assigned to each writer. Hierarchical cluster analysis was conducted on the characteristic features which were on nominal scale, and hence subjects who had similar writing habits for Arabic numerals were grouped. Pearson's χ^2 tests for independence for pairs of feature variables were constructed. The independence property allows us to estimate the probability of occurrence for certain characteristic features of a Arabic numeral. An alternative way for estimation is suggested when some of the features are not statistically independent.

1 Introduction

Writing habit, being a product of long-term adaptation to the needs and abilities of the writer, is believed to be unique. Various classification systems for handwriting have been suggested; see [3] for a review. A system for the classification of handwritten numerals have been developed by Ansell and Strach [1], and Strach [6]. Recently, computer algorithms for extracting features from scanned image of handwriting were used by Srihari et al. [5] for the analysis of individuality of handwriting.

It this paper, we analyse the characteristic features and codes of the Arabic numeral writings, i.e., $0, 1, \ldots, 9$, of 187 subjects. We give a detailed description on data collection and the methods of statistical analysis for the study. Hierarchical cluster analysis (Section 4.1) is conducted on the characteristic codes of the single numerals and the paired numerals. We define a cluster being the set of subjects having rescaled distance at the minimal level. From that, we can obtain the number of clusters in the dendrogram diagram, which is useful for measuring the variability of the numeral(s) in question.

We are also interested in investigating whether the characteristic features within each numeral are statistically independent. If the features are independent, it would provide a simple method to estimate the relative frequency or probability of occurrence of certain characteristic features, by simply multipling the marginal probabilities for individual features. χ^2 tests would be conducted for checking independence.

2 Data and methods

A sample of 187 subjects from the local Chinese population was selected to participate in the handwriting study. The subjects were asked to follow the instructions on the questionnaire provided to them. They had to write Arabic numerals, $0, 1, \ldots, 9$, as in their normal ways using their own pens. The specimens were collected together by the staff of the Hong Kong Government Laboratory for further examination. The collected numerals were then examined microscopically with the aid of a Nikon SMB-2B microscope by professional document examiners of the Hong Kong Government Laboratory. Characteristic features such as slant, direction of writing, angularity of turnings, directions of initial and/or ending strokes etc. were selected and a code was then assigned to each characteristic feature. Each feature normally has 2-5 possible assignments of characteristic codes.

Hierarchical cluster analysis is a set of statistical techniques that is particularly useful for separating a set of objects into constituent groups or clusters which minimize the variation between members of the same group [8] without making assumptions about the number of groups or the group structure. Grouping of objects into clusters is done on the basis of similarities or distances [8], [2]. The analysis is a powerful exploratory method commonly employed in many disciplines. The clustering method processes the values of the measure of similarities among pairs of objects, generating a tree or dendrogram that shows the hierarchy of similarities among all pairs of objects.

Since hierarchical cluster analysis is mainly designed for quantitative measurements, the code of the selected numeral was re-coded to a number of binary variables, with 1 referring to the presence of code status and 0 otherwise. Take for an example of numeral “0”, the original data size is 187 × 9 (number of subjects × number of characteristic features for “0”), and the recoded data has size 187 × 25 (25 codes for “0’). Proximity of dissimilarity is generated for binary data that ranges from 0 to 1 [4], [7], [2]. The dissimilarity measure is taken as the pattern difference of the fourfold table which is computed as $bc/(n^{**}2)$, where b and c refer to the diagonal cells corresponding to features present on a object but absent on the other and $n(= 187)$ is the total number of objects involving in the study [7]. Hierarchical cluster analysis measure for binary data was then employed using an algorithm that starts from all objects as apart and merge two clusters nearby until only one is left [4]. The default measure of pattern difference in SPSS was adopted for the binary codes to identify dissimilarity (notice that other similarity/dissimilarity measures have been attempted, and they give similar results to those reported in Section 4.1). Average linkage between objects was then taken to demonstrate the procedure of statistical classification. It defines the distance between two clusters as the average of the distances between all pairs of cases in which one member of the pair is from each of the clusters. This method uses information about all pairs of distances, not just the nearest or the farthest, and so it is usually preferred in cluster analysis [7]. Subjects having the similar (or the same) way of numeral writing were grouped/clustered together. A tree diagram or dendrogram was selected to present the results of cluster analysis, which is depicted horizontally with each row representing a case, and cases with high similarity are adjacent. The number of clusters and the cluster sizes were also measured.

Another question of interest is whether the quantified features are statistically independent of one another. Each feature is a nominal variable which can take different possible codes (normally 2-5). Pearson's χ^2 independence test for each pair of features is conducted. The evaluation on the probability of occurrence of certain characteristic features will be much simplified if the independence assumption is found to be valid.

3 Summary statistics

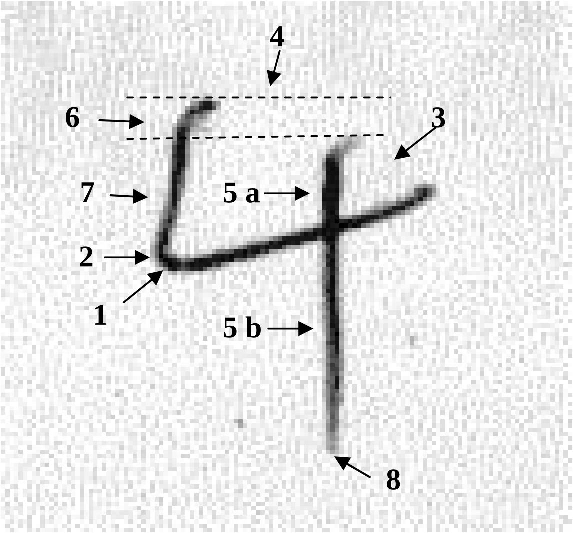
The majority of the participants were of young to middle age (20-49). Only 4% and 2% were of age < 20 and age ≥ 50 respectively. Nearly all (99%) of them were right-handed.

The assignment of characteristic features and codes is an important process in the project. Figure 1 shows an example of such an assignment for numeral 4. In this particular numeral, there are 8 characteristic features and each feature has 2-3 possible assignments of characteristic codes. Some of the features such as features 1, 2, 3, and 8 are relatively easy to distinguish, while others may need some comparisons on the length of the measurements. At the lower half of Figure 1, we have identified the code for each feature of that particular numeral. Moreover, one other code for each of the features is also provided for easy understanding. Numeral 4 has 8 characteristic features with 19 characteristic codes.

Table 1 gives an overall summary for the number of characteristic features and codes for the studied numerals; the detailed characteristic features and codes are omitted for brevity. We can see that numeral "1" is the simplest numeral as expected and it has 4 characteristic features (numeral "6" too), while numerals 5, 8 and 9 have the most features and codes.

Numeral	0	1	2	3	4	5	6	7	8	9
Feature	9	4	7	8	8	9	4	9	12	10
Code	25	11	19	20	19	30	10	23	33	31

Table 1: Numbers of characteristic features and codes for numerals 0-9.



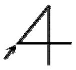
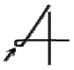
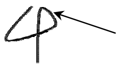
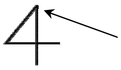
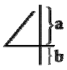
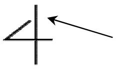
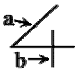
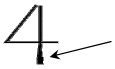
Feature	Code of the above numeral	One other code
1. Turning to the left	Round (r)	 Angular
2. Loop on the left	No (n)	 Yes
3. Connection between horizontal and vertical strokes	Open (o)	 Closed
4. Relation between slanting and vertical strokes	Open (o)	 Closed
5. Ratio of vertical stroke above & below the horizontal stroke a/b	a - Shorter (sh)	 a - Longer
6. Top part of vertical stroke relative to left slanting stroke	Shorter (sh)	 Taller
7. Left slanting stroke (a)/ portion of vertical stroke below horizontal stroke (b)	Shorter (sh)	 a - Longer
8. Ending of vertical stroke	Tapered (t)	 Blunted

Figure 1: Assignment of characteristic features and codes for numeral 4.

4 Statistical analysis

4.1 Hierarchical cluster analysis

We employ hierarchical cluster analysis for subjects grouping in the writing of numeral “4”. Figure 2 gives the dendrogram for the clustering of the subjects; for clarity, only the last 20 subjects were selected for classification. As noted from Figure 2, we have identified four tightly linked clusters namely, cluster a: subjects $(14, 20, 1, 8)$; cluster b: $(4, 7)$; cluster c: $(10, 11)$; and cluster d: $(3, 18, 9)$. The subjects within each cluster are very similar to each other and so they are grouped together. For example, we can see from the figure that subjects 14, 20, 1 and 8 are grouped together and after checking the original data we found that they in fact wrote in (exactly) the same pattern during the writing of numeral “4”. The same situation also happens to subjects 4 and 7, 10 and 11, and 3, 18 and 9. The dendrogram of the cluster analysis can give us information on the similarity/dissimilarity in the writing of “4” amongst the subjects.

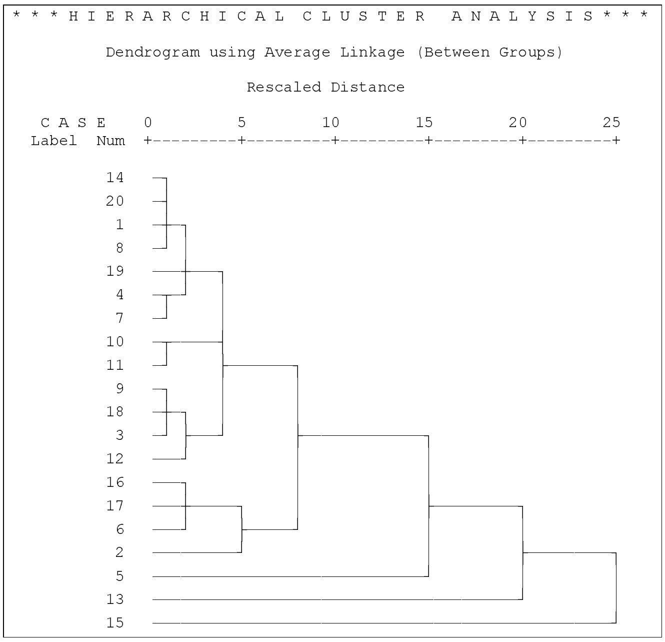


Figure 2: Clustering of the last 20 subjects for numeral 4.

We can also count the number of clusters formed in Figure 2. A cluster is defined as follows: a cluster is of size two or more subjects if its rescaled distance is at the minimal level or at the lowest part of the dendrogram and a cluster is of size 1 if otherwise. The clusters in Figure 2 are identified as, C1: (14, 20, 1, 8), C2: (19), C3: (4, 7), C4: (10, 11), C5: (9, 18, 3), and subjects

12, 16, 17, 6, 2, 5, 13 and 15 each form the remaining clusters C6, C7, ..., C13, respectively. Using the same procedure, 16 clusters are identified for the paired numerals 4 and 7 in the dendrogram of Figure 3 which is to be discussed in more details below.

Hierarchical cluster analysis is again conducted for two numerals "4" and "7" and the results are shown in Figure 3. As we can see that there are only two tightly linked clusters, less than that formed in Figure 2 for a single numeral "4". The clusters are, cluster i: subjects $(3, 18, 8, 12)$ and cluster ii: subjects $(1, 20)$. In fact, after checking the original data, we found that there were some differences in the way of writing of numerals "4" and "7" for subjects 3, 18, 8 and 12 of cluster i. The two subjects 1 and 20 of the other cluster also did not write exactly the same numerals "4" and "7".

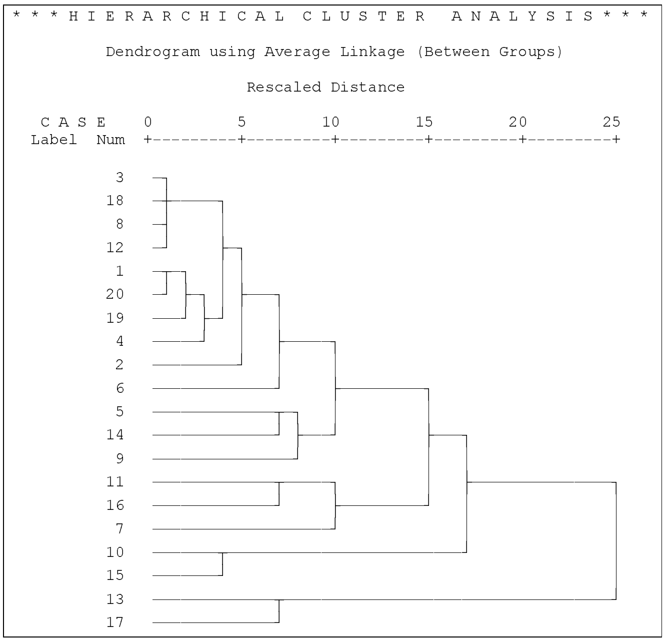


Figure 3: Clustering of the last 20 subjects for numerals 4 and 7.

Table 2 summarizes the findings obtained from hierarchical cluster analysis of Arabic numerals. The number of clusters formed and the maximum and the second maximum sizes of clusters are listed for reference. According to cluster analysis numeral "1" is the simplest handwriting character amongst all. Totally, there are only 36 clusters formed via the classification procedure with merely 8 clusters containing 5 or more subjects and the largest cluster involving 63 homogenous subjects. On the contrary, numeral "5", arming with 9 features and 30 codes, is the most informative character that help distinguish subjects' distinctiveness on handwriting.

Numerals	No. of features	No. of clusters	No. of clusters with size ≥ 5	Max. size of cluster	Second size
0	9	85	7	13	13
1	4	36	8	63	27
2	7	82	8	20	18
3	8	123	3	19	7
4	8	81	8	23	14
5	9	139	0	4	4
6	4	69	10	23	19
7	9	97	8	17	10
8	12	108	4	20	7
9	10	115	6	10	10

Table 2: Summary findings for cluster analysis of single Arabic numeral.

The combined numerals increase the number of characteristic features and codes on handwriting discrimination and enhance the heterogeneity among subjects. Table 3 summarizes the findings of cluster analysis of two Arabic numerals, 4 and others, demonstrating the dissimilarity reinforcement between subjects in handwriting. Comparing with the findings in Table 2, the combined Arabic numerals overall increases the number of clusters formed; that is, the subjects are more heterogeneous from one another than that in the writing of a single numeral. It is to be noted that for the combined numerals 4 and 5, 175 clusters (out of 187 subjects) are identified, which indicates that the handwritings of the 187 subjects for numerals 4 and 5 together are nearly all different (in one or more characteristic features).

Numerals	No. of features	No. of clusters	No. of clusters with size ≥ 5	Max. size of cluster	Second size of cluster
0, 4	17	133	2	7	5
1, 4	12	98	6	11	11
2, 4	15	147	0	4	4
3, 4	16	147	2	6	5
4, 5	17	175	1	5	2
4, 6	12	135	3	6	6
4, 7	17	140	3	7	6
4, 8	20	166	1	6	4
4, 9	18	149	1	8	4

Table 3: Summary findings for cluster analysis of two Arabic numerals.

4.2 Tests for independence and probability assessment

Next we would like to investigate whether the features within each numeral are statistically independent. Let us consider the simplest numeral "1" first. The common χ^2 test for independence is employed. There are four characteristic features measured for numeral 1, including slant, initial hook, serif and ending position. Since a large majority of the subjects (99%) wrote numeral 1 without the serif feature, this feature is excluded from the χ^2 test based on the common rule-of-five in applying the test. The other three features can form 3 sets of paired features, and the features of each pair are found to be statistically independent (each at $\alpha = 5\%$ significance level; details omitted). This sort of independence can be very important for evaluating the relative frequency or probability of occurrence of certain characteristic features. For example, we may evaluate the probability, because of independence, as

$$\begin{aligned} & P_{123}\ [(\text{Slant (S)} = \text{forward (f)}, \text{Initial Hook (IH)} = \text{right (r)}, \\ & \qquad \text{Ending Position (EP)} = \text{hook (h)}] \\ = \ & P_1(S = f) \quad P_2(IH = r) \quad P_3(EP = h) \\ = \ & 0.69 \times 0.14 \times 0.25 \\ = \ & 0.02415, \end{aligned}$$

where the marginal probabilities 0.69, 0.14 and 0.25 are obtained from the direct counting method. Of course there are many assumptions behind this estimate which might be rather crude. It is to be awared that the pairwise independence does not imply the features being all mutually independent. The estimate may also be adjusted upward if we want to make it conservative (forensic document examiners like to take the conservative approach in practice).

	1	2	3	4	5	6	7	8
1		ind	D	ind	D	D	D	ind
2			ind	ind	ind	ind	ind	ind
3				D	D	D	ind	ind
4					D	ind	ind	D
5						D	D	ind
6							D	ind
7								ind

Table 4: Results of χ^2 independence tests for features $1, \ldots, 8$ (meanings refer to text) of numeral "0", each at 5% significance level. D and ind stand for features being statistically dependent and independent respectively.

Next we investigate the numeral 0 which has nine features on (1) slant, (2) initial and ending strokes, (3) starting position, (4) ending position at right, middle or left, (5) stroke crossing position, (6) ending position being tapering or blunt, (7) shape, (8) ending position at upper half, middle or lower, and (9) the writing direction which is however omitted in our analysis because of the rule-of-five. The independence test results for feature pairs are summarized in Table 4. It is interesting to note that only feature 2, initial and ending strokes (being open or close), is statistically independent from other features considered. Feature 8 is (pairwise) independent of all other features except 4. Thus, it seems difficult to use a kind of (simple) product rule, as presented in numeral 1 where the assumption of feature independence is taken, to estimate the relative frequency or probability of occurrence of the particular characteristic features for numeral 0.

We suggest below an alternative way to (conservatively) estimate the probability of occurrence of the following feature codes for numeral 0,

$$\begin{aligned} & P_{1\,2\,3\,4\,5\,6\,7\,8}(f,c,l,m,l,t,e,u) \\ = \;& P_2(c)P_{1\,3\,4\,5\,6\,7\,8}(f,l,m,l,t,e,u) \\ < \;& P_2(c)P_{1\,3\,5\,6\,7\,8}(f,l,l,t,e,u) \\ = \;& P_2(c)P_{1\,3\,5\,6\,7}(f,l,l,t,e)P_8(u), \end{aligned}$$

where $P_2(c)$ and $P_8(u)$ can be evaluated easily, and $P_{1\,3\,5\,6\,7}(f,l,l,t,e)$ can also be estimated based on some direct counting from the sample. It is noted that similar assumptions as for numeral 1 may have to be made as well. Furthermore, we need to aware that the overall level of significance is not equal to 5%, though the individual level is set to be 5% for each paired comparison, because of multiple comparisons. Moreover, it may also not be reasonable to regard the features being statistically all independent, or all dependent.

Another question of interest is on the estimation of the probability of occurrence for characteristic feature codes of two or more numerals. We shall not attempt to answer this question due to the possible very complex dependence structure in the data.

5 Concluding remarks

We have investigated characteristic features of numerals $0,\ldots,9$. The Arabic numerals are chosen because they are commonly found in daily life. Hierarchical cluster analysis is used to classify subjects of similar handwriting features into groups. As expected, a subject is more difficult to cluster/group with others when more numerals are considered. In fact, the individuality of handwriting features may be identified in our sample when we consider 2 or 3 numerals together such as 5, 8, and 9 which have more characteristic features. This phenomenon may also be of interest to document examiners.

The χ^2 tests are constructed to see if the features are statistically pairwise independent. The features (except the serif feature) in numeral 1 are independent, while some features in numeral 0 are dependent of one another.

This dependence structure would also be found in other numerals. However, it is still possible to find some independence structure in the features such that the probability of occurrence of some characteristic features can be estimated, of which the possible limitations have to be awared. Furthermore, it is suggested that the probability should be estimated for a single numeral, and not for two or more combined numerals.

References

[1] Ansell M., Strach S.J. (1975). *The classification of handwriting numerals.* Proceedings of 7th Meeting of the IAFS, Zurich.

[2] Everitt B.S., Landau S., Leeseb M. (2001). *Cluster analysis.* 4th edition. Oxford University Press, New York.

[3] Huber R.A., Headrick A.M. (1999). *Handwriting identification: facts and fundamentals.* CRC Press, 152 – 164.

[4] Kaufman L., Rousseeuw P.J. (1990). *Finding groups in data: an introduction to cluster analysis.* Wiley, New York.

[5] Srihari S.N., Cha S.H., Arora H., Lee S. (2002). *Individuality of handwriting.* J. Forensic Sci. **47**, 1 – 17.

[6] Strach S.J. (1998). *Proposed research areas on handwriting comparison.* International J Forensic Document Examiners **4**, 312.

[7] SPSS Base 7.5 *Applications guide.* Chicago: SPSS Inc., c1997.

[8] Späth H. (1984). *Cluster analysis algorithms.* Ellis Horwood, Chichester.

Acknowledgement: The authors would like to thank a referee for helpful comments that improved the presentation of the paper, and to D.G. Clarke, Government Chemist, and S.C. Leung, Assistant Government Chemist of Hong Kong for their support and permission to use the data.

Address: W.K. Fung, C.T. Yang, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong; C.K. Li and N.L. Poon, Questioned Documents Section, Hong Kong Government Laboratory.

E-mail: wingfung@hku.hk

METHODS AND ALGORITHMS FOR ROBUST FILTERING

Ursula Gather and Roland Fried

Key words: Signal extraction, drift, edge, outlier, update algorithm.

COMPSTAT 2004 section: Robustness.

Abstract: We discuss filtering procedures for robust extraction of a signal from noisy time series. Moving averages and running medians are standard methods for this, but they have shortcomings when large spikes (outliers) respectively trends occur. Modified trimmed means and linear median hybrid filters combine advantages of both approaches, but they do not completely overcome the difficulties. Improvements can be achieved by using robust regression methods, which work even in real time because of increased computational power and faster algorithms. Extending recent work we present filters for robust online signal extraction and discuss their merits for preserving trends, abrupt shifts and extremes and for the removal of spikes.

1 Introduction

In speech recognition, video transmission and intensive care monitoring the basic task is to extract a signal from the observed noisy time series. The signal is assumed to vary smoothly most of the time with a few abrupt shifts. Besides the attenuation of normal observational noise and the removal of outlying spikes for recovering smooth sequences, the preservation of the locations and heights of shifts and local extremes is important. All this needs to be done automatically and in real time with short delays. This increases the risk of confusing outlier sequences and shifts or local extremes. For distinguishing extremes and outliers we rely on the smoothness of the underlying signal, i.e. observations which are far away from an estimated signal value are treated as outliers and not as being due to a signal peak. We can identify shifts by their duration setting a lower limit for the length of a relevant shift.

Moving averages and other linear filters are popular for signal extraction as they recover trends and are very efficient in Gaussian samples, but they are highly vulnerable to outliers and they blur level shifts. Tukey [16] suggests running medians for removing outliers and preserving level shifts, but standard medians have deficiencies in trend periods [8]. Linear median hybrid filters [10], [11] have been suggested as they are computationally more efficient than running medians, and preserve shifts similarly good or even better than these. These filters track polynomial trends, but they can only remove single isolated outliers. Modified trimmed mean filters are another compromise between running means and running medians. They choose an adaptive amount of trimming, but like running medians they also deteriorate in trend periods.

A better solution for tracking trends is to replace the median, a robust location estimator, by the estimated intercept obtained by robust regression of the data in a moving window against time. Based on a comparison of functionals with high breakdown point Davies, Fried and Gather [8] recommend Siegel's [15] repeated median because of its robustness against outliers and its stability. Since larger outliers have stronger effects on the repeated median we can add automatic rules for online trimming of outliers and construct procedures which are almost as bias-robust as filters based on least median of squares regression [13], but considerably faster and more efficient for Gaussian samples [5]. The Q_α-method [3], [14] has very nice properties for scale estimation even when a level shift occurs [9].

Robust regression also allows to construct hybrid filters which have similar benefits as linear median hybrid filters, while being considerably more robust [6]. Procedures applying adaptive trimming which do not deteriorate in trend periods can also be derived [7].

In Section 2 we introduce the filtering procedures. In Section 3 we discuss computational and other aspects. In Section 4 we propose a robust rule for the adaptive choice of the window widths. In Section 5 we analyze real and simulated data for further comparison before we give some conclusions.

2 Methods for robust filtering

We assume a component model for the sequence (x_t) of observed data

$$x_t = \mu_t + u_t + v_t,\ t \in \mathbb{Z}. \tag{1}$$

The underlying signal μ_t is the level of the time series, which is assumed to vary smoothly with a few sudden changes, while u_t is additive noise from a symmetric distribution with mean zero and variance σ^2, and v_t is impulsive (spiky) noise from an outlier generating mechanism. For online signal extraction we move a time window of width $n = 2k + 1$ through the series and use $x_{t-k}, \ldots, x_{t+k}$ to approximate μ_t. This causes a time delay of k observations. Firstly we fix k to a given value for all filters.

2.1 Filters based on robust regression

A standard median filter (running median) approximates the signal μ_t by the median of the observations $\{x_{t-k}, \ldots, x_{t+k}\}$ within a moving time window,

$$SM(x_t) = \tilde{\mu}_t = med\{x_{t-k}, \ldots, x_{t+k}\},\ t \in \mathbb{Z},$$

where μ_t is regarded as the level of the series at time point t, which is assumed to be locally constant. For tracking trends, Davies et al. [4] suggest fitting a local linear trend $\mu_{t+i} = \mu_t + i\beta_t,\ i = -k, \ldots, k$, to $\{x_{t-k}, \ldots, x_{t+k}\}$ by robust regression and recommend Siegel's [15] repeated median (RM). When applied to the data (i, x_{t+i}), $i = -k, \ldots, k$, the RM reads

$$\begin{aligned} RM(x_t) = \tilde{\mu}_t^{RM} &= med\{x_{t-k} + k\tilde{\beta}_t, \ldots, x_{t+k} - k\tilde{\beta}_t\} \\ \tilde{\beta}_t^{RM} &= med_{i=-k,\ldots,k}\left\{med_{j\neq i}\frac{x_{t+i} - x_{t+j}}{i-j}\right\}. \end{aligned}$$

2.2 Filters based on trimming

Lee and Kassam [12] suggest modified trimmed mean (MTM) filtering as a compromise between running means and running medians. MTM filters regulate the amount of trimming depending on the data. Firstly the local median $\tilde{\mu}_t$ and the local median absolute deviation about the median (MAD) $\tilde{\sigma}_t$ are calculated, then all observations farther away from the median than a multiple $q_t = d\tilde{\sigma}_t$ of the MAD are trimmed. Finally, the average of the remaining observations is taken as filter output:

$$\begin{aligned} MTM(x_t) &= \frac{1}{n_t}\sum_{i=-k}^{k} x_{t+i} \cdot 1_{[\tilde{\mu}_t - q_t, \tilde{\mu}_t + q_t]}(x_{t+i}), \\ n_t &= \#\{x_{t+i} \in [\tilde{\mu}_t - q_t, \tilde{\mu}_t + q_t], i = -k, \ldots, k\}, \\ q_t &= d \cdot c_n \cdot med\{|x_{t-k} - \tilde{\mu}_t|, \ldots, |x_{t+k} - \tilde{\mu}_t|\}. \end{aligned}$$

Here, c_n is a correction factor, which is chosen to achieve unbiasedness for Gaussian noise. For a very large window width we get $c_n = 1.483$, while e.g. for $n = 21$ we have $c_n = 1.625$. For $d = 0$, $MTM(x_t)$ is a running median, while for $d = \infty$ we get a moving average.

MTM filters implicitly assume a location model as standard median filters do. A straightforward modification is to fit a local linear trend by the repeated median and trim those observations having large residuals in this regression setting. The local variability can be estimated by applying the MAD to the regression residuals [7]. The filter output can then be derived either by least squares regression or by the repeated median of the observations with moderately large residuals. We denote the resulting filters by TRM and MRM, respectively:

$$\begin{aligned} TRM(x_t) &= \overline{x}_{J_t} - \tilde{\beta}_t^{TRM}\overline{j}_{J_t} \\ \overline{x}_{J_t} &= \frac{1}{|J_t|}\sum_{j\in J_t} x_{t+j} \\ \overline{j}_{J_t} &= \frac{1}{|J_t|}\sum_{j\in J_t} j \end{aligned}$$

$$\begin{aligned}
\tilde{\beta}_t^{TRM} &= \frac{\sum\limits_{j\in J_t}(j-\overline{j}_{J_t})(x_{t+j}-\overline{x}_{J_t})}{\sum\limits_{j\in J_t}(j-\overline{j}_{J_t})^2} \\
J_t &= \{j=-k,\ldots,k : |x_{t+j}-\tilde{\mu}_t^{RM}-j\tilde{\beta}_t^{RM}| \le q_t\} \\
MRM(x_t) &= med\{x_{t+j}-j\tilde{\beta}_t^{MRM},\ j\in J_t\} \\
\tilde{\beta}_t^{MRM} &= med_{i\in J_t}\left\{med_{j\in J_t, j\neq i}\frac{x_{t+i}-x_{t+j}}{i-j}\right\},
\end{aligned}$$

with $(\tilde{\mu}_t^{RM}, \tilde{\beta}_t^{RM})$ being the repeated median level and slope estimate for the current time window $\{x_{t-k},\ldots,x_{t+k}\}$.

2.3 Hybrid filters

Linear median hybrid filters are combinations of linear and median filters [10], [11]. Linear subfilters are applied to the input data before taking the median of their outcomes as final filter output. This reduces computation time and increases the flexibility compared to standard median filters due to the variety of linear subfilters. Linear median hybrid filters with finite impulse response, briefly FMH filters, are characterized by subfilters which respond to a finite number of impulses only.

A simple FMH filter corresponds to a location model and applies two one-sided moving averages and the current observation x_t as central subfilter for edge preservation:

$$SFMH(x_t) = med\{\Phi_1(x_t),\ x_t,\ \Phi_2(x_t)\}$$

$$\Phi_1(x_t) = \frac{1}{k}\sum_{i=1}^{k} x_{t-i}, \qquad \Phi_2(x_t) = \frac{1}{k}\sum_{i=1}^{k} x_{t+i}.$$

Predictive FMH filters correspond to a linear trend model and apply predictive FIR subfilters for one-sided extrapolation of a trend:

$$PFMH(x_t) = med\{\Phi_F(x_t), x_t, \Phi_B(x_t)\},$$

$$\Phi_F(x_t) = \sum_{i=1}^{k} h_i x_{t-i}, \qquad \Phi_B(x_t) = \sum_{i=1}^{k} h_i x_{t+i}.$$

Choosing the weights $h_i = \frac{4k-6i+2}{k(k-1)}$, $i=1,\ldots,k$ results in the minimal mean square error (MSE) predictions for a linear trend which is disturbed by white noise [11].

Combined FMH filters use predictions of different degrees,

$$CFMH(x_t) = med\{\Phi_F(x_t), \Phi_1(x_t), x_t, \Phi_2(x_t), \Phi_B(x_t)\},$$

with $\Phi_1(x_t)$, $\Phi_2(x_t)$, $\Phi_F(x_t)$ and $\Phi_B(x_t)$ being the subfilters for forward and backward extrapolation of a constant signal and a linear trend as given above.

In view of increased computational power and because of improved algorithms, computation time is nowadays not a great problem. Fried, Bernholt and Gather [6] use half-window medians and repeated medians to construct robust hybrid filters:

$$\begin{aligned} PRMH(x_t) &= med\{RM^F(x_t), x_t, RM^B(x_t)\} \\ CRMH(x_t) &= med\{RM^F(x_t), \tilde{\mu}_t^F, x_t, \tilde{\mu}_t^B, RM^B(x_t)\} \end{aligned}$$

Here, $\tilde{\mu}_t^F = med\{x_{t-k}, \ldots, x_{t-1}\}$ and $\tilde{\mu}_t^B = med\{x_{t+1}, \ldots, x_{t+k}\}$ are half-window medians, while $RM^F(x_t)$ and $RM^B(x_t)$ estimate the level at time t using the repeated median of $x_{t-k}, \ldots, x_{t-1}$ and $x_{t+1}, \ldots, x_{t+k}$, respectively:

$$\begin{aligned} RM^F(x_t) &= med\{x_{t-k} + k\tilde{\beta}_t^F, \ldots, x_{t-1} + \tilde{\beta}_t^F\} \\ \tilde{\beta}_t^F &= med_{i=-k,\ldots,-1}\left\{ med_{j=-k,\ldots,-1, j\neq i} \frac{x_{t+i} - x_{t+j}}{i-j} \right\} \\ RM^B(x_t) &= med\{x_{t+1} - \tilde{\beta}_t^B, \ldots, x_{t+k} - k\tilde{\beta}_t^B\} \\ \tilde{\beta}_t^B &= med_{i=1,\ldots,k}\left\{ med_{j=1,\ldots,k, j\neq i} \frac{x_{t+i} - x_{t+j}}{i-j} \right\}. \end{aligned}$$

3 Comparison of different filtering procedures

In the following we compare the previous filtering procedures w.r.t. computation time and their analytical properties.

3.1 Computation

The time needed for the filtering is crucial in real time applications. Fast algorithms for the update of the filter output are needed for online signal extraction. Denoting the length of the time window by n, the median of the proceeding window can be updated in logarithmic time ($O(\log n)$) using linear space if the data in the window are stored in sorted order using a red-black tree [2, Section 15.1]. This improves on the linear time needed for calculating the median from scratch.

An algorithm for the update of the repeated median in linear time using quadratic space based on a hammock graph is proposed by Bernholt and Fried [1], and another update algorithm needing only linear space running in $O(n \log n)$ average time is presented by Fried, Bernholt and Gather [6].

Updating the residuals and calculating the MAD can be done in linear time. Hence, the MTM and the TRM can both be calculated in linear time. For the MRM, however, $O(n^2)$ time is needed at least for the second repeated median. Detailed descriptions of the update algorithms can be found in Fried, Bernholt and Gather [6], [7].

The Table given below summarizes the time and the space needed for the updates of the filtering procedures. Note that the space for the repeated median and both repeated median hybrid filters can be reduced to $O(n)$, but at the expense of larger computation times.

	SM	RM	MTM	MRM	TRM	FMH	RMH
time	$O(\log n)$	$O(n)$	$O(n)$	$O(n^2)$	$O(n)$	$O(1)$	$O(n)$
space	$O(n)$	$O(n^2)$	$O(n)$	$O(n)$	$O(n^2)$	$O(n)$	$O(n)$

Table 1: Time and space needed for the update of the filters.

3.2 Analytical properties

For a discussion of the filtering procedures we concentrate on the analytical properties within a single time window when being applied to data generated from the component model (1).

Equivariance and invariance are important properties of statistical procedures. Location equivariance means that adding a constant to all observations in a window changes the filter output accordingly. Scale equivariance means that multiplication of all observations with a constant changes the estimate in the same way. All the above procedures possess these two properties.

Only some of the procedures are trend invariant, however [6], [7]. This property means that the extracted level does not change when adding a linear trend as long as the central level is fixed. The RM, the PFMH, the PRMH, the TRM and the MRM are trend invariant, while the median, the MTM, the CFMH and the CRMH are not. Therefore, for the latter methods the efficiency, the removal of spikes and the preservation of shifts are influenced by underlying trends.

Filters which are not trend invariant blur e.g. upward shifts within downward trends. Although the median and the MTM can remove k spikes completely in a single time window from a constant signal if there is no observational noise ($\sigma^2 = 0$), even a single positive outlier within a downward trend causes smearing. The predictive FMH can remove a single spike and preserve a shift exactly within a linear trend irrespectively of the directions as it is trend invariant, while the combined FMH does so only if the outlier (shift) has the same direction as the trend. The RMH filters improve on the FMH filters as they can remove up to $\lfloor k/2 \rfloor$ subsequent spikes without any effect. Furthermore, the predictive RMH preserves shifts exactly, while for the combined RMH this is true only if the shift is in the same direction as the trend, just like for the combined FMH. The RM, the TRM and the MRM can even remove $k-1$ spikes completely within a single time window irrespectively of a linear trend if $\sigma^2 = 0$.

The previous results hold when there is no observational noise. Lipschitz continuity restricts the influence of minor changes in the data due to small noise or rounding. The standard median, the FMH, the RM and the RMH filters are Lipschitz-continuous. The median is Lipschitz-continuous with constant 1 like all order statistics, while the repeated median and the repeated median hybrid filters are Lipschitz-continuous with constant $2k+1$. An FMH filter is Lipschitz-continuous with constant $\max|h_i^j|$, the maximal absolute weight given by a subfilter. MTM, MRM and TRM filters, however, are not Lipschitz-continuous, which can cause instabilities when there are small changes in the data. The discontinuity is caused by the trimming of observations. Application of continuous M-estimators is preferable for this reason, but computationally more expensive. Nevertheless, we investigate simpler trimming based methods in order to obtain information about the possible gain by further iterations.

The finite-sample breakdown point (FSBP) is the fraction of observations which have to be put into worst case positions in order to make the estimate take arbitrarily wrong values. For the median the breakdown point becomes $(k+1)/n$ when applied to $n = 2k+1$ data points, meaning that at least half of the window needs to be outlying in order to cause an arbitrarily large spike in the extracted signal. Since for the explosion of the local MAD also at least $k+1$ observations need to be modified, the MTM has the same breakdown point, while for the FMH filters two outliers are sufficient to make it break down. From the following Table we see that the RMH filters are considerably more robust than the FMH filters, and that the RM, TRM and MRM are almost as robust as the median in the sense of breakdown.

SM	MTM	RM	TRM	MRM	FMH	PRMH	CRMH
$\frac{k+1}{n}$	$\frac{k+1}{n}$	$\frac{k}{n}$	$\frac{k}{n}$	$\frac{k}{n}$	$\frac{2}{n}$	$\frac{\lfloor k/2\rfloor+1}{n}$	$\frac{\lfloor k/2\rfloor+2}{n}$

Table 2: Fraction of outliers in a window causing breakdown.

Simulations show the effect of the second step in the derivation of the TRM and the MRM on their MSE as compared to that of the RM filter. Application of least squares to the trimmed observations (TRM) increases the efficiency for Gaussian noise, but almost preserves the robustness of the repeated median, while application of the repeated median (MRM) further reduces the bias caused by outliers [7].

4 Adaptive choice of the window width

From the previous discussion we see that only the repeated median and the predictive hybrid filters PFMH and PRMH are both trend invariant and continuous, i.e. stable w.r.t. the occurrence of both trends and small changes in the data. The hybrid filters tend to preserve shifts and extremes, while the repeated median smoothes them considerably when being applied with a large window width [6], [7]. This means that on the one hand we should choose a short window width, but on the other hand a large window width is better for removing outlier patches and for the attenuation of the observational noise. This is a robust variant of the common problem of bandwidth selection in nonparametric smoothing.

Fried [5] investigates rules for online shift detection based on the most recent residuals in the time window. Similarly, we can formulate rules for the automatic choice of the window width using the regression residuals. Often least squares criteria are used to assess the local model fit and to find the bandwidth, but this is not suitable when outliers are present. Instead, a robust criterion is needed. Remembering that the median is the value which balances the signs of the residuals and that the repeated median is a regression analogue, it is natural to use the sign of the residuals. In this way we give the same weight to all observations irrespectively of their magnitude. However, note that there are always as many positive as negative residuals in the window for the repeated median fit. Therefore, we have to apply this idea to a suitable subset.

The Figure below visualizes the smoothing of a maximum by fitting a line with a too large window width. The residuals in the center will typically be positive, while most of the residuals at the start and the end of the window will be negative. These signs are simply al reverse for a minimum. Therefore it is natural to use the total number of positive residuals at the start and the end of the window for assessing the model fit. We divide the window into three sections as follows, namely the first $\lfloor (k+1)/2 \rfloor$ observations, the central $n - 2\lfloor (k+1)/2 \rfloor$ observations and the last $\lfloor (k+1)/2 \rfloor$ observations. If the total number T of positive residuals in the first and the last section is much larger than the average $\lfloor (k+1)/2 \rfloor$, we should shorten the window width since the signal slope might be decreasing substantially within the window. If T is much smaller than $\lfloor (k+1)/2 \rfloor$, the window width should also be shortened since the signal slope might be increasing.

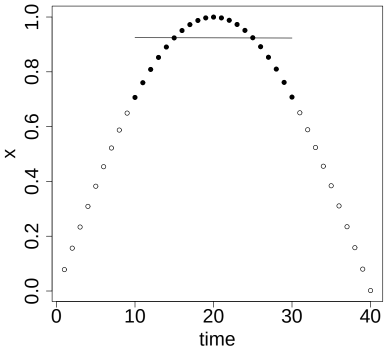


Figure 1: Smoothing of a maximum by fitting a line to the filled points.

However, this reduction should not result in a window width which is too small to resist outlying patterns. Results of previous studies [6], [7] show that the repeated median resists up to between 25% and 30% outliers without being substantially affected. Therefore, the minimal window width should be about four times the maximal length of outlier patches to be removed. For patches of length three e.g. we use the constraint $n \geq 11$. Since longer time windows allow better attenuation of observational noise and also robustness against many outliers we increase the window width after each step whenever possible.

The proposed repeated median algorithm with robust adaptive selection of the window width is as follows: Let $k_l < k_u$ be lower and upper bounds for k, and $0 \leq d_l < 1 < d_u \leq 2$ be constants. Set $k = k_l$ and $t = k + 1$.

1. Calculate the repeated median fit $(\tilde{\mu}_t, \tilde{\beta}_t)$ for $x_{t-k}, \ldots, x_{t+k}$ to obtain $RM(x_t) = \tilde{\mu}_t$.
2. Get the residuals $r_i = x_{t+i} - \tilde{\mu}_t - i\tilde{\beta}_t$, $i = -k \ldots, k$, and set $T = \#\{i = -k, \ldots, -k - 1 + \lfloor (k+1)/2 \rfloor, k + 1 - \lfloor (k+1)/2 \rfloor, \ldots, k : r_i > 0\}$.
3. If $k > k_l$ and $T < d_l \cdot \lfloor (k+1)/2 \rfloor$ or $T > d_u \cdot \lfloor (k+1)/2 \rfloor$ set $k = k - 1$ and go to 1.
4. If $k < k_u$ set $k = k + 1$.
5. Set $t = t + 1$ and go to 1.

The same or similar approaches can be used for the other robust filters. We just need to modify the window sections for the hybrid filters possibly obtaining asymmetric filters.

5 Application

We now apply the filtering procedures to two data sets. The first example is a time series simulated from an underlying sawtooth signal, which is overlaid by Gaussian white noise with zero mean and unit variance, and there are three isolated, three pairs and two triples of outliers of size -5. The Figure below shows the outputs of the CRMH and the adaptive RM filter with $k_l = 5$, $k_u = 15$, $d_l = 0.7$ and $d_u = 1.3$. The CRMH with $n = 21$ preserves the local extremes very well, but it is rather variable. The adaptive RM is almost as good at the extremes while being much smoother. Most of the time a width close to the maximal $n = 31$ is chosen, but close to the three local extremes and at about t=280 the width decreases even to the minimal $n = 11$. The PRMH not shown here is similar to the CRMH, but it is more affected by the outliers, while the ordinary RM and the median cut the extremes.

As a second example we analyze five hours of measurement of the arterial blood pressure of an intensive care patient. Figure 3 visualizes these data along with the outcomes of the MRM with a window width of $n = 21$ and of the adaptive RM filter with the same constants as before. The MRM resists some aberrant patterns very well, but it oversmoothes the local extremes at $t = 70$ and at $t = 290$. The adaptive RM again chooses the largest width $n = 31$ most of the time, but the width drops down to $n = 17$ about t=175 and t=225, and even to the minimal $n = 11$ about t=60 and t=130.

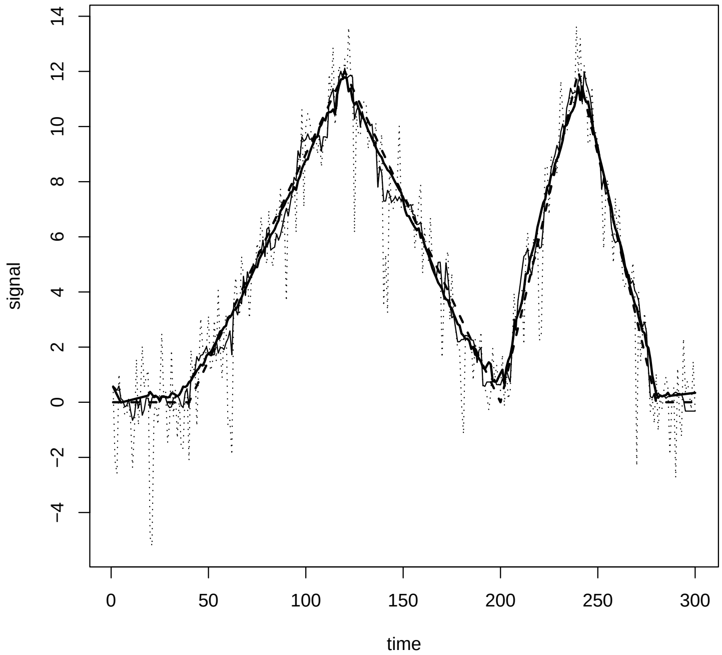


Figure 2: Simulated time series (dotted), underlying signal (dashed) and outputs of the CRMH (thin solid) and the RM with adaptive window width (bold solid).

It performs better at the extremes than the MRM, but it is affected by two subsequent outlying patterns about t=180. The RM with fixed window width also shows a spike there and performs in between the adaptive RM and the MRM at the extremes.

6 Conclusion

Improved numerical algorithms render the real time application of robust procedures for time series filtering possible. Methods for robust regression like the repeated median allow to construct filters which have similar benefits like classical linear or location based approaches when these perform well, but overcome deficiencies w.r.t. the removal of spiky noise (outliers) or the tracking of trends. We find the repeated median procedure with robust adaptive choice of the window width particularly promising. First applications show that this algorithm can be modified even for online filtering without any time delay by estimating the intercept at the right hand side of the time window, but more experience is needed to optimize the automatic choice of the window width then.

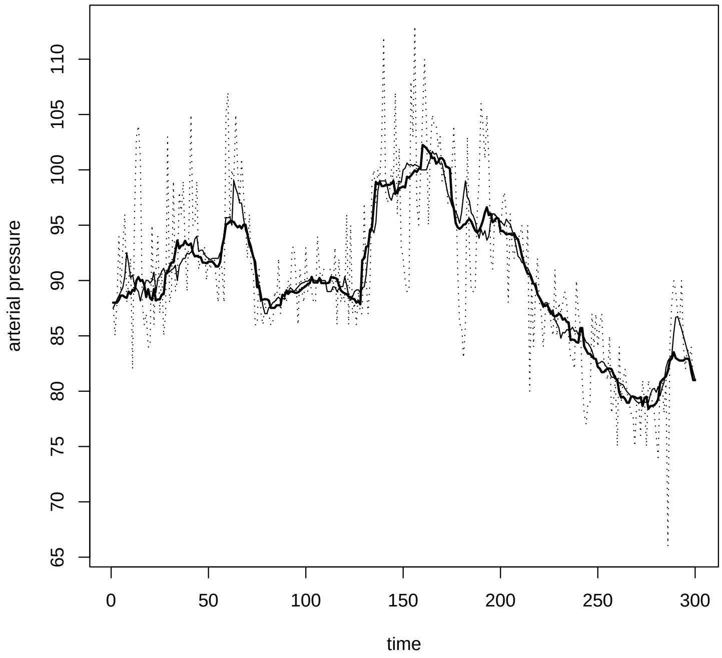


Figure 3: Arterial blood pressure (dotted) and outputs of the MRM (bold solid) and the RM with adaptive window width (thin solid).

References

[1] Bernholt T., Fried R. (2003). *Computing the update of the repeated median regression line in linear time.* Information Processing Letters **88**, 111–117.

[2] Cormen T.H., Leiserson C.E., Rivest R.L. (1990). *Introduction to algorithms.* MIT Press, Cambridge, Massachusetts, and McGraw-Hill Book Company, New York.

[3] Croux C., Rousseeuw P.J. (1992). *Time-efficient algorithms for two highly robust estimators of scale.* COMPSTAT 1992, Physica-Verlag, Heidelberg, 411–428.

[4] Davies P.L., Fried R., Gather U. (2004). *Robust signal extraction for online monitoring data.* J. Statistical Planning and Inference **122**, 65–78.

[5] Fried R. (2004). *Robust filtering of time series with trends.* Current Advances and Trends in Nonparametric Statistics, special issue of the J. of Nonparametric Statistics, to appear.

[6] Fried R., Bernholt T., Gather U. (2004a). *Repeated median and hybrid filters.* Technical Report, SFB 475, University of Dortmund, Germany.

[7] Fried R., Bernholt T., Gather U. (2004b). *Modified repeated median filters.* Preprint, Department of Statistics, University of Dortmund, Germany.

[8] Fried R., Gather U. (2002). *Fast and robust filtering of time series with trends.* COMPSTAT 2002, Physica-Verlag, Heidelberg, 367 – 372.

[9] Gather U., Fried R. (2003). *Robust estimation of scale for local linear temporal trends.* Proceedings of PROBASTAT 2002, Tatra Mountains Mathematical Publications **26**, 87 – 101.

[10] Heinonen P., Neuvo Y. (1987). *FIR-median hybrid filters.* IEEE Transactions on Acoustics, Speech, and Signal Processing **35**, 832 – 838.

[11] Heinonen P., Neuvo Y. (1988). *FIR-median hybrid filters with predictive FIR substructures.* IEEE Transactions on Acoustics, Speech, and Signal Processing **36**, 892 – 899.

[12] Lee Y., Kassam S. (1985). *Generalized median filtering and related nonlinear filtering techniques.* IEEE Transactions on Acoustics, Speech, and Signal Processing **33**, 672 – 683.

[13] Rousseeuw P.J. (1984). *Least median of squares regression.* J. American Statistical Association **79**, 871 – 880.

[14] Rousseeuw P.J., Croux C. (1993). *Alternatives to the median absolute deviation.* J. American Statistical Association **88**, 1273 – 1283.

[15] Siegel A.F. (1982). *Robust regression using repeated medians.* Biometrika **68**, 242 – 244.

[16] Tukey J.W. (1977). *Exploratory data analysis.* Addison-Wesley, Reading, Mass. (preliminary ed. 1971).

Acknowledgement: The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, “Reduction of complexity in multivariate data structures”) is gratefully acknowledged.

Address: U. Gather, R. Fried, Department of Statistics, University of Dortmund, 44221 Dortmund, Germany

E-mail: gather@statistik.uni-dortmund.de

USING GO FOR STATISTICAL ANALYSES

Robert Gentleman

Key words: Ontology, bioinformatics, graphs.

COMPSTAT 2004 section: Biostatistics.

Abstract: In this paper we use meta-data packages from the Bioconductor Project to carry out statistical analyses of gene expression data. But would like to note that the potential scope of these applications is much broader and many of the methods described here could be applied to other types of high-throughput data. To provide context we make use of data from an investigation into acute lymphoblastic leukemia.

1 Introduction

While there are a number of different definitions of an ontology we will use the notion of a restricted vocabulary as the basis for the discussions here. Ontologies and related concepts are becoming increasingly important tools for organizing and navigating information. Initiatives in biology (our main focus) as well as the semantic web are providing a variety of resources and interesting problems related to ontologies.

For genes and gene products the Gene Ontology Consortium, or GO, (`www.geneontology.org`) is an initiative that is designed to address this problem. GO provides a restricted vocabulary as well as clear indications of the relationships between terms. GO is clearly a valuable tool for data analysis, however its structure (as a DAG) and the complex nature of the relationships that it represents make appropriate use of this tool challenging.

1.1 The graph structure of GO

The GO ontologies are structured as directed acyclic graphs (DAGs) that represent a network in which each term may be a *child* of one or more *parents*. We use the expressions *GO node* and *GO term* interchangeably. Child terms are more specific than their parents. The relationship between a child and a parent can be and be either a *is a* relation or a *has a* (*part of*) relation.

Each term in the ontology is associated with a unique identifier and the relationships between the GO terms (parent/child) as well as other relevant data are provided by *GO*. The *GO* package provides six sets of mappings, two for each ontology.

In general, given a set of most specific terms of interest we can find the graph that consists of those terms and any less specific terms (parents). We will refer to this graph as the *induced GO graph* for the specific set of child nodes.

GO itself is strictly the ontology. The mapping of genes to GO terms is carried out separately. The actual mappings are provided by GOA [1] and are mappings between GO terms and LocusLink IDs which are modified to account for the multiplicity of mappings between the manufacturer IDs and LocusLink IDs.

2 An example

To demonstrate some of the tools that are included in the *GOstats* package we consider expression data from 79 samples from patients with acute lymphoblastic leukemia (ALL) that were investigated using Affymetrix GeneChip arrays [2]. The data were normalized using quantile normalization and expression estimates were computed using RMA [4]. Of particular interest is the comparison of 37 samples from patients with the BCR/ABL fusion gene resulting from a chromosomal translocation (9;22) with the 42 samples from the NEG group.

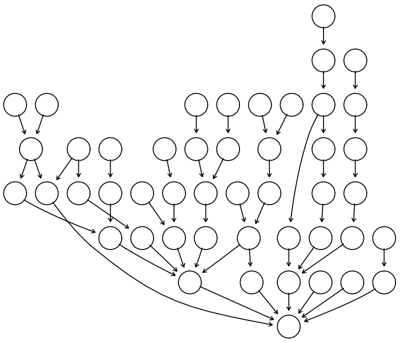


Figure 1: The induced GO graph for the selected genes.

To reduce the set of genes for consideration we applied two different sets of filters (gene filtering is considered in more detail in [5] and the interested reader is referred there). A non-specific filter was used to remove genes that showed little or no change in expression level across experiments. The resulting data set had 2391 probes remaining. To select genes whose expression values were associated with the phenotypes of interest (BCR/ABL and NEG) we used the `mt.maxT` function from the *multtest* package which computes a permutation based t-test for comparing two groups.

After adjustment for multiple testing there were only 19 probes (which correspond to 16 genes) with an adjusted p-value below 0.05. Using those genes we obtain the set of most-specific GO terms in the MF ontology that they are annotated at and compute the induced GO graph which is rendered in Figure 1. No labels have been added to the nodes in this plot since there is not sufficient room to provide informative ones. Notice that the most specific terms are at the top of the graph and that arrows go from more specific nodes to less specific ones. The node in the bottom center is the MF node. Clearly some sort of interactivity (e.g. tooltips) would be beneficial. We will return to this plot in the next section and use it to provide a more detailed view of the data.

3 Statistical analyses

3.1 Finding interesting GO terms

If genes have been partitioned into distinct sets, say by finding those with small p-values (as was done above) or by some form of clustering, then one of the questions that arises is whether genes that comprise a cluster have a common function, process or location in the cell. A second, related application of this idea is to provide meaning to a list or set of genes that were selected according to some criteria. For example, in our microarray experiment we selected genes that were differentially expressed between the BCR/ABL group and the NEG group. We might then wonder whether these genes have a common function, are involved in common processes, or perhaps are co-located in some region of the cell.

We can ask if there are more interesting genes at the node than one might expect by chance. If that is true, then that term can be thought of as being overrepresented in the data. This question can be answered using a Hypergeometric distribution. Suppose that there are N total genes annotated for the ontology of interest and that our list of interesting genes contains m distinct genes. Then we can imagine an urn with N balls in it and $N-m$ are black while m are white. If we draw k balls from the urn, where k is the number of genes annotated at a node, we are asking whether the number of white balls in that drawn sample is unusually large. Suppose that there are q white balls (interesting genes) in the drawn sample, we then ask what is the probability that $X \geq q$ where X is a Hypergeometric random variable with parameters as we have described. This probability constitutes a p-value since it is the probability of seeing something as extreme or more extreme than what was observed. This functionality is provided in the function `GOHyperG` available in the *GOstats* package.

In Figure 2, we reproduce the plot from Figure 1 except that we have now colored the nodes according to the p-value obtained from the Hypergeometric test described above. The nodes in Figure 2 are colored either red or blue depending on whether the unadjusted Hypergeometric p-value was less than 0.10 or not (for those viewing this document in black and white the nodes should be dark and light grey, respectively). The GO terms for the terms colored red are printed below. The relevant biology suggests that these are quite reasonable. We note that while the smallest p-values are associated with nodes that have few genes annotated at them there are some nodes with a reasonable number of genes annotated at them (counts) and small p-values.

	GO ID	Term	p-value	No. of Genes
1	0005148	prolactin recepto...	0.003	2
2	0005131	growth hormone re...	0.003	2
3	0005159	insulin−like grow...	0.008	5
4	0004715	non−membrane span...	0.017	11
5	0030693	caspase activity	0.019	12
6	0005126	hematopoietin/int...	0.057	37
7	0004197	cysteine−type end...	0.085	56
8	0005515	protein binding	0.095	1165
9	0008234	cysteine−type pep...	0.096	63
10	0005200	structural consti...	0.097	64
11	0003714	transcription cor...	0.097	64
12	0005198	structural molecu...	0.099	343

Table 1: GO terms, p-values and counts.

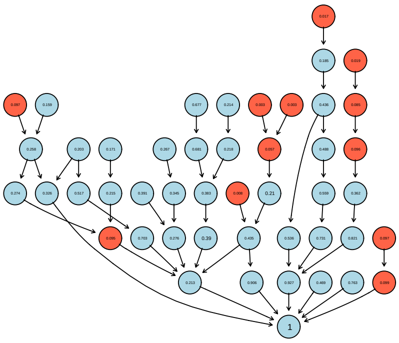


Figure 2: The induced GO graph colored according to unadjusted Hypergeometric p-values.

3.2 Selecting genes according to GO term

GO can also be used as a method of data reduction. Here one might carry out an analysis focusing on a particular subset of genes, say those associated with the GO term `transcription factor`.

Many of the effects due the BCR/ABL translocation are mediated by tyrosine kinase activity. It will therefore be of interest to examine genes that are known to have tyrosine kinase activity. We examine the set of GO terms and identify the term, `GO:0004713` from the *molecular function* portion of the GO hierarchy as referring to `protein-tyrosine kinase activity`. We see that for the Affymetrix HGU95av2 chip 230 probe sets are annotated at this particular term. Of these only 32 were selected by the non-specific filtering step. We focus our attention on these probes and carry out a permutation t-test analysis.

In this analysis of the GO–filtered data, 4 probe sets have FWER–adjusted p–values less than 0.1. They are printed below, together with the adjusted p-values from an analysis that used all probes that passed our non-specific filter and hence involved 2391 genes.

```
GO analysis

40480_s_at  2039_s_at   36643_at  2057_g_at
   0.00002    0.00025    0.02146    0.07481

All Genes"

40480_s_at  2039_s_at   36643_at  2057_g_at
     0.001      0.018      0.473      0.823
```

Due to the reduced number of tests in the analysis focused on tyrosine kinases, we are left with more significant genes after correcting for multiple testing. For instance, the probe set `36643_at`, which corresponds to the gene DDR1, was not significant in the unfocused analysis, but would be if instead the investigation was oriented towards studying tyrosine kinases *a priori*.

3.3 Using shortest paths

[6] consider some interesting applications of GO in conjunction with microarray expression data. In this section we consider a related idea and apply it to the ALL data. At their most basic level the ideas of [6] consist of forming a graph between genes (which are the nodes) based on some relevant distance. This distance might be correlation distance or it could be any other relevant distance. Then all edges in the graph that correspond to distances that are larger than some threshold are removed. Next, genes are grouped according to some specific categorization (they used GO biological process terms) and the shortest paths (using Dijkstra's algorithm) between all pairs of nodes are computed. Those shortest paths can then be examined to see whether they provide information of relevance.

In the ALL experiment we are most interested in comparing patients that have the BCR/ABL defect to those that have no measured cytogenetic abnormalities. Our adaptation of the shortest path technology is as follows. We use the output of the first filtering step described previously – that is we select genes that show some level of expression and some variation in expression across samples. We then separate the data into two sets (BCR/ABL and NEG) and within each group we define the distance between two genes as one minus the Pearson correlation (other approaches such as that use by [6] or some other robust correlation estimates). We then used an edge weight of

$d(u,v) = (1 - C_{u,v})^k$ with $k = 1$ and $\tau = 0.6$ as the cutoff for correlations, if $C_{u,v} < \tau$ then no edge exists (Zhou et. al used $k = 6$ in their analysis and some experimentation may be warranted).

Our interest in this particular example is on transcription factors. Hence we use the GO term `GO:0003700` which maps to the molecular function `transcription factor activity` to identify all genes with transcription factor activity. We used only genes for which this was a most specific annotation and obtained 814 mappings and 531 unique LocusLink ids. Of these 152 were among those probes selected for our analysis. Of these we found that there were 6 with duplicate entries. A visual inspection (not reported) suggested that the correlation between these duplicate probes was quite high and so only one of each was used in the subsequent analysis. This left us with 146 distinct transcription factors for our study.

For every pair of transcription factors we compute two quantities. The shortest path between each pair for each of the different conditions. For example in our ALL example we compute the shortest paths between all transcription factors using a graph based only on data from those with BCR/ABL and secondly the same set of values based only on data from those without any noticeable genomic defects. Then for each pair the distances are compared (plotted) and those pairs for which the distance has changed the most identified and further explored.

We first consider those transcription factors that are not connected to the others in their respective graphs. There are three sets, those that are not connected in either graph, those that are not connected in one of the two graphs but not in the other. They are reported in Table 2.

	Affymetrix ID	Symbol	Which graph
1	`34730_g_at`	TRO	Both
2	`1106_s_at`	TRA@	NEG only
3	`34850_at`	UBE2E3	NEG only
4	`1185_at`	IL3RA	BCR/ABL only
5	`32186_at`	SLC7A5	BCR/ABL only
6	`33641_g_at`	AIF1	BCR/ABL only

Table 2: Genes not connected in the different graphs.

We now consider the finite pairwise distances. First a simple t-test can be carried out to see if there is any difference between the distances in one graph versus the other. We took each pairwise distance in the NEG graph and subtracted from it the same pairwise distance computed on the BCR/ABL graph. The t-test is for whether the mean is zero and the test statistic was 0.179 with an extremely small p-value. So we see that distances in the NEG graph seem to be longer than those in the BCR/ABL. Further evidence of this difference comes from the observation that proportion of values that were larger in the NEG graph was 0.589.

We will focus our attention on those differences that are large in absolute value. We chose a value of 2.5 as our cut-off and found that there were 66 differences that were larger than 2.5. These corresponded to 26 distinct genes.

While all may be interesting and a particular investigator may want to expend considerable effort in study transcription factors that are of particular interest we will center our analysis on the set of genes that appear most frequently in this list.

There are three genes that have high counts, namely, MYC, MPO and GADD45A. This fact suggests that perhaps the expression patterns of these three different transcription factors are substantially different in the two phenotypes we are studying.

For each of the three transcription factors we can compute the average distance, separately within each graph, to all the other selected genes. We find that the results are quite consistent and that in all cases the path length is much shorter in the BCR/ABL group than it is in the NEG group. For MYC the means were 5 for NEG and 2 for BCR/ABL, and for MPO they were 4 for NEG and 2 for BCR/ABL and for GADD45A the means were 5 for NEG and 2 for BCR/ABL. It is rather interesting to observe that amongst the pairwise distances that have changed the most are those between these three specific genes.

Specific paths between transcription factors can also be examined. Recall that we compute out distance between two transcription factors based on the shortest path length between them in each of the two graphs. In our examples we focus on MYC and the distances between it and MPO and GADD45A.

We print out the different shortest paths for genes connecting MYC to both MPO and GADD45A for each of the two phenotypes, respectively (first the paths for BCR/ABL, then for the NEG samples). The MYC to MPO results are:

```
BCR/ABLE
MYC->EIF4G1->HMG20B->MPO
NEG
MYC->CDC25B->TRAP1->FLJ10326->LANCL1->EMP3->S100A4
          ->LGALS1->MPO
```

If we then make use of the results in Figure 3 we see that there are positive correlations between MYC and EIF4G1 and as well between EIF4G1 and HMG20B, but that for HMG20B and MPO the correlation is negative. Positive correlations are suggestive of shared transcriptional activity while negative correlations are suggestive of transcriptional inhibition.

The results comparing MYC to GADD45A are:

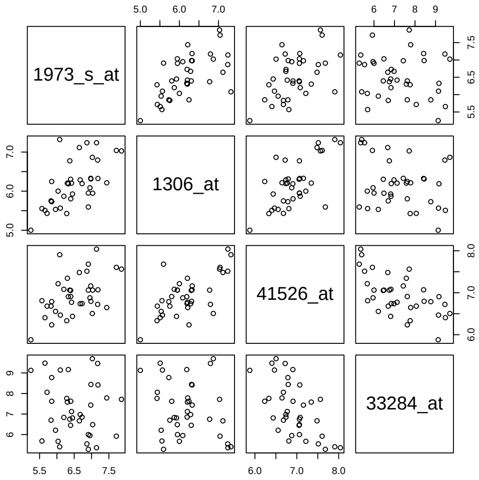


Figure 3: Pairwise scatterplots of gene expression for those genes on the shortest path between MYC and MPO from patients with the BCR/ABL translocation.

```
BCR/ABLE
MYC->UBE2A->BAZ1A->CD53->GADD45A
NEG
MYC->CDC25B->TRAP1->SSBP1->SMC1L1->TK1->HCK->
           SH3PB1->PVRL2->GADD45A
```

We do not have space to present the other pairwise scatterplots here but readers that are making use of the compendium version of this paper can easily explore those different plots on their own.

We notice that the path lengths for the NEG samples are longer (involve more genes) than those for the BCR/ABL samples. We might also want to ask whether the distances are also larger (that is that the correlations are smaller). To do this we need to obtain the edge weights from the respective graphs and compare them. We found that there appeared to be no difference (all averaged around a distance of about 0.65) but the number of edges is quite small and one might expect to see systematic differences if a larger study were undertaken.

We can check our results, at least to some extent, by examining pairwise scatterplots of the gene expressions. In Figure 3 the genes on the path from MYC to MPO are plotted. We see quite strong correlations along the diagonal and note that HMG20B and MPO have a negative correlation.

Finally, we finish our examination of these data by considering some of the specific paths between the different transcription factors. We see, in

Figures 4 the actual shortest path between the genes MYC and MPO. The two end points have been colored red, genes along the path are colored blue.

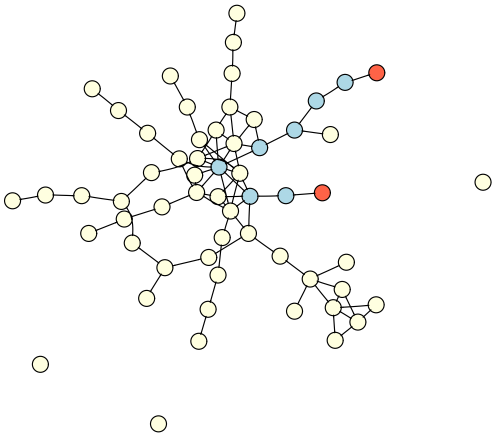


Figure 4: Shortest path between MYC and MPO in the NEG samples.

4 Discussion

GO and the mappings from genes to specific terms in each of the three ontologies provide a number of important and unique data analytic opportunities. In this paper we have considered three separate applications of these resources to the problem of analysing gene expression data and in all cases the GO related data have provided new and important insights into the data.

Using GO mappings to select certain terms for further study and reference has the possibility of providing meaning to sets of genes that have been selected according to different criteria. An equally important application is to use GOA mappings to reduce the set of genes under consideration. As the capacity of microarrays increases it is important that we begin developing tools and strategies that directly address specific questions of interest. P-value correction methods are at best a band-aid and do not represent an approach that has long term viability [5].

In our final example we adapted the method proposed by [6] to a different problem, one where we consider only transcription factors and where we are interested in understanding their interrelationships. The results are promising and in our example reflect a fundamental difference between those with the BCR/ABL translocation and those patients with no observed genetic abnormalities. Ideally these, and other observations will lead to better understanding of transcriptional regulation and from that to better understanding modalities of efficacy for drug treatments.

Perhaps more important than the statistical presentation is the fact that we have also provided software implementations for all tools described and discussed in this paper. They are available from the Bioconductor Project in the form of the *GOstats* package. *GOstats* makes substantial use of software infrastructure from the Bioconductor Project in carrying out this analysis. In particular the *graph*, *Rgraphviz* and *RBGL*, together with the different meta-data packages.

Finally, this document itself represents an approach to reproducible research in the sense discussed by [3] and it can be reproduced on any users machine equipped with R and the appropriate set of R packages. We encourage the interested reader to avail themselves of the opportunity to explore the data and the methods in more detail on their own computer.

References

[1] Camon E., Magrane M., Barrell D., Lee V., Dimmer E., Binns D., Maslen J., Harte N., Lopez R., Apweiler R. (2004). *The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology.* Nucleic Acids Research **32**, D262 – D266.

[2] Chiaretti S., Li X., Gentleman R., Vitale A., Vignetti M., Mandelli F., Ritz J., Foa R. (2004). *Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival.* Blood **103**, 2771 – 2778.

[3] Gentleman R., Temple Lang D. (2003). *Statistical analyses and reproducible research.*

[4] Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay, Y.D., Antonellis K.J., Scherf U., Speed T.P. (2003). *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics **4** 249 – 264.

[5] von Heydebreck A., Huber W., Gentleman R. (2004). *Differential expression with the bioconductor project.* In Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. John Wiley and Sons.

[6] Zhou X., Kao M.-C.J., Wong W.H. (2002). *Transitive functional annotation by shortest-path analysis of gene expression data.* PNAS **99**, 12783 – 12788.

Acknowledgement: I would like to thank Vincent Carey for many helpful discussions about these, and very many other topics. I would like to thank Drs. J. Ritz and S. Chiaretti of the DFCI for making their data available and for helping me to understand how it relates to ALL. I would like to thank J. Zhang and J. Gentry for a great deal of assistance in preparing the data and writing software in support of this research.

Address: R. Gentleman, Department of Biostatistics, Harvard University

E-mail: `rgentlem@jimmy.harvard.edu`

COMPSTAT'2004 Symposium

COMPUTATIONAL CHALLENGES IN DETERMINING AN OPTIMAL DESIGN FOR AN EXPERIMENT

Subir Ghosh

Key words: Balanced arrays, computational challenges, factorial designs, interactions, orthogonal arrays, robust designs, search designs, search linear models, search probabilities, unavailability of data.

COMPSTAT 2004 section: Design of experiments.

Abstract: In this paper we present some computationally challenging problems for finding an optimum design in an experiment. We consider the problem of finding an optimum design when one model from a set of possible models would describe the data better than other models in the set but we do not know this model a priori. We also consider the robustness of optimum designs under a model when some observations are unavailable.

1 Introduction

In the early development of designing a statistically efficient experiment, considerable attention was given to the computational simplicity of the analysis and to some desirable properties of the inferences drawn on the comparisons (parameters) of interest [2]. The concepts of orthogonality and balance in experimental designs were developed. With the progress in methodological research and the development in computing technology, the concepts of optimum designs and various optimality criteria were proposed [10]. The experiment could be performed at a single stage or at many stages over time. The data could be continuous, discrete, univariate, multivariate, time series, spatial, and other kinds or some combinations of them. Inference procedures could be parametric, nonparametric, semiparametric, frequentist, Bayesian, and others. The most amazing aspect in the design research is the enormous contributions of all kinds of researchers from extreme theorists to extreme practioners [8]. We do not attempt to make any futile effort to list all the contributors and their research. In this paper we examine some aspects of determining optimal designs and discuss some challenging problems.

2 Optimum designs

An optimum design is normally obtained by satisfying one or more optimality properties (minimizing variance, maximizing power and many others) for the comparisons (parameters) of interest under an assumed model. The choice between a best design with respect to (w.r.t.) one criterion and a best design w.r.t. another criterion is always an issue at the time of the selection of an optimum design. With the change in the computing environment, this issue has become much more complex. For example, the orthogonal fractional factorial plans may be best w.r.t. many optimality criteria but they require more runs in most situations than nonorthogonal plans and furthermore may not perform well compared to nonorthogonal plans when the assumed model is really inadequate. If we decide to give up orthogonality and opt for optimal balanced fractional factorial plans as our nonorthogonal plans, then we may cut down the cost of running the experiment as well as improve the performance when the assumed model is inadequate. Finding optimal balanced fractional factorial plans as nonorthogonal plans is always computationally challenging but it is possible to find such plans in the modern computing environment. Many such plans are already available in the design literature. The list of references is available in Ghosh and Rao [7], [8].

3 Robust designs

The unavailability of data that we often encounter in conducting an experiment should be a concern at the design stage. Ghosh [3] introduced the concept of robustness of design against the unavailability of any t (a positive integer) observations in the sense that the unbiased estimation of all the parameters of interest is still possible when any t observations are unavailable. For n observations, there are $\binom{n}{t}$ possible sets of t observations. Ghosh and Namini [5] gave several criteria and methods for determining the influential set of t observations for robust designs. There are numerous such practical issues including the presence of outliers, time trend in observations, and others in real life experiments. Such practical issues give rise to challenging computational problems in the selection of designs.

4 Model identification using search designs

The problem of finding a best design or a class of best designs satisfying one or more optimality criteria under an assumed model is a challenging task. Analytical methods are not often sufficient for resolving this task. Computational methods are very powerful in addition to the applicable analytical methods in resolving this problem. When we are not absolutely sure about the assumed model that will fit the experimental data adequately, the problem becomes daunting. In reality we are rarely sure about a particular model in terms of its effectiveness in describing the data adequately. However, we are normally sure about a set of possible models that would describe the data better than other models in the class. The pioneering work of Srivastava [13] introduced the search linear model with the purpose of searching for and identifying the best model from a set of possible models that includes the best model. We now focus on finding a best design or a class of best designs for model identification through the use of the search linear models. Computational methods are indispensable for this purpose.

In factorial experiments, the lower order effects are normally important and the higher order effects are all assumed to be negligible. In main effect plans, the main effects are important and the interaction effects are assumed to be zero. Such an assumption may or may not hold true in reality because of the possible presence of a few significant non-negligible interactions. The standard linear models cannot identify these non-negligible effects using a small number of runs or treatments considerably smaller than the total number of possible runs for an experiment. This motivates the use of search designs under the search linear model in searching for and identifying non-negligible interaction effects. We consider the problem of comparing search designs with the ability of searching for and identifying k (a positive integer) non-negligible interaction effects.

5 Search linear model

Consider the search linear model [13]

$$E(\boldsymbol{y}) = \boldsymbol{A_1}\boldsymbol{\xi_1} + \boldsymbol{A_2}\boldsymbol{\xi_2},\ V(\boldsymbol{y}) = \sigma^2\boldsymbol{I}\ , \tag{1}$$

where $\boldsymbol{y}(n \times 1)$ is the vector of observations, $\boldsymbol{A_1}(n \times \nu_1)$ and $\boldsymbol{A_2}(n \times \nu_2)$ are matrices known from the underlying design. The elements of the vector $\boldsymbol{\xi_1}(\nu_1 \times 1)$ are unknown parameters. About the elements of $\boldsymbol{\xi_2}(\nu_2 \times 1)$ we know that at most k elements are nonzero but we do not know which elements are nonzero. The k is small compared to ν_2. The goal is to search for and identify the nonzero elements of $\boldsymbol{\xi_2}$ and then estimate them along with the elements of $\boldsymbol{\xi_1}$. Such a model is called a search linear model. When $\boldsymbol{\xi_2} = \boldsymbol{0}$, the search linear model becomes the ordinary linear model. For the search linear model, we have $\boldsymbol{\xi_2} \neq \boldsymbol{0}$.

Let $\boldsymbol{A_{22}}$ be any $(n \times 2k)$ submatrix obtained by choosing $2k$ columns of $\boldsymbol{A_2}$. A design is called a search design [13] if, for every submatrix $\boldsymbol{A_{22}}$,

$$\text{Rank}[\boldsymbol{A_1}, \boldsymbol{A_{22}}] = \nu_1 + 2k. \tag{2}$$

The rank condition (2) allows us to fit and discriminate between any two models in the class of possible models described earlier. Any two models in the class have ν_1 common parameters which are the elements of $\boldsymbol{\xi_1}$ and at most $2k$ uncommon parameters which are the elements of $\boldsymbol{\xi_2}$. Note that $n \geq \nu_1 + 2k$. A search design allows us to search for and identify the nonzero elements of $\boldsymbol{\xi_2}$ and then estimate them along with the elements of $\boldsymbol{\xi_1}$.

6 Computationally challenging problems

Consider a class of $\binom{\nu_2}{k}$ linear models from (1) with the parameters as $\boldsymbol{\xi_1}$ and k elements of $\boldsymbol{\xi_2}$. The $\binom{\nu_2}{k}$ possible sets of k elements of $\boldsymbol{\xi_2}$ give rise to $\binom{\nu_2}{k}$ such models. For any two models in this class, the elements $\boldsymbol{\xi_1}$ are common parameters but in the two sets of k elements in $\boldsymbol{\xi_2}$ some common parameters may or may not be present. A search procedure identifies the model which best fits the data generated from the search design. To identify this model, the sum of squares of errors (SSE) of each model is used [13]. If SSE for the first model ($M1$) is smaller than the SSE for the second model ($M2$), then $M1$ provides a better fit and is selected over $M2$. For a fixed value of k, all $\binom{\nu_2}{k}$ models are fitted to the data and the search procedure selects the model with the smallest SSE as the best model for describing the data.

6.1 Optimal search designs

For each model in the class of $\binom{\nu_2}{k}$ linear models from (1), we consider the variance-covariance matrix of the least squares estimators of the parameters. We calculate the values of the Determinant(D), Trace(T), and Maximum Characteristic Root(MCR). So we obtain $\binom{\nu_2}{k}$ sets of values of D, T, and MCR. We calculate the arithmetic means and the geometric means of D, T, and MCR and denote them by AD, AT, AMCR, GD, GT, and GMCR. The smaller are the values of AD, AT, AMCR, GD, GT, and GMCR, the better is the search design. Note that the minimization of only D, T, and MCR represent the A- , D- , and E- optimality criteria [10]. The arithmetic mean is more meaningful than the geometric mean in some areas of application and vice versa. We use these six criteria for comparing search designs with the same number of runs. This is computationally a huge task.

For a factorial experiment with four factors each at two levels (+) and (−), suppose that $\boldsymbol{\xi_1}$ consists of the general mean and main effects and $\boldsymbol{\xi_2}$ consists of only two factor interactions. Consider two designs, d1 and d2 with 8 runs. Design d1 has the ability of searching for one nonnegligible two-factor interaction and furthermore, this plan is optimal w.r.t. the AD, GD, AT, and GT criteria. Design d2 has also the ability of searching for one nonnegligible two-factor interaction and furthermore, this plan is optimal w.r.t. the AMCR and GMCR criteria. These new plans are obtained by first finding all the search designs with 8 runs and 4 factors and then calculating their AD, AT, AMCR, GD, GT, and GMCR values. Finding of d1 and d2 is indeed a computer intensive task. Table 1 presents d1 and d2.

6.2 Search probabilities

The probability of selecting one model over another model depends on σ^2, the noise variance which we refer to as the noise. To see this dependence, we consider three cases $\sigma^2 = 0$, $\sigma^2 = \infty$, and $0 < \sigma^2 < \infty$. Let $M0$ be the true model in the class of models described above. Furthermore, let $M1$ be a competing model where $M1 \neq M0$. In the noiseless case, $\sigma^2 = 0$, the SSE for $M0$, SSE($M0$), is zero, which is always smaller than the SSE($M1$). Hence, $M0$ will definitely be selected over $M1$. Therefore, the correct nonzero interaction will always be identified with probability one. Thus, $P[SSE(M0) < SSE(M1)|M0, M1, \sigma^2 = 0] = 1$. In reality $\sigma^2 > 0$ and the SSE($M0$) may

d1				d2			
-	-	-	-	-	-	-	-
-	+	+	+	+	+	+	+
-	-	+	+	-	+	+	+
-	+	-	+	+	-	-	-
-	+	+	-	-	-	+	-
+	-	-	+	+	+	-	+
+	-	+	-	-	-	+	+
+	+	-	-	+	+	-	-

Table 1: d1 and d2 with 8 runs and 4 factors.

not be less than SSE($M1$). Therefore, $M0$ may not necessarily be selected over $M1$. Hence, the probability of correctly identifying the nonzero interaction is less than one and we write $P[SSE(M0) < SSE(M1)|M0, M1, \sigma^2 > 0] < 1$. In the case of infinite noise, $M0$ and $M1$ are equally likely to be selected and so the probability of selecting $M0$ over $M1$ is 1/2, and we write $P[SSE(M0 < SSE(M1)|M0, M1, \sigma^2 = \infty] = 1/2$. For $0 < \sigma^2 < \infty$, $P[SSE(M0) < SSE(M1)|M0, M1, \sigma^2]$ is called the *search probability* for a given $M0, M1$, and σ^2. Note that the search probability is between 1/2 and 1. Shirakura et al. [12] presented the search probability for searching one nonnegligible effect ($k = 1$) based on the normality assumption for observations under the search linear model (1).

There are many of these search probabilities to consider. We note that for a given true model $M0$, there are $(\nu_2 - 1)$ competing models of $M1$ for $k = 1$. Since the true model $M0$ is unknown, we consider all $\nu_2(\nu_2 - 1)$ possible pairs of $(M0, M1)$ and calculate all the search probabilities for a given σ^2. From these search probabilities, Ghosh and Teschmacher [9] presented a $\nu_2 \times \nu_2$ search probability matrix (SPM) where the columns correspond to the possible true models and the rows correspond to the possible competing models. The off-diagonal elements of the SPM represent the search probabilities corresponding to all possible pairs of $M0$ and $M1$ for a given σ^2. Since the true model $M0$ is different from the competing model $M1$, the diagonal elements of the SPM are not meaningful and therefore left blank.

When comparing two designs, we would like to determine which design has a greater chance of identifying the true nonzero interaction term. A method for doing this is by comparing the SPMs of the two designs for a given σ^2. The SPM for a design is dependent on a parameter, ρ, which is the ratio of the magnitude of the true unknown interaction term (signal) and σ (noise). In other words, the SPM depends on σ^2 through ρ. Let $SPM_i(\rho)$ be the SPM of the ith design for a given ρ, where the columns and rows correspond to the possible true and competing models, respectively.

Shirakura, et al. [12] proposed a criterion for comparing search designs for a specific value of ρ. This criterion is based on the minimum value of all the elements of the SPM. The higher is this minimum value, the better is the design. Ghosh and Teschmacher [9] defined the SPM, proposed two other criteria, and presented methods of comparing search designs for all values of ρ using all three criteria. One of the two proposed criteria in Ghosh and Teschmacher [9] is based on the element-by-element comparison of two $SPMs$ and the other one is based on comparing two minimum search probability vectors ($MSPVs$) whose elements are the minimum values of the columns of two $SPMs$. The comparisons are then made by using a majority rule in the sense of having the fifty percent or more elements of an SPM are greater the corresponding elements of another SPM. Similar comparisons are also made for two MSPVs. The methods proposed in Ghosh and Teschmacher [9] have opened up a new direction of computationally challenging problems for finding optimum designs.

Orthogonal designs have many well-known optimality properties under the ordinary linear model. However, balanced designs can perform better than orthogonal designs under the search linear model. Consider two search designs, $D1$ and $D2$, each with 12 runs, and 4 factors each at two levels $(-)$ and $(+)$. Design $D1$ is a balanced array of full strength and design $D2$ is an orthogonal array of strength 2 obtained from the 12-run Plackett-Burman design [11] by choosing the first four columns. Table 2 presents $D1$ and $D2$. Design $D1$ performs better than Design $D2$ under the ordinary linear model with $\boldsymbol{\xi_2} = \mathbf{0}$. However, $D2$ performs better than $D1$ under the search linear model when the vector $\boldsymbol{\xi_2}$ consists of two and three factor interactions only one of which is nonzero, so that $k = 1$. This is a really striking example illustrating the fact that an orthogonal design is not necessarily the best in all situations.

6.3 Robust designs

The optimal designs may no longer be optimal when some observations become unavailable during the experiment. Determining the robustness of optimal designs against the unavailability of data is a computationally difficult problem [3], [5]. Ghosh and Al-Sabah [6] presented some efficient composite plans for response surface experiments with surprisingly higher efficiency than existing comparable plans in the literature w.r.t. all three criteria, D, T, and MCR. For example, under the second order response surface model with ten factors, the MCR, T, and D x 10^{70} values are 7.1, 31.2, and .033 for Ghosh-Al-Sabah plan and 6791.4, 6850.0, and 1.6 for the existing Draper-Lin plan [1]. Ghosh-Al-Sabah plans were obtained while studying the robustness properties of some existing designs.

7 Conclusions

In this paper we have described some challenging computational problems in finding a best design for an experiment. Modern computing environment has

D1				D2			
+	+	+	+	+	-	+	-
-	-	-	-	+	+	-	+
-	-	-	+	-	+	+	-
-	-	+	-	+	-	+	+
-	+	-	-	+	+	-	+
+	-	-	-	+	+	+	-
-	-	+	+	-	+	+	+
-	+	-	+	-	-	+	+
+	-	-	+	-	-	-	+
-	+	+	-	+	-	-	-
+	-	+	-	-	+	-	-
+	+	-	-	-	-	-	-

Table 2: $D1$ and $D2$ with 12 runs and 4 factors.

helped us in attempting to resolve these problems. Many other challenging problems and some of their solutions are indeed available in the work of other researchers. Many new computationally challenging problems are also constantly emerging with the modern development in science and technology.

References

[1] Draper N.R., D.K.J. Lin (1990). *Small composite designs.* Technometrics **32** 187 − 194.

[2] Fisher R.A. (1935). *The design of experiments.* First Edition. Oliver and Boyd, London.

[3] Ghosh,S. (1979). *On robustness of designs against incomplete data.* Sankhya B **40**, 204 − 208.

[4] Ghosh S. (1980). *On main effect plus one plans for 2^m factorials.* Ann. Statist. **8**, 922 − 930.

[5] Ghosh S., Namini H. (1990). *Influential observations under robust designs.* IN: Coding Theory and Design Theory, Part II: Design Theory, D.K. Ray-Chaudhuri (ed.), Springer-Verlag, New York, 86 − 97.

[6] Ghosh S., Al-Sabah W.S. (1996). *Efficient composite designs with small number of runs.* J. Statist. Plann. Inference **53**, 117 − 132.

[7] Ghosh S., Rao C.R. (1996). *Design and analysis of experiments.* North-Holland, Elsevier Science B.V., Amsterdam.

[8] Ghosh S., Rao C.R. (2001). *An overview of developments in statistical designs and analysis of experiments.* In: Recent Advances in Experimental Designs and Related Topics, S. Altan and J. Singh, (eds.), Nova Science Publishers, Inc., New York, 1 − 24.

[9] Ghosh S., Teschmacher T. (2002). *Comparisons of search designs using search probabilities.* J. Statist. Plann. Inference **104**, 439 − 458.

[10] Kiefer J. (1959). *Optimum experimental designs.* J. Roy. Statist. Soc. B **21**, 272 – 319.

[11] Plackett R.L., Burman J.P. (1946). *The design of optimum multifactorial experiments.* Biometrika **33** 305 – 325.

[12] Shirakura T., Takahashi T., Srivastava J.N. (1996). *Searching probabilities for nonzero effects in search designs for the noisy case.* Ann. Statist. **24** 6, 2560 – 2568.

[13] Srivastava, J.N. (1975). *Designs for searching non-negligible effects.* In: A Survey of Statistical Design and Linear Models, J.N. Srivastava, (ed.), North-Holland, Elsevier Science B.V., Amsterdam, 505 –519.

Acknowledgement: The author would like to express his sincere gratitude to three reviewers for their critical reading of the earlier version of this paper.

Address: S. Ghosh, University of California, Riverside, CA 92521-0138,USA

E-mail: ghosh@ucrac1.ucr.edu

VISUALIZATION OF PARAMETRIC CARCINOGENESIS MODELS

Jutta Groos and Annette Kopp-Schneider

Key words: Hepatocarcinogenesis, color-shift model, maximum likelihood estimate.

COMPSTAT 2004 section: Biostatistics.

Abstract: This paper concentrate on the effective tools to compare different carcinogenesis models with respect to their ability to predict numbers and radii of foci in hepatocarcinogenesis experiments. Especially the CSM-GUI (Color-Shift graphical user interface) shows to be a powerful instrument to test a new model before starting the very time-intensive procedure of finding the maximum likelihood parameters.

1 Introduction

Hepatocarcinogenesis experiments identify focal lesions consisting of intermediate cells at different preneoplastic stages. Several hypotheses are established to describe the formation and progression of preneoplastic liver foci. A common model of hepatocarcinogenesis is the multi-stage model, which is based on the assumption that cells have to undergo multiple successive changes on their way from the normal to the malignant stage. In this model single cells change their phenotype through mutation into the next stage and proliferate according to a linear stochastic birth-death process [4] [5].

In contrast, the Color-Shift-Model (CSM) was introduced by Kopp-Schneider and colleagues [4] to describe that whole colonies of altered cells simultaneously alter their phenotype. In this model, preneoplastic foci are assumed to grow exponentially with deterministic rate and to change their phenotype ('color') after an exponentially distributed waiting time [1] [3].

To take into account that the assumption of deterministic growth rates for foci in the CSM seems to oversimplify the real process, a CSM with stochastic growth rates is introduced.

In order to compare different models with respect to their ability to predict numbers and radii of foci in a rat hepatocarcinogenesis experiment maximum likelihood estimates for the model parameters are used and the predicted and empirical distributions are vizualized.

2 Color-shift-model with stochastic growth rates in case of 2 colors

The assumption of deterministic growth rates for the foci in the CSM seems to oversimplify the real process. Therefore, a CSM with stochastic color dependent growth rates is introduced, which assumes that foci change their color when reaching a deterministic radius r_{switch} . As in the CSM, the formation of spherical foci with initial radius r_0 is described by a homogeneous Poisson process with rate μ. Let B_1 and B_2 be independent positive random variables with densities f_{B_1} and f_{B_2} . The random variables B_1 and B_2 describe the exponential growth of foci of color 1 and color 2.

Given that a focus is present at time t, the timepoint of its formation, τ_0, is a realisation of a random variable T uniformly distributed on $[0, t]$, where T , B_1 and B_2 are independent.

Consider exemplarily a focus generated at time $T = \tau_0$ which grows in color $C = 1$ with rate $B_1 = b_1$ until it reaches the radius r_{switch} , where it changes its color and grows in color $C = 2$ with rate $B_2 = b_2$.

Color 1: $R(t) < r_{switch}$:
The radius at time $t > \tau_0$, $R(t)$, is described by:

$$\boxed{R(t) = r_0 \exp(b_1(t - \tau_0))}$$

Color 2: $R(t) \geq r_{switch} \Leftrightarrow t > \frac{\ln(\frac{r_{switch}}{r_0})}{b_1} + \tau_0$.
Define

$$\tau_1 := \frac{\ln(\frac{r_{switch}}{r_0})}{b_1}$$

as the time spent in color 1 until change to color 2.
The radius of a focus of color $C = 2$ at timepoint $t > \tau_0 + \tau_1$, $R(t)$, is described by:

$$\boxed{R(t) = r_{switch} \exp(b_2(t - \tau_1 - \tau_0))}$$

So that an expression for the joint distribution of radius $R(t)$ and color $C(t) = 1$ at time t can be derived:

$$\begin{aligned}
&P(R(t) \leq r, C(t) = 1) \\
&\quad = \quad P(R(t) \leq r, R(t) \leq r_{switch}) \\
&\quad = \quad \begin{cases} 0 & r \leq r_0 \\ P(R(t) \leq r) & r \in (r_0, r_{switch}] \\ P(R(t) \leq r_{switch}) & r > r_{switch} \end{cases} \\
&\quad = \quad \begin{cases} 0 & r \leq r_0 \\ \frac{\ln(\frac{r}{r_0})}{t} \int\limits_{\frac{\ln(\frac{r}{r_0})}{t}}^{\infty} \frac{f_{B_1}(b_1)}{b_1} \, db_1 + F_{B_1}\left(\frac{\ln(\frac{r}{r_0})}{t}\right) & r \in (r_0, r_{switch}] \\ \frac{\ln(\frac{r_{switch}}{r_0})}{t} \int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty} \frac{f_{B_1}(b_1)}{b_1} \, db_1 + F_{B_1}\left(\frac{\ln(\frac{r_{switch}}{r_0})}{t}\right) & r > r_{switch} \, , \end{cases}
\end{aligned}$$

where F_{B_1} and f_{B_1} are distribution and density of the random variable B_1 .

Therefore the joint density of radius $R(t)$ and color $C(t) = 1$ at time t is:

$$\begin{aligned}
&f_{R(t),\,C(t)}(x,1) \\
&= \left[\frac{1}{xt}\int\limits_{\frac{\ln(\frac{x}{r_0})}{t}}^{\infty} \frac{f_{B_1}(b_1)}{b_1}\,db_1 - \frac{\ln(\frac{x}{r_0})}{t}\,\frac{f_{B_1}(\frac{\ln(\frac{x}{r_0})}{t})}{\frac{\ln(\frac{x}{r_0})}{t}}\,\frac{1}{xt} + f_{B_1}\left(\frac{\ln(\frac{x}{r_0})}{t}\right)\frac{1}{xt}\right] \\
&\quad \cdot 1_{(r_0,r_{switch}]}(x) \\
&= \frac{1}{xt}\int\limits_{\frac{\ln(\frac{x}{r_0})}{t}}^{\infty} \frac{f_{B_1}(b_1)}{b_1}\,db_1 \cdot 1_{(r_0,r_{switch}]}(x)\,.
\end{aligned}$$

with the indicator function:

$$1_{(a,b]}(x) := \begin{cases} 1 & x \in (a,b] \\ 0 & x \notin (a,b]\,. \end{cases}$$

The joint distribution of radius $R(t)$ and color $C(t) = 2$ at time t is:

$$\begin{aligned}
&P(R(t) \le r, C(t) = 2) \\
&= P(R(t) \le r, R(t) > r_{switch}) \\
&= \begin{cases} 0 & r \le r_{switch} \\ P(R(t) \le r | R(t) > r_{switch})P(R(t) > r_{switch}) & r > r_{switch}\,. \end{cases}
\end{aligned}$$

Therefore if $r > r_{switch}$:

$$\begin{aligned}
&P(R(t) \le r, C(t) = 2) \\
&= \int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty}\ \int\limits_{\frac{\ln(\frac{r}{r_{switch}})}{t-\tau_1}}^{\infty} \frac{\ln(\frac{r}{r_{switch}})}{b_2 t} f_{B_2}(b_2) f_{B_1}(b_1)\,db_2 db_1 \\
&+ \int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty}\ \int\limits_{-\infty}^{\frac{\ln(\frac{r}{r_{switch}})}{t-\tau_1}} \frac{t-\tau_1}{t} f_{B_2}(b_2) f_{B_1}(b_1)\,db_2 db_1 \\
&= \frac{\ln(\frac{r}{r_{switch}})}{t}\int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty}\ \int\limits_{\frac{\ln(\frac{r}{r_{switch}})}{t-\tau_1}}^{\infty} \frac{f_{B_2}(b_2) f_{B_1}(b_1)}{b_2}\,db_2 db_1 \\
&+ \frac{1}{t}\int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty} F_{B_2}\left(\frac{\ln(\frac{r}{r_{switch}})}{t-\tau_1}\right)(t-\tau_1) f_{B_1}(b_1)\,db_1,
\end{aligned}$$

where F_{B_1} and F_{B_2} , f_{B_1} and f_{B_2} are distributions and densities of the random variables B_1 and B_2 .

Hence the following expression for the joint density of radius $R(t)$ and color $C(t) = 2$ at time t is obtained:
Let $x > r_{switch}$:

$$
\begin{aligned}
f_{R(t),\,C(t)}(x,2) = {} & \frac{1}{xt} \int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty} \int\limits_{\frac{\ln(\frac{x}{r_{switch}})}{t-\tau_1}}^{\infty} \frac{f_{B_1}(b_1) f_{B_2}(b_2)}{b_2}\, db_2 db_1 \\
& - \frac{\ln(\frac{x}{r_{switch}})}{t} \int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty} f_{B_1}(b_1) \frac{f_{B_2}\left(\frac{\ln(\frac{x}{r_{switch}})}{t-\tau_1}\right)}{\frac{\ln(\frac{x}{r_{switch}})}{t-\tau_1}} \frac{1}{x\,(t-\tau_1)}\, db_1 \\
& + \frac{1}{t} \int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty} f_{B_2}\left(\frac{\ln(\frac{x}{r_{switch}})}{t-\tau_1}\right) \frac{1}{x\,(t-\tau_1)}\,(t-\tau_1)\, f_{B_1}(b_1)\, db_1 \\
= {} & \frac{1}{xt} \int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty} \int\limits_{\frac{\ln(\frac{x}{r_{switch}})}{t-\tau_1}}^{\infty} \frac{f_{B_1}(b_1) f_{B_2}(b_2)}{b_2}\, db_2 db_1 .
\end{aligned}
$$

For $x \leq r_{switch}$ is $f_{R(t),\,C(t)}(x,2) = 0$.

Therefore the joint densitity of radius $R(t)$ and color $C(t)$ at time t is:

Color 1

$$
\boxed{f_{R(t),\,C(t)}(x,1) = \frac{1}{xt} \int\limits_{\frac{\ln(\frac{x}{r_0})}{t}}^{\infty} \frac{f_{B_1}(b_1)}{b_1}\, db_1 \cdot 1_{(r_0, r_{switch}]}(x)}
$$

Color 2

$$
\boxed{\begin{aligned} f_{R(t),\,C(t)}(x,2) \quad = \quad & \tfrac{1}{xt} \int\limits_{\frac{\ln(\frac{r_{switch}}{r_0})}{t}}^{\infty} \int\limits_{\frac{\ln(\frac{x}{r_{switch}})}{t-\tau_1}}^{\infty} \tfrac{f_{B_1}(b_1) f_{B_2}(b_2)}{b_2}\, db_2 db_1 \\ & \cdot 1_{(r_{switch},\infty)}(x) \end{aligned}}
$$

3 Application to rat liver foci data

For a typical hepatocarcinogenesis experiment animals, e.g. rats, are treated with a carcinogen and liver sections are stained with special histological markers to observe foci of altered hepatocytes which are known to be precursor lesions of carcinoma. Measurements are made in two-dimensional liver sections and inference about the reality in three-dimensional liver is limited by the stereological problem. This problem is described briefly by the fact that the probability of a focus to be cut increases with its size. The model describes the three-dimensional situation. Moolgavkar and colleagues [5] suggested to translate the expressions for the distributions of size and number of foci in 3D into the corresponding expressions in 2D by the Wicksell-Transformation and to then apply the model to the two-dimensional measurements by maximum likelihood methods.

Consider that only focal transections with radii larger than ε can be detected and that one liver section per animal is evaluated. Kopp-Schneider and colleagues [4] derived the following expressions for the expected number of focal transections of color j at timepoint t in two dimensions[1],

$$\widetilde{n}_{2,j} = 2\mu t \int\limits_{\epsilon}^{\infty} \sqrt{x^2 - \epsilon^2} f_{R(t),C(t)}(x,j)\, dx, \tag{1}$$

and the density of the size distribution of focal transections of color j at timepoint t in two dimensions

$$\begin{aligned} & f_{R^{(2)}(t)|C(t)}(y|j) \\ = \; & \frac{y}{\int_{\epsilon}^{\infty} \sqrt{x^2-\epsilon^2} f_{R(t),C(t)}(x,j)\,dx} \int\limits_{y}^{\infty} \frac{1}{\sqrt{x^2-y^2}} f_{R(t),C(t)}(x,j)\,dx. \end{aligned} \tag{2}$$

Assume that foci of each animal grow and change their color independent of other foci. Let $n_{2,k}$ denote the number of focal transections of color k observed in a liver section of area A and let $r_{2,k,j}$ denote the radius of the j-th focal transection of color k. This liver section contributes the loglikelihood

$$\sum_{k=1}^{2} \left[(n_{2,k} \ln(A\widetilde{n}_{2,k}) - A\widetilde{n}_{2,k}) + \sum_{j=1}^{n_{2,k}} \ln(f_{R^{(2)}(t),C(t)}(r_{2,k,j},k)) \right] + C, \tag{3}$$

where C is a data dependent constant. Assuming that the liver sections of one experiment are independent of each other, the loglikelihood of the complete data set is the sum of the contributions of every section.

4 Example

Data from an NNM-experiment published by Weber and Bannasch in 1994 [8] are chosen to illustrate the methodology. In this study rats were treated

[1] To differentiate between the number of foci and the number of focal transections an additional index was introduced. Here the index 2 stands for two dimensions.

with 6mg NNM[2] per kg body-weight continuously during six different time-periods, 7, 11, 15, 20, 27 and 37 weeks, with each group consisting of five animals. After this time period one liver section of each rat was stained by the marker H&E [3] and different types of focal transections were observed. Here only two different types of foci are considered. The morphometric evaluation of the stained liver sections generated a data set consisting of the area of every liver section and the type and area of every focal transection detected in this section.
A Color-Shift-Model with color dependent and Beta-distributed growth rates is applied to this data set. Random variables B_1 and B_2 , which describe the exponential growth in color 1 and color 2, are Beta-distributed with parameters p_1, q_1, a_1 and p_2, q_2, a_2 . Form parameters, a_i , are introduced additionally to the parameters of the standard Beta-distribution, p_i and q_i ($p_i, q_i, a_i > 0,\ i = 1, 2$), to modify the support of the distribution function. Hence the growth rate in color i , B_i , is a positive random variable with the following density:

$$f_{B_i}(b_i) = \frac{1}{B(p_i, q_i)} \frac{b_i^{(p_i-1)}(a_i - b_i)^{(q_i-1)}}{a_i^{(p_i+q_i-1)}} \cdot 1_{[0,\, a_i]}(b_i) \quad p_i, q_i, a_i > 0,\ i = 1, 2\,,$$

where $B(p, q)$ is the Beta-function

$$B(p, q) = \int_0^1 z^{(p-1)}(1 - z)^{(q-1)}\, dz.$$

Inserting this expression into the joint density of radius $R(t)$ and color $C(t)$ at time t, double integrals are obtained in equations (1) and (2) which cannot be solved analytically. The loglikelihood function (3) depends on eight parameters.

4.1 Implementation

The MATLAB environment is used to compute the loglikelihood function, find the maximum likelihood parameters and visualize the results. Numerical double integration with singularities has to be performed for every single detected focal transection. As about 1000 focal transections are detected the computation of the likelihood is a very time-intensive procedure. Using the MEX-interface, functions for numerical double integration from the Fortran NAg library are included to improve the performance [4]. To find the maximum

[2] The chemical carcinogen N-Nitrosomorpholine (NNM) was administered in the drinking water.

[3] H&E stands for Hemalum&Eosin, a biological marker to identify acidophilic and basophilic cell structures.

[4] Subroutine D01DAF of Numerical Algorithm Groups (NAg), Fortran Library, version Mark 18 [6].

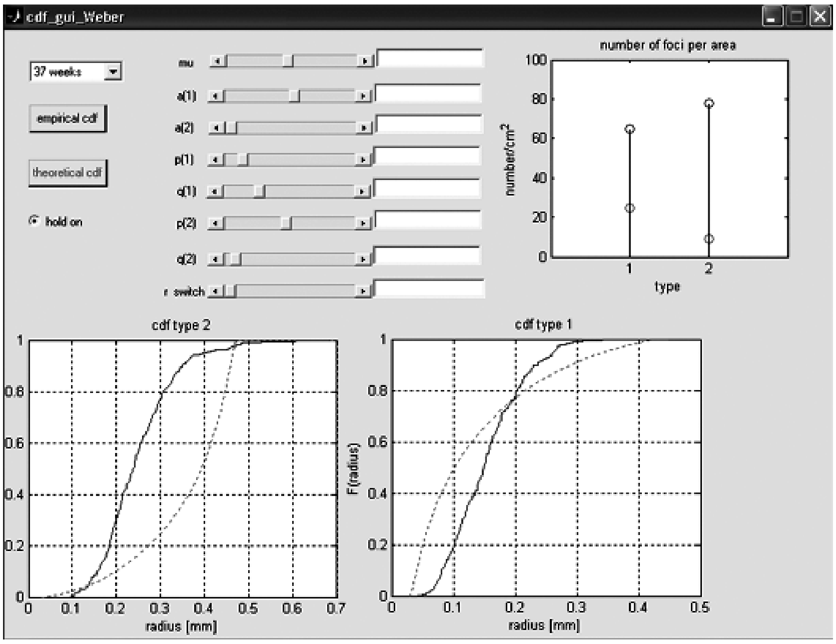


Figure 1: A time-point can be chosen over the pop-up-menue and the parameters can be varied over their corresponding sliders. Depending on the parameters the three axes show the theoretical distributions of size and number of focal transections of type 1 and 2 (dotted lines) compared with the empirical data taken from the NNM-experiment (solid lines). One slider is provided for the Poisson parameter μ, six sliders for the parameters p_1, q_1, a_1 and p_2, q_2, a_2 corresponding to the Beta-distributed growth rates in type 1 and 2 and one slider for r_{switch} .

likelihood parameters it is necessary to define a set of eight starting parameters for the *fmincon*[5] function and to define proper intervals for the range of the eight model parameters. For this purpose a graphical user interface (CSM-GUI) is implemented in MATLAB to test the theoretical distributions of size and number of focal transections in 2D under variation of parameters (Figure 1). After minimizing the negative loglikelihood by the *fmincon* function theoretical results can be compared with the empirical data.

4.2 Results

Figures 2 and 3 illustrate typical visualizations of the results of the modulation of the NNM-Experiment. The empirical size distribution is compared with the theoretical size distributions obtained from two different Color-Shift-Models using maximum likelihood estimates for the parameters. The CSM

[5] *fmincon* is a MATLAB function for nonlinear minimization under constraints used to minimize the negative loglikelihood function [7].

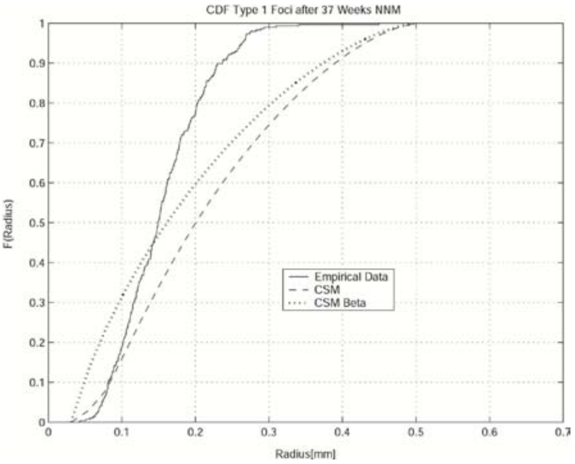


Figure 2: The result of the CSM (dashed line) and CSM with Beta-distributed Growth rates (dotted line) applied on foci of type 1 after 37 weeks. The solid line represents the empirical data.

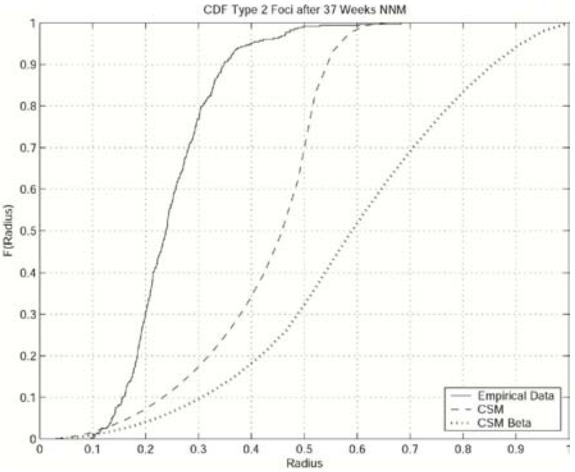


Figure 3: The result of the CSM (dashed line) and CSM with Beta-distributed Growth rates (dotted line) applied on foci of type 2 after 37 weeks. The solid line represents the empirical data.

without modifications is represented by the dashed line, the CSM with Beta-distributed growth rates is illustrated by the dotted line and the solid line stands for the empirical data from the NNM-experiment. Considering only type1-foci the modified CSM seems to predict the size distribution better than the CSM. But the visualizations for the focal transections of type 2 show that the modified CSM expects too large foci of the second type, so that there is an advantage for CSM without modification in this case. The deterministic switch-radius makes the model highly sensitive against outliers in type 1. A single large type 1 focus leads to a large estimate of r_{switch}. To make the model more robust against these outliers a model assuming a stochastic switch-radius has to be formulated.

5 Conclusions

The above mentioned forms of visualization are effective tools to compare different carcinogenesis models with respect to their ability to predict numbers and radii of foci in hepatocarcinogenesis experiments. Especially the CSM-GUI (Color-Shift graphical user interface) is a powerful instrument to test a new model before starting the very time-intensive procedure of finding the maximum likelihood parameters. To improve the CSM with stochastic growth rates a Color-Shift-Model with stochastic color dependent growth rates *and* stochastic switch-radius has to be introduced. The next step could be the integration of the whole process, finding the starting parameters, maximizing the loglikelihood function and visualizing the results in one GUI which could simplify the modulation.

References

[1] Burkholder I., Kopp-Schneider A. (2002). *Incorporating phenotype-dependent growth rates into the Color-Shift-Model for preneoplastic hepatocellular lesions*. Math. Biosci. **179**, 145.

[2] Geisler I. (2001). *Stochastische Modelle fuer den Mechanismus der Entstehung und der Progression von Krebsvorstufen in der Leber*. Doctoral thesis.

[3] Geisler I., Kopp-Schneider A. (2000). *A model for hepatocarcinogenesis with clonal expansion of three successive phenotypes of preneoplastic cells*. Math. Biosci. **168**, 167.

[4] Kopp-Schneider A., Portier C., Bannasch P. (1998). *A model for hepatocarcinogenesis treating phenotypical changes in focal hepatocellular lesions as epigenic events*. Math. Biosci. **148**, 181.

[5] Moolgavkar S., Luebeck E., de Gunst M., Port R., Schwarz M.(1990). *Quantitative analysis of enzyme-altered foci in rat hepatocarcinogenesis experiments. I. Single agent regimen.* Carcinogenesis **11**, 1271.

[6] NAg LTD (1997). NAg Fortran Library Manual, **Mark 18**.

[7] The Mathworks Inc. (2003). Matlab Documentation CD, **Release 13**.

[8] Weber E., Bannasch P. (1994). *Dose and time dependence of the cellular phenotype in rat hepatic preneoplasia induced by continuous oral exposure to N-Nitrosomorpholine*. Carcinogenesis **15**, 6.

[9] Wicksel S.D. (1925). *The corpuscle problem. A mathematical study of a biometrical problem.* Biometrica **17**, 87.

Address: J. Groos, A. Kopp-Schneider, German Cancer Research Center, Biostatistics, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany

E-mail: `j.groos@dkfz.de`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

DESIGN ASPECTS OF A COMPUTER SIMULATION STUDY FOR ASSESSING UNCERTAINTY IN HUMAN LIFETIME TOXICOKINETIC MODELS

Harald Heinzl and Martina Mittlboeck

Key words: Dioxin, indeterminability, occupational cohort, Monte Carlo simulation study.

COMPSTAT 2004 section: Biostatistics.

Abstract: The general paradigm for risk assessment of exposures to toxic agents in human environment is the identification and characterization of hazard, assessment of exposure and characterization of risk. Performed in practice risk assessment is addressing particularly the outcome of integrating the data available from epidemiology, long-term mortality and morbidity studies and mechanistic research with information on the type and extent of exposure, as well as statistical analysis properly used. Various technical and non-technical aspects of the design process of the Monte Carlo simulation study will be reported and discussed. Finally, a Monte Carlo computer simulation study was designed in order to examine in detail the influences of various sources of uncertainty and their potential implications on the risk estimates from the Boehringer cohort data is presented.

1 Introduction

The need for risk assessment of exposures to toxic agents in human environment has increased steadily over the last decades. The general paradigm for risk assessment is the identification and characterization of hazard, assessment of exposure and characterization of risk. Performed in practice risk assessment is addressing particularly the outcome of integrating the data available from epidemiology, long-term mortality and morbidity studies and mechanistic research with information on the type and extent of exposure, as well as statistical analysis properly used. A sound, scientifically based risk assessment is an essential tool for risk managers and legislators responsible for security and safety of humans.

The use of toxicokinetic models makes it possible to construct exposure indices that may be more closely related to the individual dose than traditional exposures measures. However, the process introduces a wide array of sources of uncertainty, which inevitably makes risk assessment more difficult. In addition, representing population heterogeneity in the assessment of risks and the identification of sensitive sub-population is of great concern.

The analysis of uncertainty is becoming an integral part of many scientific evaluations. For example, in the risk assessment process, an uncertainty analysis has been recognized as an important component of risk characterization by regulatory agencies [29]. Uncertainty is prevalent in the process of risk assessment of chemical compounds at various levels. Uncertainty of the exposure assessment influences dose estimates. Such effects are exaggerated further by uncertainty in dose-response modelling, mainly caused by limited knowledge about the functional dose-response relationship. Finally, uncertainty is propagated to the risk estimation procedure, which provide the basis for risk management decisions.

It is vital to distinguish uncertainty from variability: The latter is a phenomenon in the physical world to be measured, analysed and where appropriate explained. By contrast, uncertainty is an aspect of knowledge (Sir David Cox as quoted in Vose [28]. Total uncertainty is the combination of variability and uncertainty. To avoid confusion it was suggested to rename total uncertainty by indeterminability [28], a terminology adopted in our work.

Our example focusses on the risk assessment process whether 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD, “Seveso-dioxin”) is a potential human carcinogen. In 1997 TCDD was evaluated as human carcinogen [19], [22]. The decision substantially relied on empirical studies of highly exposed occupational cohorts. The so-called Boehringer cohort was amongst them, and its data were thoroughly analysed [1], [2], [13] [14], [24]. These statistical analyses were a rather delicate task as amongst other things individual lifetime TCDD-exposures starting in the 1950ies had to be reconstructed from TCDD-measurements in the 1980ies and 1990ies when such measurements became feasible and affordable. Inevitably, a lot of uncertainty remained due to lack of longitudinal physiological data, the possibility of measurement errors and workplace misclassification errors, disagreement about the appropriate statistical analysis strategy, limited knowledge about the functional dose-cancerogenic property relationship and the advent of new toxicokinetic insight - just to name a few circumstances.

Now, it is quite common that results of large-scaled statistical or epidemiological analyses will be questioned and disputed. However, the goal of an uncertainty analysis is to tell us how much we can be wrong and still be okay [7]. Therefore we designed a computer simulation study to be able to examine in detail the influences of various sources of uncertainty and their potential implications on the risk estimates from the Boehringer cohort data.

The paper ist organized as follows. In Section 2 our adopted view of uncertainty analysis is defined in brief. Section 3 is devoted to dioxin, that is, general characteristics of the compound, features of the Boehringer cohort data set and various approaches to model lifelong human toxicokinetics are described. Section 4 contains technical and non-technical design aspects of the intended computer simulation study. In Section 5 a brief discussion is given.

2 Indeterminability, variability and uncertainty

Indeterminability (or total uncertainty) denotes the inability to be able to precisely predict what the future holds. The two components of indeterminability are variability and uncertainty [28]. According to Hodges [18] a statistical, a structural and a technical part of indeterminability can be distinguished (see also [12], [17]. The statistical part corresponds to variability, whereas the other two parts correspond to uncertainty. The statistical part of indeterminability is variation given structure or in other words, residuals given a model, a common statistical technique to describe variability in a regression model.

Structural uncertainty emerges from the fact that the model itself – the assumed structure - may be uncertain either due to incomplete or insufficient knowledge about biological, physiological or toxicological mechanisms, or due to the existence of more than one way to explain a specific phenomenon, that is, there are several plausible models. A special and very important aspect of structural uncertainty is the so-called model parameter uncertainty [12], i.e. uncertainty about model assumptions and model constants. In toxicokinetic models e. g., total lipid volume of the body may be assumed non-varying over human life time or the elimination halflife of a certain toxin may be considered known in one approach, whereas it may not in another.

The third part of indeterminability in Hodges' classification is technical uncertainty, which mainly comprises the ordinary and unspectacular circumstances of everyday scientific work. It is usually neglected although occassionally it may allocate a considerable fraction of indeterminability. Examples for technical uncertainty are poor quality of raw data (e.g. typos, rounding errors), numerical estimation problems, in particular in connection with complex nonlinear models, or research limitations due to lack of resources (e.g. software, time, human expertise), which may artificially restrict the spectrum of considered scientific models or employed statistical analysis methods.

3 Dioxin at a glance

3.1 Polychlorinated dibenzodioxins and -furans (PCDD/Fs)

PCDD/Fs are highly lipophilic synthetic chemicals which arise primarily from the production and combustion process of chlorinated chemicals and as a byproduct to chlorinated bleaching and waste incineration. Environmental contamination by PCDD/Fs has been documented worldwide and is ubiquitous. In industrialised countries the PCDD/F burden of the population is assumed to result mainly from intake of contaminated food. Improvements in the analytical techniques used to measure PCDD/F concentrations have allowed for the concentration of these compounds to be assessed in reasonable amounts of human tissue, most notably in adipose tissue, blood serum and plasma. Repeated determinations in humans allow the investigation of the kinetic of these toxins.

TCDD is believed to be the most potent of the PCDD/Fs. Numerous effects in humans have been observed from exposure to TCDD; amongst them are lung cancer and soft-tissue sarcoma. Observed adverse health effects other than cancer include chloracne, altered sex hormone levels, altered development outcomes, altered thyroid function, altered immune function, cardiovascular diseases and neurological disorders to name just a few, e.g. see also the survey of Grassman et al. [15]. The establishment of a causal relationship between exposure to dioxins and diseases in humans is of outstanding significance in public health and disease prevention. To establish such a causal link is extremly difficult since chronic diseases may occur a long time after the actual exposure has ceased and this extended lag time (latency period) between exposure and disease onset may obscure a causal link. This implies the need for proper modelling of the individual intoxination process in order to construct appropriate dose metrics (like area under the concentration-time curve) for quantitative representation of the disease-exposure relationship. Obviously it is essential to relate the occurrence of diseases to dioxin levels experienced during the exposure before disease onset. Previous levels have to be estimated from present ones. Retrospective determination of dioxin levels in humans and their subsequent use in risk assessment are strongly connected to the toxicokinetics of the dioxins. Chronic environmental exposure, route of exposure, storage in adipose tissue, and mechanism of elimination are important determinants of the level of TCDD in serum years after possibly high occupational exposures. Currently available physiologically based pharmacokinetic (PBPK) models try to meet this requirements at least partly.

Occupationally exposed cohorts are an important source of information due to more pronounced effects (occupational exposures are higher in general) and improved ability to control for confounders (easier and more reliable information retrieval among workers registered in files of companies or insurance agencies). For workers in the chemical industry, where occupational exposure to dioxins has occured in past production periods, the establishment of causal relationships is also connected to insurance and compensation issues, which requires an individually-based assessment of exposure, disease onset and their relationship. In 1997 the International Agency for Research on Cancer (IARC) reevaluated TCDD as carcinogenic to humans (IARC group 1 classification) on the basis of limited evidence of carcinogenicity to humans and sufficient evidence of carcinogenicity in experimental animals [19], [22]. The most important studies, which gave evidence with respect to human carcinogenicity, were four cohort studies with adequate follow-up times of herbicide producers, one each in the United States and the Netherlands, two in Germany. The largest and most heavily exposed German cohort is the so-called Boehringer cohort [13], [14], [1], [2]. Main features of the Boehringer cohort are described in the next Subsection.

Overall, the strongest evidence for TCDD carcinogenicity is for all cancers combined, not for a specific site. Due to the lack of a clearly predominating site it was considered by the IARC that there is limited evidence in humans for the carcinogenicity of TCDD [19], [22]. This could be due to still limited power of those epidemiological studies requiring cautious appreciation, or due to an unspecific non-standard carcinogenic action of dioxin. The evidence in humans for the carcinogenicity of all other PCDDs is even more diffuse and was rated inadequate by the IARC in 1997.

3.2 The Boehringer cohort

The Boehringer cohort consists of around 1600 workers occupationally exposed to PCDD/Fs. About a quarter of the workers are women. The cohort members came from two plants operated by the C.H. Boehringer Sohn Chemical Company, one in Ingelheim and the other in Hamburg, Germany. In Ingelheim 2,4,5 trichhlorphenol (TCP) was produced from 1950 to 1954, in Hamburg TCP was produced from 1957 until contamination with dioxins was stopped in April 1983 and the plant was finally closed in October 1984 [5], [21]. Since 1984, an investigation programme independent of the C.H. Boehringer Sohn Chemical Company has been performed by the Institute of Occupational and Social Medicine of the University of Mainz [5]. Comprising 186 persons evaluable for health evaluation in the first phase from 1984 until 1989 and comprising 192 in a second medical investigation program started in 1992 biomonitoring data on TCDD and major PCDD/F congeners and severe polychlorinated biphenyl congeners have been obtained from samples from adipose tissue or blood serum lipids [3]. This cohort was further investigated in a follow-up study using dioxin concentration measurements for 88 persons [4].

The Ingelheim and Hamburg plants can be subdivided into about 20 working areas corresponding to different involvement in the production processes (e.g. bromophos production, trichlorophenol production, 2,4,5-trichlorophenoxyacetic acid production, repair, laundry, administration, etc.), believed to result in different exposures levels to dioxins. Work histories were documented using a recall questionnaire asking for the start of employment, end of employment and sojourn times in the working areas.

3.3 Available toxicokinetic models

A series of PBPK models for lifelong TCDD exposure in humans are available in the literature. Nearly all of them assume a linear elimination kinetic, they only differ in the sophistication how time-dependent physiological variables as body weight, body fat volume or liver fat volume are considered (e.g. [11], [10]; [23]; [20], [14], [26]. The model of Carrier et al. [8], [9] is an exception in terms of the elimination function which is based on a modified Michaelis-Menten function.

Of course, more biologically complex mechanistic models could be suggested. Phenomena such as TCDD absorption, distribution, binding to liver receptors, enzyme induction, and synthesis of binding proteins could be considered. However, such phenomena occur on a much faster time scale (hours to days) than TCDD elimination (years in humans), which finally justifies the assumption of an quasi-equilibrium between TCDD in lipid fraction of blood, liver and adipose tissue. Note that this assumption (or variations of it) is made either explicitly or implicitly in all of the lifelong TCDD models for humans mentioned above.

4 The computer simulation study

The planning of a large computer simulation study comprises of technical and non-technical issues. The technical issues coincide to a large extent with the problem analysis and design step of the common three-step software development process (where the third step is implementation).

The non-technical issues consist of various essential prerequisites and fundamental decisions. Treating them lightly could seriously jeopardise the success of the whole project.

4.1 Problem analysis

At first, it is necessary to analyse plausible PBPK models for human lifetime toxicokinetics of TCDD and integrate them into more comprehensive models. Among others, these model should allow for multiple exposure to different toxins with similar kinetics (PCDD/Fs instead of just TCDD alone), chronical exposure (both background and workplace) and pointwise exposure (e.g. through accidents). Usage of these models is in establishing a dose-response relationship for a proper risk assessment of TCDD. The models have also to allow the construction of individual human exposure profiles over longer time periods. Ideally, one wide-ranging model could be found, from where all others deduce as special cases.

This approach or these different approaches in modelling individual human lifetime toxicokinetics could be mechanistically compared under various realistic scenarios, e.g. temporal change in background exposure, spatial change in workplace exposure, high accidental exposure over a short time, effects of fattening and loosing weight during lifetime, lifetime effects of breastfeeding (both in contaminated women and in persons who during childhood have been breast-fed by a contaminated woman), sensitivity in model parameters, effects of congeners other than TCDD, effects of a confounder variables like smoking status (in particular effects of ignoring them), effects of ignoring interaction terms in the model (interactions among two mutually different congeners or among a congener and a confounder variable). The construction of exposure indices from individual concentration-time curves could also be studied.

The main part of the project are Monte Carlo computer simulations in order to assess uncertainty in the toxicokinetic modelling process up to its implications on risk assessment. The main issues to be studied are amongst other things:

- Uncertainty in choice of PBPK model assumptions: E.g. assume non-linear kinetic for toxin elimination to generate data and use linear kinetic for analysis. The goal is to identify those model assumptions which are particularly sensitive to dose level prediction. A sensitivity in model parameters to interindividual variation: E.g. individualise age-related changes of body fat volume.

- Uncertainty caused by measurement of toxin levels: Different laboratories report different dioxin levels for the same sample. In the Boehringer data differences of 50% or more occur frequently [12, Figure 4b].

- Uncertainty caused by workplace misclassifications: Participants of the Boehringer study have been asked about their working history. These interviews have been repeated at a later time point. Comparisons revealed that 50% of the reported working times and 30% of the reported working areas did not match between two interviews [12].

- Uncertainty caused by different approaches to model the covariance structure of repeated measurements

- Uncertainty due to choice of statistical estimation method

- Effects of missing values and unknown confounders

- Uncertainty in choice of appropriate exposure index, lag time and dose-response relationship: This form of uncertainty concerns the subsequent processing of the toxicokinetic results in dose-response models. Even if the former would yield absolutely correct values, uncertainty in the latter would still distort the results of the risk assessment process.

- Selection effects: They could have been easily occurred in the Boehringer cohort data as participation in the dioxin measurement program was on a voluntary basis. A specific form of selection bias is the so-called “healthy worker survivor effect” (see e.g. [25]).

To meet these requirements a computer program library with a flexible modular structure has to be designed and implemented (see next Subsection). Thereby note that uncertainty analysis can only shed light onto overlooked issues, underrated issues or issues which have not been known at the time of original analysis itself. It is probable that some time after the completion of the uncertainty analyses new scientific theories may evolve, e.g. a new toxicokinetic TCDD model for humans. The design of the computer program library should allow a flexible and smooth integration of currently unknown but supposable future developments.

There are numerous adequate software products available where the computer program library could be implemented so that the actual decision is mainly a matter of personal preference. In the current case the computer program library is implemented in form of SAS macros (SAS Institute Inc., Cary, NC, USA).

4.2 Program library for Monte Carlo simulations

The main goal of the simulation study is to mimick the essential features of both the Boehringer cohort and the corresponding statistical analyses. Four main computer program modules can be distinguished.

4.2.1 Simulation of whole cohort. The simulated plant is operating between 1950 and 1985. Amongst other things five main working areas with different TCDD working exposure levels are assumed. The exposure levels are assumed to follow a lognormal distribution with mean intake of 3500, 150, 40, 5 and 0 TCDD units/year, respectively. Mean background exposure is set to 1 unit/year. The mean values closely resemble the actual exposure estimates as reported in Becher et al. [1]. The highest exposure occurs solely in the 1950ies. Determination of TCDD concentrations in the simulated workers happens in 1990 and 1995. The willingness of the workers to participate in the TCDD screening programme is simulated as well. The numbers of workers in the simulated cohort and in the simulated TCDD screening programme should approximately resemble the corresponding numbers in the Boehringer cohort.

Individual change of working area, termination of work contract, retirement and death of the virtual workers are randomly simulated as well as hiring of new workers. TCDD elimination kinetic is generated according to four different scenarios, that is, simple linear kinetic with constant total lipid volume (TLV) over lifetime, simple linear kinetic with TLV varying with workers age, linear kinetic according to Thomaseth and Salvan [26] with TLV and liver lipid volume varying with workers age, and modified Michaelis-Menten kinetic with body weight varying with workers age [8], [9]. During lifetime the simulated workers are subject to develop one of two kinds of cancer. Development of cancer will increase mortality of a simulated worker and will entail his retirement. The functional dose-cancer response relationship of TCDD is modelled by increasing the hazard for the first kind of cancer proportionally to the individual TCDD exposure during lifetime. Various TCDD exposure indices can be explored (e.g. area under the concentration-time curve (AUC), lagged AUC, etc.).

Due to the hazard increase in the first out of two kinds of cancer the existence of a predominating cancer site is simulated.

4.2.2 Measurement errors. In module 4.2.1 simulated true values are recorded. These will be contaminated with TCDD measuring errors, workplace misclassification errors, etc. in order to get simulated observed values.

4.2.3 Workplace exposure backcalculation. TCDD measurements are available only a long time after the actual workplace exposure. Under plausible assumptions (concerning background exposure, fat fraction of body, form of elimination from body, etc.) the exposure levels in different working areas can be estimated by backcalculation. There have been two different main attempts to perform a backcalculation, one is described in detail by Becher et al. [1], the other is due to Portier et al. [24]. Both attempts can be compared with this program module [16].

4.2.4 Risk estimates. Extract various individual time-dependent exposure indices for all members of the simulated cohort. Assess dose-response relationship between these time-dependent exposure indices and cancer incidence and mortality by use of Cox regression models, Poisson regression models and standardised mortality ratio analyses [1]. The final results of this simulation module are cancer risk estimates which in reality would provide the decision basis for risk managers.

4.3 Miscellaneous non-technical issues

When prearranging uncertainty investigations then their time demand should be accordingly taken into account. The availability of a detailed and profound documentation of the statistical analyses in question is an important prerequisite. Risk assessment for dioxins is an interdisciplinary effort. The integration of research results from various scientific disciplines such as toxicology, molecular biology, biochemistry, medicine, epidemiology and biostatistics is required. It is self-evident that each isolated effort would be doomed to failure. An uncertainty analysis is no exception. Arrangements have to be made in order to allow the permanent discussion of assumptions and results with exponents of the other scientific disciplines.

It is an open question who should do the uncertainty analysis. Two options are obvious: the uncertainty analysis is performed within the team which did the original statistical analysis or outside this team. The pros of the former case are evident, that is, already existing knowledge of the matter will result in efficient work (and usually there will be some kind of uncertainty assessment already during the performance of a statistical analysis). However, the cons are evident as well. That is, if somebody works over a longer period of time on a certain problem, then some sort of factory blindness will be hardly avoidable. On the other hand, if somebody from outside the statistical analysis team performs the uncertainty assessment, then this person will usually have another main focus onto the research problem and new ideas may be developed due to the non-involvement in the original analysis. The cons of this approach are in the greater effort to familiarise with the subject and a possibly difficult relationship to the team members of the original analysis. These considerations should be made an integral part in the projects statistical analysis schedule from the beginning.

Here a rather traditional Monte Carlo simulation study is utilised for uncertainty assessment. It mainly consists of the exploration and evaluation of different interesting scenarios. Alternatively, an uncertainty assessment could be performed within a fully Bayesian framework (see e.g. [6], [7]. A detailed comparison of the pros and cons of both approaches is beyond the scope of this paper.

5 Discussion

Risk assessment is a vital activity in modern society because it provides the scientific basis for effort to identify and control hazards to health and life. However, risk assessment is generally subject to great uncertainty. The scientific knowledge available in this field is far from sufficient. Uncertainty in risk assessment is at present a major but largely unsolved problem to be faced with solid research.

The goal of uncertainty analysis is to provide an evaluation of the limits of our knowledge, or in other words, an uncertainty analysis should tell us how much we can be wrong and still be okay [7].

Uncertainty assessment of large-scaled statistical analyses is obviously a reasonable and essential task in the empirical research process. In our view it is useful to consider the idea of indeterminability which can be subdivided into statistical variability, structural and technical uncertainty [18], [12], [17].

Analytical approaches to assess structural and technical uncertainty will be easily limited by the complexity of the underlying problems. However, elaborate computer simulation studies have evolved as an appropriate tool for the investigation of these types of indeterminability [28].

Obviously, analysis of uncertainty comprises uncertainty itself. During an uncertainty analysis various decisions about parameter settings (e.g. constant or random, distribution type and distribution parameters, etc.) have to be made. Actually, this settings would require an uncertainty analysis of its own. That is, there would be meta-uncertainty - the uncertainty of the uncertainty analysis. And then there would be meta-meta-uncertainty, the uncertainty of the meta-uncertainty analysis such that we would built one layer of uncertainty on another and finally miss the goal. The loophole in this catch is the insight that uncertainty analyses are not done on their own, but are part of the scientific research process. Accordingly, the results of an uncertainty analysis should be communicated to the scientists who posed the research question, collected the data and performed the statistical analysis on the one hand as well as to other experts in the field on the other hand. Together these researchers will be able to assess the validity of the uncertainty analysis and to discuss the consequences of the results [17].

References

[1] Becher H., Flesch-Janys D., Gurn P., Steindorf K. (1998a). *Berichte 5/98, Krebsrisikoabschätzung für Dioxine, Risikoabschätzungen für das Krebsrisiko von polychlorinierten Dibenzodioxinen- und Furanen (PCDD/Fs) auf der Datenbasis epidemiologischer Krebsmortalitätsstudien.* Forschungsbericht im Auftrag des Umweltbundesamtes, Erich Schmidt Verlag, Berlin.

[2] Becher H., Steindorf K., Flesch-Janys D. (1998b). *Quantitative cancer risk assessment for dioxins using an occupational cohort.* Environ Health Perspect **106** (Suppl 2), 663 – 670.

[3] Beck H., Eckart K., Mathar W., Wittkowski R. (1989). *Levels of PCDD's and PCDF's in adipose tissue of occupationally exposed workers.* Chemosphere **18**, 507-516.

[4] Benner A., Edler L., Mayer K., Zober A. (1993). *Untersuchungsprogramm "Dioxin" der Berufsgenossenschaft der chemischen Industrie. Ergebnisbericht - Teil II.* Arbeitsmedizin, Sozialmedizin, Umweltmedizin **29**, 11 – 16.

[5] BG Chemie. (1990). *Untersuchungsprogramm 'Dioxin', Ergebnisbericht - Teil I. Berufsgenossenschaft der Chemischen Industrie.* BG Chemie (Ed.), Heidelberg, ISBN: 3-88338-302-9.

[6] Bois F.Y. (1999). *Analysis of PBPK models for risk characterization.* Annals of the New York Academy of Sciences **895**, 317 – 337.

[7] Bois F.Y., Diack C. (2004). *Uncertainty analysis.* In: Quantitative Methods for Cancer and Human Health Risk Assessment, Edler L., Kitsos C.P. (Eds.), Wiley, Chichester, to appear.

[8] Carrier G., Brunet R.C., Brodeur J. (1995a). *Modeling of the toxicokinetics of polychlorinated dibenzo-p-dioxins and dipenzofurans in mammalians, including humans.* I. Nonlinear distribution of PCDD/PCDF body burden between liver and adipose tissues. Toxicology and Applied Pharamcology **131**, 253 – 266.

[9] Carrier G., Brunet R.C., Brodeur J. (1995b). *Modeling of the toxicokinetics of polychlorinated dibenzo-p-dioxins and dipenzofurans in mammalians, including humans.* II. Kinetics of absorption and disposition of PCDDs/PCDFs. Toxicology and Applied Pharamcology **131**, 267- 276.

[10] Caudill S.P., Pirkle J.L., Michalek J.E. (1992). *Effects of measurement error on estimating biological half-life.* Journal of exposure analysis and environmental epidemiology **2**, 463 – 476.

[11] Craig T.O., Grzonka R.B. (1991). *A time-dependent 2,3,7,8-tetrachlorodibenzo-p-dioxin body-burden model.* Arch. Environ. Contam. Toxicol. **21**, 438 – 446.

[12] Edler L. (1999). *Uncertainty in biomonitoring and kinetic modeling.* Annals of the New York Academy of Sciences **895**, 80 – 100.

[13] Flesch-Janys D., Berger J., Gurn P., Manz A., Nagel S., Waltsgott H., Dwyer J.H. (1995). *Exposure to polychlorinated dioxins and furans (PCDD/F) and mortality in a cohort of workers from a herbicide-producing plant in Hamburg, Federal Republic of Germany.* American Journal of Epidemiology **142**, 1165 – 1175. Published erratum in American Journal of Epidemiology (1996) **144**, 716.

[14] Flesch-Janys D., Steindorf K., Gurn P., Becher H. (1998). *Estimation of the cumulated exposure to polychlorinated dibenzo-p-dioxins/furans and standardized mortality ratio analysis of cancer mortality by dose in an occupationally exposed cohort.* Environ Health Perspect **106** (Suppl 2), 655 – 662.

[15] Grassmann J.A., Masten S.A., Walker N.J., Lucier G.W. (1998). *Animal models of human response to dioxins.* Environ Health Perspect **106** (Suppl 2), 761 – 775.

[16] Heinzl H., Edler L. (2002). *Assessing uncertainty in a toxicokinetic model for human lifetime exposure to TCDD.* Organohalogen Compounds **59**, 355 – 358.

[17] Heinzl H., Edler L. (2003). *Evaluating and assessing uncertainty of large-scaled statistical analyses exemplified at the Boehringer TCDD cohort.* Proceedings of the second workshop on research methodology,. Ader H.J., Mellenbergh G.J. (Eds), VU University, Amsterdam, ISBN 90-5669-071-X, 87-94.

[18] Hodges J.S. (1987). *Uncertainty, policy analysis and statistics.* Statistical Science **2**, 259 – 291.

[19] IARC. (1997). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans.* Vol. **69**: Polychlorinated Dibenzo-para-dioxins and Polychlorinated Dibenzofurans. International Agency for Research on Cancer, Lyon.

[20] Kreuzer P.E., Csanády Gy.A., Baur C., Kessler W., Päpke O., Greim H., Filser J.G. (1997). *2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) and congeners in infants.* A toxicokinetic model of human lifetime body burden by TCDD with special emphasis on its uptake by nutrition. Arch. Toxicol. **71**, 383 – 400.

[21] Manz A., Berger J., Dwyer J.H., Flesch-Janys D., Nagel S., Waltsgott H. (1991). *Cancer mortality among workers in chemical plant contaminated with dioxin.* Lancet **338**, 959 – 964.

[22] McGregor D.B., Partensky C., Wilbourn J., Rice J.M. (1998). *An IARC Evaluation of Polychlorinated Dibenzo-p-dioxins and Polychlorinated Dibenzofurans as Risk Factors in Human Carcinogenesis.* Environ Health Perspect **106** (Suppl 2), 755 – 760.

[23] Michalek J.E., Pirkle J.L., Caudill S.P., Tripathi R.C., Patterson D.G. Jr., Needham L.L. (1996). *Pharmacokinetics of TCDD in veterans of operation ranch hand: 10-year follow-up.* Journal of toxicology and environmental health **47**, 209 – 220.

[24] Portier C.J., Edler L., Jung D., Needham L., Masten S., Parham F., Lucier G. (1999). *Half-lives and body burdens for dioxin and dioxin-like compounds in humans estimated from an occupational cohort in Germany.* Organohalogen Compounds **42**, 129 – 137.

[25] Steenland K., Deddens J., Salvan A., Stayner L. (1996). *Negative bias in exposure-response trends in occupational studies: modeling the healthy worker survivor effect.* American Journal of Epidemiology **143**, 202 – 210.

[26] Thomaseth K., Salvan A. (1998). *Estimation of occupational exposure to 2,3,7,8-tetrachlorodibenzo-p-dioxin using a minimal physiologic toxicokinetic model.* Environ Health Perspect **106** (Suppl 2), 743 – 753. Published erratum in Environ Health Perspect (1998) **106** (Suppl 4), CP2.

[27] Van der Molen G.W., Kooijman S.A.L.M., Slob W. (1996). *A generic toxicokinetic model for persistent lipophilic compounds in humans: an application to TCDD.* Fundamental and applied toxicology **31**, 83 – 94.

[28] Vose D. (2000). *Risk analysis: a quantitative guide.* 2nd ed., Wiley, Chichester.

[29] WHO. (1995). *Application of risk analysis to food standard issues.* Report of the Joint FAO/WHO Expert Consultation. World Health Organization, Geneva.

Acknowledgement: We particularly emphasise the generous support of Lutz Edler and his colleagues of the Biostatistics Unit of the German Cancer Research Center in Heidelberg, Germany. Furthermore, the study was supported in parts by grant J 1823 of the Austrian Science Fund.

Address: H. Heinzl, M. Mittlboeck, Department of Medical Computer Sciences, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

E-mail: harald.heinzl@meduniwien.ac.at,
martina.mittlboeck@meduniwien.ac.at

COMPSTAT'2004 Symposium

SIMULTANEOUS INFERENCE IN RISK ASSESSMENT; A BAYESIAN PERSPECTIVE

Leonhard Held

Key words: Risk assessment, Monte Carlo, simultaneous credible bands, simultaneous inference.

COMPSTAT 2004 section: Biostatistics.

Abstract: We consider the problem of making simultaneous inferential statements in risk assessment from a Bayesian perspective. We review a generic algorithm for computing a two-sided simultaneous credible band based on Monte Carlo samples from a multidimensional posterior distribution. A simple modification leads to an upper or lower simultaneous credible bound, which will be described. Such simultaneous credible bands and bounds have attractive properties: they are easy to calculate, completely non-parametric and invariant to monotone component-wise transformations of the variables. We illustrate the proposed approach through an example from low-dose risk estimation, previously analysed in the literature with frequentist methods.

1 Introduction

Statistical risk assessment deals with the probabilistic quantification of potential damaging effects of an environmental hazard. Of particular importance is the formulation and estimation of dose-response relationships based on data from controlled toxicological studies. This paper takes a Bayesian view to the statistical problem of estimating the dose-response relationship and derived quantities. Such an approach has at least two useful features: First, the posterior distribution of any function of the original parameters can be derived *exactly* using Monte Carlo simulation; secondly, pointwise and simultaneous credible bands and bounds can be computed *exactly* up to Monte Carlo error.

From a freqentist perspective, the calculation of simultaneous confidence bands has been developed in Pan, Piegorsch and West [8], and has been applied to risk assessment estimation in Al-Saidy et al. [1] and Piegorsch et al. [9]. Al-Saidy et al. [1] consider quantal response data with a binomial likelihood while Piegorsch et al. [9] apply the methods to continuous measurements based on a quadratic regression model. In this paper we re-analyze the data from Piegorsch et al. [9], but use a Bayesian approach based on Monte Carlo sampling. In particular, we develop methods to calculate *simultaneous credible bounds* for the *benchmark dose* at various *benchmark risks*.

The paper is organized as follows. In Section 2 we review an algorithm to calculate (two-sided) simultaneous credible bands based on Monte Carlo samples from a posterior distribution and outline a straightforward modification to obtain one-sided simultaneous credible bounds. In Section 3 we apply these methods to a problem from low-dose risk assessment and compare our results with those obtained by Piegorsch et al. [9] using frequentist methods. We close with some discussion in Section 4.

2 Monte Carlo estimation of simultaneous credible bands and bounds

2.1 Two-sided credible bands

Assume that we have a sufficiently large sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ from a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, obtained through simple Monte Carlo, or more advanced Markov chain Monte Carlo (MCMC) simulation. Here $\boldsymbol{\theta}$ is an unknown parameter of dimension p, perhaps obtained after suitable transformation of the original parameters in the model.

The approach proposed in Besag et al. [2, Section 6.3] starts with sorting and ranking the samples separately for each parameter of interest θ_i, $i = 1, \ldots, p$. Let $\theta_i^{[j]}$ denote the corresponding order statistic and $r_i^{(j)}$ the rank of $\theta_i^{(j)}$, $j = 1, \ldots, n$. Let j^* be the smallest integer such that the hyperrectangular defined by

$$[\theta_i^{[n+1-j^*]}, \theta_i^{[j^*]}], \quad i = 1, \ldots, p \tag{1}$$

contains at least k of the n values $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$. Besag et al. point out that j^* is equal to the kth order statistic of the set

$$S = \left\{ \max \left\{ n + 1 - \min_i r_i^{(j)}, \max_i r_i^{(j)} \right\}, j = 1, \ldots, n \right\}. \tag{2}$$

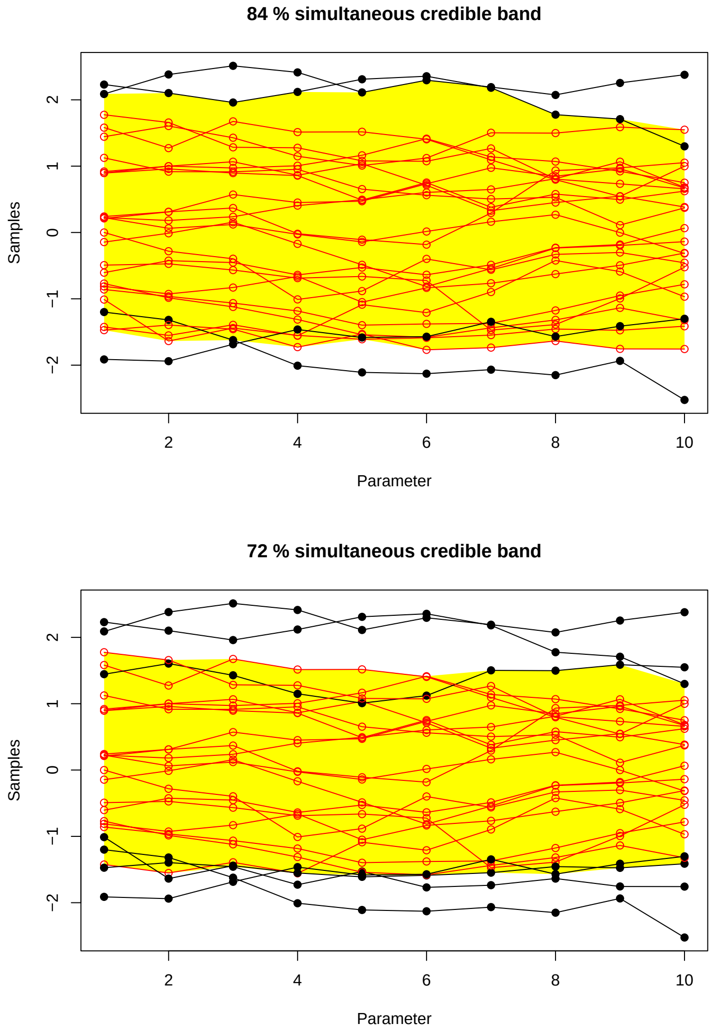
By construction, the credible region (1) will then contain (at least) $100k/n\%$ of the empirical distribution.

Figure 1 illustrates the construction of simultaneous credible bands for simulated data with $n = 25$ and $p = 10$. Each line corresponds to one sample $\boldsymbol{\theta}^{(j)}$ while each column represents a parameter θ_i. The yellow band is a simultaneous credible band of empirical coverage 84 and 72%. The set (2) is in this example

$$S = \{16, 17, 17, 18, 19, 19, 20, 20, 20, 20, 22, 22, 22, 22, 23, 23, 23, 23, 24, 24, 24, 25, 25, 25, 25\}. \tag{3}$$

It is straightforward but tedious to re-calculate (3) based on Figure 1 and formula (2).

Note that the simultaneous credible band is a product of symmetric univariate credible intervals of the same level $(2j^*/n - 1) \cdot 100\%$. Besag et al. [2] also note that the method is slightly conservative in the sense that, for n fixed, the credible region (1) will typically contain slightly more that $100k/n\%$ of the empirical distribution because of ties in the set (2); this is


84 % simultaneous credible band

72 % simultaneous credible band

Figure 1: Illustration of the construction of simultaneous credible bands for simulated data with $n = 25$ and $p = 10$.

evident from out small example where the set (3) has many ties. This problem increases to an extent with p increasing, because the number of ties will then typically increase. However, the method is still consistent as $n \to \infty$. Empirical evidence shows that these credible bands tend to get rather unstable for credibility levels close to unity. In other words, the Monte Carlo will be quite large in these circumstances, but this problem can be easily attacked by taking a larger sample. However, the method requires the storage of all samples from all components of $\boldsymbol{\theta}$ which can be prohibitive is p and n is large.

Furthermore, the sorting and ranking of the samples from each component can be computationally intensive, if n is extremely large. However, in our experience, for $n = 10,000$ samples the method gives stable estimates at the usual credibility levels (95 and 99%) in just a few seconds.

Also note that ranking and sorting has to be done only once, even if simultaneous credible bands are required on more than one level. Only the set (2), the ordered samples $\theta_i^{[j]}$ and the ranks $r_i^{(j)}$ need to be available to calculate simultaneous credible bands at additional levels. The computational effort to calculate these additional simultaneous credible bands is negligible, compared to the initial ranking and sorting.

2.2 One-sided credible bounds

Besag et al. [2] note that "one-sided and other asymmetric bounds can be constructed analogously", but do not give further details. We will now look at this problem more closely. Clearly, the general idea of the approach described above can be easily applied to calculate, say, an upper confidence bound: Let j^* be the smallest integer such that the area defined by

$$(-\infty, \theta_i^{[j^*]}], \quad i = 1, \dots, p \tag{4}$$

contains at least k of the n values $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(n)}$. This procedure thus defines a one-sided upper credible bound of credibility level $100k/n\%$. The only question remaining is if there is also an analogous formula to (2). Indeed, j^* now simply equals the kth order statistic of the set

$$\left\{ \max_i r_i^{(j)}, j = 1, \dots, n \right\} \tag{5}$$

Similarly, a lower bound can be obtained by

$$[\theta_i^{[j^*]}, \infty), \quad i = 1, \dots, p \tag{6}$$

where j^* now equals the kth order statistic of the set

$$\left\{ \min_i r_i^{(j)}, j = 1, \dots, n \right\}. \tag{7}$$

A completely equivalent way to calculate a lower simultaneous credible bound is of course to compute the negative upper simultaneous credible bound of the negative samples.

Given the general applicability of the method proposed by Besag et al. [2] described above, it is surprising how rarely it has been used in practice. We will now describe an application taken from the area of low-dose risk estimation, where simultaneous credible bounds are useful.

3 Applications in low-dose risk estimation

Here we look at a specific problem in low-dose risk estimation, where the observed data $Y(x)$ are continuous, reflecting the *adverse* effect of some toxic exposure x. In other words, $Y(x)$ is expected to decrease with increasing x.

The data come from a study originally described in Chapman et al. [4], where x is a particular concentration of copper in and $Y(x)$ is the germination tube length of giant kelp, exposed to copper at dose x. There were up to five replicate observations for each of six copper concentrations between 0 and 180 μg/L.

Let $Y(x_i) = \mu(x_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, m$ are independent. We follow Piegorsch et al. [9] and assume a simple quadratic regression model $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. It may perhaps be useful to impose a further monotonicity constraint on the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ such that the function $\mu(x)$ is decreasing with increasing x. A weaker requirement is to assume that $\mu(x)$ is monotone at least within the observed range of x values. We will comment on such modifications in the discussion but use for the moment the unconstrained model.

A key quantity in risk assessment is the so-called *risk function* $R(x) = P(Y(x) \leq \mu(0) - \delta\sigma)$, where δ is a constant and typically chosen as $\delta = 2$ or $\delta = 3$. The idea is that a response, which is more than δ standard deviations below the control mean is considered as adverse, and $R(x)$ quantifies the probability of such an event as a function of the dose x. Furthermore, the *additional risk* is defined as $R_A(x) = R(x) - R(0)$, which becomes under the normal model

$$R_A(x) = \Phi(-(\beta_1 x + \beta_2 x^2)/\sigma) - \delta) - \Phi(-\delta), \tag{8}$$

where $\Phi(.)$ is the standard normal distribution function. Finally, a key concept in risk assessment is the notion of the *benchmark risk* and *benchmark dose*. This is often used to establish a low-dose level needed, the *benchmark dose* x_B, to generate a specific additional risk R_A, the benchmark risk $z \in (0, 1)$. Hence model (8) is inverted to find the *benchmark dose* x_B for a fixed benchmark risk z, i.e. solve $R_A(x_B) = z$ for $x_B(z)$.

Piegorsch et al. [9] develop sophisticated methodology to compute a frequentist simultaneous upper confidence bound for $R_A(x)$. The established function is then inverted based on equation (8) to obtain a simultaneous lower confidence bound for $x_B(z)$. Here we will devise an alternative Bayesian approach based on Monte Carlo sampling. For notational convenience, we set $\kappa = 1/\sigma^2$. A non-informative reference prior $p(\boldsymbol{\beta}, \kappa) \propto \kappa^{-1}$ is assumed for the unknown parameters (e.g. [3]) and hence the posterior distribution is of the usual normal-gamma form, known from standard linear model theory:

$$p(\boldsymbol{\beta}, \kappa | \boldsymbol{y}) = p(\kappa | \boldsymbol{y}) p(\boldsymbol{\beta} | \kappa, \boldsymbol{y}).$$

Here $p(\kappa|\boldsymbol{y})$ is gamma distributed with parameters $(m - p)/2$ and $s^2 \cdot (m - p)/2$ where $p = 3$ is the dimension of $\boldsymbol{\beta}$ and s^2 is the classical (unbiased) estimate of the variance σ^2. Furthermore, $p(\boldsymbol{\beta}|\kappa, \boldsymbol{y})$ is normal with mean equal to the least squares estimate $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and covariance matrix $\kappa^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. We can thus easily generate independent samples from this posterior distribution by first sampling $\kappa^{(i)}$ from $p(\kappa|\boldsymbol{y})$ and then sampling $\boldsymbol{\beta}^{(i)}$ from $p(\boldsymbol{\beta}|\kappa^{(i)}, \boldsymbol{y})$.

A Bayesian approach using Monte Carlo sampling has the advantage that samples from any function of the parameters can be obtained without any need for approximations, such as, for example, the Delta method. In the current context, $R_A(x)$ as defined in (8) is a simple function of the parameters β_1, β_2 and σ^2. Hence we are able to compute the posterior distribution of $R_A(x)$ for a range of values of x, say $x_1 < x_2 < \ldots < x_M$, and then compute simultaneous credible bounds for the parameters $R_A(x_1), R_A(x_2), \ldots, R_A(x_M)$. For illustration, Figure 2 displays the first $n = 100$ samples from the posterior distribution of $R_A(x)$ for $\delta = 3$.

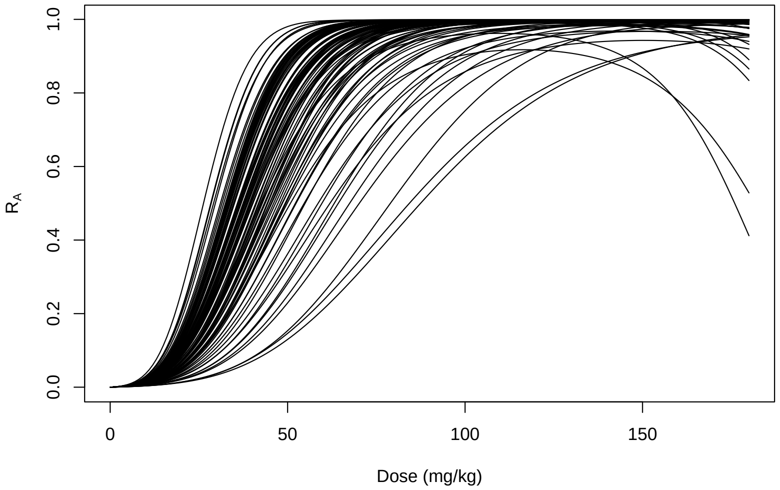


Figure 2: 100 samples from $R_A(x)$ for $x \in [0, 180]$, $\delta = 3$.

Figure 3 now displays the posterior median of $R_A(x)$, as well as the 95% simultaneous upper credible bound for $R_A(x)$, calculated using (4) and (5). Those have been obtained using $n = 10,000$ samples and 181 equally spaced values of $x \in \{0, 1, \ldots, 180\}$. For comparison, we also display the frequentist estimate of $R_A(x)$ as well as the corresponding 95% simultaneous upper confidence bound described in Piegorsch et al. [9].

Note that the Bayesian point estimates are slightly above the frequentist ones. A more pronounced difference can be seen for the simultaneous upper bound, which is again larger in the Bayesian approach.

Piegorsch et al. [9] go on to construct lower simultaneous credible bounds

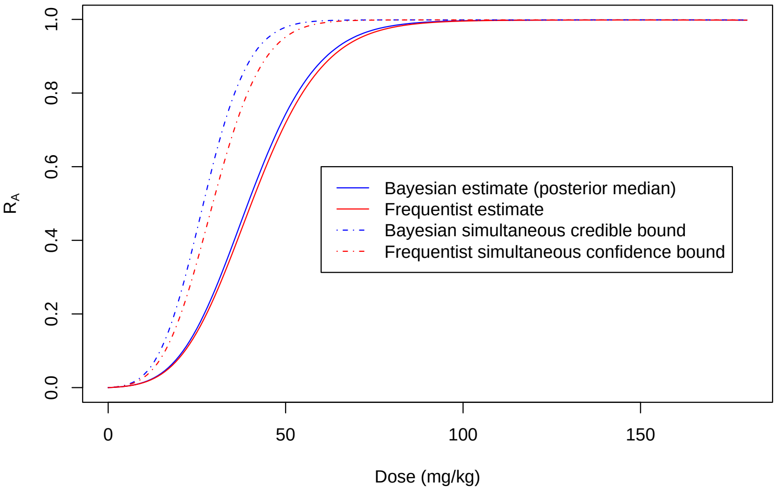


Figure 3: Estimated R_A function and simultaneous upper 95% credible bound, $\delta = 3$.

for $x_B = x_B(z)$ as a function of the benchmark dose $z \in (0, 1)$ by simply inverting the upper simultaneous bounds obtained for $R_A(x_B)$. Here we look at an alternative Bayesian sample-based solution.

Again, we view $x_B(z)$ as a function of the original parameters, i.e. for each benchmark risk z and each sample $\beta_1^{(j)}$, $\beta_2^{(j)}$ and $\sigma^{(j)}$, we solve equation (8) for $x_B^{(j)}$.The two solutions are

$$x_B(z) = \frac{1}{2\beta_2}\left(-\beta_1 \pm \sqrt{\beta_1^2 - 4C_l\beta_2\sigma}\right), \tag{9}$$

where $C_l = \delta + \Phi^{-1}(z + \Phi(-\delta))$. If there are two real positive solutions, we take the smaller value as $x_B^{(j)}$. Note that we have dropped the upper index $^{(j)}$ in (9) for simplicity. For each value of z, this defines samples from the posterior distribution of $x_B(z)$.

If $x_B(z)$ would be well-defined for every value of z and every sample of the parameters β_1, β_2 and σ, we could indeed just invert the simultaneous credible bound for $R_A(x_B)$ to obtain one for $x_B(z)$, just as in the frequentist case. However, there will not always be a positive real solution (9). Here it turns out that for $z = 0.99$, 11% of the samples do not have real solution. For smaller values of z, far less samples do not have a real solution; in fact less than 1% are "missing" for $z <= .83$. For illustration, consider Figure 5, which displays the first 100 samples from the posterior of $x(z)$.

Inverting the credible bound for $R_A(x_B)$ may hence induce bias, and we therefore consider a sample based, direct solution to construct a lower simultaneous bound for a series of benchmark doses z_l, $l = 1, \ldots, L$. We choose $L = 100$ equally-spaced values for z between 0 and 0.99. Note that the upper bound on z_L is $\Phi(\delta) < 1.0$, which in our case equals 0.9987.

We now have two options. First, we may delete all those samples, where $x^{(j)}(z_l)$ is missing for at least one benchmark risk $z_1, \ldots, z_L$ considered. Due to monotonicity properties this corresponds to deleting all those samples, where $x^{(j)}(z_L)$ is missing. However, this may induce bias in the estimation of simultaneous credible bounds, because larger values of BMD are typically missing for larger values of z. Alternatively, we may impute all missing values with their corresponding median, say, and proceed as if the sample would have been observed completely. The two different types of credible bounds are shown in Figure 5, denoted as "complete case" and "median imputed". There seems to be no substantial difference. Also displayed is the lower simultaneous confidence bound proposed by Piegorsch et al. [9], obtained by simply inverting the upper simultaneous confidence bound for $R(x)$.

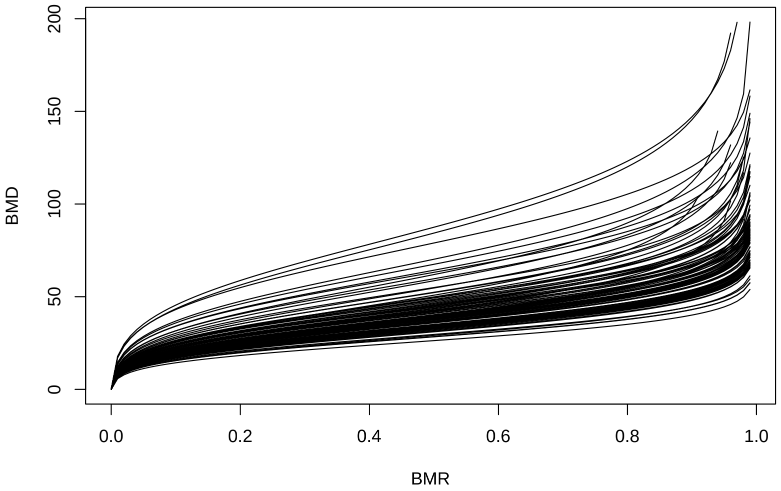


Figure 4: 100 samples of $x_B(z)$ for $\delta = 3$.

4 Discussion

The Bayesian approach to simultaneous inference in risk assessment has much to offer. It does not rely on approximations, is completely general and easy to implement. For example, it will be straightforward to calculate a Bayesian simultaneous credible bound in the application considered in Al-Saidy et al. [1], where the response variable is binomial.

We now close with two final comments. In the current application it

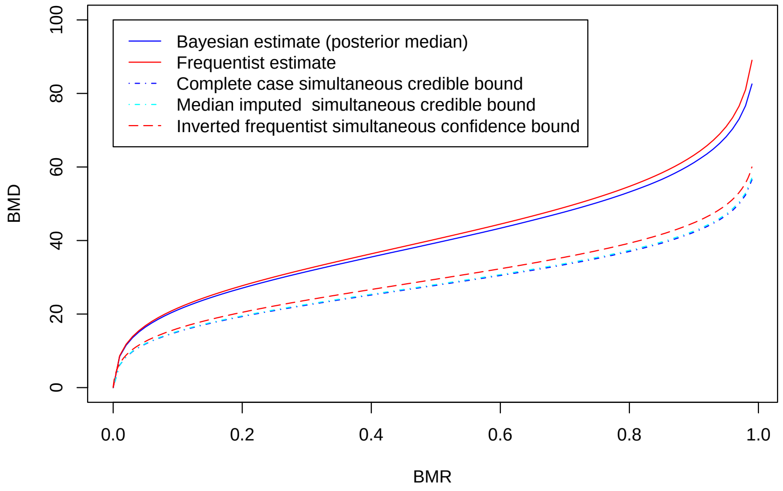


Figure 5: Estimated benchmark dose function and simultaneous lower 95% credible bound, $\delta = 3$.

turned out that a large number of samples did correspond to non-monotone dose-response relationships in the observed range of x. Certainly, the quadratic regression approach to the problem is open to question, and perhaps a monotone model, such as a logistic growth curve model or a nonparametric regression model under monotonicity constraints (e.g. [7] could have been useful. However, we should mention that we could have easily incorporated monotonicity constraints on $\boldsymbol{\beta}$ by simply ignoring all samples that do not fulfill the restriction imposed, see Gelfand, Smith and Lee [5] for more details in the context of Markov chain Monte Carlo simulation.

An alternative way to obtain simultaneous probability statements from Monte Carlo output is based on highest posterior density estimation, and has been described in Held [6]. This approach taken has the advantage that the simultaneous region does not need to be a hyper-rectangular, so is more realistic. Indeed, Held [6] has shown through examples, that simultaneous credible bands using the method described in Besag et al. [2] may include regions in the parameter space, which are not supported by the posterior at all. The difference between the two methods is related to the distinction between credible intervals based on quantiles and highest posterior density intervals in the one-dimensional case. The former intervals may include areas of low posterior density, for example if the posterior is bi-modal, whereas the latter will - by definition - only include regions of high posterior density.

However, the method by Held [6] can only be applied to calculate the posterior support for a series of *reference points*, but there is no easy way to visualize these credible regions in higher dimensions. In the current application there does not seem to be an obvious reference point for $R_A(x)$, say, so the method by Besag et al. [2] is the obvious choice for simultaneous Bayesian inference in risk assessment.

References

[1] Al-Saidy O.M., Piegorsch W.W., West R.W., Nitcheva D.K. (2004). *Confidence bands for low-dose risk estimation with quantal response data.* Biometrics, to appear. Available at `http://dostat.stat.sc.edu/bands`.

[2] Besag J.E., Green P.J., Higdon D.M., Mengersen K.L. (1995). *Bayesian computation and stochastic systems (with discussion).* Statist. Sci. **10**, 3 – 66.

[3] Box G.E.P., Tiao G.C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wiley. Reprinted by Wiley in 1992 in the Wiley Classics Library Edition.

[4] Chapman G.A., Denton D.L., Lazorchak J.M. (1995). *Short-term methods for estimating the chronic toxicity of effluents and receiving waters to West coast marine and estuarine organisms.* Technical Report EPA/600/R-95-136. U.S. Environmental Protection Agency, Cincinnati, Ohio.

[5] Gelfand A.E., Smith A.F.M., Lee T.M. (1992). *Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling.* Journal of the American Statistical Association **87**, 523 – 532.

[6] Held L. (2004). *Simultaneous posterior probability statements from Monte Carlo output.* Journal of Computational and Graphical Statistics **13**, 20 – 35.

[7] Holmes C.C., Heard N.A. (2003). *Generalized monotonic regression using random change points.* Statistics in Medicine **22**, 623 – 638.

[8] Pan W., Piegorsch W.W., West R.W. (2003). *Exact one-sided simultaneous confidence bands via Uusipaikka's method.* Annals of the Institute of Statistical Mathematics **55**, 243 – 250.

[9] Piegorsch W.W., West R.W., Pan W., Kodell R.L. (2004). *Low-dose risk estimation via simultaneous inferences.* Applied Statistics, to appear. Available at `http://dostat.stat.sc.edu/bands`.

Address: L. Held, Department of Statistics, University of Munich, Ludwigstrasse 33, 80539 Munich, Germany

E-mail: `held@stat.uni-muenchen.de`

INTERACTIVE BIPLOTS FOR VISUAL MODELLING

Heike Hofmann

Key words: Data visualisation, biplots, univariate linear models, category level points, biplot axis, visual modelling.

COMPSTAT 2004 section: Data visualisation.

Abstract: The link between statistical models and visualisation techniques is not very well explored, even though strong connections do exist. This paper describes how biplots – interactive biplots in particular - can be used for visual modelling. By slightly adjusting the way biplots are constructed they provide the means to display linear models. The goodness of fit of a particular model becomes instantly visible. This makes them a useful addition to the standard set of visualization tools for linear models.

Biplots show predicted values and residuals. This helps, firstly, to assess a model far beyond the mere statistics and to detect structural defects in it. Secondly, biplots provide a link between the modelling statistics and the original data. Additional interactive methods such as hotselection also allow the analysis of outlier effects and behaviour.

1 Introduction

Biplots are a very promising tool for visualising high-dimensional data, which include both continuous and categorical variables. The strategy of biplots is to choose a linear subspace (usually a 2-dimensional space - in order to be able to plot the result using standard techniques), which is in some respect optimal, and project the high-dimensional data onto this space. One criterion for optimality is, for instance, to minimise the discrepancy between the high- and the two dimensional representations of the data. Biplots show only one projection out of infinitely many. They therefore cannot be exact representations of the data but only approximations.

What gave the Biplots their prefix "Bi-" ($\beta\iota$ is the greek syllable for "two") is the simultaneous representation of both data points and original axes within the projection space.

The biplot axis of a continuous variable is represented by a straight line (in case of linear models, to which we will restrict ourselves) with unit points marked by small perpendicular lines. One unit of a variable X_i corresponds to one times the standard deviation of X_i. If the data matrix X is centered and standardized, these units are therefore directly comparable for all i, and the length of a unit vector gives a measure for how well a variable is represented in the chosen projection plane.

Instead of continuous axes, so called *category level points (CLPs)* are used to display a categorical variable X. Using a binary dummy variable for each of the categories of X, an imaginary axis is found as in the continuous case. A CLP is given as the 1-unit point of this axis. Each CLP therefore represents one category of X.

The different gray shades of the points in the figure is the effect of a crude graphical density estimate - light areas in the display correspond to a high number of observations.

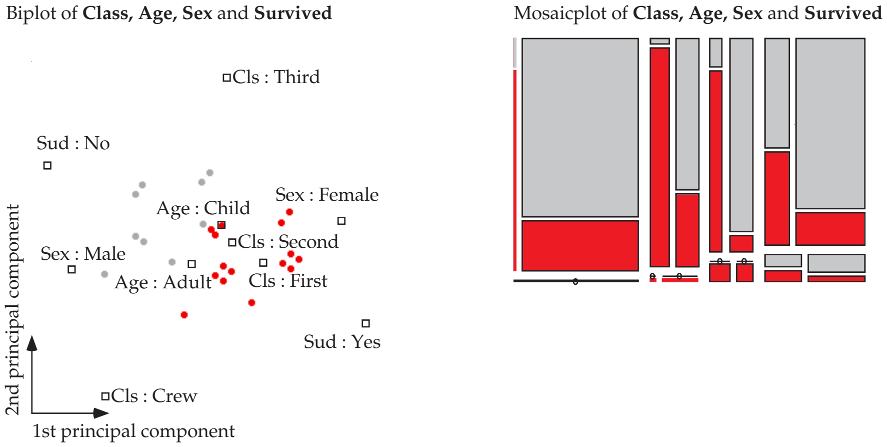


Figure 1: Biplot and corresponding mosaicplot of the Titanic Data [3]. Each dot on the left side corresponds to a cell on the right hand side. Highlighted are survivors.

Figure 1 shows a biplot of categorical variables, based on a Multiple Correspondence Analysis (MCA). Next to the biplot a mosaicplot of the same variables is drawn.

Biplots were first introduced by [5]. A recent monograph on biplots by [6] summarises different types of biplots and embeds various models in the concept. Possibilities for interactive extensions have been examined in [7].

Biplot representation

The graphical representation of a biplot is dot based. This means for categorical variables, that each combination is shown as one single dot. Of course, this does not allow conclusions about this combination's size any more. One solution to this problem is the use of density estimates. This also covers the problem of over-plotting, which, especially in large data sets, is always present in dot based representations.

The graphical representation of a biplot has two components:

- Data points are projected onto the plane spanned by the first two principal components and visualised as dots. The center of the plot is given by the projection of the p dimensional mean $(\frac{1}{n}X_1^t\mathbb{1}, \ldots, \frac{1}{n}X_p^t\mathbb{1})$.

- The unit vectors e_i^t corresponding to the (dummy) variables are also projected onto this plane.
  The graphical representation differs for continuous and categorical variables: For continuous variables, an arrow is drawn from plot center to the projection of the variable, which marks the direction of the original variables. These directions are called the *biplot axes*. The arrowheads mark the unit points on the biplot axes.
  For a categorical variable its projection on the biplot is marked by a square rectangle, the CLPs.

"Reading" a biplot

In a biplot the most important source of information is the distance between objects. The distance gives a measure of how similar or how closely related objects are.

The distance of a CLP to the plot's center (in the middle of the plot) or the length of a unit on a biplot axis reflect how good the projection of the underlying variable is, i.e. with increasing distance the goodness of fit - and with it the "importance" - of this variable increases.

The meaning of objects lying close to each other varies according to their type:

- **point - point**: close points reflect high dimensional "neighbours".
- **axis - axis**: axes with a small angle between them indicate a high positive correlation between the variables, angles near 180° indicate a high negative correlation.
- **CLP - CLP**: Neighbouring CLPs are a hint that the corresponding variables are associated, i.e. that these categories frequently occur together in the data.
- **points - axis/CLPs**: the data values for a point are found by orthogonal projection onto an axis. The axes closest to a point therefore represent the strongest influence for a data point. Accordingly, points are assigned to those categories with the closest lying CLPs. In doing so, one has to remember, that a biplot of more than just two variables cannot be anything but an approximation.

2 Interactive methods

Based on the construction and interpretation of a biplot, interactive methods have to be provided for in the display to facilitate interpretability and ease of use.

2.1 Interactive querying

Interactive querying is context sensitive - querying different objects provides different information. Examples for several querying results are given in figures 2 to 4.

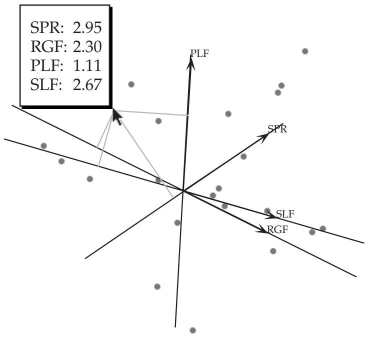


Figure 2: Querying a point or "empty space" of the plot results in drawing perpendicular lines onto the biplot axes. Estimated values of the variables are given for the point in the projection plane.

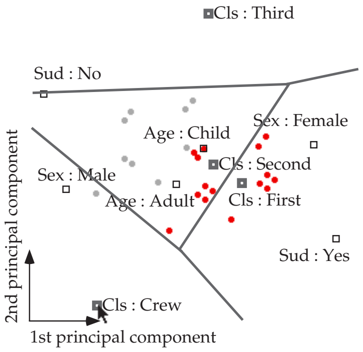


Figure 3: Querying a CLP highlights the other CLPs and the prediction regions of the underlying categorical variable.

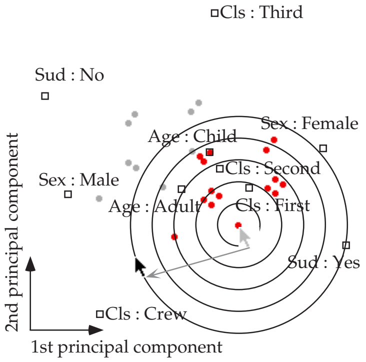


Figure 4: Drag-query: dragging from one point of the plot to another draws circles around the starting point as visual aid for estimating distances between objects.

Figure 3 shows the prediction regions corresponding to the variable'Class'. All categories corresponding to a single variable divide the biplot area in a set of mutually exclusive prediction regions. The prediction region of a CLP is defined as the space closest to the CLP, i.e. no other CLP is closer. From the prediction regions in figure3 it becomes obvious that the representation from the MCA does not fit well: almost all dots are predicted to be second class passengers - there are no combinations predicted as third class passengers.

2.2 Logical zooming and hotselection

The difference between logical and "normal" zooming lies in the fact that by logical zooming an object is not only enlarged but more details appear. Logical zooming in biplots has two main applications: logical zooming in large data sets gives a tool to drill down the data set into smaller parts, which are - hopefully - more homogeneous and therefore easier to analyze.

Another advantage of logical zooming is its possibility of excluding outliers. By focussing on the “main” part, i.e. not regarding outliers, their influence on the model becomes apparent. This is particularly useful for models with a poor behaviour with respect to outliers. If in fact the effect outliers have on a model is of fore-most interest, we will want to use hotselection [8] instead of logical zooming. The boundary between these two tools is fluent - but essentially, the concept of hotselection is less permanent than logical zooming : changes are more readily made and taken back again. In the setting of modelling, hotselection is used to compute a new model based on highlighted values only.

Figure 5 shows a biplot of a correspondence analysis taking all of the descriptive variables into account. Several clearly distinguished groups appear in the plane spanned by the first and second principal component axis. Highlighting shows poisonous mushrooms. These clusters are marked by numbers in the graphic. Using a Mosaicplot of all the descriptive variables, we want to find descriptions (as short as possible) for these groups. The following table gives a short summary of our results:

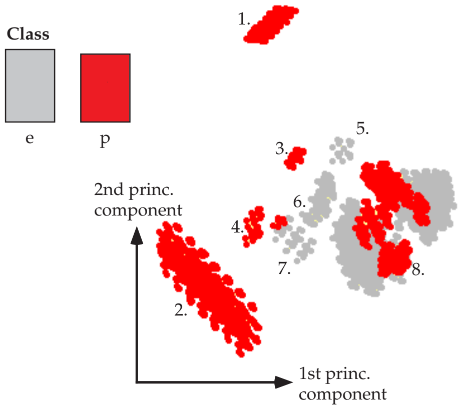


Figure 5: Biplot of an MCA of all of the mushrooms variables. Highlighted are poisonous mushrooms. Some distinct clusters appear (marked by the numbers).

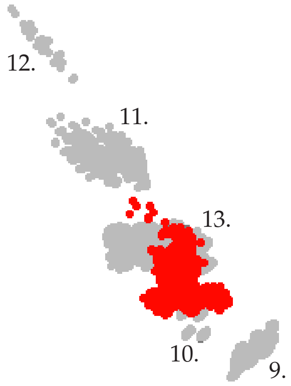


Figure 6: Zoom into group 8.

Zooming is equivalent to hierarchical clustering via MCA. The eight poisonous mushrooms in cluster 7 all have stalk color y setting them off from the rest. Figure 6 shows a zoom into the largest cluster, cluster 8. Several more groups show up in the projection plane. Clusters 9 to 12 consist of edible mushrooms only. For all of these clusters simple descriptions among the explanatory variables exist. Cluster 10 e.g. consists of mushrooms with stalk color o. All of the descriptions are only valid for the zoomed data (i.e. only in combination with all of the description for cluster 8 above). Cluster 13, consisting of 2512 mushrooms, is the only one which needs further inspection - using further logical zooming. After two more steps all poisonous mushrooms can be separated from the edible ones.

Group	Count	Class	Description
1	1296	pois.	ring type = I
2	1728	pois.	gill type: b
3	36	pois.	stalk color above ring: c (or ring type: n)
4	32	pois.	stalk surface below ring:y population: v
5	16	edible	stalk surface above ring:y stalk color above ring:n
6	120	edible	ringtype = f or ringtype = p, stalk surface:k (below and above the ring)
7	56	mixed	48 edible
8	4826	mixed	4010 edible

3 Univariate linear models with continuous response

Based on the graphical representation of a biplot and its interactive features, we will try another approach to visualise linear models among the data. The biplot representation provides a possibility to draw conclusions from a linear model in such a way, that the goodness of fit as well as the most important explanatory variables become instantly visible with one biplot representation.

Let us assume a situation, where we are dealing with a continuous response variable Y and several independent variables $X_1, \ldots, X_p$. The X_i do not necessarily have to be continuous - but we also do not work with categorical variables directly. Instead, for a categorical variable X_i a set of binary *dummy variables* is used as explained before.

Let $X_1, \ldots, X_p$ be a set of independent variables, which has been produced in this way, i.e. a variable is either continuous by default or it is a variable corresponding to a single category.

A *linear regression model* then has the form

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

where $X = (\mathbb{1}, X_1, \ldots, X_p)$ is the *design matrix* and β the *vector of parameters* β_i.

If some of the variables are dummy variables, we have to use a further condition for the parameters of these variables in order to get a unique result. Let $Z_1, \ldots, Z_I$ be the dummy variables for the categorical variable X and $\beta_1, \ldots, \beta_I$ the corresponding parameters of the linear model, then a commonly used constraint (*null-sum-coding*) on the estimates for these parameters is that they sum to zero, i.e.

$$\sum_{i=1}^{I} \hat{\beta}_i = 0,$$

or one of the categories is used as basis and the parameters of the resulting model show the influence a category has with respect to the basis. The constraint *effect-coding* on the parameters then is

$$\beta_i = 0,$$

if Z_i is the dummy variable corresponding to the basis category.

It is well known, that the *hat matrix* $H := X(X'X)^{-1}X'$ is that projection matrix, which minimizes the least squares problem of $\sum_i \epsilon_i^2$ and gives the predicted values $\hat{Y}$ as

$$\hat{Y} = HY.$$

Accordingly, the LS-estimator for β is

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

One of the favourite methods for looking for structure among the residuals $\epsilon_i = Y_i - \hat{Y}_i$ is to plot residuals versus their predicted values, i.e. the data points are projected into the plane spanned by $\hat{Y}$ and $Y - \hat{Y}$. These vectors are indeed orthogonal to each other, since the scalar product vanishes:

$$\begin{aligned}(\hat{Y}, Y - \hat{Y}) &= (HY, Y - HY) = \\ &= (Y, H'(HY - Y)) \overset{H'=H;H^2=H}{=} (Y, HY - HY) = 0.\end{aligned}$$

3.1 Finding the biplot axes and comparing effects

The biplot axes are constructed in this situation in the same way as for standard biplots of PCA or MCA. By projecting the data into the plane spanned by $\hat{Y}$ and $Y - \hat{Y}$ we get $\hat{\beta}_i$ as the coordinate of $e_i' = (0, \dots, 1, \dots, 0) \in \mathbf{R}^p$ in the direction of $\hat{Y}$.

However, while projecting e_i' in the direction of $Y - \hat{Y}$ a problem appears: generally we do not have a value along the Y-axis for any given value of X, particularly for e_i'. A direct calculation therefore is not possible. But we do know that the whole data space X is orthogonal to the direction of the residuals $Y - \hat{Y}$, since

$$X' \cdot (Y - \hat{Y}) = X'Y - \underbrace{X' \cdot X(X'X)^{-1}}_{=I} X'Y = X'Y - X'Y = 0.$$

Therefore the coordinate of the ith biplot axis in the direction of the residuals is also zero.

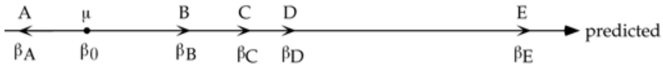


Figure 7: Axis of predicted values together with the five biplot axes for variables A, B, C, D and E.

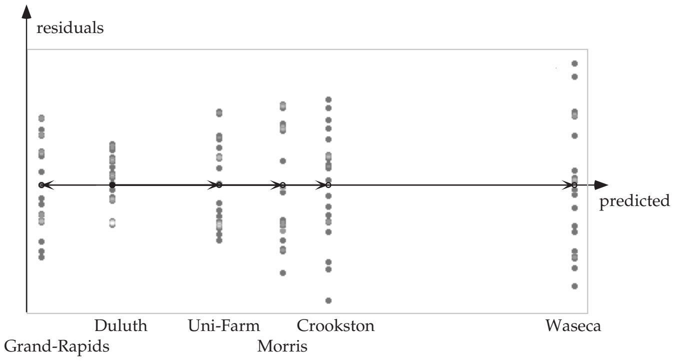
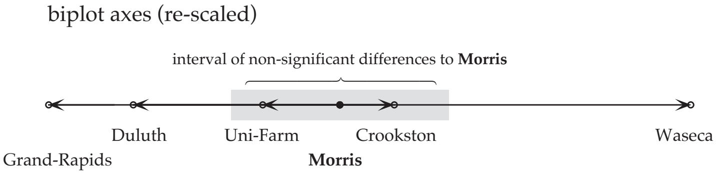


Figure 8: Analysis of variance of the Barley Data. Predicted values are plotted vs. residuals. Six different sites of barley cultivation are drawn as biplot axes. The results of Duluth are used as basis values in the Anova.

Figure 7 shows the vector of the predicted values $\hat{Y}$ together with biplot axes for five variables A, B, C, D and E. We can re-establish the relation of projected data points and their original values by orthogonal projections of the points onto the biplot axes. In the case of an *analysis of variance* this means, that we get very informative "labels" for the predicted values. Figure 8 shows an analysis of variance of the *Barley Data* [4].

We see not only parallel dot plots of the barley yields, but also a natural ordering of the six categories, even (roughly) their distance or closeness. The last point has a caveat: the lengths of their units are not directly comparable, i.e. an axis with large units is not by default a more important factor, since the "importance" of an axis also depends on the variability of $\hat{\beta}_i$. The standard test of judging, whether the ith parameter is significantly different from 0, i.e. $\beta_i = 0$ vs $\beta_i \neq 0$, uses the estimate's variability. The test statistic $\hat{\beta}_i / SE_{\hat{\beta}_i}$, where $SE^2_{\hat{\beta}_i} = \hat{\sigma}^2 e_i'(X'X)^{-1} e_i$, is approximately t distributed with $n - p - 1$ degrees of freedom.

A second choice of units on the biplot axes therefore is the term $\hat{\beta}_i / SE_{\hat{\beta}_i}$. This re-scales the biplot axes in a way that their lengths are proportional to the values of the t-statistic. More important variables in the regression model now have larger parameters, whereas biplot axes with insignificant parameters remain short. Graphically we can support this by highlighting an interval on the axis of predicted values, which corresponds to the 5% level of a t-test. See figure 9: in this example the SE_{β_i} are of the same order of magnitude, and the distances do not change compared to figure 8.



	Difference	std. err.	Prob
Morris - Crookston	-2.02000	2.309	0.978724
Morris - Duluth	7.40333	2.309	0.075985
Morris - Grand-Rapids	10.4683	2.309	0.001805
University-Farm - Morris	-2.73333	2.309	0.923061
Waseca - Morris	12.7083	2.309	0.000052

Figure 9: Comparison of effects: on the top the graphical test via the interval of non-significant values is shown, on the bottom is a table of the corresponding pairwise tests.

When setting the origin of this interval the exact coding, which we used for a categorical variable is important: if we use effect-coding, the origin of the 5% interval will be placed on the predicted value of the basis. When using a null-sum-coding the origin of the interval is set to the expected value of Y.

Figure 9 shows the (re-scaled) biplot axes of the example above. The category *Morris* is set as basis value. Around this value the interval of non-significant values is shown as a gray-shaded rectangle. The categories *Uni-Farm* and *Crookston* fall into this rectangle, indicating that these categories have parameters, which are not significantly different from the parameter for *Morris*.

Since the differences between the parameters are not affected by the choice of the coding, we may use these differences for more than one comparison (and with that, multiple tests) in each plot. From a statistical point of view this multiple test situation suggests the use of *Bonferroni-confidence intervals* for each parameter rather than the use of the above significance intervals. The difference between the above intervals and Bonferroni's intervals is essentially a factor, calculated from the level of significance and the number of comparisons made.

The price we have to pay for the re-scaling of the biplot axes with the parameter's variability is that we lose the quantitative connection between data points and biplot axes.

In order to avoid re-scaling we may try another approach to visualise the tests between the effects: the software JMP suggests the use of circles of different size around the parameter values. The size of each circle is given by the standard deviation of the parameter times $t_{\alpha/2}$. Whether two parameters are significantly different is decided by the angle: if the angle at the intersection of their circles is less than 90° the two values are not significantly different, otherwise they are (see figure 10). For a more detailed explanation of the underlying statistics see JMP's "Statistics and Graphics Guide", p.94-95.

The disadvantage of this approach is that angles have to be compared. This makes the decision between significant and not-significant differences between the parameters rather difficult visually.

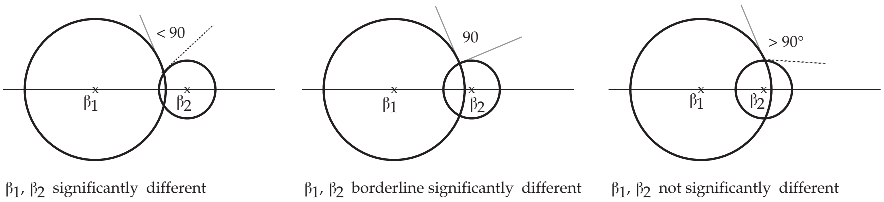

< 90
β1
β2
β1, β2 significantly different
90
β1
β2
β1, β2 borderline significantly different
> 90°
β1
β2
β1, β2 not significantly different

Figure 10: Confidence circles around parameter values. Depending on the angles at the circles' intersections the difference between the parameters is significantly different (left), borderline significantly different (middle) and not significantly different (right).

3.2 Projection of the response variable Y

Since we may write Y as the sum of the projection axes $\hat{Y}$ and $Y - \hat{Y}$,

$$Y = 1 \cdot \hat{Y} + 1 \cdot (Y - \hat{Y}),$$

Y has the coordinates (1,1) in the new coordinate system.

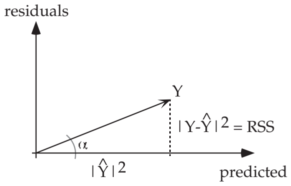


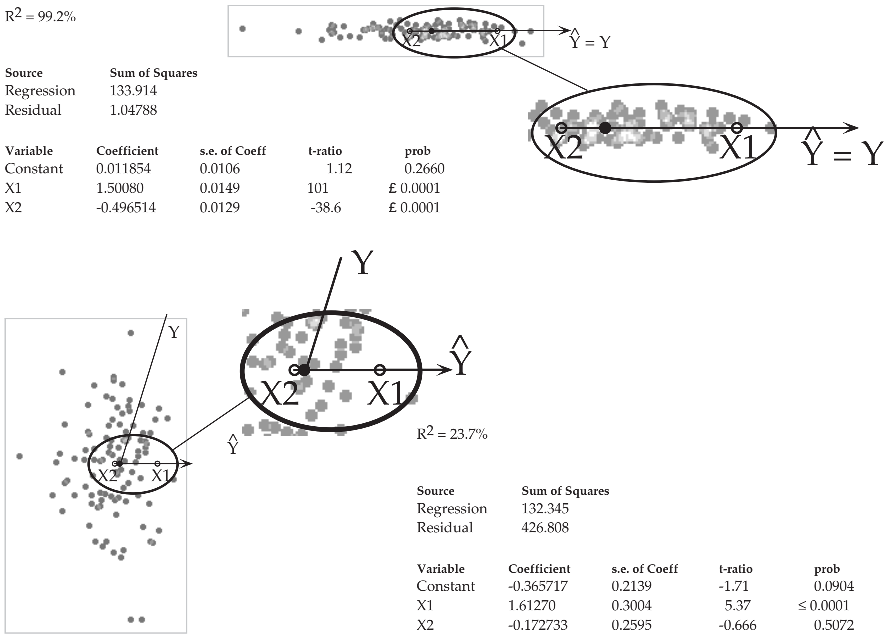
Figure 11: Response variable Y in the projection plane spanned by predicted and residual values.

The units on the projection axes are given as $|Y - \hat{Y}|$ and $|\hat{Y}|$, where $|Y - \hat{Y}|^2 = \sum_i (Y_i - \hat{Y}_i)'(Y_i - \hat{Y}_i) = RSS$ and $|\hat{Y}|^2 = TSS - RSS$. RSS is the *residual sum of squares* and TSS is the *total sum of squares*.

The coordinate of Y in direction of $Y - \hat{Y}$ shows the square root of the residual sum of squares, $\sqrt{RSS}$; the coordinate in direction of $\hat{Y}$ gives the square root of the difference between the total sum of squares, TSS, and the residual sum. The angle α between Y and $Y - \hat{Y}$ is therefore related to the goodness of fit statistics R^2 of the regression model:

$$\cos^2(\alpha) = \left(\frac{|HY|}{|Y|}\right)^2 = \frac{TSS - RSS}{TSS} = R^2,$$

i.e. the smaller α is, the better is the fit of the regression model. Of course, the angle depends on the aspect ratio of the display. By fixing the aspect ratio to 1, different plots (and thereby different models) can be compared: a plot with large width and little height indicates a good fit (the residuals are small with respect to the predicted values), while a quadratic plot or, even worse, a tall and thin plot indicates a very bad fit, see figure 12.



R^2 = 99.2%

Source	Sum of Squares
Regression	133.914
Residual	1.04788

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	0.011854	0.0106	1.12	0.2660
X1	1.50080	0.0149	101	£ 0.0001
X2	-0.496514	0.0129	-38.6	£ 0.0001

R^2 = 23.7%

Source	Sum of Squares
Regression	132.345
Residual	426.808

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	-0.365717	0.2139	-1.71	0.0904
X1	1.61270	0.3004	5.37	≤ 0.0001
X2	-0.172733	0.2595	-0.666	0.5072

Figure 12: Example of regressions with good fit (above) and bad fit (below). The goodness of fit is emphasized by the shape of the display. The angle between Y and $\hat{Y}$ also corresponds to R^2.

4 Conclusions

Biplots can be used to visualize univariate linear models. They allow, at the same time, an assessment of the model's goodness of fit. Additional interactive methods such as interactive querying provide the analytic goodness of fit statistics, too. This allows a tight link of visual display and the corresponding model. Another interactive method, hotselection, gives a way of examining the influence of single points or group of points on the model, which can be used as a very efficient way of outlier spotting.

In the paper only one-dimensional models are shown - this is just for illustration purposes. The approach itself is, of course, not limited to one dimension.

If using scatterplots for a biplot representation, biplots are restricted to a 2d display - with graphics that allow display of higher dimensionality such as a tour ([1], [2]) for example, more precise displays are possible. In a tour the described approach would mean to fix the x-axis artificially to $Y - \hat{Y}$ (equivalent to fixing Y to be fully included while touring the data) and to tour through the X space. This also allows to deal with higher-dimensional Y.

References

[1] Asimov D. (1985). *The grand tour: a tool for viewing multidimensional data.* SIAM J. Sci. Stat. Comput. **6**, 128 – 143.

[2] Buja A., Swayne D., Cook D. (1996). *Interactive high-dimensional data visualization.* Journal of Computational and Graphical Statistics **5**, 78 – 99.

[3] Dawson R.J.M. (1995). *The "unusual episode" data revisited.* Journal of Statistics Education **3**.

[4] Fisher R. (1935). *The design of experiments.* Edinburgh UK: Oliver and Boyd.

[5] Gabriel K. (1971). *The biplot graphic display of matrices with application to principal component analysis.* Biometrika **58**, 453 – 467.

[6] Gower J.C., Hand D.J. (1996). *Biplots.* London: Chapman and Hall Ltd.

[7] Hofmann H. (1998). *Interactive biplots.* In New Techniques & Technologies for Statistics (NTTS) 98, Sorrento, Italy: Eurostat, 127 – 136.

[8] Velleman P. (1995). *Data Desk 5.0, Data Description.* Ithaka, New York.

Address: Heike Hofmann, Department of Statistics, Iowa State University, Ames IA, USA

E-mail: hofmann@iastate.edu

R: THE NEXT GENERATION

Kurt Hornik

Key words: R, CRAN, S, S-PLUS, Sweave, LaTeX.

COMPSTAT 2004 section: Statistical software.

Abstract: Version 2.0 of R will be released in the course of 2004. Following the 1.0 release on 2000-02-29, the advent of this "next generation" of R mostly indicates the view of the R developers that R has now moved substantially beyond being a reference implementation of S. In this paper, we look at several of these key enhancements. We start with a review of some key facts on "R and S". Sections 2 to 5 then describe the name space mechanism, the new grid graphics system, the packaging system, and Sweave, a tool which allows to embed R code for data analysis into LaTeX documents.

1 Introduction

S is a very high level language and an environment for data analysis and graphics which has been developed at Bell Laboratories for about 30 years. In 1998, the Association for Computing Machinery (ACM) presented its Software System Award to John M. Chambers, the principal designer of S, for *"the S system, which has forever altered the way people analyze, visualize, and manipulate data . . . "*. The evolution of the S language is characterized by four books by John Chambers and coauthors, which are also the primary references for S. The "Brown Book" [1] is of historical interest only. The "Blue Book" [2] describes the "New S" language. The "White Book" [5] documents a concerted effort to add functionality to facilitate statistical modeling in S, introducing data structures such as factors, time series, and data frames, a formula notation for compactly expressing linear and generalized linear models, and a simple system for object-oriented programming in S allowing users to define their own classes and methods. Together with the Blue Book, it describes S version 3 ("S3"). [4], the "Green Book", introduces version 4 of S ("S4"), a major revision of S designed by John Chambers to improve its usefulness at every stage of the programming process, introducing in particular a new "formal" OOP system supporting multiple dispatch and multiple inheritance, and a unified input/output model via "connections". Today, a commercial implementation of the S language called "S-PLUS" is available from Insightful Corporation (`http://www.insightful.com`).

What is now the R project started in 1992 in in Auckland, New Zealand, as an experiment by Ross Ihaka and Robert Gentleman *"in trying to use the methods of LISP implementors to build a small testbed which could be used to trial some ideas on how a statistical environment might be built"* [8]. The decision to use an S-like syntax for this statistical environment, being motivated by both familiarity with S and the observation that the parse trees generated by S and LISP are essentially identical, resulted in a system "not unlike S". In fact, basing the R evaluation model on Scheme (a member of the LISP family) has given R *lexical scoping* as the most prominent difference between R and other implementation of the S language [7]. Since mid-1997 there has been a core group (the "R Core Team") who can modify the R source code CVS archive. The group currently consists of Doug Bates, John Chambers, Peter Dalgaard, Robert Gentleman, Kurt Hornik, Stefano Iacus, Ross Ihaka, Friedrich Leisch, Thomas Lumley, Martin Maechler, Duncan Murdoch, Paul Murrell, Martyn Plummer, Brian Ripley, Duncan Temple Lang, and Luke Tierney. R version 1.0, released on 2000-02-29, provided an implementation of S version 3. The key innovations in S4 were introduced in 1.x series releases (connections in 1.3, a first implementation of the S4 OOP system in version 1.4).

An R distribution provides a run-time environment with graphics, a debugger, access to to certain system functions, and the ability to run programs stored in script files, and contains functionality for a large number of statistical procedures. This "base system" is highly extensible through so-called *packages* (see Section 4) which can contain R code and corresponding documentation, data sets, code to be compiled and dynamically loaded, and so on. In fact, the R distribution itself provides its functionality via "base" packages such as **base**, **stats**, **grid**, and **methods**. The data analytic techniques described in such popular books as [23], [16], or [21] have corresponding R packages (**MASS**, **nlme**, and **survival**). In addition, there are packages for bootstrapping, various state-of-the-art machine learning techniques, and spatial statistics including interactions with GIS. Other packages facilitate interaction with most commonly used relational databases, importing data from other statistical software, and dealing with XML. Currently, more than 300 packages are available via the *Comprehensive R Archive Network* (CRAN, `http://CRAN.R-project.org`), a collection of sites which carry identical material, consisting of the R distribution(s), contributed extensions, documentation for R, and binaries.

It is important to realize that the "R Project" is really a multi-tiered large scale software development effort, with the R Core Team delivering the basic distribution which mostly provides the computational infrastructure on which others can build special-purpose data analysis solutions. In this paper, we discuss four of the key additions to this infrastructure relative to the S reference standard.

2 Name spaces

Name spaces allow package authors to control how global variables in their code are resolved. To see why this is important, suppose that package **foo** defines the function

```
  mydnorm <- function(x) 1 / sqrt(2 * pi) * exp(- x^2 / 2)
```

and has been attached to the search path so that evaluating the expression `mydnorm(0)` uses the above function when looking up a value for the symbol '`mydnorm`'. Now suppose that the user enters

```
pi <- 1
```

at the prompt, so that the symbol '`pi`' is bound to the value 1 in the R workspace ("global environment", `.GlobalEnv`). With the "usual" dynamic lookup mechanism for bindings (of symbols to values) in place, going through the collections of bindings represented by the search path and starting with the global environment, evaluating `mydnorm(0)` would not give the result that the **foo** package author had intended—namely, using the value bound to `pi` in the **base** package. More generally, top level assignments as well as attaching packages to the search path can insert shadowing definitions ahead of the ones intended. Name spaces ensure that this does not happen.

In the above example, all global variables were intended to refer to the definitions provided by the **base** package, which is always attached (at the end of the search path). Suppose that **foo** wanted to make use of functionality provided by another package **bar** which is not necessarily always attached. Traditionally, the package author would then arrange **bar** to be attached at some point. This is not only subject to shadowing as described above, but also has the effect of forcing a possibly undesired change to the search path onto the user. Using name spaces, one can *import* the required functionality (more precisely, exported variables) from other packages. Such imports then cause the other packages to be *loaded* if necessary, without attaching them.

Finally, name spaces also allow the package author to control which definitions provided by a package are visible to a package user and which ones are private and only available for internal use. By default, a definition is private; it is made public by an explicit *export* of the name of the defined variable. Similar to proving mathematical theorems, good programming practice for a high-level language such as R typically suggests providing functionality based on small building blocks which perform simple tasks and are readily comprehended. If all these blocks correspond to functions with a few lines of code, and all these functions are visible to users, these will find determining the key functionality provided by a package rather challenging. (Thus far, coding practices suggested using names starting with a '`.`' for "internal" variables, based on the fact that listing variable names in elements of the search path by default excludes names with a leading dot. This reduces clutter, but does not prevent shadowing.)

A package is given a name space by placing a NAMESPACE file containing name space directives into the top level source directory of the package. This mechanism makes it possible to obtain the information on the package code interface as part of the package meta-data, without the need of processing the package code. The main directives control export and import of variables, and superficially resemble R function calls, with the arguments being syntactic names or string constants (i.e., quoting is only necessary for non-standard names). For example, the directive

```
export(mydnorm, "[<-.myclass")
```

exports two variables. Import directives are used to import definitions from other packages with name spaces. The directive

```
import("survival")
```

imports all exported variables from package **survival**;

```
importFrom("survival", "Surv")
```

would only import its `Surv()` function. There is also a `useDynLib` directive for specifying that external code compiled into a DLL is to be loaded when the package is loaded.

As syntactic sugar, variables exported by a package with a name space can also be referenced using fully qualified references which are obtained by concatenating the package and variable name, separated by a double colon (e.g., `foo::mydnorm` in the above example). This is less efficient than a formal import and also loses the advantage of separating the dependency meta-data from the package code, so this approach is usually not recommended.

Name spaces are *sealed*. This means that once a package with a name space is loaded, one can no longer change the bindings (add or remove variables, or change the values). If it is necessary to record state information on the package level, one can use *dynamic* variables (functions allowing to get and set state information maintained in their environment). Sealing ensures that the bindings cannot be changed at run time, which has been instrumental to the development of a byte code compiler for R.

R supports both the S3 and S4 paradigms for object oriented programming. In the former, there are no "formal" data structures representing the class information, and method dispatch is based on a naming convention (methods are functions the name of which is obtained by concatenating the names of the generic and the class of the argument on which dispatch is based, separated by a period). With the advent of name spaces, this creates a problem: if a package is imported (hence loaded) but not attached to the search path, the S3 method it provides may not be found for dispatch. The name space mechanism therefore also provides facilities for registering S3 methods for dispatch. The directive

```
S3method("print", "foo")
```

registers the function `print.foo` defined in the package as the S3 method for generic `print` and class `"foo"`. (This mechanism in fact pertains to the cases where the generic is defined in a package with a name space. In this case, S3 methods only need to be registered, but not exported.) The "formal" S4 OOP paradigm provides classes and generics with more structure than their S3 counterparts, and hence conceptually allows better integration with name spaces. S4 classes are private by default; they can be made public using the `exportClasses` directive. As of writing this article, all generics for which formal methods are defined need to be declared in an `exportMethods` directive, and where the generics are formed by taking over existing functions, those functions need to be imported (explicitly unless they are defined in the base name space). These mechanisms may be different in R 2.0; the current development efforts will most likely bring the mechanisms in R more in line with those in "related" functional languages in the LISP family which provide both name spaces and a "formal" OOP system (such as Common Lisp or Dylan).

By giving package developers the tools to control the package code interface and the resolution of global variables in their code, name spaces substantially enhance the potential of R for dealing with complex data analysis tasks based on combinations of "many" extension packages, in particular providing a way of resolving conflicts among definitions in these.

3 Grid graphics

Traditional S graphics ("base graphics" in R, although now provided by package **graphics**) divides pages of graphics output into outer margins and possibly several figure regions which in turn each consist of figure margins and plot regions. This places severe limitations on the possibilities for accessing the whole graphics page, e.g. when annotating a high-level plot. (The standard example is that one cannot have arbitrarily rotated text in axis labels, as `text()` supports arbitrary rotation but can only draw inside the plot region, whereas `mtext()` can only write horizontally or vertically.) Each region has one or more coordinate systems associated with it, as controlled via "graphical parameters" (`par()`).

Grid graphics is an alternative graphics engine provided by package **grid** in the R distribution. One of its goals is to remove some of the inconvenient constraints imposed by the base graphics system. In addition, it aims at the development of functions to produce high-level graphical components which would not be very easy to produce using traditional S graphics (such as Trellis graphics [3], [6], where the more natural building block is a "panel" which consists of a plot plus one or more "strips" around it), and the rapid development of new graphics ideas. It serves these aims by providing functionality for the production of low-level to medium-level graphical components, such as lines, rectangles, data symbols, and axes, and sophisticated support for arranging graphical components. Grid does not provide high-level graphical components such as scatterplots of barplots, and hence is primarily targeted at graphics developers rather than "users", with the usual remark that in S, there is at most a gradual transition between these groups, if no such distinction at all.

In grid, there can be any number of graphics regions. A graphics region is referred to as a *viewport* and is created using the `viewport()` function. A viewport can be positioned anywhere on a graphics device (page, window, . . .), it can be rotated, and it can be clipped to. For example,

```
viewport(x = 0.5, y = 0.5, width = 0.5, height = 0.25, angle = 45)
```

describes a viewport which is centered within the page, and is half the width and one quarter of the height of the page, and rotated 45°. Note that the above is only a description of a graphics region—it is created on a graphics device only when the viewport is "pushed" onto that device using the function push.viewport(). Each device maintains a stack of viewports, with the top one being the current one. Pushing places a viewport on top of the stack, while "popping" (using pop.viewport()) removes it from there. When several viewports are pushed onto the viewport stack, later viewports are located and sized within the context of the earlier ones. Graphics output is always relative to the current viewport (on the current graphics device). Hence, selecting the region desired for output is simply a matter of pushing and popping the appropriate viewports.

The viewport mechanism makes it very simple to divide a graphics page into areas as desired. For example, to create a plot with a legend taking up 80% and 20%, respectively, of the width of the current viewport, one can simply use

```
push.viewport(viewport(x = 0, width = 0.8, just = "left"))
```

to set up the plot region, plot to it and pop it from the stack, and then use

```
push.viewport(viewport(x = 1, width = 0.2, just = "right"))
```

to set up the legend region. Grid also provides an alternative way for positioning viewports within each other based on *layouts*, which allows a simple emulation of the multi-figure array mechanism in base graphics. Any viewport pushed immediately after a viewport containing a layout may specify its location with respect to that layout.

Each viewport has a number of coordinate systems available. There are four main types: absolute (e.g., "inches"), normalized (e.g., "npc"), relative (e.g., "native"), and referential coordinates, which allow locations and sizes to be specified in terms of physical coordinates, as proportions of the page size (or the current viewport), relative to a user-defined set of x- and y-ranges, and based on the size of some other graphical object, respectively. The selection of which coordinate system to use within the current viewport is made using the unit() function, which creates an object which combines coordinate value and system information. For example,

```
viewport(x = unit(60, "native"),
         y = unit(0.5, "npc"),
         width = unit(1, "strwidth", "coordinates for everyone"),
         height = unit(3, "inches"))
```

describes a viewport which is centered at the x-value 60 of and half-way up the preceding viewport, is 3 inches high and as wide as the text "coordinates for everyone".

Grid provides a standard set of graphical primitives: lines, text, points, rectangles, polygons, and circles (the names of the corresponding functions are obtained by prefixing the names of the corresponding base graphics functions with '`grid.`'). There are also two higher-level components: x- and y-axes. These functions are mostly similar to their base counterparts, but differ in the way graphical parameters, such as line contour and thickness, are specified.

In grid, there is a much smaller set of graphical parameters, consisting of `col` (the "foreground" color for drawing lines and borders), `fill` (the "background" color for filling shapes), `lty` and `lwd` (line type and width), `fontfamily`, `fontface` (such as bold or italic), `fontsize` (the size of text in points), `lineheight` (the height of a line as a multiple of the size of text), and `cex` (multiplier applied to fontsize: the size of text is `fontsize * cex` and hence the size of a line is `fontsize * cex * lineheight`). Settings of graphical parameters are represented by `"gpar"` objects, and may be specified for both viewports and graphical objects. A setting for a viewport will apply to all graphical output within that viewport and all viewports subsequently pushed onto the viewport stack, unless the graphical object or viewport specifies a different setting. A description of graphical parameter settings is created using the `gpar()` function, which can be associated with a viewport or graphical object via their `gp` slots (as accessed by the `gp` argument to the functions creating viewports and graphical objects). The following piece of code illustrates these mechanisms.

```
push.viewport(viewport(gp = gpar(fill = "grey",
                                 fontface = "italic")))
grid.rect()
grid.rect(width = 0.8, height = 0.6, gp = gpar(fill = "white"))
grid.text(paste("This text and the inner rectangle",
                "have their own gpar settings", sep = "\n"),
          y = 0.75, gp = gpar(fontface = "plain"))
pop.viewport()
```

One of the key applications of grid graphics is in the R implementation of Trellis graphics, provided by the **lattice** package (a so-called recommended package that is available from CRAN, and included in every binary distribution of R), see [20], which illustrates how high-level graphics functionality can be built on top of grid. Because lattice consists of grid calls, it is possible to both add grid output to lattice output, and vice versa.

There is also limited support for combining base and grid graphics using functionality provided by the **gridBase** package. [14] shows how to annotate base graphics using grid (e.g., to add axis labels at arbitrary rotations), and to embed base graphics in grid viewports.

Grid provides more features not discussed here. For information on these, and examples of the use of grid, see in particular the documentation at `http://www.stat.auckland.ac.nz/ paul/grid/grid.html`.

4 Packages

The R package system provides a standardized interface to extending R's functionality. In *source* form, packages can contain

- "core" meta-information, currently serialized as a DESCRIPTION file in Debian Control File format (tag-value pairs)
- additional meta-data, such as a NAMESPACE file defining the package code interface
- code and documentation for R
- foreign code to be compiled/dynloaded (C, C++, Fortran, ...) or interpreted (Shell, Perl, Tcl, ...)
- additional material such as data sets, demos, vignettes, package-specific tests, ...

Only the core meta-information must be present. Mandatory meta-data include name and version of the package, and information on the license and the package maintainer. In a file system "representation", a source package consists of a subdirectory containing the DESCRIPTION and possibly other "top-level" files, and several pre-defined subdirectories, some of which may be missing, such as R for R code and src for foreign source code to the compiled and dynloaded.

To be available for extending R, packages must be *installed* to *libraries*, which are simply locations where R knows to find (installed) packages. Installing from source performs a variety of tasks as needed or desired, such as preformatting R documentation in plain text and HTML formats, creating DLLs from foreign code, generating a binary image of the R code, and setting up several data structures with package index information. This process is plug'n'play if the packages are "self-contained" (so that only the standard tools for processing them are required). Developers can provide configuration scripts for automatically dealing with situations where packages depend on the availability of functionality "outside of R", such as libraries for dealing with XML or accessing a database management system.

Creating packages is straightforward: developers simply need to gather the material to be packaged into the appropriate locations relative to the package source directory. If R code is the starting point, R provides a convenience function `packageSkeleton()` which creates the basic file structures as well as documentation skeletons for the R objects.

Packages are distributed as single files archiving their contents. For source packages, gzipped tar files are used. These are created via the `build` utility (currently, a Perl script) which essentially performs necessary cleanups, adds front-matter information, and creates the archive with a canonical file name obtained from the package name and version (as recorded in the DESCRIPTION file). One can also build and install *binary* packages, which are already set up for use on a particular platform (so that only mimimal processing is needed when installing). E.g., CRAN provides binary packages for the 32-bit Windows platforms, because the tools needed for processing the source packages (Make, Perl, compilers, . . .) might not be available to all users on such systems.

Packages can be distributed over the web through *repositories*, which are suitably indexed collections of packages. The package management tools provided by R allow for directly installing packages from repositories and automatically updating installed packages when newer versions are made available in the repositories. This versioning facility, together with the generality of the package mechanism, makes packages an ideal vehicle for distributing many kinds of R-related material which needs to be kept up-to-date, such as e.g. data sets or manuals (preferably implemented as package vignettes, see Section 5). The Bioconductor project (`http://www.bioconductor.org`), an open source and open development software initiative for the collective creation of extensible software infrastructure for computational biology and bioinformatics which uses R as its primary implementation language, is working on providing the next generation of client and server side tools for repository management, featuring in particular a multi-level package dependency mechanism similar to the ones found in popular GNU/Linux distributions such as Debian (`http://www.debian.org`). These tools are already available via the R extension package **reposTools** from Bioconductor, and will eventually be integrated into the R distribution.

Packages can be submitted to unit testing using the `check` utility (currently, a Perl script). When run on a package source directory, this first verifies that the package can be installed as the basic test of whether it "works", and then goes on to perform a variety of other tests, such as checking

- availability and correctness of meta-information (as recorded in the DESCRIPTION file mentioned above);
- R code, including syntactic correctness, common coding problems (e.g., when loading DLLs or defining replacement functions), consistency of S3 generics and methods, etc.;
- R documentation, including correctness (syntax, presence of all required documentation slots), consistency (of code and documentation), and completeness (all user level objects must be documented);
- whether the package is able to run the code in in the examples of its documentation (which is required). In addition, there are mechanisms for regression and certification testing of code: package maintainers can provide files with R code that will be run and if necessary compared to already certified output.

Repository maintainers can use the package testing facilities for controlling the quality of the packages in the repository, and hence the repository itself. For example, the CRAN repository tracks the R release process by only providing packages which pass the tests against the version of R being released. In addition, the effects of changes in (the development and patched version of) R and updates to contributed packages are monitored on a daily basis. It is this continuous improvement process which markedly distinguishes the R project from most other software initiatives which use repositories for distributing extensions.

Most of the testing tools used by the `check` utility are in fact implemented in R (in particular to ensure portability and availability to all users of R) and distributed in the **tools** package contained in the R distribution. It is very important to realize that whereas `check` is a rather inflexible utility for creating standardized reports on the package "quality" status, the underlying functions from package **tools** provide a flexible and extensible toolbox for computing on packages. For example, `codoc()` is a function for checking code/documentation consistency. More precisely, it analyzes the (usually, function synopsis) information of the `\usage` sections of R documentation files, and compares the documented synopses to what the code actually contains. (Currently, code and documentation for functions in a package are not generated from common sources, and hence may be inconsistent.) What `codoc()` returns is an object containing a variety of information, *including* the information on mismatches found. Printing this object gives a status report on mismatches intended for human readers; if no mismatches were found, nothing is printed. This mechanism is used by `check` to assess and report the basic codoc status. But the object returned contains additional data as well, such as information on `\usage` entries not corresponding to valid R syntax after eliminating special markup for indicating synopses for S3 or S4 methods, or on functions for which documentation was registered (via the `\alias` meta-data markup) without providing a synopsis (which "might" be a problem, and in the case of non-method functions in packages with a name space typically is one). Even though this information is not printed, it is available in the result of the codoc computations, and hence can be used for further processing.

5 Sweave

Sweave [9] is a tool that allows to embed the R code for complete data analyses in LaTeX documents. (In fact, we shall see that the underlying principles are much more general.) In the process of generating the displayed version of the document, first the code in the Sweave source file is processed (by R) and its textual or graphical output inserted as appropriate to create a LaTeX source file. Then, a DVI or PDF file is created (by `latex` or `pdflatex`).

A small Sweave source file is shown in Figure 1. The file contains two R code chunks embedded in a simple LaTeX document. At the beginning of a line, '`<< ... >>`' and '`@`' mark the start of a code and documentation chunk, respectively. Sweave translates this to a regular LaTeX document, which is then compiled to give Figure 2. The results of the Kruskal-Wallis test as well as the box plot have nicely been integrated into the final version.

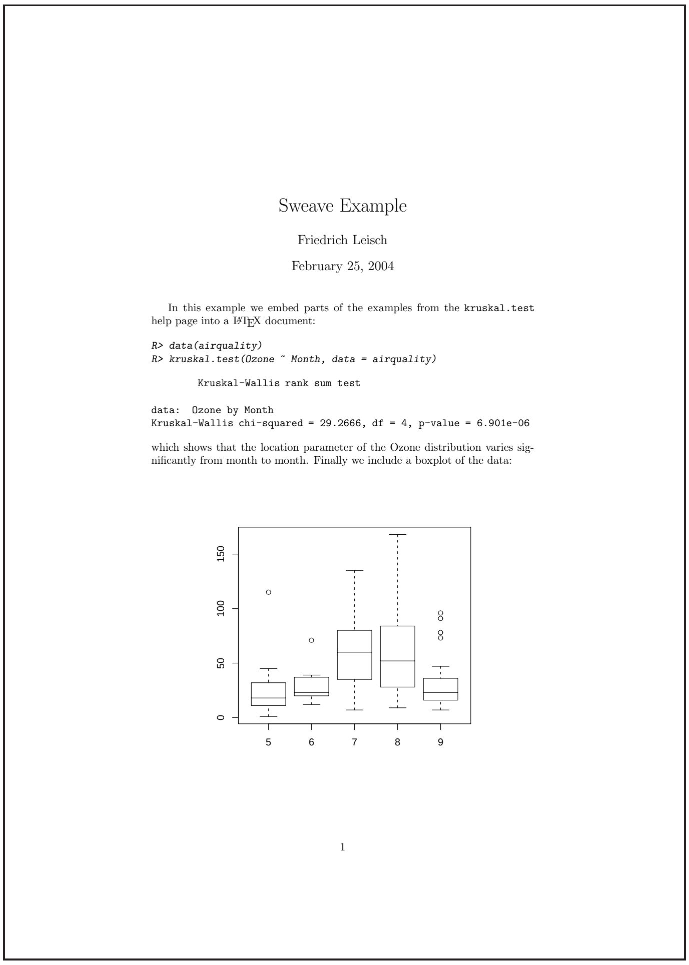
```
\documentclass[a4paper]{article}

\title{Sweave Example}
\author{Friedrich Leisch}

\begin{document}

\maketitle

In this example we embed parts of the examples from the
\texttt{kruskal.test} help page into a \LaTeX{} document:

<<>>= data(airquality) kruskal.test(Ozone ~ Month, data =
airquality) @ which shows that the location parameter of the Ozone
distribution varies significantly from month to month. Finally we
include a boxplot of the data:

\begin{center}
<<fig=TRUE,echo=FALSE>>= boxplot(Ozone ~ Month, data = airquality)
@
\end{center}

\end{document}
```

Figure 1: A small Sweave source file: example.Snw.

The Sweave source file shown in Figure 1 uses the syntax of noweb [18], a simple literate programming tool which allows to combine program source code and the corresponding documentation into a single file. This syntax is particularly useful if Emacs is used for authoring Sweave documents: then, using *ESS* [19, Emacs Speaks Statistics;], an Emacs extension package, one can connect the document to a running R process while writing it. Code chunks can be sent to R and evaluated using simple keyboard shortcuts or popup menus. Syntax highlighting, automatic indentation and keyboard shortcuts depend on the location of the pointer: in code and documentation chunks one gets the same behavior as when editing “simple” R code or LATEX files, respectively. Using Emacs or the noweb syntax is not necessary to Sweave. There is also a LATEX-based syntax, where ‘Scode’ environments are used for marking code chunks. Using this syntax, the boxplot code chunk in our example file would be typeset as

```
\begin{Scode}{fig=TRUE,echo=FALSE}
  boxplot(Ozone ~ Month, data = airquality)
\end{Scode}
```

Sweave offers fine control on how the code chunks are processed. By default, both the S code itself and its console output are inserted, inside suitable



Sweave Example

Friedrich Leisch

February 25, 2004

In this example we embed parts of the examples from the `kruskal.test` help page into a LaTeX document:

```
R> data(airquality)
R> kruskal.test(Ozone ~ Month, data = airquality)

        Kruskal-Wallis rank sum test

data:  Ozone by Month
Kruskal-Wallis chi-squared = 29.2666, df = 4, p-value = 6.901e-06
```

which shows that the location parameter of the Ozone distribution varies significantly from month to month. Finally we include a boxplot of the data:

1

Figure 2: The final document created from example.Snw.

verbatim-style environments, into the generated LaTeX file. This emulates an interactive session. One can suppress either input to or output from the R process, or indicate that output is already in LaTeX format (e.g., when using one of the CRAN extension packages **xtable** or **Hmisc** to create “pretty” tables), or completely suppress the evaluation of the code chunk. In addition, Sweave can replace S expressions inside `\Sexpr` markup in documentation chunks by their values (provided that these can be coerced into a character string).

Sweave is written entirely in S, and contained in package **utils** in the R distribution. From a user’s view, there are two basic functions. `Sweave()` translates Sweave source files into LaTeX files as described above. `Stangle()` simply extracts only the code.

As apparent from the above description, what Sweave really does is perform certain computations on integrated text documents which contain both code and documentation chunks. S4weave, a re-implementation of Sweave using S4 classes and methods currently under way, enforces this view [11]. Providing more structure also makes it possible to compute a directed graph of chunk dependencies, and hence process chunks conditionally. There is also an XML DTD for Sweave source files for document exchange with other dynamic document systems.

To assess the importance of facilities such as Sweave, one should keep in mind how reports as part of a statistical data analysis project are traditionally written. First, the data are analyzed, and afterwards the results of the analysis (numbers, graphs, ...) are used as the basis for a written report. In larger projects the two steps may be repeated alternately, but the basic procedure remains the same. The basic paradigm is to write the report around the results of the analysis. Using Sweave, one can create dynamic reports, which can be updated automatically if data or analysis change. In particular, the code is always available for *reproduce* the displayed results, which makes Sweave an ideal vehicle for disseminating reproducible research, see e.g. [13].

Sweave also greatly aids in the creation and deployment of documentation for "aggregated" functionality of S code, such as manuals for packages (where the traditional function-based S documentation methods cannot easily deliver a comprehensive view), or books on statistical analysis using S. Using Sweave, there is the additional benefit that one can always extract the code from the document (the term *vignettes* has been introduced for documents with this property) and use it for subsequent manipulating and processing. Vignettes have enough structure to allow for an integrated and interactive presentation of the code they contain. For example, `vExplorer()` from the Bioconductor **tkWidgets** package allows to view vignettes and interact with their code chunks, see e.g. [12] for more details.

6 Summary

In this paper, we have discussed four of the key innovations in the "next generation" of R. There are of course many more, including a new system for exception handling, a byte code compiler, external pointer objects, a mechanism for serialization and unserialization of R objects to and from connections, mathematical annotation of plots [15], as well as many refinements to the S language (such as a thorough distinction of the character string `"NA"` from a missing value for a character string). The NEWS file in the top-level directory of the R distribution has more information.

References

[1] Becker R.A., Chambers J.M. (1984). *S. An interactive environment for data analysis and graphics.* Monterey: Wadsworth and Brooks/Cole.

[2] Becker R.A., Chambers J.M., Wilks A.R. (1988). *The new S language.* Chapman & Hall, London.

[3] Becker R.A., Cleveland W.S., Shyu M.-J. (1996). *The visual design and control of trellis displays.* Journal of Computational and Graphical Statistics **5** 123 – 155.

[4] Chambers J.M. (1998). *Programming with data.* Springer, New York. `http://cm.bell-labs.com/cm/ms/departments/sia/Sbook/`.

[5] Chambers J.M., Hastie T.J. (1992). *Statistical models in S.* Chapman & Hall, London.

[6] Cleveland W.S. (1993). *Visualizing data.* Hobart Press, 1993.

[7] Gentleman R., Ihaka R. (2000). *Lexical scope and statistical computing.* Journal of Computational and Graphical Statistics, **9** 491 – 508. `http://www.amstat.org/publications/jcgs/`.

[8] Ihaka R. (1998). *R: Past and future history.* In S. Weisberg, (ed.), Proceedings of the 30th Symposium on the Interface, the Interface Foundation of North America, 392 – 396.

[9] Leisch F. (2002) *Sweave: Dynamic generation of statistical reports using literate data analysis.* In Wolfgang Härdle and Bernd Rönz (eds), Compstat 2002 — Proceedings in Computational Statistics, Physika Verlag, Heidelberg, Germany, 575 – 580. `http://www.ci.tuwien.ac.at/ leisch/Sweave`.

[10] Leisch F. (2002). *Sweave, part I: Mixing R and LaTeX.* R News **2** (3) 28 – 31. `http://CRAN.R-project.org/doc/Rnews/`.

[11] Leisch F. (2003). *Sweave and beyond: Computations on text documents.* In Kurt Hornik, Friedrich Leisch, and Achim Zeileis (eds), Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria. `http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/`.

[12] Leisch F. (2003). *Sweave, part II: Package vignettes.* R News **2** (2) 21 – 24. `http://CRAN.R-project.org/doc/Rnews/`.

[13] Leisch F., Rossini A.J. (2003). *Reproducible statistical research.* Chance **16** (2) 46 – 50.

[14] Murrell P. (2003). *Integrating grid graphics output with base graphics output.* R News **3** (2). `http://CRAN.R-project.org/doc/Rnews/`.

[15] Murrell P., Ihaka R. (2000). *An approach to providing mathematical annotation in plots.* Journal of Computational and Graphical Statistics **9** 582 – 599. `http://www.amstat.org/publications/jcgs/`.

[16] Pinheiro J.C., Bates D.M. (2000). *Mixed-effects models in S and S-Plus.* Springer. `http://nlme.stat.wisc.edu/MEMSS/`.

[17] R Development Core Team (2004). *Writing R extensions.* R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

[18] Ramsey N. (1998). *Noweb man page.* University of Virginia, USA, 1998. `http://www.cs.virginia.edu/ nr/noweb`. Version 2.9a.

[19] Rossini A.J., Heiberger R.M., Sparapani R., n Mächler M., Hornik K. (2004). *Emacs speaks statistics: A multi-platform, multi-package development environment for statistical analysis.* Journal of Computational and Graphical Statistics **13** (1), 1 – 15

[20] Sarkar D. (2002). *Lattice.* R News **2** (2) 19 – 23. `http://CRAN.R-project.org/doc/Rnews/`.

[21] Therneau T.M., Grambsch P. (2000). *Modeling survival data: extending the Cox model'.* Springer.

[22] Tierney L. (2003). *Name space management for R.* R News **3** (1) 2 – 6. `http://CRAN.R-project.org/doc/Rnews/`.

[23] Venables W.N., Ripley B.D. (2002). *Modern applied statistics with S. Fourth edition.* Springer. `http://www.stats.ox.ac.uk/pub/MASS4/`.

Acknowledgement: Section 2 is based on material in [22] and the *Writing R Extensions* manual [17], Section 3 on a primer on "Grid Graphics" by Paul Murrell. Section 5 draws from [10].

Address: K. Hornik, Institut für Statistik, Wirschaftsuniversität Wien, Austria

E-mail: `Kurt.Hornik@wu-wien.ac.at`

ROBUST MULTIDIMENSIONAL SCALING

Leanna L. House and David Banks

Key words: Statistical computing, data reduction, robust, multidimensional scaling.

COMPSTAT 2004 section: Statistical software.

Abstract: Modern technology enables the collection of vast quantities of data. Smart automatic data selection algorithms are needed to discover important data structures that are obscured by other structure or random noise. We suggest an efficient and flexible algorithm that chooses the "best" subsample from a given dataset. We avoid the combinatorial search over all possible subsamples and efficiently find the datapoints that describe the primary structure of the data. Although the algorithm can be used in many analysis scenarios, this paper explores the application of the method to problems in multidimensional scaling.

1 Introduction

Although modern technology enables the collection of huge amounts of data, it also exacerbates the problem of data quality control. Spurious or erroneous information caused by either the random nature of the data or human error will inevitably exist within large datasets. But the task of sifting through millions of observations and removing those that are not representative of the true population borders on the impossible. Smart, automated, data cleaning algorithms or robust analysis tools that work in tandem with the collection technologies are needed.

From a statistical perspective, robust analysis methods, including L, M, S, and R estimators, serve as appropriate means to account for contaminated data. However, such methods arguably apply only to parametric approaches and do not extend to unsupervised learning problems or multidimensional scaling. Furthermore, analyzing the data directly, without first reducing the number of observations, may exceed computer software or memory limitations.

To address this problem, we present an efficient data reduction algorithm that actively seeks the primary underlying structure of the data while removing spurious observations. Rather than use graphical methods to hunt for erroneous data as described by Karr, Sanil, and Banks [6], we systematically search among strategically chosen subsets of the collected sample. Ultimately, we find the subsample that provides the best statistical signal, as measured in terms of fit, compared to other subsets of comparable size.

The algorithm we propose does not require the evaluation of every subset within a sample. Instead, it performs a series of greedy searches that allow the method to scale to large datasets. And the algorithm is flexible since it can be applied to any situation in which there is some measure of goodness-of-fit. In this paper, we describe how the method applies in the context of linear regression and multidimensional scaling, where the measures of fit are R^2 and stress, respectively.

We understand that specifying an acceptable degree of lack-of-fit or required statistical signal for a chosen subsample is unclear. Since one is trying to cherry-pick the best possible subset of the data, we consider two options. The first entails the prespecification of the final subset size. The subset with the highest statistical signal (of the specified size) is chosen, regardless of the magnitude of the signal, or lack there of. The second approach requires the inspection of the plot, signal versus subset size. A knee in the plotted curve points to the subset size at which one is forced to include bad data.

In the context of previous statistical work, our approach is most akin to the S-estimators introduced by Rousseeuw and Yohai [9], which built upon Tukey's proposal of the shorth as an estimate of central tendency [2], [9]. Our key innovations are that instead of focusing upon parameter estimates we look at complex model fitting, and also we focus directly upon subsample selection. See [3], [4] for more details on the asymptotics of S-estimators and the difficulties that arise from imperfect identification of bad data.

In the context of previous computer science work, our procedure is related to one proposed by Li [7]. That paper also addresses the problem of finding good subsets of the data, but it uses a chi-squared criterion to measure lack-of-fit and applies only to discrete data applications. Besides offering significant generalization, we believe that the two-step selection technique described here enables substantially better scalability in realistically hard computational inference.

Section 2 describes the algorithm in detail within the context of regression. Section 3 illustrates the flexibility of the algorithm and applies it to a simulated, multidimensional scaling scenario. Section 4 concludes the paper with a discussion and a description of additional applications.

2 Proposed algorithm

Because of the wide familiarity with regression, we describe the steps of the algorithm while referring to the following scenario:

> Given n observations, $\{Y_i, \boldsymbol{X}_i\}$, we assume that the expected structure within the data is a multivariate linear model
>
> $$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i$$
>
> with independent errors terms, $\epsilon_i \sim N(0, \sigma)$. And we want to protect our analysis against the the possibility that as much as $1 - Q$ percent of the data either do not have a common linear

relationship or are random noise or follow a different functional relationship with Y. The choice of Q requires domain knowledge or a good sense of the errors in the data collection protocol.

Typical regression analyses fit all the data, and then attempt to identify outliers or high-leverage points. Some robust methods, such as S-estimation, attempt to find the best fit to some prespecified fraction of the data, but those methods do not generalize to, say, nonparametric multivariate regression. In contrast, we search among the data to find a large subset that produces good fit. This entails random selection of starting-point subsamples and the comparison of fits from subsamples of the data.

In a linear regression setting, the coefficient of determination, R^2, provides a natural choice for assessing and comparing the statistical signal of subsamples. The statistic relies on sums of squared deviations to assess lack-of-fit and does not penalize subsets for including more or less observations. Simply, a subsample with a high R^2 is better than another with a low R^2.

In general, it is desirable that the measure of fit not depend upon the size of the subsample. This is true for the coefficient of determination and also for stress in multidimensional scaling. The algorithm, however, can be modified to accommodate other situations, usually by a normalization that allows one to measure the "average" goodness-of-fit. That technique allows one to broaden the field of fit criteria to include average absolute deviation or average complexity, as measured by Mallow's C_p statistic [8] or Akaike's Information Criterion [1].

The remainder of this section describes how we randomly select a set of subsamples from which we ultimately choose the best. We do not enumerate or test all possible subsamples of size Qn. Rather, we propose starting with a series of small, randomly chosen datasets and growing each until they are of size Qn. Done properly, we can ensure that with some prespecified probability at least one of the original subsamples will eventually grow to contain nearly all good data.

2.1 Choosing the initial subsamples

To begin we select the minimum number, d, of subsamples S_i needed to guarantee, with probability c, that at least one S_i contains only "good" data; i.e., data for which the assumption of a linear model is correct. The size of the initial subsamples depends on the scenario and should equal the minimum number of observations needed to calculate the chosen lack-of-fit measure; for the case of multivariate regression in $\mathbb{R}^p$ using R^2 as the criterion, one needs $p + 2$ observations in each starting subsample.

Assuming Q percent of the data are good, then the probability of selecting (with replacement) a starting subsample that contains bad data is $1 - Q^{p+2}$. Hence, after specifying c, we may solve for d using

$$\begin{aligned} c &= \mathbb{P}[\text{ at least one of } S_1, \ldots, S_d \text{ is all good }] \\ &= 1 - \mathbb{P}[\text{ all of } S_1, \ldots, S_d \text{ are bad}] \\ &= 1 - \prod_{i=1}^{d} \mathbb{P}[\ S_i \text{ is bad }] \\ &= 1 - (1 - Q^{p+2})^d. \end{aligned}$$

For example, if we want the probability of selecting at least one good initial sample to equal 95% ($c = .95$) and we assume that 20% of the data are spurious ($Q = .8$), then we have $.95 = 1 - [1 - (.8)^{p+2}]^d$. Setting p = 1 for simple linear regression, the smallest integer greater than d is 5. Thus we need five starting-point subsamples to ensure with probability .95 that one of them will work as we want.

We assume the probability of choosing the same observation twice for one subsample is small enough to justify selecting S_i with replacement. However, one may use finite population methods if necessary (e.g., when the total sample size is small). In that case the calculation of d becomes slightly more complicated when p is very large. Such cases might necessitate the use numerical techniques to find d.

2.2 Select subsamples

Since the exact value for Q is unknown, let k equal the desired proportion of data we wish to select from the large dataset. (The value for k does not necessarily have to equal Q.) One subsample at a time, we sequentially append observations that improve (or cause littel reduction) in the goodness-of-fit measure until S_i contains the target number of kn data points.

To balance the need for computational speed against the risk of adding bad data, we suggest a two-step rule for adding observations. For the sake of creating a time efficient algorithm, we accept the risk of suboptimal selections, but we want to avoid the possibility of a "slippery slope." Specifically, we do not want a selection that only slightly increases the lack-of-fit to lower the standard so that we get a subsequent selection that also slightly increases the lack-of-fit, with the end result that a chain of marginally satisfactory selections eventually produces a subsample that contains bad data.

The addition process begins with a fast search that adds data points as the algorithm sweeps through the data (Step 1). Starting with the statistical fit measured in an original subsample, S_i, we consider the addition of each of the remaining observations in succession. If the union of an observation with S_i either increases the statistical signal or only decreases it by a minute, prespecified amount η, then the observation is added to the subsample. Hence the next candidate data point in the sequence is considered with regard to a new, slightly larger S_i. Setting n_i to represent the number of observations in the current S_i, the algorithm stops when n_i equals kn.

If after sweeping through the data one time we have $n_i < kn$, our algorithm moves to the second, significantly slower step. Here, we search over all data not already in the subsample to find the observation which, when added, reduces the goodness-of-fit by the smallest amount. We then add that observation and either improve the fit measure for S_i t best or decreases the statistical measure by the smallest possible amount (regardless of η). Notice step 2, unlike step 1, guarantees the addition of one observation on each pass through all of the data (excluding observations already in S_i). Step 2 is repeated until $n_i = kn$.

The following pseudo-code describes this two step algorithm. We use $\mathrm{GOF}(\,\cdot\,)$ to denote a generic goodness-of-fit measure.

Pseudocode for a Two-Step Selection

```
Step 1: Fast Search
Initialize: Draw $d$ random samples $S_i$ of size $p+2$ (with replacement).

Search over all observations:
        Do for all samples $S_i$:
            Do for observations $\boldsymbol{Z}_j = (Y_j, \boldsymbol{X}_j)$:
                If $\boldsymbol{Z}_j \in S_i$ goto next $j$
                If GOF($S_i$) - GOF($\boldsymbol{Z}_j \bigcup S_i$) $< \eta$ add $\boldsymbol{Z}_j$ to $S_i$.
                If $n_i = [kn]$ stop.
            Next $j$
        Next $i$.

Step 2: Slow Search
Search over all observations:
        Do for all samples $S_i$:
            Do for observations $\boldsymbol{Z}_j = (Y_j, \boldsymbol{X}_j)$:
                If $\boldsymbol{Z}_j \in S_i$ goto next $j$
                If GOF($\boldsymbol{Z}_j \bigcup S_i$) $> \max_j$ GOF($\boldsymbol{Z}_j \bigcup S_i$) add $\boldsymbol{Z}_j$ to $S_i$.
                If $n_i = [kn]$ stop.
            Next $j$
        Next $i$.
```

The algorithm requires two vital inputs: the goodness-of-fit measure and the choice of η, the tolerated increase in lack-of-fit during step 1. As mentioned previously, we recommend that the goodness-of-fit measure not depend upon the sample size; the lack-of-fit values should be comparable as n_i increases. However, the choice of η offers one way to force comparability by making it depend upon n_i as well.

If one can achieve independence between the lack-of-fit measure and sample size, then the selection of η depends upon one's willingness to accept bad observations. In the regression setting, when $\eta = 0$, step 1 only appends data points that strictly improve the R^2. On the other hand, the value of η can be determined empirically by inspection of a histogram of 100 lack-of-fit values obtained by adding 100 random data points to an initial subsample of size $p + 2$.

After repeating Step 1 and 2 for d subsamples, the final task is to select one S_i as the best or most representative of the underlying structure. If the purpose for implementing the proposed algorithm is strictly to reduce the dataset to kn, then one could select the subsample with the lowest lack-of-fit, regardless of its size. On the other hand, if the inclusion of bad observations is worrisome or the magnitude of the goodness-of-fit measure for the best subsample is unsatisfactory, then we recommend plotting the goodness-of-fit against the order of entry of the observations. Given an initial subsample with only good data, the graph should depict a long plateau with a sudden knee in the curve when bad observations begin to enter the subsample. One may choose the best size for the subsample according to the size at which the knee occurs.

Note the proposed algorithm entails a stochastic choice of starting sets, followed by a deterministic extension algorithm. Even though we can guarantee, with a specified probability, a clean starting set, we cannot make the same guarantee at the conclusion of the algorithm. Since the extension procedure depends slightly upon the order in which the cases are considered, the final result does not quite enjoy the same probabilistic properties as the initial starting sets. Nevertheless, simulation results indicate that the proposed procedure does lead, with probability near the nominal level specified in the initial calculation that determined the number of starting-point subsamples, to the selection of a subsample of good data.

3 Application: multidimensional scaling

The robustness problem in the linear regression example could have been addressed through other means, such as S-estimators, but it provides a convenient test-bed for developing and assessing the proposed methodology. Our real interest lies in more complicated problems, such as arise in nonparametric regression or classification with mislabeled data or non-metric multidimensional scaling.

Here we demonstrate the strengths of the two-step algorithm within the context of multidimensional scaling (MDS). A practical concern in using MDS is that a relatively small proportion of outliers or similar data quality problems can distort the fit into uninterpretability. Essentially, a mulitdimensional analysis attempts to force a fit that is driven largely by the bad data, and thus simple low-dimensional structure in the good data can be overlooked or not represented at all. Our procedure for cherry-picking the best sample allows the fitting procedure to ignore points that cause large increases in lack-of-fit, which in this context is most naturally measured by the stress function.

Given a clean dataset that consists of the latitudes and longitudes of 99 major cities in the eastern United States, we generated six (three groups of two) unclean datasets. The datasets differ with respect to the proportion of bad data and their degree of badness (refer to Table 1). The first set distorts one distance between two cities by 150% and 500%. The remaining sets increase the number of distortions to 10 and 30 interpoint distances. For the latter two groups, some altered distances might share one end-point. Thus we consider the percent of unclean cities, or $1 - Q$ to be greater than or equal to 2%, 10% and 30% for each set respectively.

True 1-Q (%)	Distance Distortion (%)	Original Stress	n^a	n^*	Final Stress
2	150	1.028	80	80	4.78e-12
	500	2.394	80	80	4.84e-12
10	150	1.791	80	80	4.86e-12
	500	28.196	80	80	4.81e-12
30	150	3.345	80	77	4.86e-12
	500	9.351	80	78	4.78e-12

Table 1: Compare 6 data quality scenarios for MDS.

Using Kruskal-Shephard non-metric scaling, we assess the statistical signal of a given dataset by using the stress function

$$\sum_{i \neq i'} = \sum_{i \neq i'} \frac{[g(\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|) - d_{ii'}]^2}{\sum_{j \neq j'} d_{jj'}^2}$$

where the $d_{ii'}$ are the distances between the two-dimensional embeddings of the points $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$ and the $g(\cdot)$ is an arbitrary monotonically increasing function (this implies that the fit depends only upon the ranks of the interpoint distances). The fitting is done by alternating isotonic regression to find an estimate of g with gradient descent to find an estimate of the $d_{ii'}$; our implementation used the procedure in the R software package.

For each of the six datasets, the algorithm attempts to find the subset that minimizes the stress function the most. Since the cities lie on the surface of the globe and do not embed perfectly onto a two-dimensional Euclidean space, some stress exists even within the clean dataset. The total stress measures for the perturbed datasets are listed in the third column of Table 1, whereas the stress for the original, clean dataset equals $8.42 * 10^{-12}$. Additionally, we chose to set $\eta = 1.0^{-12}$, a value slightly greater than zero and commensurate with stress in the undistorted sample.

In a real situation, Q, the percent of clean observations, is unknown. Thus, using expert information we must estimate Q in order to calculate d, the required number of starting-point subsamples. Furthermore, k, the percent by which we wish to reduce the original dataset, is typically unclear as well. In this example, for all of the datasets we assumed that $Q = .9$ and we set $k = .8$.

Table 1 describes the effect of implementing the algorithm using each dataset. The columns labeled "Original Stress" and "Final Stress" provide the stress measures for the complete datasets and the chosen subsamples respectively. The column labeled "n^a" gives the number of observations in the best subsample chosen from the direct application of the algorithm. And the column labeled "n^*" gives the number of observations in the chosen subsample after inspecting graphs that plot stress against sample size. Notice $n^a \neq n^*$ in the last two rows, when 30 interpoint distances are perturbed. This is due to the fact that k is greater than the true value of Q. Figure 1 displays the plots of stress against sample size.

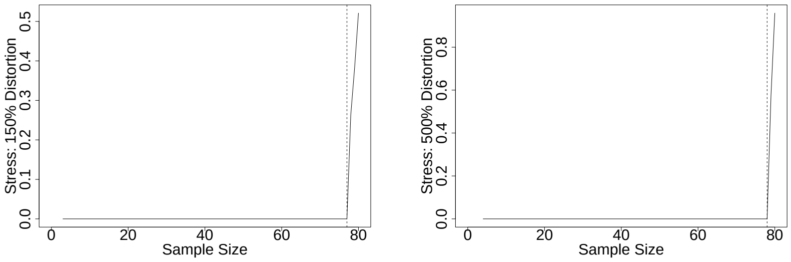


Figure 1: Plot of stress measure versus sample size (in the order of entry) when 30 distances are distorted: (left) 150% distortion; (right) 500% distortion; Notice plateau in graph while including good observations in subsample, but at sample size = 77 (left) and sample size = 78 (right) we start to append bad data.

4 Discussion

In order to take advantage of the full potential of a large dataset, we propose a straightforward method to remove bad data. In essence, we robustify the data using a two-step algorithm to select the subsample that is in best agreement with the assumed structure in the data.

We demonstrate the benefits of the algorithm within the context of multidimensional scaling. In MDS scenarios, even small proportions of bad data can entirely distort the apparent geometric relationships among the cases. Our algorithm successfully isolates the primary structure of six distorted datasets. The stress measures of the final chosen subsamples are dramatically lower than those of the corresponding original datasets.

One distinguishing feature of the algorithm is that it does not require the complete enumeration of all possible subsamples. This saves an enormous amount of computer time, and ensures that the algorithm is essentially of order $\mathcal{O}(n)$ (if one avoids or minimizes the slow-search phase). However, the spirit of our two-step algorithm could be implemented in other ways. For example, solely running the slow search in step 2 might be optimal in terms of only choosing the very best observations to include within a subsample. However, this requires $d(n-p-2)$ separate reviews of the entire pool, which is hard when n is large or the calculation of the lack-of-fit measure is complex.

The procedure we describe extends easily to almost any statistical application, requiring only some measure of fit. In fact, it can even address multiple structures within a dataset. By applying the algorithm repeatedly, each time removing the data that fit the most recently discovered underlying structure, one can retrieve disjoint subsamples representing different models. Subsequent work will extend this technique to such situations and provide a more thorough study of the performance of the search procedure.

References

[1] Akaike H. (1973). *Information theory and an extension of the maximum likelihood principle.* Second International Symposium on Information Theory, 267 – 281.

[2] Andrews D.F., Bickel P. J., Hampel F. R., Huber P. J. Rogers W.H., Tukey J.W. (1972). *Robust estimates of location: survey and advances.* Princeton University Press, Princeton, NJ.

[3] Davies P.L. (1987). *Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices.* Annals of Statistics **15**, 1269 – 1292.

[4] Davies P.L. (1990). *The asymptotics of S-estimators in the linear regression model.* Annals of Statistics **18**, 1651 – 1675.

[5] Hawkins D.M. (1993). *A feasible solution algorithm for the minimum volume ellipsoid estimator in multivariate data.* Computational Statistics **9**, 95 – 107.

[6] Karr Alan F., Sanil Ashish P., Banks David L. (2002). *Data quality: a statistical perspective.* National Institute of Statistical Sciences, Research Triangle Park, NC.

[7] Li X.-B. (2002). *Data reduction vis adaptive sampling.* Communications in Information and Systems **2**, 53 – 68.

[8] Mallows C.L. (1973). *Some comments on C_p.* Technometrics **15**, 661 – 675.

[9] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outliers detection.* Wiley, New York.

[10] Rousseeuw P.J., d Yohai V. (1984). *Robust regression by means of S-estimators.* In Robust and Nonlinear Time Series Analysis, J. Franke, W. Härdle, R.D. Martin (eds.), Lecture Notes in Statistics **26**, Springer-Verlag, New York, 256 – 272.

Address: L.L. House, D. Banks, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, 27708 U.S.A.

E-mail: `house@stat.duke.edu, banks@stat.duke.edu`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

IMPROVED JACKKNIFE VARIANCE ESTIMATES OF BILINEAR MODEL PARAMETERS

Martin Høy, Frank Westad and Harald Martens

Key words: PLSR, PCR, bilinear model, jackknife, variance.

COMPSTAT 2004 section: Partial least squares.

Abstract: This paper puts focus on some the remaining issues concerning jackknifing of centred bilinear models. A method improvement is proposed, describing how all the bilinear model parameters can be rotated in order to estimate the uncertainties of all model parameters. The mean values of centred models are also included in the rotation scheme.

The uncertainty information of the bilinear model parameters can be used to perform variable selection, variable weighting and detection of outliers.

1 Introduction

Crossvalidation [1] and especially jackknife [2] can be used in order to estimate the uncertainty of the parameters in a bilinear model [3]. This technique is currently used in commercial software (e.g. The Unscrambler) to estimate the uncertainty in the reduced-rank regression coefficients $\boldsymbol{b}_A$ in the multiple linear approximation model at rank A,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}_A + b_{0,A} \tag{1}$$

or for multiple y-variables

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{B}_A + 1\boldsymbol{b}'_{0,A} \tag{2}$$

Preliminary versions of stability information of the bilinear loadings $\boldsymbol{P}_A$, $\boldsymbol{Q}_A$ and scores $\boldsymbol{T}_A$ for the underlying bilinear regression models (see equation (6) for definitions) are also available. The uncertainty in the regression coefficients is used for e.g. variable selection while the uncertainty in the scores is used to make "stability plots" and e.g. spot sample outliers.

In this article, the method of calculating uncertainty of regression coefficients is expanded to also include the uncertainty of the bilinear model parameters, the loadings and loading weights ($\boldsymbol{p}_A$, $\boldsymbol{Q}_A$, $\boldsymbol{W}_A$) and the scores ($\boldsymbol{T}_A$). The mean value of centred models are also included in the proposed rotation scheme. This has been lacking in commercial applications, and has not yet been described in the literature.

2 Theory

2.1 Notation

Matrices are written as uppercase bold letters ($\boldsymbol{X}$), while vectors are written as lowercase bold letters ($\boldsymbol{x}$). Unless transposed (written as $\boldsymbol{x}'$), vectors are always columns. Uppercase letters (A) denotes constants, while lowercase letters are counters or indexes ($a = 1 \ldots A$).

2.2 Jackknife and segmentation

When crossvalidating or jackknifing a model, the dataset with N samples (objects) are divided into M segments.M sub-models are estimated where model $m = 1 \ldots M$ is estimated from the slightly smaller dataset where the objects in segment m are left out. In the special case of leave-one-out crossvalidation, $M = N$ with $N - 1$ samples in each subset. We have chosen to label the segment that is left out with a subscript m, and the reduced dataset with segment m missing is labelled with a subscript $-m$.

When jackknife is used in statistical literature, the data are often considered to be drawn from the same distribution, and focus is then on creating as many “independent” estimates as possible. The most common way to perform jackknife-validation is the leave-one-out, which gives N estimates of each parameter. One can also perform delete-d jackknife, where d samples are removed in each subset, giving $\binom{N}{d}$ estimates. For $d > 1$ the delete-d jackknife thus shifts the jackknife estimate towards to the bootstrap estimate, which is based on random sampling of errors or samples. The statistical formulae and properties of these estimates are well documented in statistical literature [4], [5], [6].

If the dataset is generated by e.g. a factorial design, it may contain variability on different levels. Take as (a hypothetical) example an experimenter who has tested four different levels (doses) of a treatment on 20 patients twice (two replicates), giving a total of 40 experiments. She might be interested in both the variation between the dose-levels, the variation between the patients, and the variation for a given patient over time. Traditionally, one would use ANOVA to obtain this information, but the same can be achieved by using cross-validated or jackknifed PLSR with the right segmentation of the data (see also [7]).

In this example, one could first place all samples with the same dose-level in the same segment. This would give $M = 4$ segments, and the validation would then show the ability of three of the treatments to predict the third, i.e. how different the response to the dose-levels are. One could also remove one patient at the time giving $M = 20$ segments, to validate how similar or different the patients reacted to the doses. This would be a good segmentation in order to look for outliers between the patients, i.e. whether one (or more) of the patients reacted to the treatment in a very different way than the others. Yet another possibility would be to remove one replicate at the time, giving $M = 2$ segments. The validation would then show whether the patients changed over time. One could also use the leave-one-out method giving $M = 40$ segments. The validation would then be a mix of the above, testing both the dose-levels, replicates and patients at once. These four examples of segmentation will in general give quite different estimates of the variances in the model parameters. Thus, it is very important to be aware of on what level one is validating the results [8].

Even though the jackknife-formulae for different segmentations are given in statistical literature, the authors feel the need for documenting these also in the chemometric literature. The most general expression is that of delete-d jackknife, where one explores all the combinations of data where d samples are removed, $\binom{N}{d}$. The variance of a parameter θ can then be estimated as

$$\hat{s}^2(\theta) = \frac{N-d}{d\binom{N}{d}} \sum_{m} \left(\hat{\theta}_{-m} - \bar{\theta}\right)^2 \tag{3}$$

where $\hat{\theta}_{-m}$ is the value of θ estimated when segment m is removed, and $\bar{\theta}$ is the mean value of all the estimated values.

Like in the example above with treatments and patients, we often don't explore all the combinatorial possibilities of removing d samples at the time. Instead, we only use the $M = N/d$ possible subsets given by removing each of the M segments one at the time. For $d = 1$ these two methods are the same, namely the leave-one-out validation. But for $d > 1$, we have $\binom{N}{d} \gg M$. When only M of the possible subsets is used, equation (3) reduces to

$$\hat{s}^2(\theta) = \frac{M-1}{M} \sum_{m=1}^{M} \left(\hat{\theta}_{-m} - \bar{\theta}\right)^2 \tag{4}$$

When doing significance-testing based on variance estimates from jackknife, one needs to know how many degrees of freedom to use. When using estimates from equation (4), the degrees of freedom in the variance estimate is $M - 1$. To illustrate both the correctness of equation (4) and the $M - 1$ degrees of freedom, the authors performed a Monte-Carlo simulation. The results are documented in section (3.1). The theory and results presented here are in contrast to [8], where the factor $(N - 1)/N$ is used.

In all the above, it is assumed that the size of the different segments is equal (or not very different). For segmentation schemes with unequal segment-sizes, the above formulae are more complicated.

2.3 Variance of regression coefficients

From each of the M bilinear submodels (perturbations of eqs. (1), each time using with A latent variables or factors) we estimate regression coefficients $\boldsymbol{b}_{-m,A}$, and from the complete dataset we estimate $\boldsymbol{b}_A$. One approach to estimate the uncertainty in $\boldsymbol{b}_A$ is then to sum all the squared deviations from $\boldsymbol{b}_A$ [3]:

$$\hat{s}^2(b_{ka}) = \frac{M-1}{M}\sum_{m=1}^{M}(b_{-m,ka} - b_{ka})^2 \tag{5}$$

The correction-factor outside the summation is reduced to the more well-known $(N-1)/N$ for leave-one-out crossvalidation or ordinary jackknife [5].

Note also another difference to the jackknife as described in statistical literature [4], whre each $\boldsymbol{b}_{-m,A}$-estimate is compared to the *mean* of all the M submodel estimates instead of using the value from the complete dataset. The idea behind using $\boldsymbol{b}_A$ as in equation (5) is that this is the "best" estimate we can get, using all the samples that we have available. In most cases, this is also the estimate that would be used as the final model, and we are interested in the variation around that estimate. This bias-including mean squared error estimate eliminates the mean of the perturbed submodel parameter estimates from the jackknife expressions. Since the reduced-rank PLSR models deviates from the theoretical properties of the well understood traditional full-rank OLS regression models, the authors consider the known theoretical properties in full-rank OLS regression models non-applicable for the PLSR solution. Examples will be given in the section "Results and Discussion" that substantiate this choice.

2.4 Rotation of bilinear models

It would be nice to calculate the uncertainty of all the other PCR/PLSR model parameters in the same simple way as the regression coefficients in equation (5), but this is complicated due to certain properties of the bilinear model. The bilinear model as in both PCR and PLSR can be seen as a sum of outer-products, one for each factor:

$$\boldsymbol{X} = 1\bar{\boldsymbol{x}}' + \sum_{a=1}^{A}\boldsymbol{t}_a\boldsymbol{p}'_a + \boldsymbol{E}_A \qquad \text{and} \qquad \boldsymbol{Y} = 1\bar{\boldsymbol{y}}' + \sum_{a=1}^{A}\boldsymbol{t}_a\boldsymbol{q}'_a + \boldsymbol{F}_A \tag{6}$$

where $\bar{\boldsymbol{x}}'$ and $\bar{\boldsymbol{y}}'$ contains the mean value of each variable, $\boldsymbol{t}_a$ is a vector of scores (a linear combination of the $\boldsymbol{X}$-variables), $\boldsymbol{p}_a$ and $\boldsymbol{q}_a$ are loadings for $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively and $\boldsymbol{E}_A$, $\boldsymbol{F}_A$ contains unmodelled residuals. The only difference between the PCR- and PLSR-algorithms lies in the way $\boldsymbol{t}_a$ is defined.

A property of bilinear models is that the scores and loadings have rotational freedom. We can rotate the scores in any direction, as long as the corresponding loadings are rotated the same amount in the opposite direction. The model will still contain the same information, and the regression coefficients will be the same.

Scores- and loading-vectors for the different submodels m may appear to be quite different due to trivial translations, rotation and mirroring. If e.g. the sign of each element in *both* $\boldsymbol{t}_{-m,a}$ and $\boldsymbol{p}_{-m,a}$ changes, the information explained by their product in that factor will still be the same, but it will be meaningless to compare each value in those vectors to other score- or loading-vectors with different alignment. One way to solve this problem is to *rotate* all the M sub-models toward the model calculated from the complete dataset before we compare them.

Equation (6) represents the model calculated from the complete dataset with all N samples. Rewriting that model using matrix notation, we get

$$\begin{aligned}\boldsymbol{X} &= [1 \quad \boldsymbol{T}_A]\ [\overline{\boldsymbol{x}} \quad \boldsymbol{P}_A]' + \boldsymbol{E}_A\\ \boldsymbol{Y} &= [1 \quad \boldsymbol{T}_A]\ [\overline{\boldsymbol{y}} \quad \boldsymbol{Q}_A]' + \boldsymbol{F}_A\end{aligned}$$

where

$$\boldsymbol{T} = (\boldsymbol{X} - 1\overline{\boldsymbol{x}}'\boldsymbol{W}_A](\boldsymbol{P}'\boldsymbol{W})^{-1} \tag{7}$$

and $\boldsymbol{W}_A$ is the internal loading weight matrix. For each consecutive factor, the corresponding column in $\boldsymbol{W}_A$ is defined as the first eigenvector of residual $\boldsymbol{X}-\boldsymbol{X}$ covariance (in PCR) or $\boldsymbol{X}-\boldsymbol{Y}$ covariance (in PLSR). The linear regression coefficients in eqs. (1), (2) is then defined as

$$\boldsymbol{B}_A = \boldsymbol{W}_A(\boldsymbol{P}'_A\boldsymbol{W}_A)^{-1}\boldsymbol{Q}'_A \tag{8}$$

Similarly, we can write each of the M sub-models in matrix notation, where the index $-m$ denotes that segment m has been left out.

$$\begin{aligned}\boldsymbol{X}_{-m} &= [1 \quad \boldsymbol{T}_{-m,A}]\ [\overline{\boldsymbol{x}}_{-m} \quad \boldsymbol{P}_{-m,A}]' + \boldsymbol{E}_{-m,A}\\ \boldsymbol{Y}_{-m} &= [1 \quad \boldsymbol{T}_{-m,A}]\ \left[\overline{\boldsymbol{y}}_{-m} \quad \boldsymbol{Q}_{-m,A}\right]' + \boldsymbol{F}_{-m,A}\end{aligned} \tag{9}$$

Without changing equation (9), we can insert an invertible matrix $\boldsymbol{C}$ and its inverse $\boldsymbol{C}^{-1}$, since $\boldsymbol{C}\boldsymbol{C}^{-1} = \boldsymbol{I}$.

$$\begin{aligned}\boldsymbol{X}_{-m} &= [1 \quad \boldsymbol{T}_{-m,A}]\ \boldsymbol{C}_{-m}\boldsymbol{C}_{-m}^{-1}\ [\overline{\boldsymbol{x}}_{-m} \quad \boldsymbol{P}_{-m,A}]' + \boldsymbol{E}_{-m,A}\\ \boldsymbol{Y}_{-m} &= [1 \quad \boldsymbol{T}_{-m,A}]\ \boldsymbol{C}_{-m}\boldsymbol{C}_{-m}^{-1}\ \left[\overline{\boldsymbol{y}}_{-m} \quad \boldsymbol{Q}_{-m,A}\right]' + \boldsymbol{F}_{-m,A}\end{aligned} \tag{10}$$

Comparing equation (7) and equation (10), we can define $\boldsymbol{C}_{-m}$ as a rotation matrix, where we try e.g. try to rotate $[1 \quad \boldsymbol{T}_{-m,A}]$ towards $[1 \quad \boldsymbol{T}_A]$. Similarly, we then interpret $\boldsymbol{C}_m^{-T}$ as a rotation of $[\overline{\boldsymbol{x}}_{-m} \quad \boldsymbol{P}_{-m,A}]$ towards $[\overline{\boldsymbol{x}} \quad \boldsymbol{P}_A]$. Thus, if we wanted to estimate the matrix $\boldsymbol{C}_{-m}$, we could use either the relation between the scores or the relation between one of the loadings as targets.

If the data were without noise, perfectly behaved and contained sufficient redundant information, the only difference between the submodel and the total model would be reflections and possibly reorderings (permutations) of the factors. It would then be possible to map the submodel onto the total model with a matrix $\boldsymbol{C}_{-m}$ containing only one ± 1 per column/row, and the rest of the elements 0. But when the data contains noise and insufficient redundant information, rotation at angles that are not multiples of 90° and possibly rescaling of the axis will be necessary to map the submodel perfectly onto the total model.

In order to consume as few degrees of freedom in $\boldsymbol{Y}$ as possible in the estimation of $\boldsymbol{C}$, we have chosen to use the scores matrices as targets. Since cross-validation/jackknife segment m has been removed in $\boldsymbol{T}_{-m,A}$, it has fewer rows than $\boldsymbol{T}_A$. In order to estimate $\boldsymbol{C}_{-m}$, the samples in segment m must also be removed from $\boldsymbol{T}_A$ before comparing them. This shortened version of $[1 \quad \boldsymbol{T}_A]$ is denoted as $[1 \quad \boldsymbol{T}_A]_{\setminus m}$. Since the samples in segment m is now removed from both matrices, fewer degrees of freedom in $\boldsymbol{Y}$ is consumed than if e.g. the loading matrices were to be used as targets. Note that even if the samples in segment m are not used directly when estimating $\boldsymbol{C}_{-m}$, they are not completely left out since they have been influencing $\bar{\boldsymbol{Y}}$ and $\boldsymbol{W}$ in the total model.

In order to estimate the matrix $\boldsymbol{C}_{-m}$, the criteria to be minimised is the difference between $[1 \quad \boldsymbol{T}_A]_{\setminus m}$ from the total model and the rotated $[1 \quad \boldsymbol{T}_{-m,A}]$ from the reduced model. The difference is here denoted $\boldsymbol{G}_{-m,A}$:

$$[1 \quad \boldsymbol{T}_A]_{\setminus m} = [1 \quad \boldsymbol{T}_{-m,A}]\, \boldsymbol{C}_{-m} + \boldsymbol{G}_{-m,A} \tag{11}$$

There are many possible ways to estimate $\boldsymbol{C}_{-m}$ from equation (11). To reduce the degrees of freedom consumed in the rotation, we have chosen to use an orthogonal rotation, which means that the columns in $\boldsymbol{C}_{-m}$ are orthogonal with length one. The procedure for estimating $\boldsymbol{C}_{-m}$ starts with performing an SVD:

$$\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}' = [1 \quad \boldsymbol{T}_{-m,A}]' \; [1 \quad \boldsymbol{T}_A]_{\setminus m} \tag{12}$$

and then $\boldsymbol{C}_{-m}$ is estimated as

$$\boldsymbol{C}_{-m} = \boldsymbol{U}\boldsymbol{V}' \tag{13}$$

There are many possible ways to estimate $\boldsymbol{C}_{-m}$ from equation (11) (or even without using the scores matrices), the above is just one solution. Other possible procedures are discussed in section 3.2.

2.4.1 Rotating the scores

For each left-out segment $m = 1 \dots M$, we estimate $\boldsymbol{C}_{-m}$ using equation (13). With the appropriate matrix $\boldsymbol{C}_{-m}$, we can then calculate values for the rotated versions of the scores in each submodel.

Augmenting the submodel score matrix: Since the score-matrix of submodel $-m$ is calculated with the samples of segment m left out, we would only be able to re-estimate parts of the total score-matrix by rotating the scores from submodel $-m$. In order to fix this, we first insert estimated score values of the left-out samples set m into the score-matrix of submodel $-m$ before we rotate it. These estimated values are calculated in the usual way:

$$\widehat{\boldsymbol{T}}_{m,A} = (\boldsymbol{X}_m - 1\bar{\boldsymbol{x}}')\, \boldsymbol{W}_{-m,A} \left(\boldsymbol{P}'_{-m,A}\boldsymbol{W}_{-m,A}\right)^{-1} \tag{14}$$

By inserting these values into $\boldsymbol{T}_{-m,A}$ at the right positions, we can now calculate the full rotated score-matrix of submodel $-m$. We denote the rotated matrix with a tilde, and the augmented score-matrix from submodel $-m$ is denoted with a subscript $-m$, m.

$$\left[\widetilde{1} \quad \widetilde{\boldsymbol{T}}_{-m,A}\right] = [1 \quad \boldsymbol{T}_{-m,m,A}]\,\boldsymbol{C}_{-m} \tag{15}$$

Using the rotated versions of the score-matrix as calculated in equation (15), we can estimate the variance of each element in the same way we did for the regression coefficients in equation (5). For the elements of the score-matrix, the corresponding equation is

$$\hat{s}^2(t_{ia}) = \frac{M-1}{M}\sum_{m=1}^{M}\left(\widetilde{t}_{-m,ia} - t_{ia}\right)^2 \tag{16}$$

where t_{ia} is the score-value of sample i in factor a of the total model. This equation gives an estimate of the variance of the score-value for each sample in each factor. This can be used e.g. to draw approximate confidence-regions around each sample in the score-plot, and thus determine if two samples are far enough apart to be considered different. Such an approximate confidence-region could e.g. be created by using $\pm 2\hat{s}(t_{ia})$, but it is important to emphasise that the statistical properties of the variance estimate (16) is not known, and that the "confidence-region" should be regarded as approximate. (Further improvements might be attained by degrees-of-freedom correction to compensate for the estimation of the rotation parameters.)

The rotated score-values $\widetilde{t}_{-m,ia}$ are also interesting in themselves. By plotting these values together with the score-values from the total model in the score-plot, the user gets a visual image of the stability of each sample, and such plots are often referred to as *stability plots*. Samples that are outliers will tend to get a very different score-value when they are not used in the calibration, and thus will be easily visible in the stability plot.

2.4.2 Rotating the loadings The matrices of $\boldsymbol{X}$- and $\boldsymbol{Y}$-loadings for submodel $-m$ have the same dimensions as the loading-matrices of the full model. They can therefore be rotated without augmentation.

$$\begin{aligned}\left[\widetilde{\bar{\boldsymbol{x}}}_{-m} \quad \widetilde{\boldsymbol{P}}_{-m,A}\right] &= \left[\bar{\boldsymbol{x}}_{-m} \quad \boldsymbol{P}_{-m,A}\right]\boldsymbol{C}_{-m}^{-T} \\ \left[\widetilde{\bar{\boldsymbol{y}}}_{-m} \quad \widetilde{\boldsymbol{Q}}_{-m,A}\right] &= \left[\bar{\boldsymbol{y}}_{-m} \quad \boldsymbol{Q}_{-m,A}\right]\boldsymbol{C}_{-m}^{-T}\end{aligned} \tag{17}$$

In the present case, $\boldsymbol{C}_{-m}$ is an orthogonal matrix, and thus $\boldsymbol{C}_{-m}^{-T} = \boldsymbol{C}_{-m}$. The notation in equation (17) is general, and also valid for matrices with other properties. In the same way as with the scores, the variance of each element in the loading-matrices can now be estimated:

$$\begin{aligned} \hat{s}^2(p_{ka}) &= \frac{M-1}{M}\sum_{m=1}^{M}\left(\widetilde{p}_{-m,ka}-p_{ka}\right)^2 \\ \hat{s}^2(q_{ja}) &= \frac{M-1}{M}\sum_{m=1}^{M}\left(\widetilde{q}_{-m,ja}-q_{ja}\right)^2 \qquad (18) \end{aligned}$$

As with the score-values, these variances can be used to draw approximate confidence regions in the loading plot and determine whether or not two variables are overlapping and thus contains the same information.

2.4.3 Rotation of the loading weights Rotation of the loading weights $\boldsymbol{W}_A$ (7) is a little more complicated than rotation of scores and loadings.

The rotated version of the loading weights is proposed as:

$$\widetilde{\boldsymbol{W}}_{-m,A} = \left[\bar{\boldsymbol{x}}_{-m} \quad \boldsymbol{W}_{-m,A}\boldsymbol{W}'_{-m,A}\boldsymbol{P}_{-m,A}\right]\boldsymbol{C}_{-m}^{-T}\left[0 \quad \left(\boldsymbol{P}'_A\boldsymbol{W}_A\right)^{-1}\right] \qquad (19)$$

where the column of zeros is needed because the matrix $\boldsymbol{C}_{-m}$ was estimated from equation (12), where an extra column is appended.

Similar to the other model parameters, the variance of each element in the loading weight matrices can now be estimated as:

$$\hat{s}^2(w_{ka}) = \frac{M-1}{M}\sum_{m=1}^{M}\left(\widetilde{w}_{-m,ka}-w_{ka}\right)^2 \qquad (20)$$

Having variance estimates of the individual loading weights opens up a new possibility in variable selection. It will then be possible to do a significance test of each variable k in each factor a. Values w_{ka} that are not significantly different from zero, can be forced to zero after which the vector $\boldsymbol{w}_a$ is re-orthogonalised. This procedure will then yield variable selection where it is possible to remove variables only in some of the factors, while leaving them in for other factors.

As further factors are calculated and the information left in the dataset decreases, more and more variables will become insignificant with their corresponding loading weight set to zero. Finally, the loading vector $\boldsymbol{w}_a$ will be reduced to the zero-vector, and no further factors needs to be calculated. Thus, the procedure would yield automatic selection of the number of factors to calculate, with integrated variable selection. The automatic deletion of insignificant variables is expected to yield more stable models that are also easier to interpret due to the reduced number of variables in each factor.

3 Results and discussion

3.1 Jackknife and segmentation

To confirm that equation (4) gives consistent estimates of variance with $M-1$ degrees of freedom for different segmentation sizes, a Monte-Carlo simulation was carried out. The parameter of interest in the simulation was the variance of regression coefficients in a full-rank OLS solution to MLR regression, i.e. a bilinear PCCR or PLSR model with maximum possible number of factors.

A matrix $\boldsymbol{X}$ with 300 samples and 3 variables was drawn with random, evenly distributed values between 0 and 1. The regressand $\boldsymbol{y}$ was calculated from true regression coefficients $\boldsymbol{\beta} = [0\ 1\ 2]'$ and random noise $\boldsymbol{e}$ which was drawn from the distribution $N(0, 1^2)$.

The dataset was then split up in several different ways with M ranging from 2 to 300, corresponding to the extremes of splitting in two and leave-one-out. For each value of M, the regression coefficients $\boldsymbol{b}$ were estimated and the variance of the second element in $\boldsymbol{b}$ was estimated from equation (4). The whole procedure was then repeated 500 times with different noise $\boldsymbol{e}$ added each time.

Since the true variance of the added noise ($\boldsymbol{e}$) was known, it was possible to compare the jackknife-estimated values of $s^2(\boldsymbol{b})$ with the theoretically expected values. The theoretical variance of the regression coefficients from MLR (given that $\boldsymbol{X}$ is noise-free) is

$$\sigma^2(\boldsymbol{b}) = \operatorname{diag}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right)\sigma^2(y) \tag{21}$$

Figure 1 shows the jackknife-estimated variance of the regression coefficient (5) as a function of the number of segments M, together with the theoretically expected variance value (21).

As could be expected, one can see that the variance-estimate is more uncertain when it is based only on only a few segments. But as the number of segments increases, the variance-estimate stabilises towards the theoretical value, and its own variance gets smaller.

Given that $s^2(\boldsymbol{b})$ is the estimated jackknife-variance of $\boldsymbol{b}$ based on M segment "observations", and assuming the underlying distribution is normal with variance $\sigma^2(\boldsymbol{b})$, then

$$\chi^2 = \frac{(M-1)\,s^2(\boldsymbol{b})}{\sigma^2(\boldsymbol{b})} \tag{22}$$

is chi-square distributed with $\nu = M - 1$ degrees of freedom and a variance of 2ν. Reordering this, the variance of the variance-estimate $s^2(\boldsymbol{b})$ is

$$\sigma^2\left(s^2(\boldsymbol{b})\right) = \frac{2}{\nu}\,\sigma^4(\boldsymbol{b}) \tag{23}$$

where ν is the degrees of freedom in the estimate of $s^2(\boldsymbol{b})$. Since the variance of the regression coefficient were estimated a lot of times in the Monte-Carlo simulations, it was possible to estimate also the variance of the variance-estimate, $s^2\left(s^2(\boldsymbol{b})\right)$. If we then "guess" that the degrees of freedom in $s^2(\boldsymbol{b})$ is $\nu = M - 1$, we can plot the variance of our variance-estimate as a function of $2/(M-1)$. If $M - 1$ is the correct number of degrees of freedom, this should give a straight line with intercept zero and slope σ^4.

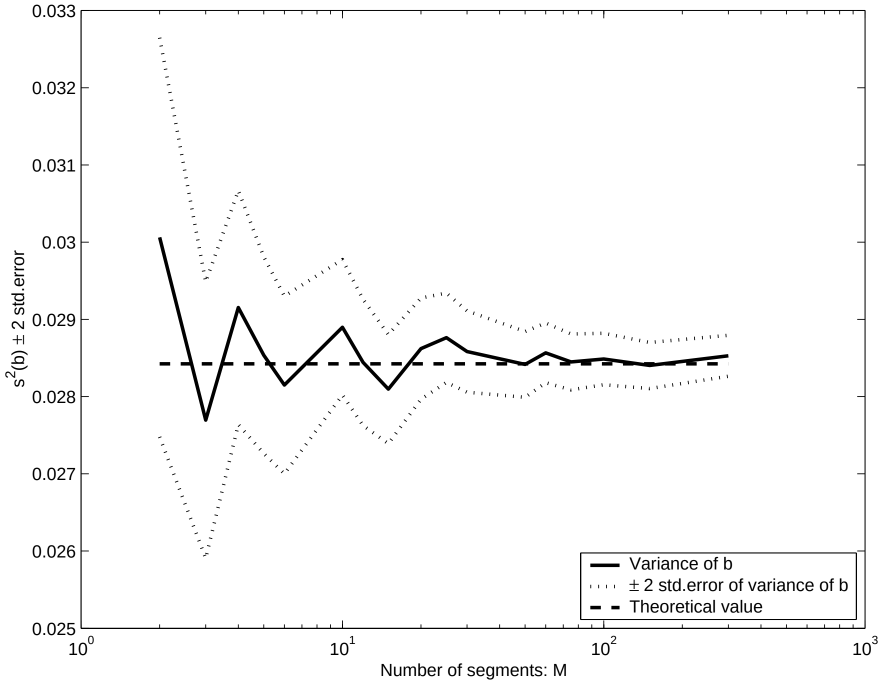


Figure 1: Variance of regression coefficient as a function of the number of segments.

As figure 2 shows, this is indeed the case. The above was also repeated with $\nu = M$ and $\nu = N$ (not shown here), but these (and other) alternatives gave a line with incorrect slope. Thus, we can conclude that equation (4) gives consistent estimates of the variance of $\boldsymbol{b}$ with $M-1$ degrees of freedom.

3.2 Alternative rotation schemes

The estimation of the orthogonal rotation matrix in equation (13) can be made even more conservative. A simpler matrix that only corrects for reflections and permutations can be calculated as

$$\boldsymbol{C}_{-m}^{\text{strict}} = \text{round}\left(\boldsymbol{C}_{-m}\right) \tag{24}$$

where the operator round () means rounding each element in $\boldsymbol{C}_{-m}$ towards the nearest integer; $-1, 0$ or 1. This approach would consume even fewer degrees of freedom than the orthogonal rotation in equation (13). When using the simple rounding procedure above, the norm of $\boldsymbol{C}_{-m}^{\text{strict}}$ must be monitored ($\boldsymbol{C}_{-m}^{\text{strict}}$ should have norm 1). If e.g. the angle between the submodel and the main model is around 45 degrees, the rounding can result in more that one element per row/column being different from zero.

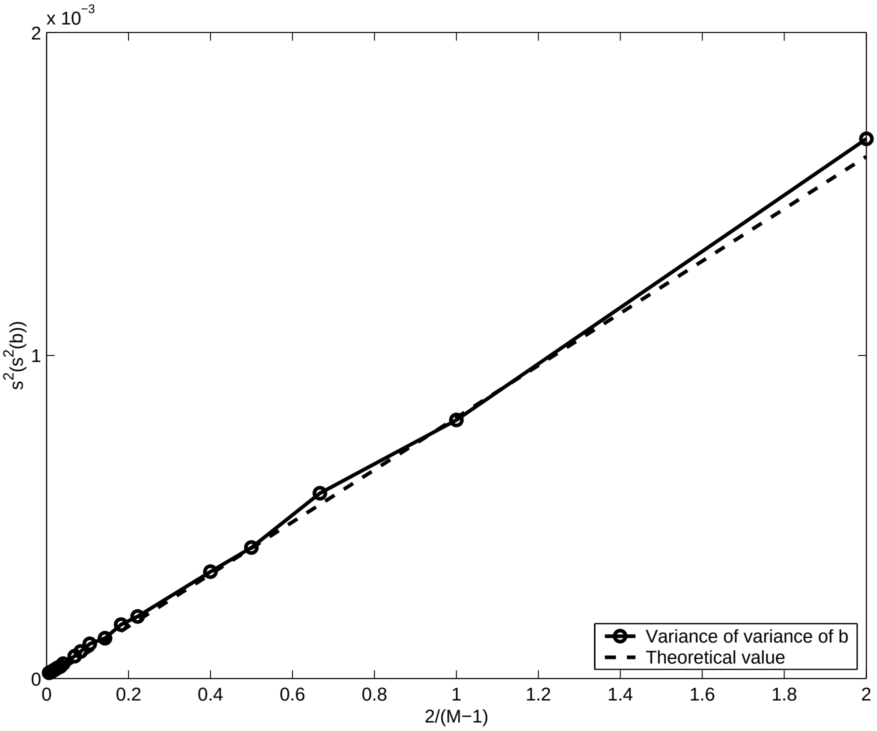


Figure 2: Variance of the variance of the regression coefficients as a function of $2/(M-1)$.

A procedure that solves this problem is to calculate the correlation between a factor in the total model and the factors in the submodel. If the highest absolute value is the diagonal element in the correlation matrix, then set the element in $\boldsymbol{C}_{-m}^{\text{strict}}$ to -1 or 1 depending on the sign of the correlation. All other elements for that factor are set to zero, both for the total model and the submodel elements. Thereafter, the highest absolute correlation for each total model with respect to the submodel is found and set to -1 or 1 in $\boldsymbol{C}_{-m}^{\text{strict}}$. This avoids that two factors in the submodel are assigned to the same factor in the total model, and yields a matrix $\boldsymbol{C}_{-m}^{\text{strict}}$ that is guaranteed to have norm one and only account for reflections and reorderings.

One could also envision other approaches that would consume more degrees of freedom. Starting from equation (11), the matrix $\boldsymbol{C}_{-m}$ could also be estimated by an OLS regression. This corresponds to projecting the total model onto the reduced model. A similar but more numerically stable approach would be to use a rank-reduced regression like PLSR instead of OLS regression in the estimation step.

3.3 The rank of the rotation

An important question regarding the rotation of submodels is *how many bilinear factors to use in the rotation?* The simulations performed suggests that the best solution is to perform the rotation of the models after A_{opt} factors have been calculated. This introduces a problem, as A_{opt} is not known *a priori* but typically estimated from a crossvalidated Root Mean Square Error of Prediction (RMSEP) curve showing estimated prediction errors in $\boldsymbol{Y}$. Thus, the current implementations starts with an ordinary PLSR in order to establish A_{opt}. Then, in a second step, the rotations are performed, and the variances are estimated.

Such a two-step procedure causes problems for the suggested "dynamic" variable selection scheme suggested in section 2.4.3. Further research in this area might solve this problem.

3.4 Large matrices with many predictor variables

If there are many predictor variables in the input data $\boldsymbol{Z}$, one might consider doing an SVD, $\boldsymbol{Z} = \boldsymbol{USV}'$, and then use the much smaller $\boldsymbol{X} = \boldsymbol{US}$ as input to the PLSR algorithm instead of $\boldsymbol{Z}$. This will greatly reduce the time consumed in the calibration, especially when doing leave-one-out crossvalidation. The variable dependent parameters from the PLSR (like $\boldsymbol{B}_A$, $\boldsymbol{P}_A$ and $\boldsymbol{W}_A$) will then have to be multiplied with $\boldsymbol{V}'$ in order to correspond to the original $\boldsymbol{X}$-variables. Since e.g. the regression coefficients are then rotated, it is necessary to estimate *covariance uncertainties* (not just variances), in order for the rotated uncertainty estimate to be applicable to the original variables. It appears that this can be done by modifying equation (5): Let $\boldsymbol{d}_{-m,A} = \boldsymbol{b}_{-m,A} - \boldsymbol{b}_A$. The covariance between the regression coefficients can then be calculated as

$$\operatorname{cov}(\boldsymbol{b}_A) = \frac{M-1}{M} \sum_{m=1}^{M} \boldsymbol{d}_{-m,A} \boldsymbol{d}'_{-m,A} \tag{25}$$

The diagonal of this covariance matrix contains the values calculated from equation (5). This covariance will be applicable to the regression coefficients $\boldsymbol{B}$ from the regression $\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{F}$. In order to be applicable to the regression with the large input matrix, $\boldsymbol{Y} = \boldsymbol{ZC} + \boldsymbol{F}$, the covariance matrix must be multiplied with $\boldsymbol{V}$:

$$\operatorname{cov}(\boldsymbol{c}_A) = \boldsymbol{V} \operatorname{cov} \boldsymbol{b}_A \boldsymbol{V}' \tag{26}$$

3.5 Examples of score-plot

A matrix of $\boldsymbol{X}$-data with seven samples and three variables were generated by sampling from a normal distribution. The $\boldsymbol{y}$-data were then calculated by multiplying $\boldsymbol{X}$ with some predefined regression coefficients, and adding

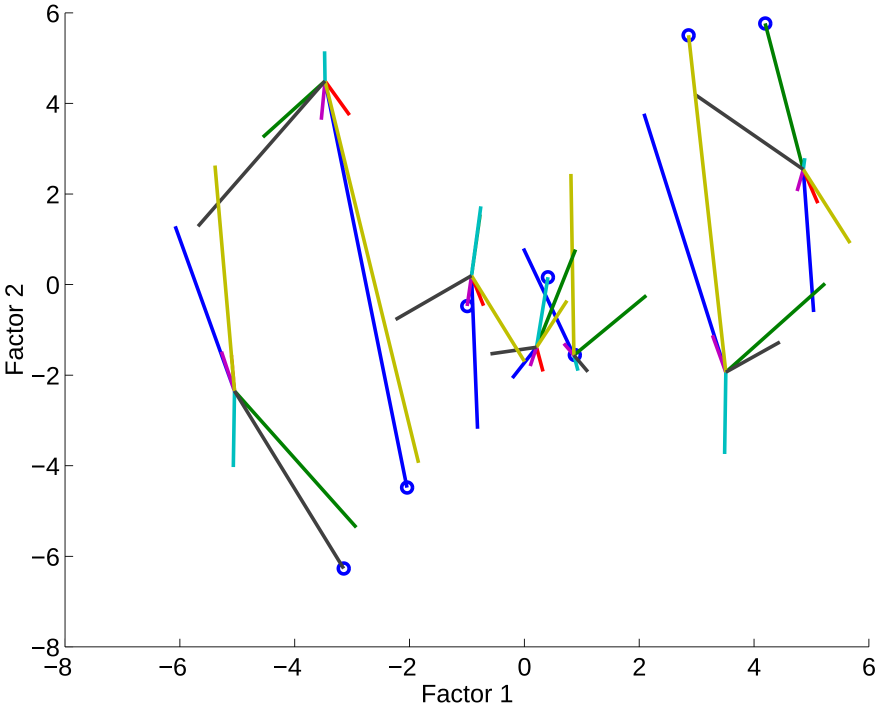


Figure 3: Original score plot with perturbations from leave one out jackknife. The centre of each "star" is the value from the complete model, and the circles denotes the value when that sample is kept out of the model calibration.

normal distributed noise with variance 0.1. Before subjecting these $\boldsymbol{X}$- and $\boldsymbol{y}$-data to a PLSR with full leave-one-out crossvalidation, normal distributed noise with variance 0.1 was also added to $\boldsymbol{X}$.

Figure 3 shows a score plot with all the values from each cross-validation segment, sometimes referred to as a *stability plot*. In the centre of each "star" are the score-values from the model calculated with all the samples. The lines going out from each "star" shows the score-value of that sample in each of the cross-validated models. The value with a circle on it denotes the value of that sample in the segment where the sample itself was left out, and thus had no influence on the model. Samples that are outliers will tend to get a very different score value when they are not included in the model, and thus the score value denoted with a circle will be further away from the centre than the other score values.

Note that several samples flip over, change sign or otherwise show large deviations that are not related to the uncertainty of the sample. This is due to the rotational freedom of bilinear models as described in the beginning of

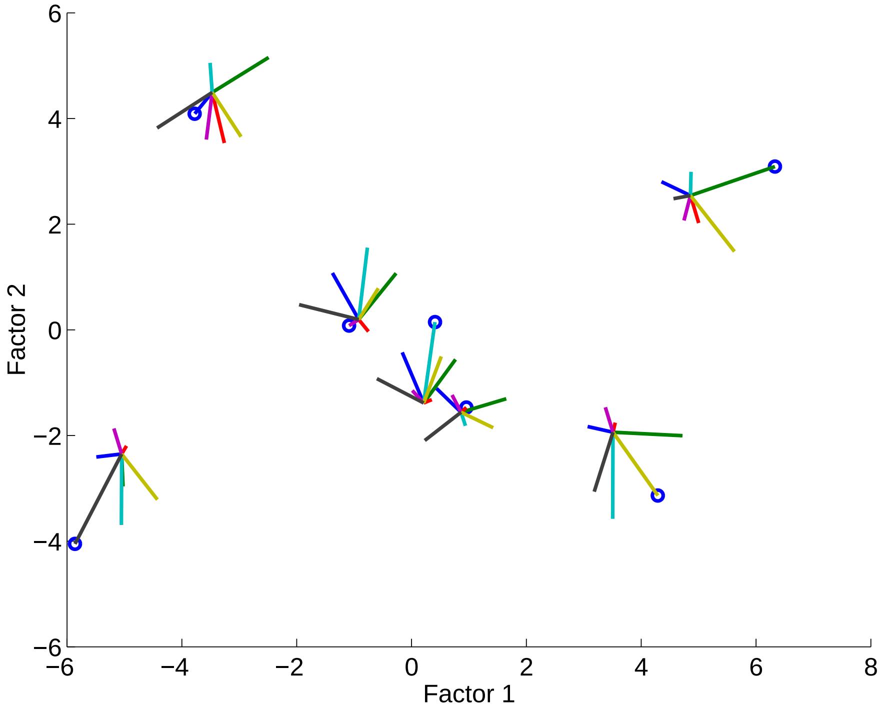


Figure 4: Rotated score plot. The centre of each "star" is the value from the complete model, and the circles denotes the value when that sample is kept out of the model calibration.

section 2.4. As a consequence, the variations between the values in figure 3 are unsuitable for calculating uncertainties.

Figure 4 shows the score plot after each of the submodels have been rotated as described in equation (15). The picture is now much clearer, and the variance left can be assumed to be due to the uncertainty of the score-values. Note that for each sample there can be a quite large difference between the mean of all the obtained values and the value from the total model. This is the rationale for choosing the total model as the reference value and not the mean (c.f. the discussion after equation (5)).

4 Conclusion

An improvement of the jackknife rotation method by Martens & Martens [3] has been proposed for estimating the uncertainty in the bilinear model parameters with the use of jackknife. The method works by rotating each of the submodels towards the main model before the values are used to estimate variances. The rotation matrix can be estimated in several ways, and some of the alternatives are discussed.

Further research is needed to establish the statistical properties of the obtained variance estimates, and alternative procedures for estimating the rotation matrix should be compared.

References

[1] Stone M. (1974). *Cross-validatory choice and assessment of statistical prediction.* J. Roy. Stat. Soc. B Met.**36** (1), 111 – 147.

[2] Efron B. (1982). *The Jackknife, the Bootstrap, and other resampling plans.* CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

[3] Martens H., Martens M. (2000). *Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR).* Food Qual. Prefer **11** (1), 5 – 16.

[4] Tukey J.W. (1958) *Bias and confidence in not quite large samples.* Ann. Math. Stat. **29**, 614.

[5] Shao J. Wu C.F.J. (1989). *A general theory for jackknife variance estimation.* Ann. Stat. **17** (3), 1176 – 1197.

[6] Efron B. Tibshirani R.J. (1998). *An introduction to the Bootstrap.* Chapman & Hall, New York.

[7] Martens H., Høy M., Westad F., Folkenberg D., Martens M. (2001). *Analysis of designed experiments by stabilised PLS Regression and jack-knifing.* Chemometr. Intell. Lab. **58** (2), 151 – 170.

[8] Martens H., Martens M. (2001). *Multivariate Analysis of Quality. An Introduction.* J.Wiley & Sons Ltd, Chichester UK.

Address: M. Høy, Norwegian Meteorological Institute, Pb 43 Blindern,
N-0313, Norway
F. Westad, Matforsk, Osloveien 1, N-1430 Ås, Norway
H. Martens, CIGENE, Norwegian Agricultural University, N-1432 Ås,
Norway

E-mail: `martin.hoy@pvv.ntnu.no`

COMPSTAT'2004 Symposium

LINE MOSAIC PLOT: ALGORITHM AND IMPLEMENTATION

Moon Yul Huh

Key words: Mosaic plot, line mosaic plot, statistical graphics, visual inference, statistical algorithms.

COMPSTAT 2004 section: E-statistics.

Abstract: Conventional mosaic plot is to graphically represent contingency tables by tiles whose size is proportional to the cell count. The plot is informative when we are well trained reading this. This paper introduces a new approach for mosaic plot called line mosaic plot which uses lines instead of tiles to represent the size of the cells in contingency tables. We also give a general straightforward algorithm to construct the plot directly from the data set while the conventional approach is to construct the plot from the cross tabulation. We demonstrate the effectiveness of this tool for visual inference using a real data set.

1 Introduction

Mosaic display introduced by Hartigan and Kleiner [6] has been generalized to multi-way tables and has been extensively worked for visual inference of independence using Mosaic plots by Friendly [4], [5]. Meyer and et. al. [11] considered visual inference of contingency tables using association plots mainly for the case of 2-way tables. Another sources for the works of Mosaic Plots are Hofmann [7], [8] and Unwin [12]. Most of the statistical packages available today have implemented mosaic displays (SAS, S-Plus, R, Minitab, and others).

Conventional mosaic plot is to graphically represent contingency tables using tiles whose size is proportional to the cell count. Figure 1 gives the mosaic plot of the Titanic data [3] as implemented in R [10]. This data will be explained in more detail in the next section. The plot is informative when we are well trained in reading this. Our experiments with the graduate students showed that the features in the mosaic plot is confusing and misleading if more than 2 variables are involved in the plot. The reason behind this could be due to the limitation of human perception. Firstly, this could be explained by the Steven's law of dimensionality. Steven's law states that perceived scale in absolute measurements is the actual scale raised to a power where the scale is as follows: for linear features, power is .9-1.1; for area feature, .6-.9; for volume, .5-.8.

Steven's Law suggests that physical relationships that are not represented in linear features can be grossly misperceived. For example, a lake represented on a map with an area graphically 10 times larger than another will be perceived as only 5 times larger as noted in Catarci, et. al. [1]. Since the

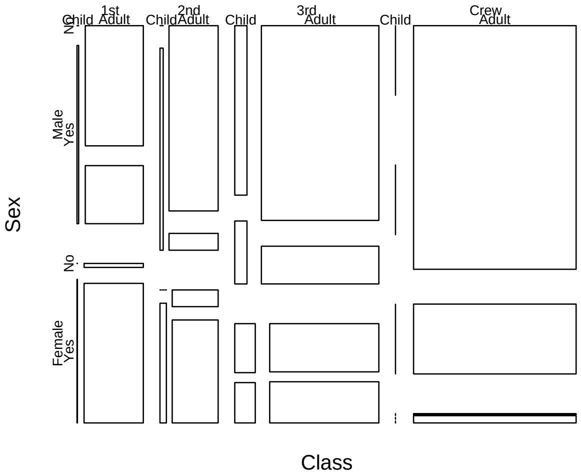


Figure 1: Conventional mosaic plot of Titanic Data using R.

mosaic plot presents all the features using 2-dimensional bars, the perceived scale of the features may be underestimated according to the law.

Secondly, the misperception of mosaic plot could be due to the fact that the columns and rows of the bars of the plot are not aligned, and make "errors in perception" as explained by Cleveland and McGill [2]. They state that the errors in perception from the graphs are in the following order.

- Position along identical, non-aligned scales.
- Length.
- Angle/Slope (though error depends greatly on orientation and type).
- Area.
- Volume.
- Color Hue, Saturation, Density (only informal testing).

Above observations suggest us to use lines instead of bars to represent the cell sizes in contingency tables and to plot the lines along a common aligned scales. Figure 2 gives the 'line' mosaic plot for the Titanic data. Details of the construction and interpretation of this plot will be given in the next section. In line mosaic plot, each cell of the contingency table is given equal sized rectangle, and the size of the frequency of each cell is represented using the total length of the lines drawn inside the rectangle. All the rectangular boxes are aligned horizontally and vertically so that the comparison of the relative size of the lengths in each rectangle can be perceived more easily.

In section 2, we give the algorithm for the line mosaic plot. In section 3, we present the implementation of the algorithm, and demonstrate the usefulness of the plot using a real data set.

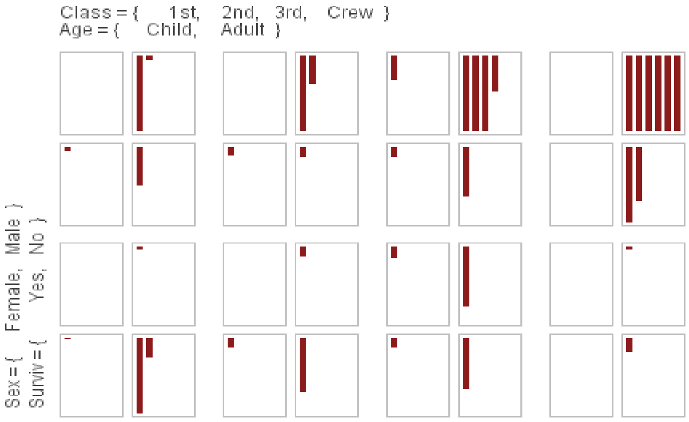


Figure 2: Line mosaic plot of Titanic Data.

2 Algorithm for mosaic array

Algorithms to generate mosaic plots have been approached in two ways as far as the author is aware of at the present time. The first approach is to construct the plot for a specific setting. In other words, suggested algorithm is to build the plot for the contingency table of a specific dimension, and apply similar method for other dimensions. Wang [13] and Friendly [4] give algorithms for 4-dimensions. Second approach is to use recursive structure as implemented in R [10]. To use these algorithms, we need contingency tables.

In this paper, we suggest a simple straightforward algorithm to construct the line mosaic plot directly from the data set. Figure 2 suggests us that line mosaic plot is simply a 2 dimensional array of the frequencies, what we call mosaic array. Mosaic array is the basic building block for our work, and in the next section, we give an algorithm to construct this array directly from the data set. Also, the algorithm for the the converse operation, or constructing data from mosaic array is given. When the problem considered is supervised learning, or when there is the target variable, mosaic array will be 3-dimensional. 3rd dimension corresponds to the target variable, and the number of levels of this dimension will be equal to the number of categories of the target variable. The construction of this case will become clear in section 3 where we give the implementation of the line mosaic plot.

We assume that all the variables are discrete, and let p be the number of variables, and $\mathbf{n}' = (n_1, \ldots, n_p)$ be the vector of the values that each variable can adopt, or number of categories for each categorical variable. Without loss of generality, we can assume that the values of each variable are transformed into integer values starting from 1. For example, the values of the variable *sex* will be $1, 2$. Also, let X be the data matrix of dimension $n \times p$ where

n is the number of observations. For convenience and for the simplicity of notation, let $\mathbf{v}$ be the $p-$ length vector denoting an observation or an instance from the data matrix X. Using this notation, we can write $v_j, j = 1, \ldots, p$ as a realization of the j^{th} variable of an instance from the data matrix X. We finally assume that the variables are ordered according to some measure of importance for mosaic plot. Hence, the first variable will be the first choice, the second one is next choice, and so on for the mosaic plot.

We now build a 2-dimensional mosaic array F which is a representation of multidimensional cross table form for the data matrix X, or the array of the form of Figure 2. The size of F will be $\Pi_{i=1}^{[p/2]} n_{2i}$ and $\Pi_{i=0}^{[(p-1)/2]} n_{2i+1}$ for row and column respectively.

An instance of X, which is denoted as $\mathbf{v}$ in the above, will add 1 to the cell $F_{I,J}$, where I and J will be determined as follows.

$$I = \Sigma_{i=1}^{[p/2]}(v_{2i} - 1)\Pi_{j=i+1}^{[p/2]} n_{2j} + v_{2[p/2]}$$

$$J = \Sigma_{i=0}^{[(p-1)/2]}(v_{2i+1} - 1)\Pi_{j=i+1}^{[(p-1)/2]} n_{2j+1} + v_{2[(p-1)/2]+1}$$

where $[x]$ denotes the integer not exceeding x.

We now give the algorithm to construct the values of the variables, $\mathbf{v}$, when an instance belongs to a cell $F(I, J)$. ¿From the row index I, variables of even indices, or $v_2, v_4, \ldots, v_{2[p/2]}$ will be constructed, and from the column index J, variables of odd indices, or $v_1, v_3, \ldots, v_{2[(p-1)/2]+1}$ will be constructed. The algorithm follows.

Values of odd indices, $v_1, v_3, \ldots, v_{2[(p-1)/2]+1}$ from J are:

$$v_i = \begin{cases} 1 + [\frac{J-1}{\Pi_{j=[(i-1)/2]+1}^{[(p-1)/2]} n_{2j+1}}], & i = 1, 3, \ldots, 2[(p-1)/2] - 1 \\ 1 + Mod\ (J-1,\ n_{2[(p-1)/2]+1}), & i = 2[(p-1)/2] + 1 \end{cases}$$

Values of even indices, $v_2, v_4, \ldots, v_{2[p/2]}$ from I are:

$$v_i = \begin{cases} 1 + [\frac{I}{\Pi_{j=[i/2]+1}^{[p/2]} n_{2j}}], & i = 2, 4, \ldots, 2[p/2] - 1 \\ 1 + Mod\ (I-1,\ n_{2[p/2]}), & i = 2[p/2] \end{cases}$$

where $Mod(x, y) = x - x \times [x/y]$. We have shown above algorithmically, a unique F is constructed for a given data set. Now, to draw a mosaic plot, we need to construct $||F||$ rectangles in total, where $||F||$ denotes the number of the cells that F makes, or is equal to $\Pi_{i=1}^{[p/2]} n_{2i} \times \Pi_{i=0}^{[(p-1)/2]} n_{2i+1}$. The rectangles are separated by some gaps between them, and it is conventional to leave larger gaps for the variables with higher hierarchy. Our implementation for the construction of the rectangles and the gaps between them are given in the following section.

To complete the algorithm, we need to consider several details. At first, we need to standardize mosaic array F according to some criterion. We can consider several options for standardization. In this work, we standardize each cell with respect to the maximum cell frequency, or we use $F(I, J)/max_{I,J}F(I, J)$. Secondly, we need to set some gaps between the rectangles, so that the plot is easier to perceive. An option for this is suggested in Friendly [4]. In this work, we apply the following method.

For horizontal direction, there will be $\Pi_{i=0}^{[(p-1)/2]} n_{2i+1} - 1$ gaps between the rectangles, and for vertical direction, there will be $\Pi_{i=1}^{[p/2]} n_{2i} - 1$ gaps between the rectangles. To implement the horizontal gaps, we leave 1 unit space between the rectangles for the lowest hierarchy, 2 'unit' space for the next hierarchy, ..., $[\frac{p+1}{2}]$ unit space for the highest hierarchy. Here, 'unit' is arbitrary. We may set 5 pixels, for example, for the unit space. For the column, we leave 0.5 unit space between the rectangles for the lowest hierarchy, 1.5 unit space for the next hierarchy, ..., $[p/2] - 0.5$ unit space for the highest hierarchy. An algorithm for the gaps is given in Figure 3.

```
For row bars:
  • Let $G \leftarrow \Pi_{i=0}^{[(p-1)/2]} n_{2i+1} - 1$, which is the total number of gaps.
  • Let the $i^{th}$ gap $g_i = 1$, for $i = 1, \ldots, G$. Let the number of variables for
    the row bars $m = [(p-1)/2]$, and initialize $d$ to 1.
  • if $(m == 1)$ break;
  • for $i = m, \ldots, 1$, step $-1$ {
    $d = d * n_{2i+1}$;
    for $j = d, \ldots, G$, step d{
    $g_j + +$;
    }}

For column bars:
  • Let $G \leftarrow \Pi_{i=1}^{[p/2]} n_{2i} - 1$, which is the total number of gaps.
  • Let $g_i = 1$, for the $i - th$ gap where $i = 1, \ldots, G$. Let the number of
    variables for the column bars $m = [p/2]$, and initialize $d$ as 1.
  • if $(m == 1)$ break;
  • for $i = m, \ldots, 1$, step $-1$ {
    $d = d * n_{2i}$;
    for $j = d, \ldots, G$, step d{
    $g_j + +$;
    }}
```

Figure 3: Algorithm for the gaps between the rectangular bars.

The above procedure works for unsupervised learning. With supervised learning, we have target variable. We assume here that the last variable, or variable p is for the target. In this case, we build F with $p-1$ variables. Frequencies in the cell (I,J), or $F(I,J)$ will be divided into n_p different categories. In this case, it will be convenient to express F in 3 dimensional form such that $F(I,J,K), K=1,\ldots,n_p$.

3 Implementation and demonstration of the line mosaic plot

We illustrate the implementation of line mosaic plot using Titanic data introduced by Dawson (1995, http://ssi.umh.ac.be/titanic.html) goes as follows. Titanic data consists of 2201 cases and 4 variables {*Class, Gender, Age* and *Survival*}. The values of each variables are: *Class*={*1st, 2nd, 3rd, crew*}*; Gender*={*male, female*}*; Age* ={*adult, child*}*; Survived* ={*yes, no*}. Hence, $p=4$, $\mathbf{n}'=(4,2,2,2)$. When a case *(1st, adult, male, yes)* is given, $\mathbf{v}'=(1,1,1,1)$, and the above algorithm give $\{I=1,\ J=1\}$. When a case is *(crew, male, child, no)*, F of Titanic data. Mosaic array F of Titanic data is given in Table 1.

57	5	14	11	75	13	192	0
118	0	154	0	387	35	670	0
140	1	80	13	76	14	20	0
4	0	13	0	89	17	3	0

Table 1: Mosaic array F of Titanic data.

Table 2 gives the mosaic array F for the Titanic data when *survive* is target variable. Implementation of this mosaic plot can be accomplished by assigning different colors for different categories. For Titanic data, we may assign *survived* as the target variable. Conventional mosaic plot and line mosaic plot of the Titanic data for this case is given in Figure 4 and Figure 5 respectively.

when $survive = yes$, or $k=1$

57	5	14	11	75	13	192	0
140	1	80	13	76	14	20	0

when $survive = no$, or $k=2$

118	0	154	0	387	35	670	0
4	0	13	0	89	17	3	0

Table 2: Mosaic array F of Titanic data with *survive* as the target variable.

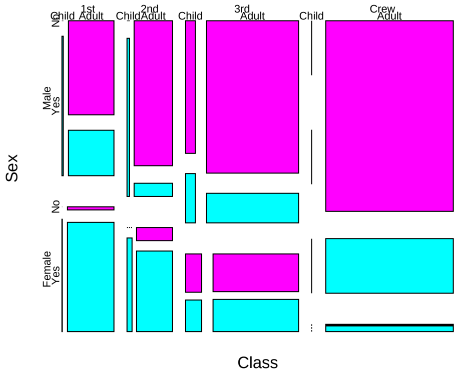


Figure 4: Mosaic plot of Titanic Data when *survived* is target variable.

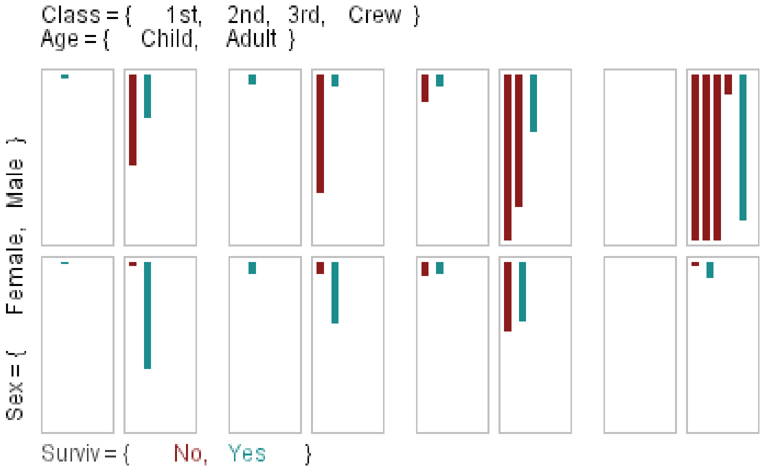
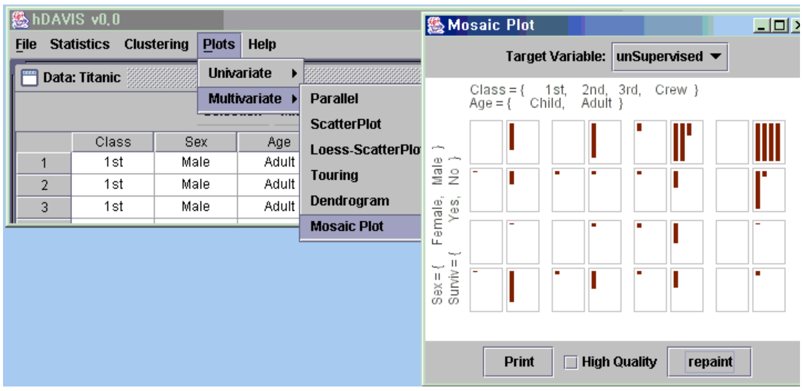


Figure 5: Line mosaic Plot of Titanic Data when *survived* is target variable.

From Figure 5, it is easy to see that most of the passengers are males, and there are very few children passengers. The largest number of passenger groups are crews, then 3*rd* class, 2*nd* class, and 1*st* class passengers are the fewest. In gender-wise, there are very few female crews, and largest class group for females is seen to be 3*rd* class, then 1*st*, and then 2*nd* class. We can visually estimate that the number of female 3*rd* group passengers is about twice the number of female 2*nd* class passengers. Turning our attention to *survive*, it is straightforward to observe that most of the *3rd* and *crew* class passengers could not survive, but most of the 1*st* and 2*nd* class female passengers survived. The proportion of survivals in the *3rd* and *crew* classes can even be estimated visually by reading the number of bars in the plot. For {*crew, adult, male*} combination, the proportion can be estimated as 2/7. For {*3rd, adult, male*} combination, the proportion is less than 1/4. For the female case, we can observe directly from the plot that the survival proportion is much higher except for the {*3rd, adult*} combination. Although there are few *child* passengers, the plot clearly shows that most of the children passengers survived except for the *3rd* class cases.

Figure 6 gives the process of obtaining a line mosaic as implemented in hDAVIS [9]. hDAVIS is freely available on the following website. `http://stat.skku.ac.kr/~myhuh/davis.html.`



	Class	Sex	Age
1	1st	Male	Adult
2	1st	Male	Adult
3	1st	Male	Adult

Target Variable: unSupervised

Figure 6: Line mosaic plot implemented in DAVIS.

References

[1] Catarci T., D'Amore F., Janecek P., Spaccapietra S. (2001). *Interacting with GIS: from paper cartography to virtual environments.* Unesco Encyclopedia on man-machine Interfaces, Advanced Geographic Information Systems, Unesco Press.

[2] Cleveland W.S., McGill R. (1985). *Graphical perception and graphical methods for analyzing scientific data.* Science **229**, 828 – 833.

[3] Dawson R.J.M. (1995). *The "unusual episode" data revisited.* J. Statistics Education **3** (3), 1 – 7.

[4] Friendly M. (1994). *Mosaic displays for multi-way contingency tables.* Journal of the American Statistical Association, **89**, 190 – 200.

[5] Friendly M. (1999). *Extending mosaic displays: marginal, partial, and conditional views of categorical data.* Journal of Computational and Graphical Statistics **8**, 373 – 395.

[6] Hartigan J.A., Kleiner B. (1981). *Mosaics for contingency tables.* Eddy, W. F., (ed.), Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, 268 – 273. Springer-Verlag, New York, NY.

[7] Hofmann H. (2000). *Exploring categorical data: interactive mosaic plots.* Metrika **51** (1), 11 – 26.

[8] Hofmann H. (2003). *Constructing and reading mosaic plots.* Computational Statistics & Data Analysis **43**, 565 – 580.

[9] Huh M.Y., Song K.R. (2002). *DAVIS: A Java-based data visualization system.* Computational Statistics, **17** (3), 411 – 423.

[10] Ihaka R., Gentleman R. (1996). *R: A language for data analysis and graphics.* Journal of Computational and Graphical Statistics **5**, 299 – 314.

[11] Meyer D., Zeileis A., Hornik K.(2003). *Visualizing independence using extended association and mosaic plots.* DSC 2003 Working Paper, Institut for Statistik & Wahrscheinlichkeitstheorie, Technische University at Wien, Institut for Statistik, Wirtschaftsuniversity at Wien.

[12] Unwin A. (2003). *Variations on mosaic plots.* Workshop on Modern Statistical Visualization and related topics(1) at ISM on 13-14, November 2003, ISM, Tokyo, Japan.

[13] Wang C.M. (1985). *Applications and computing of mosaics.* Computational Statistics & Data Analysis **3**, 89 – 97.

Acknowledgement: This work was supported by the Samsung Research Fund (2003) of Sungkyunkwan University.

Address: M.Y. Huh, Department of Statistics, Sungkyunkwan University, Chongro-Ku, Seoul, Korea

E-mail: `email: myhuh@skku.ac.kr`

GRAPHICAL DISPLAYS OF INTERNET TRAFFIC DATA

Karen Kafadar and Edward J. Wegman

Key words: Logarithmic transformation, computational methods, recursive computation, graphical displays, exploratory data analysis.

COMPSTAT 2004 section: Data visualisation.

Abstract: The threat of cyber attacks motivates the need to monitor Internet traffic data for potentially abnormal behavior. Due to the enormous volumes of such data, statistical process monitoring tools, such as those used traditionally on data in the product manufacturing departments, are inadequate. The detection of "exotic" data, which may indicate a potential attack, requires a characterization of "typical" behavior. We propose some simple graphical tools that permit ready visual identification of unusual Internet traffic patterns in "streaming" data. These methods are illustrated on a moderate-sized data set (135,605 records) collected at George Mason University.

1 Introduction

Cyber attacks on computer networks or personal computers have become major threats to nearly all operations in society. Methods to thwart such attacks are seriously needed. The problem of detecting unusual behavior in data streams occurs in many fields, such as in disease surveillance, nuclear product manufacturing, and phone and credit card use. Historically, manufacturing and financial industries have relied on conventional statistical process monitoring tools, such as control charts and process flow diagrams. Such tools are reliable and appropriate, because the data streams can be stratified into reasonably independent series. For example, monitoring a customer's credit card use relies on an analysis of the data from the customer's past charging amounts and frequencies. This data stream is a much smaller data set than the entire database, with events occurring irregularly but not frequently; moreover, one customer's data stream can be considered as independent of other customers' data streams. In contrast, Internet traffic data are virtually continuous (limited only by the resolution of the time clock that captures them), and the data for one system involve hundreds of thousands of other computer or network systems.

Tools for monitoring such data are essential. Conventional statistical analysis often assumes that data follow a mathematically tractable probability distribution function and will yield valid estimates of the parameters of this distribution. Such approaches cannot be used on millions of data points. Graphical tools for streaming data offer hope of identifying potential cyber-attacks, particularly when the tools are tailored for the application. Features of Internet traffic data are described in Section 2.

Even with novel graphical displays for massive data streams, however, a characterization of "typical" behavior is still needed, so relevant graphical tools can be made more sensitive to capturing exotic or abnormal patterns. Two approaches to the detection problem through visualization are discussed in this article. Section 3 describes a "drill-down" approach to viewing large data sets, illustrated on a data set of 135,605 records collected over a one-hour period at George Mason University. Section 4 describes a second approach, "evolutionary graphical displays", which present the data only within a narrow time window (e.g., 10 minutes); early data disappear as new, more recent data, come into view. Two examples are "waterfall diagram" and "skyline display. Section 5 offers a summary and proposals for further work.

2 Features of Internet traffic data

To monitor Internet traffic data for potential attacks, organizations will install anonymous surveillance machines outside a "firewall" to monitor incoming and outgoing traffic. For a discussion of the types of programs that monitor traffic flow, see Marchette [1, Ch. 4]. Data collected during an Internet session includes many features; key features include source and destination addresses, source and destination ports, and measures of size and duration of the session.

IP addresses

Internet traffic proceeds from one machine to another, using a protocol for data transfer known as Internet Protocol (IP), which directs the transmission of data among machines during an Internet session. The "IP header" contains several important pieces of information. Since each IP address is a 32-bit number represented in four 8-bit fields (e.g., 127.0.0.1), 2^{32} = 4,294,967,296 machines can be addressed. Multiplied by the volume of traffic during a given day, conventional static graphs cannot display such tremendous volumes of data on a system with finite resolution. The IP header captures the two addressable machines involved in an Internet session.

Transmission Control Protocol

A common communication protocol is Transmission Control Protocol (TCP). TCP implements a two-way connection between machines and contains the necessary instructions for delivering and sequencing packets. The instructions are captured in a file whose header includes the source and destination port numbers, useful for monitoring traffic flow and detecting potential attacks.

Each host machine has 2^{16} = 65,536 ports, divided into three ranges. The first range includes 1024 (2^{10}) "well-known ports" numbered 0 to 1023; for example, file transfer protocol (ftp) uses port 21; secure shell (ssh) uses port 22; telnet uses port 23; smtp mail operates from port 25; web service (http) operates from port 80; pop3 mail operates from port 110; secure web encryption (https) operates from port 443; real time stream control protocol (rtsp) uses port 554 for quick-time streaming movies. The second range consists of registered ports, numbered 1024 to 49151; for example, Sun has registered port 2049 for its network file system (nfs). The remaining 16384 (2^{14}) ports, numbered 49152 to 65536, are dynamic or private ports. Unprotected ports (source ports or destination ports) are prime candidates for intrusion; too much traffic on a given port within a short time frame may indicate a potential attack. In this data set, all ports numbered 10000 or above were coded simply as “port 10000”.

Size of session

Internet traffic data are sent in “packets”. The “size” of an Internet session can be measured in several ways: duration (e.g., number of seconds), number of packets, and number of bytes. Typically, these numbers will be correlated, but not in any specific deterministic way. However, a machine may send many packets with few bytes, or rather fewer full-sized packets; either situation may signal a potential attack on a system.

Sample data

Internet traffic data are being collected at George Mason University; a sample of ten records from a data set over the course of one hour is shown in Table 1. Column 1 labeled `time` denotes the clock time (in number of seconds from an origin) at which the Internet session began; `duration` or `len` represents the duration or length of the session in seconds; `SIP` and `DIP` are the source and destination ports, respectively; `DPort` and `SPort` are the destination and source port numbers, respectively; and `Npacket` and `Nbyte` indicate the number of packets and number of bytes transferred in the session. In the plots below, the variable `time` is shifted by 39603 seconds and scaled by 1/60, so that the first session starts at 0.01067 minutes past the start of the hour, and the last session starts at 59.971 minutes past the start of the hour. Table 2 summarizes the distribution of the values in each column with the five-number summary [4] supplemented with the 10^{th} and 90^{th} percentiles for each column (minimum, lower 10%, lower fourth, median, upper fourth, upper 10%, maximum). The “size” variables are all very highly skewed towards the upper end; the distance between the 90th percentile and the maximum is 2–3 orders of magnitude greater than the distance from the 90th percentile to the minimum. One session involved over 35 million bytes, and almost 66,000 packets, although sessions of 1,832 bytes and 12 packets were more typical. The next section provides some displays of these data, with the objective of trying to characterize “typical” behavior, so that “atypical” behavior can be noted more readily.

3 Viewing Internet traffic data

Most features collected on Internet traffic data are highly skewed, as seen for the size variables. Thus, a plot of any pair of these variables has a very high density of points in the first quadrant near the origin. By selectively

	time	duration	SIP	DIP	DPort	SPort	Npacket	Nbyte
1	39603.64	0.23	4367	54985	443	1631	9	3211
2	39603.64	0.27	18146	9675	3921	25	15	49
3	39603.65	0.04	18208	28256	1255	80	6	373
4	39603.65	1389.10	24159	17171	23	1288	845	5906
5	39603.65	373.99	60315	37727	2073	80	1759	834778
6	39603.65	0.13	28256	18208	80	1256	10	816
7	39603.65	1498.11	25699	4837	9593	80	65803	35661821
8	39603.65	0.04	18208	28256	1251	80	5	373
9	39603.66	122.38	54985	4179	1298	443	99	85559
10	39603.66	0.13	28256	18208	80	1257	10	816

Table 1: Sample of Internet traffic data from George Mason University.

	time	duration	SIP	DIP	DPort	SPort	Npacket	Nbyte
minimum	39603.64	0.00	259	259	20	20	2	0
lower 10%	39937.68	0.20	4930	4024	80	1187	9	568
lower 4th	40507.09	0.32	9765	8705	80	1369	10	860
median	41435.55	0.58	20258	25164	80	1849	12	1832
upper 4th	42326.46	3.77	41282	45900	80	3681	21	7697
upper 10%	42857.49	21.45	62754	58202	80	10000	45	25161
maximum	43201.26	3482.50	65276	65262	10000	10000	65803	35661821
#(unique values)	104268	9101	2504	5139	380	6742	1056	29876

Table 2: Summary statistics from Internet traffic data set (135,605 sessions).

"zooming in", or "drilling down" into this region, as one does on a geographical map, specific features can be better observed. An alternative to this "drill-down" approach (steps of power magnification) is a logarithmic transformation, which allows one to view the points by scanning *across* the screen rather than by magnifiying regions of the space. We describe this approach below.

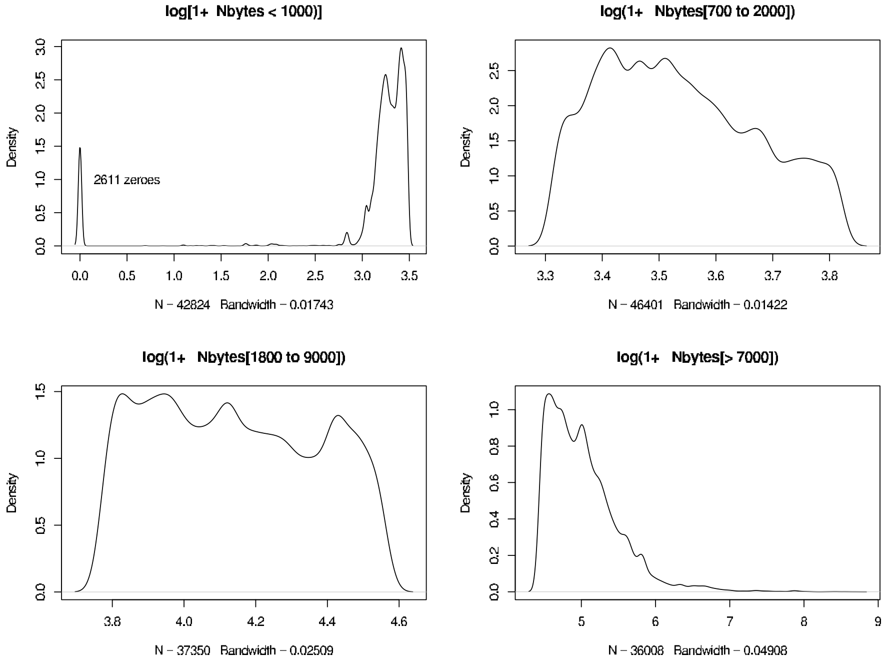


Figure 1: Kernel density estimates of $\log(1+\sqrt{Nbyte})$, four separate ranges.

Density plots

Figure 1 is a kernel density estimate [3] of `log.Nbyte` = `Nbyte*` = f(`Nbyte`), where $f(x) = \log(1+\sqrt{x})$. We use the transformation $f(x) = \log(1+\sqrt{x})$ for all three size variables to spread out their values (values of x near the low end of the scale are not spread out as far as they would be with the simple $\log(x)$ transformation; $f'(x) < 1/x$, much more so for small x). Likewise, `log.len` = f(`duration`) and `log.pkt` = f(`Npacket`). All calculations and graphs are made using the open-source software R, available from `http://www.cran.r-project.org`. A small peak at 0 reflects 2611 zeroes; the next largest byte size is 147. The data are clearly skewed, and local peaks of high density appear where `log.byte` $\approx$ 3.4, 3.8, 4.1, 4.5, and 5.1 (`Nbyte` $\approx$ 840, 1400, 3500, 8000, 26000).

Distributions of session size variables

Boxplots can be useful to display the relationship between two variables, as in Figure 2 for the two variables `log.len` = f(`duration`) (y-axis) and `log.byte` = f(`Nbyte`). The first box contains the 2911 values for which `Nbyte` is zero; the second box contains the next 1216 values where `Nbyte` ranges from 1 to 365 ($0 <$ `log.byte` ≤ 3); subsequent bins are 0.1 wide, except the last five bins. This display shows a relatively stable trend up until the last few bins, but is otherwise not very useful for outlier detection, since outliers are prevalent in each bin. The boxplot display does confirm general

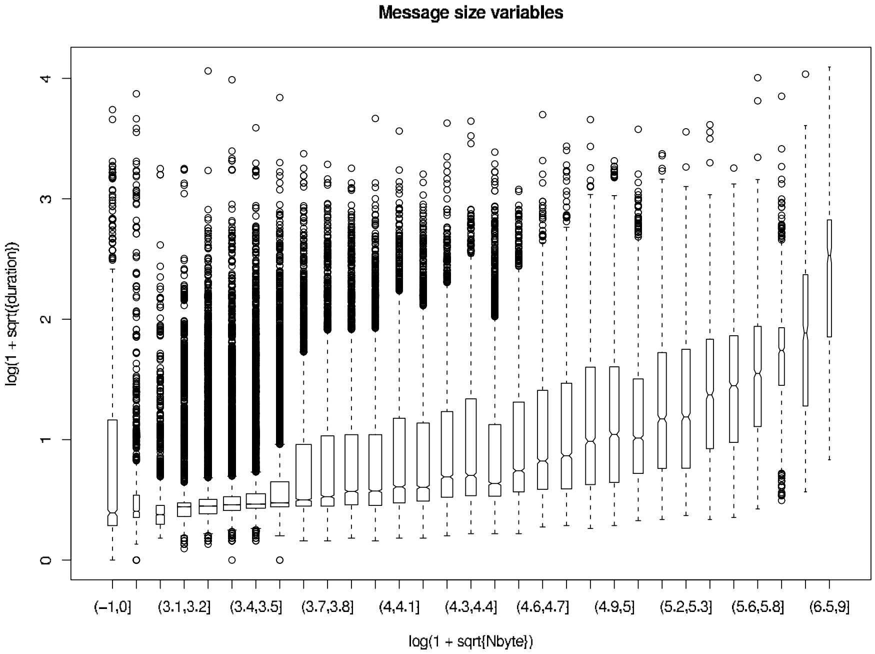


Figure 2: Boxplots of `log.duration` $= \log(1 + \sqrt{duration})$ vs `log.Nbyte` $= \log(1 + \sqrt{Nbyte})$.

trends: sessions with more bytes tend to last longer, and most sessions are short.

The preponderance of relatively short sessions can be seen in Figure 3(a), which displays the session durations as horizontal lines that extend from the start time to the end time. Because these sessions are reported in the order in which they began, the session start times range from time 0 (bottom line) to 59.971 (nearly the end of the hour). Figure 3(b) shows the same information, but each line is shifted back to 0. With continuously monitored data, the session duration lines would continue past the censoring point (illustrated as a red dotted line in Figure 3b). Relatively few sessions are “censored” (i.e., ended within the hour), reflecting the fact that most sessions are short: 93% of the sessions lasted less than 30 seconds. Figure 4 shows a barplot of the number of active sessions during each 30-second subset of this one-hour period (a time frame of 30 seconds is selected to minimize the correlation between counts in adjacent bars). The mean number of active sessions in any one 30-second interval during this hour is 923, with standard deviation 140, suggesting a rough upper “3-sigma limit” of 1343 sessions. [Because these numbers are *counts*, a square root transformation may be appropriate; see Tukey [4]. The mean and standard deviation of the square roots of the counts are 30.29 and 2.23, respectively, resulting in an approximate upper “3-sigma limit" of $(30.29 + 3 \cdot 2.23)^2 = 1367$, very close to the limit on the raw counts, since the Poisson distribution with a high mean is approximately Gaussian.] The maximum number of sessions in any one of these 120 30-second intervals is 1299, below the "3-sigma limit". This plot could be monitored continuously in time, dropping older bars off the left-side of the plot, and adding new bars on the right; the upper 3-sigma limit could depend upon hour, day, or week of the year.

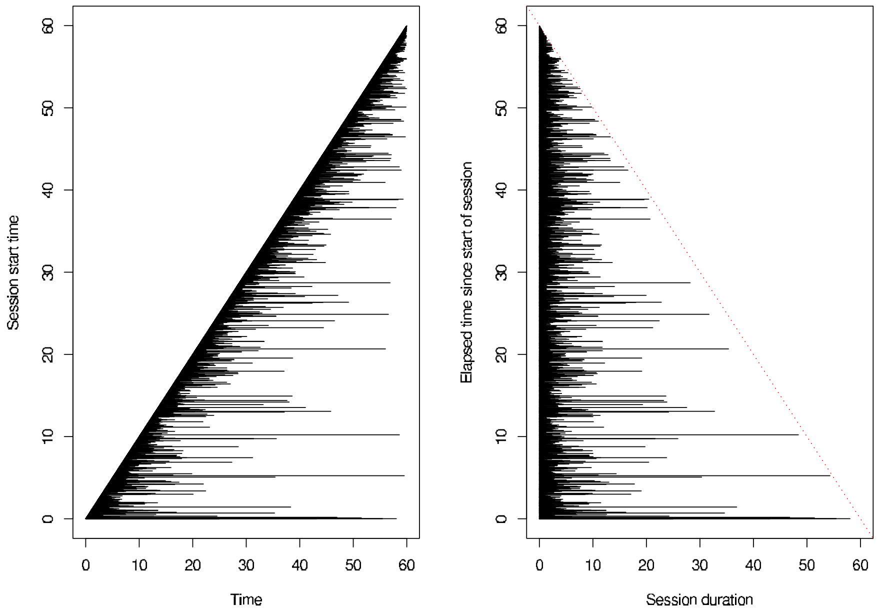


Figure 3: Session duration plot. Panel (a) displays the length of the session, starting at the actual start time. Panel (b), a shifted version of panel (a), displays the length of the session, starting at time 0.

Distribution of session duration by session start time can also be displayed as a scatter plot of points, (`time`, `duration`), as shown in Wegman and Marchette [7, p. 14]. Because the short sessions dominate this plot, Figure 5(a) shows this same plot, but using the log-transformed ordinate instead; i.e., (`time`, `log.len`). This plot shows a nearly straight line of points at `log.len` = 2.2685 between session start times of 23 to 57 minutes. Figure 5(b) expands this part of the plot and marks the identified points with red "+" symbols; they fall into two groups: an early group of 268 points (mean `duration` = 75.05 seconds), and a later group of 152 points (mean `duration` = 75.15 seconds). All 420 points are web sessions (destination port 80) and arise from a single source IP 65246, destination IP 45900, and source ports numbered 10000 or higher. A major challenge is the development of statistical "screening" algorithms to identify such "interesting" patterns in data

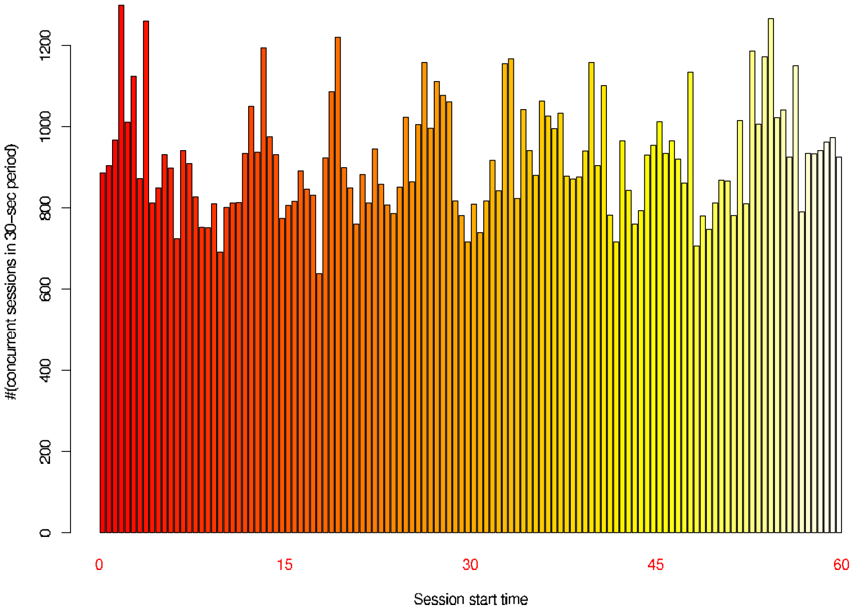


Figure 4: Barplot of the number of session in successive 30-second non-overlapping intervals during the hour (120 intervals). The mean number is 923 (standard deviation = 140), yielding an upper 3-sigma limit of approximately 1343. The maximum of these 120 counts is 1299.

plots such as this one, so that potential attacks to networks can be identified in real time. Since an infinite number of patterns can occur, a collection of likely patterns must be catalogued, so that statistical significance on their detection can be quantified. Algorithms that identify too many false negative patterns would result in an unnecessary number of shutdowns and service denials.

Relationships between pairs of size variables

Figure 6 shows a series of plots of the transformed `Nbyte` variable, `log.byte`, versus the transformed `Npacket` variable, `log.pkt` $= \log(1+\sqrt{Npacket})$, in four separate ranges. Panel (a), observations for which `log.pkt` is between 1 and 2 (`Npacket` between 3 and 40), shows a generally increasing trend, simplified in Panel (b) with boxplot displays (the labels in the x-axis are the same as those in panel (a), multiplied by 10). Panel (c) shows one line of 293 points around `log.pkt` $= 2.77$ (`Npacket` ≈ 226) and `log.byte` $=$ 5.82 (`Nbyte` $\approx$ 112,792) [all come from destination port 80 (web), source IP 23070, and destination IP 443 (`https`)]; and another set of 39 points around `log.pkt` $= 2.84$ (`Npacket` ≈ 259) and `log.byte` $= 5.64$ (`Nbyte` $\approx$ 78,341) [all have `SIP` 4837, `DIP` 56612, `DPort` 80]. Panels (c) and (d) show many points at high values of `log.pkt` along two lines with approximately unit slope but

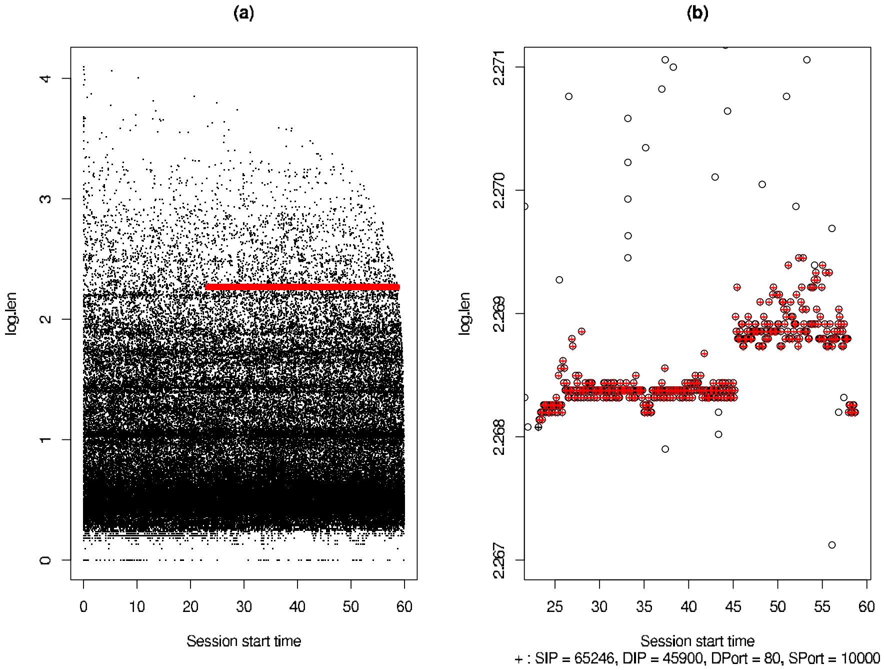


Figure 5: Plot of log-transformed `duration`, `log.len`, as a function of session start time. The left panel (all data) shows an almost perfectly horizontal line of points at `log.len` $= 2.2685$, between 23 and 57 minutes (expanded in the right panel).

with different intercepts; the upper set corresponds mainly to destination IP addresses 25 (smtp mail), 80 (web), and 443 (secure web); the lower set corresponds mostly to `DIP` 554 (rtsp). The dense set of 55 points points in Figure 6(d) ($3.4 <$ `log.pkt` < 3.8) lie near the line `log.byte` $= 3 +$ `log.pkt`, and 43 of them correspond to `DPort` 43. The extent to which such a pattern could occur by chance alone should be investigated.

The relationship between number of bytes, `NByte`, and session duration, `duration` is shown with plots of `log.byte` versus `log.len`, first for the entire data set (Figure 7), and then in 4 subranges defined by the intervals of `log.byte` and `log.len` (Figure 8). This approach to viewing the data can be considered as roughly equivalent to a “drill-down” approach, where all data are displayed in translated regions of the logged variables. Because longer sessions are associated with more bytes, and most sessions are short, plots of `log.byte` $= log(1+\sqrt{Nbyte})$ versus `log.len` $= log(1+\sqrt{duration})$ should be dense near the low end of each scale but much less dense near the upper end. In fact, the points in Figure 7 are especially dense around `log.len` $= 0.5$, then less dense until `log.len` $= 1.1$. Figure 7 also reveals

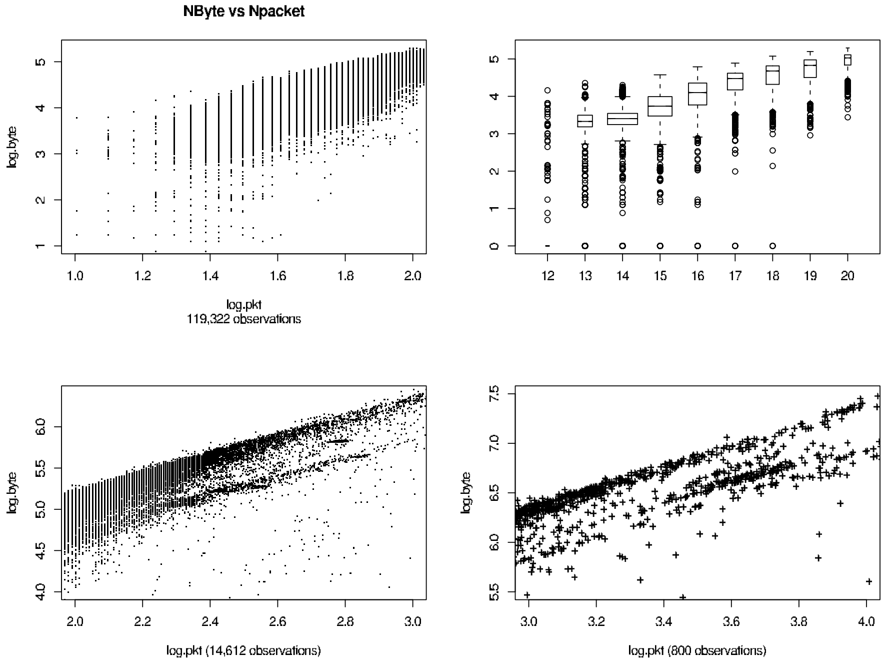


Figure 6: Plots of `log.byte` versus `log.pkt`, in 4 subranges of `log.pkt`.

a set of points in the upper right corner of the plot, 2.5 ≤ `log.len` ≤ 3, and 6 ≤ `log.byte` ≤ 7, which is discussed below in connection with Figure 9(c). Figure 8 shows three uncommonly straight lines of points: 377 points in Figures 8(a), in the region where 1.17 ≤ `log.len` ≤ 1.27 and `log.byte` ≈ 5.0; 292 points in Figure 8(b), where 1.69 ≤ `log.len` ≤ 1.79 and `log.byte` ≈ 5.8; and 60 points in Figure 8(d), where 2.7 ≤ `log.len` ≤ 2.9 and `log.byte` increases from 6.4 to 7. The points in these sets of lines have in common (1) `SIP` = 1681, `SPort` = 10000, `DPort` = 25 (smtp mail) (recall that `SPort` 10000 actually refers to all source ports numbered 10000 or higher); (2) `SIP` = 23070, `DIP` = 336, `DPort` = 80 (web); and (3) `DPort` = 554 (rtsp), `SPort` = 1276 to 2070. For a given session, initial ports are assigned at random, but subsequent ones are assigned by an incrementing pattern characteristic of the operating system. Hence, a string of `SPort` numbers may signal a potential attacker seeking information about operating system to invade.

Stratification by groups of destination ports

This hour of Internet activity involved 380 unique destination ports (Table 2). `DPort` 80 (web) is the most common, comprising 116,134 of the 135,605 records. The next most common destination port is `DPort` 443 (secure web `https`), utilized 11,627 times, followed by `DPort` 25 (mail SMTP) accessed 6,186 times. Ports 554 (rtsp), 113, 10000 (or higher), 8888 occur 200, 128, 97, 94 times, respectively. Twelve destination port numbers during

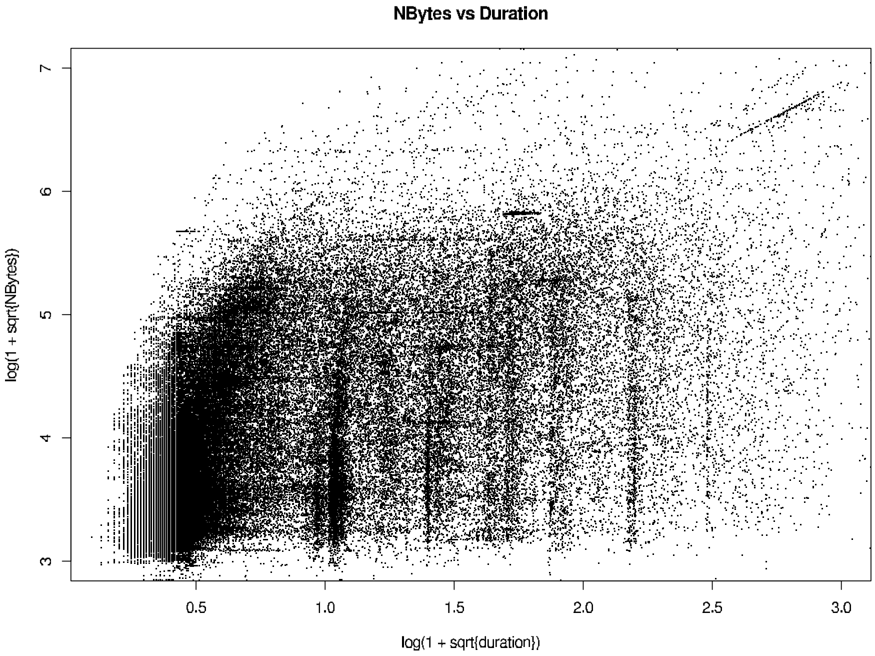


Figure 7: `log.byte` versus `log.duration`.

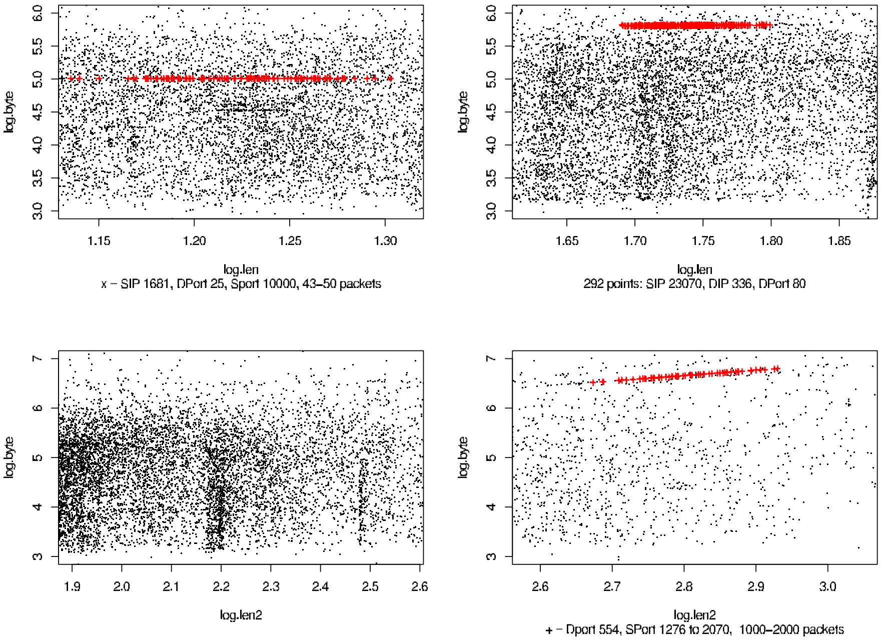


Figure 8: `log.byte` versus `log.duration`, 4 subranges.

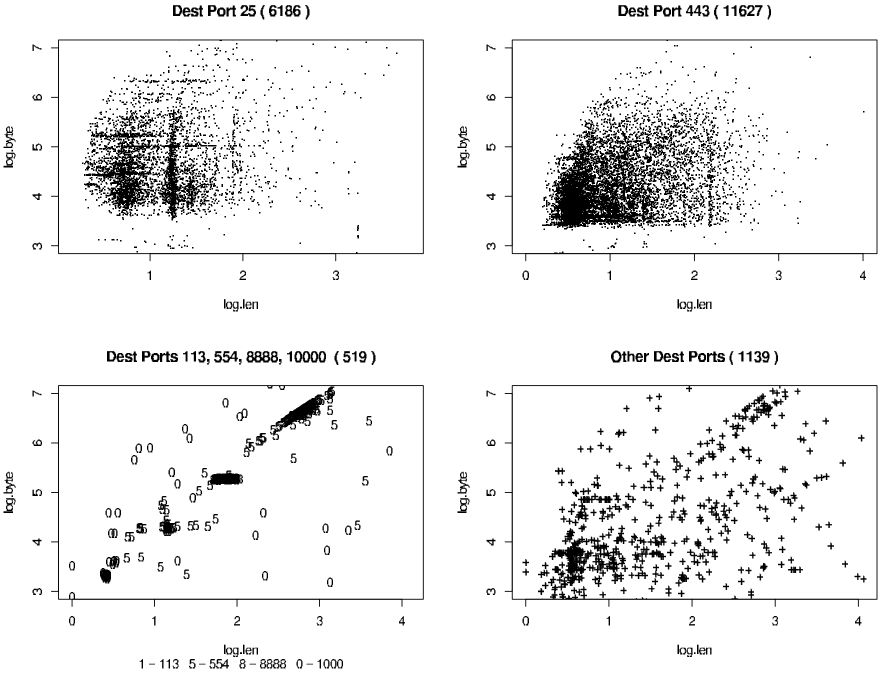


Figure 9: `log.byte` vs `log.len` for other destination ports.

this hour occurred between 5 and 29 times in the file; 5 ports occurred only 4 times, 8 occurred only 3 times, 47 destination ports occurred only twice, and 293 destination ports occurred only once. Displaying all 135,605 points on one plot is not very informative, so instead we subdivide the session records into groups according to their destination ports. Because over 85% of these data are web sessions (`DPort` = 80), a plot of `log.byte` versus `log.len` for only the web sessions looks like Figure 7 (all data). Figure 9 are scatterplots of two variables, conditioned on values of a third (non-web `DPorts`): `DPort` 25 (smtp mail) in panel (a); 443 (`https`) in panel (b); 113, 554, 8888, 10000 in panel (c), and the remaining 310 destination ports in panel (d). Panel (c) shows that the line of points in the upper right corner of Figure 7 arises from sessions with `DPort` 554 (rtsp), and that the sessions from `DPort` 8888 occur in a small cluster near `log.len` = 2 and `log.byte` = 5. Forty of the 52 points in the upper right corner of Figure 9(d), where `log.byte` $\approx 4 +$ 0.5 `log.len`, correspond to `DPort` numbers 119 and 1755, but are otherwise unrelated (some "patterns" can be spurious).

Monitoring frequency of source IP addresses

These same plots can be constructed when the data are subsetted by source IP address (`SIP`), as opposed to destination port number (`DPort`). The number of source IP addresses that may be active during a given hour

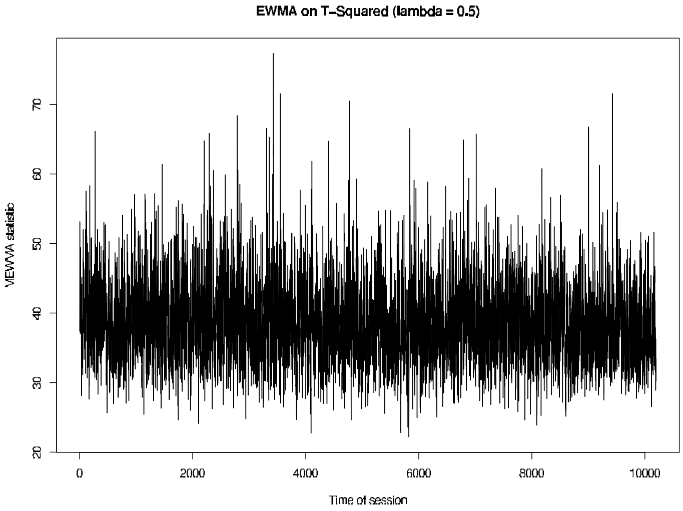


Figure 10: Multivariate EWMA plot of Hotelling's T^2, last 10,000 values.

of activity is likely to be very much higher than the number of destination ports; in this data set, only 380 unique destination ports were accessed, versus 3548 unique source ports. A plot of `log.Nbyte` versus `log.len` shows clumps of observations for certain source IP addresses, often because they correspond to heavy web traffic.

Multivariate charts

The three session "size" variables, `log.len, log.pkt, log.byte`, being somewhat correlated, are amenable to a "control chart' where the statistic being plotted is a weighted linear combination of the previously plotted variable (λ) and the current value of Hotelling's T^2 statistic $(1 - \lambda)$. Vardeman and Jobe (1999) provide tables for the optimal choices of λ. Calculating a Hotelling's T^2 statistic on three successive observations, denoted H_t, a multivariate exponentially weighted moving average (MEWMA) chart using λ = 0.5 is shown in Figure 10 (last 10,202 observations only). Most values (99.7%) are below 60; a successive run of observations above 60 might suggest abnormal session sizes. To minimize the effect of outliers on Hotelling's T^2 statistic, location and scale are estimated using medians and trimmed standard deviations instead of classical sample means and standard deviations (SDs). The SDs were estimated as 1.85, 0.34, 0.74, and the pairwise correlations are 0.53 (`log.len, log.pkt`), 0.56 (`log.len, log.pkt`), 0.90 (`log.pkt, log.byte`).

4 Evolutionary displays

Wegman and Marchette [6] advocate a new approach to visualizing massive data sets, called "evolutionary displays." Massive data sets are too large to display using graphs and plots that are designed for moderate data sets of

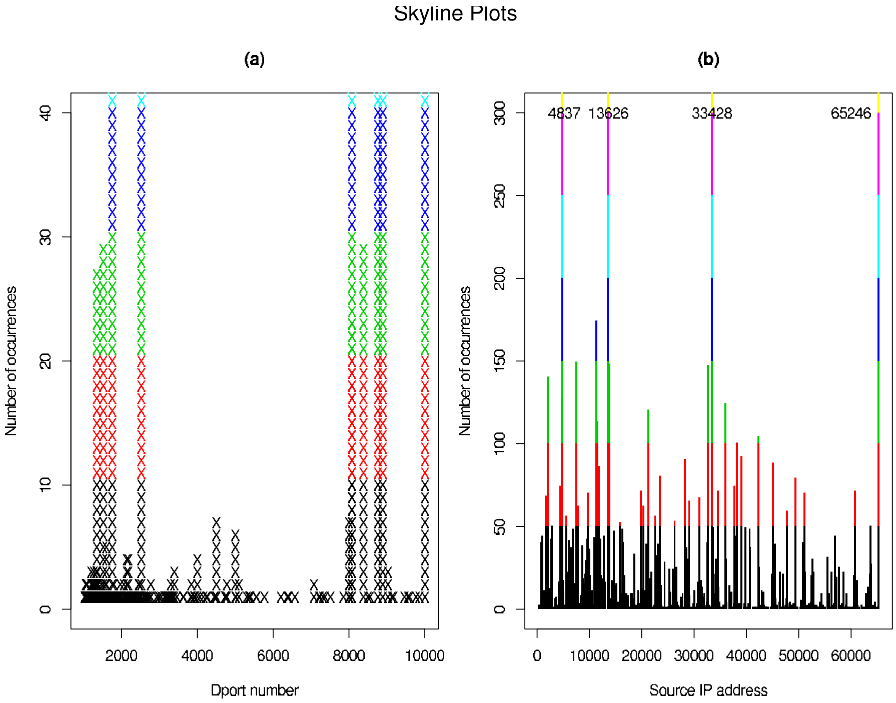


Figure 11: Skyline plots. (a): DPort access; (b): Source IP access.

fixed size. The concept behind evolutionary displays is to exhibit data within the most current time frame, dropping off old data and making room for most recent data. For example, in Figure 10, new data come in on the right as old data on the left are pushed off the screen. Wegman and Marchette [6, p. 906, Figure 4] use this concept to define a waterfall display, useful for monitoring frequency of source ports.

Skyline plots

Most destination port numbers occur only once or twice during the hour; of the 380 distinct DPorts, 293 occurred only once, 47 occurred twice, 8 occurred 3 times, 5 occurred 4 times. The remaining 27 ports occurred over 4 times; the top five are DPort 80 (web, 116,134 times), 25 (mail-smtp, 6,186 times), 443 (secure web, 11,627), 554 (rtsp; 200 times), and 113 (128 times). Setting aside the “well-known” ports 0–1023, we plot the occurrence of destination ports numbered 1024 and above, which should arise more or less at random, and flag as unusual any DPort that is referenced over 10 times. Figure 11 shows two such plots; one for DPort (color changes indicate DPort access counts greater than 10, indicative of potentially high traffic on this destination port), and one for SIP in the first 10,000 session records (color changes indicate SIP occurrences of more than 50). Four unusually frequent source IP addresses are immediately evident: 4837, 13626, 33428,and 65246, which occur 371, 422, 479, and 926 times, respectively, in the first 10,000 sessions. The construction of this plot resembles the tracing of a skyline, so we call it a “skyline plot.” Limits on skyline plots may depend upon time of day, day of week, month, or season.

5 Summary and further work

This article has highlighted several of the challenges that arise in analyzing and displaying massive data sets. Some simple statisitics based on robust quantities are useful for characterizing typical behavior (e.g., number of source and destination ports, and source and destination IP addresses, and frequency of access). These characterizations suggest graphical displays which highlight unusual usage or access. We discussed the role of “evolutionary graphics” on such data, specifically the use of “waterfall diagrams”, and proposed “skyline plots” as a means of monitoring ports and IP addresses. Future work will include massive data sets from Internet sessions and other fields.

References

[1] Marchette D.J. (2001). *Computer intrusion detection and network monitoring.* Springer.

[2] Khumbah N.-A., Wegman, E.J. (2003). *Data compression by geometric quantization.* Recent Advances and Trends in Nonparametric Statistics, M. Akritas, D.N. Politis (eds), North Holland Elsevier, Amsterdam.

[3] Silverman B.W. (1986). *Density estimation.* Chapman and Hall: London.

[4] Tukey J.W. (1977). *Exploratory data analysis.* Addison-Wesley, Reading, Massachusetts.

[5] Vardeman S.B., Jobe J.M. (1999). *Statistical quality assurance methods for engineers.* Wiley, New York.

[6] Wegman E.J., Marchette D.J. (2003). *On some techniques for streaming data: A case study of Internet packet headers.* J. Comput. Graph. Stat. **12** (4), 893 – 914.

[7] Wegman E.J.; Marchette D.J. (2004). *Statistical analysis of network data for cybersecurity.* Chance, 9 – 19.

Acknowledgement: Funding from Grant No. F49620-01-1-0274 from the Air Force Office of Scientific Research, awarded to George Mason University, is gratefully acknowledged. Part of this research was conducted during the first author’s appointment as faculty visitor at National Institute of Standards and Technology.

Address: K. Kafadar, E.J. Wegman, University of Colorado-Denver and George Mason University

E-mail: `kk@math.cudenver.edu; ewegman@galaxy.gmu.edu`

COMPSTAT'2004 Symposium

CLUSTERING ALL THREE MODES OF THREE-MODE DATA: COMPUTATIONAL POSSIBILITIES AND PROBLEMS

Henk A.L. Kiers

Key words: Cluster analysis, multiway analysis.

COMPSTAT 2004 section: Clustering.

Abstract: For the analysis of three-mode data sets (i.e., data sets pertaining to three different sets of entities) various component analysis techniques are available. These yield components that are summaries of the entities of each mode. Because such components are often interpreted in a more or less binary way in terms of the entities related strongest to them, it seems logical to actually constrain these components to have binary values only. In the present paper, such constrained models are proposed and algorithms for fitting these models are provided. In one of these variants, the components are constrained such that they correspond to nonoverlapping clusters of entities. Finally, a procedure is proposed for steering component values towards binary values, without actually imposing them to be binary, using penalties.

1 Analysis of three-mode data

Three-mode data sets are data sets pertaining to three different sets of entities. An example of a three-mode data set is a set of scores of a number of individuals, on a number of variables, each obtained under a number of different conditions. For the analysis of three-mode data, various exploratory three-way methods are available. The two most common methods for the analysis of three-mode data are CANDECOMP/PARAFAC [1], [6] and Tucker3 analysis [16], [10]. Both methods summarize the data by components for all three modes, and for the entities pertaining to each mode they yield component weights; in the case of Tucker3 analysis, in addition a so-called core array is given, which relates the components for all three modes to each other.

If we denote our $I \times J \times K$ three-mode data array by $\underline{\boldsymbol{X}}$, then the two methods can be described as fitting the model

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk} \,, \tag{1}$$

where a_{ip}, b_{jq} and c_{kr} are referred to as the component weights, which are elements of the component matrices $\boldsymbol{A}$ (for mode A), $\boldsymbol{B}$ (for mode B), and $\boldsymbol{C}$ (for mode C), of orders $I \times P$, $J \times Q$, and $K \times R$, respectively; g_{pqr} denotes the element (p,q,r) of the $P \times Q \times R$ core array $\underline{\boldsymbol{G}}$, and e_{ijk} denotes the error term for element x_{ijk}; P, Q, and R denote the numbers of components for the three respective modes. The difference between CANDECOMP/PARAFAC and Tucker3 analysis is that in CANDECOMP/PARAFAC the core is actually set equal to a superidentity array (i.e., $g_{pqr} = 1$ if $p = q = r$, $g_{pqr} = 0$ otherwise). As a consequence, in the case of CANDECOMP/PARAFAC, for all modes we have the same number of components, and (1) actually reduces to

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \tag{2}$$

Clearly, when these models are fitted to data, we end up with component matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, and, in the case of Tucker3 analysis, we also get a three-mode core array $\underline{\boldsymbol{G}}$ as outcome of the analysis.

The result of a three-mode analysis is a summary of the observation units, the variables and the conditions by means of a number of components, and possibly a core array describing the relations between them. The componentwise interpretation, however, is not very easy, because it requires one to think in dimensions along which the observation units, variables or conditions vary. Here the component weights indicate to *what extent*, for instance, the individuals can be described by the property defined by the component. Likewise, variables are related to the components for the variables to different extents. Now the interpretation of the components usually proceeds conversely: From the strengths of the relations of the variables to the components, one can interpret the meaning of the components. This interpretation is rather cumbersome if one discriminates precisely between different strengths of relations. Therefore, in practice, one tends to interpret components on the basis of the variables related strongest to it, and one tends to ignore the less related variables. In fact, thus one binarizes the relations, in sufficiently strong, and not sufficiently strong. Thus one could say that the components are interpreted as if they refer to clusters of variables consisting of those variables that have the strongest relations with them. Similar cluster based interpretations can be given to components describing individuals and conditions, if *a priori* information on the individuals and conditions is available. To enhance the interpretability of the component matrices, they are often subjected to simple structure rotations such as varimax [7], see also [8], but the clusters will always remain somewhat fuzzy (i.e., relations are never entirely binarized).

Now if, in practice, components tend to be interpreted as clusters, then would not it seem more rational to model data in terms of cluster membership, and discard the information on strengths of relations? The idea of clustering all three modes simultaneously has been pursued by various authors. Clustering approaches involving the CANDECOMP/PARAFAC model have been proposed by Chaturvedi and Carroll [3] and Leenen et al. [11], where the latter authors use Boolean products rather than ordinary products. An extension of the latter Boolean model to the Tucker3 situation, has been proposed by Ceulemans, van Mechelen and Leenen [2].

Surprisingly, except for a recent paper by Rocci and Vichi [13], straightforward (non-Boolean) generalizations of the Tucker3 model do not seem to have received attention yet, and no algorithms seem to have been published for handling this case. The present paper, therefore, focuses on that particular case. The models described here are in fact three-mode generalizations of the GENNCLUS model [4], [5] PENNCLUS, and the Double k-means clustering model by Vichi [17].

2 Clustering variants of the Tucker3 model

As has been mentioned above, in the Tucker3 model the elements of the component matrices are, in practice, often interpreted in a more or less binary way. That is, when interpreting a component for, say, the variables, for each variable it is specified whether it is associated with the component or not. Thus, a Tucker3 model that fully complies with this binary way of interpretation would simply have binary component weights for the variables: 1 for the variables associated with a component, and 0 for those not associated with the component. In fact, one might want to specify the strength of the association by a value different than 1, but if the same value is used for all variables related to a component, then one can always scale such values to 1 anyway. Therefore, it is here proposed to constrain the elements of each component matrix to be binary, that is to be equal to 0 or 1.

When all elements of the component matrices are binary, one could say that the components refer to clusters of, for example, variables. Without further constraints, such clusters may very well overlap, in the sense that some entities are associated with more than one cluster. The overlap of clusters is nonproblematic for the interpretation of the clusters themselves, but does make the overall model relatively difficult to interpret. Therefore, it can be attractive to impose a further constraint, namely the constraint that clusters do not overlap. Specifically, this constraint implies that each entity is assigned to one and only one cluster.

Models for these two constrained variants of the Tucker3 model are described below, and it is also indicated how this model can be fitted to data. In the next section, algorithms for actually carrying out such fitting procedures are given.

2.1 Tucker3 with overlapping clusters

The Tucker3 model with overlapping clusters is defined as the model

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk} \,, \tag{3}$$

where a_{ip}, b_{jq}, and c_{kr} are constrained to be binary (0 or 1). To avoid summation notation, we write the above model in terms of matrices as follows

$$\boldsymbol{X}_a = \boldsymbol{A}\boldsymbol{G}_a(\boldsymbol{C}' \otimes \boldsymbol{B}') + \boldsymbol{E}_a \,, \tag{4}$$

where $\boldsymbol{X}_a$, $\boldsymbol{G}_a$, and $\boldsymbol{E}_a$ denote the A-mode matricized versions of the three-way arrays $\underline{\boldsymbol{X}}$, $\underline{\boldsymbol{G}}$, and $\underline{\boldsymbol{E}}$ (i.e., the matrices obtained upon putting the frontal slabs next to each other, see [9], and $\otimes$ denotes the Kronecker product. To fit this model to an empirical data set, it is proposed here to minimize the sum of squared residuals, hence to minimize

$$f(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\underline{\boldsymbol{G}}) = \|\boldsymbol{X}_a - \boldsymbol{A}\boldsymbol{G}_a(\boldsymbol{C}' \otimes \boldsymbol{B}')\|^2 \,, \tag{5}$$

over $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\underline{\boldsymbol{G}}$, subject to the constraint that the elements of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are binary. Note that the core array is left fully unconstrained.

It is well known that the Tucker3 model is not unique. That is, nonsingular transformations of the component matrices can be compensated by the inverse transformations in the core, and thus do not affect the model estimates. For example, suppose we transform $\boldsymbol{A}$ by multiplying it by a nonsingular matrix $\boldsymbol{S}$, premultiplying $\boldsymbol{G}_a$ by $\boldsymbol{S}^{-1}$ yields exactly the same model estimates since $(\boldsymbol{A}\boldsymbol{S})(\boldsymbol{S}^{-1}\boldsymbol{G}_a)(\boldsymbol{C}' \otimes \boldsymbol{B}') = \boldsymbol{A}\boldsymbol{G}_a(\boldsymbol{C}' \otimes \boldsymbol{B}')$. In the case of binary constraints, this nonuniqueness is limited to those cases where nonsingular transformations do not affect the binary constraint. This is possible when there are columns in, for instance, matrix $\boldsymbol{A}$ that do not overlap: upon replacing one such column by the sum of such columns, the binary constraint will still be satisfied. Specifically, suppose $\boldsymbol{A}$ has only two columns that do not overlap (i.e., do not have unit elements at the same position), then replacing the second by the sum of the two comes down to postmultiplying $\boldsymbol{A}$ by the nonsingular matrix $S = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. Clearly, then $\boldsymbol{A}\boldsymbol{S}$ satisfies the binary constraint, and upon replacing $\boldsymbol{A}$ by $\boldsymbol{A}\boldsymbol{S}$, and $\boldsymbol{G}_a$ by $(\boldsymbol{S}^{-1}\boldsymbol{G}_a)$ we get the same estimates as with $\boldsymbol{A}$ and $\boldsymbol{G}_a$. Similar nonuniquenesses can be identified upon describing model (3) using B- or C-mode matricized versions as

$$\boldsymbol{X}_b = \boldsymbol{B}\boldsymbol{G}_b(\boldsymbol{A}' \otimes \boldsymbol{C}') + \boldsymbol{E}_b \,, \tag{6}$$

and

$$\boldsymbol{X}_c = \boldsymbol{C}\boldsymbol{G}_c(\boldsymbol{B}' \otimes \boldsymbol{A}') + \boldsymbol{E}_c \,, \tag{7}$$

where subscripts b and c indicate B- and C-mode matricized versions of the three-way arrays at hand, which are obtained by other ways of positioning slices of the three-way arrays next to each other, see Kiers [9].

2.2 Tucker3 with nonoverlapping clusters

The Tucker3 model with nonoverlapping clusters is the same model as that for overlapping clusters described above, in Section 2.1, with the *additional constraint* on the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ that in all rows one and only one element is 1, and all others are 0. The procedure to fit this model is hence to minimize (5) over $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\underline{\boldsymbol{G}}$, subject to the constraint that the elements of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are binary with exactly one unit element in each row. As a consequence of the minimization subject to these constraints, the core array now will contain the within cluster average scores in $\underline{\boldsymbol{X}}$, hence the core effectively summarizes the data in such a way that it gives the average score of the individuals in each cluster, averaged across the variables associated to the variable cluster at hand, and averaged across conditions associated with the condition cluster at hand. When the clusters can be interpreted well, then the core has a very easy interpretation too, simply in terms of 'cluster scores'.

3 Algorithm for Tucker3 with overlapping clusters

As mentioned in Section 2.1, fitting the Tucker3 model with overlapping clusters comes down to minimizing (5) over $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\underline{\boldsymbol{G}}$, subject to the constraint that the elements of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are binary. To find solutions for this minimization problem, it is proposed here to use an alternating least squares algorithm, which, starting from initial values for $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\underline{\boldsymbol{G}}$, finds updates for $\boldsymbol{A}$ keeping the other matrices fixed, then for $\boldsymbol{B}$ keeping the other matrices fixed, next for $\boldsymbol{C}$ keeping the other matrices fixed, and finally for $\underline{\boldsymbol{G}}$ keeping the other matrices fixed. After one complete cycle, the function value is evaluated, and if it has decreased considerably, a new cycle is started. This process is repeated until the function value changes no longer. Each update is found such that it decreases the function value, or, at least does not increase the function value. Because the function value is bounded below by 0, it is thus guaranteed to converge to a stable value.

3.1 Updating procedures

The choice for initial values for $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\underline{\boldsymbol{G}}$ will be discussed later. Given that such values are available, the first step is to find improved values for $\boldsymbol{A}$, keeping the other matrices fixed. Hence the problem is to minimize

$$g(\boldsymbol{A}) = \|\boldsymbol{X}_a - \boldsymbol{A}\boldsymbol{F}\|^2 \,, \tag{8}$$

where $\boldsymbol{F}$ is written for $\boldsymbol{G}_a(\boldsymbol{C}' \otimes \boldsymbol{B}')$. Now the columns of $\boldsymbol{A}$ are updated column after column, keeping the other columns of $\boldsymbol{A}$ fixed. Specifically, to update column j of $\boldsymbol{A}$, we find the minimum of

$$g(\boldsymbol{a}_j) = \left\| \boldsymbol{X}_a - \sum_{l \neq j} {}_l \boldsymbol{f}'_l - \boldsymbol{a}_j \boldsymbol{f}'_j \right\|^2 = \|\boldsymbol{X}_{-j} - \boldsymbol{a}_j \boldsymbol{f}'_j\|^2 \,, \tag{9}$$

where $\boldsymbol{X}_{-j}$ is written for $\boldsymbol{X}_a - \sum_{l \neq j} \boldsymbol{a}_l \boldsymbol{f}'_l$, $\boldsymbol{a}_j$ denotes the jth column of $\boldsymbol{A}$, and $\boldsymbol{f}'_j$ denotes the jth row of $\boldsymbol{F}$. A solution for minimizing (9) is given by Chaturvedi and Carroll [3]. A computationally slightly different procedure (with the same solution) can be derived as follows. Function (9) can be written as the sum of independent functions elaborated as

$$\begin{aligned} g(a_{ij}) &= \text{constant} - 2a_{ij}(\boldsymbol{X}_{-j}\boldsymbol{f}_j)_i + a_{ij}^2\boldsymbol{f}_j'\boldsymbol{f}_j \\ &= \text{constant} + (\boldsymbol{f}_j'\boldsymbol{f}_j - 2(\boldsymbol{X}_{-j}\boldsymbol{f}_j)_i)_{aij}, \quad i = 1, \ldots, I, \qquad (10) \end{aligned}$$

where in the second line it is used that $a_{ij}^2 = a_{ij}$ because each element of $\boldsymbol{A}$ is constrained to be binary. Each of the functions $g(a_{ij})$ is now minimized over binary a_{ij} by taking $a_{ij} = 0$ if $(\boldsymbol{f}_j'\boldsymbol{f}_j - 2(\boldsymbol{X}_{-j}\boldsymbol{f}_j)_i) > 0$, and $a_{ij} = 1$ if $(\boldsymbol{f}_j'\boldsymbol{f}_j - 2(\boldsymbol{X}_{-j}\boldsymbol{f}_j)_i) \leq 0$, hence

$$\begin{aligned} a_{ij} = 0 &\quad \text{if } 2(\boldsymbol{X}_{-j}\boldsymbol{f}_j)_i) < \boldsymbol{f}_j'\boldsymbol{f}_j \\ a_{ij} = 1 &\quad \text{if } 2(\boldsymbol{X}_{-j}\boldsymbol{f}_j)_i) \geq \boldsymbol{f}_j'\boldsymbol{f}_j \quad i = 1, \ldots, I\,. \qquad (11) \end{aligned}$$

In practice, it may happen that all elements of column j become zero by the above updates of the elements of column j. This would imply that the Tucker3 model would not use the jth A-mode component. Hence all core elements related to this component (in the jth row of $\boldsymbol{G}_a$), and therefore also the elements in the jth row of $\boldsymbol{F}$, do not have any contribution to fitting the data; in other words, then the term $\boldsymbol{a}_j\boldsymbol{f}_j' = 0$. However, in practice, this will almost never be the optimal solution for $\boldsymbol{a}_j\boldsymbol{f}_j'$, since it would imply that no contribution is better than any conceivable contribution. Furthermore, zero columns in A will cause computational problems later on in the algorithm. Therefore, whenever $\boldsymbol{a}_j = 0$, a special fixing procedure seems in order. Here we use the following. If $\boldsymbol{a}_j = 0$, first the jth row of $\boldsymbol{G}_a$ and hence also the jth row of $\boldsymbol{F}$, is multiplied by -1. This does not affect the fit, because when $\boldsymbol{a}_j\boldsymbol{f}_j' = 0$, then also $\boldsymbol{a}_j(-\boldsymbol{f}_j') = 0$. Next, $\boldsymbol{a}_j$ is updated again according to (11), and this is used as the update for $\boldsymbol{a}_j$. If it so happens that the updated $\boldsymbol{a}_j$ again is a vector with zeros only, then $\boldsymbol{a}_j$ is set back to its original values before updating column j, and likewise the core is set back to its original values.

To update matrix $\boldsymbol{B}$, a completely analogous procedure is followed. Specifically, noting that (4) has equivalently been written as (6) $\boldsymbol{X}_b = \boldsymbol{B}\boldsymbol{G}_b(\boldsymbol{A}' \otimes \boldsymbol{C}') + \boldsymbol{E}_b$, it can be seen that using this version of the model, the process of updating $\boldsymbol{B}$ is the same as that described for $\boldsymbol{A}$ above, after replacing $\boldsymbol{A}$ by $\boldsymbol{B}$, $\boldsymbol{B}$ by $\boldsymbol{C}$, $\boldsymbol{C}$ by $\boldsymbol{A}$, and $\boldsymbol{G}_a$ by $\boldsymbol{G}_b$ in the above description. Likewise, updating matrix $\boldsymbol{C}$ can be carried out by using the procedure for updating $\boldsymbol{A}$, after replacing $\boldsymbol{A}$ by $\boldsymbol{C}$, $\boldsymbol{B}$ by $\boldsymbol{A}$, $\boldsymbol{C}$ by $\boldsymbol{B}$, and $\boldsymbol{G}_a$ by $\boldsymbol{G}_c$ in the above description.

Finally updating the core array can be carried out as follows. The problem now is to minimize

$$g(\underline{\boldsymbol{G}}) = \|\boldsymbol{X}_a - \boldsymbol{A}\boldsymbol{G}_a(\boldsymbol{C}' \otimes \boldsymbol{B}')\|^2 \qquad (12)$$

over $\underline{\boldsymbol{G}}$, which in A-mode matricized form is written as $\boldsymbol{G}_a$. Because there is no constraint on $\boldsymbol{G}_a$, the solution to this problem is given by

$$\begin{aligned} \boldsymbol{G}_a &= (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{X}_a(\boldsymbol{C} \otimes \boldsymbol{B})(\boldsymbol{C}'\boldsymbol{C} \otimes \boldsymbol{B}'\boldsymbol{B})^{-1} \\ &= (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{X}_a(\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C})^{-1} \otimes \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B})^{-1}) \qquad (13) \end{aligned}$$

see [12], see also [15]. Note that, if the inverses do not exist (as may come about when any of the component matrices has incomplete rank) then the inverse is replaced by a generalized inverse.

The above described steps for updating $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\underline{\boldsymbol{G}}$ are followed by the computation of the loss function value. If this has decreased, then a new cycle of updatings is started; if it has remained the same, then the ensuing solution is considered a candidate for the minimum of the loss function. Depending on how the procedure is started, this may be a local minimum of the function rather than the global minimum. It is therefore recommended to run the algorithm from several starts. One approach is to start from (very) many random starts, hoping thus to cover a wide range of (at least) locally optimal solutions for which the chance that it contains the global minimum is high. Alternatively, or in addition, one may use a few starts that can be expected to have a high chance to lead to the global minimum. A suggestion for such 'rational' starts is given in the next subsection.

3.2 Rational starts

Because the algorithm described above very easily leads to local optima, it is important to run the algorithm from various different starts, among which preferably are starts that have a high chance of leading to the global optimum. Experience so far has indicated that a useful starting configuration can be obtained as follows. First, analyze the data by ordinary Tucker3 analysis, leading to columnwise orthonormal component matrices. Next rotate all three component matrices by means of varimax, and multiply all columns that have a negative sum of elements by -1. Then one starting configuration is obtained by setting all values that are higher than their column average to 1 and all others to 0. An alternative is to set, for each matrix, all values above a particular threshold to 1, and all others to 0. The threshold should depend on the number of elements in the component matrix at hand, and it can be varied systematically to yield different starts. By systematically varying the threshold value for $\boldsymbol{A}$ between $I^{-1/2}$ and 0 (not including 0), different starts can be obtained, which in practice seem to lead to at least reasonably good solutions; likewise for $\boldsymbol{B}$ the threshold is to be chosen between $J^{-1/2}$ and 0, and for $\boldsymbol{C}$ the threshold is to be chosen between $K^{-1/2}$ and 0. More experience is needed, however, to evaluate the usefulness of these starts.

4 Algorithm for Tucker3 with nonoverlapping clusters

To fit the Tucker3 model with nonoverlapping clusters comes down to minimizing (5) over $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\underline{\boldsymbol{G}}$, but now subject to the constraint that the elements of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are binary, and that each row of these matrices has one and only one unit element. To find solutions for this minimization problem, it is proposed to use an alternating least squares algorithm similar in set up to that for the overlapping clusters situation. The updates for the component matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are, obviously, different, while the update for the core is the same, but its computation can now be simplified somewhat. This is because the inverses in the updating formula (13) are now very easy to compute, because, due to the constraints on the component matrices, $\boldsymbol{A}'\boldsymbol{A}$, $\boldsymbol{B}'\boldsymbol{B}$, and $\boldsymbol{C}'\boldsymbol{C}$ now are diagonal matrices with on the diagonal simply the number of unit elements in the corresponding columns of the component matrices.

Below only the updating procedure for $\boldsymbol{A}$ is described. Those for $\boldsymbol{B}$ and $\boldsymbol{C}$ are obtained analogously, after letting the component matrices switch roles (compare Section 3.1), and the update for the core does not need further description.

4.1 Updating procedure for $\boldsymbol{A}$

To update $\boldsymbol{A}$ subject to the constraints at hand, we now minimize

$$g(\boldsymbol{A}) = \|\boldsymbol{X}_a - \boldsymbol{A}\boldsymbol{F}\|^2 , \tag{14}$$

over $\boldsymbol{A}$, where $\boldsymbol{F}$ is again written for $\boldsymbol{G}_a(\boldsymbol{C}' \otimes \boldsymbol{B}')$. This function can be written as the sum of the independent functions

$$g(\boldsymbol{a}_i) = \|\boldsymbol{x}_i' - \boldsymbol{a}_i'\boldsymbol{F}\|^2 = \left\|\boldsymbol{x}_i' - \sum_l a_{il}\boldsymbol{f}_l'\right\|^2 , \tag{15}$$

where $\boldsymbol{x}_i'$ and $\boldsymbol{a}_i'$ denote the ith rows of $\boldsymbol{X}_a$ and $\boldsymbol{A}$, respectively, subject to the constraint that one of the elements of $\boldsymbol{a}_i'$ is 1 and all others are 0. Thus, due to the constraint, in $\sum_l a_{il}\boldsymbol{f}_l'$ all but one term are 0, while the nonzero term (the jth) equals $\boldsymbol{f}_j'$. Hence, the problem is simply to find the value j for which $\|\boldsymbol{x}_i' - \boldsymbol{f}_j'\|^2$ is minimal, and set the associated value a_{ij} equal to 1, and all other elements of $\boldsymbol{a}_i'$ equal to 0. In formulas, the updates for the elements of $\boldsymbol{a}_i'$ are given by

$$\begin{aligned} j &= \arg\min\left(\|\boldsymbol{x}_i' - \boldsymbol{f}_j'\|^2\right) \\ a_{ij} &= 1 \\ a_{il} &= 0, \quad \text{for } l \neq j\,. \end{aligned} \tag{16}$$

If a column of $\boldsymbol{A}$ turns out to have zero elements only, a slightly modified version of the fixing procedure described for the overlapping clusters case can be used. That is, in this case *all* rows of $\boldsymbol{F}$ corresponding to zero columns in $\boldsymbol{A}$ are multiplied by -1, and the whole matrix $\boldsymbol{A}$ is updated again. If this will again result in one or more zero columns in $\boldsymbol{A}$, then $\boldsymbol{A}$ is set back to the original values of $\boldsymbol{A}$.

The problem of fitting the Tucker3 model with nonoverlapping clusters has recently been proposed also by Rocci and Vichi [13], but at the time of writing, their algorithm had not yet been published. Even more recently, Schepers and van Mechelen [14] have proposed an algorithm for fitting this model, which also has not been published yet. It is planned to compare these algorithms in the near future.

4.2 Rational starts

As possibly useful rational starts for the nonoverlapping clusters algorithm, again the results from Tucker3 analysis applied to the data, followed by varimax of the component matrices can be used. This time, after multiplying columns having negative sums by −1, starts are obtained simply by setting all rowwise highest elements to 1, and all other elements to 0. Other rational starts are used in the algorithms by Schepers and van Mechelen [14], and by Rocci and Vichi [13]. Their relative advantages are still to be studied.

5 Should we fully constrain components to be binary?

In the present paper, procedures have been described for constraining components to be binary. However, it is known that fitting models under binary constraints is very difficult, in the sense that it is very hard to find the globally optimal solution. Moreover, the constraints of binarity may for some situations be too strong. In some situations, it may be needed to allow for nonzero component weights with clearly different values within columns. For such purposes, special algorithms are needed, which, to the author's knowledge, are not yet available.

An alternative route to avoid the very strong constraint of binarity could be to require component matrices to be close to binarity rather than to exact binarity. This can be achieved by imposing the binarity constraint as a soft constraint in such a way that it penalizes (rather than prohibits) nonbinarity. In other words, soft constraints can be imposed by minimizing the ordinary Tucker3 loss function to which penalty terms are added whose values increase with increasing deviations from binarity. One procedure for attaining this is to minimize the function

$$f(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{U},\boldsymbol{V},\boldsymbol{W},\boldsymbol{G}) = \|\boldsymbol{X}_a - \boldsymbol{A}\boldsymbol{G}_a(\boldsymbol{C}'\otimes\boldsymbol{B}')\|^2 \\ +\lambda\|\boldsymbol{U}-\boldsymbol{A}\|^2+\mu\|\boldsymbol{V}-\boldsymbol{B}\|^2+\nu\|\boldsymbol{W}-\boldsymbol{C}||^2 \quad (17)$$

over arbitrary $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\underline{\boldsymbol{G}}$, and over binary auxiliary matrices $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{W}$; λ, μ and ν are penalty parameters that regulate the strength of the constraint, and that have to be specified in advance. Without further constraints, one will find degenerate solutions in which component matrices tend to 0 (thus annihilating the penalty terms), while the core elements tend to infinity in such a way that the product $\boldsymbol{A}\boldsymbol{G}_a(\boldsymbol{C}'\otimes\boldsymbol{B}')$ still fits the data well. One way to avoid such degeneracies, which in practice turned out to work reasonably well, is to constrain the auxiliary binary matrices to have at least one nonzero element in each column.

An alternating least squares algorithm for minimizing (17) has been devised and programmed. The algorithm tends to require many iterations, but does indeed give solutions with the required properties. For instance, for data constructed on the basis of component matrices that were binary up to a few elements, the method indeed singled out these elements as different from the others. However, much more experience is needed to assess its usefulness in actual practice.

6 Conclusion

The present paper has offered methods for Tucker3 analysis with the component matrices constrained to be binary, and, in a special case also such that the components have no overlap. The algorithms proposed work in the sense that they decrease the loss function value, but they appear, as usual with binary optimization problems, to be prone to hitting local optima. Some starting procedures have been proposed that worked well in some contrived examples, but the algorithms, as well as their starting procedures need further testing, as well as comparison to competitors that have been proposed recently for the nonoverlapping case.

In addition to the methods where components are constrained to be fully binary, a procedure has been proposed for weakly imposing binarity, by using penalty terms. Again, this procedure needs further testing. If it turns out to work well in practice, and if it is not very prone to hitting local optima, it could also be used for fitting the fully constrained model by gradually increasing the penalty parameters that regulate the strength of the constraints. Whether this or other procedures work best in dealing with the local optimum problem of Tucker3 with binary constraints is subject to further research.

References

[1] Carroll J. D., Chang J.-J. (1970). *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition.* Psychometrika **35**, 283 – 319.

[2] Ceulemans E., van Mechelen I., Leenen I. (2003). *Tucker3 hierarchical classes analysis.* Psychometrika **68**, 413 – 433.

[3] Chaturvedi A., Carroll J.D. (1994). *An alternating combinatorial optimization approach to fitting INDCLUS and Generalized INDCLUS models.* Journal of Classification **11**, 155 – 170.

[4] DeSarbo, W.S. (1982). *GENNCLUS: New models for general nonhierarchical clustering analysis.* Psychometrika **47**, 449 – 475.

[5] Gaul W., Schader, M. (1996). *A new algorithm for two-mode clustering.* In: Bock H.-H., Polasek W. (eds.) Data analysis and information systems. Springer, Heidelberg.

[6] Harshman R.A. (1970). *Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-mode factor analysis.* UCLA Working Papers in Phonetics **16**, 1 – 84.

[7] Kaiser H.F. (1958). *The varimax criterion for analytic rotation in factor analysis.* Psychometrika **23**, 187 – 200.

[8] Kiers H.A.L. (1998). *Joint orthomax rotation of the core and component matrices resulting from three-mode principal components analysis.* Journal of Classification **15**, 245 – 263.

[9] Kiers H.A.L. (2000). *Towards a standardized notation and terminology in multiway analysis.* Journal of Chemometrics **14**, 105-122.

[10] Kroonenberg P.M., De Leeuw J. (1980). *Principal component analysis of three-mode data by means of alternating least squares algorithms.* Psychometrika **45**, 69 – 97.

[11] Leenen I., van Mechelen I., de Boeck P., Rosenberg S. (1999). *INDCLAS: A three-way hierarchical classes model.* Psychometrika **64**, 9 – 24.

[12] Penrose R. (1956). *On best approximate solutions of linear matrix equations.* Proceedings of the Cambridge Philosophical Society **52**, 17 – 19.

[13] Rocci R., Vichi M. (2003). *Three-mode clustering of a three-way data set.* CLADAG 2003, University of Bologna, Bologna.

[14] Schepers J., Van Mechelen I. (2004). *Three-mode partitioning: Method and application.* Paper presented at the meeting of the GfKl, Dortmund, March 9-11.

[15] Ten Berge J.M.F. (1993). *Least squares optimization in multivariate analysis.* DSWO Press, Leiden.

[16] Tucker L.R. (1966). *Some mathematical notes on three-mode factor analysis.* Psychometrika **31**, 279 – 311.

[17] Vichi M. (2001) *Double k-means Clustering for simultaneous classification of objects and variables.* In: Borra S., Rocci R., Schader M. (eds.): Advances in classification and data analysis, Springer, Heidelberg.

Acknowledgement: The author is obliged to Roberto Rocci, Jan Schepers, Marieke Timmerman, Iven van Mechelen, and Maurizio Vichi.

Address: Henk A.L. Kiers, Heymans Institute, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

E-mail: `h.a.l.kiers@ppsw.rug.nl`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

FUNCTIONAL DATA ANALYSIS AND MIXED EFFECT MODELS

Alois Kneip, Robin C. Sickles and Wonho Song

Key words: Mixed effects model, functional principal component analysis, nonparametric regression.

COMPSTAT 2004 section: Functional data analysis.

Abstract: Panel studies in econometrics as well as longitudinal studies in biomedical applications provide data from a sample of individual units where each unit is observed repeatedly over time (age, etc.). In this context, mixed effect models are often applied to analyze the behavior of a response variable in dependence of a number of covariates. In some important applications it is necessary to assume that individual effects vary over time (age, etc.).

In the paper it is shown that in many situations a sensible analysis may be based on a semiparametric approach relying on tools from functional data analysis. The basic idea is that time-varying individual effects may be represented as a a sample of smooth functions which can be characterized by its Karhunen-Loève decomposition. An important application is the estimation of time-varying technical inefficiencies of individual firms in stochastic frontier analysis.

1 Introduction

Panel studies in econometrics as well as longitudinal studies in biomedical applications provide data from a sample of individual units where each unit is observed repeatedly over time (age, etc.). Statistical analysis then usually aims to model the variation of some response variable Y. In addition to its dependence on some vector of explanatory variables X, the variability of Y between different individual units is of primary interest.

For simplicity, we will assume a balanced design with T equally spaced repeated measurements per individual. The resulting observations of n individuals can then be represented in the form (Y_{it}, X_{it}), where $t = 1, \dots T$ and $i = 1, \dots, n$. The simplest form of analysis is based on mixed effect models of the form

$$Y_{it} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{itj} + u_i + \epsilon_{it} \tag{1}$$

where ϵ_{it} are i.i.d. error terms, while u_i represents individual random effects.

An important example in econometrics are stochastic frontier models. Then Y_{it} represents production output of an individual firm i in time period t, while X_{it} is a corresponding vector of production inputs. The u_i are then interpreted as technical inefficiencies. Firm i is more efficient than firm j if $u_i > u_j$.

However, in many applications it is too simple to assume constant individual effects u_i. A straightforward generalization is to suppose that $u_i \equiv u_i(t)$ is a function of t.

$$Y_{it} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{itj} + u_i(t) + \epsilon_{it} \tag{2}$$

In the following we will assume that the $u_i(t)$ can be considered as smooth random functions. In many biometrical applications, where for example t indicates age of an individual unit, smoothness can be considered as a standard assumption. In econometrics, where t usually indicates time, for a given unit i the corresponding data $\{Y_{it}, X_{it}\}$, $t = 1, \ldots, T$, represent an individual time series. In this situation model (2) assumes that the residual time series $\{Y_{it} - \beta_0 - \sum_{j=1}^{p} \beta_j X_{itj}\}$, $i = 1, \ldots, n$, can be decomposed into a smooth stochastic trend u_i and i.i.d. white noise.

Traditional analysis relies on parametric models. Very often polynomial approximations to the functions u_i are used. More generally, for some prespecified basis functions $b_1, \ldots, b_L$ the u_i are modelled by $u_i(t) = \sum_r \vartheta_{ir} b_r(t)$, where $\vartheta_{i1}, \ldots, \vartheta_{iL}$ are individual random coefficients. Analysis is then based on the well-known methodology of mixed effect models. If additionally normality is assumed and if X and ϵ are uncorrelated, likelihood estimation based on the EM algorithm is often applied. In stochastic frontier analysis such an approach has been used by Battese and Coelli [1] or Cornwell, Schmidt, and Sickles [2] in order to model time-dependent individual inefficiencies.

In this paper we consider a **nonparametric** approach based on ideas from functional data analysis as proposed by Kneip, Sickles and Song [6]. The functions u_i can be decomposed into $u_i = w_i + v_i$, where $w(t)$ is a general mean function and $v_i(t) = u_i(t) - w(t)$. Model (2) can then be rewritten in the form

$$Y_{it} = \sum_{j=1}^{p} \beta_j X_{itj} + w(t) + v_i(t) + \epsilon_{it} \tag{3}$$

Note that the constant β_0 is incorporated into $w(t)$, and that the mean of $v_i(t)$ is zero.

For a given L functional principal component analysis is then used to estimate a *best possible basis* $g_1, \ldots, g_L$ for approximating v_i by $v_i(t) = \sum_{r=1}^{L} \theta_{ir} g_r(t)$. The approach possess a number of advantages

- The basis $g_1, \ldots, g_L$ to be estimated corresponds to the best possible basis for approximating the v_i by an L-dimensional linear function space. Any approximation $v_i(t) \approx \sum_{r=1}^{L} \vartheta_{ir} b_r(t)$ based on prespecified basis functions $b_1, \ldots, b_L$ (e.g. polynomials or splines) possesses a higher systematic error.

- All $n \cdot T$ observations are used to estimate $g_1, \dots, g_L$. Compared to a completely nonparametric analysis based on simply estimating all v_i by nonparametric regression these functions can be estimated with a much higher degree of accuracy.

Functional principal components are widely used in functional data analysis (see for example [7]). It must be emphasized, however, that the present situation is different from the usual setup in this domain, since the functions v_i of interest are not directly observed. This constitutes a major complication.

The paper is organized as follows. Section 2 presents the theoretical basis of our approach relying on the Karhunen-Loève decomposition. An algorithm for determining g_r and coefficients β_j, θ_{ir} as proposed by Kneip, Sickles and Song [6] is described in Section 3. Section 3.2 presents a new procedure which may be considered as a promising alternative. Section 4 is devoted to the problem of choosing an optimal dimension L.

2 Functional principal components

Let generally $\nu_1, \dots, \nu_n$ be i.i.d. smooth random function on $L^2[0,1]$ and suppose that $E(\nu_i) = 0$. Furthermore, let $\|f\| = \sqrt{\int f(t)^2}$ denote the usual L^2-norm for $f \in L^2[0,1]$, and set $< f^*, v >= \int f^*(t) f(t) dt$. The covariance operator then is a generalization of the concept of a covariance matrix in multivariate analysis of random vectors. The so-called covariance kernel is defined as

$$\sigma(s,t) = E(\nu_i(s)\nu_i(t))$$

and the corresponding covariance operator Γ is defined by the relation

$$\Gamma v = E\left(< \nu_i, v > \nu_i\right) = \int \sigma(s,t) v(s) ds$$

for any function $v \in L^2[0,1]$. Γ is a Hilbert-Schmidt operator and possesses finite eigenvalues $l_1 \geq l_2 \geq \dots$ as well as corresponding orthonormal eigenfunctions $\gamma_1, \gamma_2, \dots$ such that $\|\gamma_r\| = 1$ and $< \gamma_r, \gamma_s > 0 =$ for $r \neq s$. A precise mathematical discussion of properties of Γ can, for example, be found in Gihman and Skorohod [4].

The well known Karhunen-Loève decomposition states that the functions ν_i can be decomposed in terms of the eigenfunctions:

$$\nu_i(t) = \sum_r \vartheta_{ir} \gamma_r(t) \tag{4}$$

where $\vartheta_{ir} =< \nu_i, \gamma_r >$. This decomposition posseses the following properties (see for example [4]):

a) $E(\vartheta_{ir}) = 0$, $r = 1, 2, \dots$, and $Var((\vartheta_{i1}) = l_1 \geq Var((\vartheta_{i2}) = l_2 \geq Var((\vartheta_{i3}) = l_3 \geq \ \dots$

b) ϑ_{ir} is uncorrelated with ϑ_{is} if $r \neq s$

c) For each $L = 1, 2, \ldots$

$$\sum_{r>L} l_r = E\left(\|\nu_i - \sum_{r=1}^{L} \vartheta_{ir}\gamma_r\|\right) \leq E\left(\min_{\alpha_{i1},\ldots,\alpha_{iL}} \|\nu_i - \sum_{r=1}^{L} \alpha_{ir} b_r\|\right) \quad (5)$$

for any possible choice of basis functions $b_1, \ldots, b_L \in L^2[0,1]$.

Uncorrelatedness of the random coefficients ϑ_{ir} for different r simplifies further analysis, which may, for example, rely on the EM algorithm. Note that this is a specific property of the Karhunen-Loève basis. For any *prespecified* basis $b_1, \ldots, b_L$ one will have to take into account that the resulting coefficients are usually correlated.

Property c) may be seen as the most important feature of (4). For any possible dimension L the decomposition provides the best possible basis $\gamma_1, \ldots, \gamma_L$ for approximating the random functions ν_i by a linear combination of L functions. Indeed, it is well-known that in many situation a relatively small number L of components is sufficient to model the underlying functions such that a model of the form

$$\nu_i(t) = \sum_{r=1}^{L} \vartheta_{ir}\gamma_r(t) \quad (6)$$

holds in a good approximation.

Of course, the major problem of (6) consists in the fact that the function γ_r as well as an appropriate dimension L are unknown. In functional data analysis it is usually assumed that n functional realizations can be observed, or at least can be approximated with a negligible error. Estimates $\hat{\gamma}_r$ can then determined from the empirical covariance operator $\Gamma_n v = \frac{1}{n}\sum_{i=1}^{n}(<\nu_i, v>\nu_i)$ Some asymptotic theory is given in Dauxois, Pousse and Romain [3]. Under some additional conditions it is shown that rates of convergence of estimated eigenvalues and empirical eigenfunctions $\gamma_{r,n}$ are of order $n^{-1/2}$.

The present situation is different, since one has to deal with $n \cdot T$ noisy observations. The major point of interest is modelling the functions $v_i(t)$ at the design points $t = 1, \ldots, T$. We may formalize smoothness of $v_1, \ldots, v_n$ by requiring that there are i.i.d. smooth random functions $\nu_1, \ldots, \nu_n \in L^2[0,1]$ with $v_i(t) = \nu_i(\frac{t}{T})$.

Discretizing (6) then leads to the model

$$v_i(t) = \sum_{r=1}^{L} \theta_{ir} g_r(t), \quad t = 1, \ldots, T, \; i = 1, \ldots, n \quad (7)$$

Empirical versions of properties a) and b) as well as of orthonormality of $\gamma_1, \gamma_2, \ldots$ are then obtained by requiring

(α) $\sum_i \theta_{i1}^2 \geq \sum_i \theta_{i2}^2 \geq \dots$

(β) $\sum_i \theta_{ir}\theta_{is} = 0$ for $r \neq s$.

(γ) $\frac{1}{T}\sum_{t=1}^T g_r(t)^2 = 1$ and $\sum_{t=1}^T g_r(t)g_s(t) = 0$ for all $r, s \in \{1, \dots, L\}$ with $r \neq s$.

Moreover, a discretized version of (5) is given by

$$\frac{1}{n}\sum_{i=1}^n \sum_{t=1}^T (v_i(t) - \sum_{r=1}^L \theta_{ir} g_r(t))^2 \leq \frac{1}{n}\sum_{i=1}^n \min_{\alpha_{i1},\dots,\alpha_{iL}} \left(\sum_{t=1}^T (v_i(t) - \sum_{r=1}^L \alpha_{ir} b_r(t))^2 \right) \tag{8}$$

for any possible choice of $b_r(t)$, $t = 1, \dots, T$, $r = 1, \dots, L$.

Note that Conditions (α) - (γ) do not impose any restriction, and they introduce a suitable normalization which ensures identifiability of the components up to sign changes (instead of θ_{ir}, g_r one may also use $-\theta_{ir}, -g_r$). If (6) holds for some suitable L, then there exist some g_r such that (7) as well as (α) - (γ) and (8) are satisfied.

Obviously the components g_r depend on the realized v_i and on the sample size n. Due to different normalization usually $g_r(t) \neq \gamma_{r,n}(\frac{t}{T})$. This does not constitute a serious drawback for an empirical analysis based on (3) and (7). In fact, in model (7) only the L dimensional linear space spanned by $g_1, \dots, g_L$ is identifiable. There are infinitely many possible choices of basis functions, and by using conditions (1) - (3) we select a particularly well-interpretable basis. Asymptotically, as $N, T \to \infty$ $g_r(t)$ as well as $\gamma_{r,n}(t)$ will both converge to $\gamma_t(t)$ in probability. Under (6) the linear subspaces of $\mathbb{R}^T$ spanned by the vectors $\{(g_r(1), \dots, g_r(T))'\}_{r=1,\dots,L}$, $\{(\gamma_{r,n}(1), \dots, \gamma_{r,n}(T))'\}_{r=1,\dots,L}$ and $\{(\gamma_r(1), \dots, \gamma_r(T))'\}_{r=1,\dots,L}$ will coincide with high probability for large samples

How to determine the functional components g_r in (7)? There are essentially two straightforward procedures which could immediately be applied if the realized functions v_i where known. These algebraic methods will serve as a basis of the practical, data-based methods to be presented in Sections 3.

Method 1: Some simple algebra shows that, if the v_i were known, the components g_r could be determined from the eigenvectors of the empirical covariance matrix Σ_n of $v_1 = (v_1(1), \dots, v_1(T))', \dots, v_n = (v_n(1), \dots, v_n(T))'$:

$$\Sigma_n = \frac{1}{n}\sum_i v_i v_i{}' \tag{9}$$

Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_T$ as well as $\gamma_1, \gamma_2, \dots, \gamma_T$ to denote the resulting eigenvalues and orthonormal eigenvectors of Σ_n. Then

$$\lambda_r = \frac{T}{n}\sum_i \theta_{ir}^2 \quad \text{for all} \quad r = 1, 2, \dots, L, \tag{10}$$

$$g_r(t) = \sqrt{T} \cdot \gamma_{rt} \quad \text{for all} \quad r = 1, \dots, L, \ t = 1, \dots, T. \tag{11}$$

Also note that $\sum_{j=L+1}^{n} \lambda_j = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}(v_i(t) - \sum_{r=1}^{L}\theta_{ir}g_r(t))^2$. If (7) holds, then obviously $\sum_{j=L+1}^{n}\lambda_j = 0$

Method 2: A second possibility is to consider the $n \times n$ matrix M_n defined by

$$(M_n)_{i,j} = \frac{1}{T}\sum_{t=1}^{T} v_i(t)v_j(t), \quad i,j = 1,\ldots,n \tag{12}$$

By using some further algebra, see for example [5], one can then deduce that all nonzero eigenvalues λ_r and h_r of the empirical covariance Σ_n and of the matrix M_n are related by $h_r = \sum_i \theta_{ir}^2 = \frac{n}{T}\lambda_r$. Moreover, the eigenvectors $\mathbf{p}_1 = (p_{11},\ldots,p_{n1})', \mathbf{p}_2 = (p_{12},\ldots,p_{n2})', \ldots$ of M_n corresponding to nonzero eigenvalues $h_1 \geq h_2 \geq \ldots$ are closely related to the parameters θ_{ir} since

$$\theta_{ir} = h_r^{1/2}\, p_{ir} \tag{13}$$

Finally, g_r can be computed from λ_r and and p_{ir}:

$$g_r(t) = h_r^{-1/2}\sum_{i}^{n} p_{ir}\, v_i(t) = \frac{\sum_i^n \theta_{ir}\, v_i(t)}{\sum_i^n \theta_{ir}^2} \tag{14}$$

3 Algorithms

When combining 3) and (7) one obtains

$$Y_{it} = \sum_{j=1}^{p}\beta_j X_{itj} + w(t) + \sum_{r=1}^{L}\theta_{ir}g_r(t)) + \epsilon_{it} \tag{15}$$

The optimal basis functions g_r satisfying (7)- (11) as well as w, β_j and θ_{ir} are unknown.

Based on the mathematical framework of Section 2 different algorithms can be applied in order to estimate the components w and g_r of the (15). In this section we will rely on a prespecified dimension L. The important question of determining an appropriate L will be considered in Section 4.

3.1 An algorithm based on estimating the covariance matrix Σ_n

In the following we will discuss a straightforward method which can be seen as a simple version of a somewhat more general algorithm proposed by Kneip, Sickles and Song [6].

The idea is easily described: In a first step partial spline methods as introduced by Speckman [8] are used to determine estimates $\hat{\beta}_j$ and $\hat{v}_i$. The mean function w is estimated nonparametrically, and then estimates $\hat{g}_r$ are determined from the empirical covariance matrix $\hat{\Sigma}_n$ of $\hat{v}_1,\ldots,\hat{v}_n$.

Let us first introduce some additional notations. Let $\bar{Y}_t = \frac{1}{n}\sum_i Y_{it}$, $\bar{Y} = (\bar{Y}_1, \dots, \bar{Y}_T)'$, $Y_i = (Y_{i1} \dots, Y_{iT})'$ and $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{iT})$. Furthermore, let $X_{ij} = (X_{i1j}, \dots, X_{iTj})'$, $\bar{X}_{tj} = \frac{1}{n}\sum_i X_{itj}$, and $\bar{X}_j = (\bar{X}_{1j}, \dots, \bar{X}_{Tj})'$. We will use X_i and $\bar{X}$ to denote the $T \times p$ matrices with elements X_{itj} and $\bar{X}_{tj}$.

The algorithm now can be described as follows:

Step 1: Determine estimates $\hat{\beta}_1, \dots, \hat{\beta}_p$ and $\hat{v}_i(t)$ by minimzing

$$\sum_i \sum_t (y_{it} - \bar{y}_t - \sum_{j=1}^{p} \beta_j (x_{itj} - \bar{x}_{tj}) - u_i(t))^2 + \sum_i \kappa \int_1^T (v_i''(s))^2 ds \tag{16}$$

where $\kappa > 0$ is a preselected smoothing parameter and v_i'' denotes the second derivative of v.

Spline theory implies that any solution $\hat{v}_i$, $i = 1, \dots, n$ of (16) possess an expansion $\hat{v}_i(t) = \sum_j \hat{\zeta}_{ji} z_j(t)$ in terms of a natural spline basis $z_1, \dots, z_T$. If Z and A denote $T \times T$ matrices with elements $z_j(t)$ and $\int_1^T z_j''(s) z_j''(t)$, the above minimization problem can be reformulated in matrix notation: Determine $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$ and $\hat{\zeta}_i = (\hat{\zeta}_{1i}, \dots, \hat{\zeta}_{Ti})'$ by minimizing

$$\sum_i \left(\|Y_i - \bar{Y} - (X_i - \bar{X})\beta - Z\zeta_i\|_2^2 + \kappa \zeta_i' A \zeta_i \right), \tag{17}$$

where $\| \cdot \|_2$ denotes the usual euclidean norm in $\mathbb{R}^T$, $\|a\|_2 = \sqrt{a'a}$.

It is easily seen that with

$$\mathcal{Z}_\kappa = Z(Z'Z + \kappa A)^{-1} Z'$$

the solutions are given by

$$\hat{\beta} = \left(\sum_i (X_i - \bar{X})'(I - \mathcal{Z}_\kappa)(X_i - \bar{X}) \right)^{-1} \sum_i (X_i - \bar{X})'(I - \mathcal{Z}_\kappa)(Y_i - \bar{Y}) \tag{18}$$

as well as

$$\hat{\zeta}_i = (Z'Z + \kappa A)^{-1} Z'(Y_i - \bar{Y} - (X_i - \bar{X})\hat{\beta}).$$

Therefore,

$$\hat{v}_i = Z\hat{\zeta}_i = \mathcal{Z}_\kappa (Y_i - \bar{Y} - (X_i - \bar{X})\hat{\beta}) \tag{19}$$

estimates $v_i = (v_i(1), \dots, v_i(T))'$.

Remarks:

- An obvious problem is the choice of κ. A straightforward approach then is to use (generalized) cross-validation procedures in order to estimate an optimal smoothing parameter $\hat{\kappa}_{opt}$. Note, however, that the goal is not to obtain optimal estimates of the $v_i(t)$ but to approximate the functions g_r in (15). Estimating g in the subsequent steps of the algorithm involves a specific way of averaging over individual data which substantially reduces variability. In order to reduce bias, a small degree of undersmoothing, i.e. choosing $\kappa < \hat{\kappa}_{opt}$, will usually be advantageous.
- Our setup is based on assuming a balanced design. However, in practice one will often have to deal with the situation that there are missing observations for some individuals. In principle, the above estimation procedure can easily be adapted to this case. If for an individual k observations are missing, then only the remaining $T-k$ are used for minimizing (16). Estimates of $\hat{v}_i(t)$ at all $t = 1, \ldots, T$ are then obtained by spline interpolation.

Step 2: An estimate $\hat{w}$ of the mean function w is calculated by minimizing

$$\sum_t \left(\bar{Y}_t - \sum_{j=1}^{p} \hat{\beta}_j \bar{X}_{tj} - w(t) \right)^2 + \kappa \int_1^T (w''(s))^2 ds.$$

Step 3: Determine the empirical covariance matrix $\hat{\Sigma}_n$ of $\hat{v}_1 = (\hat{v}_1(1), \hat{v}_1(2), \ldots, \hat{v}_1(T))', \ldots, \hat{v}_n = (\hat{v}_n(1), \hat{v}_n(2), \ldots, \hat{v}_n(T))'$ by

$$\hat{\Sigma}_n = \frac{1}{n} \sum_i \hat{v}_i \hat{v}_i'$$

and calculate its eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \hat{\lambda}_T$ and the corresponding eigenvectors $\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_T$.

Step 4: Set $\hat{g}_r(t) = \sqrt{T} \cdot \hat{\gamma}_{rt}$, $r = 1, 2, \ldots, L$, $t = 1, \ldots, T$, and for all $i = 1, \ldots, n$ determine $\hat{\theta}_{1i}, \ldots, \hat{\theta}_{Li}$ by minimizing

$$\sum_t (Y_{it} - \bar{Y}_t - (X_i - \bar{X})\hat{\beta} - \sum_{r=1}^{L} \vartheta_{ri} \hat{g}_r(t))^2$$

with respect to $\vartheta_{1i}, \ldots, \vartheta_{Li}$.

Based on this algorithm the unknown model components w and g_r in (15) can be replaced by $\hat{w}$ and $\hat{g}_r$. Further analysis may then be based on the "estimated" model

$$Y_{it} \approx \sum_{j=1}^{p} \beta_j X_{itj} + \hat{w}(t) + \sum_{r=1}^{L} \theta_{ir} \hat{g}_r(t)) + \epsilon_{it} \tag{20}$$

The algorithm automatically also yields estimates $\hat{\beta}_j$ and $\hat{\theta}_{ir}$. However, variability of these estimates may be reduced by re-estimating these coefficients by relying on (20):

Step 5: Re-estimate the coefficients $\hat{\beta}_j$ and $\hat{\theta}_{ir}$ by fitting the estimated model $Y_{it} = \sum_{j=1}^{p} \beta_j X_{itj} + \hat{w}(t) + \sum_{r=1}^{L} \theta_{ir}\hat{g}_r(t)) + \epsilon_{it}$ to the data.

Kneip, Sickles and Song [6] also study the asymptotic behavior of the resulting estimators as $n, T \to \infty$.

Let $\kappa_T = T\kappa$. If the underlying function ν_i, as discussed in Section 2, is twice continuously differentiable, then the bias in estimating v_i is of order κ_t, while variance is of order $\frac{1}{\kappa_T^{1/4} T}$. Choosing κ_T to be of order $T^{-4/5}$ then leads to the optimal *individual* rates of convergence $\frac{1}{T}\sum_t (\hat{v}_i(t) - v_i(t))^2 = O_P(T^{-4/5})$.

Under some technical assumptions (mainly concerning smoothness as well as the correlation between X_{it} and $v_i(t)$) a theorem by Kneip, Sickles and Song [6] then implies that for all $r = 1, \ldots, L$

$$T^{-1} \sum_{t=1}^{T} \left(g_r(t) - \hat{g}_r(t)\right)^2 = O_P\left(\kappa_T + \frac{1}{T^2} + \frac{1}{\kappa_T^{1/4} nT}\right) \tag{21}$$

Further results concern rates of convergence and asymptotic distributions of parameter estimates. As can be seen from (21) variance of $\hat{g}_r$ also decreases with the number n of individual units. By undersmoothing, i.e. choosing $\kappa_T = o(T^{-4/5})$, the components g_r can be estimated with better rates of convergence than those obtainable for the individual functions v_i.

In Kneip, Sickles and Song [6] finite sample performance of the estimators is additionally examined via Monte Carlo simulations. The method is then applied to the analysis of technical efficiency of the U.S. banking industry.

3.2 An algorithm based on estimating the matrix M_n

Model (3) obviously implies that

$$v_i(t) = Y_{it} - \sum_{j=1}^{p} \beta_j X_{itj} - w(t) - \epsilon_{it}$$

Hence, if the parameters β_j were known, the matrix

$$(\tilde{M}_n)_{i,j} = \frac{1}{T} \sum_{t=1}^{T} (Y_{it} - \bar{Y}_t - \sum_{j=1}^{p} \beta_j (X_{itj} - \bar{X}_{tj}), \quad i, j = 1, \ldots, n \tag{22}$$

provides an estimate of M which by Method 2) discussed in Section 2 can be used to calculate estimates of g_r.

The basic idea of the following algorithm is now easily described: Under (15) the "true" matrix M possesses only L nonzero eigenvalues, and therefore $\sum_{j=L+1}^{n} \lambda_j = 0$. Based on (22), different matrices $\tilde{M}_n(\beta)$ can be determined in dependence of all possible values of β_j. Estimates $\hat{\beta}_j$ and $\hat{M}_n$ can be obtained by minimzing the sum of the smallest $n - L$ eigenvalues of $\tilde{M}(\beta)$ with respect to β.

The precise algorithm now can be described as follows:

Step 1*: For all possible values $\tilde{\beta}_j$ of β_j, $j = 1, \ldots, p$ compute

$$(\tilde{M}_n(\tilde{\beta})_{i,j} = \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - \bar{Y}_t - \sum_{j=1}^{p}\tilde{\beta}_j(X_{itj} - \bar{X}_{tj}), \quad i, j = 1, \ldots, n$$

and its eigenvalues $h(\tilde{\beta})_1 \geq h(\tilde{\beta})_2 \geq \cdots \geq h(\tilde{\beta})_n$.

Then determine estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$ by minimizing

$$\sum_{j=L+1}^{n} h(\tilde{\beta})_j$$

with respect to $\tilde{\beta}$.

Step 2*: Set $\hat{M}_n = \tilde{M}_n(\hat{\beta})$ and determine eigenvalues $\hat{h}_1 \geq \hat{h}_2 \geq \ldots$ and corresponding orthonormal eigenvectors $\hat{p}_1, \ldots, \hat{p}_n$.

Estimates $\hat{g}_r$ are then calculated by a weighted sum of residuals:

$$\hat{g}_r(t) = \hat{\lambda}_r^{-1/2}\sum_{i}^{n}\hat{p}_{ir}\left((Y_{it} - \bar{Y}_t - \sum_{j=1}^{p}\hat{\beta}_j(X_{itj} - \bar{X}_{tj})\right) \tag{23}$$

In spite of averaging over individuals, (23) may lead to fairly noisy estimates of g_r. Some addition smoothing will usually improve the performance of the estimator. Using a spline approach, an estimate of g_r may thus alternatively be determined by minimizing

$$\sum_{t}\left(\lambda_r^{-1/2}\sum_{i}^{n}\hat{p}_{ir}(Y_{it} - \bar{Y}_t - \sum_{j=1}^{p}\hat{\beta}_j(X_{itj} - \bar{X}_{tj}) - g_r(t)\right)^2 + \kappa\int_1^T (g''(s))^2 ds$$

instead of using (23).

Step 3*: An estimate $\hat{w}$ of the mean function w is calculated by minimizing

$$\sum_{t}\left(\bar{Y}_t - \sum_{j=1}^{p}\hat{\beta}_j\bar{X}_{tj} - w(t)\right)^2 + \kappa\int_1^T (w''(s))^2 ds.$$

As in the procedure of Section 3.1 accuracy of coefficient estimates may be improved by a final re-estimation:
Step 4*: Re-estimate the coefficients $\hat{\beta}_j$ and $\hat{\theta}_{ir}$ by fitting the estimated model $Y_{it} = \sum_{j=1}^{p} \beta_j X_{itj} + \hat{w}(t) + \sum_{r=1}^{L} \theta_{ir}\hat{g}_r(t)) + \epsilon_{it}$ to the data.

Recall that that the procedure of Section 3.1 requires smoothing of the *individual* data of each of the n units in order to estimate v_i, $i = 1, \ldots, n$. An important advantage of the above algorithm thus is that it only requires some *global* smoothing over weighted averages of observations in Steps 3* and 4*. The choice of the smoothing parameter κ will thus be less critical, and a possible smoothing bias will not affect the estimates of the parameters β_j. One may expect a superior behavior of this method if the number T of repeated measurement is fairly small.

On the other hand, a drawback is the fact that already for estimating β_j in Step 1* a sensible selection of the dimension L in (15) has to be made. Indeed, usually 15) will have to be satisfied in a very good approximation in order to avoid biased estimates of the parameters. In practice, one may apply the algorithm for different values of L and choose an appropriate dimension by using some goodness-of-fit criterion.

Theoretical properties of the above algorithm have not yet been studied and remain a topic of future research.

4 Choice of dimension

Any analysis based on (15) requires a sensible choice of the dimension L. If L is too small, there may exist a large systematic error in approximating the v_i. On the other hand, if L is too large, then estimates will possess an unnecessarily large variance.

Note that for a given sample the eigenvalues of the estimated covariance matrix $\hat{\Sigma}_n$ will usually satisfy $\hat{\lambda}_r > 0$ for $r > L$. This will even be true if (15) holds exactly and if therefore the eigenvalues of true matrix Σ_n are such that $\lambda_r = 0$ for $r > L$. In other words, the noise term ϵ_{it} will “create” additional (small) components in the PCA decomposition. It is obvious that any component generated or strongly influenced by noise should *not* be included into model (15).

From this point of view one may tend to choose L in such a way that each component g_r, $r = 1, \ldots, L$, possesses an influence on the model fit which is significantly larger than that of any noise component. This idea has been adopted by Kneip, Sickles and Song [6] in order to estimate a dimension L. Under the hypothesis that (15) holds for some L, i.e. $\sum_{r=L+1}^{n} \lambda_r = 0$, they derive asymptotic approximations of mean $m(L)$ and variance $s(L)^2$ of $\sum_{r=L+1}^{h} \hat{\lambda}_r$, and it is shown that $C(L) = \frac{\sum_{r=L+1}^{h} \hat{\lambda}_r - m(L)}{s(L)}$ asymptotically possesses a standard normal distribution. For any possible value of L, $m(L)$ and $s(L)$ can be approximated from the data.

An estimate of L is then obtained by choosing the smallest $l = 1, 2, \ldots$ such that

$$C(l) \leq z_{1-\alpha},$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of a standard normal distribution.

References

[1] Battese G.E., Coelli T.J. (1992). *Frontier production functions, technical efficiency and panel data: With application to paddy farmers in India.* Journal of Productivity Analysis **3**, 153 – 169.

[2] Cornwell C., Schmidt P., Sickles R.C. (1990). *Production frontiers with cross-sectional and time-series variation in efficiency levels.* Journal of Econometrics **46**, 185 – 200.

[3] Dauxois J., Pousse A., Romain Y. (1982). *Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference.* Journal of Multivariate Analysis **12**, 136 – 154.

[4] Gihman I.I., Skorohod A.V. (1970). *The theory of stochastic processes.* New York: Springer.

[5] Good I.J. (1969). *Some applications of the singular value decomposition of a matrix.* Technometrics **11** 823 – 831.

[6] Kneip A., Sickles R.C., Song W. (2004). *On estimating the mixed effects model.* Manuscript.

[7] Ramsay J.O., Silverman B.W. (1997). *Functional data analysis.* New York: Springer.

[8] Speckman P. (1988). *Kernel smoothing in partial linear models.* Journal of the Royal Statistical Society, Series B **50**, 413 – 436.

Address: A. Kneip, Fachbereich Rechts- und Wirtschaftswissenschaften, Universität Mainz, 55099 Mainz, Germany
R.C. Sickles, W. Song, Department of Economics - MS 22, Rice University, 6100 S. Main Street, Houston, TX 77005-1892, USA

E-mail: kneip@uni-mainz.de

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

USING WEIGHTS WITH A TEXT PROXIMITY MATRIX

Angel R. Martinez, Edward J. Wegman and Wendy L. Martinez

Key words: Bigram proximity matrix, k nearest neighbors classifier, natural language processing.

COMPSTAT 2004 section: Applications, Classification.

Abstract: In previous work, we introduced a way of encoding free-form documents called the bigram proximity matrix (BPM). When this encoding was used on a corpus of documents, where each document is tagged with a topic label, results showed that the documents could be classified based on their tagged meaning. In this paper, we investigate methods of weighting the elements of the BPM, analogous to the weighting schemes found in natural language processing. These include logarithmic weights, augmented normalized frequency, inverse document frequency and pointwise mutual information. Results presented in this paper show that some of the weights increased the proportion of correctly classified documents.

1 Introduction

The bigram proximity matrix (BPM) was first developed by Martinez and Wegman [8], [9], [10] as a way of encoding text so it can be used in applications such as document clustering, classification or information retrieval. Previous studies with the BPM indicated that documents can be successfully classified using k nearest neighbors and other methods when they are encoded in this way. The objective of the current work is to define bigram weights analogous to the term weights found in natural language processing and to investigate the utility of using them in document classification.

In Section 2, we present some background information on the BPM and include an illustrative example. We then provide definitions of the bigram weights in Section 3. Section 4 contains information about the experiments that were conducted, as well as the results. Finally, we offer a summary and some comments about future work in Section 5.

2 Bigram proximity matrix

The BPM is a non-symmetric matrix that captures the number of word co-occurrences in a moving 2-word window. It is a square matrix whose column and row headings are the alphabetically ordered entries of the lexicon, plus one more element for end of sentence punctuation. The BPM matrix element ij is the number of times word i appears immediately before word j in the unit of text. The size of the BPM is determined by the size of the

	.	crowd	his	in	father	man	sought	the	wise	young
.										
crowd	1									
his					1					
in								1		
father				1						
man							1			
sought			1							
the		1							1	
wise										1
young						1				

Table 1: Example of Bigram Proximity Matrix. (Note: Zeros in empty boxes are removed for clarity.)

lexicon created by listing alphabetically the unique occurrences of the words in the text. Additionally, it should be noted that all end of sentence punctuation is replaced with a period, and the period is treated as a word. By convention, the period is designated as the first word in the ordered lexicon. It is asserted that the BPM representation of the semantic content preserves enough unique features to be semantically separable from BPMs of other thematically unrelated collections.

The rows in the BPM represent the first word in the pair, and the second word is given by the column. For example, the BPM for the sentence or text stream,

$$The\ wise\ young\ man\ sought\ his\ father\ in\ the\ crowd.$$

is shown in Table 1. We see that the matrix element located in the third row (*his*) and the fifth column (*father*) has a value of one. This means that the pair of words '*his father*' occurs once in this unit of text. It should be noted that in most cases, depending on the size of the lexicon and the size of the text stream, the BPM will be very sparse. So, while the dimensionality of the BPM can be very large, sparse matrix techniques makes the analysis fast and the storage requirements small.

3 Definition of weights

We can see from the definition of the BPM, that the elements of the matrix represent the number of times that a bigram or word pair occurs in the document. Some of the measures of semantic similarity for classification cited in Martinez [8] employed the raw frequencies, others used binary values (if the frequency is non-zero, then it is replaced with a 1), and some required conversion to probabilities or relative frequencies. In this paper, we will only be concerned with the first case, where raw bigram frequencies are compared to weighted values. Because of this, we will use one measure of semantic similarity - the *normalized correlation coefficient* (NCC). This is similar to the cosine measure used in information retrieval [7].

Let **A** represent a BPM that has been converted to a column vector by concatenating the columns, one on top of the other. We do this conversion so the usual definition of the normalized correlation coefficient can be used. Let **C** denote another BPM that has been similarly converted to a vector. The cosine of the angle between these two 'vectors' is given by

$$\mathrm{NCC} = \cos\theta_{\mathbf{AC}} = \frac{\mathbf{A}^T\mathbf{C}}{\|\mathbf{A}\|\,\|\mathbf{C}\|} = \frac{\sum_{i=1}^{M} a_i c_i}{\sqrt{\sum a_i^2}\sqrt{\sum c_i^2}}, \tag{1}$$

where $\|\mathbf{A}\|$ denotes the magnitude of vector **A**, and M is the number of words in the lexicon squared, i.e., the total number of elements in the BPM.

The NCC given in Equation 1 is a similarity measure, whose range in this case is between 0 and 1. Larger values of the NCC correspond to observations that are close together. For example, the NCC similarity between a document BPM and itself is 1. If the two document BPM 'vectors' are orthogonal to each other, then the NCC similarity is 0. We convert the NCC similarity values to Euclidean distance using the following transformation

$$d_{ij} = \sqrt{2(1 - s_{ij})}\,, \tag{2}$$

where s_{ij} represents the similarity between document i and j, and d_{ij} is the distance between document i and document j.

3.1 Local – global – document weights

We will denote the ij-th element (the ij-th bigram or word pair) of the k-th weighted BPM as a_{ijk}. We can write this in terms of local, global and document components as follows

$$a_{ijk} = l_{ijk} g_{ij} d_k, \tag{3}$$

where l_{ijk} is the local weight for bigram ij that occurs in document k, g_{ij} is the global weight for bigram ij in the corpus, and d_k is a document normalization factor. We represent the frequency or the number of times bigram ij appears in document k as f_{ijk}. We use the following to indicate the conversion of a frequency f to a binary value:

$$I(f) = \begin{matrix} 1 \text{ if } f > 0 \\ 0 \text{ if } f = 0 \end{matrix} \tag{4}$$

The two local weights we use are called the *logarithmic* and the *augmented normalized bigram frequency.* Before we define these, we make one small change in notation for ease of understanding. We denote the ij-th bigram with the subscript b, where some arbitrary order or labeling has been imposed on the bigrams (elements of the BPM). The logarithmic weight is defined as

$$l_{bk} = l = \log(1 + f_{bk}), \tag{5}$$

and the augmented normalized bigram frequency is given by

$$l_{bk} = n = \frac{I(f_{bk}) + f_{bk} \div \max_b\{f_{bk}\}}{2}. \tag{6}$$

If no local weights are used, then we denote that as just the bigram frequency

$$l_{bk} = t = f_{bk}. \tag{7}$$

Note that the letters l, t, and n are used in the information retrieval literature to denote the type of local weight [1].

We use only one global weight in this study called the *inverse document frequency* (IDF); others can be found in Berry and Browne [1]. The IDF for bigrams is defined as

$$g_b = f = \log\left(K \div \sum_{k=1}^{K} I(f_{bk})\right), \tag{8}$$

where K is the total number of documents in the corpus. When choosing a global weight, one needs to consider the state of the corpus. If the corpus changes, the BPM changes first and then the global weight must be revised. Thus, if the corpus is unstable or constantly changing, then using a global weight might not be a good idea.

We now come to the document normalization factor. The *cosine normalization* seems to be used often with term-document matrices [1], so this is what we use here. For our bigrams, this is given by

$$d_k = c = \left(\sum_{b=1}^{M} \{l_{bk} g_k\}^2\right)^{-1/2}. \tag{9}$$

This simply normalizes the BPMs, or one could think of this as ensuring that the magnitude of the BPM 'vector' is 1. We note that with the normalized correlation coefficient, the document normalization does not really qualify as a weight because this normalization would take place anyway with the distance measure. What it means is that the denominator in Equation 1 is one, so we do not need to calculate it for the similarity measure.

We can designate the weighting scheme by using a three letter code as follows:

txx	bigram frequency - no weights
nfc	augmented normalized frequency - IDF - cosine normalization
tfc	bigram frequency - IDF - cosine normalization
lfc	logarithmic - IDF - cosine normalization

3.2 Mutual information

In general, mutual information is a measure of the common information between two random variables [7]. *Pointwise mutual information* is defined on two particular points in the distributions. In natural language processing, pointwise mutual information is often calculated between elements and is used for clustering words and word sense disambiguation. We define a pointwise mutual information for bigrams, following the work by Pantel and Lin [11], where they discuss the pointwise mutual information between a word and a context (i.e., words around it). We use documents in place of contexts to define pointwise mutual information between a bigram and a document. The idea of using contexts as analogous to documents has been explored by Gale, Church and Yarowsky [5].

The pointwise mutual information between bigram b and document k is denoted as MI_{bk}. The idea is to substitute this value for each corresponding element in the document's BPM. Recall that the number of times bigram b occurs in document k is represented by f_{bk}. We then calculate the number of times bigram b occurs across *all* documents in the corpus, which is given by

$$f_{b\,\cdot} = \sum_{i=1}^{K} f_{bi} \ . \tag{10}$$

Next we need the total number of bigrams occurring in document k. This is given as

$$f_{\cdot\,k} = \sum_{i=1}^{M} f_{ik} \ . \tag{11}$$

The pointwise mutual information is defined as

$$MI_{bk} = \log\left(\frac{f_{bk} \div N}{f_{b\,\cdot} \div N \times f_{\cdot\,k} \div N}\right) = \log\left(\frac{N \times f_{bk}}{f_{b\,\cdot} \times f_{\cdot\,k}}\right) , \tag{12}$$

where N is the total number of bigrams and contexts, given by

$$N = \sum_{i=1}^{M} \sum_{j=1}^{K} f_{ij} \ .$$

One of the problems with pointwise mutual information is that it is biased toward infrequent words (bigrams) and contexts [11], so Pantel and Lin recommend multiplying Equation 12 with a discounting factor. For bigram b and document k, this is

$$C_{bk} = \frac{f_{bk}}{f_{bk} + 1} \times \frac{\min\{f_{b\cdot}\,;\, f_{\cdot\,k}\}}{\min\{f_{b\cdot}\,;\, f_{\cdot\,k}\} + 1} .$$

We did not use this factor in our research; only Equation 12 was implemented.

Topic Number	Topic Description
4	Cessna on the White House
5	Clinic Murders (Salvi)
6	Comet into Jupiter
8	Death of Kim Jong Il's Father
9	DNA in OJ Trial
11	Hall's Copter in N. Korea
12	Flooding Humble, TX
13	Justice-to-be Breyer
15	Kobe, Japan Quake
16	Lost in Iraq
17	NYC Subway Bombing
18	Oklahoma City Bombing
21	Serbians Down F-16
22	Serbs Violate Bihac
24	US Air 427 Crash
25	WTC Bombing Trial

Table 2: List of 16 topics.

4 Experiments

The goal of our experiments is to assess the usefulness of weighting the BPMs. In particular, to answer the question: Can documents be classified more successfully using weighted bigrams? In the next subsections, we describe some of the background and details of the experiments, followed by results. All experiments and analyses, including reading the documents and creating the BPMs, were done on a PC using MATLAB^{TM}, Version 6.5.

4.1 Description of corpus

We use the Topic Detection and Tracking (TDT) Pilot Corpus (Linguistic Data Consortium, Philadelphia, PA) to evaluate the utility of weighting the BPMs. This corpus of documents contains over 16,000 news stories from various wire services and were classified in terms of their meaning in the following way. A set of 25 topics were initially chosen and documents were tagged as either belonging to one of those topics (*yes*), partially belonging (*brief*) or not belonging (*no*). We chose a set of 503 documents encompassing 16 topics as shown in Table 2 and created a BPM for each one with weighting schemes as described in the previous section.

As for pre-processing the documents, we remove all punctuation (except for the end of sentences) and symbols such as hyphens, etc. As stated previously, all end of sentence punctuation is converted to a period, which is then treated as a word. We also investigate the effect of another pre-processing scheme - removing noise or stop words [8]. For the full text case, the size of the lexicon is 11,103. When noise words are removed, the lexicon contains 10,997 words.

4.2 Classification and dimensionality reduction

We are interested in seeing whether or not weighting the bigrams improves the results when we try to classify documents from the TDT corpus. To this end, we use a simple k nearest neighbor (k-nn) classifier [3]. This type of classifier works in the following way. We have a document with an unknown classification. We find its k nearest neighbors using the normalized correlation coefficient and look at their class labels. The document is assigned the class label that corresponds to the class that occurs with the highest frequency among the k nearest neighbors.

The k nearest neighbor classifier is easy to use and is suitable for high-dimensional data. It would be interesting to reduce the dimensionality of the space, so we can use some other method of investigation such as clustering or being able to visualize the data. In keeping with Martinez [8], we use the Isometric Feature Mapping or ISOMAP [12] procedure to reduce the dimensionality of the BPMs and repeat our classification experiments. This is particularly useful in our case, because it requires the interpoint distance matrix as its only input. Before we explain ISOMAP, we first briefly describe multidimensional scaling.

The purpose of multidimensional scaling is to represent points or observations in a lower dimensional space (usually 2-D or 3-D) in such a way that points that are close together in the higher dimensional space will also be close together in the lower dimensional space [2]. However, if the observations live along a lower dimensional nonlinear manifold, then the Euclidean distance between the points might not be the best measure of the distance between the points along the manifold. To illustrate this idea, we show a 2-D nonlinear manifold embedded in 3-D in Figure 1. The Euclidean distance between 2 random points on this manifold is shown in Figure 2, and we see that a better measure of the distance between them would be along this manifold.

ISOMAP seeks a mapping from a higher dimensional space to a lower dimensional one such that the mapping preserves the distances between observations, where the distance in the higher dimensional space is measured along the geodesic path of the nonlinear manifold. The first step in the ISOMAP algorithm is to convert the interpoint Euclidean distance matrix into geodesic distances. The geodesic distances are then used as input to classical multidimensional scaling. Besides the interpoint distance matrix, ISOMAP requires a value for the number of nearest neighbors (k) that is used in determining the geodesic distance. We use a value of $k = 10$ in this body of work.

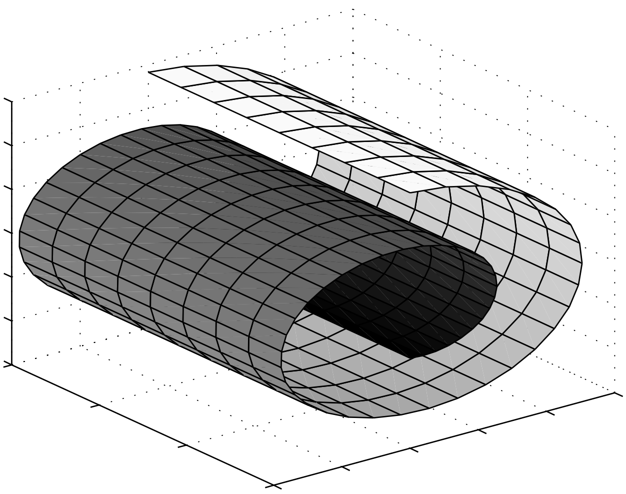


Figure 1: This illustrates a 2-D manifold (or surface) embedded in a 3-D space.

4.3 Results

To summarize, we varied the weights and other parameters and performed the following experiments with the weighted BPM.

- Text pre-processing conditions were full and denoised lexicon.
- Bigram weights were MI_{bk}, *lfc*, *nfc*, *tfc*, and *txx*.
- Dimensionality of the space for using k-nn was either full dimensionality or 4-D and 6-D from ISOMAP.
- The values of k for the k-nn classifier were $k = 1, 3, 5, 7, 10$.
- The Euclidean distance was used for the 4-D and 6-D k-nn classification.

The results from the experiments are shown in Tables 3 through 5.

Several things can be noted from these results. First, we see that in the full BPM case, using the pointwise mutual information increases the proportion of documents correctly classified. Secondly, denoising the data seems to produce poorer or similar results in the weighted case, but better results in the unweighted case (*txx*). Finally, it is interesting to note that the weighting scheme *tfc* allows us to compare the use of the IDF global weight alone. By comparing the *tfc** and *txx** entries, we see that using the IDF global weight increases the correct classification.

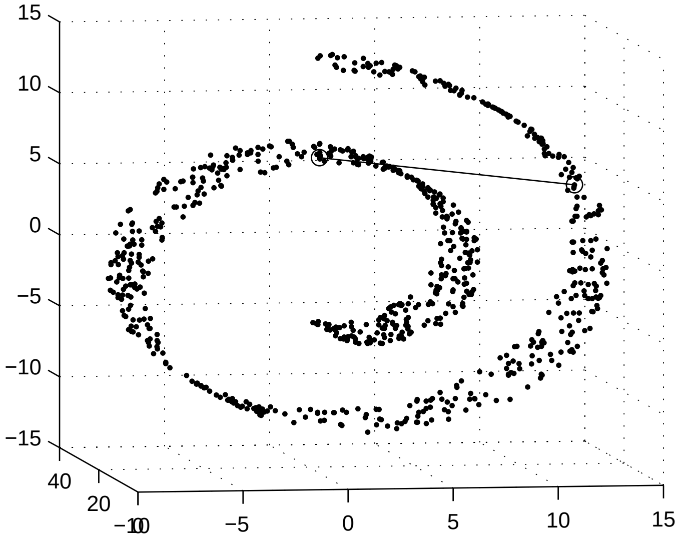


Figure 2: This is a data set randomly generated according to the manifold given in Figure 1. The Euclidean distance between two points is given by the straight line shown here. If we are seeking the neighborhood structure along the manifold, then it would be better to use the geodesic distance (the distance along the manifold or the roll) between the points.

	k = 1	**k = 3**	**k = 5**	**k = 7**	**k = 10**
lfc	0.90	0.92	0.93	0.93	0.94
lfc-den	0.87	0.87	0.86	0.87	0.87
MI	0.98	0.99	1.00	1.00	0.99
MI-den	0.98	0.98	0.99	0.99	0.99
nfc	0.99	0.99	0.99	1.00	1.00
nfc-den	0.98	0.99	0.99	0.99	0.99
tfc	0.98	0.98	0.99	0.99	0.99
tfc-den	0.99	0.98	0.98	0.99	0.98
txx	0.90	0.90	0.91	0.92	0.93
txx-den	0.93	0.93	0.93	0.93	0.92

Table 3: Proportion of documents correctly classified - full BPMs.

5 Summary

In this paper, we defined bigram weights for the BPMs that are similar to term weights used in natural language processing and information retrieval. After the BPMs are weighted, we applied the k-nn classification method to

	k = 1	**k = 3**	**k = 5**	**k = 7**	**k = 10**
lfc	0.74	0.74	0.75	0.77	0.76
lfc-den	0.71	0.71	0.73	0.73	0.72
MI	0.82	0.81	0.83	0.83	0.84
MI-den	0.81	0.83	0.85	0.87	0.86
nfc	0.84	0.84	0.85	0.86	0.85
nfc-den	0.85	0.85	0.87	0.87	0.87
tfc	0.88	0.87	0.87	0.86	0.87
tfc-den	0.86	0.86	0.87	0.86	0.86
txx	0.66	0.65	0.65	0.64	0.65
txx-den	0.73	0.72	0.74	0.73	0.75

Table 4: Proportion of documents correctly classified - BPMs reduced to 4-D.

	k = 1	**k = 3**	**k = 5**	**k = 7**	**k = 10**
lfc	0.83	0.84	0.85	0.85	0.84
lfc-den	0.78	0.79	0.81	0.80	0.80
MI	0.91	0.93	0.93	0.95	0.95
MI-den	0.91	0.93	0.93	0.93	0.93
nfc	0.92	0.92	0.92	0.93	0.93
nfc-den	0.92	0.93	0.94	0.94	0.94
tfc	0.92	0.93	0.93	0.93	0.92
tfc-den	0.92	0.91	0.94	0.92	0.90
txx	0.67	0.67	0.68	0.69	0.67
txx-den	0.83	0.81	0.83	0.82	0.82

Table 5: Proportion of documents correctly classified - BPMs reduced to 6-D.

determine whether or not weighting the BPMs improve document recognition. Results show that in some cases, where local weights were used, such as the normalized augmented frequency, did improve the classification performance. Additionally, using the pointwise mutual information, taking the context into account, significantly improved the results.

A lot of work in this area of weighting the BPMs remains to be done. One interesting possibility is to change the pointwise mutual information to include the topic. In other words, instead of using the document as the context, we might use the topic or class as the context. Of course in this case, we would have to use a training set of documents that are tagged with their topic to estimate the context. This can then be used with new untagged documents and their BPMs. Additionally, we could use the discounting factor with the mutual information. Other bigram weights can be defined and examined, such as entropy, probabilistic inverse and pivoted-cosine normalization [1]. We might also examine other real-valued measures of distance or similarity other than the NCC.

We looked at pre-processing the text by removing noise words. We could also perform some experiments using a stemmed and denoised lexicon [8], [1]. We could also examine the affect of the dimensionality reduction procedure. As stated previous, ISOMAP seeks a nonlinear manifold; we might try something like classical multidimensional scaling [2] (using the NCC similarity directly rather than the geodesic distance). Finally, we could use some other methods to analyze the reduced BPMs, such as model-based clustering [4], linear or quadratic classifiers [3], non-metric multidimensional scaling, self-organizing maps [6] etc.

References

[1] Berry M.W., Browne M. (1999). *Understanding search engines: mathematical modeling and text retrieval.* SIAM.

[2] Cox T.F., Cox M.A.A. (2001). *Multidimensional scaling, 2nd edition.* Chapman and Hall - CRC.

[3] Duda R.O., Hart P.E., Stork D.G. (2000). *Pattern classification, 2nd edition.* Wiley-Interscience.

[4] Fraley C., Raftery A.E. (1998). *How many clusters? Which clustering method? Answers via model-based cluster analysis.* The Computer Journal **41**, 578 – 588.

[5] Gale, Church and Yarowsky. (1992). *A method for disambiguating word senses in a corpus.* Computers and the Humanities **26**, 415 – 439.

[6] Kohonen, Tuevo. (2001). *Self-organizing maps, third edition.* Springer-Verlag.

[7] Manning C.D., Schütze H. 2000. *Foundations of statistical natural language processing.* The MIT Press.

[8] Martinez A.R. (2002). *A framework for the representation of semantics.* Ph.D. Dissertation, George Mason University.

[9] Martinez A.R., Wegman E.J. (2002). *A text stream transformation for semantic-based clustering.* Proceedings of the Interface.

[10] Martinez A.R., Wegman E.J. (2002). *Encoding of text to preserve meaning.* Proceedings of the Army Conference on Applied Statistics.

[11] Pantel P., Lin D. (2002). *Discovering word senses from text.* Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 613 – 619.

[12] Tenenbaum J.B., de Silva V., Langford J.C. (2000). *A global geometric framework for nonlinear dimensionality reduction.* Science **290**, 2319 – 2323.

Address: A.R. Martinez, W.L. Martinez, NAVSEA Dahlgren, USA
E.J. Wegman, School of Information Technology and Engineering, George Mason University Fairfax, Virginia

E-mail: `marinwe@onr.navy.mil`

COMPSTAT'2004 Symposium

ON CANONICAL ANALYSIS OF VECTOR TIME SERIES

Wanli Min and Ruey S. Tsay

Key words: Hankel matrix, Kronecker index, canonical correlation.

COMPSTAT 2004 section: Time series analysis.

Abstract: In this paper, we establish some asymptotic results for canonical analysis of vector linear time series when the data possess conditional heteroscedasticity. We show that for correct identification of a vector time series model, it is essential to use a modification, which we prescribe, to a commonly used test statistic for testing zero canonical correlations. A real example and simulation are used to demonstrate the importance of the proposed test statistics.

1 Introduction

Since proposed in [13], canonical correlation analysis has been widely applied in many statistical areas, especially in multivariate analysis. Time series analysis is no exception. [6] proposed a canonical analysis of vector time series that can reveal the underlying structure of the data to aid model interpretation. In particular, they showed that linear combinations of several unit-root non-stationary time series can become stationary. This is the idea of co-integration that was popular among econometricians in the 1990s after the publication of [10]. [22] applied canonical correlation analysis to develop the smallest canonical correlation method for identifying univariate ARMA model for a stationary and/or non-stationary time series. [17] introduced the concept of scalar component models to build a parsimonious VARMA model for a given vector time series. Again, canonical correlation analysis was used extensively to search for scalar component models. Many other authors also used canonical analysis in time series analysis. See, for instance, [15].

To build a model for a k-dimensional linear process, it suffices to identify the k Kronecker indexes or k linearly independent scalar component models, because we can use such information to identify those parameters that require estimation and those that can be set to zero within a dynamic linear vector model. Simply put, the Kronecker indexes and scalar component models can overcome the difficulties of curse of dimensionality, parameter explosion, exchangeable models, and redundant parameters in modelling a linear vector time series. For simplicity, we shall consider the problem of specifying Kronecker indexes in this paper. The issue discussed, however, is equally applicable to specification of scalar component models. The method of determining Kronecker indexes of a linear vector process with Gaussian innovations has been studied by [1], [7], [18], [20], among others. These studies show that canonical correlation analysis is useful in specifying the Kronecker indexes under normality. On the other hand, the assumption of Gaussian innovations is questionable in many applications, especially in analysis of economic and financial data that often exhibit conditional heteroscedasticity. See, for instance, the summary statistics of asset returns in Chapter 1 of [21]. In the literature, a simple approach to model conditional heteroscedasticity is to apply the generalized autoregressive conditional heteroscedastic (GARCH) model of [9] and [3]. We shall adopt such a model for the innovation series of multivariate time series data.

In this paper, we continue to employ canonical analysis in vector time series. However, we focus on statistical inference concerning canonical correlation coefficients when the distribution of the innovations is not Gaussian. Our main objective is to identify a vector model with structural specification for a given time series that exhibits conditional heteroscedasticity and has high kurtosis. Specifically, we study canonical correlation analysis when the innovations of the series follow a vector GARCH model.

1.1 Preliminaries

Based on the Wold decomposition, a k-dimensional stationary time series $\mathbf{Z}_t = (z_{1t}, \cdots, z_{kt})'$ can be written as $\mathbf{Z}_t = \mu + \sum_{i=0}^{\infty} \psi_i \mathbf{a}_{t-i}$, where $\mu = (\mu_1, \cdots, \mu_k)'$ is a constant vector, ψ_i are $k \times k$ coefficient matrices with $\psi_0 = \mathbf{I}_k$ being the identity matrix, and $\{\mathbf{a}_t = (a_{1t}, \cdots, a_{kt})'\}$ is a sequence of k-dimensional uncorrelated random vectors with mean zero and positive-definite covariance matrix $\mathbf{\Sigma}$. That is, $E(\mathbf{a}_t) = \mathbf{0}$, $E(\mathbf{a}_t \mathbf{a}'_{t-i}) = \mathbf{0}$ if $i \neq 0$, and $E(\mathbf{a}_t \mathbf{a}'_t) = \mathbf{\Sigma}$. The $\mathbf{a}_t$ process is referred to as the innovation series of $\mathbf{Z}_t$. If $\sum_{i=0}^{\infty} \|\psi_i\| < \infty$, then $\mathbf{Z}_t$ is (asymptotically) weakly stationary, where $\|\mathbf{A}\|$ is a matrix norm, e.g. $\|\mathbf{A}\| = \sqrt{\text{trace}(\mathbf{AA}')}$. Often one further assumes that $\mathbf{a}_t$ is Gaussian. In this paper, we assume that

$$\sup_{i,t} E(|a_{it}|^{\eta} | F_{t-1}) < \infty \quad \text{almost surely for some } \eta > 2, \tag{1}$$

where $F_{t-1} = \sigma\{\mathbf{a}_{t-1}, \mathbf{a}_{t-2}, \cdots\}$ denotes information available at time $t-1$. Writing $\psi(B) = \sum_{i=0}^{\infty} \psi_i B^i$, where B is the backshift operator such that $B\mathbf{Z}_t = \mathbf{Z}_{t-1}$, then $\mathbf{Z}_t = \mu + \psi(B)\mathbf{a}_t$. If $\psi(B)$ is rational, then $\mathbf{Z}_t$ has a VARMA representation

$$\mathbf{\Phi}(B)(\mathbf{Z}_t - \mu) = \mathbf{\Theta}(B)\mathbf{a}_t \tag{2}$$

where $\mathbf{\Phi}(B) = \mathbf{I} - \sum_{i=1}^{p} \mathbf{\Phi}_i B^i$ and $\mathbf{\Theta}(B) = \mathbf{I} - \sum_{j=1}^{q} \mathbf{\Theta}_j B^j$ are two matrix polynomials of order p and q, respectively, and have no common left factors. For further conditions of identifiability, see [8] for more details. The stationarity condition of $\mathbf{Z}_t$ is equivalent to that all zeros of the polynomial $|\mathbf{\Phi}(B)|$ are outside the unit circle.

The number of parameters of the VARMA model in Eq. (2) could reach $(p+q)k^2 + k + k(k+1)/2$ if no constraint is applied, making parameter estimation unnecessarily difficult in some applications. Several methods are available in the literature that can simplify the use of VARMA models when the innovations $\{\mathbf{a}_t\}$ are Gaussian. For instance, specification of Kronecker indexes of a Gaussian vector time series can lead to a parsimonious parametrization of VARMA representation, see [19]. In many situations, the innovational process $\mathbf{a}_t$ has conditional heteroscedasticity. In the univariate case, [3] proposed a GARCH(r_1, r_2) model to handle conditional heteroscedasticity. The model can be written as

$$a_t = \sqrt{g_t}\epsilon_t, \qquad g_t \quad = \quad \alpha_0 + \sum_{i=1}^{r_1} \alpha_i a_{t-i}^2 + \sum_{j=1}^{r_2} \beta_j g_{t-j}, \tag{3}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\{\epsilon_t\}$ is a sequence of independent and identically distributed random variables with mean zero and variance 1. It's well-known that a_t is asymptotically second order stationary if $\sum_{i=1}^{r_1} \alpha_i + \sum_{j=1}^{r_2} \beta_j < 1$. Generalization of the GARCH models to multivariate case introduces additional complexity to the modelling procedure because the covariance matrix of $\mathbf{a}_t$ has $k(k+1)/2$ elements. Writing the conditional covariance matrix of $\mathbf{a}_t$ given the past information as $\mathbf{\Sigma}_t = E(\mathbf{a}_t\mathbf{a}_t'|F_{t-1})$, where F_{t-1} is defined in Eq. (1), we have $\mathbf{a}_t = \mathbf{\Sigma}_t^{1/2}\epsilon_t$, where $\mathbf{\Sigma}_t^{1/2}$ is the symmetric square-root of the matrix $\mathbf{\Sigma}_t$ and $\{\epsilon_t\}$ is a sequence of independent and identically distributed random vectors with mean zero and identity covariance matrix. Often ϵ_t is assumed to follow a multivariate normal or Student-t distribution. To ensure the positive definiteness of $\mathbf{\Sigma}_t$, several models have been proposed in the literature. For example, consider the simple case of order (1,1). [11] consider the BEKK model $\mathbf{\Sigma}_t = \mathbf{CC}' + \mathbf{A}\mathbf{a}_{t-1}\mathbf{a}_{t-1}'\mathbf{A}' + \mathbf{B}\mathbf{\Sigma}_{t-1}\mathbf{B}'$, where $\mathbf{C}$ is a lower triangular matrix and $\mathbf{A}$ and $\mathbf{B}$ are $k \times k$ matrices. [4] discusses the diagonal model $\mathbf{\Sigma}_t = \mathbf{CC}' + \mathbf{AA}' \otimes (\mathbf{a}_{t-1}\mathbf{a}_{t-1}') + \mathbf{BB}' \otimes \mathbf{\Sigma}_{t-1}$, where $\otimes$ stands for matrix Hadamard product (element-wise product).

When GARCH effects exist, the time series $\mathbf{Z}_t$ is no longer Gaussian. Its innovations become a sequence of uncorrelated, but serially dependent random vectors. It is well-known that such innovations tend to have heavy tails, see [9] and [21], among others. The performance of canonical correlation analysis under such innovations is yet to be investigated. This is the main objective of this paper. Sections 2 & 3 review and introduce the problem considered in the paper. Section 4 establishes the statistics to specify Kronecker indexes for VARMA+GARCH process. Section 5 presents some simulation results, and Section 6 applies the analysis to a real financial time series.

2 Kronecker index and vector ARMA representation

2.1 Vector ARMA model implied by Kronecker index

For simplicity, we assume that $\mu = \mathbf{0}$. Given a time point t, define the past and future vectors $\mathbf{P}_t$ and $\mathbf{F}_t$ of the process $\mathbf{Z}_t$ as $\mathbf{P}_t = (\mathbf{Z}_{t-1}', \mathbf{Z}_{t-2}', \cdots)'$,

$\mathbf{F}_t = (\mathbf{Z}'_t, \mathbf{Z}'_{t+1}, \cdots)'$. The Hankel Matrix of $\mathbf{Z}_t$ is defined as $\mathbf{H} = E(\mathbf{F}_t\mathbf{P}'_t)$. It is obvious that for a VARMA model in Eq. (2) the Hankel matrix $\mathbf{H}$ is of finite rank. In fact, it can be shown that Rank($\mathbf{H}$) is finite if and only if $\mathbf{Z}_t$ has a VARMA model representation, see [12] and [20].

The Kronecker indexes of $\mathbf{Z}_t$ consist of a set of non-negative integers $\{K_i| \quad i = 1, \cdots, k\}$ such that for each i, K_i is the smallest non-negative integer that the $(k \times K_i + i)$th row of $\mathbf{H}$ is either a null vector or is a linear combination of the previous rows of $\mathbf{H}$. It turns out that $\sum_{i=1}^k K_i$ is the rank of $\mathbf{H}$, which is invariant under different VARMA presentations of $\mathbf{Z}_t$. In fact, the set of Kronecker indexes, $\{K_i\}_{i=1}^k$, of a given VARMA process is invariant under various forms of model representation. [20] illustrates how to construct an Echelon VARMA form for $\mathbf{Z}_t$ using the Kronecker indexes $\{K_i\}_{i=1}^k$. For a stationary process $\mathbf{Z}_t$ with specified Kronecker index $\{K_1, \cdots, K_k\}$, let $p = \max\{K_i | i = 1, \cdots, k\}$. Then $\mathbf{Z}_t$ follows a VARMA(p, p) model

$$\mathbf{\Phi}_0\mathbf{Z}_t - \sum_{i=1}^{p} \mathbf{\Phi}_i\mathbf{Z}_{t-i} = \delta + \mathbf{\Phi}_0\mathbf{e}_t - \sum_{j=1}^{p} \mathbf{\Theta}_j\mathbf{e}_{t-j}, \tag{4}$$

where δ is a constant vector, the ith row of $\mathbf{\Phi}_j$ and $\mathbf{\Theta}_j$ are zero for $j > K_i$, and $\mathbf{\Phi}_0$ is a lower triangular matrix with ones on the diagonal. Furthermore, some elements of $\mathbf{\Phi}_i$ can be set to zero based on the Kronecker indexes. A VARMA model in Eq. (4) provides a unique ARMA representation for $\mathbf{Z}_t$, see Theorem 2.5.1 in [12].

2.2 Specification of Kronecker index

If the smallest canonical correlation between the future and past vectors $\mathbf{F}_t$ and $\mathbf{P}_t$ is zero, then $X_t = \mathbf{V}'_f\mathbf{F}_t$ is uncorrelated with $\mathbf{P}_t$, i.e. Cov$(X_t, \mathbf{P}_t)$ $= \mathbf{V}'_f E(\mathbf{F}_t\mathbf{P}'_t) = \mathbf{V}'_f\mathbf{H} = \mathbf{0}$. This leads to a row dependency of the Hankel matrix so that the analysis is directly related to Kronecker index. Testing for zero canonical correlation thus plays an important role in specifying Kronecker indexes. [7] used the traditional χ^2 test to propose a modelling procedure:

Step 1: Select a large lag s so that the vector $\mathbf{P}_t = (\mathbf{Z}'_{t-1}, \cdots, \mathbf{Z}'_{t-s})'$ is a good approximation of the past vector and choose initial future sub-vector $\mathbf{F}^*_t = \{Z_{1t}\}$. If a vector AR approximation is used, then s can be selected by information criteria such as AIC or BIC.

Step 2: Let $\hat{\rho}$ be the smallest sample canonical correlation in modulus between $\mathbf{F}^*_t$ and $\mathbf{P}_t$. Denote the canonical variates by $X_t = \mathbf{V}'_f\mathbf{F}^*_t$ and $Y_t = \mathbf{V}'_p\mathbf{P}_t$, and compute the test statistics

$$S = -n\log(1 - \hat{\rho}^2) \sim \chi^2_{ks-f+1}, \tag{5}$$

where n is the number of observations, f and ks are the dimension of $\mathbf{F}^*_t$ and $\mathbf{P}_t$, respectively.

Step 3: Denote the last element of $\mathbf{F}_t^*$ as $Z_{i,t+h}$. If $H_o : \hat{\rho} = 0$ is not rejected, then the Kronecker index for the ith component Z_{it} of $\mathbf{Z}_t$ is $K_i = h$. In this case, update the future vector $\mathbf{F}_t$ by removing $Z_{i,t+j}$ for $j \geq h$. If all k Kronecker indexes have been found, the procedure is terminated. Otherwise, augment $\mathbf{F}_t^*$ by adding the next available element of the updated $\mathbf{F}_t$ and return to Step 2.

The asymptotic χ^2 distribution of the S-statistic in Eq. (5) of Step 2 is derived under the independence sampling assumption. [18] showed that the canonical correlations cannot be treated as the cross correlation of two white-noise series since the corresponding canonical variates are serially correlated. Suppose $\mathbf{F}_t^* = (Z_{1,t}, \cdots, Z_{i,t+h})'$. The smallest sample canonical correlation $\hat{\rho}$ is the lag-$(h+1)$ sample cross-correlation $\hat{\rho}_{xy}(h+1)$ of the corresponding canonical variates $X_t = \mathbf{V}_f'\mathbf{F}_t^*$ and $Y_t = \mathbf{V}_p'\mathbf{P}_t$ because Y_t is observable at time $t-1$ whereas X_t is observable at time $t+h$. Under $H_0 : \rho_{xy}(m) = 0$, the asymptotic variance of $\hat{\rho}_{xy}(m)$ is, shown in [5],

$$\text{var}[\hat{\rho}_{xy}(m)] \approx n^{-1} \sum_{\nu=-\infty}^{\infty} \{\rho_{xx}(\nu)\rho_{yy}(\nu) + \rho_{xy}(m+\nu)\rho_{yx}(m-\nu)\}. \tag{6}$$

Making use of the result mentioned above, [18] proposed a proper test statistic

$$T = -(n-s)\log(1 - \frac{\hat{\rho}^2}{\hat{d}}) \sim \chi^2_{ks-f+1} \tag{7}$$

where $\hat{d} = 1 + 2\sum_{\nu=1}^{h} \hat{\rho}_{xx}(\nu)\hat{\rho}_{yy}(\nu)$. In Eq. (7), it is understood that $\hat{d} = 1$ if $h = 0$, $\hat{\rho}_{xx}(\nu)$ and $\hat{\rho}_{yy}(\nu)$ are the lag-ν sample autocorrelations of X_t and Y_t, respectively, and n is the sample size. The Bartlett's formula in Eq. (6) is for independent Gaussian innovations $\{a_t\}$. This is not the case when the innovations follow a GARCH(r_1, r_2) model. We shall study in next section properties of sample auto-covariances in the presence of GARCH innovations. All proofs can be found in [14].

3 Sample auto-covariance functions of a linear process

Lemma 3.1. *Suppose $\{a_t\}$ is a stationary GARCH(r_1, r_2) process of Eq. (3) with finite fourth moment and ϵ_t is symmetrically distributed, then $E(a_i a_k a_j a_l) = 0$, $\forall i \leq j \leq k \leq l$ unless $i = j$ and $k = l$ both hold.*

Proposition 3.1. *Suppose $\{a_t\}$ is a GARCH(r_1, r_2) process with $E(a_t^2) = \sigma^2$ and $E(a_t^4) < \infty$ and the process X_t is defined as $X_t = \sum_{i=0}^{\infty} \psi_i a_{t-i}$ with $\sum_i |\psi_i| < \infty$, and $\sum_i i\psi_i^2 < \infty$. Let $\gamma_{xx}(0) = \sigma^2 \sum_{i=0}^{\infty} \psi_i^2$. Then the next inequality holds: $\sum_{t=1}^{\infty} \|E(X_t^2 - \gamma_{xx}(0)|\xi_0)\| < \infty$, where $\xi_0 = \sigma\{\epsilon_0, \epsilon_{-1}, \cdots\}$ and $\|Y\|$ denotes the L^2-norm of a random variable Y.*

Defining the norm of a random matrix as $\|A\| := \sqrt{E(trAA')}$, we can generalize *Prop* 3.1 to a linear process with innovational process that follows a multivariate GARCH model.

Proposition 3.2. *Assume* $\mathbf{a}_t = (a_{1t}, \cdots, a_{mt})'$ *follows a pure diagonal multivariate GARCH model, i.e.* a_{it} *follows a univariate GARCH*(r_1, r_2) *model and is stationary with finite fourth moment for each* $i = 1, \cdots, m$. *Consider the process* $X_t = \sum_{i=0}^{\infty} \mathbf{\Psi}_i' \mathbf{a}_{t-i}$ *where* $\mathbf{\Psi}_i$ *are* m*-dimensional vectors. Assume further that* $\sum_{i=0}^{\infty} \|\mathbf{\Psi}_i\| < \infty$ *and* $\sum_{i=0}^{\infty} i\|\mathbf{\Psi}_i\|^2 < \infty$. *Let* $\xi_0 = \sigma\{\mathbf{a}_0, \mathbf{a}_{-1}, \cdots\}$. *Then the next inequality holds:* $\sum_{t=1}^{\infty} \|E(X_t^2 - \gamma_{xx}(0)|\xi_0)\| < \infty$, *where* $\gamma_{xx}(0) = \sum_{i=0}^{\infty} \mathbf{\Psi}_i' \mathbf{\Sigma} \mathbf{\Psi}_i$ *and* $\mathbf{\Sigma} = E(\mathbf{a}_t \mathbf{a}_t') = diag(\sigma_1^2, \cdots, \sigma_m^2)$.

Observing $X_t Y_{t+h} = \frac{(X_t + Y_{t+h})^2 - (X_t - Y_{t+h})^2}{4}$, we have by the triangle inequality the next corollary.

Corollary 3.1. *Suppose* $X_t = \sum_{i=0}^{\infty} \mathbf{\Psi}_i' \mathbf{a}_{t-i}$ *and* $Y_t = \sum_{i=0}^{\infty} \mathbf{\Phi}_i' \mathbf{a}_{t-i}$ *both satisfy the conditions in Prop 3.2. Let* $\gamma_{xy}(h) = E(X_t Y_{t+h})$, *where* h *is an integer. We have* $\sum_{t=1}^{\infty} \|E(X_t Y_{t+h} - \gamma_{xy}(h)|\xi_0)\| < \infty$.

To generalize the result to the case that $\mathbf{X}_t$ is multivariate, we define $\text{Vec}(\mathbf{A}) = (\mathbf{A}_1', \cdots, \mathbf{A}_n')'$ for a matrix $\mathbf{A}$. We also use a Lemma in [23].

Proposition 3.3. *Let* $\mathbf{X}_t = (X_{1t}, \cdots, X_{kt})' = \sum_{i=0}^{\infty} \mathbf{\Psi}_i \mathbf{a}_{t-i}$, *where* $\mathbf{\Psi}_i$ *are matrices of dimension* $k \times m$ *and* $\mathbf{a}_t$ *is* m*-dimensional and follows a pure diagonal stationary GARCH*(r_1, r_2) *model with finite 4th moment. Further,* $\sum_{i=0}^{\infty} \|\mathbf{\Psi}_i\| < \infty, \sum_{i=0}^{\infty} i\|\mathbf{\Psi}_i\|^2 < \infty$. *Letting* $\mathbf{\Sigma} = E(\mathbf{X}_t \mathbf{X}_{t+h}')$ *where* h *is an integer, we have* $\sum_{t=1}^{\infty} \| Vec(E(\mathbf{X}_t \mathbf{X}_{t+h}'|\xi_0) - \mathbf{\Sigma})\| < \infty$

Proposition 3.4. *Let* $\mathbf{X}_t = (X_{1t}, \cdots, X_{kt})' = \sum_{i=0}^{\infty} \mathbf{\Psi}_i \mathbf{a}_{t-i}$, $\mathbf{Y}_t = (Y_{1t}, \cdots, Y_{lt})' = \sum_{i=0}^{\infty} \mathbf{\Phi}_i \mathbf{a}_{t-i}$, *where* $\mathbf{\Psi}_i$ *and* $\mathbf{\Phi}_i$ *are matrices of dimension* $k \times m$ *and* $l \times m$, *respectively. Suppose both* $\mathbf{X}_t$ *and* $\mathbf{Y}_t$ *satisfy the conditions in Proposition 3.3. Denote* $\mathbf{\Sigma}_{xy}(h) = E(\mathbf{X}_t \mathbf{Y}_{t+h}')$. *Then* $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Vec(\mathbf{X}_t \mathbf{Y}_{t+h}' - \mathbf{\Sigma}_{xy}(h)) \longrightarrow N(\mathbf{0}, \mathbf{\Sigma})$, *where* h *is any integer and* $\mathbf{\Sigma} \in R^{kl \times kl}$.

Remark 3.1. *For a causal, stationary VARMA*(p, q) *process* $\mathbf{\Phi}(B)(\mathbf{Z}_t - \mu) = \mathbf{\Theta}(B)\mathbf{a}_t$, *its* $MA(\infty)$ *representation* $\mathbf{Z}_t = \mu + \sum_{i=0}^{\infty} \mathbf{\Psi}_i \mathbf{a}_{t-i}$ *satisfies the condition* $\sum_{i=0}^{\infty} \|\mathbf{\Psi}_i\| < \infty, \sum_{i=0}^{\infty} i\|\mathbf{\Psi}_i\|^2 < \infty$ *since* $\|\mathbf{\Psi}_i\| \sim r^i$ *with* $r \in (0, 1)$

being the largest root (in magnitude) of $\mathbf{\Phi}(B^{-1})$*. Consequently, if* $\mathbf{a}_t$ *follows a pure diagonal GARCH model with finite fourth moment, the sample auto-covariance matrix of* $\mathbf{Z}_t$ *has an asymptotic joint normal distribution.*

Theorem 3.1. *Suppose that* $\mathbf{Z}_t$ *is a* k*-dimensional stationary VARMA process of model (2), where the innovation series* $\mathbf{a}_t$ *follows a GARCH*(r_1, r_2) *model with finite 4th moment. Let* $\mathbf{P}_t = (\mathbf{Z}'_{t-1}, \cdots, \mathbf{Z}'_{t-s})'$ *be a past vector with a prespecified* $s > 0$ *that contains all the information needed in predicting the future observation of* $\mathbf{Z}_t$*,* $\mathbf{F}_t = (z_{1,t}, \cdots z_{i,t+h})'$ *be the future subvector of* $\mathbf{Z}_t$ *constructed according to the procedure described in Section 2. Let* $\hat{\rho}$ *be the smallest sample canonical correlation between* $\mathbf{P}_t$ *and* $\mathbf{F}_t$*. Under the null hypothesis that the smallest canonical correlation* ρ *between* $\mathbf{P}_t$ *and* $\mathbf{F}_t$ *is zero but all the other canonical correlations are nonzero, then* $\hat{\rho}^2/var(\hat{\rho})$ *has an asymptotic* χ^2 *distribution with* $ks - f + 1$ *degrees of freedom, where* f *is the dimension of* $\mathbf{F}_t$*.*

4 Asymptotic variance of sample cross correlation

Next we consider the variance of sample cross-correlation coefficient for the case that gives rise to a zero canonical correlation between the past and future vectors of $\mathbf{Z}_t$. To this end, we make use of the Aitken's delta method. Suppose Y_t and X_t are stationary moving-average processes. More specifically, $Y_t = \sum_{i=0}^{h} \phi_i a_{t-i}$ and $X_t = \sum_{i=0}^{\infty} \psi_i a_{t-i}$ with a_t being a GARCH(r_1, r_2) process of Eq. (3). By Lemma 1, $E(a_i a_j a_k a_l) = 0\ \forall i \le j \le k \le l$ unless $i = j$ and $k = l$ both hold. Let $U = \hat{\gamma}_{xx}(0)$, $V = \hat{\gamma}_{yy}(0)$, and $W = \hat{\gamma}_{xy}(q) = \frac{1}{n-q} \sum_{t=1}^{n-q} X_t Y_{t+q}$. Given $q > h$, where h corresponds to a Kronecker index, we have $\gamma_{xy}(q) = \gamma_{yy}(q) = 0$, and on applying the delta method the following result holds:

$$\mathrm{Var}(\hat{\rho}_{xy}(q)) \approx \frac{1}{n} \sum_{|d| \le h} \left[\rho_{xx}(d)\rho_{yy}(d) + \frac{\mathrm{Cum}(X_0, X_d, Y_q, Y_{q+d})}{\gamma_{xx}(0)\gamma_{yy}(0)} \right], \qquad (8)$$

where $\mathrm{Cum}(X_0, X_d, Y_q, Y_{q+d}) = \sum_{i \ge 0} \sum_{k=0}^{h-d} \psi_i \psi_{i+d} \phi_k \phi_{k+d} \mathrm{Cov}(a_0^2, a_{q-k+i}^2)$.

Therefore, the fourth order cumulants of $\{X_t\}$ depend on the auto-covariance function of $\{a_t^2\}$. Compared to $\gamma_{xx}(d)\gamma_{yy}(d)$, $\mathrm{Cum}(X_0, X_d, Y_q, Y_{q+d})$ has a non-negligible impact on $\mathrm{Var}(\hat{\rho}_{xy}(p))$ if $\mathrm{Cov}(a_0^2, a_{q-k+i}^2)/E^2(a_0^2)$ is large. For instance, if a_t is a GARCH(1,1) process, then $\mathrm{Cov}(a_0^2, a_1^2)/\sigma^4 = 2\alpha_1 + \frac{6\alpha_1^2(\alpha_1+\beta_1/3)}{1-2\alpha_1^2-(\alpha_1+\beta_1)^2}$. This ratio is 86 given $\alpha_1 = 0.5$ and $\beta_1 = 0.2$. Considering the 4th order cumulant correction term in $\mathrm{Var}(\hat{\rho})$, one can modify the T statistic proposed by Tsay as

$$
\begin{aligned}
T^* &= -(n-s)\log(1-\frac{\hat{\rho}^2}{\hat{d^*}}) \sim \chi^2_{ks-f+1}, \qquad (9)\\
\hat{d^*} &= \sum_{|d|\le h}\left[\rho_{xx}(d)\rho_{yy}(d) + \frac{\mathrm{Cum}(X_0, X_d, Y_q, Y_{q+d})}{\gamma_{xx}(0)\gamma_{yy}(0)}\right].
\end{aligned}
$$

5 Simulations study

We conduct some simulations to study the finite sample performance of the modified test statistics. We focus on a bivariate ARMA+GARCH(1,1) model chosen to have GARCH parameters similar to those commonly seen in empirical asset returns. The model is

$$
\mathbf{Z}_t - \begin{bmatrix} 0.8 & 0 \\ 0 & 0.3 \end{bmatrix} \mathbf{Z}_{t-1} = \mathbf{a}_t - \begin{bmatrix} -0.8 & 1.3 \\ -0.3 & 0.8 \end{bmatrix} \mathbf{a}_{t-1}, \qquad (10)
$$

where $t = 1, \cdots, n$ and $\mathbf{a}_t = \mathrm{diag}(\sqrt{g_{1t}}, \sqrt{g_{2t}})\epsilon_t$ with $\epsilon_t \sim i.i.d \quad N_2(0, I)$, where g_{it} satisfy the GARCH(1,1) model $g_{it} = 0.5 + 0.2a^2_{i,t-1} + 0.7g_{i,t-1}$ for $i = 1$ and 2. For a given sample size n, each realization was obtained by generating $5n$ observations. To reduce the effect of the starting values $\mathbf{Z}_0$ and $\mathbf{a}_0$, we only use the last n observations. For this model, the two future subvectors which in theory give a zero canonical correlation are $\mathbf{F}_t(1) = (z_{1t}, z_{2t}, z_{1,t+1})'$ and $\mathbf{F}_t(2) = (z_{1t}, z_{2t}, z_{2,t+1})'$. A value of $s = 5$ was selected according to AIC criterion in a preliminary analysis using pure vector AR models. The corresponding past vector is $\mathbf{P}_t = (\mathbf{Z}'_{t-1}, \cdots, \mathbf{Z}'_{t-5})'$.

Let $S(1)$ and $S(2)$ be the test statistics $S = -n\log(1-\hat{\rho}^2)$ of [7] when the future subvectors are $\mathbf{F}_t(1)$ and $\mathbf{F}_t(2)$, respectively. Similarly, let $T(1)$ and $T(2)$ be the corresponding test statistics $T = -(n-s)\log(1-\frac{\hat{\rho}^2}{\hat{d}})$ of [18] and $T^*(1)$ and $T^*(2)$ be the test statistics $T^* = -(n-s)\log(1-\frac{\hat{\rho}^2}{\hat{d^*}})$ proposed in Eq. (9). In particular, we adopt the approach of [2] to estimate the variance of sample cross-covariance $\mathrm{Var}[\hat{\gamma}_{xy}(q)]$ by

$$
n\ \mathrm{Var}[\hat{\gamma}_{xy}(q)] \approx \hat{\sigma}^*(0) + 2\sum_{i=1}^{n-q}(1-i/n)K(ib_n)\hat{\sigma}^*(i),
$$

where $\hat{\sigma}^*(i) = \sum_t X_tY_{t+q}X_{t+i}Y_{t+i+q}/n - \hat{\gamma}^2_{xy}(q)$, $K(x) = I_{|x|\le 1}$, and $b_n = n^{-1/4}$. However, to improve the robustness of the variance estimate in finite samples, we employ a modified estimate of $\hat{\sigma}^*(i)$. The modification is to use a trimmed sequence $\{X_tY_{t+q}\}$ by trimming both the lower and upper 0.2 percentiles of X_tY_{t+q}.

As an alternative, we also applied the stationary bootstrap method of [16] to estimate $\mathrm{Var}(\hat{\rho})$. Each bootstrap step was repeated 1000 times. Let $B(1)$

Statistic	Mean	S.D	**Percentile**			Rej. at $\chi_8^2(0.95)$
			90%	95%	99%	percentage
$S(1)$	10.81	5.81	18.20	21.67	30.30	20.3
$S(2)$	11.63	8.91	20.94	25.94	37.48	22.2
$T(1)$	10.88	6.89	18.26	22.04	32.5	17.3
$T(2)$	9.14	6.66	15.94	19.27	28.66	10.8
χ_8^2	8	4	13.36	15.51	20.10	5.0
$T^*(1)$	8.13	4.29	13.82	16.35	21.60	6.5
$T^*(2)$	7.01	3.93	11.99	14.01	20.11	3.4
$B(1)$	7.72	4.1	13.05	15.32	20.81	4.8
$B(2)$	6.31	3.66	11.03	13.65	18.47	4.0

Table 1: Empirical quantiles of various test statistics for testing zero canonical correlations, based on 2,000 replications with sample size 2,000.

and $B(2)$ be the corresponding test statistics $-(n-s)\log(1-\frac{\hat{\rho}^2}{\hat{d}})$, where $\hat{d}$ is obtained from bootstraps.

Table 1 compares empirical percentiles and the size of various test statistics discussed above for the model in Eq. (10) when the sample sizes is 2000, which is common among financial data. The corresponding quantiles of the asymptotic χ_8^2 are also given in the table. Other sample size is also considered. From the table, we make the following observations. First, the T^* and bootstrap B statistics perform reasonably well when the sample size is sufficiently large. The bootstrap method outperforms the other test statistics. However, it requires intensive computation. For instance, it took several hours to compute the bootstrap tests in Table 1 whereas it only took seconds to compute the other tests. Second, the T statistics underestimate the variance of cross-correlation so that the empirical quantiles exceed their theoretical counterparts. Third, as expected, the S statistics perform poorly for both sample sizes considered. Fourth, the performance of the proposed test statistic T^* indicates that the [2] method to estimate the variance of cross-covariance is reasonable in the presence of GARCH effects provided that robust estimators $\hat{\sigma}^*(i)$ are used.

6 An illustrative example

In this section we apply the proposed test statistics to a 3-dimensional financial time series. The data consist of daily log returns, in percentages, of stocks for Amoco, IBM, and Merck from February 2, 1984 to December 31, 1991 with 2000 observations. The series are shown in Figure 1. It is well-known that daily stock return series tend to have weak dynamic dependence, but strong conditional heteroscedasticity, making them suitable for the proposed test. Our goal here is to provide an illustration of specifying a vector ARMA

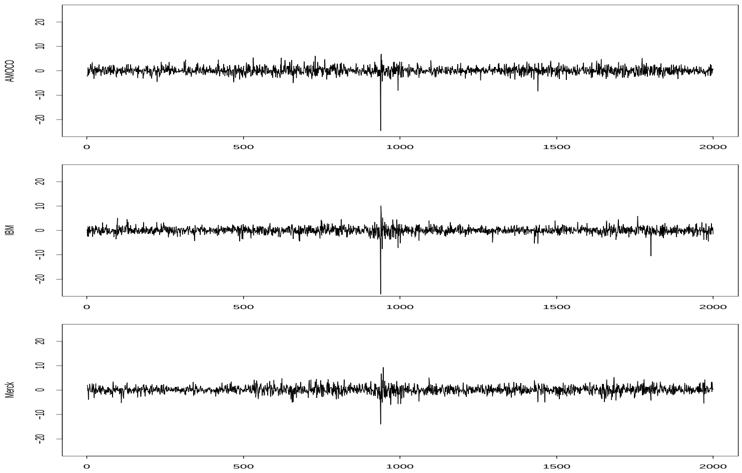


Figure 1: Time series of Amoco, IBM and Merck stocks daily return (2/2/1985–12/31/1991).

model with GARCH innovations rather than a thorough analysis of the term structure of stock returns.

Denote the return series by $\mathbf{Z}_t = (Z_{1t}, Z_{2t}, Z_{3t})'$ for Amoco, IBM, and Merck stock, respectively. Following the order specification procedure of Section 2.2, we apply the proposed test of Eq. (9), denoted by T^*, to the data and summarize the test results in Table 2. We also included the test statistics T of Eq. (7) for comparison purpose. The past vector $\mathbf{P}_t$ is determined by the AIC as $\mathbf{P}_t = (\mathbf{Z}'_{t-1}, \mathbf{Z}'_{t-2})'$. The p-value is based on a χ^2_{ks-f+1} test where $k = 3$, $s = 2$, and $f = \dim(\mathbf{F}^*_t)$.

From Table 2, the proposed test statistic T^* identified $\{1, 1, 1\}$ as the Kronecker indexes for the data, i.e. $K_i = 1$ for all i. On the contrary, if one assumes that there are no GARCH effects and uses the test statistic T, then one would identify $\{1, 1, 2\}$ as the Kronecker indexes. More specifically, the T statistic specifies $K_1 = K_2 = 1$, but finds the smallest canonical correlation between $\mathbf{F}^*_t = (Z_{1,t}, Z_{2,t}, Z_{3,t}, Z_{3,t+1})$ and $\mathbf{P}_t$ to be significant at the usual 5% level. To determine K_3, one need to consider the canonical correlation analysis between $\mathbf{F}^*_t = (Z_{1,t}, Z_{2,t}, Z_{3,t}, Z_{3,t+1}, Z_{3,t+2})'$ and the past vector $\mathbf{P}_t$. The corresponding test statistic is $T = 4.05$, which is insignificant with p-value 0.134 under the asymptotic χ^2_2 distribution. Therefore, without considering GARCH effects, the identified kronecker indexes are $(K_1 = 1, K_2 = 1, K_3 = 2)$, resulting in an ARMA(2,2) model for the data. Consequently, by correctly considering the GARCH effect, the proposed test statistic T^* was able to specify a more parsimonious ARMA(1,1) model for the data. In summary, we entertain a vector ARMA(1,1) model with diagonal GARCH(1,1) innovations for the data. The estimated VARMA-GARCH model is given below:

$$\mathbf{Z}_t - \begin{bmatrix} .0 & 2.0^{***} & .4 \\ .0 & 0.3 & .2^{*} \\ .1 & 0.9^{**} & .1 \end{bmatrix} \mathbf{Z}_{t-1} = \begin{bmatrix} -.1 \\ -.0 \\ .1^{*} \end{bmatrix} + \mathbf{a}_t + \begin{bmatrix} -.1 & 2.1^{***} & .4 \\ .1 & 0.2 & .2^{**} \\ .2^{*} & 0.9^{**} & .1 \end{bmatrix} \mathbf{a}_{t-1} \tag{11}$$

where the superscript *, **, and *** indicate significance at the 10%, 5% and 1% level, respectively, and the volatility $\mathbf{g}_t = E(\mathbf{a}_t^2 | \mathcal{F}_{t-1})$ follows the model

$$\mathbf{g}_t = \begin{bmatrix} 1.59 \\ 0.23 \\ 0.05 \end{bmatrix} + \begin{bmatrix} .28 & 0 & 0 \\ 0 & .14 & 0 \\ 0 & 0 & .06 \end{bmatrix} \mathbf{a}_{t-1}^2 + \begin{bmatrix} .00 & 0 & 0 \\ 0 & .76 & 0 \\ 0 & 0 & .91 \end{bmatrix} \mathbf{g}_{t-1}$$

where all estimates except the (1,1)th element of the coefficient matrix of $\mathbf{g}_{t-1}$ are significant at the 1% level. Model checking shows that the fitted model appears to be adequate in handling serial dependence in the data.

future subvector $\mathbf{F}_t^*$	sm.can.cor	T^*	d.f	p-value	Remark	T
$(Z_{1,t})$	.130	33.96	6	0		33.96
$(Z_{1,t}, Z_{2,t})$	.116	26.97	5	0		26.97
$(Z_{1,t}, Z_{2,t}, Z_{3,t})$	.101	20.68	4	0		20.68
$(Z_{1,t}, Z_{2,t}, Z_{3,t}, Z_{1,t+1})$	.051	5.59	3	.13	$K_1 = 1$	5.95
$(Z_{1,t}, Z_{2,t}, Z_{3,t}, Z_{2,t+1})$	.032	1.52	3	.68	$K_2 = 1$	4.48
$(Z_{1,t}, Z_{2,t}, Z_{3,t}, Z_{3,t+1})$	.055	5.98	3	.11	$K_3 = 1$	11.38

Table 2: Model specification for three daily stock Returns

References

[1] Akaike H. (1976). *Canonical correlation analysis of time series and the use of an information criterion.* Systems identification: Advances and Case Studies, eds R. K. Methra and D. G. Lainiotis. New York: Academic Press, 27 − 96.

[2] Berlinet A., Francq C. (1997). *On Bartlett's formula for non-linear processes.* Journal of Time Series Analysis **18**, 535 − 552.

[3] Bollerslev T. (1986). *Generalized autoregressive conditional heteroscedasticity.* Journal of Econometrics **31**, 307 − 327.

[4] Bollerslev T., Engle R.F., Nelson D.B. (1994). *ARCH models.* Handbook of Econometrics **IV**. Elsevces Science B.V., 2959-3-38,

[5] Box G.E.P., Jenkins G.M. (1976). *Time series analysis: forecasting and control.* San Francisco, CA: Holden-Day.

[6] Box G.E.P., Tiao G.C. (1977). *A canonical analysis of multiple time series.* Biometrika **64**, 355 − 365.

[7] Cooper D.M., Wood E.F. (1982). *Identifying multivariate time series models.* Journal of Time Series Analysis **3**, 153 − 164.

[8] Dunsmuir W., Hannan E.J. (1976). *Vector linear time series models.* Advances in Applied Probability **8**, 339 – 364.

[9] Engle R.F. (1982). *Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation.* Econometrica **50**, 987 – 1008.

[10] Engle R.F., Granger C.W.J. (1987). *Co-integration and error-correction: representation, estimation and testing.* Econometrica **55**, 251 – 276.

[11] Engle R.F., Kroner K.F. (1995). *Multivariate simultaneous generalized ARCH.* Econometric Theory **11**, 122 – 150.

[12] Hannan E.J., Deistler M. (1988). *The statistical theory of linear systems.* John Wiley, New York.

[13] Hotelling, H. (1936). *Relations between two sets of variables.* Biometrika **28**, 321 – 377.

[14] Min W.L., Tsay R.S. (2004). *On canonical analysis of multivariate time series.* Working paper, GSB, University of Chicgao.

[15] Quenouille M.H. (1957). *The analysis of multiple time series.* London: Griffin.

[16] Romano J.P., Thombs L.A. (1996). *Inference for autocorrelations under weak assumptions.* Journal of the American Statistical Association **91**, 590 – 600.

[17] Tiao G.C., Tsay R.S. (1989). *Model specification in multivariate time series (with discussion).* Journal of the Royal Statistical Society. Ser. B **51**, 157 – 213.

[18] Tsay R.S. (1989a). *Identifying multivariate time series models.* Journal of Time Series Analysis **10**, 357 – 371.

[19] Tsay R.S. (1989b). *Parsimonious parametrization of vector autoregressive moving average models.* Journal of Business and Economic Statistics **7**, 327 – 341.

[20] Tsay R.S. (1991). *Two canonical forms for vector ARMA processes.* Statistica Sinica **1**, 247 – 269.

[21] Tsay R.S. (2002). *Analysis of financial time series.* John Wiley: New York.

[22] Tsay R.S., Tiao G.C. (1985). *Use of canonical analysis in time series model identification.* Biometrika **72**, 299 – 315.

[23] Wu W.B. (2003). *Empirical processes of long-memory sequences.* Bernoulli **9**, 809 – 831.

Acknowledgement: We thank Dr. G.Tunnicliffe-Wilson for helpful comments and U.S. National Science Foundation and Graduate School of Business, University of Chicago for partial financial support.

Address: W. Min, R.S. Tsay, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicgao. IL 60637, U.S.A.

E-mail: `ruey.tsay@gsb.uchicago.edu`

LEARNING STATISTICS BY DOING OR BY DESCRIBING: THE ROLE OF SOFTWARE

Erich Neuwirth

Key words: Statistical computing, statistics education, teaching statistics.

COMPSTAT 2004 section: Teaching statistics.

Abstract: Paper discusses several key questions connected with the teaching, and learning, statistics. Among the problem covered belong: whom to teach, what type of presentation to chose, how and in which extend to use computers . . .

1 Teaching statistics: for whom?

Statistics possibly is the discipline use by most nonspecialists as part of their work. Psychologists, medical doctors, journalists, and people from many more fields they all use statistics, or at least have to be able to interpret statistical data. At election times, newspapers and TV report about opinion polls and most of the public has problems in judging the reliability of forecasts for the election based on samples. So the need of education a rather broad audience for statistics is generally accepted.

When discussing statistics education under these aspects, it is clear that we have to face different audiences with different statistical need and we also have to take into account quite different levels of formal training outside of statistics.

Statistics education may target the following knowledge levels:

- Basic statistical knowledge: understanding simple statistical summaries and graphs, numeracy.
- Basic statistical skills: selecting appropriate simple statistical methods for own analyses, ability to immediately identify misuses of statistics.
- Advanced statistical knowledge: Understanding complex methods, especially multivariate analytical and graphical methods.
- Advanced statistical skills: selecting appropriate complex methods and understanding their role in gaining insights.

We need to distinguish the level of presentation for statistics education

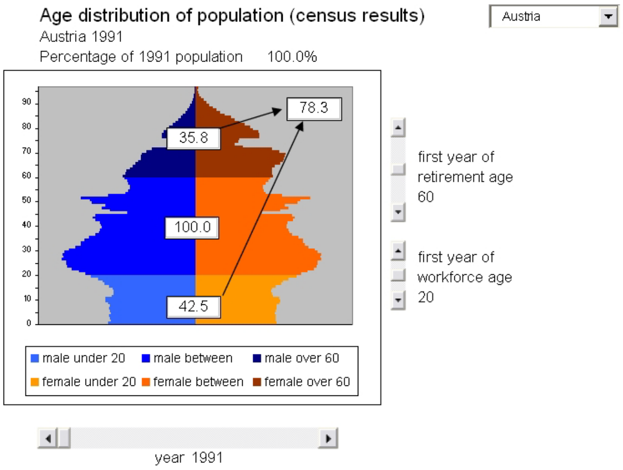
- No formal prerequisites, just data as numbers and graphs.

- Basic mathematical knowledge and skills, simple algebraic formulas admissible as tools for explaining.
- College level mathematical background

Finally, the level of computer expertise of the educatees also plays an important role in designing courses and activities for statistics education.

2 Demographical modelling: a success story

Let us begin with a success story. In Austria, like in many other countries, there is an ongoing discussion about different options of financing the retirement system. At `http://sunsite.univie.ac.at/Projects/demography/` we have published a manipulable statistical model forecasting the population's age structure for Austria for the next 30 years. The model is implemented as an Excel sheet, and it looks like this:



Age distribution of population (census results)
Austria 1991
Percentage of 1991 population 100.0%

Austria

first year of retirement age 60

first year of workforce age 20

year 1991

The most important details in this model are the "sliders"; they allow to change the graph dynamically. The horizontal slider turns the graph into a movie. The graph always displays the population pyramid for a given year; when the slider is moved, the year changes and the change of the age structure becomes dynamically visible.

The other sliders allow to change different model parameters like retirement age, and will immediately display changes in the system resulting from changes in the parameters. The model also allows to use data from different countries (currently we have Austria, Germany, USA, and Japan) to analyze how different populations structures can get.

This model can be used in two ways:

- As a ready made tool for experiments in a given framework, one might say as a demographical microworld.
  This is the “consumer mode” for the model.
- As a project to be developed by the learner. This is the “producer mode” for the model.

A nice story illustrates how the model was used for statistical education in “consumer mode”. The author received an email message from a member of the Austrian parliament, essentially stating that the MP had found the model accidentally when browsing the web. Being involved in discussions about retirement legislative questions he started playing with it and found that he could analyze some consequences of changes in retirement laws easily. The final statement was: “now I understand the problem much better”.

The author regularly teaches a course about computer based demographic modelling for sociology students. In this course, the students are shown the model at the beginning. Then there are two days of intensive computer based modelling, and at the end all the students are able to implement the model themselves. They also are guided towards further investigations. e.g. the influence of changing birth rates on demographical developments. They implement different scenarios and study possible changes with hands on modelling and parameter variation. The students really enjoy this course because the finish with the feeling that they have acquired knowledge and skills allowing them to add statistical modelling to their personal toolkit.

3 Learning statistics for data analysis: how?

The didactic success of the demography model just described is very much tied to information technology. The finished version can be downloaded on the web, the user only needs Microsoft Excel on the computer. So a very widely used tool is the computational infrastructure of the model. This also has an additional important message: serious statistical modelling can be done with software available on almost any desktop computer, quite often there is no immediate need for highly specialized software for models of higher complexity.

When the model is used in producer mode, the statistical and mathematical theory for the model is not too complicated. Mathematically speaking this is a simple linear first order difference equation. It may, however, be described only using basic arithmetic. Since it is implemented in Excel and since the students know Excel already, the important message is that serious modelling can be done with widely available general purpose software. Like in the consumer mode use of the model a case is made for modelling as a mental process and not a function of highly specialized software.

Using spreadsheet programs also has another important didactical aspect. Spreadsheets always display the data, data are not hidden. One of the most important concepts of statistics is the data matrix, also called data frame. In a spreadsheet, the data are always visible and it becomes a very physical experience that doing statistics is operating on data. This fact is much more obscured when a statistical programming language like S, R, SPSS, or SAS is used as the basic tool in statistics courses.

The main difference between the spreadsheet approach and the statistical programming language approach might be characterized as direct manipulation vs. descriptive. The programming language approach is much more formula based, the data are not as omnipresent as in the spreadsheet approach. For introductory statistics courses, this constant reminder "statistics is about data" can be quite helpful. Many students after their first course of non computer based statistics have the impression that statistics is about certain types of formulas, and not so much about data. Programming languages still somewhat support this mindset, whereas the spreadsheet approach really emphasizes the data analysis point of view. More topics about modelling with spreadsheets can be found in [6].

The direct manipulation approach is not solely restricted to spreadsheets. Programs like Fathom (available from Key Curriculum Press) also emphasize the "manipulate the data with the mouse" approach as opposed to the "write a program to manipulate the data" approach.

Spreadsheets are not the answer to all statistical problems. Excel has some flaws concerning statistics. The most inconvenient ones are some inaccuracies with distributions functions and not too high quality of random number generators, inconsistent handling of missing data, and unavailability of some of the most important types of statistical graphs (like histograms with unequal bin widths).

Therefore, it makes sense to use a more advanced statistical toolbox than just a spreadsheet program. This does now, however, imply that the spreadsheet paradigm has to be thrown overboard. The RExcel program (part of the R COM server project accessible at `http://sunsite.univie.ac.at/rcom/` and described in [5]) allows to use practically all the functionality of R from within Excel. This way, the student can still operate on the data in with the direct manipulation method, but use statistical methods not available from the spreadsheet program alone.

This also demonstrates an important message about software in general: Software should adapt to the user's needs. If possible, one should not be forced to switch programs, it is better if a standard package can be enhanced by extending its functionality.

RExcel is not the only statistical extension of Excel. PopTools (available from `http://sunsite.univie.ac.at/Spreadsite/poptools`) also is an example of how additional statistics functions can be integrated into the spreadsheet paradigm.

Statistical graphics is another extremely important concept to be discussed in the context of statistics education. [1] and [9] make a very convincing case for graphical methods. The statistics package R (available from `http://www.r-project.org` comes with many data sets, including data about age, sex, class and survival of the Titanic passengers and crew. Quite a few statistics teachers investigate this data set with mosaic plots (and without any formulas visible for the students). Again, this illustrated the point we already made: statistics should help gaining insights from data, and not be a way of just applying formulas to data. Similarly, trellis plots are a relatively new technique for multivariate analysis by using arrays of graphs arranged according to statistical variables.

So far, we have only discussed software supporting statistical education running on desktop or notebook computers. Additionally, there is a whole range of web sites for statistics education, offering course material and applets for experimenting. `http://wise.cgu.edu/` offers a good overview of such sites.

Some of these sites are just online resources, not offering much more than printable static material to support statistics courses. The more interactive sites follow a philosophy similar to the one exemplified by consumer mode use of our demography example. They offer the students opportunities to analyze data interactively. Projects like the XploRe eBooks (available from `http://www.xplore-stat.de`) combine the printed material approach and the applet approach by directly embedding applets into electronically distributed static course materials.

One of the central problems of teaching statistics mostly as a data analysis course is to find data which are interesting to analyze for students. For this purpose, the WWW is a really powerful resource. The Journal of Statistics Education at `http://www.amstat.org/publications/jse/` has an extensive collection of data especially selected for educational purposes, and StatLib (at `http://lib.stat.cmu.edu`) has a large collection of datasets cited in the statistical literature, especially in textbooks for introductory statistics.

All these datasets have the disadvantage that the students “do not connect” with them. The author therefore since 10 years collects data from his students with a questionnaire and uses these data throughout the statistics courses. The questions are what one would expect: subject area, weight, size, size of parents, grades in some school subjects and so on. The advantage of using this data set is that for each analysis each student sees his or her place in the result, and therefore feels to have learned something about a group he or she belongs to. To the author’s experience the students become quite interested in the final report they have to produce, and sometimes they take the challenge of designing statistical questions which can be analyzed with this data set.

Information technology plays an important role in collecting these data quickly. If the group is small enough, a Palm handheld calculator with questionnaire software (Pendragon forms from Pendragon Software) is used to collect the data in the classroom. At the end of the class period, the handheld is connected to a notebook computer, the data are transferred, and then immediately a first step of the analyze can be performed in front of the students. The message of doing it this way is that collecting data can be set up quite conveniently, and therefore with good planning statistics be used very quickly. For larger classes, a browser based questionnaire is used. As part of this project, students also start asking questions about the privacy of their data and so are exposed to the problems of collecting data through their own experience as part of the course.

All the projects and tools so far mostly have been concerned with analyzing data. An important area in statistics education we have not considered yet is probability. This is the topic of the next section.

4 Learning probability for statistics

As most statistic teachers have experienced, probability is important as one of the foundations of statistics, but it is rather hard to teach if the students are supposed to learn more than just a few formulas. One of the main problems is that students may misunderstand probability as a somewhat strange packaging of combinatorics. Information technology in this case allows us to add something which is not so easy without computers: experiments through Monte Carlo simulation. Chapter 7 in [6] demonstrates the basic techniques of such simulations with spreadsheets. Again, the important message is that this can be done with readily available software.

The danger when using a Monte Carlo approach to teach probability is that students only learn that "computer generated randomness" behaves like probability theory predicts, and do not connect this with "everyday" randomness. Therefore, it is very important to perform experiments using physical randomness with a device like a Galton Board (sometimes called Quincunx) and then build a Monte Carlo simulation for the same phenomenon. Comparing the outcome of "real" randomness and simulated randomness can convince the students that computer simulations are close enough to reality and therefore problems which are more or less unaccessible for real experiments can be studied with Monte Carlo simulations.

A software category we have not discussed at all so far are CAS, Computer Algebra Systems. The most known programs in this category are Mathematica, Maple, MuPAD and Derive. There are special toolkits for doing statistics and probability with CAS, see for example [2] and [8]. The approach there is somewhat different from the spreadsheet approach. The CAS program is used as a specialized programming language, and the experiments are performed by using custom made functions in this programming language.

Monte Carlo Simulations can be considered as computer implementations for the law of large numbers. A difficult topic when dealing with probability is the relation between the law of large numbers and the central limit theorem. Using computers for both Mote Carlo simulations and numerical calculations of probabilities for sums of independent random variables allows to connect numerical-analytical models with simulated randomness and show that probability is able to model randomness reasonably well.

Once the trust in simulations is built, they can be used to empirically verify facts about statistical tests and confidence intervals. Without computers, it is practically impossible to illustrate concepts like the errors of first and second kind of a test and confidence levels of confidence intervals empirically. Monte Carlo simulations once again allow us to study the empirical error rates of simulated tests and compare them with the theoretical values.

Sampling also is a very important concept in statistics. In Monte Carlo simulations, the machinery in the background produces a sequence of numbers. It does not select from a given set, it produces a new number each time it is asked for one. We might say that the random number generator is spitting out an infinite sequence of random numbers. When sampling is investigated, it is very helpful if for experimental activities we can see the whole sample space and then select the sample from this set. Spreadsheets allow us to make this process very visual. From a didactical point of view, it seems very important to clearly model the process of selecting from a given well defined finite set and not blur the lines to the production of random numbers by some unpredictable machinery.

When probability is studied, combinatorics also has to be investigated. The relationship between probability and randomness is the equal probability assumption. This is something that cannot be proved analytically. Therefore, helping to build trust in the assumption is very important for the learner. Monte Carlo experiments can play a key role for that. In this area, computers cannot only be used for simulations, that also can play an important role in better understanding combinatorics.

Just read the following description: Let us build a table. The first column is filled with 1s. The rest of the first row is filled with 0s. All the other cells contain the sum of the number above and the number above and to the left. This is a complete and completely operational description of the binomials.

This description is not only a description, it is the complete instruction to compute the binomials with a spreadsheet. Additionally, it tells that each number in each row migrates down into the next row exactly twice, once vertically and once diagonally. therefore, row sums double from row to row and this description contains the proof of the fact that the row sums of the binomials are the powers of 2.

Expressing this more formally, the binomials can be described by a two term recursion. It turns out that this kind of recursion covers most of the combinatorics problems needed for basic probability models. Therefore, the table approach to combinatorics covers most of the ground needed in an introductory course. Once again, the readily available tool spreadsheet can be used to analyze structures, and to help understand concepts, not just as a more convenient kind of pocket calculator.

5 Some final thoughts

Statistics and probability have their origin in methodology to analyze empirical data and gain insights. So at the beginning of these subjects, there often is an experiment. Without computers, it is very hard to create this experiment based situation as a general setting. Some example highlight are possible, but overall statistics courses without computers are paper and pencil based theory courses (or not too interesting computations courses for very small data sets). With computers, we can analyze real or at least realistic data sets, and we can study probability also with an experimental approach. Therefore, for many learners who are not mostly interested in theory but in methods they can apply in their daily lives, this approach is much more promising than computer free statistics. As a consequence, it might be reasonable to avoid computers as an aid to learning in some specialized areas of statistics. But overall, information technology allows to make statistical concepts and methods both more accessible and more useful for a very wide audience.

References

[1] Friendly M. (2000). *Visualizing categorical data.* SAS Institute 2000.

[2] Hastings K. (2000). *Probability with mathematica.* Lewis Publishers.

[3] Neuwirth E. (2002). *Recursively defined combinatorial functions: extending Galton's board.* Discrete Math. **239**, 33 – 51.

[4] *Embedding R in standard software, and the other way round.* In Hornik K. and Leisch, F. (eds.), DSC 2001 Proceedings, `http://www.ci.tuwien.ac.at/Conferences/DSC-2001`

[5] Neuwirth E., Baier T. (2001). *Embedding R in standard software, and the other way round.* In Hornik K., Leisch, F. (eds.), DSC 2001 Proceedings, `http://www.ci.tuwien.ac.at/Conferences/DSC-2001`

[6] Neuwirth E., Arganbright D. (2003). *The active modeler: mathematical modeling with Excel.* Brooks-Cole.

[7] Neuwirth E. *Probababilities, the US electoral college, and generating functions considered harmful.* To appear in International Journal of Computers for Mathematical Learning.

[8] Rose C., Smith D. 2002. *Mathematical statistics with mathematics.* Springer Verlag.

[9] Tufte E. (2001). *The visual display of quantitative information.* Graphics Press.

Acknowledgement: Thanks to Jaromir Antoch and Marlene Müller for their patience.

Address: E. Neuwirth, University of Vienna, Austria

E-mail: `erich.neuwirth@univie.ac.at`

EMBEDDING METHODS AND ROBUST STATISTICS FOR DIMENSION REDUCTION

George Ostrouchov and Nagiza F. Samatova

Key words: Dimension reduction, convex hull, FastMap, principal components, multidimensional scaling, robust statistics, Euclidean distance.

COMPSTAT 2004 section: Dimensional reduction.

Abstract: Recently, several non-deterministic distance embedding methods that can be used for fast dimension reduction have been proposed in the machine learning literature. These include FastMap, MetricMap, and SparseMap. Among them, FastMap, implicitly assumes that the objects are points in a p-dimensional Euclidean space. It selects a sequence of $k \leq p$ orthogonal axes defined by distant pairs of points (called pivots) and computes the projection of the points onto the orthogonal axes. We show that FastMap picks all of its pivots from the vertices of the convex hull of the data points in the original implicit Euclidean space. This provides a connection to results in robust statistics, where the convex hull is used as a tool in multivariate outlier detection and in robust estimation methods. The connection sheds a new light on some properties of FastMap and provides an opportunity for a robust class of dimension reduction algorithms that we call RobustMaps, which retain the speed of FastMap and exploit ideas in robust statistics. One simple RobustMap algorithm is shown to outperform principal components on contaminated data both in terms of clean variance captured and in terms of time complexity.

1 Introduction

Dimension reduction starts with n objects as points in a p-dimensional vector space and maps the objects onto n points in a k-dimensional vector space, where $k < p$. A more general situation arises when the point coordinates are not known and only pairwise distances (or a distance function to compute them) are available. This mapping of objects based on their distances only into a k-dimensional vector space is called finite metric space embedding [8]. Several embedding methods and their properties are discussed in [8], including FastMap, MetricMap, and SparseMap. The discussion centers mostly on whether the embeddings are contractive, a property of importance in similarity searching that guarantees no missed items. In this paper, we concentrate on FastMap and its properties that connect the technique to ideas in robust statistics.

FastMap is first introduced in [6] as a fast alternative to Multidimensional Scaling (MDS) [14] and a generalization of Principal Component Analysis (PCA) [9]. Given dimension k and Euclidean distances between n objects, FastMap maps the objects onto n points in k-dimensional Euclidean space. An implicit assumption by FastMap that the objects are points in a p-dimensional Euclidean space ($p \geq k$) is noted in [8]. Because of this assumption, FastMap is usually viewed as a dimension reduction method.

When FastMap begins with Euclidean distances between the n objects, it has time complexity $\mathrm{O}(n)$. If the Euclidean distances must be explicitly computed from a p-dimensional vector representation, FastMap time complexity is $\mathrm{O}(np)$.

We show how FastMap operates within the the implicit or explicit p-dimensional Euclidean space containing the points of a data set. FastMap selects a sequence of $k \leq p$ orthogonal axes defined by distant pairs of points (called pivots) and computes the projections of the points onto the orthogonal axes. We show that FastMap picks all of its pivots from convex hull vertices of the original data set. This provides a connection to results in robust statistics, where the convex hull is used as a tool in multivariate outlier detection and in robust estimation methods. The connection sheds a new light on some properties of FastMap, in particular its sensitivity to outliers, and provides an opportunity for a new class of dimension reduction algorithms that retain the speed of FastMap and exploit ideas in robust statistics.

We begin in Section 2 by defining the convex hull and some of its properties. In Section 3 we describe the FastMap algorithm. The main result, showing that FastMap pivots are pairs of vertices of the convex hull is in Secion 4. Section 5 discusses the implications of this result and finally Section 6 presents an algorithm, RobustMap, that results from these implications. Some further comments and conjectures about connections to QR and QLP factorizations [13] are also made.

2 Convex hull of a data set

Let S be a set of n points in p-dimensional Euclidean space. The convex hull of S, denoted by $C(S)$, is the smallest convex set (a polytope) that contains S [5], [7]. We can visualize a convex hull in two or three dimensions as a rubber band or an elastic bag stretched around the points. In higher dimensions, we must rely on more formal properties of hyperplanes, and the notion of half-space support. Our definitions below are mostly from [5], [7].

Definition 2.1. *A hyperplane is an affine subspace (a translation of a linear subspace) of* $\mathbf{R}^p$ *with dimension* $p-1$.

The set of points

$$h(u,v) = \{x \in \mathbf{R}^p : (u-v)^T(x-v) = 0\}, \text{for } u, v \in \mathbf{R}^p, \tag{1}$$

is a hyperplane perpendicular to the vector $u-v$ and passing through v. The *closed half-space* that is defined by this hyperplane and that contains u is given by

$$H(u, v) = \{x \in \mathbf{R}^p : (u - v)^T (x - v) \geq 0\}, \text{for } u, v \in \mathbf{R}^p, \tag{2}$$

Definition 2.2. *If S intersects $h(u, v)$ and S lies in $H(u, v)$ for some $u, v \in \mathbf{R}^p$, then $h(u, v)$ is a supporting hyperplane of S and $H(u, v)$ is a supporting half-space of S.*

We use Ziegler's [15, section 2.1] definition of a *face* of a polytope and state it in terms of a supporting hyperplane.

Definition 2.3. *A face of a polytope $C(S)$ is any set of the form*

$$C(S) \cap h(u, v),$$

where $h(u, v)$ is a supporting hyperplane of S for some $u, v \in \mathbf{R}^p$. Further, for a p-dimensional polytope, facets are $(p-1)$-dimensional, ridges are $(p-2)$ dimensional, edges are 1-dimensional, and vertices are 0-dimensional.

The above characterization of a vertex as a single point (a 0-dimensional face) of $C(S)$ that lies in the supporting hyperplane, will be used in Section 4 to link FastMap pivots to vertices of the convex hull.

3 FastMap overview

Given the Euclidean distance between any two points (objects) of S, k iterations of FastMap produce a k-dimensional ($k \leq p$) representation of S. Each iteration selects from S a pair of points, called *pivots*, that define an axis and computes coordinates of the S points along this axis. The pairwise distances for S can then be updated to reflect a projection of S onto the subspace (a hyperplane passing through the origin) orthogonal to this axis. The next iteration implicitly operates on the projected S in the subspace. However, these projections are accumulated and jointly performed only for the distances that are needed. In this manner, after k iterations, the S points end up with k coordinates giving their k-dimensional representation.

To provide details of the FastMap algorithm, we first introduce some notation. Let (a_i, b_i) be the pair of pivot elements from S at iteration i. Let $d_i(x, y)$ be the Euclidean distance between points x and y of S after their ith projection onto a pivot-defined hyperplane, so that $d_0(x, y)$ is the initial Euclidean distance. Also, let x_i be the ith coordinate of x in the resulting k-dimensional representation of $x \in S$.

Pivot elements are chosen by the *choose-distant-objects* heuristic shown in Fig. 1. Initially, $i = 0$. After selecting a pivot pair (a_i, b_i), the ith coordinate of each point $x \in S$ is computed as

$$x_i = \frac{d_{i-1}^2(a_i, x) + d_{i-1}^2(a_i, b_i) - d_{i-1}^2(b_i, x)}{2d_{i-1}(a_i, b_i)}. \tag{3}$$

```
Choose-distant-objects ( $S, d_i(,)$ )
  1. Choose an arbitrary object $s \in S$
  2. Let $a_{i+1}$ be the $a \in S$ that maximizes $d_i(a, s)$
  3. Let $b_{i+1}$ be the $b \in S$ that maximizes $d_i(b, s)$
  4. Report $a_{i+1}$ and $b_{i+1}$ as the distant objects.
```

Figure 1: Choose-distant-objects heuristic for iteration i.

This projection is based on the law of cosines and current distances from the two pivot points. The distances are updated whenever needed in *Choose-distant-objects* or in (3). An update for a single iteration is presented in [6] and we extend this in [1] to a combined update

$$d_i^2(x, y) = d_0^2(x, y) - \sum_{j=1}^{i} (x_j - y_j)^2. \tag{4}$$

This is based on the Pythagorean theorem and the sequence of i projections onto hyperplanes perpendicular to pivot axes.

There are k iterations, each requiring $\mathrm{O}(n)$ distance computations of $\mathrm{O}(p)$. The resulting total time complexity is $\mathrm{O}(npk)$. Note that if all the original distances are already available, the total time complexity is $\mathrm{O}(nk^2)$ due to the sum in (4). If k is a small constant compared to n and p, as is usually the case, k is dropped from the above complexity statements giving those we provided in the Introduction.

4 FastMap and vertices of the convex hull

Here we prove the main result of this paper, namely that all pivot points are selected from vertices of the convex hull of the data set. We do this in two steps. First we show that the *Choose-distant-object* heuristic pivot pair is a pair of convex hull vertices within the current working subspace. Then we show that if a point is a vertex in a subspace projection, it is also a vertex in the original p-dimensional space.

The *Choose-distant-objects* heuristic first takes an arbitrary point $b \in S$ and finds $a \in S$, the most distant point from b. Because a is the most distant point in S from b

$$(s - b)^T (s - b) \leq (a - b)^T (a - b), \forall s \in S. \tag{5}$$

Now, for any point $s \in S$ distinct from a, we have

$$\begin{aligned}
0 &< (s-a)^T(s-a) \\
&= (s-b+b-a)^T(s-b+b-a) \\
&= (s-b)^T(s-b) + 2(s-b)^T(b-a) + (b-a)^T(b-a) \\
&\leq 2(s-b)^T(b-a) + 2(b-a)^T(b-a) \qquad \text{by (5)} \\
&= 2(s-b+b-a)^T(b-a) \\
&= 2(s-a)^T(b-a) \qquad (6)
\end{aligned}$$

If we add $s = a$ in (6), we have

$$0 \leq (s-a)^T(b-a), \forall s \in S,$$

which defines a supporting half space $H(a, b)$ for all points in S. Since a is the only point in the supporting hyperplane $h(a, b)$ of S, it must be a single point face of $C(S)$. This, by Definition 2.3, is a vertex of $C(S)$.

Next, the *Choose-distant-objects* heuristic finds the point in S most distant from a. By the same argument this is again a vertex of $C(S)$. We state this as a lemma.

Lemma 4.1. *A single application of the* Choose-distant-objects *heuristic to a set of points S returns a pivot pair of points that are among the vertices of $C(S)$.*

After choosing a pair of vertices, FastMap projects the set S into a subspace orthogonal to the vector defined by the pivot pair (a, b) and repeats the *Choose-Distant-Objects* heuristic in the subspace of dimension $p - 1$. Pivot pairs and projections are computed until suitably many orthogonal vectors are extracted to be used as the principal axes of the lower dimensional representation of S. So far, we have shown that a pivot pair is a pair of convex hull vertices within its current working subspace. Are they all also vertices of $C(S)$ in the original space? The answer is yes, subject to a uniqueness caveat requiring that no pair of points (except the current pivot points) get projected onto the same point. Assuming that the points S are in sufficiently *general position* [15] takes care of this. Because we have a finite set of points, we can perturb them by an arbitrarily small amount to achieve such a general position. We show that a vertex in a subspace projection is a vertex in the original p dimensional space.

Let P_H be a symmetric projection matrix into a subspace $H \subset \mathbf{R}^p$ and let $S_H = \{P_H u : u \in S\}$ be the set of image points of S in this subspace. We also need to assume that S are in sufficiently general position so that all vertices of $C(S_H)$ are projections of distinct points of S.

Lemma 4.2. *If $P_H s$ is a vertex in the convex hull of S_H and S are in general position, then s is a vertex in the convex hull of S.*

Proof Since $P_H s$ is a vertex of $C(S_H)$, by Definition 2.3

$$P_H s = C(S_H) \cap h(u, v),$$

where $h(u, v)$ is a supporting hyperplane of $C(S_H)$ for some $u, v \in H$. Because $P_H s \in h(u, v)$, there is a $u' \in H$ such that $h(u, v) = h(u', P_H s)$. Now, $P_H s$ is the only point of S_H that is in the supporting hyperplane, so that

$$(u' - P_H s)^T (P_H x - P_H s) > 0,$$

for all $P_H x \in S_H$ distinct from $P_H s$. Because S are in general position,

$$(u' - P_H s)^T (P_H x - P_H s) > 0, \forall x \in S \text{ distinct from } s.$$

Then,

$$\begin{aligned}(u' - P_H s)^T [x - (I - P_H)x - s + (I - P_H)s] &> 0\\ (u' - P_H s)^T (x - s) - (u' - P_H s)^T (I - P_H)(x - s) &> 0.\end{aligned}$$

Since $P_H(u' - P_H s) = (u' - P_H s)$ (because $u' \in H$),

$$(u' - P_H s)^T (x - s) > 0, \forall x \in S \text{ distinct from } s.$$

Equality holds for $x = s$, so it is the unique point on this supporting hyperplane of S and thus it is a vertex of the convex hull of S. □

Letting $S_V \subseteq S$ be the vertices of $C(S)$, Lemmas 4.1 and 4.2 lead to the main result:

Theorem 4.1. *FastMap pivot pairs are a subset of the vertices of the convex hull of the data. That is,*

$$a_i, b_i \in S_V, \qquad i = 1, \ldots, k.$$

5 Implications

Convex hull computations in statistics are mostly associated with robust multivariate estimation. Loosely, an estimator of some parameter is said to be robust if it performs well even when the assumed model (implicit or explicit) is not satisfied by the data. For example, when estimating a location parameter, an implicit assumption is that the data are generated by one process that has a location. If more than one process generated the data, a robust estimator would still estimate the location of the dominant process rather than some meaningless location between the processes. The median, for example, is a robust estimator of location while the mean is not. A classic reference on robust estimation is [11].

The concept of *trimming* extremes is often used in reducing dependence on outliers in data [10]. Tukey is attributed with coining the term *peeling* as the multivariate extension of trimming [10], where one peels off the vertices of the convex hull before using the remaining points for estimating a location parameter. This is based on a generalization of the simple practice of removing the maximum and minimum before computing the mean, which dates at least to the early 19th century [10]. Here, with the aim of robustness, the very points on which FastMap depends are discarded! Clearly, FastMap is very sensitive to outliers in the data.

In situations where the data generation system is known to work smoothly, such as machine generated data, outliers may not be of concern. For example, we have recently found that in analyzing climate simulation and astrophysics simulation data, methods that are sensitive to extremes often produce the most compelling results. Here, the extremes are not outliers and may be of most interest. On the other hand, massive data sets are often the result of a long run with several checkpoint restarts where anomalies may occur. For example, in [4], instrument generated Atmospheric Radiation Measurement data [2] contains many instrument restarts that appear as zeros in data with high positive values. Although it is easy to discover these, an automated application of FastMap would be driven by the zero coordinate outliers. Clearly, there are situations where an extremes-sensitive method like FastMap is appropriate or even preferable as well as situations where it will fail.

Outlier sensitivity of FastMap is mentioned in [8] and PCA is presented as more robust. Although PCA is less sensitive to outliers than FastMap, it too is not considered a robust technique. A measure of estimator sensitivity to changes in extreme values of data is the notion of *breakdown point* [3]. Loosely speaking, the breakdown point is the smallest proportion of data that needs to be contaminated to make arbitrarily large changes to the estimator. By this definition, the breakdown point of FastMap is $\frac{1}{n}$, which is asymptotically zero. Principal Components Analysis, the most popular dimension reduction method, also has a breakdown point of $\frac{1}{n}$. In both cases, taking one point arbitrarily far in some direction will rotate the first axis in that direction. Some robust PCA methods begin by computing a robust covariance matrix estimate then proceeding with standard PCA as usual. The classical example of a high breakdown estimator is the median with a .5 breakdown point. That is, half of the data must be moved to make an arbitrarily large change in the median. A multivariate extension of the median is proposed in [12]. This extension uses the notion of half-space support to define the *depth* of a data point so that, ignoring ties, the point with maximal depth is the multivariate median.

The main lesson from robust statistics is that the most distant points are often not the best choice for defining a projection axis. The key to new fast and robust methods is a replacement of the *Choose-distant-object* heuristic by something that considers more than just the maximum distance from a point. One should back-off a little from the maximum, while considering the entire distance distribution. This distribution is already available within the $\mathrm{O}(np)$

complexity. A closer examination, even with more complex algorithms such as clustering, of the distance distribution tail can yield much more robust results, still within the O(np) complexity. In fact, such methods will be more robust than standard PCA. Clearly there are many directions that this methodology can be taken and undoubtedly many such algorithms will be proposed. We provide a simple example in the section that follows.

We would like to note another implication on an algorithm, DFastMap [1], that we recently developed for fast dimension reduction across distributed data sets. Our initial insights that lead to DFastMap produced the main idea for the present paper. Formalizing the convex hull connection to FastMap gives an explanation of why an application of DFastMap to distributed data performs as well as the serial FastMap on a centralized data set. The union of local convex hull vertices necessarily includes all convex hull vertices of the centralized data set. This assertion can be proved using arguments similar to those we used in Section 4. DFastMap centralizes the pivots, arguably a very good subset of the local convex hull vertices (see [1] for more details). This provides a key subset of the combined data convex hull vertices so that little information about extremes is lost when compared to centralizing all the data.

Finally, we also mention an implication on complexity of FastMap and convex hull computations. Because all the FastMap projection axes are computed from points in S_V, the convex hull vertices are sufficient for all distant point searches. Clearly FastMap could be faster if S_V were available. Erickson [5] reports that finding S_V by the "gift-wrapping" algorithm takes O(nf) time, where $f = |S_V|$ is the number of vertices. Since FastMap completes in O(np) time, this is not helpful as $f > p$ for any non-degenerate data sets.

6 A RobustMap algorithm

The FastMap algorithm computes all distances from one object but uses only the maximum, resulting in an outlier-sensitive method. From a statistical viewpoint the distribution of the distances contains information on potential outlier candidates. In essence, we are trimming the extremes of this distance distribution. A complication is that two objects with a similar distance to the reference object can be very far apart in the full p-dimensional space. Selecting a small number of extreme objects and clustering them in the full p-dimensional space, can provide much more information on a robust choice of a distant object. Keeping the selection of a few objects fast and their number small lets us remain within the $O(np)$ time complexity of FastMap.

We provide a simple variant of this idea. Take a constant number, say $r \ll n$, largest distances, cluster the corresponding objects, and choose a central point of the largest cluster as a pivot. This affords protection against a small number, about $r/2$, outliers. Fig. 2 gives the choose-distant-objects heuristic for RobustMap. The parameter r can be some small number that depends on the level of contamination we expect in the data. A second pa-

```
RobustMap: Choose-distant-objects ( $S, d_i(,)$ )
  1. Choose an arbitrary object $s \in S$
  2. Select $r$ largest distances in $d_i(a, s)$
  3. Cluster the $r$ corresponding objects.
  4. Let $a_{i+1}$ be the object nearest the center of the largest cluster.
  5. Similarly, choose $b_{i+1}$ as above, replacing $s$ with $a_{i+1}$.
  6. Report $a_{i+1}$ and $b_{i+1}$ as the distant objects.
```

Figure 2: RobustMap Choose-distant-objects heuristic for iteration i.

rameter controls the number of clusters. Clusters can be considered different at some fixed percentage of the largest distance. Our prototype implementation in R uses single linkage clustering, where a distance of more than 10% of the maximum distance implies a separate cluster.

To test the behavior of RobustMap, we use the Longley data in R and add an observation that blends the origin and the first observation. This is a small data set, but it allows us to move the outlier in and out of the data and quickly explore the behavior of RobustMap, PCA, and FastMap on the contaminated data. To measure the effect of the contamination, we report captured variability within the clean data, while giving the contaminated data to the algorithm. Our reference is PCA on the clean data. Fig. 3 shows typical results and we discuss how the outlier position and non-determinism of RobustMap and FastMap affect the results.

As the outlier moves farther from the data, the FastMap and PCA lines move together but remain well below RobustMap. This is reasonable, as both are highly affected by outliers. The non-determinism of RobustMap and FastMap does not change the order of the methods in Fig. 3 with RobustMap leading and FastMap coming last. Half of a 95% confidence interval around RobustMap would roughly fill the distance between the Reference and RobustMap.

Other, more complex and more robust approaches can consider a multivariate distance distributions from two or more objects. Formal tests may be developed on the basis of distributional assumptions for the objects and derivations of the resulting distance distributions. At the same time, more thorough testing is needed to explore aspects beyond capture of variability. For example, RobustMap projections are different from PCA and can provide alternate data views based on distance distributions.

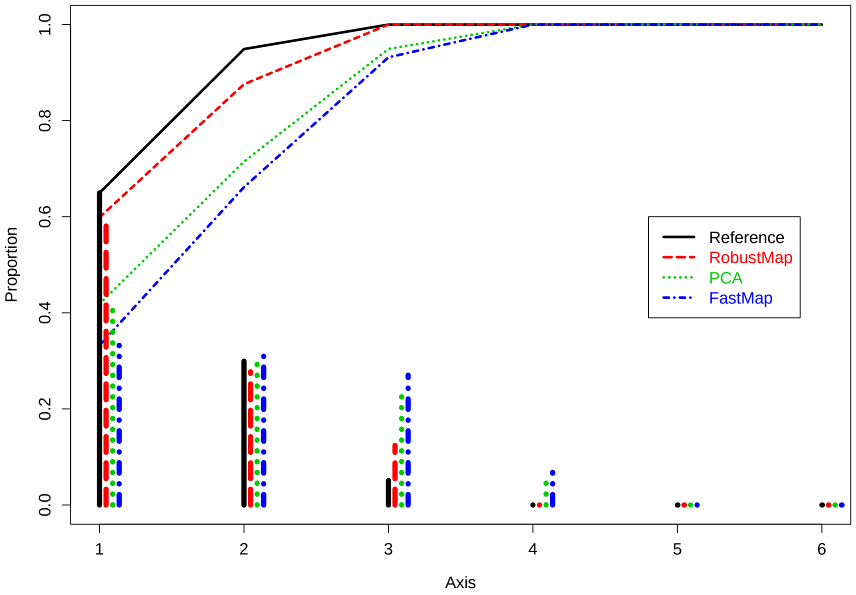


Figure 3: Proportion of clean variability captured by each component axis, when presented with contaminated data. Reference is PCA on clean data only.

We also see some preliminary evidence that these methods are related to pivoting strategies in QR factorization and the recent QLP factorization [13] that provides a fast approximation for the Singular Value Decomposition. Our prototype implementation of RobustMap and FastMap differs from the original [6] by using Householder reflections applied to the rows, somewhat like the QLP factorization. We conjecture that FastMap, RobustMap, and their connection to the convex hull provide a geometric explanation for the success of QLP factorization and may be sources of new pivoting strategies for QR factorization. This is another direction where these methods may provide new insights.

References

[1] Abu-Khzam F.N., Samatova N., Ostrouchov G., Langston M.A., Geist A. (2002). *Distributed dimension reduction algorithms for widely dispersed data.* In Parallel and Distributed Computing and Systems, ACTA Press, 174 – 178.

[2] D. O. E. (1990). *Atmospheric radiation measurement program plan.* Technical Report DOE/ER-0441, U. S. Department of Energy, Office

of Health and Environmental Research, Atmospheric and Climate Research Division, National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161.

[3] Donoho D.L., Huber P.J. (1983). *The notion of breakdown-point.* In Bickel, Doksum, and Hodges, (eds), Festschrift fur Erich L. Lehmann, Belmont, CA, Wadsworth 157 – 184.

[4] Downing D.J., Fedorov V.V., Lawkins W.F., Morris M.D., Ostrouchov G. (2000). *Large data series: Modeling the usual to identify the unusual.* Computational Statistics & Data Analysis **32** 245 – 258.

[5] Erickson J. (1999). *New lower bounds for convex hull problems in odd dimensions.* SIAM J. Comput. **28** (4), 1198 – 1214.

[6] Faloutsos C., Lin K. (1995). *FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets.* In ACM SIGMOD Conference, San Jose, CA, May 1995, 163 – 174.

[7] Gallier J.H. (2000). *Geometric methods and applications for computer science and engineering.* Springer.

[8] Hjaltason G.R., Samet H. (2003). *Properties of embedding methods for similarity searching in metric spaces.* IEEE Transactions on Pattern Analysis and Machine Intelligence **25**, 530 – 549.

[9] Hotelling H. (1933). *Analysis of a complex of statistical variables into principal components.* J. Educ. Psych. **24**, 417 – 441, 498 – 520.

[10] Huber P.J. (1972). *Robust statistics: A review.* Annals Mathematical Statistics **43** (4), 1041 – 1067.

[11] Huber P.J. (1981). *Robust statistics.* John Wiley & Sons, New York.

[12] Ruts I., Rousseeuw P.J. (1996). *Computing depth contours of bivariate point clouds.* Computational Statistics & Data Analysis **23**, 153 – 168.

[13] Stewart G.W. (1999). *The QLP approximation to the singular value decomposition.* SIAM J. Sci. Comput. **20** (4), 1336 – 1348.

[14] Torgerson W.S. (1952). *Multidimensional scaling i: Theory and method.* Psychometrika **17**, 401 – 419.

[15] Ziegler G.M. (1995). *Lectures on polytopes.* Springer-Verlag.

Acknowledgement: Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725.

Address: G. Ostrouchov, N.F. Samatova, Computer Science and Mathematics Division at the Oak Ridge National Laboratory, P.O.Box 2008, Oak Ridge, Tennessee 37831-6367, U.S.A.

E-mail: `ostrouchovg@ornl.gov`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

A GENERAL PARTITION CLUSTER ALGORITHM

Daniel Peña, Julio Rodríguez and George C. Tiao

Key words: Predictive distribution, robust estimation, SAR procedure.

COMPSTAT 2004 section: Clustering.

Abstract: A new cluster algorithm based on the SAR procedure proposed by Peña and Tiao [9] is presented. The method splits the data into more homogeneous groups by putting together observations which have the same sensitivity to the deletion of extreme points in the sample. As the sample is always split by this method the second stage is to check if observations outside each group can be recombined one by one into the groups by using the distance implied by the model. The performance of this algorithm is compared to some well known cluster methods.

1 Introduction

Finding groups in data is a key activity in many scientific fields. Gordon [8] is a good general reference. Classical Partition and Hierarchical algorithms have been very useful in many problems but they have some four main limitations. First, the criteria used are not affine equivariant and therefore the results obtained depend on the changes of scale and/or rotation applied to the data. Second, the usual heterogeneity measures based on the Euclidian metric do not work well for highly correlated observations forming elliptical clusters or when the clusters overlap. Third, we have to specify the number of clusters or decide about the criteria for choosing them. Fourth, there is no general procedure to deal with outliers. Some advances have been made to solve these problems, see [4], [5] and [16].

An alternative approach to cluster is to fit mixture models. This idea has been explored both from the classic and Bayesian point of view. Banfield and Raftery [3] and DasGupta and Raftery [6] have proposed a model-based approach to clustering which finds an initial solution by hierarchical clustering and then assumes a mixture of normals model and uses the EM algorithm to estimate the parameters. A clear advantage of fitting normal mixtures is that the implied distance is the Mahalanobis distance, which is affine equivariant. From the Bayesian point of view the parameters of the mixture are estimated by Markov Chain Monte Carlo methods and several procedures have been proposed to allow for an unknown number of components in the mixture, see [12] and [14]. A promising approach to cluster analysis, that can avoid the curse of dimensionality, is projection pursuit, where low-dimensional projections of the multivariate data are used to provide the most interesting views of the full-dimensional data. Peña and Prieto [11] have proposed an algorithm where the data is projected on the directions of maximum heterogeneity defined as those directions in which the kurtosis coefficient of the projected data is maximized or minimized. Then they used the spacings to search for clusters on the univariate variables obtained by these projections.

Finally, Peña and Tiao [9] propose the SAR (split and recombine) procedure for detecting heterogeneity in a sample with respect to a given model. This procedure is general, affine equivariant, does not require to specify a priori the number of clusters and it is well suited for finding the components in a mixture of models. The idea of the procedure is first to split the sample into more homogeneous groups and second recombine the observations one by one in order to form homogeneous clusters. The SAR procedure has two important properties, that are not shared by many of the most often used cluster algorithms, (i) it does not require an initial starting point, (ii) each homogeneous group is obtained independently from the others, so that each group does not compete with the others to incorporate an observation. The first property implies that the algorithm we propose can be used as a first solution for any other cluster algorithm, the second, that the procedure may work well even if the groups are not well separated. This paper analyzes the application of the SAR procedure to cluster analysis and it is organized as follows. Section 2 presents the main ideas of the procedure. Section 3 compares it in a Monte Carlo study to Mclust (Model Based Cluster, [7], k-means, pam (Partition around medoids, [15] and Kpp (Kurtosis projection pursuit, [11].

2 The SAR procedure

Suppose we define a measure $H(x, X)$ of the heterogeneity between an observation, x, and a set of data, X. We are going to use this measure to split the sample iteratively into homogeneous groups and to recombine observations into the groups. We assume that the heterogeneity measure $H(x, X)$ is equivariant, that is invariant to linear transformations, and is coherent with the assumed model. As the true structure of the data is unknown, we start the process by assuming that the data is homogeneous, and have been generated by a normal distribution, $N_p(\boldsymbol{\mu}, \mathbf{V})$. Then we propose a heterogeneity measure based on out of sample prediction as follows. The predictive distribution for a new observation $\mathbf{x}_f$ generated by a normal model using a Jeffrey's prior $p(\boldsymbol{\mu}, \mathbf{V}) \propto |\mathbf{V}|^{-(p+1)/2}$ is (see for instance, [2] $p(\mathbf{x}_f, \mathbf{X}) \propto \left(1 + \frac{Q_f}{n-p}\right)^{-n/2}$, where $Q_f = \frac{n}{n+1}(\mathbf{x}_f - \bar{\mathbf{x}})'\hat{\mathbf{V}}^{-1}(\mathbf{x}_f - \bar{\mathbf{x}})$ and $\bar{\mathbf{x}}$ is the sample mean and $\hat{\mathbf{V}}$ the sample covariance matrix, given by $\hat{\mathbf{V}} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}})'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}})/(n-p)$. Following Peña and Tiao [9] we will use as measure of heterogeneity of a data $\mathbf{x}_i$ with respect to a group $\mathbf{X}_{(i)}$ which does not contain this observation, the standardized predictive value given by

$$H(\mathbf{x}_i, \mathbf{X}_{(i)}) = -2 \ln \left\{ \frac{p(\mathbf{x}_i|\mathbf{X}_{(i)})}{p(\hat{\mathbf{x}}_{i(i)}|\mathbf{X}_{(i)})} \right\} = (n-1) \ln \left\{ 1 + \frac{Q_{i(i)}}{(n-1)-p} \right\}, \quad (1)$$

where $Q_{i(i)} = \frac{n-1}{n}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})'\hat{\mathbf{V}}_{(i)}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})$, and $\hat{\mathbf{V}}_{(i)}$ and $\bar{\mathbf{x}}_{(i)}$ are the covariance matrix and the mean computed using the sample $\mathbf{X}_{(i)}$ without the case i*th*. Note that $H(\mathbf{x}_i, \mathbf{X}_{(i)})$ is a monotonic function of the Mahalanobis distance $Q_{i(i)}$, which is usually used to check the heterogeneity of a point $\mathbf{x}_i$ with respect to the sample $\mathbf{X}_{(i)}$.

The splitting of the sample is made as follows. For each observation, $\mathbf{x}_i$, we define the discriminator of this point as the observation which, when deleted from the sample, makes the point $\mathbf{x}_i$ as heterogeneous as possible with the rest of the data. The discriminator of $\mathbf{x}_i$ is the point $\mathbf{x}_j$ if

$$\mathbf{x}_j = \arg\max_{x_k} H(\mathbf{x}_i, \mathbf{X}_{(ik)}) = \arg\max_{x_k} (\mathbf{x}_i - \bar{\mathbf{x}}_{(ik)})'\hat{\mathbf{V}}_{(ik)}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(ik)}),$$

where $\mathbf{X}_{(ik)}$ is the sample without the cases i*th* and k*th*.

Each sample point must have a unique discriminator, but several sample points may share the same discriminator. It can be proved (see [10]) that the discriminators are members of the convex hull of the sample. That is, a discriminator must be an extreme point. An intuitive procedure to split the sample into groups is to put together observations with share the same discriminators, as they are affected in the same way to modifications of the sample by deleting some extreme values. It is obvious that if two observations are identical they will have the same discriminator and if they are close they also will have the same discriminator. The number of points in the sample which share the same discriminator is called the order of the discriminator. We consider as special points discriminators of order larger than K, where $K = f(p, n)$ and we will put them in a special group of extreme observations. However, discriminators of order smaller than K are considered as usual points and are assigned to the group defined by all the observations that share a common discriminator. We need to define the minimum size of a set of data to be considered as a group. We will say that we have a group if we could compute the mean and covariance matrix of the group and, therefore, the minimum group size must be $n_0 = p + h$, where $h > 0$, and p is the number of variables. Usually $h = f(p, n)$ and in the examples we have taken $h = \log(n - p)$. In the procedure which follows we have considered as special points to those discriminators of order larger that K, where $K = p + h - 1$. This value seems to work well in the simulations we have made. Based on these considerations the sample is split as follows: 1) Observations which have the same discriminator are put in the same group, the discriminator is only included in the group if it has order smaller than K; 2) Discriminators of order bigger that K are allocated to a specific group of isolated points; 3) if two groups formed by the previous rules have any observation in common the two groups are joined into one group. This three rules split the sample into more homogeneous groups. Each group is now considered as a new sample and the three rules are applied again until splitting further the sample will lead to isolated points because the groups obtained are all of them of size smaller than the minimum group size n_0. A group of data is called basic group if when split will lead to subgroups of size smaller than the minimum size, $p + h$.

When the sample cannot be split further the recombining process is applied starting from any of the basic groups obtained. The recombining process is the one suggested by Peña and Tiao [9]. Each group is enlarged by incorporating observations one by one. For a given group, we begin by testing the observation outside the group which is the closest to the group in terms of the measure $H(y_f, X_g)$, where y_f is the observation outside the group formed by data X_g. If $H(y_f, X_g)$ is smaller than some cut-off value, that is the 99th percentile of the distribution of the statistic $H(y_f, X_g)$, this observation is incorporated into the group and the process of testing the closest observation to the group is repeated for the enlarged group. The enlarging process will continue until either the threshold is crossed or the entire sample is included. A similar idea of recombining points has been used for robust estimation (see for instance, [1]. We may have one of the three possible cases. First, the enlarging of all the basic groups leads to the same group which include all the observations apart from some outliers. Then we have a homogeneous sample with some isolated outliers and the procedure ends. Second, the enlarging of the basic groups leads to a partition of the sample into disjoint groups and we conclude we have some groups in the data and again the procedure ends. Third, we obtain more than a possible solution because the partition obtained is different when starting from different basic groups. Then we have more than one possible solution and the final solutions found are called possible data configurations, PDC. The selection among them is made by a model selection criterion.

3 Monte Carlo results

The properties of the algorithm have been studied in a Monte Carlo experiment, similar to the one used by Peña and Prieto [11] to illustrate the behavior of their cluster procedure. Sets of $10 \times p \times k$ random observations in dimension $p = 2, 4, 8$ have been generated from a mixture of $k = 2, 4$ components of a multivariate distributions. In all data sets the number of observations from each distribution has been determined randomly, but ensuring that each cluster contains a minimum of $p+1$ observations. The mean for each distribution is chosen at random from the multivariate normal distribution $N_p(\mathbf{0}, f\mathbf{I})$. The factor f (see Table 1) is selected to be as small as possible while ensuring that the probability of overlapping between groups is roughly equal to 0.01. We generated data sets in six different scenarios.

a) Mixture of k multivariate normal distributions. In each group the covariance matrix is generated as $\mathbf{S} = \mathbf{UDU}'$, from a random orthogonal matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{D}$ with entries generated from a uniform distribution (a1): $[10^{-3}, 5\sqrt{p}]$, so that the covariance matrices are well conditioned, and (a2): $[10^{-3}, 10\sqrt{p}]$, so that the covariance matrices are ill-conditioned.

b) Mixture of k multivariate uniform distributions with (b1) covariance generated as (a1) and (b2) covariance generated as (a2).

c) Mixture of k multivariate normal distributions generated as indicated in scenario a1), but 10% of the data are outliers (c1): generated by $N_p(\mathbf{0}, f\mathbf{I})$ and (c2): for each cluster in the data, 10% of its observations have been generated as a group of outliers at a distance $4\chi^2_{p,0.99}$ in a group along a random direction, and a single outlier along another random direction.

a1) Covariance matrices well conditioned							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**1.65**	7.33	45.35	16.73	34.98
	4	140	1.29	**0.95**	24.90	1.54	1.86
4	2	14	**4.83**	9.90	47.15	12.38	32.11
	4	20	**5.58**	9.39	27.20	6.75	10.76
8	2	12	15.43	13.13	43.29	**12.28**	55.61
	4	18	7.52	12.58	15.81	**3.75**	14.42
Average			**6.05**	8.88	33.95	8.90	24.96
a2) Covariance matrices ill-conditioned							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**1.58**	9.38	46.38	14.23	33.95
	4	140	1.00	0.61	25.14	**0.60**	1.83
4	2	14	**0.99**	4.96	48.54	11.64	32.89
	4	20	**1.39**	5.07	30.99	6.55	5.38
8	2	12	**0.64**	5.19	44.83	0.66	50.94
	4	18	**0.87**	6.01	22.92	4.36	11.01
Average			**1.08**	5.20	36.47	6.34	22.66

Table 1: Percentages of mislabeled observations for the SAR, the Kpp, the k-means, the Mclust and the `pam` procedures. Normal observations with: (a1) covariance matrices well conditioned, (a2) covariance matrices ill-conditioned. The best method in each case is indicated in boldface.

To provide better understanding of the behavior of the new procedure, in each table we compare the proposed method with Kpp, k-means, Mclust and the `pam` algorithm. The Mclust algorithm has been run with the function 'EMclust' with models EI, VI, EEE, VVV, EEV and VEV and number of cluster between 1 to 8 and the final configuration is selected by the BIC (see [7], for a description of different models used in the function 'EMclust'). The rule to select the number of clusters in the algorithm `pam` is the maximum of the silhouette statistic for $k = 1, \ldots, 8$ and in k-means the stopping rule used is the one proposed by Calinski and Harabasz.

Table 1 gives the average percentage of observations which have been labeled incorrectly in scenarios a1) and a2), obtained from 200 replications for each value in the same data sets in all procedures. In scenario a1) the SAR procedure has the best performance, and Kpp and Mclust are second having a similar behavior. In the scenario a2) when the covariance matrix is ill-conditioned, the SAR procedure is again the best followed by Kpp and Mclust. This result is quite consistent as the SAR procedure is the best in eight out of the twelve comparison included in the two scenarios of Table 1 and in the four cases in which it is not the best it is not far from the best one. The k-means and `pam` show a poor result.

b1) Covariance matrices well conditioned							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**0.45**	11.53	51.40	21.08	44.75
	4	140	**0.58**	0.38	29.25	0.84	1.16
4	2	14	**0.85**	4.81	51.71	12.48	51.41
	4	20	**1.58**	4.33	33.15	9.11	7.68
8	2	12	6.24	**5.45**	41.83	7.38	60.80
	4	18	**2.33**	4.93	20.07	5.58	16.93
Average			**2.00**	5.24	37.90	9.41	30.46
b2) Covariance matrices ill-conditioned							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**1.55**	11.78	48.65	20.53	41.95
	4	140	**0.56**	0.99	34.30	1.75	2.06
4	2	14	**0.79**	4.06	53.23	6.00	46.45
	4	20	**0.38**	3.13	34.39	7.54	7.28
8	2	12	0.34	5.76	45.96	**0.00**	62.13
	4	18	**0.46**	4.21	27.32	4.74	12.61
Average			**0.68**	4.99	40.64	6.76	28.75

Table 2: Percentages of mislabeled observations for the SAR, the Kpp, the k-means, the Mclust and the `pam` procedures. Uniform observations with: (b1) covariance matrices well conditioned, (b2) covariance matrices ill-conditioned.

Table 2 shows the outcome for scenarios b1) and b2) where we analyze the same structure that in scenarios a1) and a2) but now using mixtures of uniform distributions. Table 2 shows the percentages of mislabeled observations for both scenarios b1) and b2). The behavior of the SAR procedure is again the best as an average and the best in ten of the twelve cases. The second best behavior corresponds to Kpp, that is better than Mclust in eleven out of the twelve cases.

c1) Non concentrated contaminations							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	1.25	**0.68**	3.00	6.47	0.69
2	4	140	**0.83**	1.30	12.31	3.50	2.85
4	2	14	8.58	9.46	14.55	**6.71**	7.21
4	4	20	5.66	11.89	22.64	**5.27**	6.13
8	2	12	12.64	14.48	16.88	**12.58**	16.46
8	4	18	9.47	16.67	44.08	6.78	**4.59**
Average			6.40	9.08	18.91	6.89	**6.32**
c2) Concentrated contaminations							
p	k	f	SAR	Kpp	k-means	Mclust	`pam`
2	2	55	**0.98**	4.03	26.25	12.61	17.50
2	4	140	**0.40**	0.65	12.88	0.49	2.04
4	2	14	**3.58**	6.29	35.46	17.90	28.46
4	4	20	**3.21**	10.01	17.69	15.47	7.50
8	2	12	15.03	**13.41**	38.66	23.42	53.08
8	4	18	8.15	13.73	17.72	**6.93**	14.71
Average			**5.22**	8.02	24.78	12.80	20.55

Table 3: Percentages of mislabeled observations for the SAR, the Kpp, the k-means, the Mclust and the `pam` procedures. Normal observations with 10% the outliers: (c1) non concentrated contaminations, (c2) concentrated contaminations.

A final simulation study has been conducted (see Table 3) to determine the behavior of the methods in the presence of outliers. Scenarios c1) and c2) contain 10% of data contaminated by first, a non concentrate contamination, and second, a concentrated contamination defined in scenario c). The criterion to obtain the mislabeled observation is based only in the 90% of observations not contaminated. Table 3 shows the percentage of mislabeled observations for the scenarios c1) and c2). The maximum number of clusters k have been increase to ten in the algorithms k-means, Mclust and `pam` so that the concentrated contamination can be considered as isolated clusters. In the scenario c1) the best methods, as an average, are, with very small difference, the `pam` algorithm and the SAR procedure. However, for concentrated contamination, scenario c2), the SAR procedure is again clearly the best followed by Kpp. As a summary of this Monte Carlo study we may conclude that the SAR procedure has the smallest error classification rate in

22 out of the 36 situations considered and the best average number of mislabeled observations in 5 scenarios out of the six considered. The only scenario in which the SAR is not the best is in scenario c1) but the difference with respect to the best method, `pam`, is very small: misclassification percentage of 6.4% versus 6.32% for pam. The Kpp is the second best in five out of the six scenarios. Ordering the methods for average classification errors in all the scenarios from better to worse, the order would be: SAR, Kpp, Mclust, `pam` and k-means.

References

[1] Atkinson A.C. (1994). *Fast very robust methods for detection of multiple outliers.* Journal of the American Statistical Association **89**, 1329 – 1339.

[2] Box G.E.P., Tiao G.C. (1973). *Bayesian inference in statistical analysis.* Addison-Wesley.

[3] Banfield J.D., Raftery A. (1993). *Model-based Gaussian and non-Gaussian clustering.* Biometrics **49**, 803 – 821.

[4] Cuesta-Albertos, J. A., Gordaliza, A. C., Matrán, C. (1997). *Trimmed k-means: an attempt to robustify quantizers.* The Annals of Statistics **25**, 553 – 576.

[5] Cuevas A., Febrero, M., Fraiman R. (2000). *Estimating the number of clusters.* Canadian Journal of Statistics **28**, 367 – 382.

[6] Dasgupta A., Raftery A.E. (1998). *Detecting features in spatial point processes with clutter via model-based clustering.* Journal of the American Statistical Association **93**, 294 – 302.

[7] Fraley C., Raftery A.E. (1999). *MCLUST: Software for model-based cluster analysis.* Journal of Classification **16**, 297-306.

[8] Gordon A. (1999). *Classification.* 2nd edn. London: Chapman and Hall-CRC.

[9] Peña D., and Tiao G.C. (2003). *The SAR procedure: A diagnostic analysis of heterogeneous data.* (Manuscript submitted for publication).

[10] Peña D., Rodriguez J., Tiao G.C. (2004). *Cluster analysis by the SAR procedure* (Manuscript submitted for publication).

[11] Peña, D. and Prieto, J. (2001). *Cluster identification using projections.* Journal of the American Statistical Association **96**, 1433 – 1445.

[12] Richarson S., Green P.J. (1997). *On Bayesian analysis of mixtures with an unknown number of components.* Journal of the Royal Statistical Society B **59**, 731 – 758.

[13] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* New York: John Wiley.

[14] Stephens M. (2000). *Bayesian analysis of mixture models with an unknown number of components–an alternative to reversible jump methods.* The Annals of Statistics **28**, 40 – 74.

[15] Stuyf A., Hubert M., Rousseeuw P.J. (1997). *Integrating robust clustering techniques in S-PLUS.* Computational Statistics and Data Analysis **26**, 17 − 37.

[16] Tibshirani R., Walther G., Hastie T. (2001). *Estimating the number of clusters in a data set via the gap statistic.* Journal of the Royal Statistical Society B **63**, 411 − 423.

Address: D. Peña, Departamento de Estadística, Universidad Carlos III de Madrid, Spain

J. Rodríguez, Laboratorio de Estadística, Universidad Politécnica de Madrid, Spain

G.C. Tiao, Graduate School of Business, University of Chicago, USA

E-mail: `dpena@est-econ.uc3m.es`

ITERATIVE DENOISING FOR CROSS-CORPUS DISCOVERY

Carey E. Priebe, David J. Marchette, Youngser Park, Edward J. Wegman, Jeffrey L. Solka, Diego A. Socolinsky, Damianos Karakos, Ken W. Church, Roland Guglielmi, Ronald R. Coifman, Dekang Lin, Dennis M. Healy, Marc Q. Jacobs, Anna Tsao

Key words: Text document processing, statistical pattern recognition, dimensionality reduction.

COMPSTAT 2004 section: Dimensional reduction, Classification.

Abstract: We consider the problem of statistical pattern recognition in a heterogeneous, high-dimensional setting. In particular, we consider the search for meaningful cross-category associations in a heterogeneous text document corpus. Our approach involves "iterative denoising" — that is, iteratively extracting (corpus-dependent) features and partitioning the document collection into sub-corpora. We present an anecdote wherein this methodology discovers a meaningful cross-category association in a heterogeneous collection of scientific documents.

1 Introduction

The "integrated sensing and processing decision trees" introduced in [9] proceed according to the following philosophy. Assume that there is a heterogeneous collection of entities $\mathcal{X} = x_1, \cdots, x_n$ which can, in principle, be measured (sensed) in a large number of ways. Because the sensor cannot make all measurements simultaneously — either due to physical sensor constraints or because of the high intrinsic dimension of the complete feature collection — only a subset of the possible measurements is to be made at any one time.

Thus, for the entire entity collection $\mathcal{X}$ a first set of measurements is made. Based on the features obtained, $\mathcal{X}$ is partitioned into $\{\mathcal{X}_1, \cdots, \mathcal{X}_{J_1}\}$, each $\mathcal{X}_{j_1}$ being (presumably) more homogeneous than the original entity collection $\mathcal{X}$. Then, for each partition cell $\mathcal{X}_{j_1}$ a new set of measurements is considered. This process continues, generating branches consisting of "iteratively denoised" entity collections $\{\mathcal{X}_{j_1 1}, \cdots, \mathcal{X}_{j_1 J_2}\}$, $\{\mathcal{X}_{j_1 j_2 1}, \cdots, \mathcal{X}_{j_1 j_2 J_3}\}$, and so forth, until a collection (say, $\mathcal{X}_{j_1 j_2 j_3}$) is deemed sufficiently coherent for inference to proceed. Such collections are the leaves of the tree.

2 Iterative denoising for cross-corpus discovery

The example application we consider herein is that of discovering meaningful associations in a heterogeneous text document corpus. See, for example, [1] for a survey of text mining.

2.1 Feature extraction & dimensionality reduction

Let C be a collection of text documents. The corpus-dependent feature extraction of Lin & Pantel [6], [8] can be described as

$$\mathcal{L}_C(\cdot) : \mathcal{D}\text{ocument}\mathcal{S}\text{pace} \to [\mathcal{M}\text{utual}\mathcal{I}\text{nformation}\mathcal{F}\text{eature}]^{d_{\mathcal{L}}(C)}.$$

Both the features themselves and the number of features $d_{\mathcal{L}}(C)$ depend on the corpus C. Thus $\mathcal{L}_C(C)$ is a $|C| \times d_{\mathcal{L}}(C)$ *mutual information feature matrix*. Each of the features is associated with a word (after stemming and removal of stopper words), as follows. For document x in corpus C, and associated word w, the mutual information between x and w is given by

$$m_{x,w} = \log\left(\frac{f_{x,w}}{\sum_{\xi} f_{\xi,w} \sum_{\omega} f_{x,\omega}}\right).$$

Here $f_{x,w} = c_{x,w}/N$ where $c_{x,w}$ is the number of times word w appears in document x and N is the total number of words in the corpus C. This information is discounted to reduce the impact of infrequent words via

$$\tilde{m}_{x,w} = m_{x,w} \cdot \frac{c_{x,w}}{1 + c_{x,w}} \cdot \frac{\min(\sum_{\xi} c_{\xi,w}, \sum_{\omega} c_{x,\omega})}{1 + \min(\sum_{\xi} c_{\xi,w}, \sum_{\omega} c_{x,\omega})}.$$

The *mutual information feature vector*, then, for document x in corpus C, is given by

$$e_x = \mathcal{L}_C(x) = [\tilde{m}_{x,w_1}, \cdots, \tilde{m}_{x,w_{d_{\mathcal{L}}(C)}}].$$

Given two documents $x, y \in C$, the distance (we use the term loosely; it is in fact a pseudo-dissimilarity) employed, ρ, is given by

$$\rho(x, y) = 1 - (e_x \cdot e_y)/(||e_x||_2||e_y||_2) \in [0, 2].$$

Thus

$$\rho \circ \mathcal{L}_C(C)$$

is a $|C| \times |C|$ *interpoint distance matrix*. All subsequent processing will be based on these interpoint distances, as discussed in [7]. However, the features, and hence the interpoint distances themselves, are *corpus dependent* and so, as the iterative denoising tree is built, based on the evolving partitioning, these distances change.

Multidimensional scaling [2] is used to embed the interpoint distance matrix $\rho \circ \mathcal{L}_C(C)$ into a Euclidean space $\mathbb{R}^{d_{mds}(C)}$. Notice first that, if the feature vectors were Euclidean — that is, if we were using an actual distance in the $d_{\mathcal{L}}(C)$-dimensional space — then the features could be represented *with no distortion* in $\mathbb{R}^{d_{\mathcal{L}}(C)-1}$. Alas, they are not, and cannot be. So

$$mds \circ \rho \circ \mathcal{L}_C(C)$$

is a $|C| \times d_{mds}(C)$ *Euclidean feature matrix* representing the corpus C. The choice of $d_{mds}(C)$ represents a distortion/dimensionality tradeoff.

Finally, the Euclidean representation $mds \circ \rho \circ \mathcal{L}_C(C)$ produced by multidimensional scaling is reduced, via principal component analysis [5], to a lower dimensional space for subsequent processing. Again we face a model selection choice of dimensionality. The combination feature extraction/dimensionality reduction we propose, then, is given by

$$pca \circ mds \circ \rho \circ \mathcal{L}_C(C),$$

yielding a $|C| \times d_{pca}(C)$ *LSI feature matrix* which can be seen as akin to a (generalized) latent semantic indexing (LSI) [4].

2.2 Science news corpus

A heterogeneous corpus of text documents obtained from the Science News web site is used in this example. The Science News (SN) corpus C consists of $|C| = 1047$ documents in eight classes. Table 1 provides a breakdown of the corpus by number of documents per class. Our goal is two find two documents in different classes which have a meaningful association.

Class	**Number of Documents**
Anthropology	54
Astronomy	121
Behavioral Sciences	72
Earth Sciences	137
Life Sciences	205
Math & CS	60
Medicine	280
Physics	118

Table 1: Science News corpus.

For this Science News corpus C, feature extraction via $\mathcal{L}_C(C)$ yields a feature dimension $d_{\mathcal{L}}(C) = 10906$. That is, there are 10906 distinct meaningful words in the corpus, and the Lin & Pantel feature extraction produces a 1047×10906 feature matrix.

Multidimensional scaling (Figure 1, left panel) on the 1047×1047 interpoint distance matrix $\rho \circ \mathcal{L}_C(C)$ yields $d_{mds}(C) = 898$. (Numerical issues in the multidimensional scaling algorithm make 898 the largest dimension into which the interpoint distance matrix can be embedded. So, while Figure 1

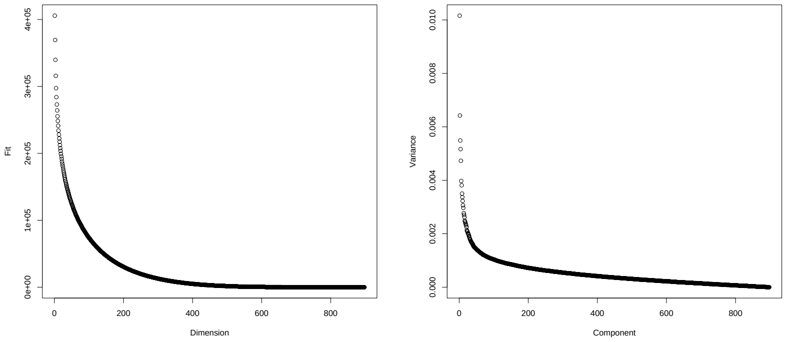


Figure 1: Multidimensional scaling (left panel) for the original 1047 10906-dimensional SN feature vectors. The largest numerically stable multidimensional scaling embedding is $d_{mds}(C) = 898$. (This left curve suggests that perhaps 200, and certainly 400 dimensions is sufficient to adequately fit the documents into Euclidean space.) Principal components (right panel) for the 898-dimensional Euclidean embedding of the original 1047 10906-dimensional SN feature vectors. (The "elbow" of this scree plot occurs, perhaps, in the range of 10-50 principal components.)

suggests that perhaps 200, and certainly 400 dimensions is sufficient to adequately fit the documents into Euclidean space, we avoid the first model selection quandary by choosing the largest numerically stable multidimensional scaling embedding.)

A subsequent principal component analysis of the 898-dimensional Euclidean features $mds \circ \rho \circ \mathcal{L}_C(C)$ yields the scree plot presented in Figure 1, right panel. This scree plot suggests that a latent semantic index dimension of perhaps 10-50 is appropriate for the SN corpus.

Figure 2 displays the projection of the data set onto the first two principal components of

$$pca \circ mds \circ \rho \circ \mathcal{L}_C(C) \tag{1}$$

for the Science News corpus. Notice that this plot suggests that the combination feature extraction/dimensionality reduction we have employed (eq. 1) has captured well some of the information concerning the eight classes, despite the fact that we are viewing just two dimensions (as opposed to, say, the 10-50 dimensions suggested by the scree plot in Figure 1). To wit: there are two groups extending from and distinguishable from the main body of documents. These two groups are dominated by medicine (the upper left arm) and astronomy (the upper right arm). Additionally, some physics documents are present in the astronomy arm and some life sciences and behavioral sciences documents are present in the medicine arm. That physics should have some similarity with astronomy, and that life sciences and behavioral sciences should have some similarity with medicine, agrees with intuition.

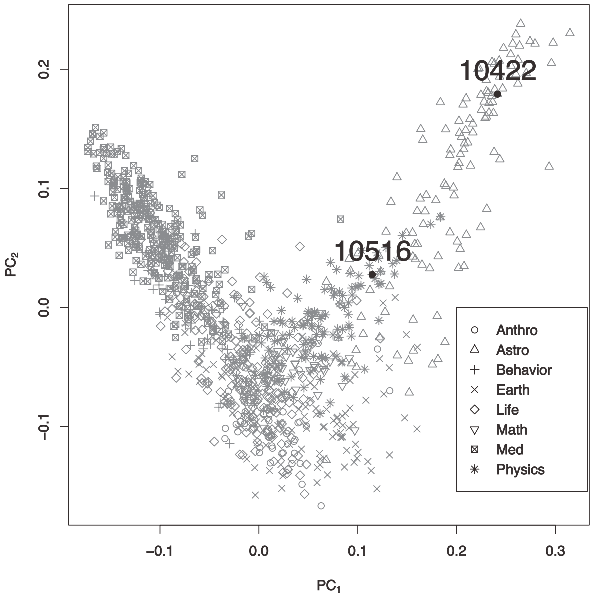


Figure 2: The first two principal components of $pca \circ mds \circ \rho \circ \mathcal{L}_C(C)$ for the Science News corpus. The eight symbols represent the eight classes; the three clusters generated via hierarchical clustering correspond roughly to the main body and the two arms. Notice that there are two groups extending from and distinguishable from the main body of documents. These two groups are dominated by medicine (the upper left arm) and astronomy (the upper right arm). The documents selected as our anecdotal "meaningful association" are indicated throughout by the solid dots and document number.

2.3 Example result

Recall that the SN corpus C has $|C| = 1047$ with class label vector

$$v = [54, 121, 72, 137, 205, 60, 280, 118].$$

The iterative denoising tree for cross-corpus discovery is illustrated on the SN corpus in Figure 3. This figure provides a coarse depiction of one path, from root to leaf, of the tree; a row-by-row description thereof follows.

Row 1: At the root, we have

$$pca \circ mds \circ \rho \circ \mathcal{L}_C(C).$$

Recall that these 1047 documents yield a feature dimension $d_{\mathcal{L}}(C) = 10906$ and an mds dimension $d_{mds}(C) = 898$. We display the first two principal components; thus the root (row 1) in Figure 3 is presented in detail in Figure 2.

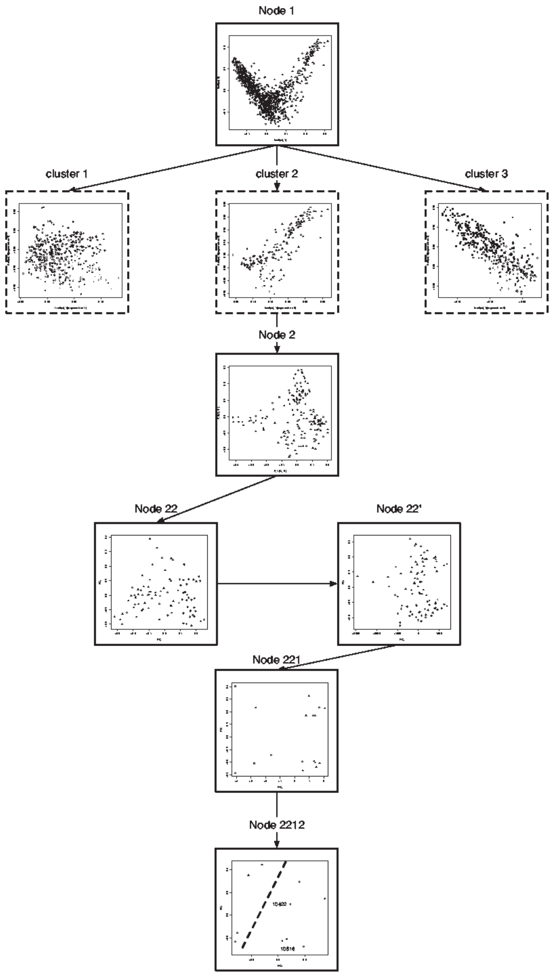


Figure 3: One path in an iterative denoising tree for the SN corpus.

Row 2: In the same space as for Row 1, we have simply split out three clusters obtained via hierarchical clustering, for display convenience.

(We choose in this manuscript to avoid model selection details; e.g., the choice of three vs. two clusters at the root. In general, we recommend that this issue be avoided by generating a *binary* tree unless user intervention is possible. In this example, the root begs for three clusters — a core and two arms.)

To illustrate an anecdotal meaningful cross-corpus discovery, we will follow cluster 2, C_2, which contains 166 documents. This subset of the original corpus is *denoised* in the sense that it is primarily physics and astronomy. The class label vector is

$$v_2 = [2, 113, 0, 10, 4, 0, 1, 36].$$

Thus, C_2 contains nearly all (113 of 121) of the astronomy documents, nearly one third (36 of 118) of the physics documents, and but a smattering from the other classes. So while the original feature extraction was done in the context of a corpus containing medicine, behavioral sciences, and mathematics documents, these topics are not a part of the context for the feature extraction for C_2 and this feature extraction can therefore focus on features germane to physics and astronomy.

Row 3: Here we display

$$pca \circ mds \circ \rho \circ \mathcal{L}_{C_2}(C_2).$$

(See Figure 4 for more detail.) These 166 documents yield a feature dimension $d_{\mathcal{L}}(C_2) = 3037$ and an mds dimension $d_{mds}(C_2) = 162$. Since $\mathcal{L}$ involves *corpus-dependent* feature extraction, this display is different than the "cluster 2" display in Row 2. This difference is due to denoising. The indicated partition represents the clusters generated via hierarchical clustering. Notice that one of the clusters (C_{22}, lower right, containing 91 documents) contains approximately half of C_2's astronomy documents (52 of 113) and nearly all of C_2's physics documents (35 of 36). In continuing pursuit of our anecdotal meaningful cross-corpus discovery, we follow C_{22}.

Row 4: The class label vector for C_{22} is

$$v_{22} = [0, 52, 0, 1, 2, 0, 1, 35].$$

The left display in Row 4 (see Figure 5 for more detail) depicts

$$pca \circ mds \circ \rho \circ \mathcal{L}_{C_{22}}(C_{22}).$$

These 91 documents yield a feature dimension $d_{\mathcal{L}}(C_{22}) = 1981$ and an mds dimension $d_{mds}(C_{22}) = 89$. Again, recall that the feature extraction is corpus-dependent. Now consider altering the geometry via the document subset

$$S_{22} = \{10500, 10651\} \subset C_{22}.$$

(These documents were chosen arbitrarily, for the purposes of illustration: they consist of a Physics document about neutrinos and an Astronomy document about black holes.) In the display, the two black squares represent S_{22}.

The right display in Row 4 (see Figure 6 for more detail) depicts the altered geometry after consideration of S_{22}. That is, here we have added

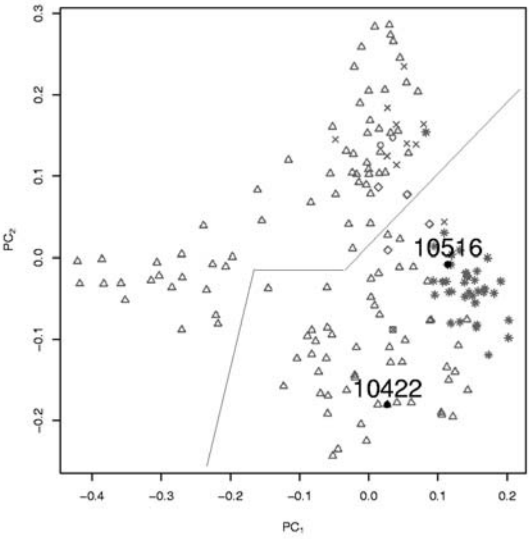


Figure 4: Node N_2 in the iterative denoising tree for the SN corpus.

a new (90th) feature $K_c d(\cdot, S_{22})$ to the 89 multidimensional scaling features, and are displaying

$$pca \circ [(mds \circ \rho \circ \mathcal{L}_{C_{22}}(C_{22}))\,; K_c d(\cdot, S_{22})]\,.$$

In the display, the two black squares again represent S_{22}. The distance-to-subset used for the additional "tunnelling" feature (see, for instance, [3]) $d(\cdot, S_{22})$, is the minimum Euclidean distance to an element of the subset in the LSI-space defined by the selected principal components; in this case, the scree plot suggests $d_{pca}(C'_{22}) = 20$. The coefficient K_c used for the tunnelling feature is obtained by scaling the values $d(\cdot, S_{22})$ so that the variance for the tunnelling feature $K_c d(\cdot, S_{22})$ is some pre-specified positive multiple c of the maximum multidimensional scaling feature variance. We use $c = 10000$ in this example so that this new feature dominates the multidimensional scaling features in the subsequent principal component analysis. (Note that the scale presented in N'_{22} in Figure 6 is such that the ordinate has no impact on the subsequent clustering; the abscissa dominates.) Rather than use the automatic clustering (depicted), we illustrate user intervention via manual clustering based on a vertical line (recall that the abscissa dominates) at 700 in N'_{22}. We follow the rightmost cluster obtained thusly, C_{221}.

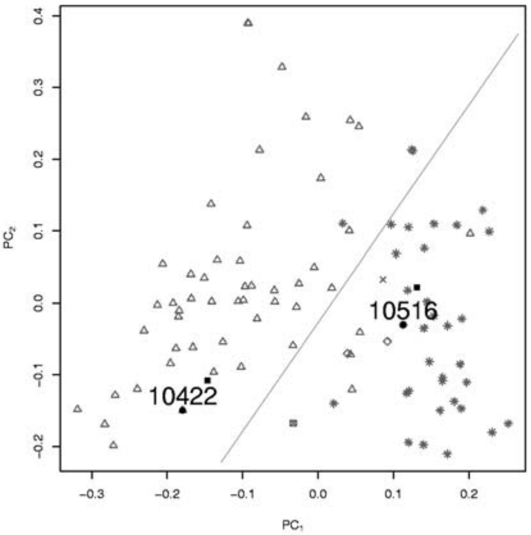


Figure 5: Node N_{22} in the iterative denoising tree for the SN corpus.

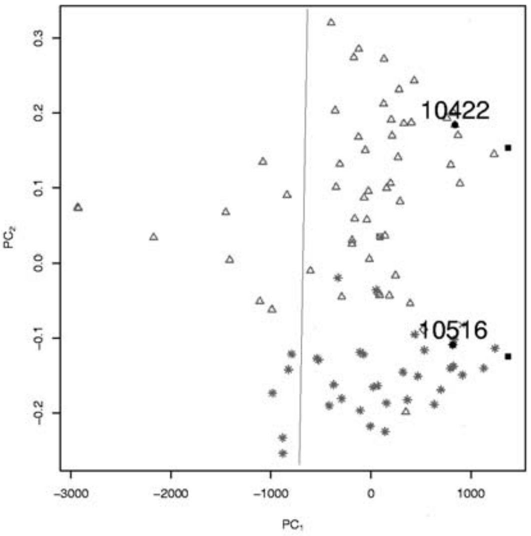


Figure 6: Node N'_{22} in the iterative denoising tree for the SN corpus.

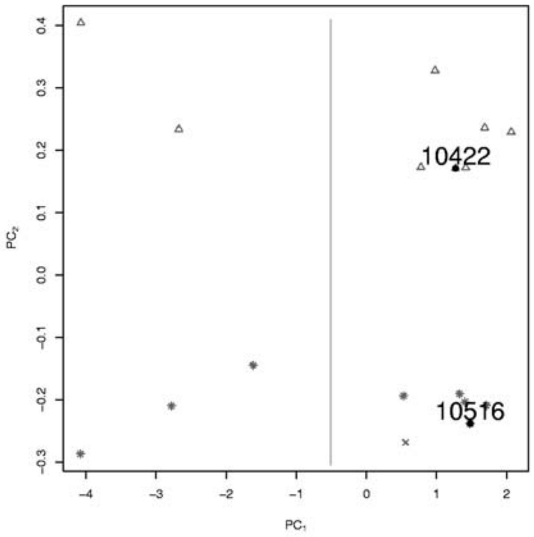


Figure 7: Node N_{221} in the iterative denoising tree for the SN corpus.

Row 5: The document collection C_{221} is, again, almost entirely astronomy and physics, with

$$|C_{221}| = 17$$

and

$$v_{221} = [0, 8, 0, 1, 0, 0, 0, 8].$$

These 17 documents yield a feature dimension $d_{\mathcal{L}}(C_{221}) = 367$ and an mds dimension $d_{mds} = 16$. After recalculating the features for C_{221}, we display

$$pca \circ [(mds \circ \rho \circ \mathcal{L}_{C_{221}}(C_{221}))\,; K_{c'}d(\cdot, S_{22})]\,.$$

(See Figure 7 for more detail.) (A value of $c' = 100$ is used here; the impact of the tunnelling feature is lessened.)

Row 6: Here we consider one of the two clusters, C_{2212}, from N_{221} via

$$pca \circ [(mds \circ \rho \circ \mathcal{L}_{C_{2212}}(C_{2212}))\,; K_{c'}d(\cdot, S_{22})]\,.$$

$$|C_{2212}| = 12$$

and

$$v_{2212} = [0, 6, 0, 1, 0, 0, 0, 5].$$

These 12 documents yield a feature dimension $d_{\mathcal{L}}(C_{2212}) = 215$ and an mds dimension $d_{mds} = 11$. This, in turn, clusters into C_{22121} and C_{22122}.

Let us finally consider C_{22121}. This leaf contains eight documents, with class label vector

$$v_{22121} = [0, 4, 0, 0, 0, 0, 0, 4].$$

Pairs of documents from different classes which fall to the same leaf of the iterative denoising tree are candidate associations. Thus this example yields 16 candidate associations, at least one of which (astronomy #10422 = "X-Ray Universe: Quasar's jet goes the distance" by R. Cowen, Science News Online, Feb. 16, 2002 & physics #10516 = "Glimpses inside a tiny, flashing bubble" by I. Peterson, Science News Online, Oct. 5, 1996) is plausibly a *meaningful* association.

3 Conclusion

We have presented an anecdote — not an experiment! — suggesting that an iterative denoising methodology can be a useful tool in discovering meaningful cross-corpus associations. Corpus-dependent feature extraction is an essential part of the methodology, providing features which are iteratively fine-tuned to ever more homogeneous subsets of documents as one progresses down the tree. The specific approaches to feature extraction, dimensionality reduction, and partitioning may be profitably altered within the framework of the general methodology. The adaptive geometry provided by employing distance-to-subset "tunnelling" features allows the user to alter the details of tree growth. Experimental design to allow for statistical evaluation of the performance of the methodology provides some interesting hurdles, and will be reported elsewhere.

Finally, we note that the methodology described is not specific to text document processing, and may have application in many disparate discovery scenarios. The fundamental idea, as in [9], is to address the problem of there being more measurements that can be made than should be made at any one time.

References

[1] Berry M.W., editor (2004). *Survey of text mining: clustering, classification, and retrieval.* Springer-Verlag.

[2] Borg I., Groenen P. (1997). *Modern multidimensional scaling: theory and applications.* Springer-Verlag.

[3] Cowen L.J., Priebe C.E. (1997). *Randomized nonlinear projections uncover high–dimensional structure.* Advances in Applied Mathematics **9**, 319 − 331.

[4] Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. (1990). *Indexing by latent semantic analysis.* Journal of the American Society for Information Science **41** (6), 391 − 407.

[5] Jolliffe I.T. (1986). *Principal component analysis.* Springer-Verlag.

[6] Lin D., Pantel P. (2002). *Concept discovery from text.* In Proceedings of Conference on Computational Linguistics 2002, Taipei, Taiwan, 577 – 583.

[7] Maa J.-F., Pearl D.K., Bartoszynsky R. (1996). *Reducing multidimensional two-sample data to one-dimensional interpoint comparisons.* The Annals of Statistics **24**, 1069 – 1074.

[8] Pantel P., Lin D. (2002). *Discovering word senses from text.* In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002, Edmonton, Canada, 613 – 619.

[9] Priebe C.E., Marchette D.J., Healy D.M. (2004). *Integrated sensing and processing decision trees.* IEEE Trans. PAMI, to appear.

Acknowledgement: Sponsored by the Defense Advanced Research Projects Agency under "Novel Mathematical and Computational Approaches to Exploitation of Massive, Non-physical Data", ARPA Order No. P246, Program Code 3E20. Issued by DARPA/CMO under Contract No. MDA972-03-C-0014 to AlgoTek, Inc. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either explicitly or implied, of DARPA or the U.S. Government. Approved for Public Release, Distribution Unlimited.

Address: C.E. Priebe, E.J. Wegman, D.A. Socolinsky, K.W. Church, R. Guglielmi, R.R. Coifman, D. Lin, M.Q. Jacobs, A. Tsao, AlgoTek, Inc., 3811 N. Fairfax Dr., Suite 700
D.J. Marchette, J.L. Solka, NSWCDD B10, Dahlgren, VA
Y. Park, D. Karakos, Johns Hopkins U., Balt., MD
D.M. Healy, DARPA, Arlington, VA 22203

E-mail: cep@jhu.edu

COMPSTAT'2004 Symposium

FROM DATA TO DIFFERENTIAL EQUATIONS

Jim O. Ramsay

Key words: Functional data analysis, differential equation, smoothing, nonparametric regression, lupus.

COMPSTAT 2004 section: Functional data analysis, Spatial statistics.

Abstract: Differential equations are the natural way to model systems with functional inputs and functional outputs. They allow us to study the system's dynamics in the sense of explicitly modelling how the output changes in response to sudden changes in input. For example, engineers developing control systems for industrial processes routinely use DIFE's as modelling tools.

A new method is described for going directly from noisy discrete data, not necessarily sampled at equally spaced times, to a system of differential equations of arbitrary orders, linear or nonlinear, that describes the data. The method involves a generalization of nonparametric curve estimation in which the penalty functional rather than the smoothing functions is estimated.

Examples are drawn from chemical engineering and medicine.

1 Introduction: Three main themes

Differential equations (here shortened to DIFE's) make explicit the relation between one or more derivatives and the function itself. For example, the general first order equation for function $x(t)$,

$$Dx = f(t, x),$$

defines a dependency of the first derivative Dx on the function x as well as, possibly, other direct dependencies on argument t.

The talk for which this paper is a summary aims to make three general points:

- DIFE's are powerful tools for modeling data. Indeed, they are already routinely used in the chemical, physical and biological sciences as well as in engineering. They are important primarily because they model the *dynamics* of an observed process; that is, rates of change are modeled along the observed function. This is especially important in input/output systems where how the system responds to an abrupt change in input can be as important as the long-term change that results.
- We have new methods for fitting differential equations or dynamic models to raw noisy data that appear to be substantially more effective than

existing techniques. These methods are based on developments in *functional data analysis* [1], [2], a collection of methods for the analysis of curves and images as data.

- Some important applications are outlined to show the potential for these developments in chemical engineering and medicine. In chemical engineering, there are many possibilities for the use of these techniques in process control applications. In medicine, the methods are applied to some data on lupus, where the dynamics of the disease are the central issue for developing effective treatments.

2 Why consider differential equation models?

The behavior of a derivative is often of more interest than the function itself. The classic example is mechanics, where Newton's second law for position $x(t)$ as a function of mass m, $x(t) = mD^2x(t)$, as well as it's descendent, $e = mc^2$, shows that energy exchange takes place at the level of acceleration, not position. How rapidly a system responds rather than its final level of response is often what matters.

Since a DIFE links the behavior of a derivative to the behavior of the function, it implies that derivatives will exhibit the same smoothness and regularity that characterizes the function. Consequently, a DIFE is can be an important method for computing stable estimates of derivatives.

Natural scientists often deliver theory to biologists and engineers in the form of DIFE's. Moreover, many other fields, such as pharmacokinetics and industrial process control, routinely use DIFE's as models for real-life systems. DIFE's are especially important when feedback systems must be developed to control the behavior of systems.

Although linear differential equations are much easier to work with then nonlinear systems, nonlinear DIFE's are often compact and elegant models for systems exhibiting exceedingly complex behavior, especially in the biosciences. Indeed, chaotic systems and systems exhibiting catastrophic changes are usually modelled with nonlinear dynamics.

It is the business of statistics to model random variation. We usually model random behavior in functions by assuming a fixed underlying process with superimposed noisy variation. But a DIFE allows a much richer range of ways in which stochastic behavior can be introduced:

- random coefficient functions
- random forcing functions
- random initial, boundary and other constraints
- system time t unfolding at a random rate

3 A simple input/output system

We begin by looking at a first order linear DIFE for a single output function $x(t)$ and a single input function $u(t)$, although our ultimate goal is to link multiple outputs to multiple inputs.

Figure 1 is an example: The fluid level in a tray within a distillation column of an oil refinery is shown as a function of the flow of a fluid into the tray. We must explain two things: By how much does the fluid level ultimate change in response to the change in input flow indicated, and how rapidly does this change take place?

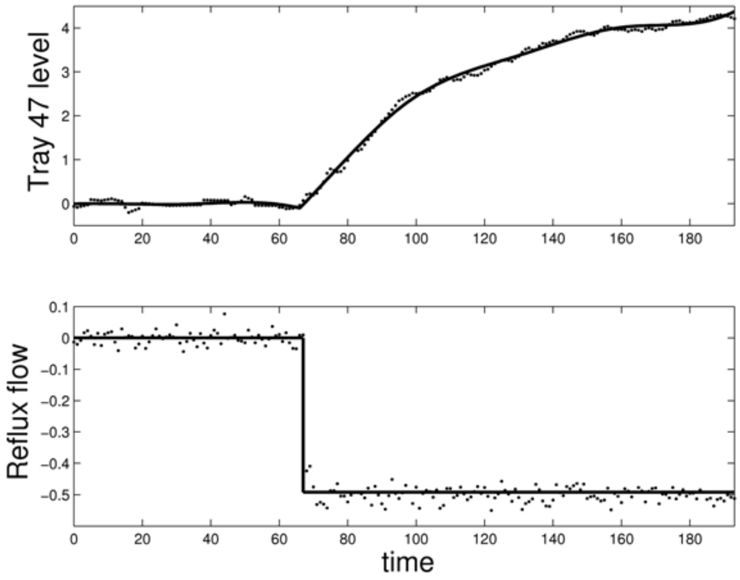


Figure 1: The upper panel shows the level of material in a tray of a distillation column in an oil refinery, and the lower level shows the flow of material being distilled into the tray. The points are measured values, and the solid lines are smooths of the data using regression splines.

The DIFE has the general form

$$Dx(t) = -\beta(t)x(t) + \alpha(t)u(t) \tag{1}$$

The *homogeneous* part of the equation

$$Dx(t) = -\beta(t)x(t)$$

describes the *endogenous* or internal dynamics of the system. the *forcing function u* is an exogenous functional independent variable that perturbs these internal dynamics. The functions α and β are the coefficient functions that define the DIFE. The system is linear in these coefficient functions, and also in the input and output functions.

One way to understand the separate roles of α and β is to study a simpler *constant coefficient* model with an input that steps from 0 to 1 at time 1 and for which $x(0) = 1$. The solution to the equation in this case is

$$\begin{aligned} x(t) &= e^{-\beta t}, 0 \le t \le 1, \\ &= e^{-\beta t}[1 - (\alpha/\beta)e^{-\beta(t-1)}], 1 \le t. \end{aligned}$$

We see that β controls the rate of change and that the ultimate level or *gain* is α/β. We can compare α to the volume control on a radio playing a song carried by radio signal u; the bigger α, the louder the sound. The bass/treble control, on the other hand, corresponds to β; the larger β, the higher the frequency of what we hear.

4 Fitting a differential equation to data

The basic idea is to use *profiled least squares* to estimate unknown parameters, such as α and β in the above example. We do this by replacing the smoothing function x used to smooth a sequence of noisy functional observations $y_j, i = 1, \ldots, n$ by the equations defining the fit to the data conditional on a roughness penalty defined by a differential operator L. Then we optimize the fit with respect to only the unknown parameters; the fitted values $x(t_j)$ are computed as byproduct of the process, but do not themselves require additional parameters.

Focussing for simplicity on the first order linear equation (1), and defining the coefficient functions to be constants, we define the linear differential operator $L_{\alpha,\beta}$ to be

$$L_{\alpha,\beta}x(t) = -\beta x(t) + Dx(t) - \alpha u(t)$$

Function x is a solution to (1) if and only if $L_{\alpha,\beta}x = 0$.

Now define the penalized least squares fitting criterion as

$$\texttt{PENSSE}(y|\lambda, \alpha, \beta) = \sum_j [y_j - x(t_j)]^2 + \lambda \int [L_{\alpha,\beta}x]^2. \tag{2}$$

If x has the basis function expansion $x(t) = \mathbf{c}'\boldsymbol{\phi}(t)$, where $\boldsymbol{\phi}(t)$ is a functional vector of basis functions of length K, then criterion (2) is minimized with respect to coefficient vector $\mathbf{c}$ by

$$\mathbf{c}(\alpha, \beta) = [\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda\mathbf{R}(\alpha, \beta)]^{-1}[\boldsymbol{\Phi}'\mathbf{y} + \lambda\mathbf{s}(\alpha, \beta)] \tag{3}$$

where

- $\boldsymbol{\Phi}$ is the n by K matrix of basis function values $\phi_k(t_j)$

- λ is a smoothing parameter
- $\mathbf{y}$ is the vector of noisy observations to be smoothed
- penalty matrix $\mathbf{R}(\alpha, \beta)$ is

$$\mathbf{R}(\alpha, \beta) = \int (L_{\alpha,\beta}\boldsymbol{\phi})(L_{\alpha,\beta}\boldsymbol{\phi})'$$

- penalty vector $\mathbf{s}(\alpha, \beta)$ is

$$\mathbf{s}(\alpha, \beta) = \int (L_{\alpha,\beta}\boldsymbol{\phi})u$$

Substituting (3) into (2), we may now minimize the un–penalized profiled error sum of squares

$$\texttt{PROFSSE}(y|\lambda, \alpha, \beta) = \sum_j [y_j - x(t_j|\alpha, \beta)]^2 \tag{4}$$

with respect to parameters α and β. Our experience is that the smoothing parameter λ can usually be selected by minimizing the generalized cross-validation (GCV) criterion.

This process may be extended to equations of an arbitrary order, nonlinear equations, and systems of equations.

5 Two simulated data examples

5.1 Twenty tilted sinusoids

How well can we recover derivatives using this process? Consider a tilted sinusoid

$$x_i(t) = c_{i1} + c_{i2}t + c_{i3}\sin(6\pi t) + c_{i4}\cos(6\pi t)$$

that is annihilated by the operator

$$Lx_i = (6\pi)^2 D^2 x + D^4 x$$

We generated $N = 20$ of these by randomly generating coefficients from $N(0, 1)$ and adding noise to x_i from the same distribution.

As a point of comparison, when we smoothed with $L = D^4$, best results were obtained with $\lambda = 10^{-10}$ and the integrated root-mean-squared errors for the function and the first two derivatives were 0.32, 9.3and 315.6, respectively.

When we estimated all four constant coefficients for the order four linear differential operator L, best results were obtained for $\lambda = 10^{-5}$ and the integrated root-mean-squared errors for the function and the first two derivatives were 0.18, 2.8 and 49.3, respectively. These represent improvements in precision of estimation by factors of 1.8, 3.3 and 6.4, respectively. We estimated β_2 to be 353.6, whereas the right value was 355.3.

The most dramatic improvement in derivative estimation occurred at the boundaries. Estimating the linear differential operator virtually eliminated the usual instability of derivative estimates in these regions because these estimates are linked by the DIFE to the behavior of the function values, which are only mildly more unstable at the boundaries than within the interior. But even in the interior, for example, the precisions of the estimates of D^1x and D^2x were at least doubled.

5.2 A single forced harmonic

How well does the method do when applied to a single functional observation? A second order equation with coefficients $\beta_0 = 4.04$ and $\beta_1 = 0.4$ was forced by a step function u that was zero up to $t = 2\pi$ and one after, and multiplied by coefficient $\alpha = -2.0$. Noise sampled from $N(0, 0.04)$ was added. One hundred trials were conducted, and in each λ was chosen by minimizing GCV.

The mean estimates of coefficients β_0, β_1 and α were $4.041 \pm 0.007, 0.397 \pm 0.005$ and -1.998 ± 0.0009, respectively, indicating no detectable bias.

6 The oil refinery data

After some experimentation with first and second order models, and with constant and varying coefficient models, the clear conclusion was that the constant coefficient model $Dx = -0.02x - 0.19u$ was preferred. The standard error for β was estimated be 0.0004 by both the bootstrapping and the delta methods. The corresponding estimates for the standard error of α were 0.0024 and 0.0025, respectively.

Figure 2 shows the data for the tray level along with the fit to the data implied by the differential equation.

7 The lupus data

7.1 The disease

Systemic lupus erythematosus (SLE), or simply “lupus”, is an auto–immune disease in the same family as rheumatoid arthritis. The body’s immune system attacks itself, producing a wide spectrum of symptoms and affecting many organs. These attacks, called *flares*, occur suddenly and unpredictably, last for varying periods, and then disappear, sometimes for long periods. “Erythematosus” means reddening, referring to a characteristic skin rash by which it was first identified.

The disease is incurable. Around 9 times as many women as men get the disease, and blacks and some Asian groups more susceptible. Incidence

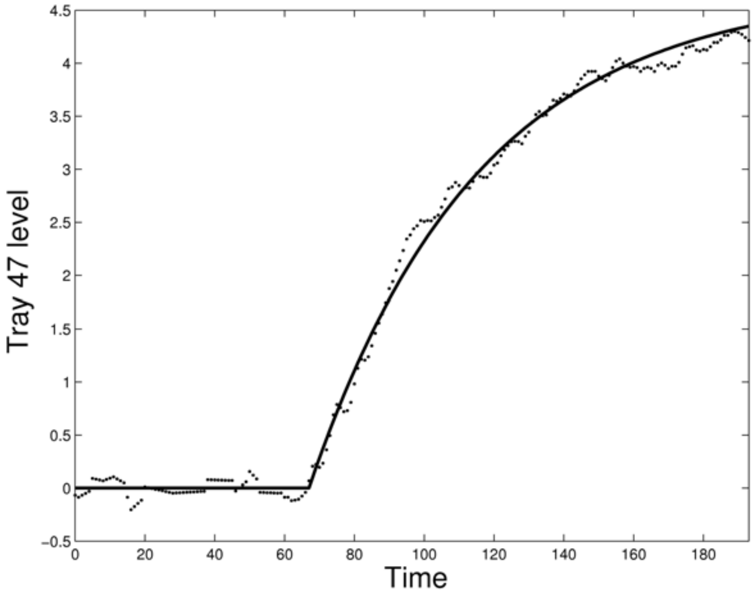


Figure 2: The fit to the data defined by the differential equation is shown as a solid line, and the data as points.

ranges from 3 to 400 per 100,000. Lupus can appear at any age, and the earlier it appears, the more severe it tends to be. Lupus is on the increase, and in some places is now more common than rheumatoid arthritis. Genetic, environmental, and hormonal factors are all involved. Exposures to chemicals and ultra-violet light are suspected triggers for flares.

Symptoms range from mild to severe, and can cause permanent damage or be fatal. A rash on the face and chest, pain and swelling in the joints and fatigue are common and early signs of a flare. The kidneys are often affected, with swelling and loss of function, and end–stage renal failure is a real risk. The heart, arteries, lungs, eyes and central nervous system may also be involved; and the psychological effects of lupus are receiving more and more attention. A typical flare goes from just noticeable to acute in the order of ten days or less.

The variation in the nature and severity of symptoms combined with the unpredictability of flares makes treating this disease a huge challenge. Mild symptoms are treated with anti–inflammatory drugs (aspirin, etc.), and more severe symptoms require the use of corticosteroids, usually *prednisone*. The response time to an increase in prednisone dose is usually of the order of a few days. However, corticosteroids are toxic if taken over long periods at high doses, with common side effects being weight gain, sleeplessness and osteoporosis. Sudden decreases in dose can trigger a new flare; consequently, high dose levels must be tapered down gradually.

Patients are assessed at regular intervals. Although lupus symptoms are multidimensional, long term treatment requires some overall measure of disease severity. A number of symptom severity scales have been proposed, and the SLEDAI scale is now widely used. SLEDAI is a check list of 24 symptoms, each given a numerical weight ranging from 1 for fever to 8 for seizures.

A flare has been defined by an international committee as a SLEDAI score increase of 3 or more to a level of 8 or higher. During flares SLEDAI scores of 25 to 30 are common.

A joint McGill/University of Toronto team headed by Dr. Paul Fortin has complete histories for about 300 patients spanning, in many cases, around 20 years. This is one of the largest and highest quality set of patient records in the world.

Figure 3 shows the data for a single patient over a three-year period. Notice the strong flare that coincides with the reduction in prednisone dose just after the seventh year.

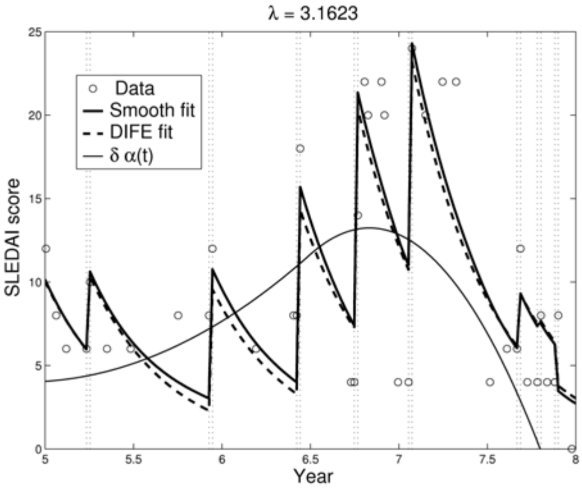


Figure 3: The data for a single patient over a three-year period. Heavy lines join times and values of SLEDAI measurements. A flare is indicated by a solid heavy line joining the first SLEDAI measurement within the flare to the previous measurement extended to a time 0.02 years back. The light solid line joins times and values at which prednisone doses were fixed.

7.2 The statistical challenges

We require a model for:

- flare timings (a point process)
- flare intensities (a marked point process)
- flare durations (a marked interval process)
- flare dynamics: rate of onset and rate of recovery
- how flare characteristics depend on prednisone level and
- prednisone dynamics or rate of change
- individual differences in all of the above

7.3 Data issues

The SLEDAI scale score has limited reliability. The dates at which these scores are assessed are themselves haphazard. Some data may be actually missing, eg: does SLEDAI = 0 always mean "no symptoms"?

We can, however, work closely with the physicians who work with these patients to identify flare characteristics, including flare onset times, flare durations, and to answer some questions. For example, a SLEDAI score may not change, but the fact that prednisone was increased at that point suggests that the disease has nonetheless become acute. We can also return to patient records to retrieve other information as required.

7.4 A simple model for flare dynamics

Let $u(t)$ be an indicator function for when lupus is in its active state and a flare is taking place.

- $u(t)$ takes only values 0 and 1.
- The times t_i at which $u(t)$ becomes positive can be estimated directly from the data, and therefore assumed known.
- The duration of an active state will be δ, and may vary from flare to flare.

We might propose to model symptom level $s(t)$ as a first order differential equation:

$$Ds(t) = -\beta s(t) + \alpha(t)u(t) \tag{5}$$

This, however, is too simple in an important way. It predicts that the rate of increase in symptom level is equal to its rate decrease when $u(t)$ returns to zero. In fact, however, symptoms rise far more rapidly than they decay.

We can imagine that the disease also affects the body's capacity to respond to the disease itself, as well as it's capacity to recover. That is, β is also affected by the disease, and therefore must be replaced by the function $\beta(t)$. When the patient is healthy between flares, $\beta(t)$ is high, leading to rapid response to the onset of the disease. When the patient is experiencing a flare, $\beta(t)$ is near zero, implying a slow recovery.

We tried this differential equation for $\beta(t)$

$$D\beta(t) = -\gamma\beta(t) + \theta[1 - u(t)]$$

When $u(t)$ switches on, $\beta(t)$ decays to zero, and $Ds(t)$ tends to equal $\alpha u(t)$; that is, $s(t)$ increases linearly while $u(t) = 1$. When $u(t)$ switches off, $\beta(t)$ returns to the level of its gain, θ/γ, and $s(t)$ tends to decay exponentially with rate equal to $\beta(t)$'s gain.

This gives us the general shape of a lupus flare. The increase in symptoms is essentially linear because $\beta(t)$ decays rapidly to 0 inside a flare; when $\beta(t) \approx 0$, the gain becomes $\alpha\delta$. But after a flare, when $u(t)$ returns to zero, β returns to its healthy level, and there is an exponential decrease in symptoms.

Actually, preliminary results indicated large values for rate parameter γ, implying that $\beta(t)$ moved extremely rapidly between virtually zero and its maximum value, defined by θ. We decided to simplify the differential equation for $\beta(t)$ to

$$D\beta(t) = \theta[1 - u(t)]$$

This implies linear increase within a flare episode, and exponential decrease afterwards with a rate constant θ.

7.5 The data analysis

Order 4 B-spline basis functions, with a knot at every data point, and three coincident knots at the times of onset and offset of flares were used to represent symptom function $s(t)$. Coincident knots allow the first derivative to be discontinuous at flare boundaries, as required by the model. Coefficient $\alpha(t)$ was made nonconstant, and represented as a basis function expansion in terms of four order 3 B-splines.

Figure 4 shows the fitting function s, the solution to the differential equation and coefficient function α for smoothing parameter $\lambda = 10^{-0.5}$. For this value of λ, s fits the data quite well everywhere, and certainly adequately given the reliability of the SLEDAI score. But the solution of the differential equation climbs with each successive flare because the estimated rate constant $\theta = 0.61$ is too small to allow enough decay in symptoms between flares.

Figure 5 shows these results for the higher smoothing parameter $\lambda = 10^{0.5}$. Now we see that the fit $s(t)$ and the solution to the differential equation are very close. The estimate of θ is now 2.24, and this rate constant is sufficient

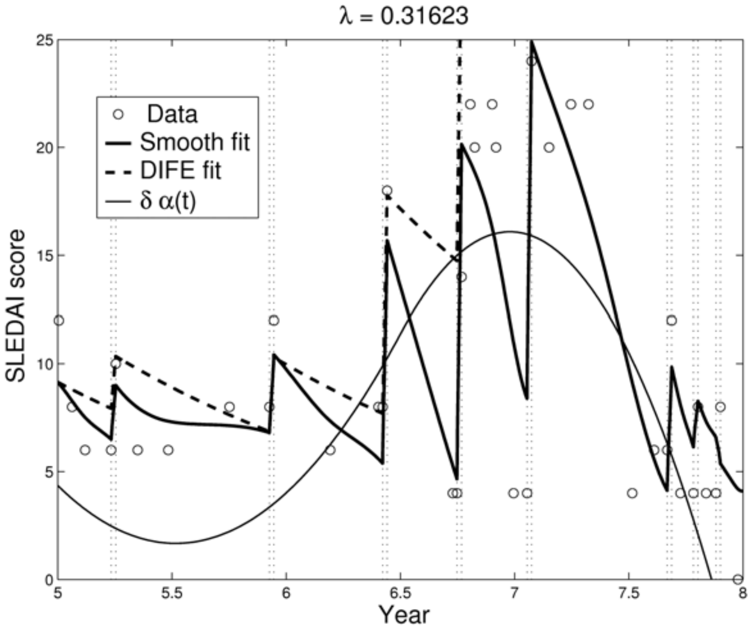


Figure 4: Results for the analysis of the data in Figure 3 using a smoothing parameter $\lambda = 10^{-0.5}$. The circles are SLEDAI measurements. The heavy solid line is the fit to the data $s(t)$ that minimizes criterion (4) and the dashed line is the solution to the differential equation (5). The light solid line plots the value of $\delta\alpha(t)$.

for the recovery from a flare before the next flare begins. What we lose, however, is the capacity to fit lower values of SLEDAI; the range of variation within a flare is too limited to permit this.

On the whole, however, these fits are quite satisfactory and capture well the main dynamic features of this segment of a lupus record.

References

[1] Ramsay, J. O. and Silverman, B. W. (1997). *Functional data analysis*. New York: Springer.

[2] Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis*. New York: Springer.

Acknowledgement: The statistical investigations of these data are funded by the Canadian Institute for Health Research (CIHR). The co-investigators on the lupus project are Dr. Michal Abrahamowicz, McGill University, and Dr. Paul Fortin, University of Toronto. I would also like to recognize the

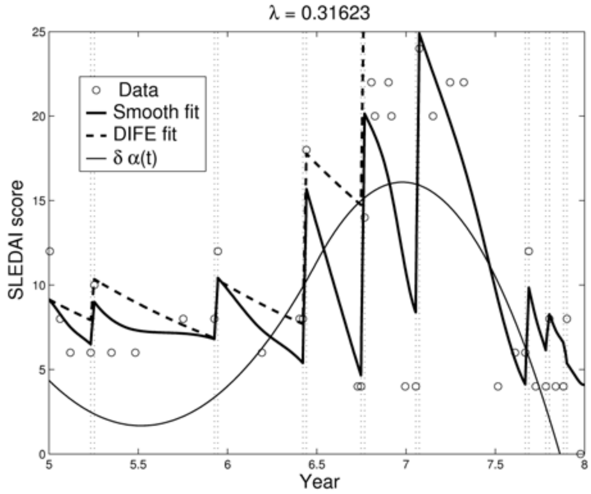


Figure 5: Results for the analysis of the data in Figure 3 using a smoothing parameter $\lambda = 10^{0.5}$. The circles and lines are as in Figure 4.

contributions of my graduate students, Mr. Jiguo Cao, Ms. Carlotta Fok and Ms. Wen Zhang. The example from chemical engineering was supplied by Dr. James McLellan of Queen's University, and this research also benefited from the research collaboration with Andrew Poyton, a graduate student at Queen's.

Address: J.O. Ramsay, McGill University, 1205 Dr. Penfield Ave., Montréal, Québec, Canada H3A 1B1

E-mail: `ramsay@psych.mcgill.ca`

SIMPLE SIMULATIONS FOR ROBUST TESTS OF MULTIPLE OUTLIERS IN REGRESSION

Marco Riani and Anthony Atkinson

Key words: Forward search, large data sets, simultaneous inference, trimmed estimators..

COMPSTAT 2004 section: Robustness.

Abstract: The null distribution of the likelihood ratio test for outliers in regression depends on the distributional properties of trimmed samples. Approximations to the distribution of the statistic that are simple to simulate are described and applied to three examples.

1 Introduction

Tests of outliers in regression need estimates of both the parameters of the linear model and of the error variance σ^2. If the outliers are included in the set used for estimation, inconsistent estimates of the parameters will be obtained and the existence and the effect of the outliers will be masked. We therefore consider procedures in which the observations are divided into two groups: those believed to be 'good' and the outliers. The good observations are used to provide estimates of the parameters to be used in the test for outliers.

Let there provisionally be m good observations out of n. We are interested in the null distribution of the outlier test. We therefore need to perform our calculations as though there were no outliers. If we were interested in the simplest case when, instead of regression, the focus is the location parameter of a random sample from a symmetrical distribution, we would base our estimates on the m central observations, trimming the remaining $m - n$. The properties of our estimators would then be those coming from this trimmed sample of n observations, rather than from m observations taken at random from the parent population. We use this insight to provide excellent approximations to the distribution of the outlier test in regression.

The literature on the detection of outliers in regression is vast. The test we study here is the likelihood ratio test, that is the test based on the prediction residuals used, for example, by Hadi and Simonoff [13], for the detection of multiple outliers. Two useful surveys of methods for multiple outliers in regression are Beckman and Cook [9] and Barnett and Lewis [8]. An important point is that, if several outliers are present, single deletion methods (for example, Cook and Weisberg [12], Atkinson [1]) may fail. Hawkins [14] argues for exclusion of all possibly outlying observations, which are then tested sequentially for reinclusion. This corresponds to our description in which m observations are used for estimation.

The drawback to Hawkins's procedure is that it is unclear how many observations should be deleted, and, because of masking, which ones, before reinclusion and testing begin. However, the forward search is an objective procedure of this type: it starts from a small, robustly chosen, subset of the data and fits subsets of increasing size. Each newly introduced observation can be tested for outlyingness before it is included in the fitted subset.

The use of the forward search in regression is described in Atkinson and Riani [4] where, as in Atkinson [2], the emphasis is on informative plots and their interpretation. The extension to multivariate data is described by Atkinson [3], with a book length treatment in Atkinson, Riani and Cerioli [7]. Although the forward search is a powerful general method for the detection of multiple outliers and unidentified clusters, the references do not describe inferential procedures based on the quantities plotted. Atkinson and Riani [6] use the forward search as a means of generating a series of outlier tests with decreasing amounts of trimming; m increases from slightly more than the number of parameters to n. The values of the statistics are assessed by simulation and by analytical approximations to the robust tests. The interest in the present paper is in the application of the tests. We use both simulations of forward searches and two simple simulated approximations to the distribution to analyse three sets of data. As a result we are able to combine the power of the forward search with precise statistical procedures.

The paper is organised as follows: in §2 we briefly review the forward search and robust estimation; both depend on estimators from trimmed samples. In §3 we write the outlier test explicitly in terms of such samples and show how simulations using samples from trimmed distributions can be used to approximate the distribution of the statistic. Examples in §4 show how well our approximation works. The final section briefly describes further work.

2 Least squares and outlier detection

2.1 Least squares

In the regression model

$$y = X\beta + \epsilon, \tag{1}$$

y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^T, and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$.

With $\hat{\beta}$ the least squares estimator of β the vector of least squares residuals is

$$e = y - \hat{y} = y - X\hat{\beta} = (I - H)y, \tag{2}$$

where $H = X(X^TX)^{-1}X^T$ is the 'hat' matrix, with diagonal elements h_i and off-diagonal elements h_{ij}. The mean square estimator of σ^2 can be written

$$s^2 = e^Te/(n-p) = \sum_{i=1}^{n} e_i^2/(n-p). \tag{3}$$

We define the standardized residuals

$$q_i = e_i/\sqrt{(1-h_i)} = (y_i - \hat{y}_i)/\sqrt{(1-h_i)}. \tag{4}$$

Like the errors ϵ_i, the q_i are distributed $N(0, \sigma^2)$, although they are not independent.

The likelihood ratio test for agreement of a new observation y_{new} observed at x_{new} with the sample of n observations providing $\hat{\beta}$ and s^2 is the prediction residual

$$d_i^* = \frac{y_{\text{new}} - x_{\text{new}}^T\hat{\beta}}{s\sqrt{\{1 + x_{\text{new}}^T(X^TX)^{-1}x_{\text{new}}\}}}, \tag{5}$$

which, when the observation y_{new} comes from the same population as the other observations, has a t distribution on $n-p$ degrees of freedom.

2.2 The forward search

Let $\mathcal{M}$ be the set of all subsets of size m of the n observations. The forward search fits subsets of observations of size m to the data, with $m_0 \le m \le n$. We discuss the starting point of the search in §2.3.

Let $S_*^{(m)} \in \mathcal{M}$ be the optimum subset of size m. Least squares applied to this subset yields parameter estimates $\hat{\beta}(m^*)$ and $s^2(m^*)$, the mean square estimate of σ^2 on $m-p$ degrees of freedom. Residuals can be calculated for all observations including those not in $S_*^{(m)}$. The n resulting standardized residuals can from (4) be written as

$$q_i(m^*) = \frac{y_i - x_i^T\hat{\beta}(m^*)}{\sqrt{\{1 - h_i(m^*)\}}}. \tag{6}$$

The notation $h_i(m^*)$ serves as a reminder that the leverage of each observation depends on $S_*^{(m)}$. The search moves forward with the subset $S_*^{(m+1)}$ consisting of the observations with the $m+1$ smallest absolute values of the e_i, that is the numerator of $q_i(m^*)$.

In order to simulate the distribution of the outlier test of §2.4 we need a simple way of simulating variables with the same distribution as the $q_i(m^*)$. When $m = n$ these residuals are those in (4) and the distribution is $N(0, \sigma^2)$. But with $m < n$ the estimates of the parameters are based on only those observations giving the central m residuals: $\hat{\beta}(m^*)$ and $s^2(m^*)$ are calculated from truncated samples.

2.3 Robust estimation and the start of the search

The search starts from a subset of p observations $S_*^{(p)}$ that is chosen to provide a very robust estimator of the regression parameters. For example, if Least Median of Squares (LMS, Rousseeuw [16]) is used, the subset of p observations is found minimizing the scale estimate

$$\sigma^2(h) = e^2_{[h]}(p^*), \tag{7}$$

where $e^2_{[k]}(p^*)$ is the kth ordered squared residual and h is the integer part of $(n+p+1)/2$ and corresponds to 'half' the observations when allowance is made for fitting. Typically the search either examines all subsets of size p, if this is not too large, or several thousand subsets are examined at random. These starting methods destroy masking; any remaining outliers are then removed in the initial steps of the search. Consequently, the search is insensitive to the exact starting procedure. What is important for our present purpose is that the search again uses parameter estimates based on a central part of the sample.

2.4 Testing for outliers

Let the observation "nearest" to those constituting $S_*^{(m)}$ be $i_{\min}$ where

$$q_{\text{imin}}(m^*) = \min |q_i(m^*)| \quad \text{for} \quad i \notin S_*^{(m)}, \tag{8}$$

the observation with the minimum prediction residual among those not in $S_*^{(m)}$. If observation $i_{\min}$ is an outlier, so will be all other observations not in $S_*^{(m)}$.

To test whether observation $i_{\min}$ is an outlier we use the predictive residual (5). The test for agreement of the observed and predicted values is

$$|d_{\text{imin}}| = \left| \frac{y_{\text{imin}} - x^T_{\text{imin}}\hat{\beta}(m^*)}{s(m^*)\sqrt{\{1 + h_{\text{imin}}(m^*)\}}} \right|. \tag{9}$$

It is the distribution of this statistic that is the subject of this paper. In (5), when all observations were used in fitting and a new observation was being tested, the distribution was t_{n-p}. Now the estimates $\beta(m^*)$ and $s(m^*)$ are based on the central part of the distribution. Even under the null hypothesis that the sample contains no outliers, the distribution is no longer t.

3 Simulating the distribution

The empirical distribution of the series of test statistics can be found by repeated simulations of forward searches. In this section we describe this method and then describe two alternative simulation-based methods. The first replaces the series of simulations and forward searches with independent simulations for each value of m. The second uses a series of orderings of simulated data, but avoids the forward search.

Both of these methods are for the statistics calculated for simple samples. In §3.4 we introduce a correction for the dependence of the distribution of the statistics on p.

3.1 The empirical distribution

In order to find the distribution of the test statistic during the forward search the most straightforward method is to simulate samples of all n observations and repeat the forward search a number of times. In order to capture any special features of the hat matrix, the matrix of explanatory variables is that of the data under study. Observations are simulated using the fitted values at the end of the search, that is $x_i^T\hat{\beta}(n)$, and the estimated standard deviation $s(n)$.

3.2 Method 1: Truncated samples

We are interested in approximations to the null distribution of (9) for given m which can easily be found. The statistic is a function of the m residuals $q_i(m^*) \in S_*^{(m)}$ and of $q_{\text{imin}}(m^*)$. In the absence of outliers, these will be the observations with the $m+1$ smallest values of $|q_i(m^*)|$. Since the $q_i(m^*)$ are residuals, their distribution does not depend on the parameters β of the linear model. They have also been standardised to have constant variance, which is then estimated. To find the required distribution we therefore simulate from a truncated normal distribution and calculate the value of the outlier test for such samples. The steps are:

Step 1. Obtain a random sample of $m+1$ observations U_i from the uniform distribution on $[0.5-(m+1)/2n, 0.5+(m+1)/2n]$.

Step 2. Use the inversion method to obtain a sample of $m+1$ from the truncated normal distribution:

$$z_i = \Phi^{-1}(U_i), \tag{10}$$

where Φ is the standard normal c.d.f.

Step 3. Find the most outlying observation:

$$z_{\text{imin}} = \max |z_i| \quad i = 1, \dots, m+1. \tag{11}$$

Then $S_z^{(m)} = \{z_i\}, i \neq i_{\text{min}} = 1, \dots, m+1$.

Step 4. Estimate the parameters. Let $\bar{z}(m)$ be the mean of the m observations in $S_z^{(m)}$ and $s_z^2(m)$ be the mean square estimate of the variance.

Step 5. Calculate the simulated value of the outlier test in (9):

$$d_{\text{imin}}^z = \frac{z_{\text{imin}} - \bar{z}(m)}{s_z(m)\sqrt{\{(m+1)/m\}}}. \tag{12}$$

The simulation of the truncated normal distribution using the inversion method in Steps 1 and 2 is straightforward in S-Plus or R.

3.3 Method 2: Ordered observations

In the forward search the n observations are ordered for each value of n. In the absence of outliers we might expect that this order would not change much during the search. As a second method of approximating the distribution of the statistics, we simulate sets of n observations from the normal distribution, correct for the mean and order the absolute values of the observations. For our calculations for each value of m we use the m smallest absolute residuals to estimate the parameters. The procedure is repeated several times, typically 1,000, to give the empirical distribution of the statistics.

3.4 Adjustment for regression

In both Method 1 and Method 2 we estimate the sample mean, rather than a regression model, so $h_{\mathrm{imin}}(m) = 1/m$. Simulations show that the resulting upper percentage points of the distribution are too small when we are analysing regression data. Good agreement is obtained by using the adjusted statistic

$$|d_{\mathrm{imin}}| = \left| \sqrt{\frac{m+\theta p}{m}} \frac{y_{\mathrm{imin}} - x_{\mathrm{imin}}^T \hat{\beta}(m^*)}{s(m^*)\sqrt{\{1 + h_{\mathrm{imin}}(m^*)\}}} \right|, \qquad (13)$$

with $\theta = 0.7$. As m increases, the effect of the correction becomes less.

4 Examples

4.1 Hawkins's data

This set of simulated data was analysed by Atkinson and Riani [4], §3.1. There are 128 observations and nine explanatory variables. The data were intended by Hawkins to be misleading for standard regression methods. Figure 1 shows a forward plot of the minimum deletion residual among observations not in the subset, that is the outlier test statistic (13), together with two sets of simulated percentage points of the distribution, both based on 1,000 simulations. We first consider these simulation envelopes.

The envelopes plotted with continuous lines in the figure are the 1, 2.5, 5, 50, 95, 97.5 and 99% points of the empirical distribution of the outlier test during forward searches simulated without outliers. The dotted lines are from our second approximate simulation method in which random samples of observations are ordered once. Agreement between the two envelopes is excellent during the second half of the search; agreement between the two sets of upper envelopes is also good during the first half of the search for $m > 20$. The envelopes are of a kind we shall see in all simulations. Initially they are very broad, corresponding to distributions with high trimming and

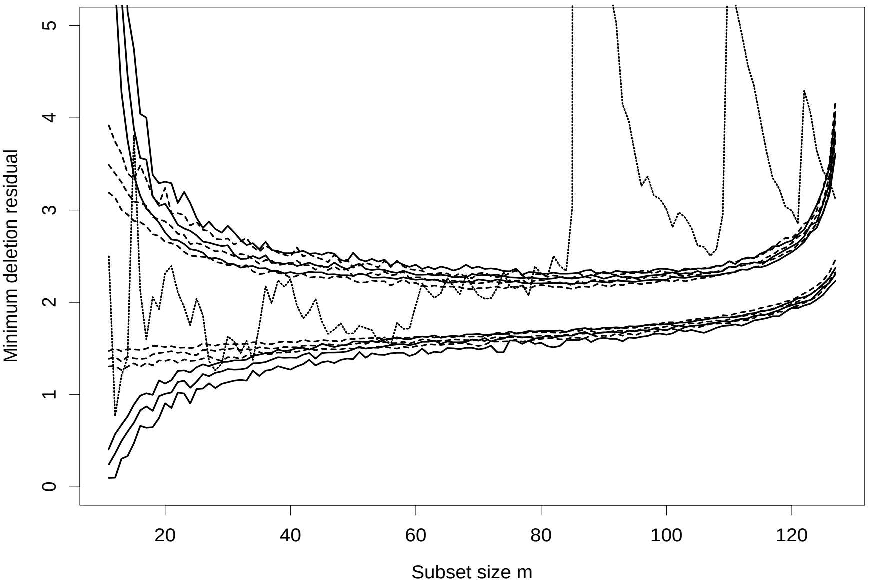


Figure 1: Hawkins's Data: forward plot of minimum deletion residuals (the outlier test). The four groups of observations are clearly separated by the three large peaks signalling the first observation from each new group immediately before it enters the subset. The dotted lines are envelopes simulated by Method 2.

few degrees of freedom for the estimation of error. In the central part of the search the band is virtually horizontal and gradually narrows. Towards the end of the search there is rapid increase as we test the few largest residuals.

The continuous line showing the plot of the outlier test in the figure reveals all the features that Hawkins put in the data. There are 86 observations with very small variance. The plot shows a huge jump in the value of the statistic when the first observation of the next group enters. This process is repeated two more times, clearly identifying the four separate groups of data that are present, the decline after each peak being due to the effect of masking. The forward plot of this test statistic is the same as that in the lower panel of Figure 3.6 of Atkinson and Riani [4]; the new confidence bands calibrate inferences about the significance of the peaks.

The envelopes rise rapidly at the end of the search and we can see that the outlier test finishes up being non-significant. Thus Hawkins has succeeded in constructing a data set with many outliers all of which are masked. The curve of the statistic starts to rise just before $m = 86$. If we take only the first 86 observations and provide simulation envelopes for them, the envelopes rise at the end as the envelopes do here from m around 125. The last few observations do not then lie outside the simulation bands for this reduced set of data.

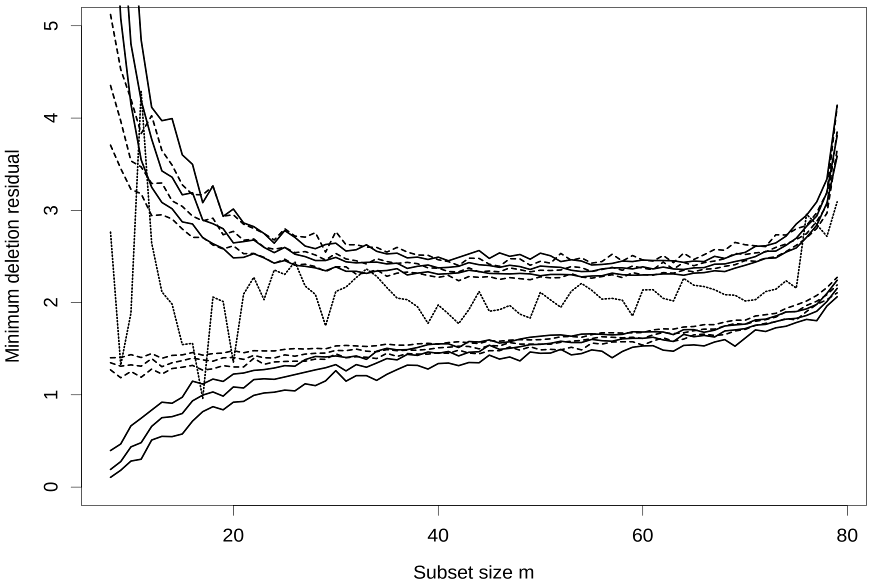


Figure 2: Ozone Data: forward plot of minimum deletion residuals (the outlier test). There are some mild outliers towards the end of the search and some evidence of masking. The dotted lines are envelopes simulated by Method 1.

4.2 Ozone data

Hawkins's data are a synthetic example in which there are many outliers. We now consider two examples of real data.

The first is the ozone data from Breiman and Freedman [10] which give readings of ozone concentration on 300 consecutive days. The results for the first 80 days were extensively analysed by Atkinson and Riani [4], §3.4. Here we follow their analysis.

As a result of the use of the forward search combined with response transformation the final model found by Atkinson and Riani had a logged response with five of Breiman and Freedman's original variables augmented by a linear trend in time. Figure 2 shows a forward plot of the outlier test for this model together with simulation envelopes from the forward search (continuous lines) and the approximate envelopes from the first method, of sampling from a truncated distribution. The agreement between the two sets of envelopes is again good, particularly for the upper envelope.

The evidence from this plot is much less dramatic than that of Figure 1. Apart from the very beginning of the search, the plot lies near or within the bounds for all values of m up to the introduction of the 76th observation. Thereafter there seem to be four mild outliers, a conclusion in line with the forward plot of residuals in Figure 3.37 of Atkinson and Riani [4]. At

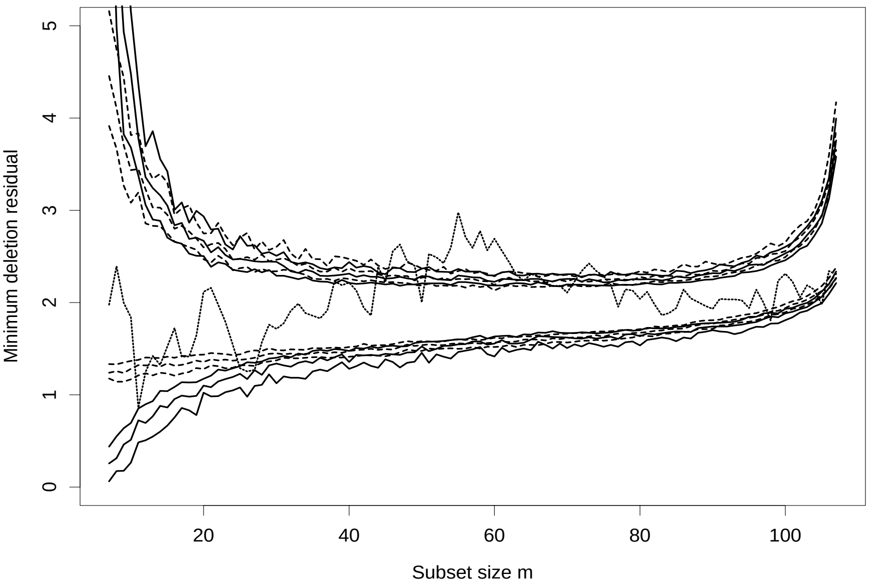


Figure 3: Surgical Unit Data: forward plot of minimum deletion residuals (the outlier test). The appreciable maximum of the statistic in the centre of the search suggest there may be two equal sized groups of observations that differ in some systematic way. The dotted lines are envelopes simulated by Method 1.

$m = 76$ this plot shows four appreciable residuals, three negative and one positive: these lie apart from the general cloud of residuals throughout the whole search. The plot also shows some evidence of masking, the residuals decreasing somewhat in magnitude at the end of the search. The effect of masking is also evident in Figure 2, where the test statistic lies within the simulation envelopes for the last two steps of the search. Although the masking here is not as misleading about the structure of the data as that in Figure 1, there are again outliers whose presence would be overlooked by an analysis based on all the data, or on single deletion diagnostics.

4.3 Surgical unit data

Neter, Kutner, Nachtsheim and Wasserman [15] introduce, on p.334, data on the survival time of 54 patients undergoing liver surgery, together with four explanatory variables that may be used to predict survival time. Their preferred model regresses y on three of the explanatory variables, x_4 being excluded. On p.437 another 54 observations are introduced to check the model fitted to the first 54. Their Table 10.9 compares parameter estimates from the two sets for the preferred regression model. The conclusion is that there is no systematic difference between the two sets and that the same model is acceptable for all the data.

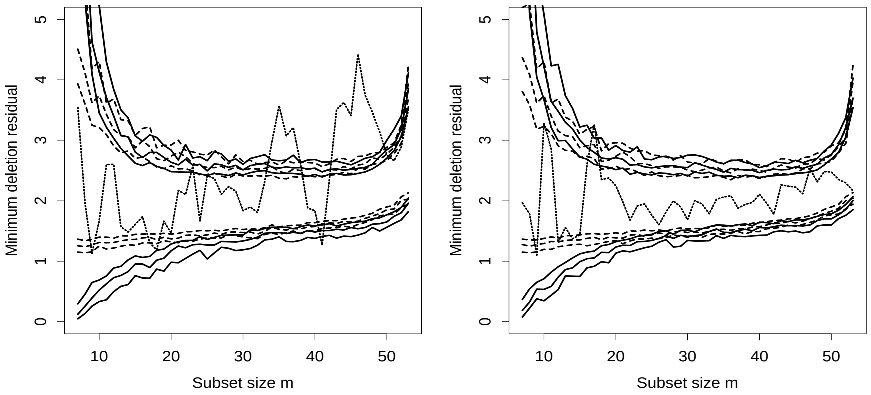


Figure 4: Surgical Unit Data: forward plot of minimum deletion residuals (the outlier test) for the first and second 54 observations. There is strong evidence that here are three groups amongst the first 54 observations. The dotted lines are envelopes simulated by Method 1.

Atkinson and Riani [5] analysed the combined set of all 108 observations using the forward search to assess the influence of individual observations on the estimated regression coefficients. They also conclude that a logged response and a linear model in $x_1 - x_3$ adequately describes the data. Because we will shortly be augmenting the set of explanatory variables, we work with all four original variables.

Figure 3 is a forward plot of the test for outliers for all 108 observations, together with simulation envelope and the approximation found by our first method. This surprising plot seems to show evidence of two groups - the extreme value of the statistic, well outside the boundaries is at the entre of the search, after which there is a gradual decline in the values. At the end of the search the statistic is nudging the lower envelope, a stronger version of the effect of masking noticed in the two previous figures.

Since the maximum value of the statistic is at $m = 55$, we examine those units that enter after this value, to see whether they might belong to a second cluster. Detailed analysis of the results of the forward search show that, after $m = 57$ nearly all the patients entering have unit numbers greater than 54 and so come from the group of confirmatory observations.

This figure suggests the group of confirmatory observations may be different from the original 54 units. Accordingly, we introduce a dummy variable for the two sets and repeat the analysis. This variable is highly significant, with a t value of -7.83 at the end of the search. However, the resulting forward plot still has a slight peak in the centre, although this is much reduced from that in Figure 3. Some remaining structure is indicated.

To take the analysis further we consider the two groups separately. Figure 4 gives the forward plots of the test for outliers. The plot for the second group of observations in the right-hand panel, suggests that the group is homogeneous. However, that in the left-hand panel strongly indicates that the first group contains at least one identifiable subgroup that needs to be disentangled before further analysis is undertaken. A next stage in the analysis would be to extend the scatterplot matrix of the data in Figure 8.3 of Neter et al. [15] to include different plotting symbols for the tentative groups.

5 Discussion

The previous examples are comparatively small and the many plots from the forward search can easily be interpreted. However, as the number of units increases, plots for individual units, such as forward plots of residuals, can become messy and uninformative due to overplotting. Atkinson and Riani [6] analyse 500 observations on the behaviour of customers with loyalty cards from a supermarket chain in Northern Italy. Despite the larger number of observations the forward plot of the test for outliers is as easily interpreted as those in this paper and shows an unsuspected group of 30 very different customers.

There are two further general methodological matters that deserve comment. The first is that the envelopes presented in this paper were all found by simulation. An alternative, investigated by Atkinson and Riani [6], is to calculate the percentage points directly using analytical results on order statistics and the variance of truncated normal distributions. The other point is that, however the envelopes are calculated, the probability statements refer to pointwise exceedance of the bands. To find, for example, the probability of at least one transgression of a specified envelope somewhere during a particular region of the search, for example the second half, requires calculation of the simultaneous probability of transgression at any of the stages of the search within that region. Computationally feasible methods are described by Buja and Rolke [11].

Atkinson and Riani [6] may be viewed at `www.lse.ac.uk/collections/statistics/research/`

References

[1] Atkinson A.C. (1985). *Plots, transformations, and regression.* Oxford University Press, Oxford.

[2] Atkinson A.C. (1994). *Fast very robust methods for the detection of multiple outliers.* Journal of the American Statistical Association **89**, 1329 – 1339.

[3] Atkinson A.C. (2002). *The forward search.* In W. Härdle and B. Rönz, editors, COMPSTAT 2002: Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, 587 – 592.

[4] Atkinson A.C., Riani M. (2000). *Robust diagnostic regression analysis.* Springer–Verlag, New York.

[5] Atkinson A.C., Riani M. (2002). *Forward search added variable t tests and the effect of masked outliers on model selection.* Biometrika **89**, 939 – 946.

[6] Atkinson A.C., Riani M. (2004). *Distribution theory and simulations for tests of outliers in regression.* Submitted.

[7] Atkinson A.C., Riani M., Cerioli A. (2004). *Exploring multivariate data with the forward search.* Springer–Verlag, New York.

[8] Barnett V., Lewis T. (1994) *Outliers in statistical data (3rd edition).* Wiley, New York.

[9] Beckman R.J., Cook R.D. (1983) *Outlier detection (with discussion).* Technometrics **25**, 119 – 163.

[10] Breiman L., Friedman J.H. (1985). *Estimating optimal transformations for multiple regression and transformation (with discussion).* Journal of the American Statistical Association **80**, 580 – 619.

[11] Buja A., Rolke W. (2003). *Calibration for simultaneity: (re)sampling methods for simultaneous inference with applications to function estimation and functional data.* Technical report, The Wharton School, University of Pennsylvania.

[12] Cook R.D., Weisberg S. (1982). *Residuals and influence in regression.* Chapman and Hall, London.

[13] Hadi A.S., Simonoff J.S. (1993). *Procedures for the identification of multiple outliers in linear models.* Journal of the American Statistical Association **88**, 1264 – 1272.

[14] Hawkins D.M. (1983). *Discussion of paper by Beckman and Cook.* Technometrics **25**, 155 – 156.

[15] Neter J., Kutner M.H., Nachtsheim C.J., Wasserman W. (1996). *Applied linear statistical models, 4th edition.* McGraw-Hill, New York.

[16] Rousseeuw P.J. (1984). *Least median of squares regression.* Journal of the American Statistical Association **79** 871 – 880.

Address: M. Riani, Dipartimento di Economia, Università di Parma, Italy
A. Atkinson, Department of Statistics, London School of Economics, UK
E-mail: `mriani@unipr.it, a.c.atkinson@lse.ac.uk`

THE ST@TNET PROJECT FOR TEACHING STATISTICS

Gilbert Saporta and Marc Bourdeau

Key words: Teaching, statistics, information society.

COMPSTAT 2004 section: Teaching statistics.

Abstract: This paper describes the design and development of St@tNet, an Internet environment for the teaching of basic Applied Statistics. St@tNet has been developed by a consortium of French-speaking universities. After some general considerations on education for the Information Society, and more specifically for the teaching of Statistics, we will present our product in its present state of development.

1 Means and ends

The title of this session is about teaching Statistics for the Information Society. Well, the Information Society began with the invention of the printing press with moveable type, and that has profoundly modified the formal education process. In essence, it has permitted widespread knowledge dissemination. For a few centuries things have stayed more or less the same, until the invention of mass media. Starting with the radio, then television, it became apparent that the world had once more profoundly changed. Their consequences in the formal and informal education processes were no doubt far-reaching, but now that we have entered the computer age, we have passed into speed *Warp Five* to speak StarTrek lingo; in the last few years the Internet development has brought about a genuine revolution in education thinking, actually a totally new *zeitgeist*.

Neil Postman [1931-2003], one of the keener observers of the evolution of education, and of society in general, has fully explored the consequences of this information revolution [4], [5]. He reports, and it is a common observation, that the situation of teachers and professors has become precarious: they are worried, even anxious, about their role and their immediate future in the Information Society.

Topping all this, governing bodies in many of the developed countries have nowadays become obstinate in drastically reducing budgets, with the elusive hope that the new technologies will give rise to an unprecedented increase in productivity: lesser means, greater expectations... Illusion, reality, who can tell? And what is the end of education finally?

At first sight then, it might appear that the new Information technologies (ITs) could lead to the end of the profession: at this journey's end, all the transmission of knowledge would originate from a few specialized quarters far away from students, pedagogical encounters would be virtual with the Internet being the sole communication channel. Universities and colleges would supply themselves for knowledge transmission and certification from those virtual hyper-classrooms.

ITs could secure huge savings for education boards, but could entail the disappearance of most teachers and professors.

From an overview of some recent and very successful pedagogical experiments in Québec universities using ITs, one can suspect that things will not be that simple [1]. The same situation, it is easy to confirm, is prevalent the world over. Actually, getting an education is a form a travelling. And quality travelling often imply personal guides, at least human encounters, not just guidebooks and TV documentaries though they can be illuminating and irreplaceable. In our experience, all the pedagogies devised with the ITs in mind have always implied more personal contacts with students, less mass dispensing of knowledge[1]

With the Internet, we have perhaps entered an era of renaissance of the true pedagogical relation, not the opposite. As we will explain, this has far reaching implications for teachers and students reciprocal relations.

1.1 Teaching statistics in the information age

Concerning Statistics and Data analysis, there is no gloom and doom scenario in view: there is a huge increase of informations that have to be processed. As John Wilder Tukey [1915-2000] has so correctly noted “The best thing about being a Statistician is that you get to play in everybody’s backyard.”

Better tools of analysis are badly needed, and, since there is already a widespread availability of data sources and an increased appetite for synthetic information, an important increase in Statistics literacy is urgently needed for an ever increasing number of people. Think, among other things, of the amount of information stored and available in national Statistics Offices the world over. All newspapers and mass media are now replete with reports of polls, of official statistics on the economy and society in general. Think also of the huge amount of business informations stored in Data Warehouses that come with an abundance of *Data Mining* softwares recently marketed. Making sense out of this “chaos” [8] is a huge undertaking. We are heading towards a knowledge-based society where statisticians will be ever more in demand.

We report here on the education material for the teaching of Statistics produced in our universities. See Saporta [6] for an overview of some of the web facilities for the teaching of Statistics[2]. See also the remarkable paper by Velleman & Moore for the ins and outs for the use of ITs in the teaching of Statistics [9].

[1] All the relevant documents upon which rests this assertion and that have been used for [1], are located on the following web pages `http://www.mgi.polymtl.ca/marc.bourdeau/InfAgeTeaching` ...

[2] Also available in the web pages just referred to.

2 The St@tNet project

The St@tNet project is developed at the *Conservatoire National des Arts et Métiers* (Cnam), a major public institution for continuous education and an integral part of the French Ministry of Education, Research and Technology. The Cnam was founded in 1794 to "enlighten ignorance that does not yet know, and poverty which cannot afford knowledge." More than 70 000 adult students attend its courses each year in numerous fields, two-thirds of them have already had two years of 'higher' education, one third are women.

Courses are given mainly in the evenings and in Saturday classes for credits leading towards a degree, as well as through in-service training during working hours, and, finally, through distance-learning. The Cnam links a network of 150 towns and is organized around a 'main' complex in Paris, 22 regional centers, plus some centers in overseas territories. One can begin a program anywhere in the network and continue in any other center. Graduate studies leading to Masters and PhDs are available in many disciplines.

St@tNet follows a series of previous developments of teaching materials for introductory Statistics that date back to the early nineties. Previous courses were available on diskettes and CD-roms [7]. The actual web-course version was financed by the *Agence Universitaire de la Francophonie* (*AUF*) and the French *Ministère de l'Éducation Nationale*. It is operational since 2002, and can be obtained also on a CD-rom version.

St@tNet is the only web resource proposed at the Cnam for distance learning for the much needed Introductory Statistics. It is freely accessible[3]. Indeed, having been financed by public funds, and for the advancement of public learning in conformity with its founding principles that go back to the Enlightenment Age, the decision of the free access of St@tNet was finally agreed upon after fierce debates, but registering at a cost of 250 euros is mandatory for certification purposes and the use of usual facilities: tutorship (one tutor per 25 students), an Internet access on a virtual teaching environment (VTE), an e-mail, etc. This fee comprises the CD-rom that avoids most of the Internet costs and waiting times, especially in distant locations. St@tNet is now also implemented on the virtual campuses of the *Agence Universitaire de la Francophonie* where it is one of the two most popular resources for self-education. Starting in the Fall of 2004, the Cnam will organize a certification system for the AUF courses. St@tNet is a complementary resource for the *École Militaire*, it is also recommended by the French association of mathematics teachers as an aid to school teachers who have to adapt themselves to new curricula that include elements of Probability and Statistics.

With its network of institutions, the Cnam is an ideal ground for the development of pedagogy and teaching material using ITs. Modern teaching of Applied Statistics requires the use of specialized software, and should be

[3] `http://www.agro-montpellier.fr/cnam-lr/statnet/`.

data based, centered on case studies for more advanced material and hands-on training. Applied Statistics is indeed much more than a set of mathematical formulae: its learning implies the development of "statistical thinking", requires the understanding of difficult concepts such as variation, randomness, laws of chance — a difficult oxymoron at first glance —, probable errors, risks, etc. Animations and various graphical tools provide efficient means of learning.

Depending on the level, one can think of various designs for the Internet environments and interactions. Up to now, there are two stages planned in the St@tNet project, the first one is fully operational, the second in development, but with partial versions tested in ordinary classrooms.

For the first stage, at the very basic level of statistical knowledge, St@tNet has opted for a complete Html environment. The advantage of this choice is that interactions of the students with the environment are quite easy to realize: this course is by no means a paper-course translated into Html, as one can still see quite often, but a full-fledged Html environment with frequent short interactions inserted by design into the course.

For higher levels of knowledge, where short interactions are much less needed, St@tNet has opted for a downloadable Latex-Pdf text, with full hyper-referencing possibilities, and many of the hyper-references are internal.

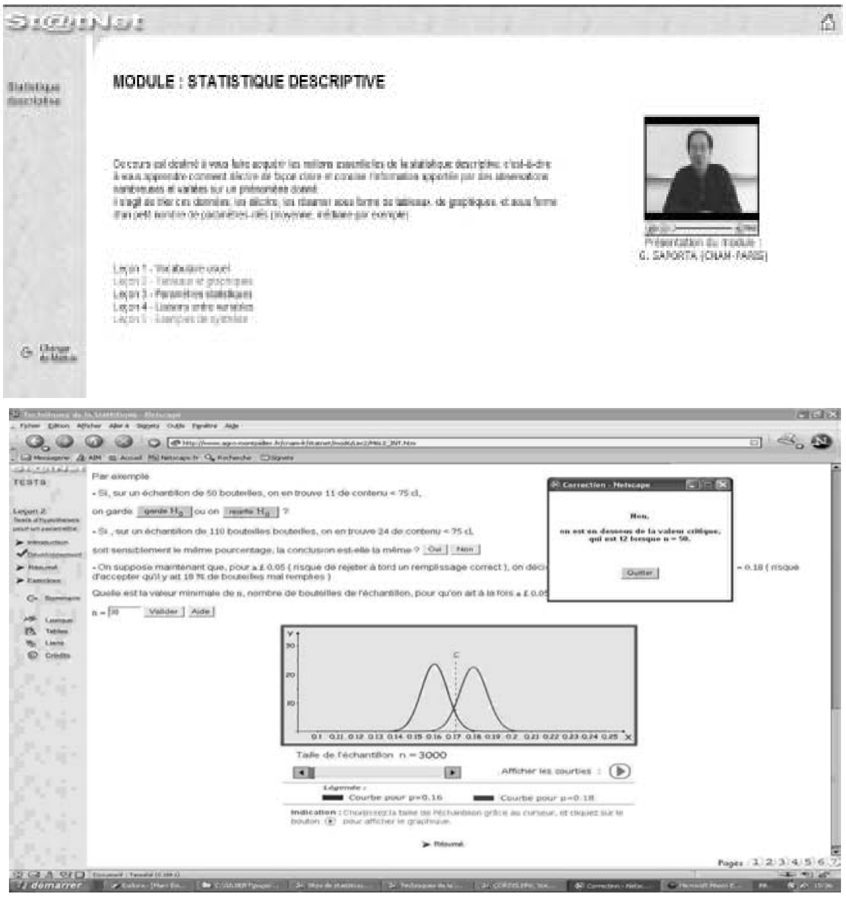
2.1 First stage: the basics

The first stage of the project, the one for the really basic knowledge, is now fully operational. It consists of six modules: data description, probability, random variables, sampling and estimation, tests, basic linear regression. Each of the modules is introduced by a video file (Figure 1, upper part) and is composed of lessons, all of which are of the same structure: Introduction, development, synopsis, exercices. A glossary of terms is accessible within each lesson, as well as all the necessary Statistical Tables and Internet links.

Once in a module, and after viewing its presentation video, the user can pick a *lesson* of his choice: indeed, the learning progression is not designed with a linear structure in mind. Most of our students detest such a progression that do not correspond to their needs.

The lower part of Figure 1 shows part of a page of the *Développement* (development) section of *Leçon* 1 (Lesson 1) of the module *Tests* (tests), with the shown pop-up window that is produced when a wrong answer is given by the reader. Upon a wrong answer, the reader can either correct his answer or get the right one with a short explanation.

Similarly to what is represented in this last Figure, lessons are interspersed with questions to the reader to check if the elements of learning have been correctly assimilated, as well as with some Flash animations and some hyperlinks to Java applets. All lessons end with a page of summary (Figure 2), and a few more elaborate exercices, again with answers given directly on the page, with pop-ups for feedback. A pop-up Glossary, the same for all lessons is hyper-referenced, and, finally, a page of links is available, with some of them referring to external Java applets useful for the learning.



MODULE : STATISTIQUE DESCRIPTIVE

Ce cours est destiné à vous faire acquérir les notions essentielles de la statistique descriptive, c'est-à-dire à vous apprendre comment décrire de façon claire et concise l'information apportée par des observations nombreuses et variées sur un phénomène donné.
Il s'agit de trier ces données, les décrire, les résumer sous forme de tableaux, de graphiques, et sous forme d'un petit nombre de paramètres-clés (moyenne, médiane par exemple).

Leçon 1 - Vocabulaire usuel
Leçon 2 - Tableaux et graphiques
Leçon 3 - Paramètres statistiques
Leçon 4 - Liaisons entre variables
Leçon 5 - Exemples de synthèse

Présentation du module :
G. SAPORTA (CNAM-PARIS)

Par exemple

- Si, sur un échantillon de 50 bouteilles, on en trouve 11 de contenu < 75 cl,

on garde [garde H_0] ou on [rejette H_0] ?

- Si , sur un échantillon de 110 bouteilles bouteilles, on en trouve 24 de contenu < 75 cl,

soit sensiblement le même pourcentage, la conclusion est-elle la même ? [Oui] [Non]

- On suppose maintenant que, pour a £ 0.05 (risque de rejeter à tord un remplissage correct), on déci = 0.18 (risque d'accepter qu'il y ait 18 % de bouteilles mal remplies)

Quelle est la valeur minimale de n, nombre de bouteilles de l'échantillon, pour qu'on ait à la fois a £ 0.05

n = 30 [Valider] [Aide]

Correction - Netscape

Non,
on est en dessous de la valeur critique, qui est 12 lorsque n = 50.

[Quitter]

Taille de l'échantillon n = 3000

Afficher les courbes :

Légende :
Courbe pour p=0.16
Courbe pour p=0.18

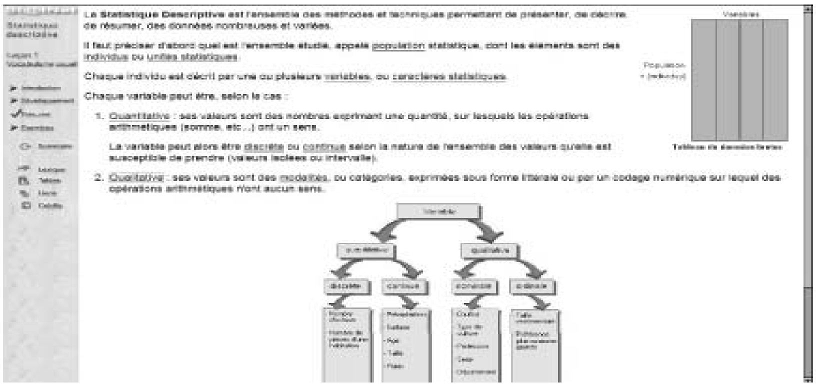
Indication : Choisissez la taille de l'échantillon grâce au curseur, et cliquez sur le bouton pour afficher le graphique.

Résumé

Pages : 1 2 3 4 5 6 7

Figure 1: Upper: The entry for the module *Statistiques descriptives* (Descriptive statistics), with its introductory video. Lower: Part of the development section for Lesson 1 of the module *Tests* (Tests), with a pop-up window obtained with a wrong answer.

A new audience has been reached by this approach, and the rate of retention and success is better than for traditional courses. This last point might be the consequence of the type of students (a "sampling bias" !) interested in such an environment.



La **Statistique Descriptive** est l'ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer, des données nombreuses et variées.

Il faut préciser d'abord quel est l'ensemble étudié, appelé population statistique, dont les éléments sont des individus ou unités statistiques.

Chaque individu est décrit par une ou plusieurs variables, ou caractères statistiques.

Chaque variable peut être, selon le cas :

1. Quantitative : ses valeurs sont des nombres exprimant une quantité, sur lesquels les opérations arithmétiques (somme, etc...) ont un sens.

   La variable peut alors être discrète ou continue selon la nature de l'ensemble des valeurs qu'elle est susceptible de prendre (valeurs isolées ou intervalle).

2. Qualitative : ses valeurs sont des modalités, ou catégories, exprimées sous forme littérale ou par un codage numérique sur lequel des opérations arithmétiques n'ont aucun sens.

Figure 2: The summary page from Lesson 1 of the module *Statistiques descriptives* (Descriptive Statistics).)

2.2 Second stage: applied linear models

After the first stage of the project was carried out, and after a decision was made to embark on a large project concerning applied linear models, consisting of the standard curriculum completed by methodologies for categorical data, like the logistic and log-linear models, reflection was given as to what format would be appropriate for more advanced learning.

The advanced learner of a given discipline, especially at the Cnam, has very different needs than the learner of the elements. More often than not, a first course in Probability-Statistics is mandatory. A second course is taken by those who feel a greater relevance of the material taught to their actual work. Hence a truer motivation. In any case, to get the attention of a student, any student, one has to pay heed to its needs, to speak his language.

In Applied Statistics, the actual practice requires the continuous use of a Statistics software on real data – real as opposed to simulated, with all the complexities then of reality –, and an important part of the work consists of careful questioning from the analyst and writing of the facts found during the process.

All this points to a pedagogy that rests principally on case studies, probably the natural points of entry to the curriculum for many students in the engineering and management sciences to whom this course is destined. Theory, the mathematical derivations – and they show a complexity far beyond that found in the elements –, are seen as answers to specific questioning on their part. Thus bigger and more mathematical chunks of material in more advanced studies, instead of the tidbits of the elements.

Another important point in our view of things, there should be a constant preoccupation from the designers of Internet courses, all courses for that matter, to instill into the students the art of questioning. We refer here also to Postman in his last essay ([4] p.161 *seq*): "(...) question-asking is the most significant intellectual skill available to human beings," and it is extremely strange that, especially in the Sciences whether hard or applied, it is not taught in schools!

Finally, and this also harks back to Postman in all his books on Education, we have written historical notes on all the principal aspects on the origin of the need of statistical models for reality. It is a fact that with History notes there is a sort of holographic phenomenon: even when one starts from hard sciences' bits of knowledge, exploring how things came to be, where ideas came from and how we came by them, provides, if propelled by a sense of questioning, an insight on the whole of societies, on all of Human nature. This constitutes an essential part for any formation. Education after all is not only about information, but first and foremost about the formation or casting of minds, young ones in particular.

In summary, due to the mathematical sophistication of this material there is a need for textbook typography, as well as, as usual, a need for a complete system of inner referencing and outer or hyper-referencing facilities. This leaves nowadays almost no choice: such a course must be written in Latex-Pdf typeset. The Pdf-files are virus-proof, they can be readily printed on paper with textbook color quality, their use on computer screens is very confortable, moreover providing some annotating facilities, and, finally, inner links and hyperlinks are manipulated with extreme ease.

This second stage of St@tNet is not, as yet, fully operational, but a demo-version ia available, and parts of the material, especially some case studies, were tested with great success in standard classrooms[4]. In the following pages we present some of its highlights.

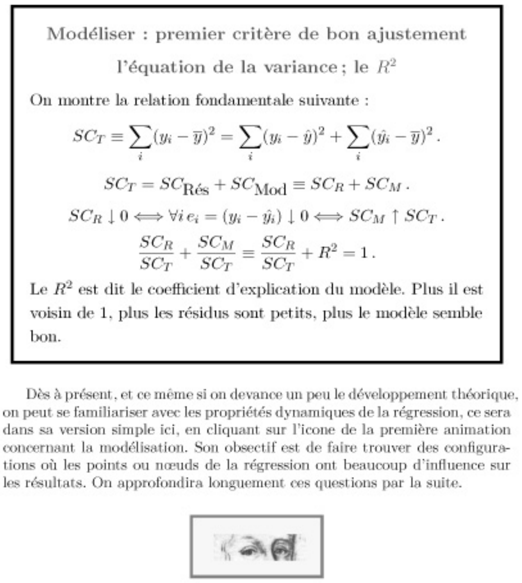
In Figure 3, we can see part of ordinary page of the course file. At the bottom of the page an icon referring to a Flash animation, an image of which appears on Figure 4.

The reader can flip back and forth from any page giving internal links to an equation, a table, a figure. He can also, if he subscribes to an Internet server, readily access a certain number of hyper-links to whatever sites deemed interesting by the authors. These pages will be added automatically at the end of the pdf-file file which can be saved with the added information. The Adobe-reader provides also various facilities to annotate the file pages.

Like in the first stage of St@tNet, many Flash animations are also included in the text. They constitute a remarkable tool to ease the learning. The development of a Flash animation is fairly easy, they are space efficient, and the Flash plug-in is very light and widespread. Furthermore, these animations are upward compatible, and can be readily updated.

Each one of our animations comes with a certain number of controllable buttons one of which is an audio file. In the example (Figure 4), the nodes of

[4] `http://www.mgi.polymtl.ca/marc.bourdeau/InfAgeTeaching` .



Modéliser : premier critère de bon ajustement

l'équation de la variance ; le R^2

On montre la relation fondamentale suivante :

$$SC_T \equiv \sum_i (y_i - \overline{y})^2 = \sum_i (y_i - \hat{y})^2 + \sum_i (\hat{y}_i - \overline{y})^2 .$$

$$SC_T = SC_{\text{Rés}} + SC_{\text{Mod}} \equiv SC_R + SC_M .$$

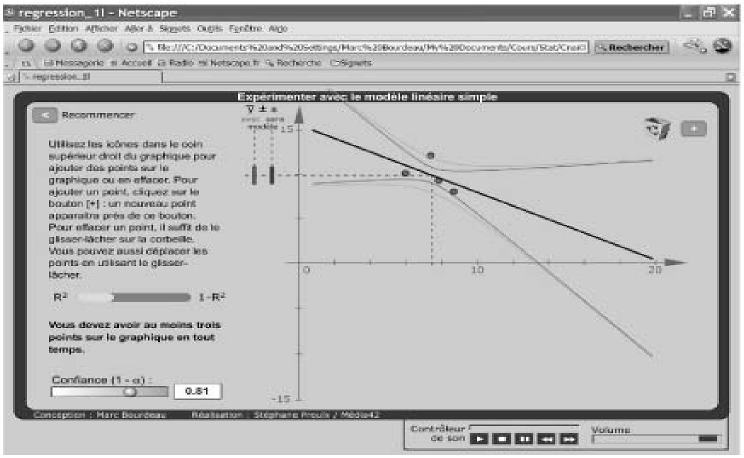
$$SC_R \downarrow 0 \Longleftrightarrow \forall i\, e_i = (y_i - \hat{y}_i) \downarrow 0 \Longleftrightarrow SC_M \uparrow SC_T .$$

$$\frac{SC_R}{SC_T} + \frac{SC_M}{SC_T} \equiv \frac{SC_R}{SC_T} + R^2 = 1 .$$

Le R^2 est dit le coefficient d'explication du modèle. Plus il est voisin de 1, plus les résidus sont petits, plus le modèle semble bon.

Dès à présent, et ce même si on devance un peu le développement théorique, on peut se familiariser avec les propriétés dynamiques de la régression, ce sera dans sa version simple ici, en cliquant sur l'icone de la première animation concernant la modélisation. Son obsectif est de faire trouver des configurations où les points ou nœuds de la régression ont beaucoup d'influence sur les résultats. On approfondira longuement ces questions par la suite.

Figure 3: Part of a typical page in stage two, with the icon referring to a Flash animation.



regression_1l - Netscape
Fichier Edition Afficher Aller à Signets Outils Fenêtre Aide
file:///C:/Documents%20and%20Settings/Marc%20Bourdeau/My%20Documents/Cours/Stat/Cnal
Recherche
Messagerie Accueil Radio Netscape.fr Recherche Signets
regression_1l
Expérimenter avec le modèle linéaire simple
Recommencer
$\overline{y} \pm s$
avec sans
modèle
15
0
10
20
-15
+
R²
1-R²
Confiance (1 - α) :
0.81
Conception : Marc Bourdeau
Réalisation : Stéphane Proulx / Média42
Contrôleur de son
Volume

Utilisez les icônes dans le coin supérieur droit du graphique pour ajouter des points sur le graphique ou en effacer. Pour ajouter un point, cliquez sur le bouton [+] : un nouveau point apparaitra près de ce bouton. Pour effacer un point, il suffit de le glisser-lâcher sur la corbeille. Vous pouvez aussi déplacer les points en utilisant le glisser-lâcher.

Vous devez avoir au moins trois points sur le graphique en tout temps.

Figure 4: A page from one of the course Flash animations, with its controllable buttons, one of which (bottom) is for an audio file.

the regression are mobile and new nodes can be added, the confidences bands resulting from the least squares results have a button to control their level, and whenever a change is made, directly with the mouse on the computer screen, the new regression line and confidence bands with the other numerical parameters promptly appear on the screen. The audio file provides instructions for the use of the animation, a few explanations, and always, this is very important, a questioning that the animation brings out.

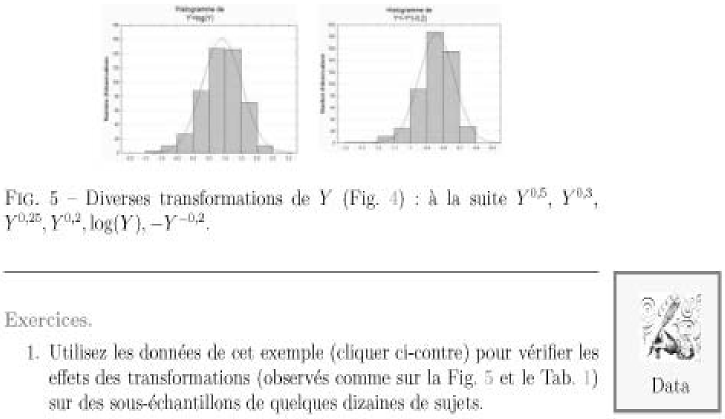


FIG. 5 – Diverses transformations de Y (Fig. 4) : à la suite $Y^{0,5}$, $Y^{0,3}$, $Y^{0,25}$, $Y^{0,2}$, $\log(Y)$, $-Y^{-0,2}$.

Exercices.

1. Utilisez les données de cet exemple (cliquer ci-contre) pour vérifier les effets des transformations (observés comme sur la Fig. 5 et le Tab. 1) sur des sous-échantillons de quelques dizaines de sujets.

Data

Figure 5: A typical page of a case study, with an icon to import the data.

In Figure 5, we show a typical page of a case study. Remember: case studies are the backbone of our pedagogy. A case study is usually several pages long and is built along a certain questioning on a data set that is usually quite complicated. Its flow is very progressive, and generally requires a few dozen hours of work with the writing of a roughly 20 page report.

This task imperatively implies team work. And true collaboration is necessary: a case study is not composed of a certain number of unrelated problems, like the standard homeworks found in most curricula, but has a synthetic character where each part responds or resonates with other ones. There is a wide landscape built in every case study, several chapters of Statistics are brought to bear. This precludes the "usual" split up of the work... The work required is similar to actual data analysis required by engineers and scientists, and ordinary work for that matter.

In a standard classroom, students often have a natural peer group and team formation is quite easy – though there are more optimal ways for this selection –, but for distance learning things are not that easy. However, in most organizations nowadays, the new ITs (chats, forums, etc.) already allow team virtual meeting and working that take up a very large, if not the larger part of the work process!

Students tend to use all the modern hyper-communicating computer facilities that are usually available on all recent computers, as well as within the VTEs in use in most universities. In most North American universities students are in constant Internet contact with their peer-groups, almost day and night, sending each other files of their works, of their thoughts, comments on the courses, etc. And Internet real time voice communication facilities are rapidly spreading. Writing, however, thanks to the Internet, has regained much luster: writing constitutes indeed an essential tool for the unveiling of one's real thoughts. Team work is constant. On a less bright side, homeworks and exams tend to be freely accessible to all...

Professors must adapt to this situation. The Internet has and will profoundly change the learning world. Thus ITs can compensate the isolation of students of the past in distance learning, as they have already done so in the traditional classroom. But our type of pedagogy implies a much greater contact not only horizontally, from students to students as we just noted, but also vertically. For professors of Statistics, it is not the transmission of bits of knowledge that will constitute their main task, it is the statistical cast of mind itself that will be more and more the focal point of teaching. And statistical thinking, like all casts of mind, is best transferred through apprenticeship.

3 Conclusion

The Information Age offers mind-boggling perspectives and cannot but have a profound impact on the pedagogy of whatever discipline there is, but first and foremost for those that present a technical character, and Statistics is one of them. All the presentations in this session will no doubt show the diversity of options.

The end of the journey for teachers? At first sight, it might appear that all these new facilities lead to the disappearance of teachers and professors. But many very successful pedagogical experiments have shown that human pedagogical guides are more necessary than ever, and that ITs provide an indispensable structure for more interactions between them and the students. Il would not be surprising that the new pedagogical paradigm would be that of apprentices and masters. In all the pedagogical experiences we have seen, not only in Statistics, not only our own, there is a greater need than ever for human personal transmission. The role of professors becomes more and more that of a *personne ressource*, a guide so to speak, and less and less that of a knowledge dispenser. Pure knowledge transmission is not the principal role of professors anymore: this has now been more or less automated thanks to the new ITs. Transmission is required now at a much higher cognitive level. And written words for their precision, as well as oral contacts, play — ITs in the background again! — a crucial role. On the Internet, all courses tend to become tutorials! And this is the expensive form of teaching... That may explain why the pedagogical interaction has become so much more demanding than ever.

We do not imply however that teaching becomes a student driven process for the development of curricula and material. But a clear result of our experience with teaching case-based Statistics courses to engineers with an active pedagogy approach, amply confirms Parr's [3] and Moore's [2] experiences: we obtain a much more efficient knowledge transmission, as well as a more positive attitude towards the discipline, than what we observed through years of teaching with the traditional approach.

On the other hand, many governments nowadays tend easily to believe that education and other public services are not of primary importance and cost too much. For reasons of globalization and so forth, they preside over decreasing public spending. The Internet Age can readily provide very low quality and very low cost formative material – garbage-in, garbage-out –, as well as higher than ever quality education. The latter being the kind needed in an increasingly complex world. But the wheel of Fortune spins faster than ever, and, it has always been the case, the outcomes are not totally random: the better educated no doubt will reap the profits.

The question of what's in store for pedagogues in the future will in any case be with us for some time.

What about St@tNet's journey? The conception, development and implementation of St@tNet required considerable resources, human as well as financial. The end product could constitute a complete curriculum in French for Applied Statistics.

The first stage was conceived with a playful spirit in mind, to which the elementary concepts of Probability and Statistics lend themselves fairly easily. But putting it into service required a considerable amount of work, so much more than the writing of a standard chalk and blackboard course, or of a set of telegraphic computer slides. The second stage is much more difficult to conceive if one does not care for a standard run of the mill product, but strives after something more pedagogically efficient. The deeper one goes into the discipline, the more difficult the task.

The question is not whether or not there will be a need for St@tNet or its successors in the future, but what form they will take, and what resources will it be necessary to put into action? A knowledge-based society will indeed bring no dearth of work for statisticians and teachers of the discipline (cf. the 6th European research program FP6). However, at the same time that technology's pace shows no sign of slowing down and that the demand is growing rapidly, the human and financial resources might become more difficult to muster... International cooperation and sharing of the new IT products catering to the needs of students of Statistics as well as greater imagination and dedication on the part of teachers of Statistics will no doubt be necessary.

References

[1] Bourdeau M. (2003). *L'enseignement supérieur et les TICE (Technologies de l'information et des communications en enseignement). Big bang ou méga flop ?* Conférence pour l'inauguration de la mission TICE, Université de Bretagne Sud, 4 mars 2003. `http://www.mgi.polymtl.ca/marc.bourdeau/InfAgeTeaching`.

[2] Moore David S. (1997). *New pedagogy and new content: The case of statistics.* International Statistical Review, **65**(2), 123 – 137.

[3] Parr William C. & Smith Marlene A. (1998). *Developing case-based business statistics courses.* The American Statistician, **52**(4), 330 – 337.

[4] Postman N. (1999). *Building a bridge to the 18th century. How the past can improve our future.* Alfred A. Knopf, New York.

[5] Postman N. (1995). *The end of education. Redefining the value of school.* Alfred A. Knopf, New York.

[6] Saporta G. (2002). *St@tNet, an Internet based software for teaching introductory statistics.* Proceedings Icots 6, Sixth International Conference on Teaching Statistics, Capetown, July 8-12 2002.

[7] Saporta G., Morin A. (1996). *Interactive software for learning statistics.* Compstat 96, Barcelona, 26-30 August 1996.

[8] Serban A.N., Luan J. (2002). *Overview of knowledge management.* New Directions for Institutional Research, **2002**(113), 5 – 12.

[9] Velleman Paul F., Moore David S. (1996). *Multimedia for teaching statistics: promises and pitfalls.* The American Statistician, **50**(3), 217 – 225.

Acknowledgement: St@tNet was developed with the generous support of the *Agence Universitaire pour la Francophonie* and of the French *Ministère de l'Éducation Nationale.*

Deep gratitude to all those working on the project:
`http://www.mgi.polymtl.ca/marc.bourdeau/InfAgeTeaching/credits.pdf`.

Address: G. Saporta, Conservatoire National des Arts et Métiers, 292, rue Saint-Martin, F-75003 Paris, France, `http://cedric.cnam.fr/~saporta`
M. Bourdeau 'Ecole Polytechnique de Montréal,
`http://www.mgi.polymtl.ca/marc.bourdeau`

E-mail: `Saporta@cnam.fr, Marc.Bourdeau@polymtl.ca`

AN AUTOMATIC THRESHOLDING APPROACH TO GENE EXPRESSION ANALYSIS

Michael G. Schimek and Wolfgang Schmidt

Key words: Empirical Bayes, microarray, multiple testing, R, sparse sequence, statistical computing, threshold.

COMPSTAT 2004 section: Bayesian methods.

Abstract: The statistical problems of gene expression analysis based on the two popular array readout methods, cDNA and Affymetrix, are addressed. As an alternative to multiple frequentist statistical testing the empirical Bayes methodology is introduced. An empirical Bayes thresholding approach is described and its relevance for microarray data analysis is shown. Finally two data sets, one of cDNA-type and the other of Affymetrix-type, are analyzed with the new automatic and computationally efficient thresholding technique.

1 Introduction

In recent years the new technology of microarrays has made it feasible to measure expression of thousands of genes to identify changes between different biological states. Statisticians are requested to design methods which help to quantify the relevance of these experimentally obtained changes.

In such biological experiments (for an introduction see [12]) we are confronted with the problem of high-dimensionality because of thousands of genes involved and at the same time with small sample sizes (due to limited availability of cases and for reasons of cost). This makes it a statistically and computationally demanding task. The complexity of the diseases, the poor understanding of the underlying biology and the imperfection of the measurements (many different sources of noise) are additional problems.

There are two dominating DNA array readout methods, cDNA ([3], p. 17ff, [12]) and Affymetrix GeneChips ([2], [12]). In the first the data are read from a fluorescent signal and in the last the data are recorded from a radioactive signal. The statistical method described in this paper can be applied in both instances.

At first we portray popular techniques for gene expression analysis. Then the idea of empirical Bayes methods is introduced and an empirical Bayes thresholding (EBT) approach is described in some detail. Its practical relevance is demonstrated for colon data from one of our laboratories (cDNA) and for the so-called Golub data set (Affymetrix) from [8].

2 Fold change and classical inference methods

Let us assume that for each of n genes $(i = 1, \ldots, n)$ we have measurements over J experimental conditions $(j = 1, \ldots, J)$ on K slides (arrays) per experiment $(k = 1, \ldots, K)$. These measurements may be intensity readings from a spotted cDNA or probe-level intensity signals from an Affymetrix oligonucleotide system. The expression data are either log-ratios of intensities or log-intensities.

For each gene the fundamental question is whether the level of expression is substantially different between two $(J = 2)$ or more $(J > 2)$ situations. One approach commonly used in early publications (e.g. [13]) - always limited to control versus treatment designs - has been a simple fold approach. This means that a gene is labeled significantly changed if its average expression level varies by more than a constant factor, typically two, between the two experimental conditions, the so-called "twofold rule". This ad hoc approach has one severe disadvantage: a factor of two does not mean the same in different regions of the spectrum of intensities, especially when extreme values are concerned.

The standard statistical approach taken is significance testing which implies that fold change is replaced by significance. The null hypothesis for each gene is that the data we observe have some common distributional parameter among the conditions, usually the mean of the expression levels. For each gene we calculate a statistic that is a function of the data. For instance a t-test statistic is applied (despite an unrealistic distributional assumption). Yet there is not much gain compared to the fold approach because the difference between two logarithmic expression levels is the logarithm of their ratio.

Which errors are committed at each particular gene when testing for differential expression? Apart from the type I error (false positive) and the type II error (false negative) there is the complication of testing multiple hypotheses simultaneously. Each gene has individual type I and II errors and it is nothing but clear how to measure the overall error rates. In the recent literature several compound error measures have been suggested, such as the false discovery rate ([4]) and the positive false discovery rate ([14]). However their calculation is not straight forward and the selection of a test statistic in this situation is far from trivial.

Suppose n genes $(i = 1, \ldots, n)$ have been measured over two experimental conditions $(j = 1, 2)$ on K_1 arrays of condition 1 and K_2 arrays of condition 2 and $K_1 + K_2 = K$. Let $\bar{x}_{i1}$ and $\bar{x}_{i2}$ be the mean gene expression for gene i under conditions 1 and 2, and let s_i be the pooled standard deviation for gene i,

$$s_i = \sqrt{\left(\frac{1}{K_1} + \frac{1}{K_2}\right) \frac{\sum_{k_1=1}^{K_1} (x_{ik_1} - \bar{x}_{i1})^2 + \sum_{k_2=1}^{K_2} (x_{ik_2} - \bar{x}_{i2})^2}{K - 2}}.$$

As pointed out already a test statistic for assessing differential gene expression is the standard t-test

$$t_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{s_i}.$$

Alternatively a rank-sum statistic can be adopted. Suppose r_{ik} be the rank of the kth expression level within gene i. Then the rank-sum statistic for gene i is

$$r_i = \sum_{k_1=1}^{K_1} r_{ik_1}.$$

An extreme r_i value in either direction would indicate a difference in gene expression. The t-statistic as introduced above tests for difference in the mean whereas the rank statistic tests for difference in distribution.

Under the assumption of only two experimental conditions and no correlation between measurements it is possible to derive the null distribution: either permutation ([17]) or bootstrap ([6]) techniques can be applied. In both cases the computational demand is quite high.

If the null distribution is calculated individually for each gene, this has two disadvantages. The first is what is known as granularity problem: the null distribution has a resolution on the order of the number of permutations. With n genes and m permutations the resolution is on the order of $1/m$ for individual null distributions, but $1/(nm)$ for a pooled null distribution. For instance, if we test 3000 genes with 100 permutations, then we can expect to reject 30 at a time. The second problem is that we are not in the position to construct better rejection regions. With individual null distributions, each gene is treated as a different experiment. For each "experiment" we have m observations from the null distribution and one from the original measurements. It is not possible to compare the null distribution to the observed statistic to derive more powerful, asymmetric rejection regions ([15], p. 277f). This means loss of power.

In the SAM ("Significance Analysis of Microarray") method ([16]) another approach has been taken: there the test statistics are pooled and considered to follow a mixture distribution. As a consequence, many observations from the mixture of the null and affected distributions, as well as from the pure null distribution are available, leading to improved rejection regions. The pitfall of using different distributions for the estimation of the overall error rate (due to pooling of the null statistics) is sufficiently controlled in SAM according to its authors. In SAM expression is evaluated by a combination of test and thresholding steps for the purpose of non-symmetric rejection regions. This approach improves the decision process when the numbers of overexpressed and underexpressed genes are substantially different (usually the case in practice). The cutoff for test significance is tuned via a user-specified parameter connected to the false discovery rate (the number of false positives is limited this way). Hence SAM is not an automatic approach.

3 Empirical Bayes methods

Empirical Bayes methods have been around in statistics for thirty years, beginning with [5]. One cannot say that they have enjoyed much attention so far. Why are they attractive for gene expression analysis? Empirical Bayes methods are well-suited for high-dimensional decision problems. In contrast to techniques as discussed above, where inference is performed separately for different genes, in empirical Bayes information among genes is shared. In most microarray experiments involving thousands of genes but only a small number of microarrays, the amount of information per gene is quite low. The idea is to combine related inference problems which means that the evaluation of the expression level of one gene is influenced by the overall expression levels. Here advantage is taken of the quantification of the variability characterizing the bulk of genes (still assuming independent measurements).

The general framework of empirical Bayes methods is quite flexible. Probability distributions are specified in several layers that account for multiple sources of variation. Then based on a mixture model posterior probabilities are computed. This allows comparison among multiple conditions.

Empirical Bayes methods have been applied for the first time to gene expression analysis in [11] and [7]. In these papers the concepts of fold change respectively significance with respect to the frequentist false discovery rate are generalized. In the next section we consider an empirical Bayes thresholding approach that does not require a concept of fold change or significance.

3.1 A fully automatic thresholding approach

Here we describe an empirical Bayes approach for the estimation of possibly sparse sequences observed with white noise (modest correlation is tolerable). A sparse sequence consists of a relatively small number of informative measurements (in which the signal component is dominating) and a very large number of noisy zero measurements. This is the typical situation found in image processing. Gene expression profiling can be seen along the same lines.

Johnstone and Silverman (2004) have proposed a method that can handle such sparse sequences by means of thresholding without any users-specified parameters apart from distributional assumptions ([10]). It is called empirical Bayes thresholding (EBT). As will be seen later on, the choice of the threshold is the most critical aspect, both in terms of signal extraction and computational burden.

In empirical Bayes threshold models the object of interest is a sequence of parameters θ_i on each of which we have a single observation X_i subject to some noise ϵ_i, such that

$$X_i = \theta_i + \epsilon_i.$$

For the estimation of the θ_is additional assumptions are required. In [10] the θ_is are assumed to be medians and that the observation $X = (X_1, \ldots, X_n)$ satisfies

$$X_i = \mu_i + \epsilon_i,$$

where the ϵ_is are $N(0, \sigma^2)$ random variables, not too highly correlated. Further let $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ be a vector of medians (means are also feasible but not of interest in this paper).

Obviously the μ_is will not be exactly zero in most applications. The p-norm of μ, $\|\mu\|_p = (\sum |\mu_i|^p)^{1/p}$ allows for a more subtle characterization of sparsity of μ (assuming small p). In other words, the quantification of sparsity corresponds to bounds on the p-norm of μ for $p > 0$. Consider the sum of squares of a vector with $\|\mu\|_p = 1$ for some small p. If only one of the components of μ is nonzero, then the energy will be 1. If on the other hand, all of the components are equal, then the energy will be $n^{1-2/p}$ and is tending to zero as $n \to \infty$ if $p < 2$, tending rapidly to zero if p is near zero. Consider the case of p small. Then the only way for a signal in an l_p ball with small p to have large energy (sum of squares) is to consist of a few large components, as opposed to many small components of roughly equal magnitude. Among all signals with a given energy, the sparse ones are those with small l_p norm.

Some measure of sparsity is needed because sparsity of a signal is not solely a matter of the proportion of μ_i that are zero or near zero, but also of subtle ways in which the energy of the signal μ is distributed among the various components. For our purposes it is sufficient that the number of indices i for which μ_i is nonzero is bounded. In engineering such a parameter μ is called a "nearly black signal". For some η this is

$$l_0\,[\eta] = \left\{ \mu : \frac{1}{n} \sum_{i=1}^{n} I\,[\mu_i \neq 0] \leq \eta \right\}, \tag{1}$$

where I denotes an indicator function. Assuming the signal is sparse in the sense of belonging to an l_p norm ball of small radius η, we have

$$l_p\,[\eta] = \left\{ \mu : \frac{1}{n} \sum_{i=1}^{n} |\mu_i|^p \leq \eta^p \right\}. \tag{2}$$

For (1) and (2) it is possible to derive minimax squared error properties. It can be shown that EBT adapts automatically to the degree and character of sparsity of the signal with the minimax rate (i.e. the optimum rate for such signals; for details see [10]). It is worth mentioning that the minimax properties are the same as in the false discovery rate approach in [1].

Suppose the errors ϵ_i are independent. Within the Bayesian context sparsity is equivalent to suitable prior distributions for the θ_is we are interested in. The notion that many or most of the θ_is are near zero is captured by assuming that the elements θ_i have independent prior distributions each given by the mixture

$$f_{prior}(\theta) = (1 - \omega)\delta_0(\theta) + \omega\gamma(\theta). \tag{3}$$

The nonzero part of the prior, γ, is assumed to be a fixed unimodal symmetric density. γ is traditionally assumed to be a normal density, Here ([10]) it is recommend to use a heavier-tailed prior. For the mixing prior in (3) it is favorable to use for γ the Laplace density with scale parameter $a > 0$

$$\gamma_a(u) = \frac{1}{2} a \exp(-a\,|u|)$$

or the mixture density

$$(\mu|\Theta = \theta) \sim N(0, \theta^{-1} - 1) \text{ with } \Theta \sim \text{ Beta}\,(\alpha, 1)\,.$$

The mixture density for μ has tails that decay as $\mu^{-2\alpha-1}$. For $\alpha = \frac{1}{2}$ the tails have the same weight as those of the Cauchy distribution.

In both cases the posterior distribution of μ given an observed X, and the marginal distribution of X, are tractable. This makes it feasible to adopt marginal maximum likelihood for the ω selection as well as to estimate μ by the posterior median.

Further assumptions required for the nonzero part of the prior γ are (i) a fixed unimodal symmetric density, (ii) tails to be exponential or heavier, and (iii) a mild regularity condition.

The key feature of this empirical Bayes approach is the threshold. If the absolute value of a particular X_i exceeds some threshold t then it is taken to correspond to a nonzero μ_i, estimated simply by X_i itself. Otherwise the coefficient μ_i is estimated zero. The problem is, that the threshold t (or rather t_is) needs to be tuned to the sparsity of the signal. If a threshold appropriate for dense singnals is applied to a sparse signal, or vice versa, the result is of no use at all. Hence a good threshold selection method needs (i) to be adaptive between sparse and dense signals, (ii) to be stable to small changes in the data, and (iii) to be tractable to compute. The approach in [10] comprises all these properties.

Let us now discuss the choice of the mixing weight ω, or equivalently, of the threshold $t(\omega)$. Assume that the X_i are independent. For any value of the weight ω consider the posterior distribution of μ given $X = x$ under the assumption that $X \sim N(\mu, \sigma^2)$. Let $\mu(x;\omega)$ be the median of this distribution. For fixed $\omega < 1$, $\mu(x;\omega)$ is a monotonic function of x with the following threshold property

$$\exists t(\omega) > 0 \text{ such that } \mu(x;\omega) = 0 \Leftrightarrow |x| \leq t(\omega).$$

Let $g = \gamma * \phi$ denote the convolution of the density γ with the standard normal density ϕ. The marginal density of the observations X_i is then

$$(1 - \omega)\phi(x) - \omega g(x).$$

The marginal maximum likelihood estimator $\tilde{\omega}$ of ω is defined as the maximizer of the marginal log-likelihood

$$l(\omega) = \sum_{i=1}^{n} log \left\{(1 - \omega)\,\phi\,(X_i) + \omega g\,(X_i)\right\},$$

subject to the constraint on ω that the threshold satisfies $t(\omega) \leq \sqrt{2\log n}$ (the threshold takes values from 0 to $\sqrt{2\log n}$).

What is the posterior probability that μ is nonzero? Let us define

$$\beta(x) = \frac{g(x)}{\phi(x)} - 1. \tag{4}$$

Then the posterior probability $\omega_{post}(x) = P(\mu \neq 0|X = x)$ will satisfy

$$\omega_{post}(x) = \frac{\omega g(x)}{\omega g(x) + (1-\omega)\phi(x)} = \frac{1+\beta(x)}{\omega^{-1} + \beta(x)}$$

As a result it can be found using function (4) alone.

To find the posterior median $\hat{\mu}(x;\omega)$ of μ given $X = x > 0$, we need the cumulative distribution

$$\tilde{F}_1(\mu|x) = \int_{\mu}^{\infty} f_1(u|x)du,$$

where f_1 is a density. If $x > 0$, we can find $\hat{\mu}(x;\omega)$ via the following properties:

$$\begin{array}{ll} \widehat{\mu}(x;\omega) = 0 & \text{if } \omega_{post}\tilde{F}_1(0|x) < \frac{1}{2} \\ \tilde{F}_1(\widehat{\mu}(x;\omega)|x) = (2\omega_{post}(x))^{-1} & \text{otherwise.} \end{array}$$

For $\omega_{post}(x) \leq \frac{1}{2}$ the median is necessarily zero (no need to evaluate $\tilde{F}_1(0|x)$). For $x < 0$ the antisymmetric property $\hat{\mu}(-x,\omega) = -\hat{\mu}(x,\omega)$ can be used.

The Bayes factor threshold is related to the posterior median. It is a value $\tau(\omega)$ such that $P(\mu > 0|X = \tau(\omega)) = 0.5$. This is to say that $\tau(\omega)$ is the largest value of the sequence for which the estimated μ will be zero, if the estimate is obtained from the posterior median.

How can we find the estimate $\tilde{\omega}$ of ω or the scale parameter a of the Laplace density? Maximization of the marginal maximum likelihood l gives the solution. Let us define the score function $S(\omega) = l'(\omega)$. Because of smoothness and monotonicity of $S(\omega)$ it is possible to find the estimates by a binary search, or an even faster algorithm. The obtained values are then plugged back into the prior and the parameters μ_i are evaluated via these estimates, either by using the posterior median itself, or by using some other threshold rule with the same threshold $t(\omega)$.

The threshold is obtained from the posterior median $\hat{\mu}$, mainly by use of the following properties:

(i) shrinkage rule: $0 \leq \widehat{\mu} \leq x$ for $x \geq 0$

(ii) threshold rule: there exists $t(x) > 0$ such that $\widehat{\mu}(x) = 0$ if and only if $|x| \leq t(\omega)$

(iii) bounded shrinkage: there exists a constant b such that for all ω and x $|\widehat{\mu}(x;\omega) - x| \leq t(x) + b$.

This approach is quite unique in combining features of excellent theoretical properties and efficient computation. According to [10] the results

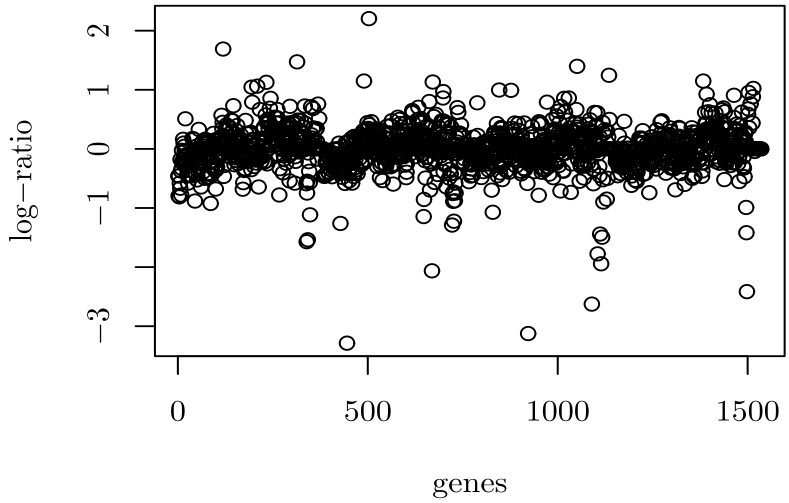


Figure 1: Colon data after preprocessing.

proven for white noise errors still hold for modestly correlated errors, at least in an approximate sense. This generalization is important for microarray applications because some of the measurements are usually replicates. The EBT approach was implemented in the R language ([9]) by Iain M. Johnstone and Bernard W. Silverman. The master function of the EBT algorithm is *ebayesthresh*(). This function as well as the others required to analyze sparse sequences can be downloaded freely for academic purposes from `http://www.stats.ox.ac.uk/∼silverma/ebayesthresh/`. Relevant documentation is found there too. After having been sourced to R, the EBT algorithm can be used as any other function in R.

4 Two examples

Our first example uses cDNA measurements and our second example is based on Affymetrix measurements. In both techniques we have many different sources of noise such as variation in hybridization time, variation in reagent concentrations, leak of external light during chip reading, inhomogeneities in chip preparation, variations in laser intensity during chip reading, trace contamination with cross-hybridizing oligonucleotides, etc. EBT is an ideal method to handle a decision problem under sparsity due to substantial noise.

4.1 The colon data set

The data are of cDNA-type from a colon carcinoma experiment which encompasses a set of 13 colon carcinoma patients. None of these patients was treated neoadjuvant. The invasionfront of the tumors was investi-

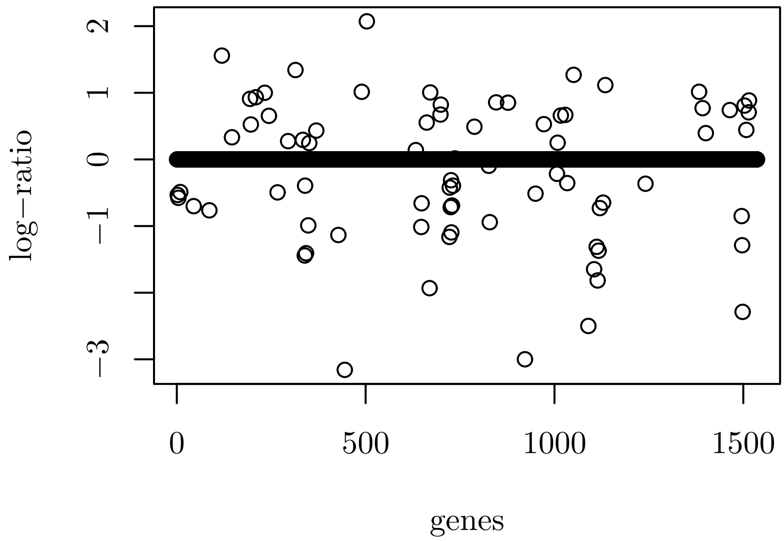


Figure 2: EBT result for colon data.

gated and hybridization was made against a pool of 4 probes of normal colon tissue. Standard protocol was used and quality control ensured by the Institute of Pathology, Medical University of Graz. The experiments contain $n = 1536$ different genes, among them some replicates. The cDNA chips where then scanned with microarray image analysis software Imagene from BioDiscovery producing two text files. These files where then imported into R ([9]) using the object-oriented microarray analysis library *com.braju.sma* (can be obtained from the University of California at Berkeley, `http://www.maths.lth.se/help/R/com.braju.sma/`) The following preprocessing steps were applied: (i) background subtraction, (ii) transformation into $M = \log_2 (Red/Green)$ (M is further referred to as *log-ratio*), (iii) normalizing within slide using scaled print-tip method, (iv) few values which were not detected on a subset of slides, hence NAs, were set to zero in order to allow further processing, and (v) the experiments where merged using the median. The data after preprocessing are displayed in Fig. 1.

The EBT algorithm was applied using following parameters: $prior =$ "*laplace*" and $a = NA$, so that the scale parameter a is estimated by marginal maximum likelihood. $bayesfac = T$ means that whenever a threshold is explicitly calculated, the Bayes factor threshold will be used. Having $sdev = NA$, the standard deviation is estimated via the median absolute deviation from zero ($mad(x, center = 0)$). Finally, with $threshrule =$ "*median*" the posterior median is chosen.

Fig. 2 shows the genes from the colon data that are informative after administering the EBT algorithm. Finally we obtained $n_1 = 37$ overexpressed and $n_2 = 39$ underexpressed genes. These could be verified by pathologists.

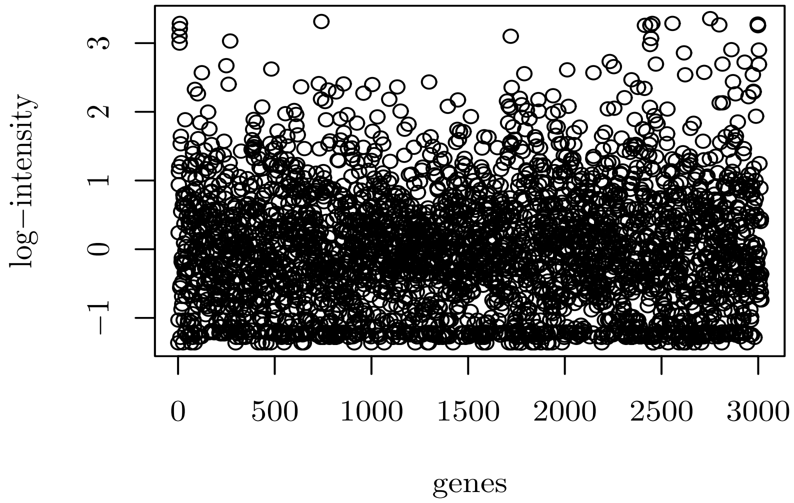


Figure 3: Subclass ALL of Golub data after preprocessing.

4.2 The Golub data set

The Golub data set ([8]) is well-known and has been re-analyzed by many authors. The data originate from a gene expression study with patients suffering from two types of acute leukemia. Here we consider only a subset of it (i.e. the learning set) with data from 27 acute lymphoblastic leukemia (ALL) patients and 11 acute myeloid leukemia (AML) patients. The intensities were measured using Affymetrix high-density oligonucleotide chips. The data comprise the expression of $n = 6817$ human genes and can be obtained from the Web site `http://www-genome.wi.mit.edu/mpr/data_set_ALL_AML.html`.

The following preprocessing steps were administered using R functions: (i) thresholding of the expressions (floor of 100 and ceiling of 16000), (ii) filtering by excluding genes with expressions $max/min \leq 5$ or$(max - min) \leq 500$, (iii) a base 10 logarithmic transformation, (iv) scaling the matrix using the R command *scale* from the package *base*, which is a generic function whose default method centers and/or scales the columns of a numeric matrix, and (v) merging the experiments by subclass ALL or AML respectively. Due to the preprocessing an accumulation at a minimum value was observed that had to be eliminated in order to comply with the requirements of the EBT algorithm. This bottom line was eliminated by removing the respective gene during the processing step. It is re-inserted at the preceding index location of the result matrix with a value of zero after the transformation. This filtering omits 35 genes for the merged ALL data and 117 genes for the merged AML data. For the subclass ALL the preprocessed data are displayed in Fig. 3.

Applying the EBT algorithm with the same parameters as specified for

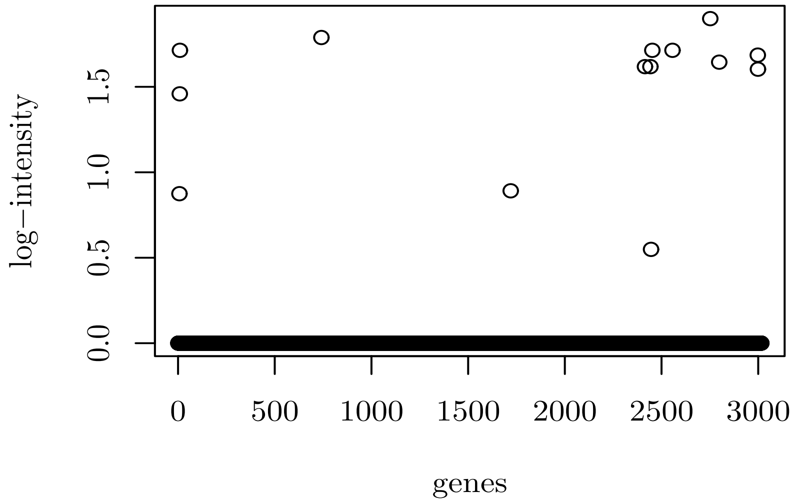


Figure 4: EBT result for subclass ALL of Golub data.

the above colon data cDNA experiment, 2 genes are detected for ALL, 1 gene is detected for AML and 12 genes are expressed in both subclasses (for a plot of the all together 14 overexpressed genes of the ALL subclass see Fig. 4). This is a substantial reduction in the number of informative genes. Whether the identified genes that are overexpressed in one subclass while not in the other (i.e. having a zero estimate) – obviously of discriminative power – are of high biological relevance needs to be answered by future leukemia research and is not a statistical matter.

References

[1] Abramovich F., Benjamini Y., Donoho D.L., Johnstone I.M. (2002). *Adapting to unknown sparsity by controlling the false discovery rate.* Preprint.

[2] Affymetrix (1999). *Affymetrix microarray suite user guide.* Affymetrix, Santa Clara, CA.

[3] Baldi P., Hatfield G.W. (2002). *DNA microarrays and gene expression. From experiments to data analysis and modeling.* Cambridge University Press, Cambridge.

[4] Benjamini Y., Hochberg Y. (1995). *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* J. Royal Statist. Soc., **B 85**, 289 – 300.

[5] Efron B., Morris C. (1973). *Combining possibly related estimation problems* (with discussion). J. Royal Statist. Soc. **B 35**, 379 – 421.

[6] Efron B., Tishirani R.J. (1993). *An introduction to the Bootstrap.* Chapman & Hall, London.

[7] Efron B., Tishirani R.J., Storey J.D., Tusher V. (2001). *Empirical Bayes analysis of a microarray experiment.* J. Amer. Statist. Assoc. **96**, 1151–1160.

[8] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S. (1999). *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science **286**, 531–537.

[9] Ihaka R., Gentleman R. (1996). *R: A language for data analysis and graphics.* J. Computat. Graph. Statist. **5**, 299–314.

[10] Johnstone I.M., Silverman B.W. (2004). *Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences.* To appear in Annal. Statist.

[11] Newton M.A., Kendziorski C.M., Richmond C.S., Blattner F.R. (2001). *On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data.* J. Computat. Biol. **8**, 37–52.

[12] Nguyen D.V., Arpat A.B., Wang N., Carroll R.J. (2002). *DNA microarray experiments: Biological and technological aspects.* Biometrics **58**, 701–717.

[13] Schena A.M., Shalon D., Davis R.W., Brown P.O. (1995). *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science **270**, 467–470.

[14] Storey J. D. (2002). *A direct approach to false discovery rates.* J. Royal Statist. Soc. **B 64**, 479–498.

[15] Storey J.D., Tibshirani R. (2003). *SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays.* In Parmigiani G., Garrett E.S., Irizarry R.A., Zeger S.L. (ed.) The analysis of gene expression data. Methods and software. Springer-Verlag, New York, 272–290.

[16] Tusher V., Tibshirani R., Chu C. (2001). *Significance analysis of microarray applied to transcriptional responses to ionizing radiation.* Proceedings of the National Academy of Sciences **98**, 5116–5121.

[17] Westfall P.H., Young S.S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment.* Wiley, New York.

Acknowledgement: Thanks to K. Wagner for the cDNA hybridisation of the colon data and to Dr. M. Asslaber for making the data set available to us; both from Institute for Pathology, Medical University of Graz, Austria (Lore Saldow Project and Gen-AU). The second author acknowledges funding from Gen-AU (”A Comprehensive Disease Bank for Functional Genomics”).

Address: Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2, A-8036 Graz, Austria

E-mail: `michael.schimek@meduni-graz.at`

KERNEL METHODS FOR MANIFOLD ESTIMATION

Bernhard Schölkopf

Key words: Kernel methods, support vector machines, quantile estimation.

COMPSTAT 2004 section: Neural networks and machine learning.

Abstract: We describe methods for estimating manifolds in high-dimensional spacs. They work by mapping the data into a reproducing kernel Hilbert space and then determining regions in terms of hyperplanes.

1 Kernel algorithms for pattern recognition

Suppose we are given empirical data

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}. \tag{1}$$

Here, the domain $\mathcal{X}$ is some nonempty set that the inputs x_i are taken from; the $y_i \in \mathcal{Y}$ are called *targets*. Here and below, $i, j = 1, \ldots, m$.

We have made no assumptions on the domain $\mathcal{X}$ other than it being a set. In order to study the problem of learning, we need additional structure. In learning, we want to be able to *generalize* to unseen data points. In the case of pattern recognition, given some new input $x \in \mathcal{X}$, we want to predict the corresponding $y \in \{\pm 1\}$. Loosely speaking, we want to choose y such that (x, y) is *similar* to the training examples. To this end, we need similarity measures in $\mathcal{X}$ and in $\{\pm 1\}$. The latter is easier, as two target values can only be identical or different.[1] For the former, we require a similarity measure

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \qquad (x, x') \mapsto k(x, x') \tag{2}$$

with the property that there exists a map Φ into a Hilbert space $\mathcal{H}$ such that for all $x, x' \in \mathcal{X}$,

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle . \tag{3}$$

Such a function k is called a *positive definite kernel* [1], [10], [8], $\mathcal{H}$ is the *reproducing kernel Hilbert space (RKHS)* associated with it, and Φ is called its *feature map*. A popular example, in the case where $\mathcal{X}$ is a normed space, is the Gaussian

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\,\sigma^2}\right), \tag{4}$$

where $\sigma > 0$.

[1] In the case where the outputs are taken from a general set $\mathcal{Y}$, the situation is more complex, cf. [11].

The advantage of using a positive definite kernel as a similarity measure is that it allows us to construct algorithms in Hilbert spaces. For instance, consider the following simple classification algorithm, where $\mathcal{Y} = \{\pm 1\}$. The idea is to compute the means of the two classes in the RKHS, $\mathbf{c}_1 = \frac{1}{m_1}\sum_{\{i:y_i=+1\}}\Phi(x_i)$, and $\mathbf{c}_2 = \frac{1}{m_2}\sum_{\{i:y_i=-1\}}\Phi(x_i)$, where m_1 and m_2 are the number of examples with positive and negative target values, respectively. We then assign a new point $\Phi(x)$ to the class whose mean is closer to it. This can be shown [8] to lead to

$$y = \operatorname{sgn}\left(\langle\Phi(x),\mathbf{c}_1\rangle - \langle\Phi(x),\mathbf{c}_2\rangle + b\right) \tag{5}$$

with $b = \frac{1}{2}\left(\|\mathbf{c}_2\|^2 - \|\mathbf{c}_1\|^2\right)$. Rewritten in terms of k, this reads

$$y = \operatorname{sgn}\left(\frac{1}{m_1}\sum_{\{i:y_i=+1\}} k(x,x_i) - \frac{1}{m_2}\sum_{\{i:y_i=-1\}} k(x,x_i) + b\right) \tag{6}$$

and $b = \frac{1}{2}\left(\frac{1}{m_2^2}\sum_{\{(i,j):y_i=y_j=-1\}} k(x_i,x_j) - \frac{1}{m_1^2}\sum_{\{(i,j):y_i=y_j=+1\}} k(x_i,x_j)\right)$.

Let us consider one well-known special case of this type of classifier. Assume that the class means have the same distance to the origin (hence $b = 0$), and that k can be viewed as a density, i.e., it is positive and has integral 1 (assuming the integral exists), $\int_X k(x,x')dx = 1$ for all $x' \in \mathcal{X}$. Then (6) corresponds to the Bayes decision boundary separating the two classes, subject to the assumption that the two classes are equally likely and were generated from two probability distributions that are correctly estimated by the Parzen windows estimators of the two classes,

$$p_1(x) := \frac{1}{m_1}\sum_{\{i:y_i=+1\}} k(x,x_i), \qquad p_2(x) := \frac{1}{m_2}\sum_{\{i:y_i=-1\}} k(x,x_i). \tag{7}$$

The classifier (6) is quite close to the *Support Vector Machine (SVM)* that has recently attracted much attention [10], [8]. It is linear in the RKHS (see (5)), while in the input domain, it is represented by a kernel expansion (6). It is example-based in the sense that the kernels are centered on the training examples, i.e., one of the two arguments of the kernels is always a training example. This is a general property of kernel methods, due to the Representer Theorem [5], [8]. The main point where SVMs deviate from (6) is in the selection of the examples that the kernels are centered on, and in the weight that is put on the individual kernels in the decision function. The SVM decision boundary takes the form

$$y = \operatorname{sgn}\left(\sum_{i=1}^{m}\lambda_i k(x,x_i) + b\right), \tag{8}$$

where the coefficients λ_i and b are computed by solving a convex quadratic programming problem such that the margin of separation of the classes in the RKHS is maximized. It turns out that for many problems this leads to sparse solutions, i.e., often many of the λ_i take the value 0. The x_i with nonzero λ_i are usually called *Support Vectors*.

Using methods from statistical learning theory [10], one can bound the generalization error of SVMs. In a nutshell, statistical learning theory shows that it is imperative that one uses a class of functions whose *capacity* (e.g., measured by the VC dimension) is matched to the size of the training set. In SVMs, the capacity measure used is the size of the margin, which is inversely proportional to the RKHS norm of the SVM parameter vector.

The SV algorithm has been generalized to problems such as regression estimation [10], mappings between general sets of objects [11], and single class problems. As the latter algorithm is closely related to the one to be proposed in the present paper, we will describe it in the next section.

2 Single-class SVMs

Let us assume we are given unlabelled data $x_1, \ldots, x_m \in \mathcal{X}$ generated i.i.d. according to some underlying distribution P. We would like to estimate quantiles C of P using kernel expansions as $C \approx \{x \in \mathcal{X} | f(x) \in I\}$. Here, I is an interval, and $f = \sum_{i=1}^{m} \lambda_i k(x, x_i)$.

In the case of $I = [\rho, \infty[$ (where $\rho \in \mathbb{R}R$), an approach to compute such an estimator f is the single-class SVM [7]. It approximately computes the smallest set $C \in \mathcal{C}$ containing a specified fraction of all training examples, where smallness is measured in terms of a regularizer corresponding to the norm in the RKHS associated with k, and $\mathcal{C}$ is the family of sets corresponding to half-spaces in the RKHS. When choosing a suitable kernel, this notion of smallness will coincide with the intuitive idea that the quantile estimate should not only contain a specified fraction of the training points, but it should also be sufficiently smooth so that we can be confident that this statement will also be approximately true for previously unseen points sampled from P (for an analysis, see [7]).

Let us briefly describe the main ideas of the approach. The training points are mapped into the RKHS using the feature map Φ associated with k, and then it is attempted to separate them from the origin with a large margin by solving the following quadratic program: for $\nu \in (0, 1]$,[2]

$$\underset{\mathbf{w}\in\mathcal{H}}{\text{minimize}}\,, \boldsymbol{\xi} \in \mathbb{R}^m, \rho \in \mathbb{R} \qquad \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu m}\sum_i \xi_i - \rho \tag{9}$$

$$\text{subject to} \qquad \langle \mathbf{w}, \Phi(x_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0. \tag{10}$$

Since nonzero slack variables ξ_i are penalized in the objective function, we can expect that if $\mathbf{w}$ and ρ solve this problem, then the decision function,

$$f(x) = \operatorname{sgn}\left(\langle \mathbf{w}, \Phi(x)\rangle - \rho\right), \tag{11}$$

[2] Here and below we follow the convention that bold face greek character denote vectors, e.g., $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)^\top$.

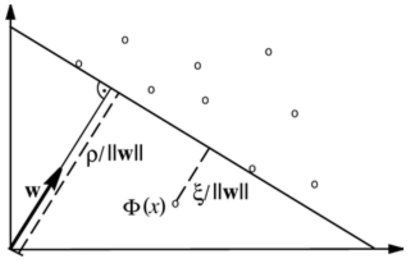


Figure 1: In the 2-D toy example depicted, the hyperplane $\langle \mathbf{w}, \Phi(x) \rangle = \rho$ separates all but one of the points from the origin. The outlier $\Phi(x)$ is associated with a slack variable ξ, which is penalized in the objective function (9). The distance from the outlier to the hyperplane is $\xi / \|\mathbf{w}\|$; the distance between hyperplane and origin is $\rho / \|\mathbf{w}\|$. The latter implies that a small $\|\mathbf{w}\|$ corresponds to a large margin of separation from the origin (from [8]).

will equal 1 for most examples x_i contained in the training set,[3] while the regularization term $\|\mathbf{w}\|$ will still be small. For an illustration, see Figure 1. The trade-off between these two goals is controlled by a parameter ν.

One can show that the solution takes the form

$$f(x) = \operatorname{sgn}\left(\sum_i \alpha_i k(x_i, x) - \rho \right), \tag{12}$$

where the α_i are computed by solving the dual problem,

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^m}{\text{minimize}} \quad \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) \tag{13}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu m} \quad \text{and} \quad \sum_i \alpha_i = 1. \tag{14}$$

Note that due to (14), the training examples contribute with nonnegative weights $\alpha_i \geq 0$ to the solution (12). One can show that asymptotically, a fraction ν of all training examples will have strictly positive weights, and the rest will be zero.

3 Implicit manifold estimation

A richer class of solutions, where some of the weights can be negative, is obtained if we change the geometric setup. In this case, we estimate a region which is a slab in the RKHS, i.e., the area enclosed between two parallel hyperplanes (see Figure 2).

[3] We use the convention that $\operatorname{sgn}(z)$ equals 1 for $z \geq 0$ and -1 otherwise.

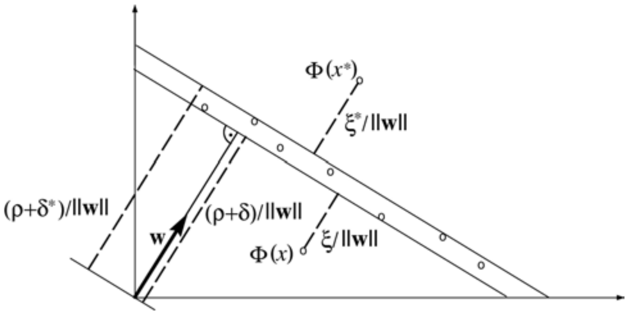


Figure 2: Two parallel hyperplanes $\langle \mathbf{w}, \Phi(x)\rangle = \rho + \delta^{(*)}$ enclosing all but two of the points. The outlier $\Phi(x^{(*)})$ is associated with a slack variable $\xi^{(*)}$, which is penalized in the objective function (15).

To this end, we consider the following modified program:[4]

$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{H}}{\text{minimize}}\,, \boldsymbol{\xi}^{(*)} \in \mathbb{R}^m, \rho\in\mathbb{R} \qquad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu m}\sum_i (\xi_i + \xi_i^*) - \rho && (15)\\ \text{subject to} \qquad & \delta - \xi_i \le \langle \mathbf{w}, \Phi(x_i)\rangle - \rho \le \delta^* + \xi_i^* && (16)\\ \text{and} \qquad & \xi_i^{(*)} \ge 0. && (17) \end{aligned}$$

Here, $\delta^{(*)}$ are fixed parameters. Note that strictly speaking, one of them is redundant: one can show that if we subtract some offset from both, then we obtain the same overall solution, with ρ offset by the same amount. Hence, we can generally set one of them to zero, say, $\delta = 0$. In the simulations shown below, this is the case; nonetheless, we prefer to keep the δ in the optimization problem.

Before we compute the dual problem, let us discuss the relationship of this convex quadratic optimization problem to other approaches.

- For $\delta = 0$ and $\delta^* = \infty$ (i.e., no upper constraint), we recover the single-class SVM (9)–(10).
- If we drop ρ from the objective function and set $\delta = -\varepsilon$, $\delta^* = \varepsilon$ (for some fixed $\varepsilon \ge 0$), we obtain the ε-insensitive support vector regression algorithm [10], for a data set where all output values $y_1, \dots, y_m$ are zero. Note that in this case, the solution is trivial, $\mathbf{w} = 0$. This shows that the ρ in our objective function cannot be dropped and plays an important role.
- For $\delta = \delta^* = 0$, the term $\sum_i (\xi_i + \xi_i^*)$ measures the distance of the point $\Phi(x_i)$ from the hyperplane $\langle \mathbf{w}, \Phi(x_i)\rangle - \rho = 0$ (up to a scaling

[4] Here and below, the superscript $(*)$ simultaneously denotes the variables with and without asterisk, e.g., $\boldsymbol{\xi}^{(*)}$ is a shorthand for $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$.

of $\|\mathbf{w}\|$). As ν tends to zero, this term will dominate the objective function. Hence, in this case, the solution will be a hyperplane that approximates the data well in the sense that the points lie close to it in the RKHS norm.

Let us now compute the dual optimization problem. Here are all constraints, along with the Lagrange multipliers that we will use for them:

$$\xi_i - \delta + \langle \mathbf{w}, \Phi(x_i) \rangle - \rho \geq 0, \qquad \alpha_i \geq 0 \tag{18}$$
$$\xi_i^* + \delta^* + \rho - \langle \mathbf{w}, \Phi(x_i) \rangle \geq 0, \qquad \alpha_i^* \geq 0 \tag{19}$$
$$\xi_i^{(*)} \geq 0, \qquad \beta_i^{(*)} \geq 0 \tag{20}$$

This leads to the Lagrangian

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\xi}^{(*)}, \rho, \boldsymbol{\alpha}^{(*)}, \boldsymbol{\beta}^{(*)}) &= \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu m}\sum_i (\xi_i + \xi_i^*) - \rho \\ &\quad - \sum_i \alpha_i [\xi_i - \delta + \langle \mathbf{w}, \Phi(x_i) \rangle - \rho] \\ &\quad - \sum_i \alpha_i^* [\xi_i^* + \delta^* + \rho - \langle \mathbf{w}, \Phi(x_i) \rangle] \\ &\quad - \sum_i \beta_i \xi_i - \sum_i \beta_i^* \xi_i^* . \end{aligned} \tag{21}$$

The solution of our primal problem (15)–(17) is known to be a saddle point of L. To find it, we need to minimize w.r.t. the primal variables $\mathbf{w}, \boldsymbol{\xi}^{(*)}, \rho$ and maximize w.r.t. the dual variables $\boldsymbol{\alpha}^{(*)}, \boldsymbol{\beta}^{(*)}$. Setting the derivatives w.r.t. the primal variables equal to zero, we obtain

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \iff \mathbf{w} = \sum_i (\alpha_i - \alpha_i^*)\Phi(x_i) \tag{22}$$
$$\frac{\partial L}{\partial \xi_i^{(*)}} = 0 \iff \frac{1}{\nu m} - \alpha_i^{(*)} - \beta_i^{(*)} = 0 \tag{23}$$
$$\frac{\partial L}{\partial \rho} = 0 \iff \sum_i (\alpha_i - \alpha_i^*) = 1. \tag{24}$$

Substituting these conditions into the Lagrangian leads to the dual problem,

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^m}{\text{minimize}} \quad \frac{1}{2}\sum_{ij} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i, x_j) - \delta \sum_i \alpha_i + \delta^* \sum_i \alpha_i^* \tag{25}$$
$$\text{subject to} \quad 0 \leq \alpha_i^{(*)} \leq \frac{1}{\nu m} \tag{26}$$
$$\text{and} \quad \sum_i (\alpha_i - \alpha_i^*) = 1, \tag{27}$$

where the box constraints on $\alpha_i^{(*)}$, (26), have been derived from (23) by taking into account that $\alpha_i^{(*)}, \beta_i^{(*)} \geq 0$.[5]

The dual problem can be solved using standard quadratic programming packages. Alternatively, custom methods such as variants of SMO (cf. [7]) can be used. The offset ρ can be computed from the value of the corresponding variable in the double dual, or using the Karush-Kuhn-Tucker (KKT) conditions, just as in other support vector methods [8]. Once this is done, we can evaluate for each test point x whether it satisfies $\delta \leq \langle \mathbf{w}, \Phi(x) \rangle - \rho \leq \delta^*$. In other words, we have an implicit description of the region in $\mathcal{X}$ that corresponds to the region in between the two hyperplanes in the RKHS. For $\delta = \delta^*$, this is a single hyperplane, corresponding to a manifold in $\mathcal{X}$.[6] See Figure 3 for some toy examples of the algorithm in action.

We now analyze how the parameter ν influences the solution. To this end, we introduce the following shorthands for the sets of SV and outlier indices:

$$\begin{aligned} SV &:= \{i \mid \langle \mathbf{w}, \Phi(x_i) \rangle - \rho - \delta \leq 0\} && (28) \\ SV^* &:= \{i \mid \langle \mathbf{w}, \Phi(x_i) \rangle - \rho - \delta^* \geq 0\} && (29) \\ OL^{(*)} &:= \{i \mid \xi_i^{(*)} > 0\} && (30) \end{aligned}$$

It is clear from the primal optimization problem that for all i, $\xi_i > 0$ implies $\langle \mathbf{w}, \Phi(x_i) \rangle - \rho - \delta < 0$ (and likewise, $\xi_i^* > 0$ implies $\langle \mathbf{w}, \Phi(x_i) \rangle - \rho - \delta^* > 0$), hence $OL^{(*)} \subset SV^{(*)}$. The difference of the SV and OL sets are those points that lie precisely on the boundaries of the constraints.[7]

Below, $|A|$ denotes the cardinality of the set A.

Proposition 3.1. *The solution of (15)–(17) satisfies*

$$\begin{aligned} \frac{|SV|}{m} - \frac{|OL^*|}{m} &\geq \nu, && (31) \\ \frac{|OL|}{m} - \frac{|SV^*|}{m} &\leq \nu. && (32) \end{aligned}$$

Two notes before we proceed to the proof:

- The above statements are not symmetric with respect to exchanging the quantities with asterisks and their counterparts without asterisk.

[5] As an aside, note that due to (27), the dual solution is invariant with respect to the transformation $\delta^{(*)} \to \delta^{(*)} + const.$ — such a transformation only adds a constant to the objective function, leaving the solution unaffected.

[6] subject to suitable conditions on k

[7] The present usage differs slightly from the standard definition of SVs (support vectors), which are usually those that satisfy $\alpha_i^{(*)} > 0$. In our definition, SV are those points where the constraints are active. However, the difference is marginal: (i) It follows from the KKT conditions that $\alpha_i^{(*)} > 0$ implies that the corresponding constraint is active. (ii) while it can happen in theory that a constraint is active and nevertheless the corresponding $\alpha_i^{(*)}$ is zero, this almost never occurs in practice.

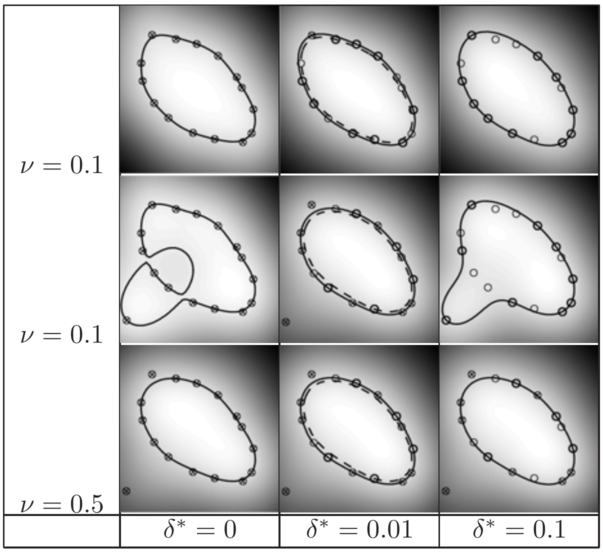


Figure 3: Toy examples of (25)–(27), showing the training points (circles), SVs lying exactly on the hyperplanes (bold circles), and outliers marked by crosses (depicted area $[-1,1]^2$, kernel (4), parameter settings $\sigma = 0.5, \delta = 0$). Lines correspond to hyperplanes constructed in the RKHS (see text); the dashed line is the hyperplane corresponding to the constraint with the $\boldsymbol{\xi}^*$ variables. For $\delta^* = 0$, the two hyperplanes coincide (note that due to finite accuracy, the points do not lie exactly on the hyperplane and are thus marked as outliers); for $\delta^* = 0.1$, the dashed hyperplane is sufficiently far away from the data to reduce the algorithm to the single-class SVM (9)–(10). The *top row* shows a simple toy data set, which in the *middle row* is contaminated with an outlier. The *bottom row* shows how $\nu = 0.5$ handles the outlier.

This is due to the sign of ρ in the primal objective function. If we used $+\rho$ rather than $-\rho$, we would obtain almost the same dual, the only difference being that the constraint (27) would have a "-1" on the right hand side. In this case, the role of the quantities with and without asterisks would be reversed in Proposition 3.1.

- The "ν-property" of single class SVMs is obtained as the special case where $OL^* = SV^* = \{\}$.

Proof. Assume that $(\mathbf{w}, \boldsymbol{\xi}^{(*)}, \rho)$ is a solution of (15)–(17). Thus it is optimal w.r.t. all primal variables, in particular $\boldsymbol{\xi}^{(*)}$ and ρ, i.e., keeping $\mathbf{w}$ fixed. In that case, the problem takes the form

$$\underset{\boldsymbol{\xi}^{(*)} \in \mathbb{R}^m, \rho \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{m} \sum_i (\xi_i + \xi_i^*) - \nu\rho \tag{33}$$

$$\text{subject to} \quad \delta - \xi_i \leq \langle \mathbf{w}, \Phi(x_i) \rangle - \rho \leq \delta^* + \xi_i^* \quad \text{and} \quad \xi_i^{(*)} \geq 0. \tag{34}$$

If we increase ρ by a small $\epsilon > 0$ (cf. Figure 2),[8] (33) decreases proportionally to ν *plus* the fraction of points with $\xi_i^* > 0$ (since these slack variables can be shrunk by the same ϵ without violating the constraints) *minus* the fraction of SVs (remember that all SVs either have $\xi_i > 0$ or lie exactly on the hyperplane $\langle \mathbf{w}, \Phi(x_i)\rangle - \rho - \delta = 0$ — in both cases, an increase of ρ by ϵ will lead to the same increase in the ξ_i variables, in order to satisfy the constraints). If the overall decrease were positive, i.e., if $\nu + \frac{|OL^*|}{m} - \frac{|SV|}{m} > 0$, then we could get a strict decrease in (33) by changing ρ, violating the assumption that we are already at the optimum. Therefore, we have $\frac{|SV|}{m} - \frac{|OL^*|}{m} \geq \nu$.

If, on the other hand, we decrease ρ by an $\epsilon > 0$, the objective function will decrease proportionally to $\frac{|OL|}{m} - \frac{|SV^*|}{m} - \nu$. As above, this quantity cannot by strictly positive, since we are already optimal. Therefore we have $\frac{|OL|}{m} - \frac{|SV^*|}{m} \leq \nu$. □

If in addition we make certain assumptions on the distribution generating the data and on the kernel,[9] then asymptotically, the two inequalities in the proposition become equalities with probability 1. The main idea of the proof can be given in a nutshell: if the capacity of the function class that we are using is well behaved (which it is, since we are regularizing using the RKHS norm $\|\mathbf{w}\|$), then asymptotically, the set of points which lie exactly on the hyperplanes is negligible. Hence, loosely speaking, we have $SV^{(*)} = OL^{(*)}$, and thus $\nu \leq \frac{|SV|}{m} - \frac{|SV^*|}{m} = \frac{|OL|}{m} - \frac{|OL^*|}{m} \leq \nu$. For details, see [8], [9].

To conclude this section, note that an approximate description of the data as the zero set of a function is not only useful as a compact representation of the data. It can also potentially be used in tasks such as denoising and image super-resolution. Given a noisy point x, we can map it into the RKHS and then project it onto the hyperplane(s) that we have learnt. We then compute an approximate pre-image under Φ to get a noise-free version of x. A similar statistical denoising technique has been used in conjunction with kernel PCA (to be described next) with rather encouraging results [8], [4].

4 Other kernel approaches for manifold estimation

There exist several other possibilities to use machine learning methods employing positive definite kernels for estimating manifolds. One of them is known as the RPM algorithm (see [8]); two other ones, to be described below, build on the kernel PCA algorithm.

Kernel PCA The kernel method for computing dot products in an RKHS is not restricted to SV machines. It can be used to develop nonlinear generalizations of any algorithm that can be cast in terms of dot products, such

[8] We choose ϵ small enough so that all constraints that are not active will also not be active after adding the ϵ; it is easy to see that such an ϵ exists.

[9] Essentially, we need to require that the distribution have a density w.r.t. the Lebesgue measure, and that k is analytic and non-constant (cf. [8], [9]).

as principal component analysis. Given data $x_1, \dots, x_m \in \mathcal{X}$, kernel principal component analysis (kPCA) [8] computes the principal components of the points $\Phi(x_1), \dots, \Phi(x_m)$. Since $\mathcal{H}$ may be infinite-dimensional, the PCA problem needs to be transformed into a problem that can be solved in terms of the kernel k. To this end, we consider an estimated covariance matrix in $\mathcal{H}$,

$$\mathbf{C} := \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i)\Phi(x_i)^\top, \tag{35}$$

where $\Phi(x_i)^\top$ denotes the linear form mapping $\mathbf{v} \in \mathcal{H}$ to $\langle \Phi(x_i), \mathbf{v} \rangle$. To diagonalize $\mathbf{C}$, we first observe that all solutions to $\mathbf{Cv} = \lambda \mathbf{v}$ with $\lambda \neq 0$ must lie in the span of Φ-images of the training data (as can be seen by substituting (35) and dividing by λ). Thus, we may expand the solution $\mathbf{v}$ as $\mathbf{v} = \sum_{i=1}^{m} \alpha_i \Phi(x_i)$, thereby reducing the problem to that of finding the α_i. The latter can be shown to take the form $m\lambda\boldsymbol{\alpha} = K\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$ and $K_{ij} = k(x_i, x_j)$. Absorbing the m factor into the eigenvalue λ, one can moreover show that the p-th feature extractor takes the form

$$\langle \mathbf{v}^p, \Phi(x) \rangle = \frac{1}{\sqrt{\lambda^p}} \sum_{i=1}^{m} \alpha_i^p k(x_i, x). \tag{36}$$

This is derived by computing the dot product between a test point $\Phi(x)$ and the p-th eigenvector in the RKHS; the $\frac{1}{\sqrt{\lambda^p}}$ factor ensures that $\langle \mathbf{v}^p, \mathbf{v}^p \rangle = 1$.

When evaluated on the training example x_n, (36) takes the form

$$\langle \mathbf{v}^p, \Phi(x_n) \rangle = \frac{1}{\sqrt{\lambda^p}} (K\alpha^p)_n = \frac{1}{\sqrt{\lambda^p}} (\lambda^p \alpha^p)_n = \sqrt{\lambda^p} \alpha_n^p. \tag{37}$$

In (35), we have implicitly assumed that the data in the RKHS have zero mean. If this is not the case, we need to subtract the mean $(1/m)\sum_i \Phi(x_i)$ from all points. This leads to a slightly different eigenvalue problem, where we diagonalize

$$K' = (\mathbf{1} - ee^\top) K (\mathbf{1} - ee^\top) \tag{38}$$

(with $e = m^{-1/2}(1, \dots, 1)^\top$) rather than K.

The kPCA algorithm can be used to obtain an implicit description of a manifold containing the data as follows. The principal directions with the smallest eigenvalues (sometimes called "minor components") characterize directions in the RKHS such that when projected onto these directions, the data set has the smallest possible variance which can be obtained in any direction which is in the span of the mapped data.[10] Generally, we are interested in low variance directions which lie in the span of sets of inputs points (e.g., the training set) mapped into the RKHS, as these lead to implicit

[10] Note that for some kernels, the RKHS will be infinite dimensional. In that case, there are infinitely many *zero* variance directions which do not lie in the span of the data.

function expansions in terms of kernel functions. If we consider expansions in terms of the training set, the functions take the form

$$f(x) = \left\langle \mathbf{V}^p, \Phi(x) - \frac{1}{m}\sum_i \Phi(x_i) \right\rangle. \tag{39}$$

A tighter description of the desired manifold may be obtainable by intersecting several such surfaces, e.g., using (39) for values of p corresponding to several small eigenvalues λ_p.

LLE and Laplacian Eigenmaps Kernel PCA can also be used for manifold learning in a rather different way. In this case, the manifold is not learnt as the zero set of a kernel expansion. Rather, we will obtain a low dimensional coordinate embedding of data sampled from the manifold (”dimensionality reduction”).

It turns out that *locally linear embedding (LLE)* [6], currently a rather popular algorithm for nonlinear dimensionality reduction, is a special case of kPCA [3]: The LLE algorithm first constructs W to be the matrix whose row i (summing to 1) contains the coefficients to of the minimal squared error affine reconstruction of x_i from its p nearest neighbors. Denote $M := (\mathbf{1} - W)(\mathbf{1} - W^\top)$, with maximal eigenvalue $\lambda_{\max}$. One can show that M's smallest eigenvalue is 0 and the corresponding uniform eigenvector is e. In LLE, the coordinate values of the m-dimensional eigenvectors $m-d, \ldots, m-1$ give an embedding of the m data points in $\mathbb{R}^d$. If we define $K := (\lambda_{\max}\mathbf{1} - M)$, then by construction, K is a positive definite matrix, its leading eigenvector is e, and the coordinates of the eigenvectors $2, \ldots, d+1$ provide the LLE embedding. Equivalently, we can use the eigenvectors $1, \ldots, d$ of the matrix obtained by projecting out the subspace spanned by e, i.e., $(\mathbf{1} - ee^\top)K(\mathbf{1} - ee^\top)$. Note that this is identical to the centered kernel matrix (38) used in kPCA. We thus know that the coordinates of the leading eigenvectors of kPCA performed on K yield the LLE embedding. This, together with (37), shows that the LLE embedding is identical to the kPCA projections up to a whitening multiplication with $\sqrt{\lambda^p}$.

As shown in [3], several other approaches can be viewed as special cases of kPCA, including certain spectral methods. Many of these methods are based on the computation of a weighted adjacency matrix W on the data, e.g., using the kernel (4) on neighboring points (where several definitions of neighborhood are possible).[11] Define the *graph Laplacian* L by $L_{ii} := d_i$, $L_{ij} = -W_{ij}$ if x_i and x_j are neighbors, and 0 otherwise, where $d_i = \sum_{j\sim i} W_{ij}$ is the degree of the ith vertex. It turns out that similar to LLE, the bottom eigenvectors of the Laplacian can provide a low-dimensional representation of the data [2], and again, a link to KPCA can be established [3].

[11] This local similarity measure can also take into account invariances of the data.

5 Conclusion

Kernel methods have a solid foundation in statistical learning theory and functional analysis. They let us interpret (and design) learning algorithms geometrically in an RKHS, and combine statistics and geometry in an elegant way. The present article has described several methods for using this approach for the estimation of manifolds.

References

[1] Aizerman M.A., Braverman É.M., Rozonoér L.I. (1964). *Theoretical foundations of the potential function method in pattern recognition learning.* Automation and Remote Control **25** 821 – 837.

[2] Belkin M., Niyogi P. (2003). *Laplacian eigenmaps for dimensionality reduction and data representation.* Neural Computation **15** (6) 1373 – 1396.

[3] Ham J., Lee D., Mika S., Schölkopf B. (2004). *A kernel view of the dimensionality reduction of manifolds.* In Proceedings of ICML (in press).

[4] Kim K.I., Franz M.O., Schölkopf B. (2004). *Kernel Hebbian algorithm for single-frame super-resolution.* In Statistical Learning in Computer Vision Workshop, Prague.

[5] Kimeldorf G.S., Wahba G. (1971). *Some results on Tchebycheffian spline functions.* Journal of Mathematical Analysis and Applications **33** 82 – 95.

[6] Roweis S., Saul L. (2000). *Nonlinear dimensionality reduction by locally linear embedding.* Science **290**, 2323 – 2326.

[7] Schölkopf B., Platt J., Shawe-Taylor J., Smola A.J., Williamson R.C. (2001). *Estimating the support of a high-dimensional distribution.* Neural Computation **13** 1443 – 1471.

[8] Schölkopf B., Smola A.J. (2002). *Learning with kernels.* MIT Press, Cambridge, MA.

[9] Steinwart I. (2004). *Sparseness of support vector machines—some asymptotically sharp bounds.* In S. Thrun, L. Saul, and B. Schölkopf, (eds), Advances in Neural Information Processing Systems **16**. MIT Press, Cambridge, MA.

[10] Vapnik V.N. (1995). *The nature of statistical learning theory.* Springer Verlag, New York.

[11] Weston J., Chapelle O., Elisseeff A., Schölkopf B., Vapnik V. (2003). *Kernel dependency estimation.* In S. Becker, S. Thrun, and K. Obermayer, (eds), Advances in Neural Information Processing Systems **15**, Cambridge, MA, USA. MIT Press.

Address: B. Schölkopf, Max-Planck-Institut für biologische Kybernetik, Spemannstr. 38, Tübingen, Germany

E-mail: `bernhard.schoelkopf@tuebingen.mpg.de`

COMPSTAT'2004 Symposium

OUTLIER DETECTION AND CLUSTERING BY PARTIAL MIXTURE MODELING

David W. Scott

Key words: Minimum distance estimation, robust estimation, exploratory data analysis.

COMPSTAT 2004 section: Statistical software.

Abstract: Clustering algorithms based upon nonparametric or semiparametric density estimation are of more theoretical interest than some of the distance-based hierarchical or ad hoc algorithmic procedures. However density estimation is subject to the curse of dimensionality so that care must be exercised. Clustering algorithms are sometimes described as biased since solutions may be highly influenced by initial configurations. Clusters may be associated with modes of a nonparametric density estimator or with components of a (normal) mixture estimator. Mode-finding algorithms are related to but different than gaussian mixture models. In this paper, we describe a hybrid algorithm which finds modes by fitting incomplete mixture models, or partial mixture component models. Problems with bias are reduced since the partial mixture model is fitted many times using carefully chosen random starting guesses. Many of these partial fits offer unique diagnostic information about the structure and features hidden in the data. We describe the algorithms and present some case studies.

1 Introduction

In this paper, we consider the problem of finding outliers and/or clusters through the use of the normal mixture model

$$f(\mathbf{x}) = \sum_{k=1}^{K} w_k \, \phi(\mathbf{x} \,|\, \mu_k, \Sigma_k) \,. \qquad (1)$$

Mixture models afford a very general family of densities. If the number of components, K, is quite large, then almost any density may be well-approximated by this model. Aitkin and Wilson [1] first suggested using the mixture model as a way of handling data with multiple outliers, especially when some of the outliers group into clumps. They used the EM algorithm to fit the mixture model. Assuming that the "good" data are in one cluster and make up at least fifty percent of the total data, then it is easy to see that we have introduced a number of "nuisance parameters" into the problem (to model the outliers).

Implementing this idea in practice is challenging. If there are just a few "clusters" of outliers, then the number of nuisance parameters should not pose too much difficulty. However, as the dimension increases, the total number of parameters grows quite rapidly, especially if a completely general covariance matrix, Σ_k, is used for each component. The most directly challenging problem is finding an appropriate choice of the number of components, K, and initial guesses for the many parameters. An obvious first choice is to use a clustering algorithm such as k-means [15] as an approach to find an initial partition, and then compute the relative size, means, and covariances of each group to use as initial guesses for the EM algorithm.

It is abundantly clear that for many of our fits, we will in fact be using the wrong value of K. Furthermore, even if we happen to be using the appropriate value for K, there may be a number of different solutions, depending upon the specific initialization of the parameters. Starting with a large number of initial configurations is helpful, but as the dimension and sample size increase, the number of possibilities quickly exceeds our capabilities.

However, the least discussed and least understood problem arises because so little is generally known about the statistical distributions of the clusters representing the outliers. It certainly seems more reasonable to know something about the distribution of the "good" data; however, one is on much less firm ground trying to claim the same knowledge about the distributions of the several non-informative clusters. Even in the situation where the "good" data are in more than one cluster, sometimes little is known about the distribution in one or more of those "good" clusters.

In this paper, we discuss how an alternative to the EM algorithm can provide surprisingly useful estimates and diagnostics, even when K is incorrect. Such technology is especially interesting when K is too small, since in this situation the number of parameters to be estimated may be a small fraction of the number in the full, correct model. Furthermore, this technology is of special interest in the situation where little is known about the correct distribution of many of the clusters. This latter capability is of growing importance and interest in the analysis of massive datasets typically encountered in data mining applications.

2 Mixture fits with too few components

We examine some empirical results to reinforce these ideas. One well-known trimodal density in two dimensions is the lagged Old Faithful Geyser duration data, $\{(x_{t-1}, x_t),\ t = 2, \ldots, 298\}$; see [2] and [27]. Successive eruptions were observed and the duration of each eruption, $\{x_t,\ t = 1, \ldots, 299\}$, recorded to the nearest second. A quick count shows that 23, 2, and 53 of the original 299 values occurred exactly at $x_t = 2$, 3, and 4 minutes, respectively. Examining the original time sequence suggests that those measurements are clumped; perhaps accurate measurements were not taken after dark. We modified the data as follows: the 105 values that were only recorded to the nearest minute were blurred by adding uniform noise of 30 seconds in duration. Then all of the data were blurred by adding uniform noise, $U(-.5, .5)$, seconds, and then converted back into minutes.

In Figure 1, maximum likelihood estimates (MLE) of a bivariate normal and three two-component bivariate normal mixture fits are shown. Each bivariate normal density is represented by 3 elliptical contours at the 1, 2, and 3-σ levels. Figure 1 provides some examples of different solutions, depending upon the value of K selected and the starting values for the parameters chosen. In two dimensions, your eye can tell you what is wrong with these fits. In higher dimensions, diagnostics indicating a lack of fit leave unclear if a component should be split into two, or if the assumed shaped of the component is not correct.

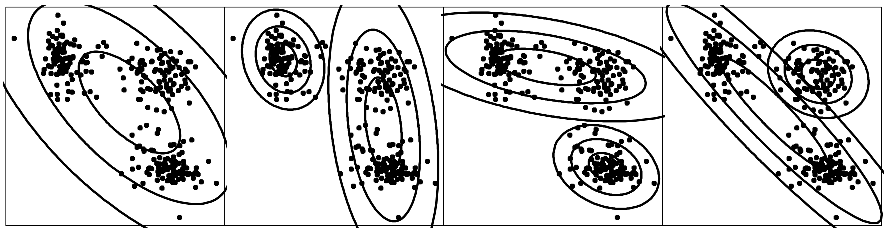


Figure 1: Maximum likelihood bivariate normal mixture fits to the lagged Old Faithful geyser eruption data with $K = 1$ and $K = 2$. The weights in each frame from L to R are (1.0), (.350, .650), (.645, .355), and (.728, .272). Each bivariate normal component is represented by 3 contours at the 1, 2, and 3-σ levels.

3 The L2E criterion

Minimum distance estimation for parametric modeling of $f_\theta(x) = f(x|\theta)$ is a well-known alternative to maximum likelihood; see [7]. In practice, several authors have suggested modeling the data with a nonparametric estimator (such as the histogram or kernel method), and then numerically finding the values of the parameters in the parametric model that minimize the distance between f_θ and the curve; see [6] and [9], who considered Hellinger and L2 distances, respectively. Using a nonparametric curve as a target introduces some choices, such as the smoothing parameter, but also severely limits the dimension of the data and the number of parameters that can be modeled. (Precise numerical integration is quite expensive even in two dimensions. Numerical optimization algorithms require very good accuracy in order to numerically estimate the gradient vectors.)

Several authors have discovered an alternative criterion for parametric estimation in the case of L2 or integrated squared error (ISE); see [25], [13], [5], [20], [21], [22], for example. (This idea follows from the pioneering work of Rudemo [18] and Bowman [8] on cross-validation of smoothing parameters in nonparametric density estimates.) In particular, Scott [20], [21] considered estimation of mixture models by this technique. Given a true density, $g(x)$, and a model, $f_\theta(x)$, the goal is to find a fully data-based estimate of the L2 distance between g and f, which is then minimized with respect to θ. Expanding the L2 criterion

$$d(\hat{f}_\theta, g) = \int \left[\hat{f}_\theta(x) - g(x)\right]^2 dx\,, \tag{2}$$

we obtain the three integrals

$$d(\hat{f}_\theta, g) = \int \hat{f}_\theta(x)^2 dx - 2\int \hat{f}_\theta(x)\, g(x)\, dx + \int g(x)^2 dx\,. \tag{3}$$

The third integral is unknown but is constant with respect to θ and therefore may be ignored. The first integral is often available as a closed form expression that may be evaluated for any posited value of θ. Additionally, we must add an assumption on the model that this integral is always finite, i.e. $f_\theta \in L_2$. The second integral is simply the average height of the density estimate, given by $-2\,\mathrm{E}[\hat{f}_\theta(X)]$, where $X \sim g(x)$, and which may be estimated in an unbiased fashion by $-2n^{-1}\sum_{i=1}^n \hat{f}_\theta(x_i)$. Combining, the L2E criterion for parametric estimation is given by

$$\hat{\theta} = \arg\min_\theta \left[\int \hat{f}_\theta(x)^2 dx - \frac{2}{n}\sum_{i=1}^n \hat{f}_\theta(x_i)\right]. \tag{4}$$

For the multivariate normal mixture model in Equation 1,

$$\int_{\Re^d} \hat{f}_\theta(x)^2 dx = \sum_{k=1}^K \sum_{\ell=1}^K w_k\, w_\ell\, \phi(0\,|\,\mu_k - \mu_\ell, \Sigma_k + \Sigma_\ell). \tag{5}$$

Since this is a computationally feasible closed-form expression, estimation of the normal mixture model by the L2E procedure may be performed by use of any standard nonlinear optimization code; see [20], [21]. In particular, we used the *nlmin* routine in the Splus library for the examples in this paper.

Next, we return to the Old Faithful geyser example. Using the same starting values as in Figure 1, we computed the corresponding L2E estimates, which are displayed in Figure 2. Clearly, both algorithms are attracted to the same (local) estimates, which combine various clusters into one (since $K < 3$). However, there are interesting differences. First we compare the estimated weights: in Figure 1, the MLE weight of the larger component in each frame is 1, 0.65, 0.65, and 0.73, respectively, while in Figure 2 the corresponding L2E weights are 1, 0.74, 0.72, and 0.71. Of more interest, the L2E covariance matrices are either tighter or smaller. Since the (explicit) goal of L2E is to find the most normal fit (locally), observe that a number of points in the smaller clusters fall outside the 3-σ contours in frames 2 and 3 of Figure 2. The MLE covariance estimate is not robust and is inflated by those (slight) outliers. These differences are likely due to the inherent robustness properties of any minimum distance criterion; see [12]. Increasing the covariance matrix to "cover" a few outliers results in a large increase in the integrated squared or L2 error, and hence those points are largely ignored.

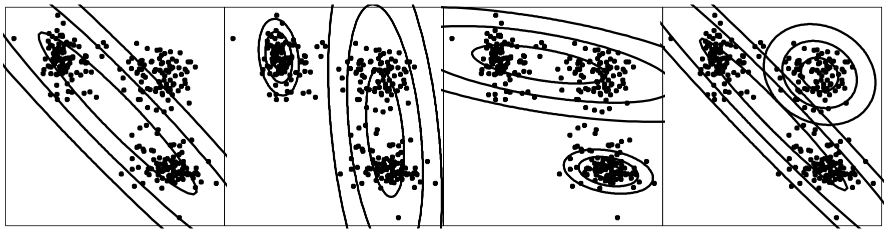


Figure 2: Several L2E mixture fits to the lagged Old Faithful geyser eruption data with $K = 1$ and $K = 2$; see text. The weights in each frame are (1.0), (.258, .742), (.714, .286), and (.711, .289).

4 Partial mixture modeling

The two-component L2E estimates above were computed with the constraint that $w_1+w_2 = 1$. Is this constraint necessary? Can the weights w_1 and w_2 be treated as unconstrained variables? Certainly, when using EM or maximum likelihood, increasing the weights increases the likelihood without bound, so that the constraint is necessary (and active). However, *the L2E criterion does not require that the model $\hat{f}_\theta$ be a density.* The second integral in Equation 3 measures the average height of the density model, but a careful review of the argument leading to Equation 4 confirms the fact that only $g(x)$ is required to be a density, not $\hat{f}_\theta(x)$; see [22].

With this understanding, when we fit a L2E mixture model with $K = 2$, we are only assuming that the true mixture has at least 2 components. That is, we explicitly use our model for the local components of "good" data (local in the sense of our initial parameter guesses), but make no explicit assumption about the (unknown) distribution of the remaining data, no matter how many or few clusters they clump into. Our algorithm is entirely local. Different starting values may lead to quite different estimates.

Thus, we re-coded our L2E algorithm treating all of the weights in Equation 5 as *unconstrained* variables. In Figure 3, we display some of the "unconstrainted" L2E mixture estimates, using the same starting values as in Figure 2. These estimates are qualitatively quite similar to those in Figure 2, with some interesting differences. Comparing the first frames in Figures 2 and 3, the covariance matrix has narrowed as the weight decreased to .783. The sums of the (unconstrained) weights in the final three frames of Figure 3 are 0.947. 0.966, and 1.048. In the first two cases, the total probability modeled is less than unity, suggesting a small fraction of the data are being treated/labeled as outliers with respect to the fitted normal mixture model. The fact that the third total probability exceeds unity is consistent with our previous observation that the best fitting curve in the L2 or ISE sense often integrates to more than 1, when there is a gap in the middle of the data.

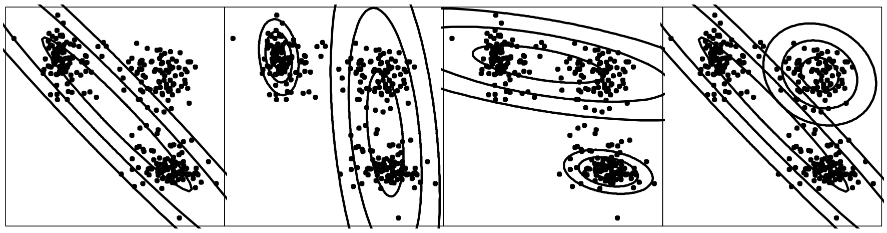


Figure 3: Several L2E partial mixture fits to the lagged Old Faithful geyser eruption data with $K = 1$ and $K = 2$, but without any constraints on the weights; see text. The weights in each frame are $(.783)$, $(.253, .694)$, $(.683, .283)$, and $(.751, .297)$.

Since there are potentially many more local solutions, we display four more L2E solutions in Figure 4. Some of these estimates are quite unexpected and deserve careful examination. The first frame is a variation of a $K = 1$ component which captures 2 clusters. However, the $K = 2$ estimates in the last 3 frames each capture two individual clusters, while completely ignoring the third. Comparing the contours in the last three frames of Figure 4, we see that exactly the same estimates appear in different pairs. Looking at the weights in Figures 3 and 4, we see that the smaller isolated components are almost exactly reproduced while entirely ignoring the third cluster. This feature of L2E is quite novel and we conclude that many of the local L2E results hold valuable diagnostic information as well as quite useful estimates of the local structure of the data.

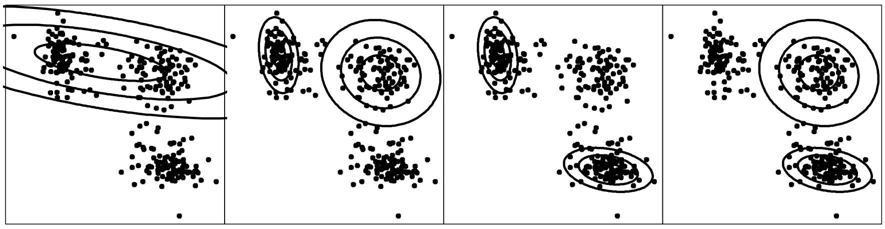


Figure 4: Same as Figure 3 but different starting values; see text. The weights in each frame are $(.683)$, $(.253, .316)$, $(.253, .283)$, and $(.316, .283)$.

Finally, in Figure 5, we conclude this investigation of the geyser data by checking a number of $K = 1$ unconstrained L2E solutions. In this case, the three individual components are found one at a time, depending upon the initial parameter values. Notice that the weights are identical to those in the previous figure. Furthermore, these weights are less than 50%, which is the usual breakdown point of robust algorithms; see [17]. However, the L2E algorithm is local and different ideas of breakdown apply.

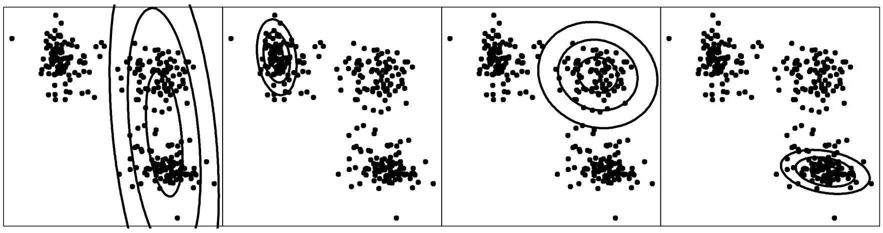


Figure 5: Four more $K = 1$ partial mixture fits to the geyser data; see text. The weights in each frame are (.694), (.253), (.316), and (.283).

5 Other examples

5.1 Star data

Another well-studied bivariate dataset was discussed by Rousseeuw and Leroy [17]. The data are measurements of the temperature and light intensity of 47 stars in the direction of Cygnus. For our analysis, the data were blurred by uniform $U(-.005, .005)$ noise. Four giant stars exert enough influence to distort the correlation of a least-squares or maximum likelihood estimate; see the first frame in Figure 7. In the second frame, a $K = 2$ MLE normal mixture is displayed. Notice the four giant stars are represented by one of the two mixture components and has a nearly singular covariance matrix. The third frame shows a $K = 1$ partial component mixture fit by L2E, with $\hat{w} = 0.937$. The shape of the two covariance matrices of the "good" data is somewhat different in these three frames. In particular, the correlation coefficients are -0.21, 0.61, and 0.73, respectively.

These data were recently re-analyzed by Wang and Raftery [26] with nearest-neighbor variance estimator (NNVE), an extension of the NNBR estimator [10]. They compared their covariance estimates to the minimum volume ellipsoid (MVE) of Rousseeuw and Leroy [17] as well as the (non-robust) MLE. In Figure 7, I have overlaid these 4 covariance matrices (at the 1-σ contour level) with that of the partial density component (PDC) estimate obtained by L2E shown in the third frame of Figure 6. For convenience, I have centered these ellipses on the origin. The NNVE and NNBR ellipses are virtu-

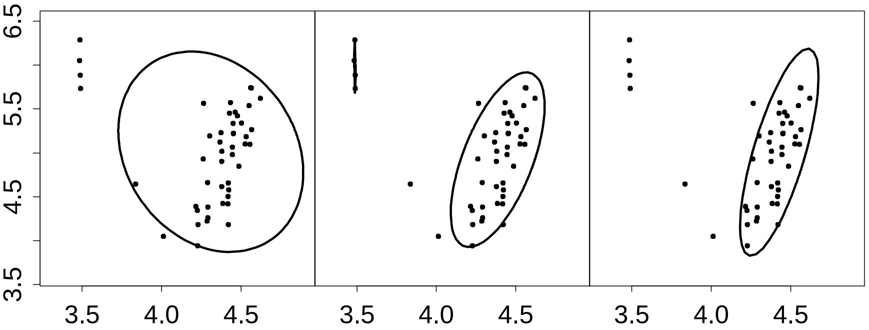


Figure 6: Two-σ contours of MLE ($K = 1$), MLE mixture ($K = 2$), and partial L2E mixture ($K = 1$) fits to the blurred star data.

ally identical, while the MVE ellipse is slightly rotated and narrower. These three are surrounded by the slightly elongated L2E PDC ellipse. Of course, the MLE has the wrong (non-robust) orientation. The correlation coefficients for NNVE and NNBR are 0.65 versus 0.73 for MVE and L2E. Observe that L2E does not explicitly require a search for the good data. The other three algorithms require extensive search and/or calibration of an auxiliary parameter. L2E is driven by the choice of the shape of the mixing distribution. One might choose instead to use t_ν components, as suggested by McLachlan and Peel [16], although the degrees of freedom must be specified. In either case, L2E provides useful diagnostic information as a byproduct of the estimation, rather than as a follow-on step of analysis.

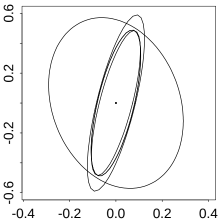


Figure 7: Ellipses representing the 2-σ contours of five estimates of the covariance matrix of the star data; see text.

5.2 Australian athlete data

For our final example, we consider four variables from the AIS data on Australian Athletes [11]. These data are available in the R package with the command `data(ais,package='sn')`. Following Wang and Raftery [26], we selected the variables body fat (BFAT), body mass index (BMI), red cell count (RCC), and lean body mass (LBM). (Wang and Raftery also included ferritin in their analysis.) We blurred the data then standardized each variable.

We fit a $K = 1$ L2E starting with the maximum likelihood estimate. The result was $\hat{w}_1 = 0.98$. A pairwise scatterdiagram of the 202 points is shown in Figure 8, together with contours of the fitted 4-dimensional ellipse. A careful examination of this plots suggests some clusters. In fact, the first 100 measurements are of female athletes and the last 102 measurements are of male athletes.

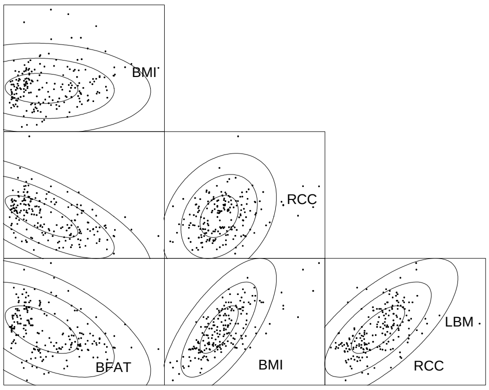


Figure 8: Ellipses representing the (1,2,3)-σ contours of a L2E partial mixture estimate of the Australian athlete data; see text.

Starting with the MLE values for the female athletes, we re-fit a $K = 1$ L2E. Now $\hat{w}_1 = 0.41$ (somewhat less than the 49.5% female population). The contours of the fitted 4-dimensional ellipse are superimposed upon the scatter matrix in Figure 9. The L2E is clearly modeling a large fraction of the female athletes.

Finally, we started the L2E with the male values. However, L2E found a smaller subset of the data lying in a subspace. (L2E is just as susceptible at MLE at being attracted to singular mixture components, depending upon initial guesses. That is why blurring was applied in all our examples to remove trivial singularities due to rounding.) Further experimentation would be interesting.

6 Discussion

We have shown how a minimum distance criterion and a mixture model with only one or two partial components can provide useful estimates and diagnostics. In particular, the value of $\hat{w}_1 + \hat{w}_2$ provides an indication of the

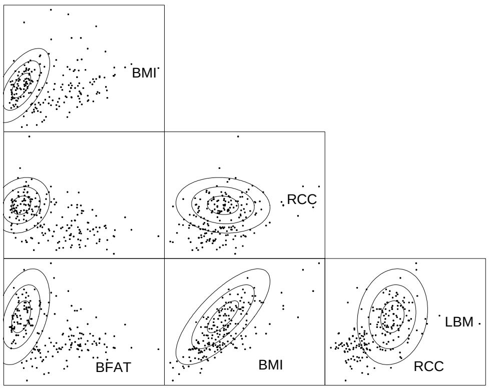


Figure 9: Ellipses representing the (1,2,3)-σ contours of a second L2E partial mixture estimate of the Australian athlete data; see text.

fraction of the data being modeled by a $K = 2$ mixture. In our experience, the proportion of solutions that are interesting when $K = 2$ and the parameters are initialized by some random process is quite small. Further research on this question is open. However, many of the $K = 1$ solutions following random initialization are quite useful. The systematic use of these ideas for clustering is explored further in [23].

Alternatively, Banfield and Raftery [4] allow a number of outliers to be modeled as a spatial Poisson process. It would be interesting to apply that model with $K = 2$ to these data, where the noise is not Poisson, and to compare the parameter estimates.

The identification of outliers without an explicit probability model should always be viewed as preliminary and exploratory. If a probability model is known, then the tasks of parameter estimation and outlier identification can be more rigorously defined. However, even probability models are usually known only approximately at best, and hence outliers so identified are still subject to certain biases.

The general topic of outlier detection is discussed in [3]. Robust estimation is described by Huber [14]. Coupled with a good exploratory such as XGobi [24], the L2E PDC has much potential for helping unlock information in complex data.

References

[1] Aitkin M., Wilson, G.T. (1980). *Mixture models, outliers, and the EM algorithm.* Technometrics **22**, 325 – 331.

[2] Azzalini A., Bowman A.W. (1990). *A look at some data on the old faithful geyser.* Applied Statistics **39**, 357 – 365.

[3] Barnett V., Lewis T. (1994). *Outliers in statistical data.* John Wiley & Sons, New York.

[4] Banfield J.D., Raftery A.E. (1993). *Model-based Gaussian and non-Gaussian clustering.* Biometrics **49**, 803 – 821.

[5] Basu A., Harris I.R., Hjort H.L., Jones M.C. (1998). *Robust and efficient estimation by minimising a density power divergence.* Biometrika **85**, 549 – 560.

[6] Beran R. (1977), *Robust location estimates.* The Annals of Statistics **5**, 431 – 444.

[7] Beran R. (1984). *Minimum distance procedures.* In Handbook of Statistics Volume 4: Nonparametric Methods, pp. 741 – 754.

[8] Bowman A.W. (1984). *An alternative method of cross-validation for the smoothing of density estimates.* Biometrika **71**, 353 – 360.

[9] Brown L.D., Hwang J.T.G. (1993). *How to approximate a histogram by a normal density.* The American Statistician **47**, 251 – 255.

[10] Byers S., Raftery A.E. (1998). *Nearest-neighbor clutter removal for estimating features in spatial point processes.* Journal of the American Statistical Association **93**, 577 – 584.

[11] Cook R.D., Weisberg S. (1994). *An introduction to regression graphics.* Wiley, New York.

[12] Donoho D.L., Liu R.C. (1988). *The 'automatic' robustness of minimum distance functional.* The Annals of Statistics **16**, 552 – 586.

[13] Hjort H.L. (1994). *Minimum L2 and robust Kullback-Leibler estimation.* Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, P. Lachout and J.Á. Víšek (eds.), Prague Academy of Sciences of the Czech Republic, pp. 102 – 105.

[14] Huber P.J. (1981). *Robust statistics.* John Wiley & Sons, New York.

[15] MacQueen J.B. (1967). *Some methods for classification and analysis of multivariate observations.* Proc. Symp. Math. Statist. Prob 5th Symposium **1**, 281 – 297, Berkeley, CA.

[16] McLachlan G.J., Peel D. (2001). *Finite mixture models.* John Wiley & Sons, New York.

[17] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* John Wiley & Sons, New York.

[18] Rudemo M. (1982). *Empirical choice of histogram and kernel density estimators.* Scandinavian Journal of Statistics **9**, 65 – 78.

[19] Scott D.W. (1992). *Multivariate density estimation: theory, practice, and visualization.* John Wiley, New York.

[20] Scott D.W. (1998). *On fitting and adapting of density estimates.* Computing Science and Statistics, S. Weisberg (Ed.) **30**, 124 – 133.

[21] Scott D.W. (1999). *Remarks on fitting and interpreting mixture models.* Computing Science and Statistics, K. Berk and M. Pourahmadi, (Eds.) **31**, 104 – 109.

[22] Scott D.W. (2001). *Parametric statistical modeling by minimum integrated square error.* Technometrics **43**, 274 − 285.

[23] Scott D.W., Szewczyk W.F. (2001). *The stochastic mode tree and clustering.* Journal of Computational and Graphical Statistics, under revision.

[24] Swayne D.F., Cook D., Buja A. (1998). *XGobi: Interactive dynamic data visualization in the X Window system.* Journal of Computational and Graphical Statistics **7**, 113 − 130.

[25] Terrell G.R. (1990). *Linear density estimates.* Proceedings of the Statistical Computing Section, American Statistical Association, 297 − 302.

[26] Wang N., Raftery A.E. (2002). *Nearest-neighbor variance estimation: Robust covariance estimation via nearest-neighbor cleaning.* Journal of the American Statistical Association **97**, 994 − 1019.

[27] Weisberg S. (1985). *Applied linear regression.* John Wiley, New York.

Acknowledgement: This work was supported in part by NSF grant number DMS 02-04723 and NSF Digital Government contract number EIA-9983459.

Address: D.W. Scott, Rice University, Department of Statistics, MS-138, POBox 1892, Houston, TX 77251-1892 USA

E-mail: `scottdw@rice.edu`

INTERDATABASE AND DANDD

Ritei Shibata

Key words: Working with data, environment, database, Internet.

COMPSTAT 2004 section: E-statistics.

Abstract: InterDatabase is a comfortable environment for working simultaneously with different types of data which are scattered over various databases or files on a network or on the Internet. This paper reports an implementation of the InterDatabase based on DandD which is a system for Data and Description. Due to the high level of data abstraction established in a long run DandD project, the implemented InterDatabase is, in fact, a flexible environment which covers almost all fields of science and which can be used for data acquisition, data cleaning, data organisation, data visualisation, data analysis and data modelling.

1 Environmental changes of statistics

Using highly developed computers and networks, it has become faster and easier to get various data from different sources or databases. Some of these are publicly available on the Internet. A major challenge facing statisticians is the creation of powerful environments for working with such a variety of data. InterDatabase is one such environment ([9]). Here, by "Inter" we mean both "Inter-databases" to utilise different databases and "Databases on the Internet" to utilise scattered databases over the Internet. Both are closely related and it does not seem so meaningful to distinguish them once Internet access has been established. Therefore, we can simply say that InterDatabase is an environment for utilising different databases simultaneously. Before describing InterDatabase, we review other related approaches.

2 Approaches

Let us quickly review the following different approaches to the environment for working with various type of data; NetCDF, DDI and MetBroker.

2.1 NetCDF

NetCDF (Network Common Data Form, [4]) is a data abstraction for storing and retrieving multidimensional data and is distributed as a software library which provides a concrete implementation of the abstraction. The software has been developed under the Unidata Program sponsored by the US National Science Foundation to support research and education in the atmospheric sciences. This approach is closely related to our InterDatabase, but not the same. NetCDF is targeted only for multidimensional or array data. However InterDataBase is not restricted to data following such a neat format. Another point is that NetCDF requires reformatting all the data so that it accords with a common rule, called Common Data Language (CDL). This can be a burden unless the data processing procedures have been established from the beginning in each field of application.

2.2 DDI and NESSTAR

DDI (Data Documentation Initiative, [7]) is also close to InterDatabase in concept. It aims to create a universally supported metadata standard for the social science community and is implemented as an XML document. The DDI-tree contains five main branches.

1. Document Description
2. Study Description
3. Data File Description
4. Variable Description
5. Other Study-Related Materials

DDI is mainly concerned with table data such as the results of surveys. The data is therefore assumed to be a simple set of realizations of several variables, so that it would not be an easy job to describe complicated relations among variables, or to include array data in which each axis corresponds to an explanatory variable and the value of the array corresponds to the response variable. Also description procedures are not yet formalised well. For example, description of the sampling design is left free.

NESSTAR is a metadata-driven system which can be used in conjunction with DDI metadata. It searches or navigates data files, based on what is written in the DDImetadata and can display a simple summary. However the system is not aiming at manipulation of several data sets to combine it into one.

2.3 MetBroker

MetBroker([3]) is middleware which provides consistent access to heterogeneous weather databases. It is a mediator that sits between agricultural models and various sources of online data. This approach resolves data heterogeneity problems by writing a suitable program. It is efficient for meeting the needs of specific tasks, but it would be laborious to rewrite the program to meet the needs of users or to accommodate structural changes in databases.

3 Our approach

As was mentioned before, our goal is to provide a good environment to work with different types of data which might be scattered over networks. Our approach is probably closer in concept to DDI or NESSTAR. In InterDatabase, the DandD Client Server System ([8]) is driven by a DandD instance and provides a similar environment. A major difference of InterDatabase from DDI and NESSTAR is the unified general approach to providing such an environment. A high level of data abstraction is necessary to retain such generality and an intimate linkage between the abstraction and development of support softwares is indispensable.

DandD (Data and Description) is a long run project started around 1990. Preliminary works can be found, for example, in [10]. The aim of this project was to establish a formal rule of description of data. A hope was to make it possible to do an automatic analysis of data as well as to make it easier to exchange data with enough description for the aim of analysis. In the first part of the project, data abstraction was a main concern. The basic model had been established to construct necessary number of structures, relational or array, by quoting data vectors which are simple sequences of numbers. All necessary attributes are classified into three levels. The bottom level is for each data vector, the middle level is for each structure constructed, and the top level is for the whole data. The rule had been implemented by a LISP like own language, and some experimental supporting softwares had been developed.

A breakthrough had been occurred by an introduction of XML as a media for implementation of DandD rule in 1997. The project grew to cover various data which are not necessarily included as a body of the XML document. This led to an introduction of the concept of External Data Vector and further led to the idea of InterDatabase. Therefore InterDatabase is a natural extension of our original idea of DandD and it is now a part of DandD together with its support system, DandD server client system ([8]). Let us focus our attention into the closely related features of DandD to InterDatabase.

4 DandD

DandD is a generic name for a system consisting of the following three elements.

1. DandD rule: A syntax and semantics for describing data. The syntax is currently written as a DTD.

2. DandD instance: A document implemented along the DandD rule for the data. Currently it takes the form of XML document.

3. DandD client server system: A software system that provides a suitable environment for handling DandD instances.

As has been mentioned before, the data itself is not necessarily a part of a DandD instance, and it allows us to implement InterDatabase.

4.1 Data vector

In DandD, any data are decomposed into several data vectors. The data vector here is defined as a simple sequence of numbers, which becomes a body of elements denoted by the tag `<DataVector>` in a DandD instance. To keep consistency, only sequences of numbers are allowed as the vector. Therefore, for example, categorical data is always converted to a sequence of numbers and the coding information is attached to the elements as an attribute `Code`. If the body is empty, the three attributes `Access`, `Protocol` and `PostProcessing` tell us all the necessary information to get the body from outside DandD.

The attribute `Access` tells us where to access by its `IPAddress` attribute, and any information needed for the access, for example, user I.D. or password by other attributes. The attribute `Protocol` tells us the physical network protocol for the access and other information such as, for example, a query sentence which is needed to extract the data from a database. The attribute `PostProcessing` is for converting the data obtained from a data server to a simple sequence of numbers. Then the sequence can be regarded as if it were the body of the `DataVector` in the DandD instance. The following is an example of `DataVector`, where the body is empty and should be obtained from a relational database system.

Example 1

```
<DataVector  Id="i1"  LongName="Year" Access="a1"
@Protocol="b1" PostProcessing="c1"/>

     ...

<Access Id="a1" IPAddress="131.113.65.1" UserId="dandd"/>
<Protocol Id="b1" Physical="TCP"/>
<JDBC DatabaseServerType="postgresql" DatabaseName="KobeQuake">
@select year from kobequake
</JDBC>
</Protocol>
<ScanFormat Id="c1">
%s-%*s-%*s
</ScanFormat>
```

The attributes `Access` and `Protocol` of the `DataVector` with I.D. `i1` refer to the elements with I.D.s `a1` and `b1`, respectively. The two attributes `IPAddress` and `UserId` of the Access `a1` tell us the IP address and the user I.D., respectively. The attribute `Physical` of the `Protocol` with I.D. `b1` tells us that the physical access protocol to the data server is `TCP`. Currently the only other available protocol is `UDP`, but more protocols will be introduced according to need. This element has a sub element `JDBC` which further tells us the software protocol to communicate with the data server. The JDBC is an interface to access a database through Java language([5][6]), which absorbs differences of database servers. Other available protocols allowed here are `FTP` and `HTTP`. The attributes `DatabaseServerType` and `DatabaseName` of `JDBC` tell us that the database server is PostgreSQL and that the name of the database to be accessed is KobeQuake, respectively. In fact, the database is a record of the disastrous earthquake that occurred in the Kobe area in Japan on 17 January 1995, and the example above is a part of a DandD example instance `KobeQuake.dad` which is available from the DandD project home page

`http://www.stat.math.keio.ac.jp/DandD`.

The body of `JDBC` is a Structured Query Language (SQL) sentence which gets a column `year` from the table `kobequake` in the relational database `KobeQuake`.

The last element `ScanFormat` is referred to by the I.D. `c1` in the attribute `PostProcessing` of the `DataVector` specifies the processing method after receiving a response from the database server. The response is not necessarily a sequence of numbers and often has to be converted to fit the DandD requirement that the body of `DataVector` is a sequence of numbers. In this example, the response to the query is a sequence of dates of the form of YY-MM-DD and what is needed as the body of this `DataVector` is only the YY part. The body of `ScanFormat` specifies the extraction method by a formula. The syntax is the same as that of the function `scanf` in the language C. Other elements which can be referred to, together with or in place of `ScanFormt`, are `PrintFormat`, `Arithmetic`, `Media` and `Movie`. The `PrintFormat` is used for adding something to each element of the sequence. The syntax is the same as that of the function `printf` in the C language as well. Although this `PrintFormat` element does not appear in this example, it becomes necessary, for example, to add the prefix 20 to all YY to make a four digit representation of the year for consistency with other data vectors obtained from different data sources. The `Arithmetic` is used for a more complicated manipulation, arithmetic operation on each element of the sequence. This is used, for example, when the conversion of the unit from centimetre to metre, or sexagesimal to digit is necessary. The other two functions are experimental and support the case when the response from data server is an image or a movie. In the attribute `PostProcessing` of the `DataVector`, several such manipulations of the element of the response can be referred. Each manipulation is assumed to be applied in order.

Besides the elements mentioned above, the `Code` element can be also referred together in `PostProcessing`. The primary role of the `Code` element was to provide coding information for categorical data. The body is a sequence of quoted strings, which provides codes for natural numbers in the body of `DataVector`, and it is usually referred to in the attribute `Code` of the `DataVector`. If it is referred to in the attribute `PostProcessing`, it means that each element of the current sequence is matched to the body of the `Code` and converted to the matched index. This `Code` is used as a default `Code` attribute of the `DataVector` as well. The conversion of the labels of the levels of categorical data can be described if two different `Code`s are given to the attributes `PostProcessing` and `Code` of a `DataVector`. Then, the code given in `PostProcessing` indicates the code used for the database or for the data file, and the code given in the attribute `Code` indicates the code which should be used in the DandD instance. If both are missing, the obtained sequence is regarded as a sequence of numbers.

The reason why we provided such functionalities at the level of data acquisition from a data server comes from our design principle. The principle of InterDatabase is to provide a good flexible environment for working with various types of data, and so it is better to do any necessary conversions at the stage of data acquisition outside DandD. We are then free from the differences of the data sources. An alternative would be to modify the existing databases or data files according to the needs of the user or to create a new database. However, this is not only laborious but also inefficient for the case of huge data sets which are rarely used as a whole. In InterDatabase, no modification is necessary to the existing data sources. This principle is close to that of DDI or MetBroker, but InterDatabase provides a more general and flexible way of resolving such differences than others since it is free from any particular software. It is sufficient to write explicitly any necessary information in the form of XML, which is needed for processing, analysis, modelling and its utilisation.

4.2 Data

The `DataVector`s defined in the element `DataBody` are organised into several structures within the element `Data`. Two types of structures are available; `Relational` or `Array`. The relational model is general enough to represent any relations among data ([2]). The relational database under the frame work of the relational model is now a standard for database systems because of its generality and ease of system maintenance. A relation is a collection of variables and the realization is a collection of data vectors each of which is a sequence of realized values of each variable. The realization looks like a table and it is usually called a *table* in the Relational Database Management System (RDBMS).

Caution is necessary when using the word *table*. A contingency table or the result of a designed experiment is also called a table in statistics. However such a table is not a table in the sense of RDBMS. In the relation model, each row in the *table* is regarded as a point in the value space of the variables, so that the *table* is nothing more than a set of such points. Therefore, the position of each row in the *table* has no specific meaning. This is in contrast to, for example, a two dimensional contingency table, in which two hidden variables exist, say row index and column index variables, besides the variable for the values in the table. Therefore, it should be reorganised as a *table* in RDBMS, of two index variables and a variable for the values of the table. Each index variable then repeatedly takes the same value as many times as the number of rows or columns. To avoid such a redundancy, we allow an array structure besides the relational structure in DandD, since such a table or multidimensional array frequently appears as a neat data structure and it becomes cumbersome to represent it as a relation. Example 2 gives a practical example of a relational structure in DandD.

Example 2

```
<Data>
 <Relational Id="Futures" LongName="TSLongNameE TSLongNameJ"
      MainKey="dly dlm dld cmdty dvy dvm mkt"
      Control="dly dlm dld"  Nominal="cmdty dvy dvm mkt">
   <Value Id="dly" LongName="Dealing Year"
      RefId="Dealing_Year02 Dealing_Year03" Systems="t1"/>
   <Value Id="dlm"  LongName="Dealing Month"
      RefId="Dealing_Month02 Dealing_Month03" Systems="t1"/>
   <Value Id="dld"  LongName="Dealing Day"
      RefId="Dealing_Day02 Dealing_Day03"  Systems="t1"/>
   <Value Id="cmdty" LongName="Commodity Dealt"
      RefId="Commodity02 Commodity03"/>
   <Value Id="dvy" LongName="Delivery Year"
      RefId="Delivery_Year02 Delivery_Year03" Systems="t2"/>
   <Value Id="dvm" LongName="Delivery Month"
      RefId="Delivery_Month02 Delivery_Month03" Systems="t2"/>
   <Value Id="mkt" LongName="Dealing Market"
      RefId="Market02 Market03"/>
   <Value Id="op" LongName="Opening Price of a Day"
      RefId="S_price02 S_price03" Systems="i1"/>
   <Value Id="hp" LongName="Highest Price in a  Day"
      RefId="H_price02 H_price03" Systems="i1"/>
   <Value Id="lp" LongName=Lowest Price in a Day"
      RefId="L_price02 L_price03" Systems="i1"/>
   <Value Id="cp" LongName="Closing Price of a Day"
       RefId="E_price02 E_price03" Systems=i1"/>
   <Value Id="sp" LongName="Settlement Price of a Day"
       RefId="B_price02 B_price03" Systmes="i1"/>
   <Value Id="amt" LongName="Amount of Dealings in a Day"
       RefId="Amount02 Amount03"/>
   <Value Id="oint"  LongName="Amount of Open Interest"
       RefId="OpenInterest_Amount02 OpenInterest_Amount03"/>
  </Relational>
  <Time Id="t1">
   <Year RefId="dly"/>
   <Month RefId="dlm"/>
   <Day RefId="dld"/>
  </Time>
  <Time Id="t2">
    <Year RefId="dvy"/>
    <Month RefId="dvm"/>
  </Time>
  <Interval Id="i1">
    <Min RefId="lp"/>
    <Max RefId="hp"/>
    <Other RefId="op"/>
    <Other RefId="cp"/>
    <Other RefId="sp"/>
  </Interval>
</Data>
```

This is a part of an example of DandD instance, `Futures2002-2003.dad`, describing the record of daily prices of various commodity futures from December 2002 to January 2003. The record is obtained from the site

`//ftp.tokyoweb.or.jp/tocomftp/pub/`

through FTP as a CSV (comma separated values) file for each month. In the example, relational data is defined by the tag `<Relational>` and the sub elements `<Value>` define the columns of the relational data. The reason why two data vectors are referred in the attribute `RefId` of any `Value` is that the records in the site are separated into two files, `2002-12.csv` for December and `2003-01.csv` for January. Moreover, the site changed the record format after the 1 January 2003 and the records before that day are stored in a directory `past` and newer records are stored in a directory `now`. Therefore, as in Example 1, we need to adjust the old format to the newer one. The following example illustrates a few of the definitions of such data vectors. Here we have omitted some attributes which are not essential for understanding the key points.

Example 3

```
<DataVector Id="Delivery_Year02" Access="acc1" Protocol="prt1"
          PostProcessing="Delivery_Year02scan am1"/>
<DataVector Id="Delivery_Year03" Access="acc1" Protocol="prt2"
          PostProcessing="Delivery_Year03scan"/>

     ...

<Access Id="acc1" IPAddress="ftp.tokyoweb.or.jp"
     UserId="anonymous"/>
<Protocol  Id="prt1" Encoding="Shift_JIS" Physical="TCP">
```

```
     <FTP Id="ftp1" Suffix="/tocomftp/pub/past/2002-12.csv"/>
</Protocol>
<Protocol  Id="prt2" Encoding="Shift_JIS" Physical="TCP">
     <FTP Id="ftp2" Suffix="/tocomftp/pub/now/2003-01.csv"/>
</Protocol>
<Arithmetic Id="am1">
 x+2000
</Arithmetic>
<ScanFormat Id="Dealing_Year02scan">
%4d%*2d%*2d,%*s,%*d,%*d,%*d,%*d,%*d,%*d,%*d,%*d,%*d
</ScanFormat>
<ScanFormat Id="Dealing_Year03scan">
%4d%*2d%*2d,%*s,%*d,%*f,%*f,%*f,%*f,%*f,%*d,%*d,%*d
</ScanFormat>
```

Two data vectors are defined in this example, sharing the same attribute `Access`. The attribute `PostProcessing` of the first vector says that the `ScanFormat` with I.D. `Dealing_Year02scan` and `Arithmetc` with I.D. `am1` should successively be applied to each of the lines returned by an execution of FTP protocol. We need such two step processing because the last two digits of a year are only recorded in the file `2002-12.csv` but the full four digits are recorded in the file `2003-01.csv`. To adjust the format to the newer one, the last two digits are extracted from each line of the CSV file `2002-12.csv` and converted to the four digit year representation. The attribute `Encoding` of the `Protocol` indicates that character code of the lines obtained by FTP is the shift JIS code. The reader may guess other differences of the formats of those two files. Many other different formats for the files are possible.

Consider Example 2 again. The attribute `MainKey` of the `Relational` tells us the main key of the relational data. This is the same idea as in RDBMS. Each record is identified by the combination of the indicated `Values`. This attribute together with the `ForeignKey` attribute enables us to make links between several relational data. Other attributes `Control` and `Nominal` indicate which `Value`s are factors. Possible other factor attributes are `Variable`, `Block`, `Latent` and `Auxiliary`. The concept of factor type is useful not only for applying a model like ANOVA, but also for the visualisation of data. In the example above, the attribute `Control` suggests that the specified variables constitute an x-axis of a plot and the attribute `Nominal` suggests that separate visualisations should be organised according to the values of the variables. This is an example showing how the semantics of variables can be described in a formal way. Such a formal description plays an important role in an automatic visualisation or a semi automatic data analysis.

It is crucial to describe several relations among *variables* by the `Systems` attribute of `Value`. Note that the *relation* here is not the same as that in the relational model, which is a relation of the given records as a set of points in a value space. In Example 2, two `Time` relations and an `Interval` relation are defined. The Time indicates that the specified Values constitute a calendar system. The Interval indicates that the four Values are closely related, constituting an interval given by Min and Max with several aggregated values, the opening price, the closing price and the settlement price given by Other. Note that futures price moves time by time in a day.

5 InterDatabase implemented by DandD

We have briefly explained some important aspects of a DandD instance, in conjunction with InterDatabase. One of the advantages of our implementation is that any number of DandD instances can be created for the same set of data. The data view is not unique and changes from stage to stage and from person to person. For example, it would be natural to organise data reflecting the way it has been collected, describing the attributes and background information as precisely as possible for possible future needs. However, we may find a better data organisation and a necessary and sufficient description after browsing the data or even after modelling. Also, the data view heavily depends on the aims of data utilisation. For example, the choice of response variable depends on the aim. It would not be realistic to fix such a view, and rather natural to allow different views for a set of data. In DandD, different views can be easily created. It is usually enough to modify an existing DandD instance. No modification of data sources is necessary. As a consequence, many DandD instances will be produced in the course of data acquisition and modelling. Thus a mechanism is necessary to organise different data views or DandD instances.

5.1 Network of DandD instances

A mechanism Relatives is available in DandD instance. This make it possible to construct a network of DandD instances, to trace changing data views. The element Relatives has sub elements Parent, Child and Sibling. Each element indicates which DandD instances are the parents, children or siblings through its URL. Then, other related views can be easily searched on the Internet and accumulated as experiences, without maintaining any library. We hope that such an accumulation leads to more productive work with data. Also, other mechanisms Model or Summary should help the user obtain a better understanding of the data.

5.2 Data updates

Data is updated day by day or time by time. As a consequence it is important to have a mechanism to pursue the changes in a database. Further work is necessary to implement such a mechanism, although our InterDatabase is robust enough to accommodate such changes.

References

[1] DandD Project. (2004). *DandD Home Page.* `http://www.stat.math.keio.ac.jp/DandD/`.

[2] Date C. (2003). *An introduction to database systems.* 8th Edition, Addison-Wesley, Boston.

[3] Laurenson M., Otuka A., Ninomiya S.(2002). *Developing agricultural models using MetBroker mediation software.* Journal of Agricultural Meteorology **58** (1), 1−9.

[4] Rew R., Davis Gren.(1990). *NetCDF: An interface for scientific data access.* IEEE Computer Graphics and Applications, **10** (4), 91−99.

[5] Sun Micro Systems. (2004). *Java Technology.* `http://java.sun.com/`.

[6] Sun Micro Systems. (2004). *JDBC Technology.* `http://java.sun.com/`

[7] The Norwegian social science data services. (1999). *Providing global access to distributed data through metadata standardisation - the parallel stories of NESSTAR and The DDI.* Working Paper 10, UN/ECE Work Session on Statistical Metadata, Geneva.

[8] Yokouchi D. (2004). *DandD client server system.* Compstat 2004, Physica-Verlag, Prague.

[9] Yokouchi D, Shibata R.(2001). *InterDatabase - DandD instance as an agent on the Internet* - (in Japanese). Proceedings of the Institute of Statistical Mathematics, **49** (2), 317−331.

[10] Shibata R., Sibuya M. (1987). *Formal description of data type for statistical analysis.* Proceedings of the first IASC world conference, 203−212.

Address: R. Shibata, Keio University Yokohama, Japan

E-mail: `shibata@stat.math.keio.ac.jp`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

EXPLORATORY VISUAL ANALYSIS OF GRAPHS IN GGOBI

Deborah F. Swayne and Andreas Buja

Key words: GGobi, graphical software, system R, covariate displays.

COMPSTAT 2004 section: Data visualisation.

Abstract: Graphs have long been of interest in telecommunications and social network analysis, and they are now receiving increasing attention from statisticians working in other areas, particularly in biostatistics. Most of the visualization software available for working with graphs has come from outside statistics and has not included the kind of interaction that statisticians have come to expect. At the same time, most of the exploratory visualization software available to statisticians has made no provision for the special structure of graphs.

Graphics software for the exploratory visual analysis of graph data should include the following: graph layout methods; a variety of displays and methods for exploring variables on both nodes and edges, including methods that allow these covariate displays to be linked to the network view; methods for thinning or otherwise trimming a large graph. In addition, the power of the visualization software is greater if it can be smoothly linked to an extensible and interactive statistics environment.

In this paper, we will describe how these goals have been addressed in GGobi through its data format, architecture, graphical user interface design, and its relationship to the R software [7].

1 Introduction

A graph consists of nodes and edges; the edges connect pairs of nodes. In social network analysis, the nodes frequently represent people or institutions; the edges represent interactions such as conversations or trading relationships. The graphs encountered in telecommunications are similar: the nodes typically represent telephone numbers or IP (Internet Protocol) addresses; the edges capture telephone calls or exchanges of packets.

For a data analyst studying graph data, the description of the graph is often only part of the story, because the nodes and the edges may each correspond to multivariate data. For example, if the graph captures a set of telephone numbers and telephone calls, we may have demographic data or usage data about the bill-payer for each telephone number, and we may also know the time and duration of phone calls. We therefore observe variables on nodes and on edges.

How do exploratory data analysts approach such data? First, we need to visualize the graph, that is, to lay it out by using node positions that have been calculated to help us interpret the graph structure. This is not a well-defined objective, but often the distance between nodes in the layout should reflect their distance from one another according to some distance metric on the graph. Another guideline is that minimizing edge crossings usually makes a graph more readable by cutting down on clutter. Still, there is no "best" layout method, or even a best layout for a particular graph: for example, one layout may clarify a graph's overall structure while deemphasizing local structure, while in another layout, a local region of interest may be clearly drawn but the overall structure looks like spaghetti. Graph layout in an interactive context, then, should offer several layout algorithms and a lot of interaction methods for tuning and exploration.

The layout algorithms should be fast enough to be used in real time. For example, we might draw only straight-line edges, and we might not sacrifice any time to choose the perfect position for node labels. The suite of layout algorithms should include methods for laying out graphs in 3D (or higher-D), which we can rotate to shift our viewpoint and focus on local structure.

Other important interaction methods include the following:

- We should be able to tune the layout by moving nodes interactively.
- We should be able to pan and zoom the display of the graph.
- We should have a variety of ways to thin or subset the graph by eliminating or collapsing nodes and edges. At times, we may not want to eliminate nodes, but to find ways to highlight nodes and edges of interest while "downlighting" the rest. In that way, we retain context as we focus on a subset of interest.

So far, we have considered only the structure of the graph, ignoring the multivariate data associated with the nodes and edges. Once the layout is displayed, one wants to explore the data together with the graph, to investigate the relationships between the variables and the shape of the graph. The use of linked views, by now a standard feature of interactive data visualization software, is well suited to this goal. The graph view can be linked to displays of multivariate data on both nodes and edges.

These additional views can be used to highlight, label or paint nodes and edges in the graph view according to variable values, so that we can explore the distribution of data values in the graph (see Fig. 2). Equally, we can highlight data in the covariate views. For example, we might want to thin the graph according to covariate values. In the case of telephone calls, we could erase the edges corresponding to the shortest calls, and then erase all the nodes that no longer have edges.

Finally, this software will be more powerful and more extensible if it can be programmed using some scripting language, and if it is connected to a software system for data analysis that includes a library of standard graph algorithms.

Graph drawing is an active research area in computer science with a long history [2]. The layouts produced are highly tuned and often beautiful. Since they are not produced within the context of data analysis, the graphics are typically not interactive, and the programmers have not adopted the linked views approach. Some tools (e.g. Pajek [1]) offer a library of graph algorithms in addition to layout, and some can even be extended with plugins (e.g. Tulip, `www.tulip-software.org`). Still, the designers clearly do not have exploratory data analysis (EDA) in mind.

Within the field of statistics, graph visualization has not gotten very much attention. A notable exception is the work of [12], which has never been released to the public. Even the social network analysis community, which combines an interest in graph drawing with an interest in multivariate data analysis, has not to our knowledge produced tools which combine both sets of visualization capabilities. We therefore feel that there exists a gap in current software offerings for the exploration of graph data. GGobi [10] is our attempt to fill this gap.

This paper is structured as follows. Section 2 introduces GGobi, the software which will be discussed in the rest of the paper. Section 3 describes GGobi's methods for graph layout. Section 4 describes some of GGobi's methods for manipulating displays, especially graph views. Section 5 explains how GGobi can be embedded in other software, and what this design offers for graph data analysis. Section 6 describes the data format that is used to specify relationships between nodes and edges, graph elements and variables. We use a real telecommunications dataset for illustration throughout the paper. The meaning of its variables has been masked to protect the privacy of the customers.

2 GGobi

GGobi is general-purpose multivariate data visualization software, designed to support EDA. GGobi displays include scatterplots, scatterplot matrices, barcharts, time series plots, and parallel coordinate plots. All displays can be linked for color and glyph brushing as well as for point and edge labeling. GGobi is known for its powerful projection facilities for high-dimensional rotations. Among GGobi's many other manipulations are panning and zooming, subsampling, and interactive moving of points and groups of points in data space.

GGobi can be easily extended, either by being embedded in other software or by the addition of plugins; either way, it can be controlled using an Application Programming Interface (API). An illustration of its extensibility is that it can be embedded in R.

GGobi is a direct descendent of a data visualization system called XGobi [9] that has been in use since the early 1990's. XGobi supported the specification and display of graphs, but it did not include any graph layout methods. Graph data was an afterthought with XGobi, while it was a consideration in the GGobi design process from the beginning.

GGobi supports a plain ASCII format involving multiple input files (as in

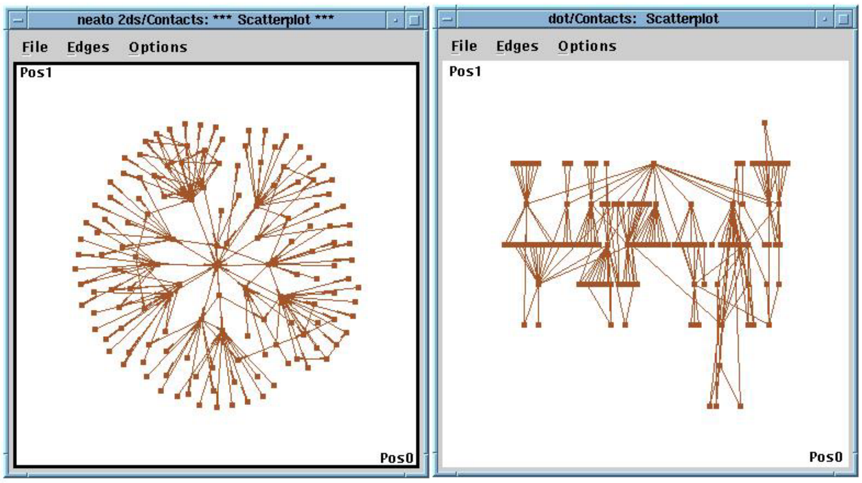

neato 2ds/Contacts: *** Scatterplot ***
File
Edges
Options
Pos1
Pos0
dot/Contacts: Scatterplot
File
Edges
Options
Pos1
Pos0

Figure 1: These two displays show layouts of the *snetwork.xml* data generated by the GraphViz layout methods. On the left is a 2-D “neato” layout; on the right a “dot” layout.

XGobi) for the simplest data specifications, but an XML (Extensible Markup Language) file format has to be used for anything richer, and graphs are an example. The format is briefly described in Section 6.

3 Graph layout

We have used GGobi’s plugin mechanism to add graph layout. Because this is specialized software, it is convenient that this functionality can be optional. There are two plugins available for GGobi that can be used for laying out graphs.

3.1 The graph layout plugin

The simplest plugin is called GraphLayout. It includes three layout methods, two of which rely on the library included with GraphViz [6], a freely available collection of tools for manipulating graph structures and generating graph layouts. All three methods work by generating a new dataset on the fly and making it available through the GGobi interface, so scatterplots of the new position variables can be displayed, and edges added to them.

The three layout methods are:

Radial: The radial layout [12] places a designated node at the center, and arranges the rest of the nodes in concentric circles around it. The resulting layout is a tree arranged radially, with any extra edges added. If the underlying graph is not very tree-like, the layout can result in a great many edge crossings, and the layout doesn't do anything to minimize these crossings. In addition to the two position variables, the method generates a few other variables, such as the number of steps between node j and the center.

Dot: "Dot" produces hierarchical layouts of directed graphs in 2D; the other layout methods ignore edge direction. It first finds an optimal rank assignment for each node, then sets the vertex order within ranks, and finally finds optimal coordinates for the nodes.

Neato: The "neato" layout algorithm produces "spring" model layouts of undirected graphs. In spring models, the graph is modelled as a set of objects connected by springs, assuming both attractive and repulsive forces, and an iterative solver is used to find a low-energy configuration. Only the positions at the final configuration are returned by the algorithm. Neato is the most general-purpose method of the three. Further, neato can generate layouts in spaces from 2D to 10D, and edge weights can be used to further tune the layout.

The first layout method is illustrated in Fig. 2; the latter two are illustrated in Fig. 1.

There is a manual for the plugin which describes its use in more detail. The dot and neato layout methods are described in the GraphViz documentation, which can be found on
`www.research.att.com/sw/tools/graphviz/refs.html`.
The GraphViz software can be obtained from `www.graphviz.org`.

3.2 The ggvis plugin: multidimensional scaling

The "ggvis" plugin is a reimplementation of XGVis [4] a multidimensional scaling (MDS) tool which is part of the XGobi software. MDS is a method for visualizing data where objects are characterized by dissimilarity values for all pairs of objects. It interprets these dissimilarities as distances and constructs maps in R^k. It was originally developed as a data analysis method in the social sciences, but it is also used to lay out graphs.

Like neato, ggvis computes layouts through iterative optimization, but unlike neato, the display is redrawn at each iteration, so we can watch the layout take shape. We can also intervene during the optimization process, by moving points interactively when they are trapped in local minima, or by adjusting parameters of the MDS objective function.

GGVis puts a large number of parameters under interactive user control. As a consequence, ggvis layouts are highly tunable. One of the most useful ggvis parameters is the exponent of a power transformation of the target distances; lowering it below one lets the short distances dominate, while exponents greater than one expose the long distances. This lets us decide whether we want to spread the leaves out, highlighting the structure in the leaves, or to collapse them, revealing the connectivity in the interior of the graph.

In addition to parameters, we can make use of color and glyph groupings of the nodes. We may subselect one group at a time for layout, or we may lay out the groups simultaneously but as unconnected graphs. Or we may lay out a subgroup and use it as an anchor set for laying out the remaining nodes.

There is also a diagnostic plot that permits us to judge how closely the pairwise distances in the layout match the target distances.

3.3 Multiple edge sets

Sometimes one wants to compare different edge sets for the same set of nodes. In the case of telephone calls, for instance, the extended community associated with a phone number changes from week to week, with changes both in the set of phone numbers in the community, and in the total length of the conversations between any pair of nodes.

One strategy to compare these different edge sets is to start by determining a layout based on the union of all nodes and the union of all edges. Since any of the edge sets can be associated with the set of nodes used to determine the layout, it's easy to compare them: Open multiple scatterplots of the nodes in the graph view, and assign a different edge set to each one. That technique could even be the basis for an animation of edges and edge variables over time.

4 Graph exploration

Once the layout has been produced and the graph is displayed, a great deal of exploration is possible without using any further plugins. Most of this functionality depends on using linked views. As one would expect, nodes in the graph view are linked to points in scatterplots of node variables, or to bars in a barchart; this is a familiar style of linking. It is perhaps less obvious that an edge in the graph view and a point in a scatterplot of edge variables are also linked: these are just different ways of rendering the same record. Here are some of the manipulations available in GGobi:

Move Points: In this mode, any point can be moved to manually tune the layout. To move a group of points, one brushes them with a common glyph and color; by moving any member of the group, one moves the whole group. Under certain circumstances, point motion can be linked across plots of layouts, namely, when the nodes are shared across graphs that differ only in edge sets and share a single layout in separate windows.

Edit Edges: To edit the graph interactively, add nodes (by clicking the mouse where you want the new node to appear) and edges (by pressing down the mouse button at the source node and dragging the edge to the destination). To view or modify the default properties (such as record label or variable values), use the left button; to simply have the new record added quickly, use the right or middle button. To delete nodes or edges, use "shadow" brushing as described below.

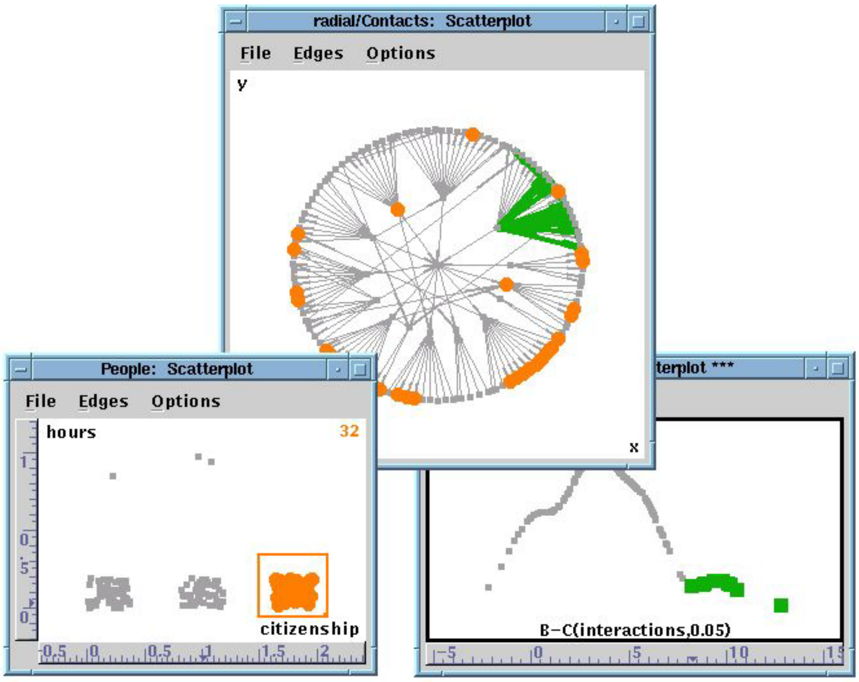


Figure 2: An illustration of linked brushing with graphs. The nodes in the graph are linked to the data in the scatterplot at the lower left; the edges to the data in the scatterplot at the lower right.

Identify: When the identification mode is active, bringing the cursor near a point causes a label to be displayed, both in the current display and in other displays. By default, this is the case label supplied in the data file (or the row number), but it can also be a list of variable name - value pairs or an id. If edge identification is selected, the nearest edge will be labelled instead of the nearest point.

Brushing (interactively): Linked brushing is probably the most familiar use of linked views. In the case of graphs, it is probably clear by now that it can be used in at least two ways. First, a plot of node data is linked to a graph view such that brushing points in one plot causes the same points to change color or glyph in the other. Second, a plot of edge data is linked to the graph view such that brushing points in the edge data plot affects the edges in the graph view, and vice versa. This latter functionality is an innovative feature of ggobi.

One brushing style allows a point or an edge to be “shadow” brushed, so that it’s drawn in a faint color and can later be removed from the displays altogether.

Fig. 2 shows linking between a radial layout of the *snetwork.xml* data and two scatterplots. Two rectangular arrays of data are involved, one for the nodes and the other for the edges. The window at the lower right contains a 1-D plot (an ASH, or Average Shifted Histogram ([8]) of a transformation of one of the edge variables, *interactions*. The highest values have been brushed with large green rectangles (rendered in dark gray in the gray-scale printed version of this paper), and the corresponding edges in the radial layout view are wide and green. All the green edges are connected to a single node, which tells us that a single individual participates in all of the longest interactions in the data. The window at the lower left contains a jittered scatterplot of *hours vs citizenship*, the two variables recorded for each person. The points representing the people with the highest values of the citizenship variable (visa holders) have been brushed with large orange circles (rendered as large medium-gray circles in gray scale), and the corresponding points are brushed in the graph view. A couple of subgraphs contain no visa holders at all, and a couple of other subgraphs are dominated by visa holders, but we also see a great deal of interaction between visa holders and other people in the data. (Recall that the data is actually about telephone calls, but that its meaning has been thoroughly obscured to protect customer privacy.)

The line characteristics (color, type and thickness) are implied when the point characteristics (color, type and size) are specified in the *Choose color & glyph* panel.

One of the options available in the brushing mode is shadow brushing [3]; that is, to select points or edges to be drawn in a “shadow” color, close to the color of the background. This is especially appealing for graph visualization because clutter is often severe, yet we often don’t want to lose sight of the graph structure when viewing a subset of the data. (Sometimes, of course, we don’t want to draw those points at all, even as shadows, and then we exclude them using the *Color & glyph groups* tool.)

Coloring by variables: Since interactive brushing of continuous variables can be tedious, an automatic scheme is available as part of the *Color schemes* tool. In the *snetwork.xml* data, one of the edge variables (*interactions*) is continuous, so we can choose a sequential color scale and apply it to the “Contacts” edge set using the *interactions*. (Since the distribution of that variable is highly skewed, we might also apply a transformation first.)

Panning and Zooming: It is essential to be able to zoom in on interesting regions of the graph view, and that functionality is available in GGobi’s scale mode. (GGobi displays are not linked for scaling.)

All these methods are described in more detail in the GGobi manual, available on `www.ggobi.org`.

4.1 The graph manipulation plugin

All of the interactive methods just listed are useful for multivariate data, not just for graphs. In addition to those methods, we have added a plugin for methods of exploration that are peculiar to graphs. It has two functions as of this writing, both of them designed for focussing on contiguous subsets of the graph.

The first function responds to a button click by shadow-brushing leaf nodes and the edges connected to them recursively until no leaf nodes are highlighted. It can be a useful way to quickly hide a lot of clutter in a a messy graph, and get a look at the center.

The second is a method for focussing on a node and its nearest neighbors. It is used in conjunction with the *Identification* mode in GGobi. Move the cursor near a point of interest, and then click a mouse button. All points will be shadow brushed with the exception of the nearest point and its neighbors within one or two steps. In this way, one can walk around the graph, focussing on one small neighborhood at a time.

5 Graphs in GGobi's API

While GGobi is a stand–alone application, it has been designed and constructed as a programming library and can be embedded within other applications. It has a large, and still evolving, Application Programming Interface (API) which developers can use to integrate the GGobi functionality with other code. For data analysts, GGobi becomes much more powerful once it is embedded in a statistics environment with an extension language.

Our most developed example is the Rggobi package, which allows GGobi to be embedded in the R process. Users can then launch GGobi (using R data frames or data files outside R), and then read and set data values and case attributes (such as color and glyph), and even add event handlers which cause R to respond to GGobi events. Edge sets can also be added, and the attributes of edges (color, line type and line thickness) are handled exactly like point attributes.

In this first simple example, we create a matrix to represent the nodes, and open it in ggobi. We next create an empty data set, dimensioned to hold six records. Finally we create a 6 by 2 array to define the edges as 6 rows of source - destination pairs, named in terms of the node labels, and add the edge set to the running ggobi.

```
x <- matrix(c(0,0,2,1, 0,2,0,1, 0,0,0,1), 4, 3,
     dimnames = list(c("a", "b", "c", "d"), c("X", "Y", "Z")))
gg <- ggobi(x)

d2 <- gg$createEdgeData(6, name="edges")
e2 <- rbind(c("a","b"), c("b","c"), c("a","c"),
            c("a","d"), c("b","d"), c("c","d"))
gg$setEdges(e2, edgeset = d2)
```

In the second example, we deal with a more complex case, in which there are variables corresponding to the edges as well as to the nodes. We start again, using the matrix x just described. Next we add a second dataset, 3 by 2, composed of the data corresponding to the edges. Finally we add three edges to the second dataset.

```
gg <- ggobi(x)

z <- matrix(c(1,2,1, 1,2,2), 3, 2,
     dimnames = list(letters[10:12], c("X", "Y")))
d2 <- gg$setData(z, name="z")

e1 <- rbind(c("a", "b"), c("b", "c"), c("a", "d"))
gg$setEdges(e1, edgeset=gg[["z"]])
```

We plan to extend the API and the Rggobi package so that they can work with other graph packages currently under development as part of the Bioconductor project (www.bioconductor.org).

6 Data format: Specifying graphs in XML

GGobi relies on XML (the Extensible Markup Language) for everything beyond the simplest of input data. The use of XML has allowed us to design a system of mark-ups or tags that describe one or more datasets in great detail within a single file, even specifying the relationships between records in different datasets.

We based GGobi's XML format on a pre-existing XML format designed for the Omegahat project (www.omegahat.org) and the S language (R and S-Plus). Some of the information that can be specified in the GGobi XML file includes variable types and axis ranges, the symbol and color corresponding to a record, and multiple data sets and the rules for linking them.

GGobi's XML format is described elsewhere [11], so we will only explain here how the specification of data records is used to describe graphs. A data record specification may be as simple as this:

```
<record> 1.0 2.5 </record>
<record> 1.7 2.2 </record>
```

This is a pair of records for a dataset with two variables. If we want to identify these records as nodes, we must also give them unique ids. (Ids can also be used for linking and identification, but that usage is described elsewhere.)

```
<record id="Macbeth"> 1.0 2.5 </record>
<record id="Banquo">  1.7 2.2 </record>
```

If we want a set of edges to be drawn on a scatterplot or a graph view of these nodes, we need a second dataset. If there is to be an edge from "Macbeth" to "Banquo," the second dataset must contain a record like this:

```
<record source="Macbeth" destination="Banquo"> </record>
```

If there are variables corresponding to that edge, they are specified within the record, just as they are for nodes.

```
<record source="Macbeth" destination="Banquo">
   27 42 4.6
</record>
```

As we implied in Section 3.3, it's possible to specify more than one edge set corresponding to the same node set within the same XML file, and that offers a way to compare related edge sets.

There are graph specification languages in XML under development, and we expect it will be easy to translate between those formats and GGobi's, though those other languages probably won't fully support multivariate data.

For the interested reader, the GGobi distribution includes several graph datasets in XML. Some include position variables so that additional layout isn't required: *buckyball.xml* and *cube6.xml* describe geometric objects, with no additional variables. Another, *snetwork.xml,* is fully multivariate and does not include variables that can be used for displaying the graph; that is the dataset that served as an example throughout this paper.

7 Conclusions

As more statisticians become interested in graph data analysis, they approach this area with the expectations and expertise acquired in working with general multivariate data. They expect first of all to be able to work in environments like R, with a set of algorithms, a variety of static display methods, and a scripting language. This set of goals is being pursued in the Bioconductor project and elsewhere.

Second, statisticians and other data analysts who have come to rely on direct manipulation graphical methods will want to use them with this form of data as well: to quickly update plots, changing variables and projection, to pan and zoom displays, and to use linked views to explore the graph and the distribution of multivariate data in the graph. GGobi's data format supports describing the graph and the data together, and its architecture allows the addition of plugins, so it's natural to extend GGobi, applying all its functionality to graph data.

Finally, we want to integrate the direct manipulation graphics, algorithms and scripting language so that we can use them all together. This expectation is not yet as automatic as the first two: People often still imagine building a single monolithic application that can do everything. As the example of graph data shows, however, there are many specialized problems that are often overlooked, so no monolithic piece of software can satisfy the needs of all users. If instead it's possible to integrate complementary software tools, and to extend them with plugins and packages, then even the most unusual cases can be handled without too much trouble.

The GGobi software and documentation, including several plugins and the Rggobi package, are available on the web site www.ggobi.org.

References

[1] Batagelj V., Mrvar A. (1998). *Pajek - program for large network analysis.* Connections **21**, 47 − 57.

[2] Battista G.D., Eades P., Tamassia R., Tollis I. (1994). *Annotated bibliography on graph drawing algorithms.* Computational Geometry: Theory and Applications **4**, 235 − 282.

[3] Becker R.A., Cleveland W.S. (1987). *Brushing scatterplots.* Technometrics **29**, 127 − 142.

[4] Buja A., Swayne D.F. (2002). *Visualization methodology for multidimensional scaling.* Journal of Classification **18**, 7 − 43.

[5] Chen C.-H., Chen J.-A. (2000). *Interactive diagnostic plots for multidimensional scaling with applications in psychosis disorder data analysis.* Statistica Sinica **10**, 665 − 691.

[6] Gansner E.R., North S.C. (2000). *An open graph visualization system and its applications to software engineering.* Software – Practice and Experience **30** (11), 1203 − 1233.

[7] Ihaka R., Gentleman R. (1996). *R: A language for data analysis and graphics.* Journal of Computational and Graphical Statistics **5**, 299 − 314.

[8] Scott D.W. (1985). *Average shifted histograms: effective non–parametric density estimation in several dimensions.* Annals of Statistics **13**, 1024 − 1040.

[9] Swayne D.F., Cook D., Buja A. (1998). *XGobi: Interactive dynamic data visualization in the X Window System.* Journal of Computational and Graphical Statistics **7** (1), 113 − 130.

[10] Swayne D.F., Temple Lang D., Buja A., Cook D. (2003). *GGobi: evolving from XGobi into an extensible framework for interactive data visualization.* Computational Statistics & Data Analysis **43**, 423 − 444.

[11] Temple Lang D., Swayne D. F. (2001). *The ggobi XML input format.* www.ggobi.org.

[12] Wills G. (1999). *NicheWorks – interactive visualization of very large graphs.* Journal of Computational and Graphical Statistics **8** (2), 190 − 212.

Acknowledgement: We thank the reviewer who pointed out to us that the ggvis plugin would be a good environment for implementing the Interactive Diagnostic plots for MDS as described in [5].

Address: D.F. Swayne, AT&T Labs – Research
A. Buja, The Wharton School, University of Pennsylvania Duncan Temple Lang, University of California, Davis

E-mail: dfs@research.att.com

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

PLS REGRESSION AND PLS PATH MODELING FOR MULTIPLE TABLE ANALYSIS

Michel Tenenhaus

Key words: Multiple factor analysis, PLS regression , PLS path modeling, generalized canonical correlation analysis.

COMPSTAT 2004 section: Partial least squares.

Abstract: A situation where J blocks of variables are observed on the same set of individuals is considered in this paper. A factor analysis logic is applied to tables instead of individuals. The latent variables of each block should well explain their own block and in the same time the latent variables of same rank should be as positively correlated as possible. In the first part of the paper we describe the hierarchical PLS path model and remind that it allows to recover the usual multiple table analysis methods. In the second part we suppose that the number of latent variables can be different from one block to another and that these latent variables are orthogonal. PLS regression and PLS path modeling are used for this situation. This approach is illustrated by an example from sensory analysis.

1 Introduction

We consider in this paper a situation where J blocks of variables $X_1, \ldots, X_J$ are observed on the same set of individuals. The problem under study is completely symmetrical as all blocks of variables play the same role. All the variables are supposed to be standardized. We can follow a factor analysis logic on tables instead of variables. In the first section of this presentation we suppose that each block X_j is multidimensional and is summarized by m latent variables plus a residual E_j. Each data table is decomposed into two parts: $X_j = t_{j1}p_{j1}^{'} + \cdots + t_{jm}p_{jm}^{'} + E_j$. The first part of the decomposition is $t_{j1}p_{j1}^{'} + \cdots + t_{jm}p_{jm}^{'}$. The latent variables ($t_{j1}, \ldots, t_{jm}$) should well explain the data table X_j and in the same time the latent variables of same rank $h(t_{1h}, \ldots, t_{Jh})$ should be as *positively* correlated as possible. The second part of the decomposition is the residual E_j which represents the part of X_j not related to the other block, i.e. the specific part of X_j.

We show that the PLS approach allows to recover the usual methods for multiple table analysis. In section two we suppose that the number of latent variables can be different from one block to another and that these latent variables are orthogonal. PLS regression and PLS path modeling are used for this situation. This approach is illustrated by an example from sensory analysis in the last section.

2 Multiple Table Analysis: a classical approach

In Multiple Table Analysis it is usual to introduce a super-block X_{J+1} merging all the blocks X_j. This super-block is summarized by m latent variables $t_{J+1,1}, \ldots, t_{J+1,m}$ also called auxiliary variables. The causal model describing this situation is given in Figure 1. This model corresponds to the hierarchical model proposed by Wold [16].

The latent variables $t_{j1}, \ldots, t_{jm}$ should well explain their own block X_j. In the same time the latent variables of same rank $(t_{1h}, \ldots, t_{Jh})$ and the auxiliary variable $t_{J+1,h}$ should be as *positively* correlated as possible. In the usual Multiple Table Analysis (= MTA) methods, as Horst's [6] and Carroll's [1] Generalized Canonical Correlation Analysis, orthogonality constraints are imposed on the auxiliary variables $t_{J+1,h}$ and the latent variables t_{jh} related to block j have no orthogonality constraints. We define for the super-block X_{J+1} the sequence of blocks $E_{J+1,h}$ obtained by deflation: each block $E_{J+1,h}$is defined as the residual of the regression of X_{J+1} on the latent variables $t_{J+1,1}, \ldots, t_{J+1,h}$. Figure 2 corresponds to step h. For computing the latent variables t_{jh} and the auxiliary variables $t_{J+1,h}$ we use the general PLS algorithm [16] defined as follows for step h of this specific application:

External estimation:

- Each block X_j is summarized by the latent variable $t_{jh} = X_j w_{jh}$
- The super-block $X_{J+1,h}$ is summarized by the latent variable $t_{J+1,h} = E_{J+1,h-1} w_{J+1,h}$

Internal estimation:

- Each block X_j is also summarized by the latent variable $z_{jh} = e_{jh} t_{J+1,h}$, where e_{jh} is the sign of the correlation between t_{jh} and $t_{J+1,h}$. We will however choose $e_{jh} = +1$ and show that the correlation is then positive.
- The super-block $E_{J+1,h-1}$ is summarized by the latent variable $z_{J+1,h} = \sum_{j=1}^{J} e_{J+1,j,h} t_{jh}$, where $e_{J+1,j,h} = +1$ when the centroid scheme is used, or the correlation between t_{jh} and $t_{J+1,h}$ for the factorial scheme, or furthermore the regression coefficient of t_{jh} in the regression of $t_{J+1,h}$ on $t_{1h}, \ldots, t_{Jh}$ for the path weighting scheme.

We can now describe the PLS algorithm for the J-block case. The weights w_{jh} can be computed according to two modes: Mode A or B.

In Mode A simple regression is used:

$$w_{jh} \propto X_j^{'} t_{J+1,h}, j = 1 \text{ to } J, \text{ and } w_{J+1,h} \propto E_{J+1,h-1}^{'} z_{J+1,h} \qquad (1)$$

where $\propto$ means that the left term is equal to the right term up to a normalization.

For Mode B multiple regression is used:

$$w_{jh} \quad \propto \quad (X_j^{'}X_j)^{-1}X_j^{'}t_{J+1,h}, j = 1 \text{ to } J,$$

$$\text{and } \; w_{J+1,h} \propto (E_{J+1,h-1}^{'}E_{J+1,h-1})^{-1}E_{J+1,h-1}^{'}z_{J+1,h} \qquad (2)$$

The normalization depends upon the method used. For some method w_{jh} is of norm 1. For other methods the variance of t_{jh} is equal to 1.

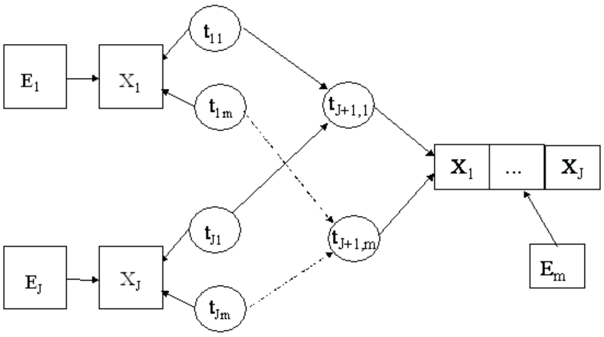


Figure 1: Path model for the J-block case.

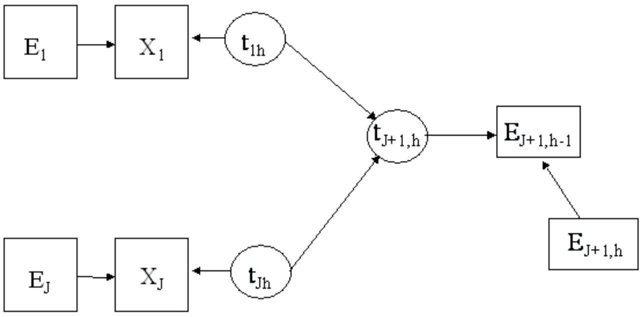
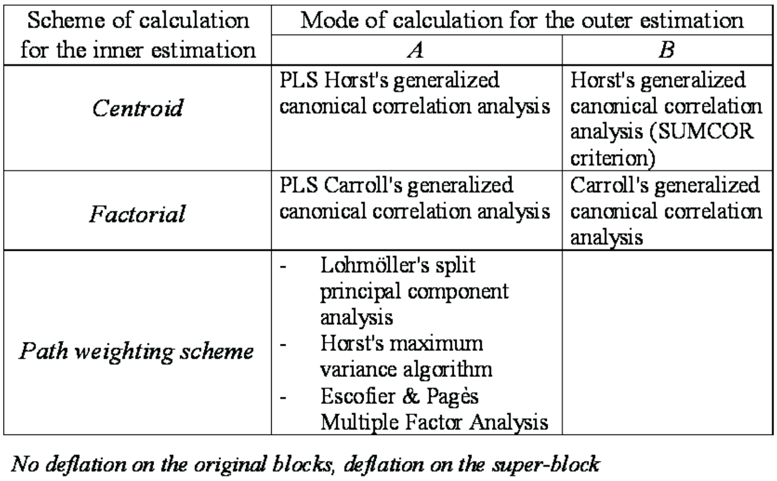


Figure 2: Path model for the J-block case : Step h.

It is now easy to check that the correlation between t_{jh} and $t_{J+1,h}$ is always positive: $t_{J+1,h}^{'}t_{jh} = t_{J+1,h}^{'}X_jw_{jh} \propto t_{J+1,h}^{'}X_jX_j^{'}t_{J+1,h} > 0$ when Mode A is used. The same result is obtained when Mode B is used. This justifies the replacement in both (1) and (2) of the internal estimation $z_{j,h}$ by the external estimation $t_{J+1,h}$.

The PLS algorithm can now be described. We begin by an arbitrary choice of the weights w_{jh}. We get the external estimations of the latent variables, then the internal ones. Using the equations (1) or (2) we get new weights. This procedure is iterated until convergence always verified in practice, but only mathematically proven for the two-block case.

The various options of PLS Path Modeling (Mode A or B for external estimation; centroid, factorial or path weighting schemes for internal estimation) allow to find again many methods for Multiple Table Analysis: Generalized Canonical Analysis (the Horst's one [6] and the Carroll's one [1], Multiple Factor Analysis [4], Lohmöller's split principal component analysis [9], Horst's maximum variance algorithm [7]. The links between PLS and these methods have been demonstrated in [9] or [11] and studied on practical examples in [5] and [10]. These various methods are obtained by using the PLS algorithm according to the options described in Table 1. The super-block only is deflated; the original blocks are not deflated.



Scheme of calculation for the inner estimation	Mode of calculation for the outer estimation	
	A	*B*
Centroid	PLS Horst's generalized canonical correlation analysis	Horst's generalized canonical correlation analysis (SUMCOR criterion)
Factorial	PLS Carroll's generalized canonical correlation analysis	Carroll's generalized canonical correlation analysis
Path weighting scheme	- Lohmöller's split principal component analysis - Horst's maximum variance algorithm - Escofier & Pagès Multiple Factor Analysis	

No deflation on the original blocks, deflation on the super-block

Table 1: Multiple Table Analysis and PLS algorithm.

Discussion on the orthogonality constraints

There is some advantage on imposing orthogonality constraints only on the latent variables related to the super-block: no dimension limitation due to block sizes. If orthogonality constraints were imposed on the block latent variables, then the maximum m of latent variables would be the size of the smallest block. The super-block X_{J+1} is summarized by m orthogonal latent variables $t_{J+1,1}, \ldots, t_{J+1,m}$. Each block X_j is summarized by m latent variables $t_{j1}, \ldots, t_{jm}$. But these latent variables can be highly correlated and consequently do not reflect the real dimension of the block. In each block X_j the latent variables $t_{j1}, \ldots, t_{jm}$ represent the part of the block correlated with the other blocks. A principal component analysis of these latent variables will give the actual dimension of this part of X_j.

It can be preferred to impose orthogonality on the latent variables of each block. But we have to remove the dimension limitation due to the smallest block. This situation is going to be discussed in the next section.

3 Multiple Table Analysis: new perspectives

We will describe in this section a new approach more focused on the blocks than on the super-block. This approach is called PLS-MTA : a PLS approach to Multiple Table Analysis.

We now suppose a variable number of common components in each block:

$$X_j = t_{j1}p_{j1}^{'} + \cdots + t_{jm_j}p_{jm_j}^{'} + E_j \tag{3}$$

A two steps procedure is proposed to find these components.

Step 1

For each block X_j we define the super-block $X_{J+1,-j}$ obtained by merging all the other blocks X_i for $i \neq j$. For each j we carry out a PLS regression of $X_{J+1,-j}$ on X_j. So we obtain m_j orthogonal and standardized PLS components $\tilde{t}_{j1}, \ldots, \tilde{t}_{jm_j}$which represent the part of X_j related with the other blocks. The choice of the number m_j of components is determined by cross-validation.

Step 2

One of the procedures described in Table 1 is used on the blocks $\tilde{T}_j = \left\{\tilde{t}_{j1}, \ldots, \tilde{t}_{jm_j}\right\}$ for h = 1. We obtain the rank one components $t_{11}, \ldots, t_{J1}$ and $t_{J+1,1}$. Then, to obtain the next components we only consider the blocks with $m_j > 1$. For these blocks we construct the residual $\tilde{T}_{j1}$ of the regression of $\tilde{T}_j$ on t_{j1}. A MTA is then applied on these blocks and we obtain the rank two components $t_{12}, \ldots, t_{J2}$ (for j with m_j *> 1*) and $t_{J+1,2}$. The components t_{j1} and t_{j2} are uncorrelated by construction, but the auxiliary variables $t_{J+1,1}$ and $t_{J+1,2}$ can be slightly correlated as we did not impose orthogonality constraint on these components. This research of components is iterated until the various m_j common components are found. These components can finally be expressed in term of the original variables.

There is a great advantage on imposing orthogonality constraints on each block components: the new m_j orthogonal and standardized components $t_{j1}, \ldots, t_{jm_j}$ are deduced from the m_j orthogonal and standardized PLS components $\tilde{t}_{j1}, \ldots, \tilde{t}_{jm_j}$ by a rotation. That means that

$$[t_{j1}, \ldots, t_{jm_j}] = [\tilde{t}_{j1}, \ldots, \tilde{t}_{jm_j}]A_j \tag{4}$$

where A_j is an orthogonal (rotation) matrix.

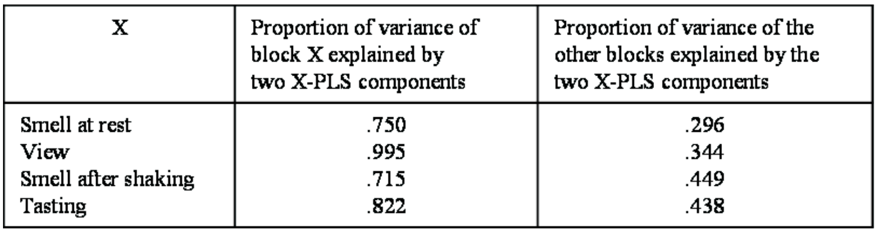
4 Application

We are going to use PLS-MTA on wine data which has been collected by C. Asselin and R. Morlat and are fully described in [3]. A set of 21 red wines with Bourgueil, Chinon and Saumur origins are described by 27 variables distributed in four blocks: X_1 = Smell at rest = [smell intensity at rest, aromatic quality at rest, fruity note at rest, floral note at rest, spicy note at rest], X_2 = View = [visual intensity, shading (from orange to purple), impression of surface], X_3 = Smell after shaking = [smell intensity, smell quality, fruity note, floral note, spicy note, vegetable note, phelonic note, aromatic intensity in mouth, aromatic persistence in mouth, aromatic quality in mouth], X_4 = Tasting = [intensity of attack, acidity, astringency, alcohol, balance (acidity, astringency, alcohol), mellowness, bitterness, ending intensity in mouth, harmony]. Another variable describing the global quality of the wine will be used as an illustrative variable.

We now describe the application of PLS-MTA methodology on these data.

Step 1

PLS regressions of $[X_2, X_3, X_4]$ on X_1, $[X_1, X_3, X_4]$ on X_2, $[X_1, X_2, X_4]$ on X_3, and $[X_1, X_2, X_3]$ on X_4 all lead to two PLS components when we decide to keep a component if it is significant (Q^2 is larger than 0.05). The X- and Y- explanatory powers of these components are given in table 2.



X	Proportion of variance of block X explained by two X-PLS components	Proportion of variance of the other blocks explained by the two X-PLS components
Smell at rest	.750	.296
View	.995	.344
Smell after shaking	.715	.449
Tasting	.822	.438

Table 2: Proportion of X and Y variances explained by the first two X-PLS components.

Then the "smell at rest" block $\widetilde{T_1} = \{\widetilde{t}_{11}, \widetilde{t}_{12}\}$, the "view" block $\widetilde{T}_2 = \left\{\widetilde{t}_{21}, \widetilde{t}_{22}\right\}$, the "smell after shaking" block $\widetilde{T}_3 = \{\widetilde{t}_{31}, \widetilde{t}_{32}\}$, and the "tasting" block $\widetilde{T_4} = \{\widetilde{t}_{41}, \widetilde{t}_{42}\}$ are defined with the standardized PLS X-components.

Step 2

The PLS components being orthogonal, it is equivalent to use Mode A or B for the left part of the causal model given in Figure 3 (PLS-Graph output [2]. Due to the small number of observations Mode A has to be used for the right part of the causal model of Figure 3. We use the centroid scheme for the internal estimation. We give in Figure 3 the MTA model for the first rank components and in Table 3 the correlations between the latent variables.

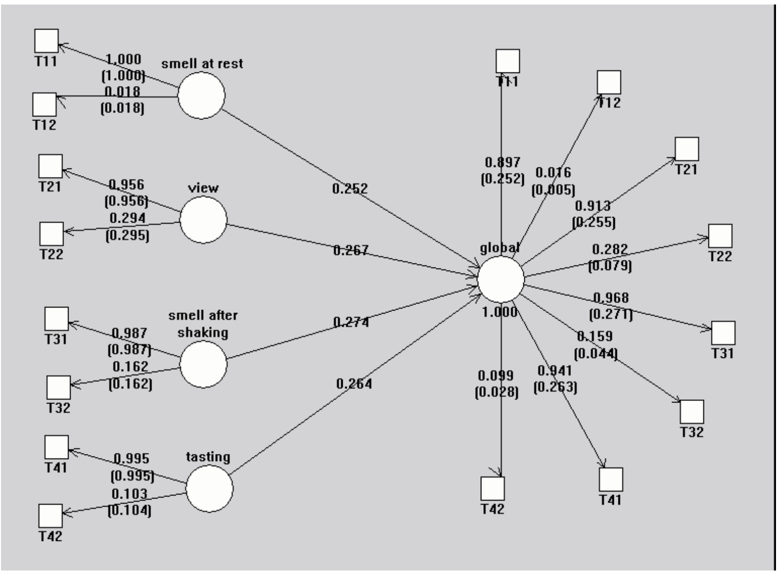
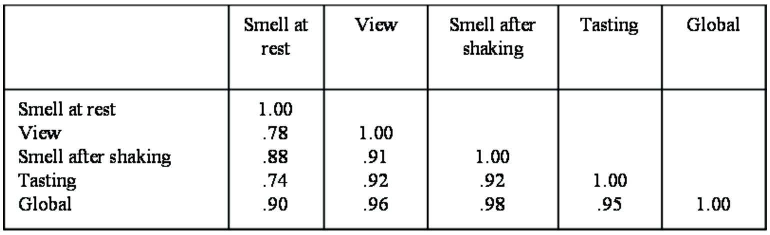


Figure 3: Path model for the first rank components (PLS-Graph output).



	Smell at rest	View	Smell after shaking	Tasting	Global
Smell at rest	1.00				
View	.78	1.00			
Smell after shaking	.88	.91	1.00		
Tasting	.74	.92	.92	1.00	
Global	.90	.96	.98	.95	1.00

Table 3: Correlations between the rank 1 latent variables.

In Figure 3 the figures above the arrows are the correlation loadings and the figures in brackets below the arrows are the weights applied to the standardized variables. Correlations and weights are equal on the left side of the path model because the PLS components are uncorrelated.

Rank one components are written as:

$$\begin{aligned}
t_{11} &= .9998 \times \widetilde{t}_{11} + .0176 \times \widetilde{t}_{12} \\
t_{21} &= .9558 \times \widetilde{t}_{21} + .2950 \times \widetilde{t}_{22} \\
t_{31} &= .9869 \times \widetilde{t}_{31} + .1619 \times \widetilde{t}_{32} \\
t_{41} &= .9947 \times \widetilde{t}_{41} + .1042 \times \widetilde{t}_{42} \\
t_{51} &= .2516 \times \widetilde{t}_{11} + .0045 \times \widetilde{t}_{12} + .2552 \times \widetilde{t}_{21} + .0788 \times \widetilde{t}_{22} + .2707 \times \widetilde{t}_{31} \\
&\quad + .0445 \times \widetilde{t}_{32} + .2628 \times \widetilde{t}_{41} + .0276 \times \widetilde{t}_{42}
\end{aligned}$$

We may note that the rank one components are highly correlated to the first PLS components $\widetilde{t}_{11}, \widetilde{t}_{21}, \widetilde{t}_{31}$ and $\widetilde{t}_{41}$.

To obtain the rank two components it is now useful to use equation (4) which here becomes:

$$[t_{j1}, t_{j2}] = [\tilde{t}_{j1}, \tilde{t}_{j2}] \begin{bmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta j \end{bmatrix} \tag{5}$$

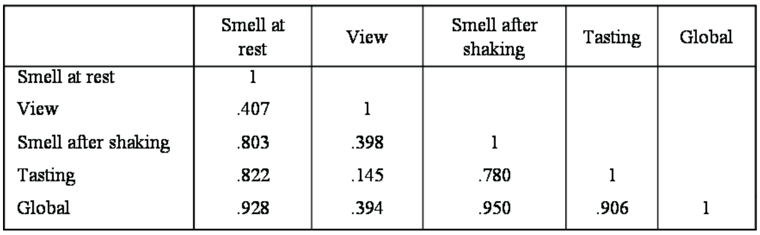
as

$$A_j = \begin{bmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta j \end{bmatrix} \tag{6}$$

is the orthogonal rotation matrix in the plan with an angle θ_j. For each of the new components $t_{11}, \ldots, t_{41}$ it can be checked that the squares of the coefficients of the PLS components $\tilde{t}_{j1}, \tilde{t}_{j2}$ sum up to one. It is then easy to get the rank two components:

$$\begin{aligned} t_{12} &= -.0176 \times \widetilde{t}_{11} + .9998 \times \widetilde{t}_{12} \\ t_{22} &= -.2950 \times \widetilde{t}_{21} + .9558 \times \widetilde{t}_{22} \\ t_{32} &= -.1619 \times \widetilde{t}_{31} + .9869 \times \widetilde{t}_{32} \\ t_{42} &= -.1042 \times \widetilde{t}_{41} + .9747 \times \widetilde{t}_{42} \end{aligned}$$

However, to get the external latent variable t_{52} for the super-block we need to apply the complete algorithm. We first regress each block $\tilde{T}_j = \left\{\tilde{t}_{j1},\ \tilde{t}_{j2}\right\}$ on t_{j1}. Then the path model used for rank one components is used on the standardized residual tables $\tilde{T}_{j1} = \{\tilde{t}_{j11},\ \tilde{t}_{j21}\}$. The results are given in Figure 4.



	Smell at rest	View	Smell after shaking	Tasting	Global
Smell at rest	1				
View	.407	1			
Smell after shaking	.803	.398	1		
Tasting	.822	.145	.780	1	
Global	.928	.394	.950	.906	1

Table 4: Correlations between the rank two latent variables.

It is more clear to express the rank two component in term of the original standardized variables. We then get the previous expressions for $t_{12}, \ldots, t_{42}$ and the following one for t_{52}:

$$\begin{aligned} t_{52} \quad = \quad & -.005 \times \widetilde{t}_{11} + .288 \times \widetilde{t}_{12} + .014 \times \widetilde{t}_{21} - .045 \times \widetilde{t}_{22} - .078 \times \widetilde{t}_{31} \\ & +.463 \times \widetilde{t}_{32} - .029 \times \widetilde{t}_{41} + .295 \times \widetilde{t}_{42} \end{aligned}$$

In Table 4 we give the correlations between the rank two components. The sensory components of rank one and two are uncorrelated by construction. The global components are also practically uncorrelated (r = -.000008).

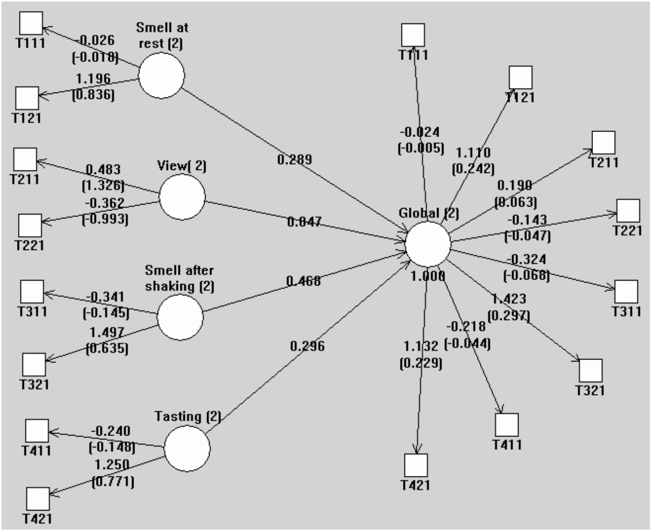


Figure 4: Path model for the second rank components in term of residuals.

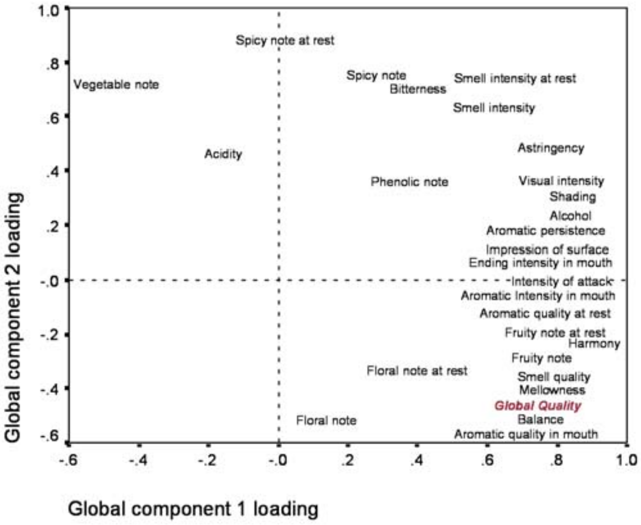


Figure 5: Variable loadings with the global components.

5 Discussion

PLS-MTA comes to carry out a kind of principal component analysis on each block and on the super-block such that the components of same rank are as positively correlated as possible. So, for each dimension h, the interpretations

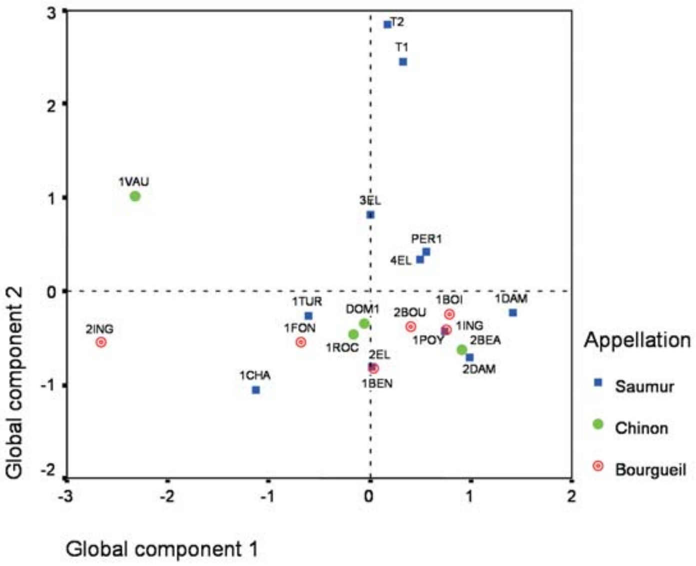


Figure 6: Wine visualization in the global component space.

of various block components t_{hj}, $j = 1, \ldots, J+1$ can be related. In Figure 5 the "Smell at rest", "View", "Smell after shaking" and "Tasting" loadings with the global components are displayed. It makes sense as the correlations of the variables with the block components and the global components are rather close. The global quality judgement on the wines has also been displayed as an illustrative variable. In Figure 6 the wines are also displayed using the global components. The best wines are located in the south-eastern quadrant.

References

[1] Carroll J.D. (1968). *A generalization of canonical correlation analysis to three or more sets of variables.* Proc. 76th Conv. Am. Psych. Assoc., 227 – 228.

[2] Chin W.W. (2003). *PLS-graph user's guide.* C.T. Bauer College of Business, University of Houston, USA.

[3] Escofier B., Pagès J. (1988). *Analyses factorielles simples et multiples.* Dunod, Paris.

[4] Escofier B., Pagès J. (1994). *Multiple factor analysis.* (AFMULT package), Computational Statistics and Data Analysis **18**, 121 – 140.

[5] Guinot C., Latreille J., Tenenhaus M. (2001). *PLS Path modellind and multiple table analysis. Application to the cosmetic habits of women in Ile-de-France.* Chemometrics and Intelligent Laboratory Systems **58**, 247 − 259.

[6] Horst P. (1961). *Relations among m sets of variables.* Psychometrika **26**, 126 − 149.

[7] Horst P. (1965). *Factor analysis of data matrices.* Holt, Rinehart and Winston, New York.

[8] Hotelling H. (1936). *Relations between two sets of variates.* Biometrika **28**, 321 − 377.

[9] Lohmöller J.-B. (1989). *Latent variables path modeling with partial least squares.* Physica-Verlag, Heildelberg.

[10] Pagès J., Tenenhaus M. (2001). *Multiple factor analysis combined with PLS path modeling. Application to the analysis of relationships between physico-chemical variables, sensory profiles and hedonic judgements.* Chemometrics and Intelligent Laboratory Systems **58** 261 − 273.

[11] Tenenhaus M. (1999). *L'approche PLS.* Revue de Statistique Appliquée, **47**, (2), 5 − 40.

[12] Tenenhaus M., Esposito Vinzi V., Chatelin Y.-M., Lauro C. (2004). *PLS path modeling.* Computational Statistics an Data Analysis (to appear).

[13] Tucker L.R. (1958). *An inter-battery method of factor analysis.* Psychometrika **23** (2), 111 − 136.

[14] Van den Wollenberg A.L. (1977). *Redundancy analysis: an alternative for canonical correlation.* Psychometrika **42**, 207 − 219.

[15] Wold H. (1982). *Soft modeling: the basic design and some extensions.* In Systems under indirect observation, Part 2, K.G. Jöreskog & H. Wold (Eds), North-Holland, Amsterdam, 1 − 54.

[16] Wold H. (1985). *Partial least squares.* In Encyclopedia of Statistical Sciences, Kotz, S. & Johnson, N.L. (Eds), John Wiley & Sons, New York **6**, 581 − 591.

[17] Wold S., Martens H., Wold H. (1983). *The multivariate calibration problem in chemistry solved by the PLS method.* In: A. Ruhe and B. Kågström (Eds), Proc. Conf. Matrix Pencils. Lectures Notes in Mathematics, Springer-Verlag, Heidelberg.

Address: M. Tenenhaus, HEC School of Management, 78351 Jouy en Josas, France

E-mail: `tenenhaus@hec.fr`

1001 GRAPHICS

Martin Theus

Key words: Statistical graphics, defaults, rendering, interaction.

COMPSTAT 2004 section: Data visualisation.

Abstract: Statistical graphics, or in more modern terms, data visualization, is not a new discipline. Whereas in the early days the construction of a graph was technically not easy and usually even required some artistic capabilities, generating statistical graphs is very easy in today's statistical software packages. This obviously leads to a less careful construction of these plots. In an object oriented software package like R we can call the generic function `plot` with almost any arbitrary object as argument, and some plot method will render this object, whether it makes sense or not.

This paper investigates how well chosen plot defaults and rendering techniques can guarantee much better results in a graphical data analysis. Furthermore, standard plots and examples of plot ensembles are presented which are suitable for analyzing variables of a specific structure.

1 Introduction

Everybody knows the phrase "A picture can be worth a 1000 words". Advocates of statistical graphical methods and data visualization sometimes use this phrase to support their position. Whereas everyone knows that there are many examples which prove that they are right, there is a far greater number of examples (although less quoted) which prove the opposite. All positive examples are usually very well thought out. E.g. Minard's visualization of Napoleon's march on Moscow is a very popular example for the power of a good visualization. The power of Minard's graph lies in the well chosen combination of spatial plotting of time series information, not to mention several artistic and aesthetic considerations, which are not that obvious at a first glance. This brings us back to the phrase "A picture is worth a 1000 words", which only holds true if the picture is really well chosen. Today, where the next statistical graphics is only one keystroke or mouse-click away, we tend to produce many graphs which would probably need more than a 1000 words to be interpreted.

In the next Section of this paper we will investigate the influence of the right choice of plot defaults on the quality, i.e. the interpretability and usability, of a graph. This should make us more alert to default plot settings, which are often inappropriate for the solution sought initially. The final section of the paper goes beyond single graphs, and shows strategies for analyzing multivariate data with ensembles of standard statistical graphs.

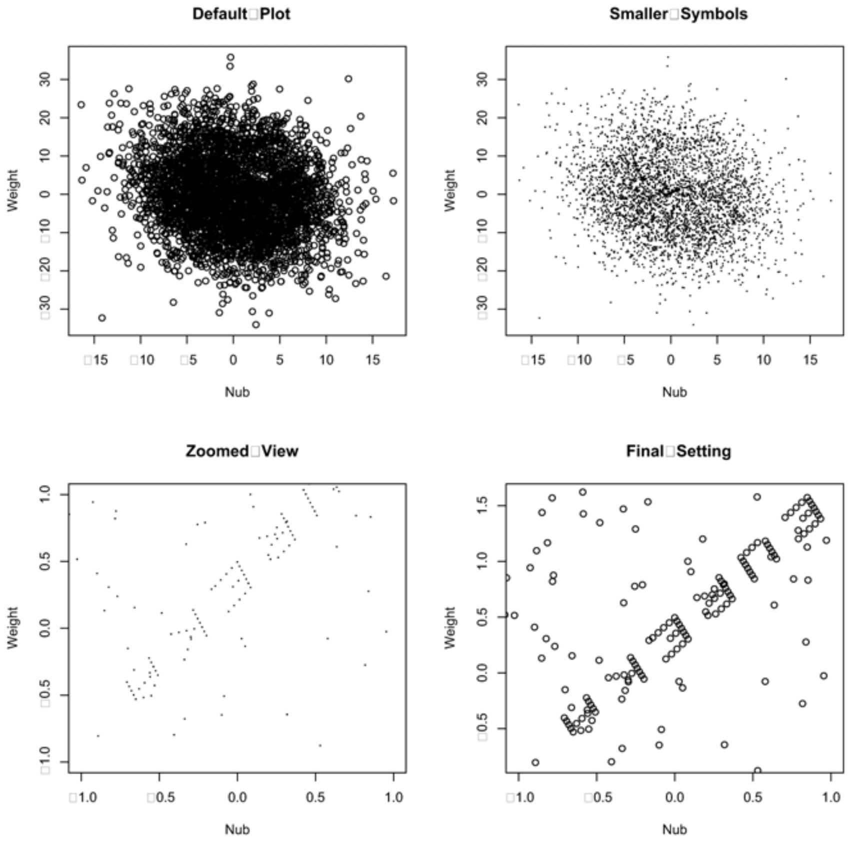


Figure 1: The pollen data plotted in R.

2 On plot defaults

2.1 The scatterplot — less can be more

A scatterplot of two quantitative variables is probably the most elementary and fundamental plot in statistics. At a first glance there do not seem to be many degrees of freedom to choose parameters to improve a scatterplot. Reviewing Cleveland [1] and [2] the only thing we can do with scatterplots is to change scales and plot symbols. Obviously Cleveland's work was written at times where pen plotter and amber CRTs were the latest technology. Furthermore, datasets with more than just a few hundred observations were very uncommon. Today's problems often look much different. A couple of thousand points are often regarded as rather small, but would have used up a whole ink cartridge of a pen plotter 25 years ago. This calls for new, advanced rendering strategies.

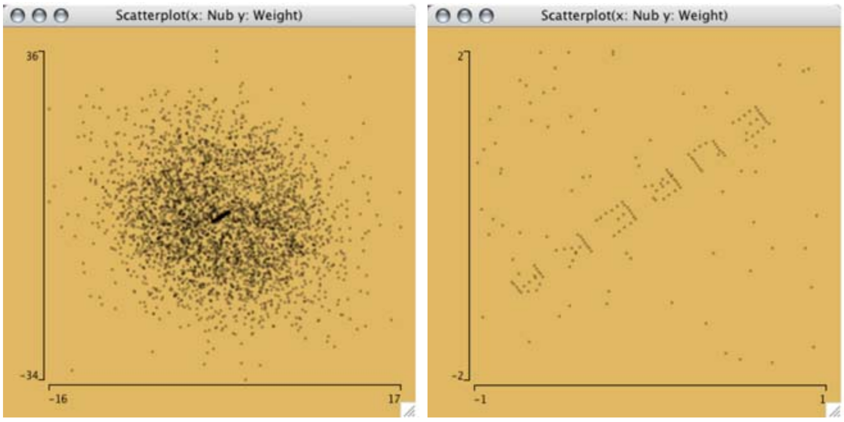


Figure 2: The pollen data plotted in Mondrian.

Figure 1 shows an example where the default plot symbol is unsuitable to find the interestring structure in a dataset. The dataset is the so called "pollen" data from the 1985 ASA data competition, and consists of a 5-dim. normal distribution with the word "EUREKA" added to the center of the data. The upper left plot shows the data plotted with the default setting of the R `plot` function. The '`o`' which is used as the default plot symbol is only suitable for small datasets with less than 100 points. The upper right plot shows the same data now plotted with '`.`' as plot symbol and reveals — by squeezing your eyes — the unusually high density in the center of the plot. The plots in the lower row show how we isolate the feature by zooming in. The corresponding R-code is:

```
> names(pollen)
[1] "Ridge"   "Nub"     "Crack"   "Weight"  "Density" "Number"
> attach(pollen)
> par(mfrow=c(2,2))
> plot(Nub, Weight, main="Default Plot")
> plot(Nub, Weight, pch=".", main="Smaller Symbols")
> plot(Nub, Weight, pch=".", xlim=c(-1,1), ylim=c(-1,1),  ...
> plot(Nub, Weight, xlim=c(-1,1), ylim=c(-0.8,1.6), ...
```

Figure 2 shows the same data plotted in Mondrian [6]. The default scatterplot in Mondrian uses $\alpha-$transparency to cope with overplotting. $\alpha-$transparency allows us to use suitably sized points in a scatterplot, without losing the information about density in the scatterplot. The amount of transparency gets bigger with the number of points to plot. In Figure 2 the unusual feature is immediately visible without the need to optimize plot parameters. More information on how to plot scatterplots can be found in Cook et al. [3].

2.2 The histogram — yet another optimal representation?

The histogram is probably number two in the list of most often used statistical graphs. There exist dozens of rules (cf. D. Scott [5]), which number of bins is the "best" under which circumstances. "Best" usually means, that the sum of the squared differenceS between the true density and the estimation via the histogram is minimized with some variance constraint.

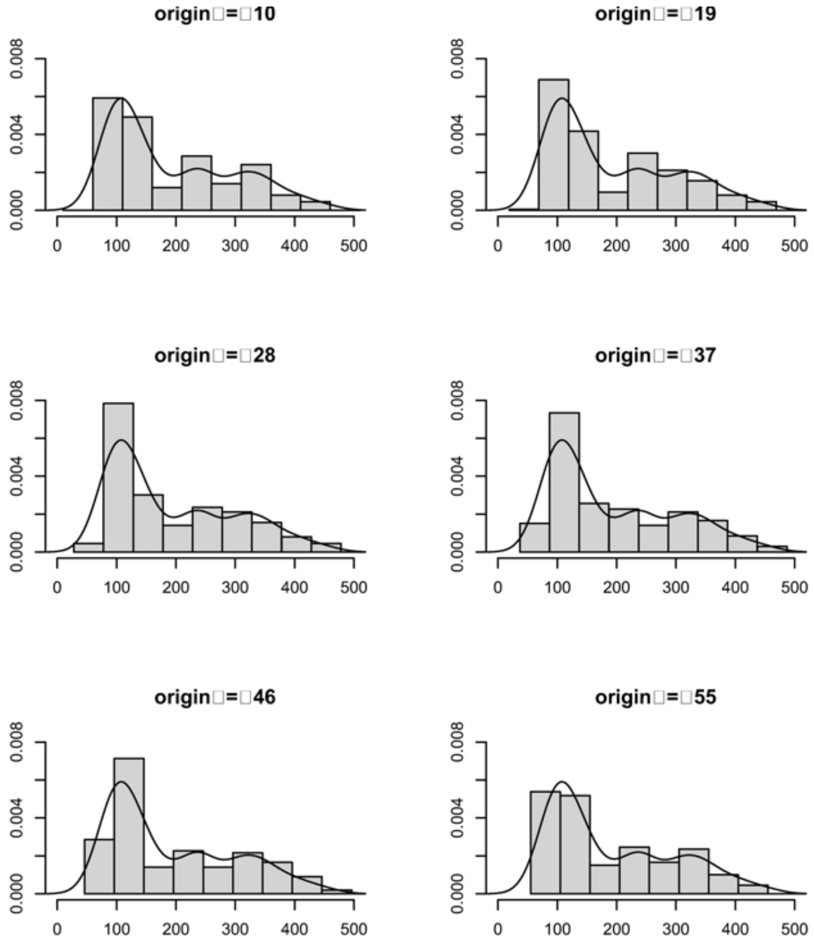


Figure 3: 6 histograms with superposed density estimatorS for the variable "displacement" of the "mpg-auto" dataset from the UCI ML repository. The number of bins has been determined according to "Sturges Rule".

In cases where the data comes from a single generating process following a continuous, only mildly skewed random variable, these rules will deliver sufficiently nice results[1]. The more critical situation arises, when the data is a mixture of several generating processes from both continuous and discrete random variables. In these situations, we have to cope with gaps, discrete patterns and accumulation points. Unfortunately real data usually comes from the latter kind of process.

Figure 3 shows an example of six histograms for the variable "displacement" of the "mpg-auto" dataset from the UCI Machine Learning Repository with origins at 10, 19, 28, 37, 46 and 55. The number of bins has been determined according to "Sturges Rule". The bin width has been "beautified" to 50 within the R `hist` function. Obviously non of the six origins gives us a satisfying estimation of the underlying density, nor does the kernel density estimator. The explanation is not too hard to find. Most cars in the dataset have only a very small displacement of 80 to 160. Bigger cars — all 6 cylinder engines in the dataset — form another mode at 220 to 260. Two discrete spikes can be found at 300 and 340, with some larger outliers, all corresponding to 8 cylinder engines.

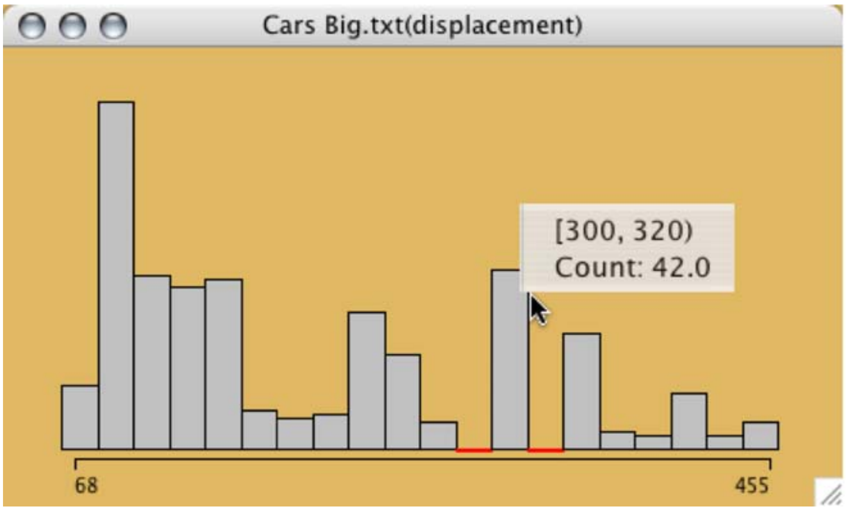


Figure 4: A histogram starting at 60 with bin wiDth 20, yielding 20 binS for the variable "displacement".

Figure 4 shows a histogram starting at 60 with bin width 20, yielding 20 bins for the variable "displacement", showing all of the above features. Finding a parameter setting revealing these features is easy in an interactive environment, but harder in a command line interface, where each new setting

[1] Although in these cases almost any origin and bin width will lead to almost optimal results.

must be retyped, until a satisfying setting is found. Finding explanations for the above described structural features can be done most conveniently within an interactive environment, which allows linked highlighting. This leads to the next section.

Plotting subgroups in histograms

It is common practice to color a subgroup in a histogram. Usually this should answer the question, whether this subgroup is any different from the whole population or not.

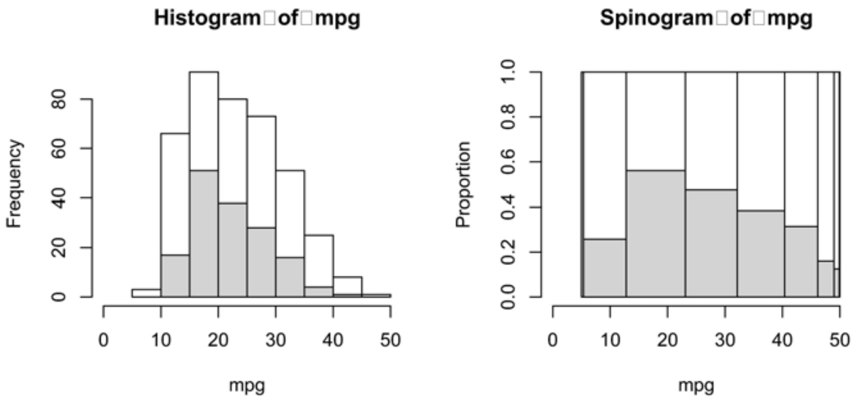


Figure 5: Left: A histogram for the variable “mpg” with model years 74–78 highlighted. Right: A Spinogram, showing the same data.

Figure 5 shows an example of this situation. The left histogram has all model years from 74 to 78 highlighted. At first glance we would expect that the selected subgroup has approximately the same distribution as the whole population. To verify this, we use a spinogram.

A spinogram is a histogram, where all bars have the same height. In order to keep the proportionality of the area of a bar and the number of cases in the bar, the width is adjusted, i.e. whereas in a histogram with equally spaced bins the height of a bar is proportional to the number of cases in this group, in a spinogram the width is proportional. Obviously the x-axis of a spinogram then is transformed to a no longer linear but still continuous scale. This puts more visual weight on areas with high density and less weight on areas with low density. The highlighting in a spinogram is still done from bottom to top. This allows the comparison of proportions of the highlighted cases across the whole range of the underlying variable. Whereas this comparison is easily possible, the comparison of proportions in highlighted histograms is almost impossible. This is due to the fact that our visual system is well able to compare positions along a common scale, but almost incapable of judging length or position in different scales (cf. Cleveland [1] 262pp). Coming back to the example in Figure 5 the spinogram reveals that the cars in the years 74 – 78 mostly have mpg-values close to the overall mean, i.e. the tails of the distribution of this group are less populated than in the rest of the sample.

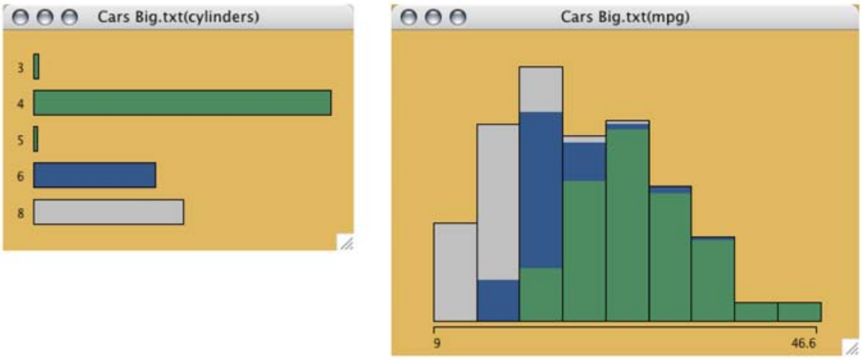


Figure 6: A histogram of the variable “mpg” colored according to the number of cylinders.

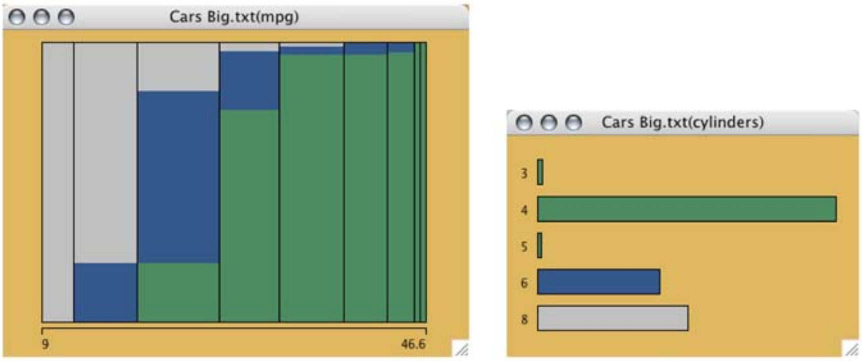


Figure 7: The same data as in Figure 6, now plotted as a spinogram.

Spinograms also allow you to look at the conditional distribution of more than one highlighted group. Figure 6 shows a histogram of “mpg” color brushed according to the number of cylinders of the engine (cars with 3-5 cylinders are joined in one group). Again, the histogram suffers from the differently scaled proportions and is hard to read. Figure 7 shows the corresponding spinogram, which makes the comparison across bars much easier. This kind of display is especially useful in classification problems, which need to assign more than two groups. With multiple groups, the stacking order of the groups in the spinogram becomes an important issue. A more comprehensive illustration of how to visualize conditional distributions can be found in Hofmann and Theus [4].

2.3 Mosaic plots — but which one?

Mosaic plots have been adopted more and more in the statistics community over the last 10 years. They form a very powerful framework to visualize multidimensional categorical data. Mosaic plots are especially good at visualizing associations between 2, 3, 4 or even 5 variables at a time. They are weaker for looking at only few variables, each having many categories.
Figure 8 shows a mosaic plot for the "mpg-auto" data for "Model year" and "Cylinder". Due to the strong variation in the variable "cylinders" over the different years, it is quite hard to read across the years while following a particular number of cylinders. The same problem arises when labeling the categories of the conditioned variable, i.e. "Cylinders". In Figure 8 an equidistant labelling was chosen, which does not fit any particular year, but should be a good estimate for all years. In this situation a fluctuation diagram, as shown in Figure 9, is much more appropriate to display the data. In a fluctuation diagram all cells get the same space assigned in a grid like layout.
The area which is filled by a tile within a cell is still proportional to the number of observations in this cell. Thus the only cell which is completely filled with a tile is the cell with the maximum cell count. The advantage of this kind of display is obvious. Using the grid like layout it it now easy to follow a particular category of a variable throughout the whole plot. Comparing Figures 8 and 9 we can see the structure in the data more clearly in the fluctuation diagram. The number of 4 cylinder cars is steadily growing over the 13 years, whereas the 8 cylinder cars seem to disappear in the early 80s. The number of 6 Cylinder cars is relatively stable over the years, whereas 3 and 5 cylinder cars are only found rarely.

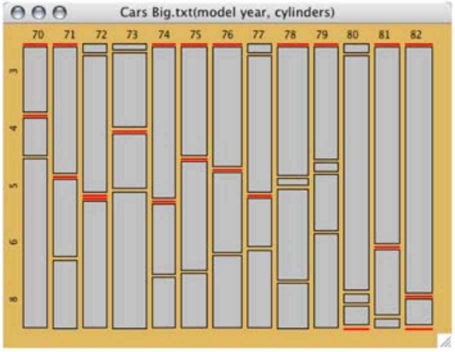


Figure 8: A mosaic plot for "Model year" and "Cylinder".

Besides fluctuation diagrams two other variations of the standard mosaic

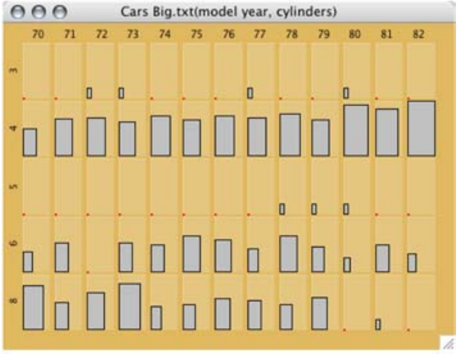


Figure 9: A fluctuation diagram for “Model year” and “Cylinder”.

plot have proven to be useful. In the *same bin size* display all tiles are of equal size, which is useful to detect empty cells in high dimensional datasets and the *multiple barchart* view which scales the size of the tiles along only one axis.

3 Plot ensembles

The last section gave some hints on how to choose the right plot parameters and/or plot types, in order to get meaningful plots. This helps to optimize a single plot or view.

In an exploratory data analysis process we often try to answer statistical questions with graphics. E.g. looking at the “mpg-auto” data we might be interested in the influence of the originating country or continent and the number of cylinders on the gas consumption of a car. This relationship between two categorical and one continuous variable can be investigated by using an ensemble of 4 linked plots.

The plot ensemble in Figure 10 features a barchart for cylinders and origin, a mosaic plot of the two variables and a boxplot of “mpg” conditioned on number of cylinders (alternatively we also could use a boxplot of “mpg” conditioned on the originating country). In this ensemble we see the interaction structure of the two influencing variables in the mosaic plot, as well as their marginal distribution in the two barcharts. The boxplot shows the distribution of “mpg” for each cylinder group and via highlighting we can investigate the interaction structure of the “origin” and “cylinders” on “mpg”. In Figure 10 the group of all Japanese cars has been highlighted.

The next example in Figure 11 shows how we can look at the temporal distribution of spam e-mails. In the barchart of the classification variable “spam” all spam e-mails have been selected. In the barchart for “Day of Week”, as well as the corresponding spineplot, we see the absolute and relative distribution of spam e-mails over the course of a week. Whereas the

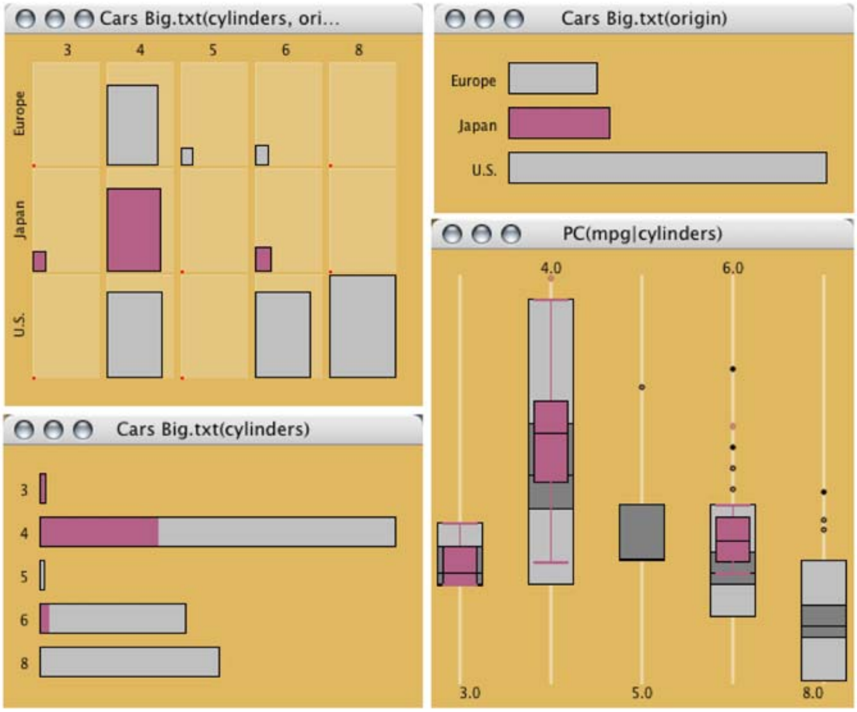


Figure 10: An ensemble of four plots to investigate the influence of country and cylinder on "mpg".

absolute amount of spam e-mails grows towards the middle of the week, the relative amount is highest at the weekends. In the histogram of "Time of Day" we see an almost constant amount of spam mails over the 24 hours of a day, whereas due to the small number of ordinary e-mails outside business hours, the proportion of spam is very high during the night.

The ensembles in Figure 10 and 11 are only two examples which show that a specific question in an exploratory data analysis can be answered with ensembles of (linked) plots. If statistical packages do not offer the whole suite of basic plots users can not plot data in the most suitable way. If for instance a package only offers point plots for quantitative data, these plots are used to try to visualize discrete data.

4 Conclusion

The rise of computers with graphical capabilities has lead to new graphical data analysis possibilities, but also caused an inflation in the use of statistical graphics. Only well designed graphics can be "worth a 1000 words".

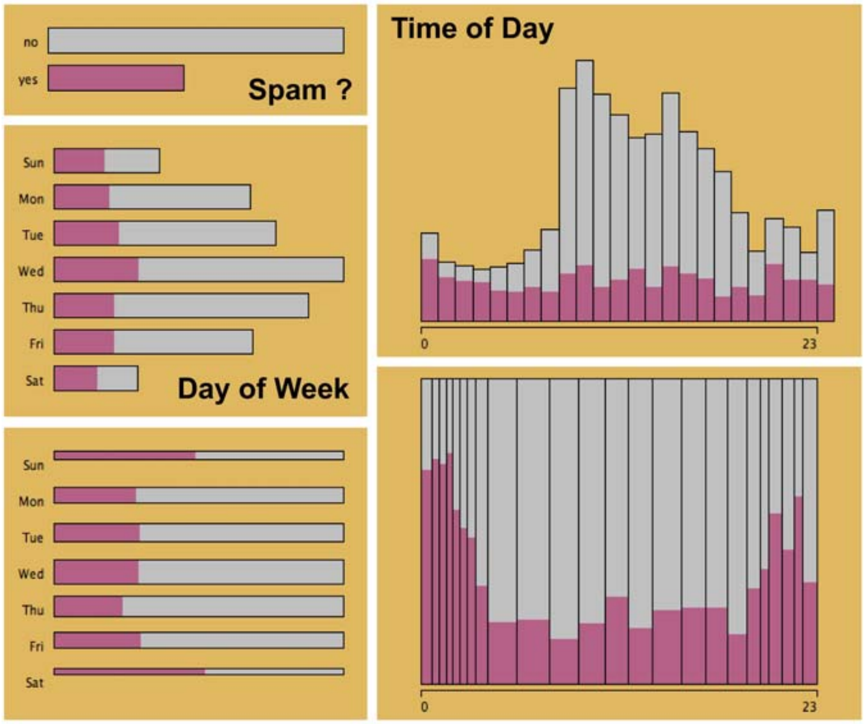


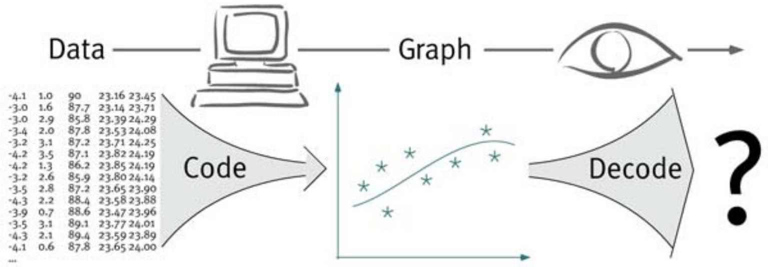
Figure 11: An ensemble of plots to investigate the temporal distribution of spam e-mails.

Many statistical software packages do not take care over default settings. This deficit can often be explained by the fact that the underlying code and graphical model is quite old, and was not adapted to modern data problems and rendering methods yet.

Using $\alpha-$channel transparency can help a lot when trying to avoid overplotting problems in scatterplots and parallel coordinate plots. The histogram as a means of density estimation is an example of a plot where “no default” is the only good default[2]. Spinograms are a good choice when trying to visualize a sub-population of a continuous variable. A histogram, which is often used instead, is not useful for this task. Mosaic plots are complemented by three variations to build a suite of plots, which can visualize multivariate discrete data. Where the one plot is good, the other one fails.

Generally, for a comprehensive graphical data exploration, we need a wide range of plots, which can be applied exactly for the purpose they serve best. No craftsman would enter a construction site with a toolbox consisting of just a single type of tool.

[2] In a recent talk an expert an Support Vector Machines (SVM) noted that he would suggest that all implementations of SVMs should always force the user to explicitly specify parameters, since there is nothing such as a default parameter setting which would generally yield acceptable results



-4.1	1.0	90	23.16	23.45
-3.0	1.6	87.7	23.14	23.71
-3.0	2.9	85.8	23.39	24.29
-3.4	2.0	87.8	23.53	24.08
-3.2	3.1	87.2	23.71	24.25
-4.2	3.5	87.1	23.82	24.19
-4.2	1.3	86.2	23.85	24.19
-3.2	2.6	85.9	23.80	24.14
-3.5	2.8	87.2	23.65	23.90
-4.3	2.2	88.4	23.58	23.88
-3.9	0.7	88.6	23.47	23.96
-3.5	3.1	89.1	23.77	24.01
-4.3	2.1	89.4	23.59	23.89
-4.1	0.6	87.8	23.65	24.00
...				

Figure 12: Statistical graphics code information on data in an abstract form. A successful decoding by a human is only possible, if the abstraction is suitable for the kind of data coded.

Figure 12 illustrates the process of coding information in a statistical graph. Given some data we code — and often condense — the information about this data via a computer based procedure into an abstract representation. The crucial part is the decoding process by the human observer. A successful decoding by a human is only possible, if the abstraction is suitable for the kind of data coded.

Additionally we must keep in mind that the human visual system has many limitations as basically described in Cleveland's [1] overview in the context of graph reading. His investigations have been limited to the state of statistical graphics in the early 80s. Today's rendering techniques offer new possibilities and challenges.

References

[1] Cleveland W.S. (1985). *The elements of graphing data.* Wadsworth, Monetrey, CA.

[2] Cleveland W.S. (1993). *Visualizing data.* Hobart, Summit, NJ.

[3] Cook D., Theus M., Hofmann H. *Scatterplots for massive datasets.* Journal of Computational and Graphical Statistics, submitted.

[4] Hofmann H., Theus M. *Visualizing conditional distributions.* Journal of Computational and Graphical Statistics, submitted.

[5] Scott D. (1992) *Multivariate density estimation – theory, practice, and visualization.* Wiley, New York.

[6] Theus M. (2002). *Interactive data visualization using mondrian.* Journal of Statistical Software **7** (11).

Address: M. Theus, Department of Computational Statistics and Data Analysis, Augsburg University, Universitätsstr. 14, 86135 Augsburg, Germany

E-mail: `martin.theus@math.uni-augsburg.de`

COMPSTAT'2004 Symposium

FITTING BRADLEY TERRY MODELS USING A MULTIPLICATIVE ALGORITHM

Ben Torsney

Key words: Bradley Terry model, discrete data, factorial structure, general equivalence theorem, maximum likelihood estimation, multiplicative algorithm, optimal design theory, paired comparisons.

COMPSTAT 2004 section: Design of experiments.

Abstract: We consider the problem of estimating the parameters of a Bradley Terry Model by the method of maximum likelihood, given data from a paired comparisons experiment. The parameters of a basic model can be taken to be weights which are positive and sum to one. Hence they correspond to design weights and optimality theorems and numerical techniques developed in the optimal design arena can be transported to this estimation problem. Furthermore extensions of the basic model to allow for a factorial structure in the treatments leads to an optimisation problem with respect to several sets of weights or distributions. We can extend techniques to this case. In section 1 we introduce the notion of paired comparisons experiments and the Bradley Terry Model. In section 2 the parameter estimation problem is outlined with optimality results and a general class of multiplicative algorithms outlined in sections 3 and 4 respectively. A specific algorithm is applied to the Bradley Terry log-likelihood in section 5 and treatments with a factorial structure are considered in section 6. Finally in section 7 extensions to triple comparisons and to extended rankings are briefly outlined.

1 Paired comparisions

1.1 Introduction

We consider paired comparison experiments in which J treatments or products are compared in pairs. In a simple form a subject is presented with two treatments and asked to indicate which he/she prefers or considers better. In reality the subject will be an expert tester; for example, a food taster in examples arising in food technology. The link with optimal design theory (apart from the fact that a specialised design, paired comparisons, is under consideration) is that, the parameters of one model, Bradley Terry model, for the resultant data are like weights. Hence the theory characterising and the methods developed for finding optimal design weights can be applied to characterising and finding the maximum likelihood estimators of these Bradley Terry weights.

1.2 The data

In a simple experiment a set of such testers is available and each is presented with one pair from a set of J treatments, say $T_1, T_2, \ldots, T_J$. The number of comparisons, n_{ij} of T_i to T_j, we assume has been predetermined. Sufficient summary data comprises the set $\{O_{ij} : i = 1, 2, \ldots, J;\ i = 1, 2, \ldots, J;\ i < j$ or $i > j\}$, where O_{ij} is the observed frequency with which T_i is preferred to T_j. Of course $O_{ij} + O_{ji} = n_{ij}$.

1.3 Models

1.3.1 A general model In the absence of other information the most general model here is to propose:

$$O_{ij} \sim B_i(n_{ij}, \theta_{ij}) \tag{1}$$

where

$$\theta_{ij} = P(T_i \texttt{ is prefered to } T_j)$$

Apart from the constraint $O_{ij} + O_{ji} = n_{ij}$, independence between frequencies is to be recommended. So apart from the constraint $\theta_{ij} + \theta ji = 1$, these define unrelated binomial parameters. The maximum likelihood estimator of θ_{ij} is O_{ij}/n_{ij} (the proportion of times T_i is preferred to T_j in these n_{ij} comparisons), and formal inferences can be based on the asymptotic properties of these.

1.3.2 Bradley Terry Model This is a more restricted model in that it imposes interrelations between the θ_{ij}. It proposes that:

$$\theta_{ij} = \frac{p_i}{p_i + p_j} \tag{2}$$

where $p_1, p_2, \ldots, p_J$ are positive parameters. See [1].
These can be viewed as indices or quality characteristics, one for each treatment. These are only unique up to a constant multiple, since θ_{ij} is invariant to proportional changes in p_i and p_j. A constraint needs to be imposed for uniqueness. One possibility is:

$$\sum p_i = 1$$

This implies $0 < p_i < 1$. We return to this later.

1.3.3 Motivation for Bradley Terry Model However we can show that θ_{ij} is uniquely determined by a latent difference. Let $p_i = \exp(\lambda_i)$. Then:

$$\theta_{ij} = \frac{\exp(\delta_{ij})}{1 + \exp(\delta_{ij})} \tag{3}$$

where $\delta_{ij} = \lambda_i - \lambda_j$.

Thus θ_{ij} is uniquely determined by the difference in the transformed quality characteristics λ_i, λ_j, while it is invariant to shifts in their values.
Further $\theta_{ij} = F(\delta_{ij})$, where $F(\delta)$ is the logistic distribution function. If we assume that the difference in quality, between the two treatments, has a logistic distribution, then θ_{ij} is the probability of a difference of at most δ_{ij}; or the difference in quality is given by:

$$\delta_{ij} = F^{-1}(\theta_{ij}) = F^{-1}\{p_i/(p_i + p_j)\}$$

See [6]. Other choices of $F(.)$ can lead to alternative models with parameters similar to $p_1, p_2, \ldots, p_J$.

2 Parameter estimation

In terms of the original parameters the likelihood of the data is a product of binomial likelihoods, namely:

$$L = \prod\prod_{r<s} (\theta_{rs})^{O_{rs}} (\theta_{sr})^{O_{sr}} \tag{4}$$

Let $p = (p_1, p_2, \ldots, p_J)$ and , for convenience, let $O_{ii} = 0$, $i = 1, 2, \ldots, J$, and $O_{i.} = \sum_j O_{ij}$.
Then the likelihood of the data under the Bradley Terry model is given by making the substitutions $\theta_{rs} = p_r/(p_r+p_s)$, $\theta_{sr} = p_s/(p_r+p_s)$, $,O_{rs}+O_{sr} = n_{rs}$, to yield:

$$L(p) = \left\{\prod_i (p_i)^{O_{i.}}\right\} \left\{\prod\prod_{r<s} (p_r + p_s)^{(-n_{rs})}\right\} \tag{5}$$

We wish to choose p $(p > 0)$ to maximise $L(p)$. Since θ_{ij} is invariant to proportional changes in the p_i's, so is $L(p)$. In fact $L(p)$ is a homogeneous function of degree zero in p; i.e. $L(cp) = L(p)$, where c is a scalar constant. It is constant on rays running out from the origin. It will therefore be maximised all along one specific ray. We can identify this ray by finding a particular optimising p^*. This we can do by imposing a constraint on p. Possible constraints are $\sum p_i = 1$ or $\prod p_i = 1$, or $g(p) = 1$ where $g(p)$ is a surface which cuts each ray exactly once. In the case $J = 2$ a suitable $g(p)$ is defined by $p_2 = h(p_1)$, where $h(.)$ is a decreasing function which cuts the two main axes, as in the case of $h(p_1) = 1 - p_1$, or has these as asymptotes, as in the case of $h(p_1) = 1/p_1$. In general a suitable choice of $g(p)$ is one which is positive and homogeneous of some degree h. Note that other alternatives are $\sum p_i = C$ or $\prod p_i = C$, where C is any positive constant; e.g. $C = J$.

The choice of $\prod p_i = 1$, being equivalent to $\sum \ln(p_i) = 0$, confers on $\alpha_i = \ln(p_i)$ the notion of a main effect. We will opt for the choice of $\sum p_i = 1$, which conveys the notion of p_i as a weight. We wish to maximise the likelihood or log-likelihood subject to this constraint and to non-negativity too. This is an example of the following general problem:

Problem (P):
Maximise $\phi(p)$ subject to $p_i \geq 0$, $\sum p_i = 1$.

We wish to maximise $\phi(p)$ with respect to a probability distribution.
Here we will take $\phi(p) = \ln\{L(p)\}$.

There are many examples of this problem arising in various areas of statistics, especially in the area of optimal regression design. We can exploit optimality results and algorithms developed in this area. The feasible region is an open but bounded set. Thus there should always be a solution to this problem allowing for the possibility of an unbounded maximum, multiple solutions and solutions at vertices (*i.e.* $p_t = 1$, $p_i = 0$, $i \neq t$).

3 Optimality conditions

We can define optimality conditions in terms of the point to point directional derivative defined by Whittle [19]. The directional derivative of $F_\phi(p,q)$ of a criterion $\phi(.)$ at p in the direction of q is the limit as $\epsilon \downarrow 0$ of:

$$[\phi\{(1-\epsilon)p + \epsilon q\} - \phi(p)]/\epsilon$$

i.e.

$$F_\phi(p,q) = dg/d\epsilon \mid \epsilon = 0^+$$

where $g(\epsilon) = \phi\{(1-\epsilon)p + \epsilon q\}$.

This derivative exists even if $\phi(.)$ is not differentiable; but if $\phi(.)$ is differentiable then:

$$F_\phi(p,q) = (q-p)^T d$$

where $d = \partial\phi/\partial p$.

Let $F_i = F_\phi(p, e_j)$, where e_j is the j^{th} unit vector in $\Re^J$. Then:

$$F_j = d_j - p^T d = d_j - \sum p_i d_i, \text{ where } d_j = \partial\phi/\partial p_j.$$

We call F_j the j^{th} vertex directional derivative of $\phi(.)$ at p. Note that $\sum p_j F_j = 0$, so that, in general some F_j are negative and some are positive.

If $\phi(.)$ is differentiable at p^*, then a necessary condition for $\phi(p^*)$ to be a local maximum of $\phi(.)$ in the feasible region of Problem (P) is:

$$F_j^* = F_\phi\{p^*, e_j\} = 0 \texttt{ for } p_j^* > 0,$$

$$F_j^* = F_\phi\{p^*, e_j\} \leq 0 \texttt{ for } p_j^* = 0,$$

If $\phi(.)$ is concave on its feasible region, then these first order stationarity conditions are both necessary and sufficient. This is the general equivalence theorem in optimal design. See [19], [5].

It is clear that all p_j^* must be positive in the case of the Bradley Terry likelihood, so that the second condition is redundant.

4 Algorithms

4.1 A multiplicative algorithm

Problem (P) has a distinct set of constraints, namely the variables $p_1, p_2, \ldots,$ p_J must be nonnegative and sum to 1. An iteration which neatly submits to these and has some suitable properties is the multiplicative algorithm:

$$p_j^{(r+1)} = \frac{p_j^{(r)} f(d_j^{r)})}{\sum p_i^{(r)} f(d_i^{(r)})} \tag{6}$$

where $d_j^{(r)} = \partial\phi/\partial p_j \mid p = p^{(r)}$ while $f(d)$ is positive and strictly increasing in d and may depend on one or more free parameters.

This type of iteration was first proposed by [13], taking $f(d) = d^\delta$, with $\delta > 0$. This, of course, requires positive derivatives. Subsequent empirical studies include Silvey et al [11], which is a study of the choice of δ when $f(d) = d^\delta$, $\delta > 0$; Torsney [15], which mainly considers $f(d) = e^{\delta d}$ in a variety of applications, for which one criterion $\phi(.)$ could have negative derivatives; Torsney and Alahmadi [16] who consider other choices of $f(.)$; Torsney and Mandal [18] who consider objective choices of $f(.)$; and [8] who explore developments of the algorithm based on a clustering approach in the context of a continuous design space. Torsney and Mandal [17] and Mandal et al [9] also apply these algorithms to the construction of constrained optimal designs.

Titterington [12] describes a proof of monotonicity of $f(d) = d^\delta$ in the case of D-optimality. Torsney [14] explores monotonicity of particular values for δ for particular $\phi(p)$. Torsney [14] also establishes a sufficient condition for monotonicity of $f(d) = d^\delta$, $\delta = 1/(t+1)$, when the criterion $\phi(p)$ is homogenous of degree $-t, t > 0$ with positive derivatives and proves this condition to be true in the case of linear design criteria such as the c-optimal and the A-optimal criteria, for which $t = 1$, so that $\delta = 1/2$. In other cases the value $\delta = 1$ can be shown to yield an EM algorithm, which is known to be monotonic and convergent; see [13]. Beyond this there are minimal results on convergence, although this will depend on the choice of $f(.)$ and of parameters like δ. See [11] for some empirical results. In principal the choice of $f(.)$ is arbitrary but objective bases for choices are addressed in the formal properties now listed.

4.2 Properties of the algorithm

Under the conditions imposed on $f(.)$, the above iterations possess the following properties which are considered in more detail in [15], [16] and [7]:

1. $p^{(r)}$ is always feasible.
2. $F_\phi\{p^{(r)}, p^{(r+1)}\} \geq 0$, with equality when the d_j's corresponding to nonzero p_j's have a common value d $(= \sum p_i d_i)$, in which case $p^{(r)} = p^{(r+1)}$.

3. An iterate $p^{(r)}$ is a fixed point of the iteration if the derivatives $d_j^{(r)}$ corresponding to nonzero $p_j^{(r)}$ are all equal; equivalently if the corresponding vertex directional derivatives $F_j^{(r)}$ are zero. Thus a solution to Problem (P) is a fixed point of the iteration. So also are solutions subject to setting a given subset of weights to zero; see [15].
4. We mentioned that $f(.)$ may depend on one or more free parameters. Torsney and Alahmadi [16] explore methods for choosing a single positive parameter δ for various given choices of $f(.)$. Torsney and Mandal [18] explore methods for choosing $f(.)$, which can accommodate negative partial derivatives or for which (positive) partial derivatives can be replaced by vertex directional derivatives. A further paper is in preparation on choosing $f(.)$ when the criteria has positive derivatives.

5 Fitting Bradley Terry Models

Our criteria is:

$$\phi(p) = \ln\{L(p)\} = \sum_i O_{i.} \ln(p_i) - \sum\sum_{r<s} n_{rs} \ln(p_r + p_s) \tag{7}$$

Since $L(p)$ is a homogeneous function of degree zero $\sum p_i d_i = 0$. In fact $d_j = F_j$. So there are always positive and negative d_j unless all are zero. We require a function $f(d)$ which is defined for positive and negative d, where we take d to represent a partial derivative. Noting that all p_j^* must be positive a suitable choice of $f(.)$ should be governed by the fact that at the optimum $d_j^* = 0,\ j = 1, 2, \ldots, J$.

This suggests that suitable function is one that is 'centred' on zero and changes reasonably quickly about $d = 0$. It should also be desirable to treat positive and negative derivatives symmetrically. Torsney and Mandal [18] start from a function $h(x)$ defined on $\Re$, such that $h(x) > 0,\ h'(x) > 0,\ h(0) = 1$. They propose:

$$f(x) = \begin{cases} h(x) & : \ x < 0 \\ 2 - h(-x) & : \ x > 0 \end{cases}$$

i.e.

$$f(x) = (1 + s) - sh(-sx), \quad s = sign(x)$$

Clearly $f(x)$ is increasing, while for $y > 0$, $(y,\ f(y))$ and $(-y,\ f(-y))$ are reflections of each other in the point $(0, 1) = (0, f(0))$; i.e. $f(-y) = [2-f(y)]$. Equivalently $f'(y)$ is symmetric about zero. Note that $0 < f(x) < 2$, so that $f(x)$ is bounded; also $f(0) = 1$.

Torsney and Mandal [18] consider various choices of $h(x)$, including $h(x) = 2H(\delta x)$, where δ is a positive parameter and $H(.)$ is a cumulative distribution function such that $H(0) = 1/2$. Here we opt for $H(.) = \Phi(.)$, so that iterations prove to be:

$$p_j^{(r+1)} = \frac{p_j^{(r)}\Phi(\delta d_j^{r)})}{\sum p_i^{(r)}\Phi(\delta d_i^{(r)})}$$

Example: We use this algorithm in two examples.

Example 1:
In this case $J = 8$ coffee types were compared through 26 pairwise comparisons on each pair, yielding a total of N = 728 observations; i.e. $\sum\sum O_{ij} = 728$. A suitable δ is $\delta = 1/N$. In effect we are standardising through replacing observed by relative frequencies in the log-likelihood, and then taking $\delta = 1$. Starting from $p_j^{(0)} = 1/J$, the numbers of iterations needed to achieve $\max|d_j| = \max|F_j| \leq 10^{-n}$, $n = 0, 1, \ldots, 7$ respectively are 17, 21, 25, 32, 38, 45, 51, 59. The optimal p^* is: (0.190257, 0.122731, 0.155456, 0.106993, 0.091339, 0.149406, 0.080953, 0.102865). Iterations were monotonic.

Example 2:
In this example $J = 9$ quality of life dimensions were compared in pairs by each of 50 patients with early signs of rheumatoid arthritis (RA). The 9 dimensions were: ability to physically function, pain, stiffness, ability to work, fatigue, depression, interference with social activities, side effects, and financial burden. This data arose from the Consortium of Practicing Rheumatologists long-term observational multi-center study of early severe RA. Patients entered in this additive cohort had less than 1 year of symptom onset. The responses were obtained at their first telephone interview. Formed in 1992, the Consortium prospectively followed them to delineate early outcome and factors, such as treatment, functional, radiographic, psychosocial, and economic outcomes. Data on disease severity, functional status, psychosocial health, cost, radiographic damage, laboratory serologies and acute phase reactants were recorded at Baseline and at 6 months, 1 year, and annually thereafter. As a chronic illness, RA impacts every dimension of quality of life. Even among RA patients, however, differences in life situations, clinical presentation, and disease course can be striking, leading to varying patient rankings of the importance of difference disease and life factors. The 9 factors were selected to represent aspects of RA that patients could easily identify and compare.

There were a total of $N = 1800$ comparisons; i.e. $\sum\sum O_{ij} = 1800$. In 8 cases there were ties. These were split 50:50 between the relevant treatments. Again a suitable δ is $\delta = 1/N$. Starting from $p_j^{(0)} = 1/J$, the numbers of iterations needed to achieve $\max|d_j| = \max|F_j| \leq 10^{-n}$, $n = 0, 1, \ldots, 6$ respectively are 28, 42, 56, 69, 84, 96, 110. The optimal p^* is: (0.265361, 0.172154, 0.151644, 0.059151, 0.123506, 0.030753, 0.037740, 0.055038, 0.104653), the order of the components corresponding to the order of the dimensions as listed above. Iterations were monotonic.

There is a further issue here. These 1800 responses have been obtained from only 50 patients. Each patient has responded on each pairwise comparison. We have assumed independence between the resulting 36 observations. Dittrich et al [3] also contemplate this 'independent decisions' model, an independence which allows for inconsistent responses by a patient. However they extend it to a 'dependent decisions' model. For an individual patient's comparison of T_i and T_j let $Y_{ij} = 1$ if he/she records that T_i is preferred to T_j and $Y_{ij} = 0$ otherwise. In the case of three dimensions their model is:

$$P(y_{12}, y_{13}, y_{23}) =$$

$$C\{[(\sqrt{(p_1/p_2)})^{y_{12}}(\sqrt{(p_1/p_3)})^{y_{13}}(\sqrt{(p_2/p_3)})^{y_{23}}][(\omega_1^{y_{12}y_{13}})(\omega_2^{y_{12}y_{23}})(\omega_3^{y_{13}y_{23}})]\}$$

where C is a normalising constant.

All parameters must be positive. A constraint is still needed on the p_j as above, but none are needed on the ω_j. However we could transform to $q_j = \omega_j/(\omega_1 + \omega_2 + \omega_3)$, so that $(q1 + q2 + q3) = 1$, while $\alpha = (\omega_1 + \omega_2 + \omega_3)$ is a free positive parameter, which could be treated like the variable q arising in models allowing for ties discussed in section 7 below. The above class of algorithm could then be used to find the optimal values of both the p_i's and q_j's. This would need the individual responses on each pair of dimensions from each respondent.

If $\omega_1 = \omega_2 = \omega_3 = 1$, we recover the independence model.

Furthermore extensions of Bradley Terry models are available when respondents record consistent rankings; see below. However there is scope for extending this work.

6 Treatments with a factorial structure

In Example 1 the 8 coffees comprised the 8 combinations arising from 2 brew strengths, 2 roast colours, and 2 brands. Simpler versions of the Bradley Terry Model have been proposed in terms of definitions of main effects and possibly low order interactions. We consider main effects only for the moment in the case of 3 factors.

Suppose that we have $J = KLM$ treatments arising from the KLM factor level combinations of 3 factors, denoted by α, β, γ with K, L and M levels respectively. We have treatments T_{klm}, $k = 1, \ldots, K$; $l = 1, \ldots, L$; $m = 1, \ldots, M$, with associated Bradley Terry parameters p_{klm} , such that T_{klm} is preferred to T_{qrs} with probability $\{p_{klm}/(p_{klm} + p_{qrs})\}$. This is allowing for main effects and interactions of all orders.

A main effects or additive model corresponds to:

$$p_{klm} = \alpha_k \beta_l \gamma_m$$

i.e.

$$\ln(p_{klm}) = \ln(\alpha_k) + \ln(\beta_l) + \ln(\gamma_m)$$

where $\alpha_k,\ \beta_l,\ \gamma_m\ >\ 0$.

The likelihood is again a homogeneous function of degree zero in each of the three sets of main effect parameters. Constraints need to be imposed on each of them. Various choices can be considered as above with appropriate extensions of the above algorithm. If we opt for the constraints

$$\sum \alpha_k = \sum \beta_l = \sum \gamma_m = 1$$

we wish to maximise the log-likelihood with respect to several distributions. At the optimum all partial derivative should be zero. (Note that alternatives could be $\sum \alpha_k = K,\ \sum \beta_l = L,\ \sum \gamma_m = M$).

A suitable set of iterations are:

$$\alpha_k^{(r+1)} = \frac{\alpha_k^{(r)} f_\alpha(d_k^\alpha)}{\sum \alpha_t^{(r)} f_\alpha(d_t^{(\alpha)})}$$

$$\beta_l^{(r+1)} = \frac{\beta_l^{(r)} f_\beta(d_l^\beta)}{\sum \beta_t^{(r)} f_\beta(d_t^{(\beta)})}$$

$$\gamma_m^{(r+1)} = \frac{\gamma_m^{(r)} f_\gamma(d_m^\gamma)}{\sum \gamma_t^{(r)} f_\gamma(d_t^{(\gamma)})}$$

where $f_\alpha(.),\ f_\beta(.),\ f_\gamma(.)$ are positive increasing functions and $d_k^{(\alpha)} = \partial\phi/\partial\alpha_k$ at $\alpha = \alpha^{(r)}$ etc.

This set of iterations enjoys the same properties as those for a single distribution, including $F_\phi(\lambda^{(r)}\ \lambda^{(r+1)}) \geq 0$, where $\lambda = (\alpha^T, \beta^T, \gamma^T)^T$. See [18]. In our example $K = L = M = 2$. Taking $\delta = 1/N, f_\theta(d^{(\theta)}) = \Phi(\delta d^{(\theta)})$ and $\theta_j^{(0)} = 1/2, j = 1, 2,$ for $\theta = \alpha, \beta, \gamma,$ representing brew strength, roast colour and coffee brand respectively, the numbers of iterations needed to achieve $\max|d_j^{(\theta)}| = \max|F_j^{(\theta)}| \leq 10^{-n}$, for $n = 0, 1, \ldots, 7$ respectively are 7, 12,15,19, 23,27, 31, 36. Optimal values are: $\alpha^* = (0.574904, 0.425096)$, $\beta^* = (0.551050, 0.448950)$, $\gamma^* = (0.504887, 0.495113)$;
and the optimal p^*: is (0.159949, 0.156852, 0.130313, 0.127790, 0.118269, 0.115980, 0.096356, 0.094491).
Iterations were monotonic.

Notes:

1. Other variations of the above iterations are possible. One is to cycle through the three sets of main effect parameters running the iterations for each one in turn, while keeping the others fixed.
2. Obviously the approach is extendable to any number of factors.

3. There are extensions of the Bradley Terry model which allow for interactions and the above iterations can be extended to these too. For example a model including an interaction between brew strength and roast colour corresponds to:

   $$p_{klm} = \alpha_k \beta_l \gamma_m (\alpha\beta)_{kl}$$

   i.e.

   $$\ln(p_{klm}) = \ln(\alpha_k) + \ln(\beta_l) + \ln(\gamma_m) + \ln((\alpha\beta)_{kl})$$

   where $(\alpha\beta)_{kl} > 0$.
   The likelihood is now additionally homogenous in two sets of respects; namely, it is invariant to proportional changes in the terms $(\alpha\beta)_{kl}$ when the constant of proportionality either varies with α or with β. Several sets of consistent constraints are needed. One possibility is the set

   $$\sum(\alpha\beta)_{kl} = K,\ l = 1, 2, \ldots, L;\ \sum(\alpha\beta)_{kl} = L,\ k = 1, 2, \ldots, K$$

   or sums could be replaced by products.

   Further development of our class of iterations is needed. For each α and each β the $(\alpha\beta)_{kl}$ in effect define a set of probability distributions except that the probabilities are scaled to add to a constant differing from 1. One option would be to alternate between iterations (appropriately modified to satisfy these re-scaling constraints) for each set. An alternative derives from Linear Programming Theory. The non-negativity and equality constraints imply that the set $\{(\alpha\beta)_{kl} : k = 1, 2, \ldots, K;\ l = 1, 2, \ldots, L\}$ belongs to a bounded convex polyhedron whose vertices are Basic Feasible Solutions. The convex weights defining the $(\alpha\beta)_{kl}$ defines one distribution. An extra set of equations for updating these can be added to the sets for main effects.

7 Extensions of the Bradley Terry Model

There are extensions of the basic Bradley Terry Model which can be fitted using the above methods. These include:
(a) Models allowing a 'no-preference' option. Two possibilities are:
(i)

$P(T_i \texttt{ is preferred to } T_j) = p_i/(p_i + p_j + p_0)$

$P(T_j \texttt{ is preferred to } T_i) = p_j/(p_i + p_j + p_0)$

$P(\texttt{No preference}) = p_0/(p_i + p_j + p_0)$

One extra parameter has been introduced p_0 which must be positive and these probabilities and hence the likelihood are homogenous of degree zero in $p_0, p_1, \ldots, p_J$. Finding maximum likelihood estimates of these defines another example of Problem (P).

(ii) Rao and Kupper [10] proposed :

$P(T_i$ `is preferred to` $T_j) = p_i/(p_i + qp_j)$

$P(T_j$ `is preferred to` $T_i) = p_j/(p_j + qp_i)$

where $q > 1$.
This model has a latent logistic distribution motivation since $P(T_i$ `is preferred to` $T_j) = F(\lambda_i - \lambda_j - \tau),\ \tau \geq 0$, where $F(.)$ is the logistic distribution function and $p_i = \exp(\lambda_i),\ q = \exp(\tau)$.

(iii) Davidson [2] proposed:

$P(T_i$ `is preferred to` $T_j) = p_i/(p_i + p_j + q(p_i p_j)^{1/2})$

$P(T_j$ `is preferred to` $T_i) = p_j/(p_i + p_j + q(p_i p_j)^{1/2})$

$P($`No preference`$) = q(p_i p_j)^{1/2}/(p_i + p_j + q(p_i p_j)^{1/2})$.

where $q > 1$.

Each of (ii) and (iii) lead to likelihoods which are homogenous of degree zero in the p_i's. Also note that $\{A/(A + qB)\} = \{r_1 A/(r_1 A + r_2 B)\}$, where $r_1 = q^{-1/2}$ and $r_2 = 1/r_1$. This is homogenous of degree zero in r_1 and r_2. Hence we could impose the constraint $r_1 + r_2 = 1$. However $r_1 \geq r_2$. A further transformation is $s_1 = r_1 - r_2,\ s_2 = 2r_2$. Now constraints are $s_1, s_2 > 0,\ s_1 + s_2 = 1$. We can now maximise the likelihood with respect to two distributions using our family of algorithms. To determine q we need to re-scale to $r_1 r_2 = 1$.

Henery [4] replaces q in (ii) by $q_{ij} = q_i * q_j$ with $q_i * q_j \geq 1$. The latter condition implies that at most one q_i can be less than 1, (the minimum in fact). If none satisfy this condition then the above transformation could be applied to each q_i, leading to an optimisation of the likelihood with respect to $(J+1)$ distributions. If the minimum was known to be less than 1, and it's subscript i were known too, then an appropriate variation of the approach takes $r_{1i} = (1/q_i)^{1/2}$, where r_{1i} is the value of r_1 for this particular q_i.

Kuk [6] considers applications to the outcome of football matches and extends the model to include two sets of the parameters $\{p_i\}$ and two sets of the parameters $\{q_j\}$, one each for 'home' and 'away' games. The likelihood is homogenous of degree zero in the two sets of p_i's as a whole and in three sets of variables which are based on transformations of the q_j's similar to that defining r_1 and r_2 above. Thus we wish to maximise the likelihood with respect to four distributions.

(b) Triple Comparisons.
An extension of pairwise comparisons is to invite subjects to place three treatments in order of preference . Let:

$$\theta_{ijk} = P(T_i \texttt{ is preferred to } T_j \texttt{ and } T_j \texttt{ is preferred to } T_k)$$

Various possible extensions of the Bradley Terry Model include:

$\theta ijk = p_i p_j / \{(p_i + p_j + p_k)(p_j + p_k)\}$;

$\theta_{ijk} = (p_i)^2 p_j / D,$

where $D = (p_i)^2 p_j + (p_j)^2 p_i + (p_i)^2 p_k + (p_k)^2 p_i + (p_j)^2 p_k + (p_k)^2 p_j$

(c) Extended Rankings.
The latter model extends to rankings of more than three treatments, while both models define likelihoods which are homogenous of degree zero in $p_1, p_2, \ldots, p_J$, each of which must be positive. Maximum likelihood estimation of these is equivalent to another example of Problem (P). Equally if the treatments have a factorial structure, the likelihood can be expressed as a function of several distributions and optimised with respect to these using the algorithms described.

8 Discussion

The primary focus of this paper is one of cross fertilisation, an arguably somewhat limited even simple minded one. It is to point out that a class of maximum likelihood estimation problems could be attacked using tools for solving optimal design problems because in each case one or several sets of optimising weights or distributions are sought. Hence the equivalence theorems characterising optimality in the optimal design arena and related algorithms can be transported over to the parameter estimation arena. This is one new contribution of this work. One other is using a new version of the above mentioned algorithms, one which can accomodate negative derivatives.

References

[1] Bradley R.A., Terry M.E. (1952). *The rank analysis of incomplete block designs I, The method of paired comparisons.* Biometrika **39**, 324 – 345.

[2] Davidson R.R. (1970). *On extending the Bradley Terry model to accommodate ties in paired comparisons experiments.* J.Am. Statist. Ass. **65**, 317 – 328.

[3] Hittich R., Hatzinger R., Katzenbeisser W. (2002). *Modelling dependencies in paired comparisions data- a log-linear approach.* Computational Statistics & Data Analysis **40**, 39 – 57.

[4] Henery R.J. (1992). *An extension of the Thurstone-Mosteller model for chess.* Statistician **41**, 559 – 567.

[5] Kiefer J. (1974). *General equivalence theory for optimum designs (approximate theory).* Annals of Statistics **2**, 849 – 879.

[6] Kuk A.C.Y. (1995). *Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players.* Statistician **44**, 523 – 528.

[7] Mandal S., Torsney B. (2000). *Algorithms for the construction of optimising distributions.* Communications in Statistics (Theory and Methods) **29**, 1219 – 1231.

[8] Mandal S., Torsney B. (2004). *Construction of optimal designs using a clustering approach.* (Under revision for J. Stat. Planning & Inf.)

[9] Mandal S., Torsney B., Carriere K.C. (2004). *Constructing optimal designs with constraints.* Jounal of Statistical Planning and Inference (to appear).

[10] Rao P.V., Kupper L.L. (1967). *Ties in paired comparison experiments: a generalisation of the Bradley Terry model.* J. Am. Statist. Ass. **62**, 192 – 204.

[11] Silvey S.D., Titterington D.M., Torsney B. (1978). *An algorithm for optimal designs on a finite design space.* Communications in Statistics A **14**, 1379 – 1389.

[12] Titterington D.M. (1976). *Algorithms for computing D-optimal designs on a finite design space.* Proc. 1976 Conf. On Information Sciences and Systems, Dept. of Elect. Eng., Johns Hopkins Univ. Baltimore, MD, 213 – 216.

[13] Torsney B. (1977). *Contribution to discussion of 'Maximum Likelihood Estimation via the EM Algorithm' by Dempster, Laird and Rubin.* J. Royal Stat. Soc. (B) **39**, 26 – 27.

[14] Torsney B. (1983). *A moment inequality and monotonicity of an algorithm.* Lecture Notes in Economics and Mathematical Systems, A.V. Fiacco, K.O. Kortanek (Eds.), Springer Verlag **215**, 249 – 260.

[15] Torsney B. (1988). *Computing optimizing distributions with applications in design, estimation and image processing.* In: Optimal Design and Analysis of Experiments, Y. Dodge, V.V. Fedorov, H.P. Wynn (Eds.), North Holland., 361 – 370.

[16] Torsney B., Alahmadi A.M. (1992). *Further developments of algorithms for constructing optimizing distributions.* In Model Oriented data Analysis, V. Fedorov, W.G. Muller, I.N. Vuchkov (Eds), Proceedings of 2nd IIASA-Workshop, St. Kyrik, Bulgaria, 1990, Physica Verlag, 121 – 129

[17] Torsney. B., Mandal S. (2000). *Construction of constrained optimal designs.* In: Optimum Design 2000, A. Atkinson, B. Bogacka, A. Zhiglavsky (Eds), Proceedings of Design, held in honour of 60th Birthday of Valeri Fedorov, Cardiff, Kluwer, 141 – 152.

[18] Torsney B., Mandal S. (2004). *Multiplicative algorithms for constructing optimizing distributions mODa 7.* Advances in Model Oriented Design and Analysis, 143 – 150.

[19] Whittle P. (1973). *Some general points in the theory of optimal experimental design.* J. Roy. Statist, Soc. B **35**, 123 – 130.

Address: B. Torsney, Department of Statistics, University of Glasgow, Glasgow G12 8QW, U.K.

E-mail: `B.Torsney@stats.gla.ac.uk`

COMPSTAT'2004 Symposium

MODELLING MULTIPLE TIME SERIES: ACHIEVING THE AIMS

Granville Tunnicliffe-Wilson and Alex Morton

Key words: Cross-spectral analysis, extended autoregression, prediction, transfer functions.

COMPSTAT 2004 section: Time series analysis.

Abstract: We review the traditional aims and methodology of multiple time series modelling, and present some recent developments in the models available to achieve these aims, in the context of both regularly and irregularly sampled data. These models are analogues of the vector autoregressive process, based on the generalised shift, or Laguerre, operator. They form a subclass of vector autoregressive moving-average processes; they retain many of the attractive features of the standard vector AR model, but have an added dimension of flexibility, that leads to improvements in predictive ability.

1 Reviewing the objectives and methodology

The aims of time series analysis are revealed in the titles of some of the early books on the subject. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, by Wiener [19], has a comprehensive title, but *Prediction and Regulation* by Whittle [18], and *Time Series analysis, Forecasting and Control* by Box and Jenkins [5] make more explicit the application to control, which was undoubtedly one of Wiener's objectives. The objectives are also clearly stated in the title of *Statistical analysis and Control of Dynamic Systems* by Akaike and Nakagawa [2], the original publication of which, in Japanese, took place in 1972.

The time series model may itself be the immediate objective of the modelling, as in predator-prey systems and a host of other scientific applications, where an understanding of the mechanisms of interaction between time series variables is required. However, prediction is an encompassing objective. The model is generally identified by its predictive capacity, whatever the aim of the application. Smoothing, or more generally signal extraction, depends on the structure identified by the predictive model. Control applications rely on the ability to predict an output series from an input series.

Methodology developed in the early years is still widely used. Spectral analysis has tended to give way to time domain methods, particularly in econometric forecasting. There is, however, one context, that of modelling causal (or one-sided) dependency, in which cross-spectral analysis, is currently an under-used tool. It is generally very efficient, both statistically and in terms of the time and effort required to obtain useful results. Early books which presented this methodology, such as Jenkins and Watts [10], are now, fortunately, supplemented by some recent, well received texts. These cover the use of spectral analysis for identifying the transfer function coefficients ν_k, by which a dependent series y_t is related to lagged values of the explanatory series x_t:

$$y_t = \nu_0 x_t + \nu_1 x_{t-1} + \nu_2 x_{t-2} + \cdots + n_t. \tag{1}$$

Although cross-spectral analysis is based on frequency domain regression, its results can be expressed as estimates, over an appropriate lag window, of the transfer function coefficients. We illustrate this with a simple example, partly to encourage the re-introduction of such methods, but also to demonstrate, in part, why the subject moved away from them.

Figure 1(a) shows temperatures measured every minute by sensors in the cab and trailer of a transport vehicle. It is clear that the cab temperature lags the trailer temperature. Figure 1(b) shows the transfer function coefficients in this relationship, as estimated by cross spectral analysis. The estimates were produced almost automatically, with little user intervention. Limits on the plot show that significant values are spread over lags 0 to 4, with a peak at lag 2. This represents a one-sided, or causal, relationship, that may be used to predict the cab temperature from the trailer temperature as shown in Figure 2(a).

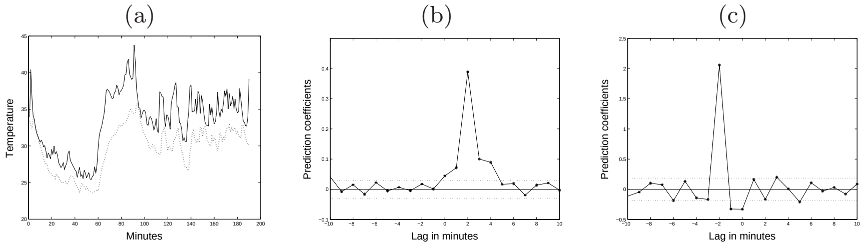


Figure 1: (a) Graphs of temperatures inside a transport vehicle trailer (solid line) and in the cab (dotted line), (b) lagged prediction coefficients obtained by cross-spectral analysis, for predicting cab temperature from trailer temperature, and (c) for predicting trailer temperature from cab temperature.

However, the desired aim was to predict trailer temperatures from the sensor in the cab. Figure 1(c) shows the estimated transfer function coefficients when the roles of the series are reversed. The significant values are spread over lags 0 to -2. The relationship is no longer causal and these coefficients cannot be used for prediction. But reasonable linear predictions of the trailer temperature from the cab temperature can still be constructed, as shown in Figure 2(b).

In general, cross-spectral estimation of prediction coefficients is limited to one-sided or causal relationships. It can, therefore, be used successfully to estimate input-output relationships in open loop systems, but the estimates are distorted when applied to input-output data gathered under closed loop, feedback control, conditions. A solution to this problem was presented by

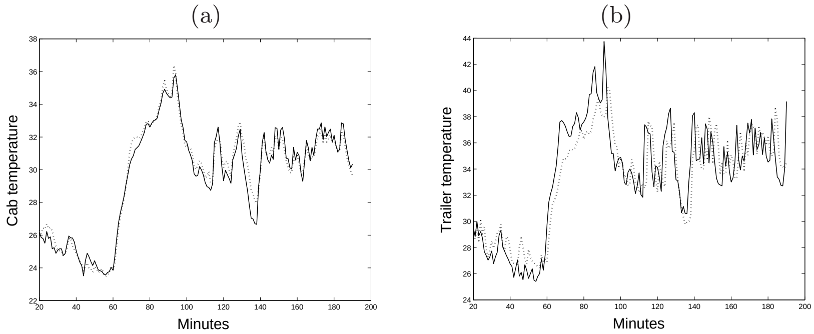


Figure 2: Predictions of transport vehicle temperatures: (a) The cab temperatures (solid line) with values predicted (dotted line) from the trailer temperatures, (b) the trailer temperatures (solid line) with values predicted (dotted line) from the cab temperatures.

Akaike and Nakagawa in the industrial context of designing a cement kiln controller. It was based on multivariate autoregressive modelling of the records of plant variables. The identified model could also be used directly in plant control by expressing it in state space form. The predictions in Figure 2(b) were obtained in this way.

From that point on, time domain methods, and, particularly in the multivariate context, empirical autoregressive modelling, have dominated the methodology for time series analysis. However, the spectacular success in the univariate context, of autoregressive moving-average (ARMA) models and their extensions to integrated and seasonal processes, has not carried over to the multivariate context. Despite the fact that multivariate ARMA models were formulated many years ago [15], and much effort has been been put into procedures for their identification, see for example Tiao and Tsay [16], there are very few examples of real applications compared with those of the multivariate (pure) autoregressive model. More successful has been the state space identification of multivariate time series models, see for example Aoki [3], in which the states are selected to form a basis of the multivariate time series prediction space. Although these state models have a multivariate ARMA representation, this is not required for their application to prediction and control.

In the econometric literature, the multivariate (or vector) autoregressive model is still dominant. Structural forms have been used to incorporate economic constraints, and Bayesian formulations to incorporate prior beliefs, as in Doan, Litterman and Sims [7]. The use of the concept of co-integration to characterise and test for persistence in the relationships between multivariate series, has depended very much on vector autoregressions to account for any residual autocorrelation in the error correction model. The reason for this dominance must be, in large part, the simplicity of the multivariate autoregressive model, and its convenience for order selection, estimation and theoretical analysis. It also has the potential, by choice of a sufficiently high order, to approximate closely any linear process.

The question is therefore, whether the multivariate autoregressive model does provide, essentially, for all our requirements in the world of linear multiple time series modelling. In asking this we will leave aside the problem of seasonality, and restrict the question to non-seasonal series, because seasonality can often be removed or modelled separately. The answer, we believe, is yes in many cases. But there are important reasons why, in practice, the multivariate autoregressive model is *not* fully adequate. The fact that ARMA models are used for univariate series suggests that pure autoregressive models may be less than adequate. The reason may simply be parsimony. The autoregressive approximation may require rather more coefficients than an ARMA model, to achieve the same predictive accuracy. If a criterion such as AIC [1] is used to select the order automatically, then the penalty on the number of parameters may compromise this predictive accuracy, particularly at high lead times, when the series length is small.

The number of coefficients in a multivariate autoregressive model will generally be much greater, for a given order (maximum lag) of model, than for a univariate model. The loss of predictive ability that results from the requirement to choose a relatively low order model may therefore be much more important. The class of models we describe in the next two sections provides one possible, and simple, way to mitigate this loss of predictive ability, without foregoing most of the attractive features of the standard multivariate autoregression.

2 A basis for prediction

In both the discrete and the continuous case, the same idea underlies the models that we formulate in the next section. A chosen, finite, number p of weighted functions of the present and past values of the process will be used as linear predictors of future values. We will call these the *ZAR states* in the discrete case, and *CZAR states* in the discrete case. For continuous time series the models are expressed in terms of a continuous record of the process, but they are also very useful in applications to irregularly sampled data, or, in the case of multiple time series, when different series are recorded regularly but at different sampling rates. In these contexts, the state space form of the model is integrated to determine the state transition from the time of each observation to the next.

A discrete model, very closely related to the univariate form of the discrete model which we describe, was presented by Wahlberg and Hannan [17]. A continuous model, exactly equivalent to the univariate form of the continuous model which we describe, was presented by Belcher, Hampton and Tunnicliffe Wilson [4].

In the case of a discrete process x_t, $t = \ldots, 1, 2, 3, \ldots$, the ZAR states

$x_{t,k}$, at time t, are defined for orders $k = 0, 1, \ldots, p-1$, by

$$x_{t,k} = W^k x_t \tag{2}$$

where the operator W is known as the generalised shift operator, and is defined in terms of the backward shift operator B and a specified smoothing coefficient, or discount factor, θ, by

$$W = \frac{B-\theta}{1-\theta B} = -\theta + (1-\theta^2)(B + \theta B^2 + \theta^2 B^3 + \ldots). \tag{3}$$

In practice W is applied by the recursive calculations,

$$x_{t,k+1} = x_{t-1,k} - \theta x_{t,k} + \theta x_{t-1,k+1}, \tag{4}$$

taking $x_{t,0} = x_t$. The choice of θ is in the range $0 \leq \theta < 1$, and in the case $\theta = 0$, the state $x_{t,k}$ reduces to the lagged value x_{t-k}.

For the continuous time process $x(t)$, the CZAR states $x_k(t)$, at time t, are defined, for order $k = 0, 1, \ldots p-1$, by

$$x_k(t) = Z^k x(t) \tag{5}$$

where the operator Z is defined formally in terms of the Laplace (or differential) operator s, and a decay rate constant κ in the range $\kappa > 0$, by

$$Z = \frac{1 - s/\kappa}{1 + s/\kappa} = \frac{\kappa - s}{\kappa + s}. \tag{6}$$

There is, however, no requirement of differentiability placed upon a series to which this operator is applied, because it is well defined as

$$Zx(t) = -x(t) + 2\kappa \int_{\tau=0}^{\infty} \exp(-\kappa\tau) x(t-\tau) d\tau, \tag{7}$$

for any second order stationary process $x(t)$.

The operators are equally well defined when x_t or $x(t)$ is a vector process of dimension m, though we note that a set of mp scalar functions of the present and past is then defined.

Figure 3 shows the weight functions applied to present and past values for the orders $k = 1, \ldots, 5$ for the discrete operator, taking $\theta = 0.5$, and orders $k = 1$, 3 and 5 for the continuous case, taking, without loss of generality, $\kappa = 1$.

In each case, if we were to let $p \to \infty$, we would obtain a basis for the present and past values of the series (taking time t as the present). The idea is that if we are to limit the number p, of linear functions of the present and past, that we use for predicting the future, then the states defined above give us greater flexibility in the discrete case, than the simple choice of lagged values x_t, x_{t-1}, $\ldots x_{t-p+1}$. The effective range of past values that are weighted

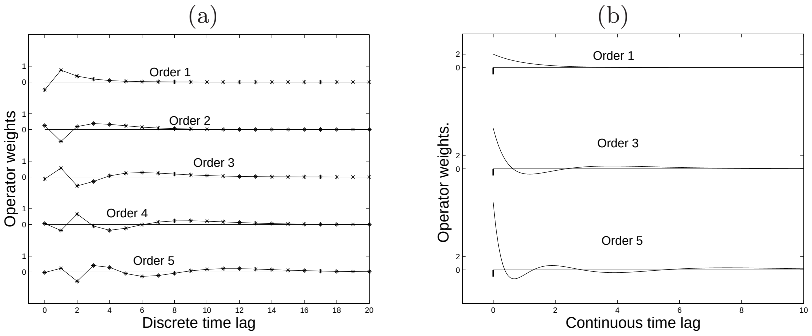


Figure 3: (a) Discrete weights for the first 5 orders of the ZAR operator, (b) continuous weights for orders 1, 3 and 5 of the CZAR operator.

into the predictors is approximately $p(1+\theta)/(1-\theta)$, rather than p. In the continuous case the effective range is approximately $2p/\kappa$.

There is no guarantee that, for a given discrete process, the choice of $\theta > 0$ will define better predictors. However, consider a continuous process $x(\tau)$, that is sampled at times $\tau = \delta t$, to give the discrete process x_t. Defining the ZAR states of x_t by setting $\theta = 1 - \kappa\delta$, these will converge, appropriately, to the CZAR states of $x(\tau)$, as $\delta \to 0$. The consequence of using the simple lagged states x_{t-k}, regardless of how small δ might become, would lead in the limit to states that were equivalent to $x(\tau)$ *and its derivatives to order* $p-1$. There is in general no guarantee that these would exist. That is why the pure autoregressive model in continuous time, that uses these derivatives as its states, *is unable to approximate* an arbitrary continuous time stationary process, though the order is increased indefinitely.

For this reason, the advantage of the CZAR model, proposed in the next section, over the standard continuous time autoregressive (CAR) model is undeniable, in terms of empirical approximation. The success of the univariate application of the CZAR model lead us to consider the discrete ZAR form. The foregoing argument suggests that whenever a discrete process might be considered to be a sampled continuous process, the discrete ZAR model should be preferred to the standard AR model, for its approximation.

The weight functions that we use to define the ZAR and CZAR states are closely related to the respective discrete and continuous Laguerre functions, which have the possible advantage of providing orthogonal bases of the past and present. Partington [14] describes a variety of similar weight functions that could be used to define a basis of the past observations of a discrete process. Bray [6] uses a basis that differs from the Laguerre functions, but may be orthogonalised to provide a similar basis.

Our use of the operator Z was developed from the application of the Cayley-Hamilton transformation to reparameterisation of continuous time models by Belcher et al. [4]. This transformation has been widely used to map from continuous time to discrete time systems. Most famously, Wiener [19] solved the prediction problem for continuous time series by transforming it to that of prediction for a discrete parameter process. The exposition by Doob [8, p. 582] sets this out clearly. The operator W may be motivated as the discrete analogue of Z, in which the Möebius transformation of the unit disk to itself replaces the Cayley-Hamilton transformation.

3 Extended autoregressive models

We propose models for zero mean stationary processes based on the previously introduced concepts. These are readily extended by the addition of a constant term or other fixed regressors, to processes with non-zero mean. In the following, θ and κ are taken to be pre-specified coefficients, with α_i and φ_i model parameters. The ZAR(p, θ) model for a discrete vector process x_t, that is implied by the use at time $t-1$ of the linear predictors defined by (2), is

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-1,1} + \cdots + \alpha_p x_{t-1,p-1} + e_t, \tag{8}$$

where e_t is white noise with variance σ_e^2. When this model is true, e_t is the linear innovation in x_t. This is the most convenient form for many purposes, such as model estimation and prediction, and we call it the *predictive form*, but we also present an algebraically equivalent form of this model, which we term the *natural form*, as follows:

$$x_t = \varphi_1 x_{t,1} + \varphi_2 x_{t,2} + \cdots + \varphi_p x_{t,p} + n_t, \tag{9}$$

where n_t follow the AR(1) model:

$$n_t = \theta n_{t-1} + \varepsilon(t), \tag{10}$$

$\varepsilon(t)$ being white noise with variance σ_ε^2. We also write (9) as

$$\varphi(W) x_t = \left(1 - \varphi_1 W - \varphi_2 W^2 - \cdots - \varphi_p W^p\right) x_t = n_t. \tag{11}$$

We describe (9) as the natural form of the model because the process defined, for any fixed t, by

$$y_k = W^{-k} x_t, \tag{12}$$

is also a stationary process, and 9) is just a *standard* autoregressive approximation of y_k.

We also note that (9) is equivalent to an ARMA$(p, p-1)$ model with a pre-specified moving average operator $(1 - \theta B)^{p-1}$. The model presented by Wahlberg and Hannan [17], and the model of Morton and Tunnicliffe Wilson [12], are very similar, except that they have ARMA(p, p) representations.

The CZAR(p, κ) model for a continuous vector process $x(t)$ is analogous. The predictive form of model is

$$dx(t) = [\alpha_1 x(t) + \alpha_2 x_1(t) + \cdots + \alpha_p x_{p-1}(t)]\, dt + dB(t), \tag{13}$$

when $B(t)$ is Brownian motion with diffusion variance σ_b^2. The natural, algebraically equivalent, form of this model is

$$x(t) = \varphi_1 x_1(t) + \varphi_2 x_2(t) + \cdots + \varphi_p x_p(t) + n(t), \tag{14}$$

where $n(t)$ now follows the continuous time AR(1) model, or CAR(1) model:

$$dn(t) = -\kappa n(t)dt + dH(t), \tag{15}$$

where $H(t)$ is Brownian motion with variance σ_h^2. We also write (14) as

$$\varphi(Z)x(t) = \left(1 - \varphi_1 Z - \varphi_2 Z^2 - \cdots - \varphi_p Z^p\right) x(t) = n(t). \tag{16}$$

We describe (14) as the natural form of the model because the process defined, for any fixed t, by

$$y_k = Z^{-k} x(t), \tag{17}$$

is also a stationary process, and (14) is just a *standard* autoregressive approximation of y_k. We note that (14) is equivalent to a CARMA$(p, p-1)$ model with moving average operator $(\kappa + s)^{p-1}$.

4 Examples

Our first example illustrates the effect on predictions of using a discrete trivariate ZAR model for the three series of monthly flour prices that were modelled by Tiao and Tsay [16].

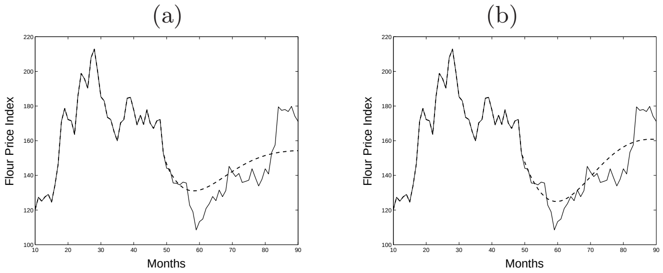


Figure 4: Predictions of monthly flour prices at Buffalo, using past values of three series of flour prices, at Buffalo, Minneapolis and Kansas City: (a) predictions (dotted line) from a standard AR(2) model, (b) predictions from a ZAR(2, 0.5) model.

In Figure 4 we see forecasts of just one of the three series, but made using two trivariate models. Using the AIC [1], a standard AR(2) model and a ZAR(2,0.5) model were selected. This example illustrates the fact that forecasts made using the ZAR model tend to show less damped behaviour. Although these are in-sample forecasts, and too much must not be read into one such example, the ZAR model forecasts tend to predict better the turning points of the irregular cyclical behaviour of the series.

The three flour price series were very similar in nature, and it is natural to represent them by a symmetric vector autoregression. Our second example is very different; the data arise from what is clearly an input-output system. The rainfall is measured at two locations in a river catchment, and the river-flow from the catchment is also measured. Figure 5 shows the hourly measurements over a period slightly in excess of four days. The river-flow record is much more slowly changing than the rainfall record and visual inspection shows that the response from input to output is spread over a period of several hours, possibly with a range of time constants reflecting some relatively rapid, and some relatively slow runoff. The objective is to use the rainfall record to predict the river flow. The transfer function of this response is difficult to estimate using spectral analysis because it is so dispersed over many lags. The use of the ZAR model is appropriate here because of this dispersed response. Using the AIC a standard AR(2) model was selected for the three series, whereas a ZAR(6, 0.75) was selected. The choice of 0.75 for the smoothing parameter is not critical, but was chosen because the low frequency delay in the W operator is about $1.75/0.25 = 7$ hours.

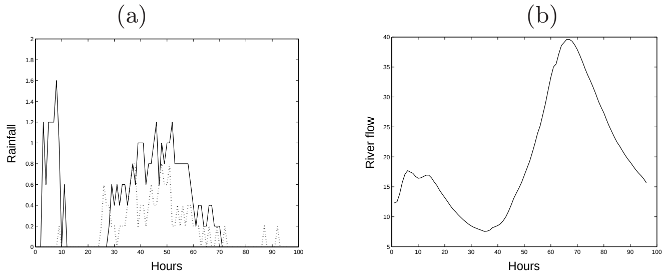


Figure 5: Hourly records of rainfall and river-flow in a single catchment: (a) the solid and broken lines show the rainfall at two gauges in the catchment, (b) the river flow .

A model of relatively low order can then capture a response covering a period of more than one day. In fact the AR(2) model gave very poor in-sample predictions, whereas the ZAR(6, 0.75) produced extremely good in-sample predictions. A fair comparison is illustrated using models of the same order, an AR(3) and a ZAR(3,0.75) model.

Figure 6(a) shows 'predictions' of the river flow from hour 20 using the fitted AR(3) model. The model parameters are estimated using the first 80 values of all three series. Given these parameters, the predictions of river-flow shown from hour 20 are constructed using the rain-fall series alone over that period. A state space representation is used with the Kalman filter to compute this. The peak river-flow is substantially under-predicted. Figure 6(b) shows the corresponding predictions using the ZAR(3,0.75) model.

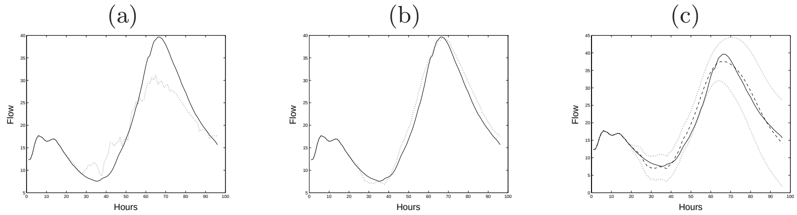


Figure 6: Predictions of the river flow (solid line) using different models and information: (a) the dotted line shows predictions using a trivariate AR(3) model, based on river flow information up to hour 20, and full knowledge of the rainfall throughout the record, (b) similar predictions using a trivariate ZAR(3,0.75) model, (c) predictions (broken line) are obtained as in (b), except that the known rainflow is used only up to hour 50, and thereafter all the series are predicted: the dotted lines show 90% probability limits for the forecasts.

These are very close to the actuality. Figure 6(c) is constructed using the same ZAR(3,0.75) model, but no observations of rainfall or river-flow are used beyond hour 50. The prediction limits are shown on this figure and rapidly widen beyond that hour, but they provide a realistic and useful bound on the peak river flow many hours later. The last 20 observations were not used in model estimation, so their predictions are genuinely out-of sample. In this example the ZAR model reveals its potential.

Our first example of the CZAR model relates to discrete time series with different, and varying, sampling intervals. Figure 7(a) shows monthly Claimant Count (CC) figures that have been long used as a measure of unemployment. A more recent measure of unemployment has been the Labour Force Survey (LFS) estimate, which is shown in the same figure. The LFS estimate was recorded annually, then quarterly. In the figure, the quarterly measurements have been interpolated monthly. These series were analysed by Harvey and Chung [9], in which one of the aims was to estimate the slope of the LFS series by using a bivariate model to 'borrow' information from the more frequently observed CC series. A continuous time model is natural for such series, and we estimated the bivariate CZAR(2,0.5) model. We report the use of this model for slope estimation, in Morton and Tunnicliffe Wilson [12]. Here, we illustrate its application to prediction. Figure 7(b) shows forecasts of the LFS unemployment and their error limits obtained from this model. The bivariate model enables good monthly forecasts to be produced, from a point where only 8 annual values have been recorded.

Our final example is a bivariate model of data which is truly sampled irregularly. Kirchner and Weil [11] present a compendium of marine fossil records which indicate the pattern of extinctions and originations of marine animals over the past 545 million years (Myrs). The records are arranged into 108 stratigraphic intervals which vary in length from 2.5 to 12.5 Myrs,

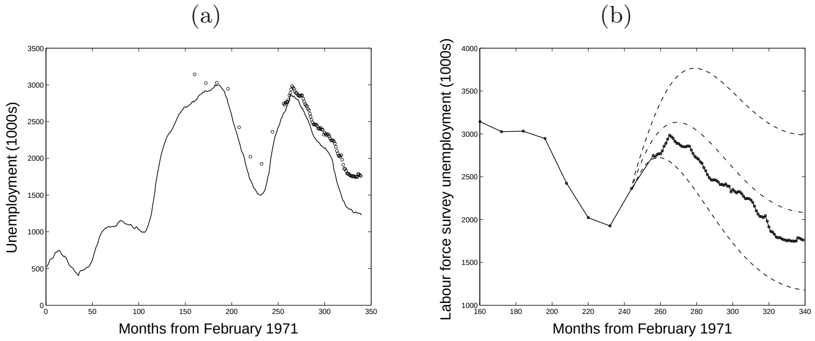


Figure 7: (a) The Claimant Count (solid line) unemployment series, and the Labour Force Survey (small circles) unemployment series, (b) the Labour Force Survey series (solid line) with forecasts and forecast error limits (broken lines) .

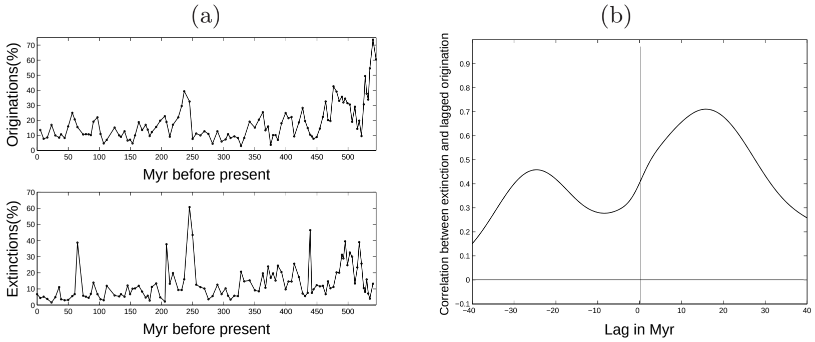


Figure 8: (a) The series of originations and extinctions of genera, (b) the estimated lagged cross-correlation function between these series .

and for each of these the number of families and genera of marine animals to appear and disappear is documented.

The objective is to investigate the relationship between the series, and in particular, the recovery of species following mass extinctions. Figure 8(a) shows the series of genera. We fitted a bivariate CZAR(5,0.5) model to the logarithms of these series. Figure 8(b) shows the cross-correlation function derived from this model. The peak is at the lag of 16 Myr, which is similar to that obtained by Kirchner and Weil using other methods.

References

[1] Akaike H. (1973). *A new look at statistical model identification.* IEEE Transactions on Automatic Control **AC-19**, 716 − 723.

[2] Akaike H., Nakagawa T. (1988). *Statistical analysis and Control of Dynamic Systems*, Kluwer, Dordrecht.

[3] Aoki M. (1990). *State space modelling of time series.* Springer-Verlag, Berlin.

[4] Belcher J., Hampton J.S., Tunnicliffe Wilson G. (1994). *Parameterisation of continuous time autoregressive models for irregularly sampled time series data.* J. Royal Statist. Soc. B **56**, 141 – 155.

[5] Box G.P., Jenkins G.M. (1970). *Time series analysis: forecasting and control.* Holden-Day, San Francisco.

[6] Bray J. (1971). *Dynamic equations for economic forecasting with the G.D.P. - unemployment relation and the growth of G.D.P. in the United Kingdom as an example.* J. Royal. Statist. Soc. A **134**, 167 – 227.

[7] Doan T., Litterman R., Sims C. (1984). *Forecasting and conditional projections using realistic prior distributions.* Econometric Reviews **3**, 1 – 100.

[8] Doob J.L. (1953). *Stochastic processes.* Wiley.

[9] Harvey A.C., Chung C. (2000). *Estimating the underlying change in unemployment in the UK.* J. Royal Statist. Soc. A **163**, 303 – 340.

[10] Jenkins G.M., Watts G. D. (1969). *Spectral Analysis and its Applications.* Holden-Day, San Francisco.

[11] Kirchner J.W., Weil A (2000). *Delayed biological recovery from extinctions throughout the fossil record.* Nature **44**, 177 – 180.

[12] Morton A.S., Tunnicliffe Wilson G. (2001). *Extracting economic cycles using modified autoregressions.* The Manchester School **69**, 574 – 585.

[13] Morton A.S., Tunnicliffe Wilson G. (2003). *A class of modified high order autoregressive models with improved resolution of low frequency cycles.* J. Time Series Analysis, (to appear).

[14] Partington J.R. (1997). *Interpolation, identification, and sampling.* Clarendon Press, Oxford.

[15] Quenouille M.H. (1957). *The analysis of multiple time series.* Griffin, London.

[16] Tiao G.C., Tsay R.S. (1989). *Model specification in multivariate time series (with discussion).* J. Royal Statist. Soc. B **51**, 157 – 213.

[17] Wahlberg B., Hannan, E.J. (1993). *Parametric signal modelling using laguerre filters.* The Annals of Applied Probability **3**, 467 – 496.

[18] Whittle P. (1963). *Prediction and regulation.* English Universities Press. London.

[19] Wiener N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series.* Cambridge, New York.

Address: G. Tunnicliffe-Wilson, A. Morton, Dept. of Mathematics and Statistics, Lancaster University, UK

E-mail: `G.Tunnicliffe-Wilson@lancaster.ac.uk`

TOTAL LEAST SQUARES AND ERRORS-IN-VARIABLES MODELING: BRIDGING THE GAP BETWEEN STATISTICS, COMPUTATIONAL MATHEMATICS AND ENGINEERING

Sabine Van Huffel

Key words: Total least squares, errors-in-variables, orthogonal regression, singular value decomposition, numerical algorithms.

COMPSTAT 2004 section: Numerical methods for statistics.

Abstract: The main purpose of this paper is to present an overview of the progress of a modeling technique which is known as Total Least Squares (TLS) in computational mathematics and engineering, and as Errors-In-Variables (EIV) modeling or orthogonal regression in the statistical community. The basic concepts of TLS and EIV modeling are presented. In particular, it is shown how the seemingly different linear algebraic approach of TLS, as studied in computational mathematics and applied in diverse engineering fields, is related to EIV regression, as studied in the field of statistics. Computational methods, as well as the main algebraic, sensitivity and statistical properties of the estimators, are discussed. Furthermore, generalizations of the basic concept of TLS and EIV modeling, such as structured TLS, Lp approximations, nonlinear and polynomial EIV, are introduced and applications of the technique in engineering are overviewed.

1 Introduction and problem formulation

The Total Least Squares (TLS) method is one of several linear parameter estimation techniques that has been devised to compensate for data errors. The basic motivation for TLS is the following: Let a set of multidimensional data points (vectors) be given. How can one obtain a linear model that explains these data? The idea is to modify all data points in such a way that some norm of the modification is minimized subject to the constraint that the modified vectors satisfy a linear relation. Although the name "total least squares" appeared in the literature only 25 years [15] ago, this method of fitting is certainly not new and has a long history in the statistical literature, where the method is known as "orthogonal regression", "errors-in-variables regression" or "measurement error modeling". The univariate line fitting problem was already discussed since 1877 [2]. More recently, the TLS approach to fitting has also stimulated interests outside statistics. One of the main reasons for its popularity is the availability of efficient and numerically robust algorithms in which the Singular Value Decomposition (SVD) plays a prominent role [15].

Another reason is the fact that TLS is an application oriented procedure. It is suited for situations in which all data are corrupted by noise, which is almost always the case in engineering applications. In this sense, TLS and EIV modeling are a powerful extension of classical least squares and ordinary regression, which corresponds only to a partial modification of the data.

A comprehensive description of the state of the art on TLS from its conception up to the summer of 1990 and its use in parameter estimation has been presented in [33]. While the latter book is entirely devoted to TLS, a second [34] and third book [35] present the progress in TLS and in the broader field of errors-in-variables modeling respectively from 1990 till 1996 and from 1996 till 2001.

The problem of *linear parameter estimation* arises in a broad class of scientific disciplines such as signal processing, automatic control, system theory and in general engineering, statistics, physics, economics, biology, medicine, etc. It starts from a model described by a linear equation:

$$\xi_1\beta_1 + \ldots + \xi_p\beta_p = \eta \tag{1}$$

where $\xi_1, \ldots, \xi_p$ and η denote the variables and $\beta = [\beta_1, \ldots, \beta_p]^T \in \mathbb{R}^p$ plays the role of a parameter vector that characterizes the specific system. A basic problem of applied mathematics is to determine an estimate of the *true* but *unknown* parameters from certain measurements of the variables. This gives rise to an *overdetermined* set of n linear equations $(n > p)$:

$$X\beta \approx y \tag{2}$$

where the ith row of data matrix $X \in \mathbb{R}^{n\times p}$ and vector $y \in \mathbb{R}^n$ contain respectively the measurements of the variables $\xi_1, \ldots, \xi_p$ and η.

In the classical least squares approach, as commonly used in ordinary regression, the measurements X of the variables ξ_i are assumed to be free of error and hence, all errors are confined to the observation vector y. However, this assumption is frequently unrealistic: sampling errors, human errors, modeling errors and instrument errors may imply inaccuracies of the data matrix X as well. One way to take errors in X into account is to introduce *perturbations also in* X. Therefore, the following TLS problem was introduced in the field of computational mathematics [14], [15] ($R(X)$ denotes the range of X and $\|X\|_F$ its Frobenius norm [16]):

Definition 1.1 (Total Least Squares problem). *Given an overdetermined set of n linear equations $X\beta \approx y$ in p unknowns β. The total least squares problem seeks to*

$$\min_{\widehat{\Delta},\widehat{\epsilon},\widehat{\beta}} \| \, [\widehat{\Delta} \; \widehat{\epsilon}] \, \|_F \quad \textit{subject to} \;\; (X - \widehat{\Delta})\widehat{\beta} = y - \widehat{\epsilon} \tag{3}$$

$\widehat{\beta}$ is called a TLS solution and $[\widehat{\Delta} \; \widehat{\epsilon}]$ the corresponding TLS correction.

This paper is organized as follows. Section 2 describes the univariate EIV regression problem from a statistical point of view. Section 3 then formulates the TLS problem from a computational point of view and shows the relationship with univariate EIV regression. Next, Section 4 presents the SVD based basic TLS algorithm, while Section 5 describes major properties of the TLS approach. Furthermore, extensions of the technique are discussed in Section 6 while Section 7 overviews the many applications of TLS in engineering fields. Finally, Section 8 gives the conclusions.

2 Univariate EIV regression: a statistical approach

2.1 Model formulation

For the simplest EIV model, the goal is to estimate from bivariate data a straight line fit between 2 variables, both of which are measured with error.

Definition 2.1 (Univariate Ordinary Regression). *For a sample size of n, $(\xi_i, y_i), i = 1, \ldots, n$, the standard regression model with one explanatory variable is given by*

$$\beta_0 + \xi_i\beta_1 + \epsilon_i = y_i, \quad i = 1, \ldots, n \tag{4}$$

where the independent variable ξ_i is either fixed or random and the error ϵ_i has zero mean and is uncorrelated with ξ_i.

The unknown intercept β_0 and slope β_1 are usually estimated using a Least-Squares (LS) approach for reasons of computational efficiency.

Definition 2.2 (Univariate EIV Regression). *For a sample size of n, (x_i, y_i), $i = 1, \ldots, n$, the univariate EIV regression model is defined as follows. The unobservable true variables (ξ_i, η_i) satisfy*

$$\beta_0 + \xi_i\beta_1 = \eta_i, \quad i = 1, \ldots, n \tag{5}$$

however, one observes $(x_i, y_i), i = 1, \ldots, n$, which are the true variables plus additive errors (δ_i, ϵ_i):

$$\xi_i = \xi_i + \delta_i \quad \textit{and} \quad y_i = \eta_i + \epsilon_i, \quad i = 1, \ldots, n \tag{6}$$

Assume that $\delta_i, \epsilon_i, \quad i = 1, \ldots, n$, all have finite variances, zero mean (without loss of generality), and are uncorrelated, i.e., $\mathrm{E}(\delta_i) = \mathrm{E}(\epsilon_i) = 0$, $\mathrm{var}(\delta_i) = \sigma_\delta^2$, $\mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$ for all i, $\mathrm{cov}(\delta_i, \delta_j) = \mathrm{cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, $\mathrm{cov}(\delta_i, \epsilon_j) = 0$ for all i, j. Depending on the assumption about ξ_i, three different models are defined. If the ξ_i are unknown constants, then the model is known as a functional relationship. If the ξ_i are independent identically distributed (i.i.d.) random variables and independent of the errors, the model is called a structural relationship and we have: $E(\xi_i) = \mu$ and $\mathrm{var}(\xi_i) = \sigma^2$. A generalization of both models is the ultrastructural relationship which assumes that the ξ_i are independent random variables but not identically distributed, i.e. having possibly different means μ_i and common variance σ^2.

EIV regression looks like standard regression if one rewrites Eqs. (5-6) as

$$\beta_0 + x_i\beta_1 + \zeta_i = y_i \text{ with } \zeta_i \equiv \epsilon_i - \delta_i\beta_1, \quad i = 1, \ldots, n. \tag{7}$$

However, this is not the usual regression model, x_i is random and is correlated with the error term ζ_i: $\operatorname{cov}(x_i, \zeta_i) = -\beta_1\sigma_\delta^2$. This covariance is only zero when $\sigma_\delta^2 = 0$, which is the regression model, or when $\beta_1 = 0$, which is the trivial case. If one attempts to use ordinary regression estimates (least squares) on EIV regression modeled data, one obtains inconsistent estimates.

The seemingly minor change between model (4) and model (5)-(6) has important practical and theoretical consequences. One of the most important differences between both models concerns model identifiability. It is common to assume that all random variables in the EIV regression model are jointly normal. In this case, the structural and functional model are not identifiable [7]. Side conditions need to be imposed, the most common of which are the following: (1) the ratio of the error variances, $\lambda \equiv \sigma_\epsilon^2/\sigma_\delta^2$, is known; (2) σ_δ^2 is known; (3) σ_ϵ^2 is known; (4) both of the error variances, σ_δ^2 and σ_ϵ^2, are known. The first assumption is the most popular and is the one with the most published theoretical results, dating back to Adcock [2], [3]. It also leads to the commonly known Orthogonal Regression (OR) estimator. Indeed, if λ is known, the data can be scaled so that $\lambda = 1$. In this case, the maximum likelihood solution of the normal EIV regression problem is OR, which minimizes the sum of squares of the orthogonal distances from the data points to the regression line instead of the sum of squares of the vertical distances, as in standard regression (see Figure 1).

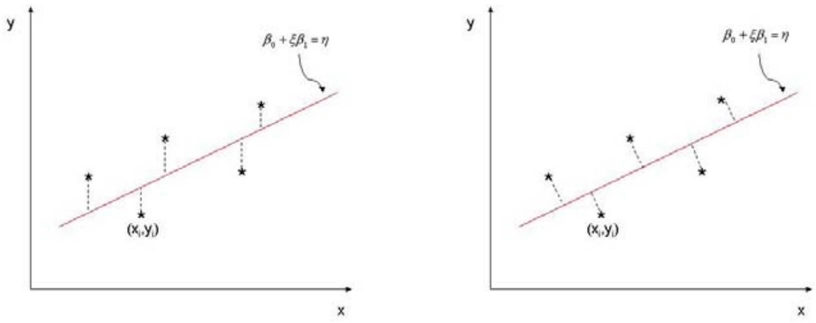


Figure 1: Standard regression (LS) Orthogonal regression (TLS)

2.2 Parameter estimation

Assume that the data have been properly scaled so that $\lambda = 1$. For the functional relationship, the likelihood function is $L(\beta_0, \beta_1, \sigma_\delta^2, \xi_1, \ldots, \xi_n) \propto$

$$\sigma_\delta^{-2n} \exp\left[\frac{-1}{2\sigma_\delta^2}\left\{\sum_{i=1}^{n}(x_i - \xi_i)^2 + (y_i - \beta_0 - \xi_i\beta_1)^2\right\}\right] \tag{8}$$

Note that $\delta_i = x_i - \xi_i$ and $\epsilon_i = y_i - \beta_0 - \xi_i\beta_1$ so that maximizing (8) requires minimizing $\sum(\delta_i^2 + \epsilon_i^2)$, which means that the sum of squares of the orthogonal distances from the data points to the line is minimized. Adcock [2], [3] considered the appropriate estimator to be orthogonal regression, which has been rediscovered many times during the first half of the 20th century. Lindley [23], however, considered a weighted least squares approach to the model (7) as follows. Estimate β_0, β_1 by taking both errors ϵ_i and δ_i into account to minimize a sum of weighted squared residuals, where the weights are proportional to the reciprocal of the variance of the errors ζ_i, i.e., $\sigma_\epsilon^2 + \sigma_\delta^2\beta_1^2$. Thus, one minimizes:

$$Q(\beta_0, \beta_1) = \frac{1}{n(\sigma_\epsilon^2 + \sigma_\delta^2\beta_1^2)}\sum_{i=1}^{n}(y_i - \beta_0 - \xi_i\beta_1)^2. \tag{9}$$

This minimization problem is solved when λ is known or both σ_ϵ^2 and σ_δ^2 are known. If $\lambda = 1$, the denominator reduces to $1 + \beta_1^2$ and amounts to orthogonal regression. Weighted least squares has drawn much attention in the literature; see [7] for references. Since Sprent [28], the name has standardized to generalized least squares. The success of generalized LS might give the impression that it is *the* LS method for the EIV regression model. Since generalized LS estimation only works for the no-equation-error model with the error covariance matrix known up to a scalar multiple, a unified approach for modifying LS to suit all different assumptions on the error covariance structure is called for. Modified LS is such an approach. The normality assumption on the errors (and on the true variables for the structural and ultrastructural relationships) is not needed, only the existence of second moments. From Eq. (7) it is clear that ζ_i are i.i.d. random variables with zero mean and variance $\sigma_\epsilon^2 + \sigma_\delta^2\beta_1^2$ regardless of the type of relationship. Cheng [7] developed modified LS estimators for β_0 and β_1 by minimizing an unbiased and consistent estimator of the appropriate unknown error variance. The estimators are a function of the residuals. Assuming λ known, an appropriate modified LS estimator for the unknown error variance σ_δ^2 is obtained by minimizing

$$Q(\beta_0, \beta_1) = \frac{1}{n(\lambda + \beta_1^2)}\sum(y_i - \beta_0 - \xi_i\beta_1)^2. \tag{10}$$

Minimizing Q with respect to β_0 and β_1 yields:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{(where } \bar{v} \text{ denotes the mean of a vector } v\text{)} \tag{11}$$

$$\hat{\beta}_1 = \frac{s_{yy} - \lambda s_{xx} + [(s_{yy} - \lambda s_{xx})^2 + 4\lambda s_{xy}^2]^{\frac{1}{2}}}{2 s_{xy}}, \text{ provided } s_{xy} \neq 0 \tag{12}$$

with $s_{xx} = \frac{1}{n} \sum (x_i - \bar{x})^2$, $s_{yy} = \frac{1}{n} \sum (y_i - \bar{y})^2$ and $s_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$ the sample variances and covariance. In summary, the statistical approach seeks for estimators of the EIV regression model with optimal statistical properties (such as maximum likelihood, unbiasedness, consistency, etc.), mostly reflecting asymptotic behaviour as $n \to \infty$. If $p > 1$ explanatory variables ξ are considered, the problem formulation can be extended but the estimator β of dimension p can no longer be found analytically, as derived above, but via an eigenvalue-eigenvector approach [12, 13] or an SVD approach (see further).

3 TLS and EIV regression: a computational approach

3.1 Model formulation

In computational mathematics, measurement errors in linear models are tackled from a geometrical point of view, as explained in Section 1. To enlighten the difference with the statistical approach , we consider the univariate model and first assume that the intercept is zero, i.e. $\beta_0 = 0$. It is assumed that the true variables satisfy a compatible linear relationship, given by Eqs. (5)-(6). The TLS approach then aims to find minimal corrections (in a LS sense) $\hat{\delta}_i$ and $\hat{\epsilon}_i$ to the measured data x_i, y_i such that the corrected data $x_i - \hat{\delta}_i, y_i - \hat{\epsilon}_i$ satisfy exactly the unobserved relationship, i.e.

Definition 3.1 (Univariate TLS problem). *Given* $(x_i, y_i), i = 1, \ldots, n$ *satisfying Eqs. (5)-(6). Find corrections* $\hat{\delta}_i$ *and* $\hat{\epsilon}_i$ *and a slope estimate* $\hat{\beta}_1$ *by minimizing*

$$\min_{\hat{\delta}_i, \hat{\epsilon}_i, \hat{\beta}_1} \sum_{i=1}^{n} (\hat{\delta}_i^2 + \hat{\epsilon}_i^2) \text{ subject to } (x_i - \hat{\delta}_i)\hat{\beta}_1 = y_i - \hat{\epsilon}_i, \quad i = 1, \ldots, n \tag{13}$$

Solving this seemingly different minimization problem leads to the same slope estimator $\hat{\beta}_1$, called the TLS solution, as given in (12). If the underlying relationship is an intercept model, as given by Eqs. (5)-(6), the same TLS approach can be used provided the centered data $x_i - \bar{x}$ and $y_i - \bar{y}$ are used. Alternatively, a mixed LS-TLS approach [33] can be applied to the original data:

Definition 3.2 (Univariate mixed LS-TLS problem). *Given* (x_i, y_i), $i = 1, \ldots, n$ *satisfying Eqs. (5-6). Find corrections* $\hat{\delta}_i$ *and* $\hat{\epsilon}_i$, *an intercept* *estimate* $\hat{\beta}_0$ *and a slope estimate* $\hat{\beta}_1$ *by minimizing*

$$\min_{\hat{\delta}_i, \hat{\epsilon}_i, \hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (\hat{\delta}_i^2 + \hat{\epsilon}_i^2) \ \textit{subject to} \ \hat{\beta}_0 + (x_i - \hat{\delta}_i)\hat{\beta}_1 = y_i - \hat{\epsilon}_i, \quad i = 1, \ldots, n \quad (14)$$

This approach is called mixed LS-TLS because the underlying relationship between the true variables is equivalent with

$$\omega_i \beta_0 + \xi_i \beta_1 = \eta_i, \quad i = 1, \ldots, n \qquad (15)$$

where η_i, ξ_i are unobservable, as expressed by Eq. (6), and $\omega_i \equiv 1 \ \forall i$ is exactly known. Therefore, no corrections are needed for the observations ω_i in contrast to the corresponding observations x_i, y_i of ξ_i, η_i. Hence, the best estimates are found via a mixture of a LS and TLS approach, see [33]. Solving this mixed LS-TLS minimization problem leads to the same slope estimators $\hat{\beta}_0, \hat{\beta}_1$, called the mixed LS-TLS solution, as given in (11)-(12). Hence, for the univariate case, TLS in its simplest version is just orthogonal regression. For $p > 1$ explanatory variables, the TLS problem formulation is generalized as given in definition 1. Further extensions are discussed in Section 6.

3.2 Historical remarks

Although the name 'total least squares' appeared only recently in the literature [14], [15], this method of fitting is certainly not new and has a long history in the statistical literature where the method is known as orthogonal regression or errors-in-variables regression. Indeed, the univariate line fitting problem ($p = 1$) was already discussed since 1877 [2]. Some well-known contributors are Adcock [2], [3], Pearson [26], Koopmans [17], Madansky [24] and York [37] (see [4], [7] for a list of references). The method of orthogonal regression has been rediscovered many times, often independently. About thirty years ago, the technique was extended to multiple regression problems ($p > 1$) and later to multivariate problems which deal with more than one observation vector y, e.g., [29], [13].

More recently, the TLS approach to fitting also stimulated interest outside statistics. In the field of numerical analysis, this problem was first studied by Golub and Van Loan [14], [15]. Their analysis, as well as their algorithm, is strongly based on the SVD. Geometrical insight into the properties of the SVD brought Staar [30] independently to the same concept. Van Huffel and Vandewalle [32] generalized the algorithm of Golub and Van Loan to all cases in which their algorithm fails to produce a solution, described the properties of these so-called nongeneric TLS problems and proved that the proposed generalization still satisfies the TLS criteria if additional constraints are imposed on the solution space. This seemingly different linear algebraic approach is actually equivalent to the method of multivariate EIV regression analysis, studied by Gleser [13]. Gleser's method is based on an eigenvalue-eigenvector analysis, while the TLS method uses the SVD which is numerically more robust in the sense of algorithmic implementation. Furthermore, the TLS algorithm computes the minimum norm solution (called minimum norm TLS) whenever the TLS problem lacks a unique minimizer. These extensions are not considered by Gleser.

In engineering fields, e.g., experimental modal analysis, the TLS technique (more commonly known as the H_v technique), was also introduced about 20 years ago [21]. In the field of system identification, Levin [22] first studied the problem. His method, called the eigenvector method or Koopmans-Levin method [10], computes the same estimate as the TLS algorithm whenever the TLS problem has a unique solution. Compensated least squares was yet another name arising in this area: this method compensates for the bias in the estimator, due to measurement error, and is shown to be asymptotically equivalent to TLS [31]. Furthermore, in the area of signal processing, the minimum norm method was introduced and shown to be equivalent to minimum norm TLS [9]. Finally, the TLS approach is tightly related to the maximum likelihood Principal Component Analysis (PCA) method used in chemometrics [36].

4 Basic TLS algorithm and computational issues

We now analyze the TLS problem by making substantial use of the SVD.

Definition 4.1 (Singular Value Decomposition). *The singular value decomposition (SVD) of the $n \times (p+1)$ matrix $[X\ y]$ is defined by*

$$[X\ y] = U\Sigma V^T \tag{16}$$

where $U = [u_1, \ldots, u_n]$, $u_i \in \mathbb{R}^n$, $U^T U = I_n$ and $V = [v_1, \ldots, v_{p+1}]$, $v_i \in \mathbb{R}^{p+1}$, $V^T V = I_{p+1}$ contain respectively the left and right singular vectors, and $\Sigma = diag(\sigma_1, \ldots, \sigma_r)$, $r = \min\{n, p+1\}$, $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$, are the singular values in decreasing order of magnitude.

To solve Eq. (2) with TLS, bring the set into the form:

$$[X\ y][\beta^T; -1]^T \approx 0 \tag{17}$$

If $\sigma_{p+1} \neq 0$, $[X\ y]$ is of rank $p+1$ and the space S generated by the rows of $[X\ y]$ coincides with $\mathbb{R}^{p+1}$. There is no nonzero vector in the orthogonal complement of S, hence the set of equations (17) is incompatible. In order to obtain a solution, the rank of $[X\ y]$ must be reduced to p. Using the Eckart-Young-Mirsky theorem [16], the best rank p TLS approximation $[\widehat{X}\ \widehat{y}]$ of $[X\ y]$, which minimizes the deviations in variance, is obtained by setting the smallest singular value σ_{p+1} of $[X\ y]$ to zero. The following theorem gives conditions for the **uniqueness** and **existence** of a TLS solution (v_{ij} denotes the (i, j)th entry of matrix V):

Theorem 4.1. Solution of the basic TLS problem $X\beta \approx y$.
Let(16) be the SVD of $[X\ y]$ *and* $\sigma_{\min}(X)$ *the smallest singular value of* X*. If* $\sigma_{\min}(X) > \sigma_{p+1}$*, the rank 1 TLS correction solves the TLS problem (3)*

$$[\Delta\widehat{X}\ \Delta\widehat{y}] = [X\ y] - [\widehat{X}\ \widehat{y}] = \sigma_{p+1}u_{p+1}v_{p+1}^T$$

with $[\widehat{X}\ \widehat{y}] = U\widehat{\Sigma}V^T$, $\widehat{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_p, 0)$ *and the TLS solution*

$$\widehat{\beta} = -\frac{1}{v_{p+1,p+1}}[v_{1,p+1}, \ldots, v_{p,p+1}]^T \tag{18}$$

exists and is the unique solution to $\widehat{X}\beta = \widehat{y}$.

Note the equivalence: $\sigma_{\min}(X) > \sigma_{p+1} \Leftrightarrow \sigma_p > \sigma_{p+1}$ and $v_{p+1,p+1} \neq 0$.

The following algorithm computes (if possible) a TLS solution $\widehat{\beta}$ of $X\beta \approx y$ such that $(X - \Delta\widehat{X})\widehat{\beta} = y - \Delta\widehat{y}$ and $\|[\Delta\widehat{X}\ \Delta\widehat{y}]\|_F$ is minimal.

Algorithm 4.1. Basic TLS solution of $X\beta \approx y$**.** Given $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$.
Step 1: Compute the SVD (16), i.e. $[X\ y] = U\Sigma V^T$
Step 2: If $v_{p+1,p+1} \neq 0$ then $\widehat{\beta} = -\frac{1}{v_{p+1,p+1}}[v_{1,p+1}, \ldots, v_{p,p+1}]^T$

For the univariate case ($p = 1$), one easily proves, using the basic properties of eigenvalue and singular value decompositions, that the SVD based TLS solution, given by $\widehat{\beta}_1 = -v_{12}v_{22}^{-1}$, equals the analytical solution in Eq. (12).

The conditions $\sigma_{\min}(X) > \sigma_{p+1}$, or equivalently $\sigma_p > \sigma_{p+1}$ and $v_{p+1,p+1} \neq 0$, ensure that algorithm 4.1 computes the **unique** TLS solution of $X\beta \approx y$. These conditions are generically satisfied provided X is of full rank and the set $X\beta \approx y$ is not too conflicting. Hence, most TLS problems which arise in practice can be solved by means of algorithm 4.1, in which the TLS solution is obtained by a simple scaling of the right singular vector of $[X\ y]$ corresponding to its smallest singular value.

Extensions of this basic TLS problem to multivariate TLS problems $XB \approx Y$ having more than one right hand side vector, to problems in which the TLS solution is no longer unique or fails to have a solution altogether and to mixed LS-TLS problems that assume some of the columns of X to be error-free, are considered in detail in [33]. In addition, it is shown how to speed up the TLS computations directly by computing the SVD only partially or iteratively if a good starting vector is available. More recent advances, e.g. recursive TLS algorithms, neural based TLS algorithms, rank-revealing TLS algorithms, regularized TLS algorithms, TLS algorithms for large scale problems, etc., are reviewed in [34], [35].

5 TLS properties

Under specific conditions, the TLS solution, as introduced in numerical analysis, computes optimal parameter estimates in models with *only measurement* *error*, referred to as classical *errors-in-variables* (EIV) models. This is shown for the univariate case in Sections 2 and 3. These models are characterized by the fact that the true values of the observed variables satisfy one or more unknown but exact linear relations of the form (1). In particular, in case of one underlying linear relation, we define:

Definition 5.1 (Multiple EIV regression model). *Assume that the n measurements in X, y are related to p unknowns β by :*

$$\Xi\beta = \eta \qquad\qquad X = \Xi + \Delta \;\; \textit{and} \;\; y = \eta + \epsilon \tag{19}$$

where Δ, ϵ represent the measurement errors and all rows of $[\Delta\ \epsilon]$ are i.i.d. with zero mean and covariance matrix $\mathcal{C}$, known up to a scalar multiple σ_ν^2.

If additionally $\mathcal{C} = \sigma_\nu^2 I$ is assumed with I the identity matrix (i.e. Δ_{ij} and ϵ_i are uncorrelated random variables with equal variance) and $\lim_{n\to\infty} \frac{1}{n}\Xi^T\Xi$ exists and is positive definite, then it can be proven [12, 14] that the TLS solution $\widehat{\beta}_{TLS}$ of $X\beta \approx y$ estimates the *true* parameter values β, given by $(\Xi^T\Xi)^{-1}\Xi^T\eta$, *consistently*, i.e. $\widehat{\beta}_{TLS}$ converges to β as $n \to \infty$. This TLS property does not depend on any assumed distributional form of the errors. It should be noted that the TLS correction $[\widehat{\Delta}\ \widehat{\epsilon}]$, being of rank 1 as shown in Theorem 1, can not be considered as an appropriate estimator for the true measurement errors Δ and ϵ, added to the data [33], [15]. Note also that the LS estimates are inconsistent in this case. In these cases, TLS gives better estimates than does LS, as confirmed by simulations [33]. This situation may occur far more often in practice than is recognized. It is very common in agricultural, medical and economic science, in humanities, business and many other data analysis situations. Hence TLS should be a quite useful tool to data analysts. In fact, the keyrole and importance of LS in regression analysis is the same as that of TLS in EIV regression. Nevertheless, a lot of confusion exists in the fields of numerical analysis and statistics about the principle of TLS and its relation to EIV modeling. In particular, the name "Total Least Squares" is still largely unknown in the statistical community, while inversely the concept of EIV modeling did not penetrate sufficiently well in the field of computational mathematics and engineering. Roughly speaking, TLS is a special case of EIV estimation and, as such, TLS is reduced to a method in statistics but, on the other hand, TLS appears in many other fields, where mainly the data modification idea is used and explained from a geometric point of view, independently from its statistical interpretation.

Let us now discuss some of the main properties of the TLS method by comparing them with those of LS. First of all, a lot of insight can be gained by comparing their analytical expressions, given by:

$$\text{LS:} \quad \widehat{\beta}_{LS} = (X^TX)^{-1}X^Ty \tag{20}$$

$$\text{TLS:} \quad \widehat{\beta}_{TLS} = (X^TX - \sigma_{p+1}^2 I)^{-1}X^Ty \tag{21}$$

with X of full rank and σ_{p+1} the smallest singular value of $[X\ y]$.

From a numerical analyst's point of view, these formulas tell us that the TLS solution is more ill-conditioned than the LS solution since it has a higher condition number. This implies that errors in the data more likely affect the TLS solution than the LS solution. This is particularly true under worst case perturbations. Hence, TLS can be considered as a kind of *de*regularizing procedure. However, from a statistical point of view, these formulas tell us that TLS is doing the right thing in the presence of i.i.d. equally sized errors since it removes (asymptotically) the bias by subtracting the error covariance matrix (estimated by $\sigma_{p+1}^2 I$) from the data covariance matrix $X^T X$.

Secondly, while LS minimizes a sum of squared residuals, TLS minimizes a sum of *weighted* squared residuals, expressed as follows:

$$\text{LS:} \qquad \min_{\beta} \|X\beta - y\|^2 \tag{22}$$

$$\text{TLS:} \qquad \min_{z=[\beta^T \; -1]^T} \frac{\|[X \; y]z\|^2}{\|z\|^2} = \min_{\beta} \frac{\|X\beta - y\|^2}{\|\beta\|^2 + 1} \tag{23}$$

From a numerical analyst's point of view, we say that TLS minimizes the Rayleigh quotient. From a statistical point of view, we say that we weight the residuals by multiplying them with the inverse of the corresponding error covariance matrix (up to a scaling factor) to derive consistent estimates.

Other properties of TLS, which were studied in the field of numerical analysis, are its sensitivity in the presence of errors on all data [33]. Differences between the LS and TLS solution are shown to increase when the ratio $\sigma_p([X \; y])/\sigma_{\min}(X)$ is growing. This is the case when the set of equations $X\beta \approx y$ becomes less compatible, when the vector y is growing in length and when X tends to be rank-deficient. Assuming i.i.d. equally sized errors, the improved accuracy of the TLS solution compared to that of LS is maximal when the orthogonal projection of y is parallel with the pth singular vector of X, corresponding to $\sigma_{\min}(X)$. Additional algebraic connections and sensitivity properties of the TLS and LS problem, as well as many more statistical properties of the TLS estimators, based on knowledge of the distribution of the errors in the data, have been described, see [33], [34] for an overview.

6 TLS extensions

The statistical model that corresponds to the basic TLS approach is the no-equation-error EIV regression model with the restrictive condition that the measurement errors on the data are i.i.d. with zero mean and common error covariance matrix, equal to the identity matrix up to an unknown scalar. Most published TLS algorithms just handle this case while other more useful EIV regression estimators did not receive enough attention in computational mathematics. To relax these restrictions, several extensions of the TLS problem have been investigated. In particular, the *mixed LS-TLS* problem formulation allows to extend consistency of the TLS estimator in EIV models, where some of the variables ξ_i are measured without error. The *data least* *squares* problem refers to the special case in which all variables except η are measured with error and was introduced in the field of signal processing by DeGroat and Dowling [8] in the mid nineties. Whenever the errors are independent but *unequally* sized, *weighted TLS* problems should be considered using appropriate diagonal scaling matrices in order to maintain consistency. If, additionally, the errors are also correlated, then the *generalized TLS* problem formulation allows to extend consistency of the TLS estimator in EIV models, provided the corresponding error covariance matrix is known up to a factor of proportionality (see definition 7). More general problem formulations, such as *restricted TLS*, which also allow the incorporation of equality constraints, have been proposed, as well as equivalent problem formulations using other $\mathrm{L_p}$ norms and resulting in the so-called *Total* $\mathrm{L_p}$ *approximations* (see [33] for references). The latter problems proved to be useful in the presence of outliers. Robustness of the TLS solution is also improved by adding regularization, resulting in the *regularized TLS* methods [11], [27], [35]. In addition, various types of bounded uncertainties have been proposed in order to improve robustness of the estimators under various noise conditions and algorithms are outlined [34], [35].

Furthermore, *constrained TLS* problems have been formulated. Arun [5] addressed the unitarily constrained TLS problem, i.e., $XB \approx Y$, subject to the constraint that the solution matrix B should be unitary. He proved that this solution is the same as the solution to the orthogonal Procrustes problem [16, p.582]. Abatzoglou et al [1] considered yet another constrained TLS problem, which extends the classical TLS problem (3) to the case where the errors $[\Delta\ \epsilon]$ in the data $[X\ y]$ are algebraically related. However, if there is a linear dependence among the error entries in $[\Delta\,\epsilon]$, then the TLS solution no longer has optimal statistical properties (e.g. maximum likelihood in case of normality). This happens, for instance, in dynamic system modeling, e.g., in system identification when we try to estimate the impulse response of a system from its input and output by discrete deconvolution. In these so-called *structured TLS* problems, the data matrix $[X\ y]$ is structured, typically block Toeplitz or Hankel. In order to preserve maximum likelihood properties and consistency of the solution [1], [18], the TLS problem formulation, given in definition 1, must be extended with the additional constraint that any (affine) structure of X or $[X\ y]$ must be preserved in $\widehat{\Delta}$ or $[\widehat{\Delta}\ \widehat{\epsilon}]$, where $\widehat{\Delta}$ and $\widehat{\epsilon}$ are chosen to minimize the error in the discrete L_1, L_2 and L_∞ norm. For L_2 norm minimization, various computational algorithms have been presented, as surveyed in [34], [35], and shown to reduce the computation time by exploiting the matrix structure in the computations. In addition, it is shown how to extend the problem and solve it, if latency or equation errors are included. Recently, robustness of the structured TLS solution has been improved by adding regularization, see e.g. [25].

Yet, another important extension is the *elementwise-weighted TLS* (EW-TLS) estimator, which computes consistent estimates in linear EIV models, where the measurement errors are elementwise differently sized or, more generally, where the corresponding error covariance matrices may differ from row to row. Some of the variables are allowed to be exactly known (observable) [19], [35]. Mild conditions for weak consistency of the EW-TLS estimator are given and an iterative procedure to compute it is proposed.

Finally, we mention the important extension to *nonlinear EIV* models, nicely studied in the book of Caroll, Ruppert and Stefanski [6]. In these models, the relationship between the variables ξ_i and η is assumed to be nonlinear. It is important to notice here that the close relationship between nonlinear TLS and EIV stops to exist. Indeed, consider the bilinear EIV model $XBG \approx Y$, in which X, G, and Y are affected by measurement errors. Applying TLS to this model leads to the following bilinear TLS problem:

$$\min_{\widehat{\Delta}_X, \widehat{\Delta}_G, \widehat{\Delta}_Y, B} \|[\widehat{\Delta}_X \ \widehat{\Delta}_G \ \widehat{\Delta}_Y]\|_F^2 \, \text{s.t.} \, (X - \widehat{\Delta}_X)\, B\, (G - \widehat{\Delta}_G) = Y - \widehat{\Delta}_Y$$

However, solving this problem yields inconsistent estimates of B [12]. A consistent estimate can be obtained [20] using the adjusted LS estimator (the full rank case is considered here for reasons of simplicity):

$$\widehat{B}_{ALS} = (X^T X - V_X)^{-1} (X^T Y G^T)(G G^T - V_G)^{-1} \tag{24}$$

with $V_X = E(\Delta_X^T \Delta_X)$, $V_G = E(\Delta_G \Delta_G^T)$ and Δ_X and Δ_G represent the errors on X and G respectively. Corrections for small samples have been derived and shown to give superior performance for small sized problems. Various other types of nonlinear EIV models, including bilinear, polynomial, nonlinear functional, semi-linear and Cox's proportional Hazards models, have been considered and consistent estimators are derived, see [35] for an overview.

7 Applications in engineering fields

Since the publication of the SVD based TLS algorithm [15], many new TLS algorithms have been developed and, as a result, the number of applications in TLS and EIV modeling has increased exponentially in the last decade, because of its emergence in new fields such as computer vision, image reconstruction, speech and audio processing, and its gain in popularity in fields as signal processing, modal and spectral analysis, system identification and astronomy. In [34], [35], the use of TLS and errors-in-variables models in the most important application fields, such as signal processing and system identification, are surveyed and new algorithms that apply the TLS concept to the model characteristics used in those fields are described. In these fields, the structured TLS approach is important. In particular, a lot of common problems in ststem identification and signal processing can be reduced to special types of structured TLS problems, including block Hankel or Toeplitz matrix structures, the essence of which is the LS approximation of a given matrix by a rank-deficient one. For example, in system identification the well-known Kalman filtering is extended to the errors-in-variables context in which noise on the inputs as well as on the outputs is taken into account thereby improving the filtering performance. In the field of signal processing, in particular in-vivo magnetic resonance spectroscopy and audio coding, new state-space based methods have been derived by making use of the TLS approach for spectral estimation with extensions to decimation and multichannel data quantification. In addition, it has been shown how to extend the least mean squares (LMS) algorithm to the EIV context for use in adaptive signal processing and various noise environments. Finally, TLS applications also emerge in other fields, including information retrieval, image reconstruction, multivariate calibration, astronomy, and computer vision. It is shown in [35] how the TLS approach and its generalizations, including structured, regularized and generalized TLS, can be successfully applied.

This list of applications of TLS and EIV modeling is certainly not exhaustive and clearly illustrates the increased interest of TLS and EIV modeling in engineering over the past 20 years.

8 Conclusions

The basic principle of TLS is that the noisy data $[X\ y]$, while not satisfying a linear relation, are modified with minimal effort, as measured by the Frobenius norm, in a 'nearby' matrix $[\widehat{X}\ \widehat{y}]$ which is rank-deficient so that the set $\widehat{X}\beta = \widehat{y}$ is compatible. This matrix $[\widehat{X}\ \widehat{y}]$ is a rank-one modification of the data matrix $[A\ b]$. The solution to the TLS problem can be determined from the SVD of the matrix $[X\ y]$. A simple algorithm outlines the computations of the solution of the basic TLS problem. By 'basic' is meant that only one right-hand side vector y is considered and that the TLS problem is solvable (generic) and has a unique solution. Extensions of this basic TLS problem are discussed. Much of the literature concerns the classical TLS problem $X\beta \approx y$, in which all columns of X are subject to errors, but more general TLS problems, as well as other problems related to classical TLS, have been proposed and are briefly overviewed here.

Engineering applications of the Total Least Squares (TLS) technique have been overviewed. TLS has its roots in statistics where it can be defined as a special case of classical Errors-in-Variables (EIV) regression in which all measurement errors on the data are i.i.d. with zero mean and equal variance. Due to the development of a powerful algorithm based on the SVD in computational mathematics the method became very popular in engineering applications. This is a nice example of interdisciplinary work. However, the danger exists that researchers will focus their attention on the wrong problems which are either unreasonable from a statistical point of view (e.g. biased, inconsistent, not efficient) or not practically useful from an engineering point of view (e.g. assumptions never satisfied). This paper invites any reader to open the frontiers of its own discipline and look over the border into neighbouring areas so that the any engineering problem, dealing with measurement error, is studied in a correct way.

References

[1] Abatzoglou T.J., Mendel J.M. and Harada G.A. (1991). *The constrained total least squares technique and its applications to harmonic superresolution.* IEEE Trans. Acoust., Speech & Signal Processing **39**, 1070 – 1087.

[2] Adcock R.J. (1877). *A problem in least squares.* The Analyst **4**, 183 – 184.

[3] Adcock R.J. (1878). *A problem in least squares.* The Analyst **5**, 53 – 54.

[4] Anderson T.W. (1984). *The 1982 Wald memorial lectures : Estimating linear statistical relationships.* Ann. Statist. **12**, 1 – 45.

[5] Arun K.S. (1992). *A unitarily constrained total least-squares problem in signal-processing.* SIAM J. Matrix Anal. Appl. **13**, 729 – 745.

[6] Carroll R.J., Ruppert D. and Stefanski L.A. (1995). *Measurement error in nonlinear models*, Chapman & Hall/CRC, London.

[7] Cheng C.-L. and Van Ness J.W. (1999). *Statistical regression with measurement error.* Arnold, London.

[8] Degroat R.D. and Dowling E.M. (1993). *The data least squares problem and channel equalization.* IEEE Trans. Sign. Process. **41**, 407 – 411.

[9] Dowling E.M. and Degroat R.D. (1991). *The equivalence of the total least-squares and minimum norm methods.* IEEE Trans. Sign. Process. **39**, 1891 – 1892.

[10] Fernando K.V. and Nicholson H. (1985). *Identification of linear systems with input and output noise : the Koopmans-Levin method.* IEE Proc. D **132**, 30 – 36.

[11] Fierro R.D., Golub G.H., Hansen P.C. and O'Leary D.P. (1997). *Regularization by truncated total least squares.* SIAM J. Sci. Comp. **18**, 1223 – 1241.

[12] Fuller W.A. (1987). *Error measurement models.* John Wiley, New York.

[13] Gleser L.J. (1981). *Estimation in a multivariate "errors in variables" regression model : Large sample results.* Ann. Statist. **9**, 24 – 44.

[14] Golub G.H. (1973). *Some modified matrix eigenvalue problems.* Siam Review **15**, 318 – 344.

[15] Golub G.H. and Van Loan C.F. (1980). *An analysis of the total least squares problem.* SIAM J. Numer. Anal. **17**, 883 – 893.

[16] Golub G.H. and Van Loan C.F. (1996). *Matrix computations.* 3rd ed., The Johns Hopkins Univ.Press, Baltimore.

[17] Koopmans T.C. (1937). *Linear regression analysis of economic time series.* De Erven F. Bohn, N.V. Haarlem.

[18] Kukush A., Markovsky I. and Van Huffel S. (2004). *Consistency of the structured total least squares estimator in a multivariate model.* Journal of Statistical Planning and Inference, to appear.

[19] Kukush A. and Van Huffel S. (2004). *Consistency of elementwise-weighted total least squares estimator in a multivariate errors-in-variables model AX=B.* Metrika **59**, issue 1, to appear.

[20] Kukush A., Markovsky I. and Van Huffel S. (2003). *Consistent estimation in the bilinear multivariate errors-in-variables model.* Metrika **57**, 253 – 285.

[21] Leuridan J., De Vis D., Van Der Auweraer H. and Lembregts F. (1986). *A comparison of some frequency response function measurement techniques.* Proc. 4th Int. Modal Analysis Conf., Los Angeles, CA, Feb. 3-6, 908 – 918.

[22] Levin M.J. (1964). *Estimation of a system pulse transfer function in the presence of noise.* IEEE Trans. Automat. Contr. **9**, 229 – 235.

[23] Lindley D.V. (1947). *Regression lines and the linear functional relationship.* J.R. Statist. Soc. Suppl. **9**, 218 – 244.

[24] Madansky A. (1959). *The fitting of straight lines when both variables are subject to error.* J. Amer. Statist. Assoc. **54**, 173 – 205.

[25] Mastronardi N., Lemmerling P. and Van Huffel S. (2004). *Fast regularized structured total least squares algorithm for solving the basic deconvolution problem.* Numer. Lin. Alg. with Appl., to appear.

[26] Pearson K. (1901). *On lines and planes of closest fit to points in space.* Philos. Mag. **2**, 559 – 572.

[27] Sima D., Van Huffel S. and Golub G.H. (2004). *Regularized Total Least Squares based on quadratic eigenvalue problem solvers.* BIT, to appear.

[28] Sprent P. (1966). *A generalized least squares approach to linear functional relationships.* J.R. Statist. Soc. B **28**, 278 – 297.

[29] Sprent P. (1969). *Models in regression and related topics.* Methuen & Co. ltd., London, UK.

[30] Staar J. (1982). *Concepts for reliable modelling of linear systems with application to on-line identification of multivariable state space descriptions.* PhD thesis, Dept. EE, K.U.Leuven, Leuven, Belgium.

[31] Stoica P. and Söderström T (1982). *Bias correction in least squares identification.* Int. J. Control **35**, 449 – 457.

[32] Van Huffel S. and Vandewalle J. (1988). *Analysis and solution of the nongeneric total least squares problem.* SIAM J. Matrix Anal. Appl. **9**, 360 – 372.

[33] Van Huffel S. and Vandewalle J. (1981). *The total least squares problem: computational aspects and analysis*, SIAM, Philadelphia.

[34] Van Huffel S., editor, (1997). *Recent advances in total least squares techniques and errors-in-variables modeling*, SIAM Proceedings series, SIAM, Philadelphia.

[35] Van Huffel S. and Lemmerling , editors, (2002). *Total least squares and errors-in-variables modeling: Analysis, Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht.

[36] Wentzell P.D., Andrews D.T., Hamilton D.C., Faber K. and Kowalski B.R. (1997). *Maximum likelihood principal component analysis*. J. Chemometrics **11**, 339 – 366.

[37] York D. (1966). *Least squares fitting of a straight line*. Can. J. of Physics **44**, 1079 – 1086.

Acknowledgement: Dr. Sabine Van Huffel is a full professor at the Katholieke Universiteit Leuven, Belgium. Research supported by the KU Leuven research council (GOA-Mefisto 666), the Flemish Government (FWO projects G.0078.01, G.0269.02, G.0270.02, research communities ICCoS, ANMMM), and the Belgian Federal Government (IUAP V-22).

Address: S. Van Huffel, Katholieke Universiteit Leuven, Department of Electrical Engineering, Division ESAT-SCD, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

E-mail: `sabine.vanhuffel@esat.kuleuven.ac.be`

BINARY DATA IN THE PRESENCE OF MISCLASSIFICATIONS

Jorge Alberto Achcar, Edson Zangiacomi Martinez and Francisco Louzada-Neto

Key words: Binary data, sensitivity and specificity, Bayesian analysis.

COMPSTAT 2004 section: Bayesian methods.

Abstract: This contribution concentrates on the sensitivity and specificity of medical tests in the presence of misclassifications

1 Introduction

In many applications of binary data, we have the presence of covariates and misclassifications (see, for example, Geng and Asano [5], or Soares and Paulino [8]). As a special situation, we could have the presence of false positives or false negatives in diagnostic medical tests. In medical terminology, we denote sensitivity as the probability of positive test given that the patient really has the disease and specificity as the probability of a negative test given that the patient has not the disease. Let p be the prevalence of disease in the population and D the disease status, where $D = 1$ (or simply, D) denotes an individual with the disease and $D = 0$ (or $\overline{D}$) denotes a free-disease individual. Thus, $p = P(D)$. Let T be a random variable related to the diagnostic test results, where $T = 1$ (or T) denotes a positive test and $T = 0$ (or $\overline{T}$) denotes a negative test. The sensitivity of the diagnostic test is given by $S_E = P(T|D)$ and the specificity is given by $S_P = P(\overline{T}|\overline{D})$. Observe that $1 - S_E$ and $1 - S_P$ are the probabilities of misclassifications. In the presence of a vector of covariates $\mathbf{X}_i' = (X_{0i}, X_{1i}, \ldots, X_{Li})$, the probability of a positive test (sucess) is given by $\eta_i = p_i S_E + (1 - p_i)(1 - S_P)$, where $i = 1, \ldots, n$. Different parametric choices for p_i could be considered. We assume a logistic regression model given by $e^{\mathbf{x}'\beta} / \left(1 + e^{\mathbf{x}'\beta}\right)$. Observe that we also could assume S_E and S_P dependent on the covariates.

2 Bayesian analysis, all individuals are unverified

Let us assume that all individuals are unverified about the real disease status after an application of a medical test. Assuming a Bernoulli distribution for the test result t_i with sucess probability given by η_i, let $D_i = (t_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ be the data where $t_i = 1$ (T) or 0 ($\overline{T}$) and $\mathbf{X}_i$ is a covariate vector associated to each individual. Also assume the logit link for the probability of prevalence of disease for the individual in the population. The likelihood function for $\theta' = (S_E, S_P, \beta_{p0}, \beta_{p1}, \ldots, \beta_{pL})$ is given by

$$L(\theta) = \prod_{i=1}^{n} \left[p_i S_E + (1 - p_i)(1 - S_P)\right]^{t_i} \left[p_i (1 - S_E) + (1 - p_i) S_P\right]^{1 - t_i}. \tag{1}$$

For a Bayesian analysis of the model, let us assume the following prior distributions for S_E, S_P and $\beta_p' = (\beta_{p0}, \beta_{p1}, \ldots, \beta_{pL}) : S_E \sim Beta(a,b)$, $S_P \sim Beta(c,d)$, $\beta_{pj} \sim N(e_j; f_j^2)$, a, b, c, d, e_j, f_j known, where $j = 0, 1, \ldots, L$, $Beta(a,b)$ denotes a Beta distribution with mean $a/(a+b)$ and variance $ab/\left[(a+b)^2(a+b+1)\right]$ and $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. We also assume prior independence among the parameters. To obtain better performance for the Gibbs sampling algorithm (see, for example, Gelfand and Smith [4]), we introduce latent variables given by $\mathbf{D}_i^* = (\mathbf{D}_{i1}^*, \mathbf{D}_{i2}^*)$ where $\mathbf{D}_{i1}^* = (D_{i1}^*, \overline{D}_{i1}^*)$ given $T_i = 1$ and $\mathbf{D}_{i2}^* = (D_{i2}^*, \overline{D}_{i2}^*)$ given $T_i = 0$ for $i = 1, \ldots, n$. That is,

(a) $\mathbf{D}_{i1}^* = (D_{i1}^*, \overline{D}_{i1}^*)$ given $T_i = 1$ where $D_{i1}^* + \overline{D}_{i1}^* = 1$ and D_{i1}^* is a random variable (disease status) with a Bernoulli distribution with sucess probability given by $h_{1i} = P\left(D_{i1}^* | T_i\right) = p_i S_E \left[p_i S_E + (1-p_i)(1-S_P)\right]^{-1}$.

(b) $\mathbf{D}_{i2}^* = (D_{i2}^*, \overline{D}_{i2}^*)$ given $T_i = 0$ where $D_{i2}^* + \overline{D}_{i2}^* = 1$ and D_{i2}^* is a random variable (disease status) with a Bernoulli distribution with sucess probability given by $h_{2i} = P\left(D_{i2}^* | T_i\right) = p_i\left(1 - S_E\right)\left[p_i\left(1 - S_E\right) + \left(1 - p_i\right) S_P\right]^{-1}$.

In this way, the joint posterior distribution for θ is given by

$$\begin{aligned}\Pi\left(\theta | \mathbf{t}, \mathbf{d}^*\right) \quad \propto \quad & S_E^{a+\sum_{i=1}^n t_i d_{i1}^* - 1} \left(1 - S_E\right)^{b + \sum_{i=1}^n (1-t_i) d_{i2}^* - 1} \times \\ & \times S_P^{c + \sum_{i=1}^n (1-t_i)(1-d_{i2}^*) - 1} \left(1 - S_P\right)^{d + \sum_{i=1}^n t_i (1 - d_{i1}^*) - 1} \times \\ & \times \prod_{j=0}^L \exp\left[-\left(2 f_j^2\right)^{-1} \left(\beta_{pj} - e_j\right)^2\right] \times \\ & \times \prod_{i=1}^n p_i^{t_i d_{i1}^* + (1-t_i) d_{i2}^*} \left(1 - p_i\right)^{t_i (1 - d_{i1}^*) + (1-t_i)(1-d_{i2}^*)}. \quad (2)\end{aligned}$$

The conditional distributions for the Gibbs sampling algorithm are given by

$$S_E | \theta_{(S_E)}, \mathbf{t}, \mathbf{d}^* \sim Beta\left(a + \sum_{i=1}^n d_{i1}^* t_i, b + \sum_{i=1}^n d_{i2}^* \left(1 - t_i\right)\right), \quad (3)$$

$$S_P | \theta_{(S_P)}, \mathbf{t}, \mathbf{d}^* \sim Beta\left(c + \sum_{i=1}^n \left(1 - d_{i2}^*\right)\left(1 - t_i\right), d + \sum_{i=1}^n t_i \left(1 - d_{i1}^*\right)\right), \quad (4)$$

$$\Pi\left(\beta_{pj} | \theta_{(\beta_{pj})}, \mathbf{t}, \mathbf{d}^*\right) \propto \exp\left[-\left(2 f_j^2\right)^{-1} \left(\beta_{pj} - e_j\right)^2\right] \psi(\theta), \quad (5)$$

where

$$\psi(\theta) = \exp\left\{\beta_{pj} \sum_{i=1}^n x_{ji} \left[t_i d_{i1}^* + \left(1 - t_i\right) d_{i2}^*\right] - \sum_{i=1}^n \ln\left(1 + e^{\sum_{j=0}^L \beta_{pj} x_{ji}}\right)\right\},$$

$x_{0i} = 1$, $i = 1, \ldots, n$ and $\theta_{(r)} = \left(\theta_1, \ldots, \theta_{r-1}, \theta_{r+1}, \ldots, \theta_L\right)$. Observe that we should simulate samples for β_{pj}, $j = 0, 1, \ldots, L$ considering the Metropolis-Hastings algorithm. Starting with initial values $\theta^{(0)}$, we simulate samples of the joint posterior distribution (2) following the steps:

(i) generate a sample of $\mathbf{D}^*_{i1} = (D^*_{i1}, \overline{D}^*_{i1})$, $i = 1, \ldots, n$, from the conditional Bernoulli distribution with success probability h_{1i} given $T_i = 1$;

(ii) generate a sample of $\mathbf{D}^*_{i2} = (D^*_{i2}, \overline{D}^*_{i2})$, $i = 1, \ldots, n$, from the conditional Bernoulli distribution with success probability h_{2i} given $T_i = 0$;

(iii) generate a sample from the conditional distributions
$\Pi\left(\theta_1|\theta_{(\theta_1)}, \mathbf{t}, \mathbf{d}^*\right), \ldots, \Pi\left(\theta_L|\theta_{(\theta_L)}, \mathbf{t}, \mathbf{d}^*\right)$.

A special case is given when we do not have the presence of covariates [7]. In this case, we assume the same Beta prior distributions for S_E and S_P and a Beta distribution for p with hyperparameters e and f considering e and f known. Also assuming the introduction of the latent variables $\mathbf{D}^*_{i1}$ given $T_i = 1$ and $\mathbf{D}^*_{i2}$ given $T_i = 0$, the conditional posterior distributions for the Gibbs sampling algorithm are given by (3), (4) and $p|\theta_{(p)}, \mathbf{t}, \mathbf{d}^*$ has a Beta distrbution with parameters $e + \sum_{i=1}^n d^*_{i1} t_i + \sum_{i=1}^n d^*_{i2}\left(1 - t_i\right)$ and $f + \sum_{i=1}^n t_i\left(1 - d^*_{i1}\right) + \sum_{i=1}^n\left(1 - t_i\right)\left(1 - d^*_{i2}\right)$.

3 Model formulation assuming verified and unverified individuals

In medical applications, if we consider only the verified cases (see for example, Begg [1]; Begg and Greenes [2] or Zhou [9], we could obtain biased estimators for S_E and S_P. This occurs when only part of the sampled individuals are verified about their real disease status by a procedure generically denominated by "gold standard". Let V be a random variable related to the verification by a gold standard, where $V = 1$ (or V) denotes a verified individual and $V = 0$ (or $\overline{V}$) denotes an unverified individual. In Table 1, we have the probabilities in the cross-tabulation of the variables V, D and T considering verified and unverified individuals by a gold standard.

	verified (V)		
	D	$\overline{D}$	unverified $(\overline{V})$
T	$p\lambda_{11}S_E$	$(1-p)\,\lambda_{01}\,(1-S_P)$	$p\,(1-\lambda_{11})\,S_E+$ $+\,(1-p)\,(1-\lambda_{01})\,(1-S_P)$
$\overline{T}$	$p\lambda_{10}\,(1-S_E)$	$(1-p)\,\lambda_{00}S_P$	$p\,(1-\lambda_{10})\,(1-S_E)+$ $+\,(1-p)\,(1-\lambda_{00})\,S_P$

Table 1: Probabilities in the cross-tabulation of the variables V, D and T.

Let us assume that D and V are independent random variables. In Table 1, we have $\lambda_{11} = P\left(V|TD\right)$, $\lambda_{01} = P\left(V|T\overline{D}\right)$, $\lambda_{10} = P\left(V|\overline{T}D\right)$, $\lambda_{00} = P\left(V|\overline{TD}\right)$, $S_E = P(T|D)$ and $S_P = P(\overline{T}|\overline{D})$. For the calculation of the probabilities of Table 1, observe that $P\left(T = 1, D = 1, V = 1\right) = P\left(TDV\right) = P\left(D\right)P\left(V|TD\right)P\left(T|D\right) = P\lambda_{11}S_E$. In the same way, we find the other probabilities given in Table 1. Considering the data given by $\mathcal{D}_i = \left(d_i, t_i, v_i\right)$, $i = 1, \ldots, n$, where $d_i = 1$ (D) or 0 $(\overline{D})$; $t_i = 1$ (T) or 0 $(\overline{T})$, and $v_i = 1$ (V) or 0 $(\overline{V})$, the likelihood function for $\theta'_1 = (\lambda_{11}, \lambda_{10}, \lambda_{01}, \lambda_{00}, S_E, S_P, p)$ is

$$\begin{aligned} L\left(\theta_1\right) \;=\; & \prod\nolimits_{i=1}^{n}\left(p\lambda_{11}S_E\right)^{d_i t_i v_i}\left[p\lambda_{10}\left(1-S_E\right)\right]^{d_i(1-t_i)v_i}\times \\ & \times\left[(1-p)\,\lambda_{01}\left(1-S_P\right)\right]^{(1-d_i)t_i v_i}\left[(1-p)\,\lambda_{00}S_P\right]^{(1-d_i)(1-t_i)v_i}\times \\ & \times\left[p\left(1-\lambda_{11}\right)S_E+(1-p)\left(1-\lambda_{01}\right)\left(1-S_P\right)\right]^{t_i(1-v_i)}\times \\ & \times\left[p\left(1-\lambda_{10}\right)\left(1-S_E\right)+(1-p)\left(1-\lambda_{00}\right)S_P\right]^{(1-t_i)(1-v_i)}. \qquad (6) \end{aligned}$$

In the presence of a vector of covariates $\mathbf{X}_i' = (X_{0i}, X_{1i}, \ldots, X_{Li})$, let us assume the logit links for p_i, S_{E_i}, S_{P_i}, λ_{11_i}, λ_{01_i}, λ_{10_i}, and λ_{00_i} given by $\upsilon_{l_i} = \exp\left(\sum_{j=0}^{L}\beta_{lj}x_{ji}\right)\left[1+\exp\left(\sum_{j=0}^{L}\beta_{lj}x_{ji}\right)\right]^{-1}$, for $l = 1, 2, \ldots, 7$; $X_{0i=1}$; $\upsilon_{1_i} = p_i$; $\upsilon_{2_i} = S_{E_i}$; $\upsilon_{3_i} = S_{P_i}$; $\upsilon_{4_i} = \lambda_{11_i}$; $\upsilon_{5_i} = \lambda_{01_i}$; $\upsilon_{6_i} = \lambda_{10_i}$ and $\upsilon_{7_i} = \lambda_{00_i}$, $i = 1, \ldots, n$. In this way, we have a vector of parameters given by $\theta_2' = (\beta_1, \beta_2, \ldots, \beta_7)$, where $\beta_1' = (\beta_{10}, \beta_{11}, \ldots, \beta_{1L})$, $\beta_2' = (\beta_{20}, \beta_{21}, \ldots, \beta_{2L})$,..., $\beta_7' = (\beta_{70}, \beta_{71}, \ldots, \beta_{7L})$.

4 A Bayesian analysis in the presence of covariates

Let us assume the model given in Table 1 assuming the verified and the unverified cases and the presence of a vector of covariates $\mathbf{X}_i'$, $i = 1, \ldots, n$, associated to each individual. Assuming prior independence among the parameters, consider normal $N\left(a_{lj}, b_{lj}^2\right)$ prior distributions for β_{lj}, where a_{lj}, b_{lj} known, $l = 1, 2, \ldots, 7$ and $j = 0, 1, \ldots, L$. We also assume the introduction of latent variables $\mathbf{D}_i^* = (\mathbf{D}_{i1}^*, \mathbf{D}_{i2}^*)$ (see Section 2) where $\mathbf{D}_{i1}^* = (D_{i1}^*, \overline{D}_{i1}^*)$ given $V_i = 0$ and $T_i = 1$ and $\mathbf{D}_{i2}^* = (D_{i2}^*, \overline{D}_{i2}^*)$ given $V_i = 0$ and $T_i = 0$, $i = 1, \ldots, n$. That is,

(a) $\mathbf{D}_{i1}^* = (D_{i1}^*, \overline{D}_{i1}^*)$ given $V_i = 0$ and $T_i = 1$ where $D_{i1}^* + \overline{D}_{i1}^* = 1$, $i = 1, \ldots, n$, and D_{i1}^* is a random variable with a Bernoulli distribution with success probability given by

$$h_{1i} = P\left(D_{i1}^* | \overline{V}_i T_i\right) = \frac{p_i\left(1-\lambda_{11_i}\right)S_{E_i}}{p_i\left(1-\lambda_{11_i}\right)S_{E_i} + \left(1-p_i\right)\left(1-\lambda_{01_i}\right)\left(1-S_{P_i}\right)}. \qquad (7)$$

(b) $\mathbf{D}_{i2}^* = (D_{i2}^*, \overline{D}_{i2}^*)$ given $V_i = 0$ and $T_i = 0$ where $D_{i2}^* + \overline{D}_{i2}^* = 1$, $i = 1, \ldots, n$, and D_{i2}^* is a random variable with a Bernoulli distribution with success probability given by

$$h_{2i} = P\left(D_{i2}^* | \overline{V}_i \overline{T}_i\right) = \frac{p_i\left(1-\lambda_{10_i}\right)\left(1-S_{E_i}\right)}{p_i\left(1-\lambda_{10_i}\right)\left(1-S_{E_i}\right) + \left(1-p_i\right)\left(1-\lambda_{00_i}\right)S_{P_i}}. \qquad (8)$$

The joint distribution for $\theta_2' = (\beta_1, \beta_2, \ldots, \beta_7)$ is given by

$$\begin{aligned}
&\Pi\left(\theta_2|\mathcal{D},\mathbf{t},\mathbf{d}^*\right) \propto \Pi\left(\theta_2\right)\left[\prod\nolimits_{i=1}^{n} \lambda_{00_i}^{r_i^{(001)}}\left(1-\lambda_{00_i}\right)^{s_{2_i}^{(000)}}\right] \times \\
&\times\left[\prod\nolimits_{i=1}^{n} \lambda_{10_i}^{r_i^{(101)}}\left(1-\lambda_{10_i}\right)^{s_{2_i}^{(100)}}\right]\left[\prod\nolimits_{i=1}^{n} \lambda_{01_i}^{r_i^{(011)}}\left(1-\lambda_{01_i}\right)^{s_{1_i}^{(010)}}\right] \times \\
&\times\left[\prod_{i=1}^{n} \lambda_{11_i}^{r_i^{(111)}}\left(1-\lambda_{11_i}\right)^{s_{1_i}^{(110)}}\right]\left[\prod_{i=1}^{n} S_{P_i}^{r_i^{(001)}+s_{2_i}^{(000)}}\left(1-S_{P_i}\right)^{r_i^{(011)}+s_{1_i}^{(010)}}\right] \times \\
&\times\left[\prod\nolimits_{i=1}^{n} S_{E_i}^{r_i^{(111)}+s_{1_i}^{(110)}}\left(1-S_{E_i}\right)^{r_i^{(101)}+s_{2_i}^{(100)}}\right] \times \\
&\times\left[\prod\nolimits_{i=1}^{n} p_i^{r_i^{(111)}+r_i^{(101)}+s_{1_i}^{(110)}+s_{2_i}^{(100)}}\left(1-p_i\right)^{r_i^{(011)}+r_i^{(001)}+s_{1_i}^{(010)}+s_{2_i}^{(000)}}\right], \quad (9)
\end{aligned}$$

where $\Pi\left(\theta_2\right)$ is the joint prior distribution for θ_2, $r_i^{(111)} = d_i t_i \upsilon_i$, $r_i^{(011)} = \left(1-d_i\right) t_i \upsilon_i$, $r_i^{(101)} = d_i\left(1-t_i\right)\upsilon_i$, $r_i^{(001)} = \left(1-d_i\right)\left(1-t_i\right)\upsilon_i$, $s_{1_i}^{(110)} = d_{i1}^* t_i \times \times\left(1-\upsilon_i\right)$, $s_{1_i}^{(010)} = \left(1-d_{i2}^*\right) t_i\left(1-\upsilon_i\right)$, $s_{2_i}^{(100)} = d_{i2}^*\left(1-t_i\right)\left(1-\upsilon_i\right)$, $s_{2_i}^{(000)} = \left(1-d_{i2}^*\right)\left(1-t_i\right)\left(1-\upsilon_i\right)$, $i = 1,\ldots,n$. We simulate samples from the joint distribution for $\theta_2' = \left(\beta_1,\beta_2,\ldots,\beta_7\right)$ using the Metropolis-Hastings algorithm.

In the situation where we do not have the presence of covariates, we have a vector of parameters given by $\theta_1' = (\upsilon_1,\upsilon_2,\ldots,\upsilon_7)$, where $\upsilon_1 = p$, $\upsilon_2 = S_E$, $\upsilon_3 = S_P$, $\upsilon_4 = \lambda_{11}$, $\upsilon_5 = \lambda_{01}$, $\upsilon_6 = \lambda_{10}$ and $\upsilon_7 = \lambda_{00}$. Assuming prior independence among the parameters, let us consider $Beta\left(a_{1j},b_{1j}\right)$ prior distributions for υ_j, $j = 1,\ldots,7$, with a_{1j} and b_{1j} known hyperparameters. Also considering the introduction of the latent variables $\mathbf{D}_{i1}^* = (D_{i1}^*,\overline{D}_{i1}^*)$, $i = 1,\ldots,n$, the joint posterior distribution for θ_1 is given by

$$\begin{aligned}
&\Pi\left(\theta_1|\mathcal{D},\mathbf{t},\mathbf{d}^*\right) \propto p^{a_{11}+T_1-1}(1-p)^{b_{11}+T_2-1} S_E^{a_{12}+T_3-1}(1-S_E)^{b_{12}+T_4-1} \times \\
&\times S_P^{a_{13}+T_5-1}(1-S_P)^{b_{13}+T_6-1}\lambda_{01}^{a_{14}+T_7-1}(1-\lambda_{01})^{b_{14}+T_8-1}\lambda_{10}^{a_{15}+T_9-1} \times \\
&\times(1-\lambda_{10})^{b_{15}+T_{10}-1}\lambda_{00}^{a_{16}+T_{11}-1}(1-\lambda_{00})^{b_{16}+T_{12}-1} \times \\
&\times\lambda_{11}^{a_{17}+T_{13}-1}(1-\lambda_{11})^{b_{17}+T_{14}-1}, \qquad (10)
\end{aligned}$$

where $T_1 = \sum_{i=1}^n d_i\upsilon_i + \sum_{j=1}^2\sum_{i=1}^n(1-\upsilon_i)d_{ij}^*$, $T_2 = \sum_{i=1}^n\left(1-d_i\right)\upsilon_i + \sum_{j=1}^2\sum_{i=1}^n(1-\upsilon_i)\left(1-d_{ij}^*\right)$, $T_3 = \sum_{i=1}^n d_i t_i\upsilon_i + \sum_{i=1}^n t_i(1-\upsilon_i)d_{i1}^*$, $T_4 = \sum_{i=1}^n d_i\left(1-t_i\right) + \sum_{i=1}^n(1-t_i)(1-\upsilon_i)d_{i2}^*$, $T_5 = \sum_{i=1}^n\left(1-d_i\right)\left(1-t_i\right)\upsilon_i + \sum_{i=1}^n\left(1-t_i\right)\left(1-\upsilon_i\right)\left(1-d_{i2}^*\right)$, $T_6 = \sum_{i=1}^n\left(1-d_i\right)t_i\upsilon_i + \sum_{i=1}^n t_i(1-\upsilon_i)\times \times\left(1-d_{i1}^*\right)$, $T_7 = \sum_{i=1}^n\left(1-d_i\right)t_i\upsilon_i$, $T_8 = \sum_{i=1}^n t_i(1-\upsilon_i)\left(1-d_{i1}^*\right)$, $T_9 = \sum_{i=1}^n d_i\left(1-t_i\right)\upsilon_i$, $T_{10} = \sum_{i=1}^n(1-t_i)(1-\upsilon_i)d_{i2}^*$, $T_{11} = \sum_{i=1}^n\left(1-d_i\right) \times \times\left(1-t_i\right)\upsilon_i$, $T_{12} = \sum_{i=1}^n\left(1-t_i\right)\left(1-\upsilon_i\right)\left(1-d_{i2}^*\right)$, $T_{13} = \sum_{i=1}^n d_i t_i\upsilon_i$, and $T_{14} = \sum_{i=1}^n t_i\left(1-\upsilon_i\right)d_{i1}^*$. Considering that, for instance, $\theta_{(p)}$ is the vector θ_1 without the parameter p, the conditional distributions for the Gibbs sampling algorithm are given by

$$\begin{aligned}
p|\theta_{(p)},\mathcal{D},\mathbf{d}^* &\sim Beta\left(a_{11}+T_1,b_{11}+T_2\right),\\
S_E|\theta_{(S_E)},\mathcal{D},\mathbf{d}^* &\sim Beta\left(a_{12}+T_3,b_{12}+T_4\right),\\
S_P|\theta_{(S_P)},\mathcal{D},\mathbf{d}^* &\sim Beta\left(a_{13}+T_5,b_{13}+T_6\right),\\
\lambda_{01}|\theta_{(\lambda_{01})},\mathcal{D},\mathbf{d}^* &\sim Beta\left(a_{14}+T_7,b_{14}+T_8\right),\\
\lambda_{10}|\theta_{(\lambda_{10})},\mathcal{D},\mathbf{d}^* &\sim Beta\left(a_{15}+T_9,b_{15}+T_{10}\right),\\
\lambda_{00}|\theta_{(\lambda_{00})},\mathcal{D},\mathbf{d}^* &\sim Beta\left(a_{16}+T_{11},b_{16}+T_{12}\right) \text{ and}\\
\lambda_{11}|\theta_{(\lambda_{11})},\mathcal{D},\mathbf{d}^* &\sim Beta\left(a_{17}+T_{13},b_{17}+T_{14}\right). \qquad (11)
\end{aligned}$$

5 An example

Aiming to get estimates for the measures of the performance of cervical cytology (Papanicolaou test) and hybrid capture II (HC-II) in detecting neoplastic and pre-neoplastic cervical lesions, without a gold standard, a study was conducted using a sample of 807 women who visited two different public health units in Campinas, Brazil. Table 2 displays how these test results were distributed among study participants. Considering that smoking is a known risk factor for the disease, Table 2 shows also the frequency distribution of test outcomes by smoking status.

tests			smoking status (*)		
cervical cytology	hybrid capture II	total	1	2	3
negative	negative	653	408	126	119
positive	negative	21	13	5	3
negative	positive	102	68	24	10
positive	positive	31	17	11	3

(*) 1: was never a smoker; 2: current smoker; 3: smoker in the past

Table 2: Frequency distribution of tests findings by smoking status.

Using the proposed methodology, we estimated the sensitivity and specificity measures of cervical cytology and HC-II, without a reference test (gold standard). In this model, X is a covariate related to smoking status, coded as “was never a smoker”, “current smoker” or “was a smoker in the past”. To to introduce this variable in the model, we use a set of dummy variables X_1 and X_2, where X_1 and X_2 are both equal to zero when a woman was never a smoker; X_1 is equal to 1 and X_2 is equal to 0 when a woman is a current smoker; and when a woman was a smoker in the past, we used $X_1 = 0$ and $X_2 = 1$. From the respective conditional posterior distributions we generated a chain of 100,000 iterations, and in order to diminish some effect of the initial parameters values, we discarted the first 20,000 elements of each chain. For each parameter we considered every 50^{th} draw. The results for a Bayesian analysis without the covariate.are provided in Table 3. The choice of hyperparameter values for the prior distributions of $S_{E_1}, S_{E_2}, S_{P_1}$ and S_{P_2} were based in results introduced in articles from the medical literature. The hyperparameter values for the prior distributions of p were subjectively choosen, and a small sensitivity analysis was made by choosing other hyperparameter values, but their choice do not modify substantially the results presented below and the correspondent results are omitted here.

		posterior summaries		95% credible	
parameter	prior distribution	mean	SD	interval	
S_{E_1}	$Beta(27.46, 26.38)$	0.496	0.0689	0.363	0.630
S_{P_1}	$Beta(7.53, 0.15)$	0.987	0.0121	0.963	0.999
S_{E_2}	$Beta(10.55, 1.86)$	0.758	0.1183	0.557	0.969
S_{P_2}	$Beta(9.48, 1.42)$	0.904	0.0199	0.868	0.947
p	$Beta(1, 1)$	0.111	0.0329	0.056	0.180

Table 3: Posterior summaries (S.D.: standard deviation).

parameter	mean	SD	95% credible interval	
β_{10}	−0.01996	0.41418	−0.76299	0.82661
β_{20}	1.64439	0.46609	0.76556	2.58763
β_{30}	3.88043	0.32590	3.31424	4.61192
β_{40}	2.15986	0.22578	1.77946	2.66919
β_{50}	−2.37024	0.27468	−2.88650	−1.83678
β_{11}	0.19941	0.44036	−0.64108	1.09630
β_{21}	0.05828	0.45221	−0.79778	1.01135
β_{31}	−0.02367	0.43427	−0.88744	0.87934
β_{41}	−0.01321	0.33745	−0.62645	0.73420
β_{51}	0.38630	0.34090	−0.30118	1.01960
β_{12}	0.01599	0.47539	−0.91589	0.94134
β_{22}	−0.05165	0.50753	−1.04197	0.93975
β_{32}	0.08058	0.42796	−0.74023	0.93379
β_{42}	0.47367	0.34229	−0.16628	1.16486
β_{52}	−0.33981	0.36796	−1.05416	0.36621

Table 4: Posterior summaries (S.D.: standard deviation).

Table 4 shows the posterior summaries for the parameters in the Bayesian model that considers the smoking status as a covariate. The parameters β_{11} to β_{51} are related to category “current smoker” versus “was never a smoker”, and the parameters β_{12} to β_{52} are associated to category “was a smoker in the past” versus “was never a smoker”. We note that the 95% credible intervals for β_{11} to β_{51} and β_{12} to β_{52} included the value 0, suggesting that the effect of smoking status on the S_{E_1}, S_{E_2}, S_{P_1}, S_{P_2} and p measures is not important. Posterior mean estimates of S_{E_1}, S_{E_2}, S_{P_1}, S_{P_2} and p, calculated from simulated values for the vector θ_2, are given by: (1) (was never a smoker)

$\widehat{S_{E_1}} = 0.495$, $\widehat{S_{E_2}} = 0.979$, $\widehat{S_{P_1}} = 0.828$, $\widehat{S_{P_2}} = 0.895$ and $\widehat{p} = 0.088$; (2) (current smoker) $\widehat{S_{E_1}} = 0.540$, $\widehat{S_{E_2}} = 0.977$, $\widehat{S_{P_1}} = 0.829$, $\widehat{S_{P_2}} = 0.890$ and $\widehat{p} = 0.127$; and (3) (was a smoker in the past) $\widehat{S_{E_1}} = 0.499$, $\widehat{S_{E_2}} = 0.979$, $\widehat{S_{P_1}} = 0.811$, $\widehat{S_{P_2}} = 0.929$ and $\widehat{p} = 0.067$. These results suggest again that the smoking status is not related to prevalence of cervical lesions and the performance measures of the tests. We used the software SAS (proc IML) to perform all the simulations. The convergence of the Gibbs samples was monitored by standard existing methods [6] available in CODA package [3].

References

[1] Begg C.B. (1987). *Biases in the assessment of diagnostic tests.* Statistics in Medicine **6**, 411 − 423.

[2] Begg C.B., Greenes R.A. (1983). *Assessment of diagnostic tests when disease verification is subject to selection bias.* Biometrics **39**, 207 − 215.

[3] Best N.G., Cowles M.K., Vines S.K. (1995). *CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, version 0.3.* MRC Biostatistics Unit, Cambridge.

[4] Gelfand A.E., Smith A.F.M. (1990). *Sampling based approaches to calculating marginal densities.* Journal of the American Statistical Association **85**, 398 − 409.

[5] Geng Z., Asano C. (1989). *Bayesian estimation methods for categorical data with misclassification.* Communications in Statistics **8**, 2935 − 2954.

[6] Geweke J. (1992). *Evaluating the accuracy of sampling-based approaches to calculating posterior moments.* In: Bayesian Statistics 4, Bernardo J.M, Berger J.O., Dawid A.P. and Smith A.F.M. (eds), Clarendom Press: Oxford, 169 − 194.

[7] Joseph L., Gyorkos T.W., Coupal L. (1995). *Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard.* American Journal of Epidemiology **141**, 263 − 272.

[8] Soares P., Paulino C.D. (2001). *Incomplete categorical data analysis: A Bayesian perspective.* Journal of Statistical Computation and Simulation **69**, 157 − 170.

[9] Zhou X. (1993). *Maximum likelihood estimators of sensitivity and specificity corrected for verification bias.* Communications in Statistics Theory and Methods **22**, 3177 − 3198.

Address: J.A. Achcar, E.Z. Martinez, Departamento de Medicina Social, FMRP USP, Universidade de São Paulo, Ribeirão Preto, SP, Brazil
F. Louzada-Neto, Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, SP, Brazil

E-mail: `achcar@fmrp.usp.br`

COMPSTAT'2004 Symposium

MULTIPLE CORRESPONDENCE SPLINE ANALYSIS FOR GRAPHICALLY REPRESENTING NONLINEAR RELATIONSHIPS BETWEEN VARIABLES

Koichi Adachi

Key words: Multiple correspondence analysis, principal component analysis inter-variable nonlinear relationships, trajectories of variables, smoothing splines.

COMPSTAT 2004 section: Data visualization.

Abstract: To deal with difficulties which may arise when multiple correspondence analysis (MCA) or principal component analysis (PCA) is used for finding inter-variable nonlinear relationships, we propose a variant of MCA. In this method, variables are represented as trajectories in a low-dimensional space, where the trajectories are defined using smoothing splines. In a simulation study, the proposed method was found to recover true trajectories better than MCA and PCA.

1 Introduction

We often encounter data including variables related nonlinearly to each other. Such a data set is illustrated in Table 1(A). There, the first variable, though correlated linearly to the second, has nonlinear relation to the third variable, showing that cakes of medium sweet are preferred, but less sweet or too sweet cakes are not preferred. To represent such nonlinear relationships graphically in a low-dimensional space, we propose a variant of MCA (multiple correspondence analysis) in this paper. This proposal is motivated by noting difficulties which may arise when MCA or PCA (principal component analysis) is used for exploring inter-variable nonlinear relations. To describe this motivation, we begin with formulating PCA and MCA within a common framework.

Let $\mathbf{X} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_m]$ denote an n-objects by m-variables data matrix and the mutually different K_j values in $\tilde{\mathbf{x}}_j$ be collected into $\mathbf{y}_j{}' = [y_{j1}, \ldots, y_{jK_j}]$: for example, $\mathbf{y}_1{}' = [10, 25, 40]$ for $\tilde{\mathbf{x}}_1$ in Table 1(A). Let $\mathbf{G}_j = (g_{ijk})(n \times K_j)$ be the indicator matrix transformed from $\tilde{\mathbf{x}}_j$ as illustrated in Table 1(B): g_{ijk} (the ikth element of $\mathbf{G}_j$) takes one if $x_{ij} = y_{jk}$ and zero otherwise, with x_{ij} the ith element of $\tilde{\mathbf{x}_j}$. We further let $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ be the $n \times n$ centering matrix with $\mathbf{I}_n$ the $n \times n$ identity matrix and $\mathbf{1}_n$ the $n \times 1$ vector of ones.

According to Gifi [3], PCA for centered data matrix $\mathbf{JX}$ is formulated as minimizing

(A) Data matrix $\mathbf{X}$

Cake	Sales $\tilde{\mathbf{x}}_1$	Name value $\tilde{\mathbf{x}}_2$	Sweet $\tilde{\mathbf{x}}_3$	...
1	10	1	5	...
2	10	1	1	...
3	25	2	1	...
4	40	3	3	...
5	40	4	3	...

(B) Matrix $\mathbf{G}$ transformed from $\mathbf{X}$

	Sales $\mathbf{G}_1$			Name value $\mathbf{G}_2$				Sweet $\mathbf{G}_3$		...
10	25	40	1	2	3	4	1	3	5	
1			1						1	...
1				1			1			...
	1				1				1	...
		1			1			1		...
		1				1		1		...

$$LP(\mathbf{F}, \mathbf{A}) = \sum_{j=1}^{m} \left\| \mathbf{F} - \mathbf{J}\tilde{\mathbf{x}}_j \mathbf{a}_j' \right\|^2 \tag{1}$$

over $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]'$($n$-objects by p-dimensions) and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]'$($m \times p$) under the following normalization conditions.

$$\mathbf{F} = \mathbf{JF}. \tag{2}$$

$$n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_p. \tag{3}$$

In MCA, a value in a data set, i.e., y_{jk} is treated not as a quantity but just as a nominal category to be given score vector, which we express as $\mathbf{w}_{jk}(p \times 1)$. To obtain score vectors $\mathbf{W}_j = [\mathbf{w}_j 1, \ldots, \mathbf{w}_{jkj}]'$($K_j$-categories by p-dimensions),

$$LM(\mathbf{F}, \mathbf{W}) = \sum_{j=1}^{m} \|\mathbf{F} - \mathbf{G}_j \mathbf{W}_j\|^2 \tag{4}$$

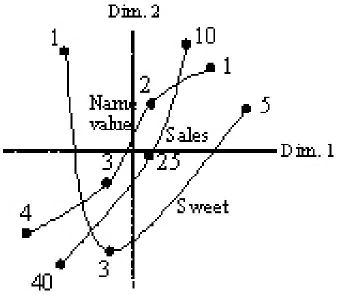
is minimized over $\mathbf{W} = [\mathbf{W}_1{}', \ldots, \mathbf{W}_m{}']'(\sum_{j=1}^{m} K_j \times p)$ and $\mathbf{F}(n \times p)$ subject to (2) and (3) [3]. Loss function (4) is derived from the following principle.

> *Homogeneity principle.* Let k_{ij}^* denote the category associated with object i for variable j, with $x_{ij} = y_{j,k_{ij}^*}$ (e.g., $k_{31}^* = 2$ and $x_{31} = y_{12} = 25$ in Table 1). Then, $\mathbf{w}_{j,k_{ij}^*}$ (the score vector of the category for object i) should be close (or homogeneous) to $\mathbf{f}_i$ (the vector of object i).

The departure from this principle can be expressed as $\sum_i \sum_j \|\mathbf{f}_i - \mathbf{w}_{k,k^*_{ij}}\|^2$, which equals (4).

Noticing that $\mathbf{D}_j = \mathbf{G}'_j\mathbf{G}_j$ is the diagonal matrix having category frequencies on the diagonal, $\mathbf{D}_j\mathbf{1}_{K_j} = \mathbf{G}'_j\mathbf{1}_n, \mathbf{G}_j\mathbf{1}_{K_j} = \mathbf{1}_n$, and $\tilde{\mathbf{x}}_j = \mathbf{G}_j\mathbf{y}_j$, we find $\mathbf{J}\mathbf{G}_j = \mathbf{G}_j\mathbf{E}_j$ and $\mathbf{J}\tilde{\mathbf{x}}_j = \mathbf{J}\mathbf{G}_j\mathbf{y}_j = \mathbf{G}_j\mathbf{E}_j\mathbf{y}_j$ with $\mathbf{E}_j = \mathbf{I}_{K_j} - n^{-1}\mathbf{1}_{K_j}\mathbf{1}'_{K_j}\mathbf{D}_j$. Thus (1) is rewritten as

$$LP(\mathbf{F}, \mathbf{A}) = \sum_{j=1}^{m} \|\mathbf{F} - \mathbf{G}_j\mathbf{E}_j\mathbf{y}_j\mathbf{a}'_j\|^2. \tag{5}$$



Comparing this with (4), we can find that PCA is a constrained version of MCA with $\mathbf{W}_j = \mathbf{E}_j\mathbf{y}_j\mathbf{a}'_j$.

For the data in Table 1, MCA might yield a result illustrated in Figure 1. There, a trajectory associated with variable j connects the score vectors $\mathbf{w}_{jk}$ and $\mathbf{w}_{j,k+1}$ which are represented as points labeled with the corresponding y_{jk} and $y_{j,k+1}$. This configuration allows us easily to grasp inter-variable correlations: nonlinear relations of variable "sweet" to the others are found by the curved trajectory for "sweet". Such relations cannot be found in results of PCA, since $\mathbf{W}_j$ is constrained as $\mathbf{W}_j = \mathbf{E}_j\mathbf{y}_j\mathbf{a}'_j$, i.e., $\mathbf{w}_{jk} = z_{jk}\mathbf{a}_j$ with z_{jk} the kth element of $\mathbf{z}_j = \mathbf{E}_j\mathbf{y}_j$, and PCA yields straight trajectories extending in the direction of vector $\mathbf{a}_j$. Nishisato [5] has illustrated the above advantage of MCA with a numerical example. This advantage owes to that y_{jk} is regarded just as a category "k" to be given score vector $\mathbf{w}_{jk}$.

However, MCA may not work well if K_j is large, since the number of unknown parameters $\mathbf{w}_{jk}$ increases so that the resulting $\mathbf{w}_{jk}$ may be unstable and give too zigzag trajectories. As a remedy for this difficulty, Nishisato [5] has suggested classifying $y_{j1}, \ldots, y_{jK_j}$ into a fewer categories. However, how to classify them has not been clarified. In this paper, we consider another method, in which trajectories are required to be smooth. This requirement is attained by defining a loss function as

$$L(\mathbf{F}, \mathbf{W}) = LM(\mathbf{F}, \mathbf{W}) + s \times LS(\mathbf{W}), \tag{6}$$

i.e., by combing (4) and $LS(\mathbf{W})$ expressing loss of the smoothness of trajectories, with s a positive constant.

Function $LS(\mathbf{W})$ and minimization of (6) are detailed in section 2. A simulation study is described in section 3 and the method is compared with related techniques in the final section.

2 Proposed method

We regard the trajectory for variable j (illustrated in Figure 1) as a function of continuous variable Y_j and express this function as $\mathbf{w}_j(Y_j)$ with its lth element $w_{jl}(Y_j)$. According to this notation, score vector $\mathbf{w}_{jk}$ is expressed as $\mathbf{w}_{jk} = \mathbf{w}_j(y_{jk})$, that is, $\mathbf{w}_j(y_{jk})$ is the point at $Y_j = y_{jk}$ on trajectory $\mathbf{w}_j(Y_j)$. We require $w_{jl}(Y_j)$ to change smoothly with Y_j. The departure from the smoothness can be expressed as the integral of the squared second derivative of $w_{jl}(Y_j)$:

$$LS_{jl} = \int \left(\frac{d^2 w_{jl}(Y_j)}{dy^2} \right)^2 dY_j. \tag{7}$$

We define $LS(\mathbf{W})$ in (6) as the sum of (7) over variables j and dimensions l.

It is known that (7) takes a lower value when $w_{jl}(Y_j)$ is the smoothing spline (natural cubic spline) function of Y_j with knots $y_{j1}, \ldots, y_{jK_j}$, than when $w_{jl}(Y_j)$ is any other functional [6]. We thus let $w_{jl}(Y_j)$ be the smoothing spline. Then, integral (7) can be rewritten in the form $LS_{jl} = \tilde{\mathbf{w}}'_{jl}\mathbf{Q}_j\mathbf{R}_j^{-1}\mathbf{Q}'_j\tilde{\mathbf{w}}_{jl}$ (e.g., Green and Silverman [4]). Here, $\tilde{\mathbf{w}}_{jl}$ is the l-th column of $\mathbf{W}_j$, and $\mathbf{Q}_j$ and $\mathbf{R}_j$ are $K_j \times (K_j - 2)$ and $(K_j - 2) \times (K_j - 2)$ tridiagonal matrices, respectively, defined using $h_{jk} = y_{j,k+1} - y_{jk}$ as follows. With $q_{jkk'}$ the kk'th elements of $\mathbf{Q}_j$, its non-zero elements are expressed as $q_{jkk} = h_{jk}^{-1}, q_{j,k+1,k} = -h_{jk}^{-1} - h_{j,k+1}^{-1}$, and $q_{j,k+2,k} = -h_{j,k+1}^{-1}$ for $k = 1, \ldots, K_j - 2$. With $r_{jkk'}$ the kk'th elements of $\mathbf{R}_j$, its non-zero elements are defined as $r_{jkk} = 3^{-1}(h_{jk} + h_{j,k+1})$ for $k = 1, \ldots, K_j$ and $r_{j,k-1,k} = r_{j,k,k-1} = 6^{-1}h_{jk}$ for $k = 2, \ldots, K_j - 2$. The sum of (7) is thus written as

$$LS(\mathbf{W}) = \sum_{j=1}^{m}\sum_{l=1}^{p} LS_{jl} = \sum_{j=1}^{m} \mathrm{tr}\, \mathbf{W}'_j\mathbf{Q}_j\mathbf{R}_j^{-1}\mathbf{Q}'_j\mathbf{W}_j. \tag{8}$$

Substituting (7) and (8) into (6), we have

$$L(\mathbf{F}, \mathbf{W}) = \sum_{j=1}^{m} \left(\|\mathbf{F} - \mathbf{G}_j\mathbf{W}_j\|^2 + s \times \mathrm{tr}\, \mathbf{W}'_j\mathbf{Q}_j\mathbf{R}_j^{-1}\mathbf{Q}'_j\mathbf{W}_j \right). \tag{9}$$

This is minimized over $\mathbf{W}$ and $\mathbf{F}$ subject to (2) and (3), for given p and $s > 0$. We refer to this method as multiple correspondence spline analysis (MCSA) in this paper.

Loss function (9) is rewritten as

$$L(\mathbf{F}, \mathbf{W}) = nmp = \mathrm{tr} \sum_{j=1}^{m} \left(\mathrm{tr}\, \mathbf{W}'_j\mathbf{C}_j\mathbf{W}_j - 2\mathrm{tr}\, \mathbf{F}'\mathbf{J}\mathbf{G}_j\mathbf{W}_j \right). \tag{10}$$

Here, (2) and (3) are used and $\mathbf{C}_j = \mathbf{D}_j + s\mathbf{Q}_j\mathbf{R}_j^{-1}\mathbf{Q}_j'$. We supposed $\mathbf{D}_j$ to be positive definite. Then, $\mathbf{C}_j$ is also positive definite, since $\mathbf{Q}_j\mathbf{R}_j^{-1}\mathbf{Q}_j'$ is nonnegative definite and $s > 0$. Solving $\partial L(\mathbf{FW})/\partial \mathbf{W}_j = \mathbf{0}$ (with $\mathbf{0}$ denoting a null matrix or vector of an appropriate size), we find that the optimal $\mathbf{W}_j$ satisfies

$$\hat{\mathbf{W}}_j = \mathbf{C}_j^{-1}\mathbf{G}_j'\mathbf{J}\mathbf{F}. \tag{11}$$

The substitution of (11) into $\mathbf{W}_j$ in (10) yields

$$L(\mathbf{F}, *) = nmp - \operatorname{tr} \mathbf{F}'\mathbf{JGC}^{-1}\mathbf{G}'\mathbf{JF}$$

with $\mathbf{G} = [\mathbf{G}_1, \ldots, \mathbf{G}_m]$ $(n \times \sum_{j=1}^m K_j)$ and $\mathbf{C}$ the $(\sum_{j=1}^m K_j) \times (\sum_{j=1}^m K_j)$ block diagonal matrix whose jth diagonal block is $\mathbf{C}_j$. Minimizing $L(\mathbf{F}, *)$ subject to (2) and (3) is attained using the SVD (singular value decomposition) $\mathbf{JGC}^{-\frac{1}{2}} = \mathbf{NTM}'$. Here, $\mathbf{N}'\mathbf{N} = \mathbf{M}'\mathbf{M} = \mathbf{I}_g$ with g the rank of $\mathbf{JGC}^{-\frac{1}{2}}$ and $\mathbf{T}$ is the diagonal matrix whose lth diagonal element is the lth largest singular value of $\mathbf{JGC}^{-\frac{1}{2}}$. Let $\mathbf{T}_p$ be the first $p \times p$ diagonal block of $\mathbf{T}$, and $\mathbf{N}_p$ and $\mathbf{M}_p$ contain the first p columns of $\mathbf{N}$ and $\mathbf{M}$, respectively, with $p \le g$. The optimal $\mathbf{F}$ is then given by $\hat{\mathbf{F}} = n^{\frac{1}{2}}\mathbf{N}_p$. Substituting this and the above SVD into (11), we find the optimal $\mathbf{W}$ is given by $\hat{\mathbf{W}} = n^{\frac{1}{2}}\mathbf{C}^{-\frac{1}{2}}\mathbf{M}_p\mathbf{T}_p$. That is, the solution is given explicitly using SVD.

However, we may encounter cases where the size of $\mathbf{JGC}^{-\frac{1}{2}}$ to be decomposed, i.e., n or $\sum_{j=1}^m K_j$, is so large that we cannot obtain solution easily. We thus use another algorithm without the decomposition of large-sized matrices. This is an alternating least squares algorithm, in which the following two steps are iterated alternately until $L(\mathbf{F}, \mathbf{W})$ is judged to converge to its minimum.

Step 1. Given $\mathbf{F}, L(\mathbf{F}, \mathbf{W})$ is minimized over $\mathbf{W} = [\mathbf{W}_1', \ldots, \mathbf{W}_m']'$. The optimal $\mathbf{W}$ is obtained with (11).

Step 2. Given $\mathbf{W}, L(\mathbf{F}, \mathbf{W})$ is minimized over $\mathbf{F}$ subject to (2) and (3). The optimal $\mathbf{F}$ is given by $\hat{\mathbf{F}} = n^{\frac{1}{2}}\mathbf{KL}'$ with $\mathbf{JGW} = \mathbf{KSL}'$ the SVD of $\mathbf{JGW}(n \times p)$. $\mathbf{K}$ and $\mathbf{L}$ can be obtained with the eigenvalue-decomposition of $\mathbf{V} = \mathbf{W}'\mathbf{G}'\mathbf{JGW}$. This size is of $p \times p$, and p is usually far smaller than n or $\sum_{j=1}^m K_j$.

3 Simulation

To evaluate the proposed method MCSA, we performed a small simulation study, in which artificial data synthesized from true scores were used for assessing how well they are recovered by the method.

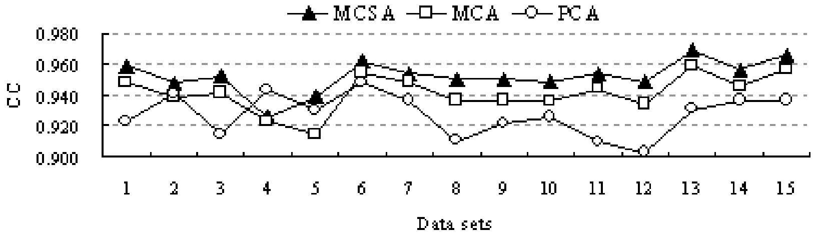
Setting $p = 2,\ n = 200,\ m = 10,\ K_j = 10$, and $y_{jk} = k$ for $j = 1, \ldots, m$, we had different fifteen sets of true scores and data matrices, i.e., 15 sets of $\{\mathbf{F}, \mathbf{W}$ and $\mathbf{X}$ (or $\mathbf{G}$)$\}$. Each set was generated through the following three stages.

Stage 1. $\mathbf{f}_i$ $(i = 1, \ldots, n)$ were sampled independently from the bivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}_2$. This sampling allows the resulting $\mathbf{F}$ approximately to satisfy (2) and (3), but not exactly.

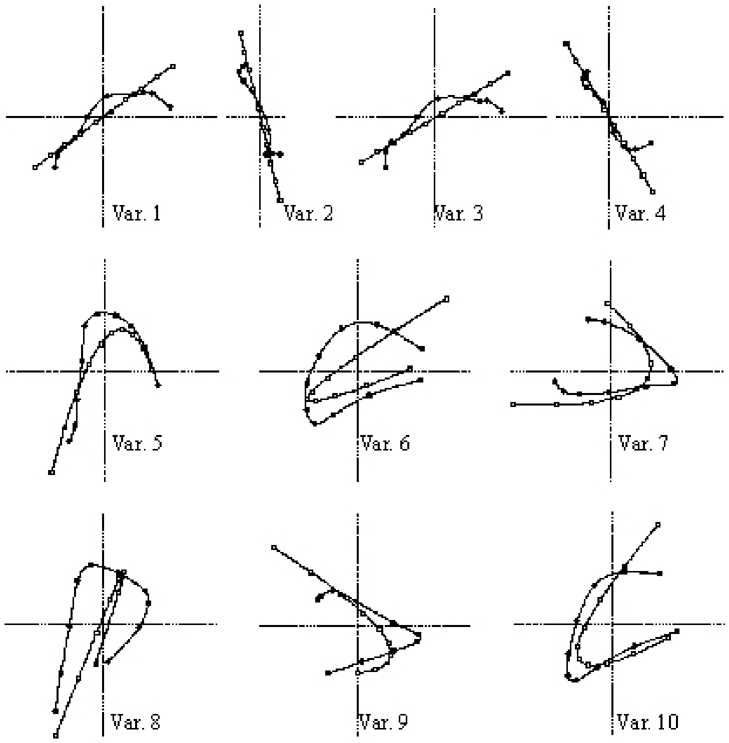
Stage 2. To obtain true trajectories $\mathbf{w}_j(Y_j)$ which are linear for $j \leq 4$ and quadratic for $j \geq 5$ (see Figure 3), we first generated *raw* trajectories $\mathbf{w}^*_j(Y_j)$ and then transformed them into $\mathbf{w}_j(Y_j)$ which yield true score vectors $\mathbf{w}_{jk} = \mathbf{w}_j(y_{jk})$. Using orthogonal polynomials, the lth element of $\mathbf{w}^*_j(Y_j)$ was chosen as $w^*_{jl} = b_{0l}u_{1j}(Y_j)$ for $j \leq 4$ and $w^*_{jl}(Y_j) = b_{1l}u_{1j}(Y_j) + b_{2l}u_{2j}(Y_j)$ for $j \geq 5$, where $u_{1j}(Y_j) = 1-2Y_j/(K_j-1), u_{2j}(Y_j) = 1-6Y_j/(K_j-1)+6Y_j(Y_j-1)/[(K_j-1)(K_j-2)]$, and b_{0l}, b_{1l} and b_{2l} were independently sampled from the uniform distribution ranging from -1 to 1. Then, $w^*_{jl}(Y_j)$ was linearly transformed into $w_{jl}(Y_j)$ so that the average and the variance of $w_{jl}(y_{jk})$ over $k = 1, \ldots, K_j$ became zero and unit, respectively. This standardization allows the distribution of score vectors to be similar to that of $\mathbf{f}_i$.

Stage 3. We generated data $\mathbf{X} = (x_{ij})$ in such a manner that they are consistent to the homogeneity principle which implies that k^*_{ij} (a category to be associated with object i) should be the category whose score vector is close to $\mathbf{f}_i$. That is, we set $k^*_{ij} = \arg\min_{1 \leq k \leq K_j} \|\mathbf{f}_i - \mathbf{w}_{jk}\|$ to generate $x_{ij} = y_{j,k_ij^*}$. This is simplified as $x_{ij} = \arg\min_k \|\mathbf{f}_i - \mathbf{w}_{jk}\|$, since of $y_{jk} = k$, in this simulation.

Each of the $15\mathbf{X}'s$ or the corresponding $\mathbf{G}'s$ is analyzed by MCSA, MCA and PCA. In order to assess the recovery of score vectors, we obtained a congruence coefficient (CC). This is defined as the cosine between $100 \times 99/2$ dimensional vectors $\mathbf{d}$ and $\hat{\mathbf{d}}$, where $\mathbf{d}$ contains the distances between the 100(= 10 variables $\times$10 categories) true $\mathbf{w}_{jk}$'s and $\hat{\mathbf{d}}$ contains the corresponding distances for the resulting $\hat{\mathbf{w}}_{jk}$. At the highest CC takes one which shows complete recovery. The results are shown in Figure 2. There, the CC's for MCSA are found higher than those of PCA and MCA for 14 data sets (excluding the fourth set), which shows the superiority of MCSA in recovering score vectors.



To illustrate the recovery of true trajectories by MCSA, the result for the third data set giving the median CC = 0.952 over CC's for 15 sets, is shown in Figure 3. There, the pair of the true and resulting trajectories for each variable is depicted separately in each panel, for ease of grasping recovery. Though the resulting trajectories deviate somewhat from true ones, the recovery is thought satisfactory in that we can find true inter-variable relationships from the result fairly well.



4 Final remarks

In order to represent inter-variable relations graphically in a low-dimensional configuration, we proposed a variant of MCA named MCSA. In MCSA, variables are represented as trajectories defined using smoothing splines and the objective function to be minimized is formed by combining the loss function for MCA with the loss of the smoothness of trajectories.

A fixed effect curvilinear model (FECM) proposed by Besse and Ferraty [2] is related to MCSA, in that FECM is an extension of the nonmetric PCA that is a variant of MCA [3] and smoothing splines are also used in FECM. According to our notation, the loss function for FECM may be written as

$$LF(\mathbf{F}, \mathbf{A}, \mathbf{Y}) = \sum_{j=1}^{m} \|\mathbf{G}_j \mathbf{E}_j \mathbf{y}_j - \mathbf{F}\mathbf{a}_j\|^2 + \sum_{j=1}^{m} s_j \mathbf{y}_j' \mathbf{Q}_j \mathbf{R}_j^{-1} \mathbf{Q}_j \mathbf{y}_j, \qquad (12)$$

which is minimized over $\mathbf{F}, \mathbf{A}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]$ under normalization conditions (2), (3) and $\mathbf{y}_j \mathbf{E}_j' \mathbf{D}_j \mathbf{E}_j \mathbf{y}_j = n$, for given $s_j > 0$. Note that $\mathbf{y}_j$ is unknown in (12). Since the minimization of $\|\mathbf{G}_j \mathbf{E}_j \mathbf{y}_j - \mathbf{F}\mathbf{a}_j\|^2$ is found equivalent to that of $\|\mathbf{F} - \mathbf{G}_j \mathbf{E}_j \mathbf{y}_j \mathbf{a}_j'\|^2$ under the normalization conditions, FECM can also be formulated as minimizing

$$LF(\mathbf{F},\mathbf{A},\mathbf{Y}) = \sum_{j=1}^{m} \|\mathbf{F} - \mathbf{G}_j\mathbf{E}_j\mathbf{y}_j\mathbf{a}_j'\|^2 + \sum_{j=1}^{m} s_j\mathbf{y}_j'\mathbf{Q}_j\mathbf{R}_j^{-1}\mathbf{Q}_j\mathbf{y}_j. \qquad (13)$$

The first term in the right side of (13) is identical to (5) and can thus be viewed as a constrained version of (4) with $\mathbf{W}_j = \mathbf{E}_j\mathbf{y}_j\mathbf{a}_j'$. It shows that, differently than our method, FECM yields linear trajectories of variables extending in the direction of $\mathbf{a}_j$. However, $\mathbf{y}_j$ is unknown in FECM, which differs from PCA.

A variant of MCA with smoothing splines has also been presented by Adachi [1]. However, in this method splines are used for the smoothing on **F** (the scores of objects), while they are used for **W** (variables) in our MCSA.

In the simulation study, we chose one as the value of s, i.e., as the weight for the loss of smoothness. However, another choice might yield a better result. It thus remains for future study to consider a method for choosing the optimal value of s.

References

[1] Adachi K. (2002). *Optimal quantification of a longitudinal indicator matrix: Homogeneity and smoothness analysis.* Journal of Classification 19, 215 – 248.

[2] Besse P. C. and Ferraty F. (1995). *A fixed effect curvilinear model.* Computational Statistics 10, 339 – 351.

[3] Gifi A. (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

[4] Green P. J. and Silverman B. W. (1994). *Nonparametric regression and generalized linear models: A roughness penalty Approach.* London: Chapman & Hall.

[5] Nishisato S. (2003). *Geometric perspectives of dual scaling for assessment of information in data.* In H. Yanai, A. Okada, K. Shigemasu, Y. Kano & J.J. Meulman (Eds.), New developments in psychometrics, 453 – 462. Tokyo: Springer.

[6] Schoenberg I. J. (1964). *Spline functions and the problem of graduation.* Proceedings of the National Academy of Sciences of the United States, 52, 947 – 950.

Acknowledgement: Thanks to my family for their generous support.

Address: K. Adachi, Department of Psychology, Ritsumeikan University, Tohjiin-Kitamachi 56-1, Kita-ku, Kyoto 603-8577, Japan

E-mail: `k-adachi@lt.ritsumei.ac.jp`

MODELLING SHORT TERM VARIABILITY INTERACTIONS IN ECG: QT VERSUS RR

Rute Almeida, Ana Paula Rocha, Paules E. Pueyo, Juan Pablo Martínez and Pablo Laguna

Key words: Parametric modelling, QT variability, HRV.

COMPSTAT 2004 section: Time series analysis.

Abstract: QT and RR series interactions were explored by a dynamic linear approach using AR and ARARX models with automatic orders selection. Validation with simulated data and application to real records are presented. An important QTV fraction was found to be not linearly driven by HRV.

1 Introduction

The electrocardiogram (ECG) analysis is extensively used as a diagnostic tool to provide information on the heart function. Each cardiac beat (Figure 1) is typically associated to a sequence of five principal waves denoted by P, Q, R, S and T, whose characteristics are clinically relevant. In particular, the time interval between the onset of the QRS complex and the T wave end, known as QT interval, is considered to express the duration of ventricular repolarization. Abnormal QT values have been associated with ventricular pro-arrythmicity and its beat-to-beat variations are, to some extent, driven by the autonomic nervous system through the RR interval (measured as the time interval between consecutive beats). However it has not been yet clearly quantified which fraction of QT variability (QTV) is effectively correlated with RR beat-by-beat variations (Heart Rate Variability - HRV).

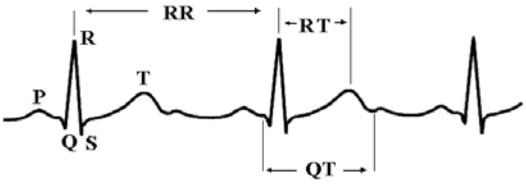


Figure 1: Schematic representation of relevant information in a cardiac beat.

The determination of RR and QT sequences requires the detection and delineation of ECG waves and limits. A wavelet transform based delineation system has proven to be quite robust against noise and morphological variations [3], even in the problematic T wave. Problems in delineation of T end lead to uncertainty in QTV measures which, allied to its smaller amplitude compared to HRV, represents the main difficulties in exploring this relation.

Many authors used alternative measures such as the RT interval (time between the peaks of the R and T waves); however, in spite of being easier to measure, the RT presents even shorter length than QT interval, additionally penalising the variability measures. A linear dynamic parametric approach was proposed by Porta et al [7] to express the interactions between the RR and RT intervals and allowing to quantify the fraction of the RT variability driven by RR. In previous work [1] we used a linear low order model similar to the one proposed by Porta to explore the short term HRV and QTV relations. A generalized and improved version of that model including automatic orders selection is now proposed and validated, defining an approach to quantify the fraction of QTV not driven by HRV.

2 Methods

2.1 Model formulation

Our approach, based on Porta [7], expresses RR and QT variability interactions in an open loop linear model (Figure 2) where A_{11}, A_{12}, A_{22} and D are polynomials in z^{-1} with coefficients $a_{11}[k]$, $a_{12}[k]$, $a_{22}[k]$ and $d[k]$, respectively. The series $W_{RR}[n]$ and $W_{QT}[n]$ are uncorrelated stationary zero-mean white noises with variances λ^2_{RR} and λ^2_{QT} and n denotes beat number.

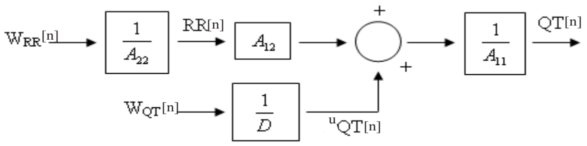


Figure 2: Schematic representation of the QTV versus HRV model.

$RR[n]$ series was modelled as an AR_p stationary random process given by

$$RR[n] = -\sum_{k=1}^{p} a_{22}[k]RR[n-k] + W_{RR}[n] \tag{1}$$

The QT was assumed to result from two uncorrelated sources, one driven by heart rate and other resulting of an exogenous input ($ARARX_q$ model [2])

$$\begin{aligned} QT[n] &= \sum_{k=0}^{q} a_{12}[k]RR[n-k] - \sum_{k=1}^{q} a_{11}[k]QT[n-k] +^u QT[n], \\ {}^uQT[n] &= -\sum_{k=1}^{q} d[k]\,{}^uQT[n-k] + W_{QT}[n] \end{aligned} \tag{2}$$

Therefore, the model accounts for the possible dependence on its past values and those of the RR interval (as shown in recent studies [8]). For simplicity, the same order q was assumed for all ARARX model polynomials, while a possible different order p is allowed for the AR model. This is a generalization from previous approaches [1], [7] where the same order was considered for all polynomials in the model. In fact there is no reason to constrain the QT and RR sequences to the same memory of its own past.

The assumption of uncorrelated sources allows to compute the Power Spectral Density (PSD) of QT ($S_{QT}(f)$) as the sum of the partial spectra that express each one of the contributions

$$S_{QT/W_{RR}}(f) = \overline{RR}\lambda_{RR}^2 \left| \frac{A_{12}(z)}{A_{11}(z)A_{22}(z)} \right|^2_{z=\exp(j2\pi f\overline{RR})} \tag{3}$$

$$S_{QT/W_{QT}}(f) = \overline{RR}\lambda_{QT}^2 \left| \frac{1}{A_{11}(z)D(z)} \right|^2_{z=\exp(j2\pi f\overline{RR})} \tag{4}$$

where f is the frequency in Hz. As both $QT[n]$ and $RR[n]$ series are unevenly sampled the mean RR interval ($\overline{RR}$) was used as sampling rate for estimating the PSD functions, what has been shown acceptable for low frequencies far from the Nyquist frequency [4]. As usual in HRV studies, the spectral energy within each frequency band (*band*) was measured taking the areas (P^{band}) below the spectra, S,

$$P_E^{band} = \int_{f\in band} S_E(f)df; \tag{5}$$

with $E \in \{QT, QT/W_{QT}, QT/W_{RR}\}$. The ratios between $P_{QT/W_{QT}}^{band}$ and in total power P_{QT}^{band} represent the relative contribution of the QTV not driven by RR in the frequency band *band*.

2.2 Model identification and order selection

From the $RR[n]$ and $QT[n]$ interval series corrected from the mean, the polynomial A_{11} was estimated using least squares, while the ARARX model parameters were iteratively obtained using a generalized least squares methodology [2]. For adequate orders the convergence to white noise residual W_{QT} is expected in a reasonable small number of iterations and a large enough SNR guarantees that the minima of the square residue are global [2].

From $p, q \in \{6, 8, 10, 12, 14, 16, 18\}$, an order was considered to be adequate for modelling a given segment of data if the normalized autocorrelations of the residual ($W_{RR}[n]$ or $W_{QT}[n]$) satisfied a 5% significance bilateral test, both in lags lower than 40 beats and considering all lags. The optimal p and q were automatically selected from the adequate orders as the ones that better satisfied a common criteria such as FPE or AIC [2]. The uncorrelation between $W_{RR}[n]$ and $W_{QT}[n]$ was also verified for the same 5% significance.

2.3 Simulation set-up and performance evaluation

The validation of the model was based on simulated $RR[n]$ and $QT[n]$ series with known QTV fraction correlated with RR ($QT_{W_{RR}}$).

The $RR[n]$ sequences were simulated using a model IPFM (integral pulse frequency modulation) [4] following a AR_{10} modulating signal. Two models (RR1 and RR2) with different main frequency components (Figure 3) were used to simulate uncorrelated RR series realizations.

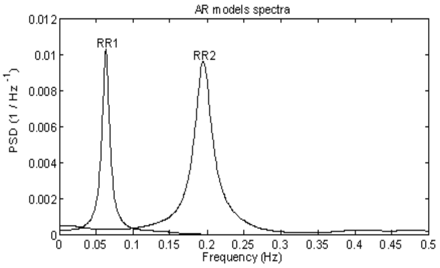


Figure 3: Spectra of the AR_{10} models used to generate data.

To obtain realistic QT series from RR sequences we considered a constant QT value qt_0, extracted from a real beat, and used the classical $Bazett's$ formula as a static relation between a QT and the previous RR [8]: $QTj[n] = qt_0\sqrt{RRj[n]}$, for $j = 1, 2$. The test data was defined considering 3 cases:

A: QT and RR correlated: $RR1_i[n]$ vs $QT1_i[n]$ and $RR2_i[n]$ vs $QT2_i[n]$;

B: QT and RR uncorrelated: $RR1_i[n]$ vs $QT2_i[n]$ and $RR2_i[n]$ vs $QT1_i[n]$;

C: Mixture of the dependencies: $RR1_i[n]$ vs $QT1_i[n]+QT2_i[n]-\overline{QT2_i}$ and $RR2_i[n]$ vs $QT1_i[n]+QT2_i[n]-\overline{QT1_i}$; were $\overline{QTj_i} = (\sum_{n=0}^{N} QTj_i[n])/N$.

were i denotes realization. The $QT[n]$ fraction linearly driven by $RR[n]$ is denoted as $QT_{RR}[n]$ and calculated for each pair of test data as the projection of ($QT_i[n]$-$\overline{QT_i}$) over the subspace generated by the corresponding ($RR_i[n]$-$\overline{RR_i}$) and its delayed vectors up to order 10 (in accordance with RR simulation). The ratio between the power variability measures of this projection and of the total QTV corresponds to the fraction correlated with HRV. The reference variability measures $\tilde{P}_E^{band}$ were obtained from AR_{10} spectral model, analogously as P_E^{band} and the errors calculated as $P_E^{band} - \tilde{P}_E^{band}$.

After identification of the model (figure 2), from the estimated coefficients and the residues $W_{RR}[n]$ and $W_{QT}[n]$, we calculated explicitly the signals $QT_{W_{RR}}[n]$ and $QT_{W_{QT}}[n]$ corresponding to the two uncorrelated driving sources in QT ($QT[n] - \overline{QT} = QT_{W_{QT}}[n] + QT_{W_{RR}}[n]$). The similarity of $QT_{W_{RR}}[n]$ and the reference projection $QT_{RR}[n]$ was evaluated from the coherence between them and the same was applied to $QT_{W_{QT}}[n]$ versus the difference between ($QT[n] - \overline{QT}$) and the reference projection (corresponding to the QTV fraction uncorrelated to HRV). Both spectral coherences were calculated using an non parametric approach (Welch method with a Hanning window).

2.4 Real data set

ECG recordings of young normal subjects from POLI/MEDLAV and Politecnico Ca' Granda databases [6] were used in this study (3 leads at 500 Hz) and each lead was processed by the delineation system in [3]. Only segments with minimum length of 315 consecutive beats with valid RR and QT intervals were considered in the subsequent analysis: anomalies in RR series were identified [5] and QT intervals out of a 3-standard deviation band were rejected as possible outliers. Longer segments were carved up respecting the minimum length admitted, what allowed to obtain 29 segments from POLI/MEDLAV database and 135 segments from Politecnico Ca' Granda database, with a mean length of 415 and 402 beats (≈ 292.46 and 329.24 sec), respectively.

3 Results and discussion

The methods were implemented using MATLAB and the facilities of the System Identification Toolbox. All the results are relative to the orders chosen by FPE but analogous ones were obtained using AIC. To evaluate whether the uncorrelated fraction differed for different frequencies, the measures were estimated considering separately low frequency (band = LF: 0.04-0.15 Hz) and high frequency (band = HF: 0.15-0.4Hz), frequency bands typically used in HRV studies. Total power (band = TP) was considered as the band from 0.04 Hz to the highest frequency present in each spectrum.

3.1 Simulated data

We simulated 50 uncorrelated RR realizations ($i = 1, ..., 50$) with 348 beats at 500 Hz, resulting on a test data of 300 pairs of RR vs QT series.

As expected the orders selected (figure 4, left) were mainly the lowest, as an AR_{10} was used to generated $RR[n]$ series and the IPFM model does not change revelantly the frequency components, for the considered frequency bands [4]. The errors in the calculated ratios between $P^{band}_{QT/W_{QT}}$ and P^{band}_{QT} were lower than 5% for more than 75% of the series in LF and HF frequency bands and for about 96% of the segments considering TP. In the right panel of figure 4 are presented the distribution of the errors for each case.

The mean and standard deviation of the errors in the estimated QTV fraction uncorrelated to RR can be found in the table below, for each case of simulated series and for all the data set, considering each frequency band.

Considering all data sets, the mean errors were lower that 4% for all

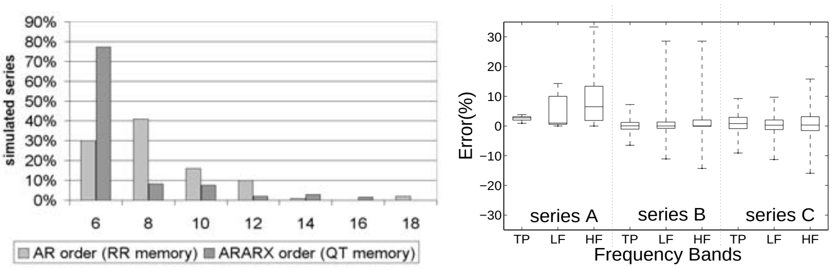


Figure 4: Simulated data: left) selected orders; right) box and whisker plots of errors in the ratios between $P^{band}_{QT/W_{QT}}$ and P^{band}_{QT}.

	series A	series B	series C	all data sets
$P^{TP}_{QT/W_{QT}}$	2.53 ± 0.76	-0.06 ± 2.11	0.88 ± 3.12	1.12 ± 2.46
$P^{LF}_{QT/W_{QT}}$	4.56 ± 5.21	1.01 ± 5.99	0.63 ± 3.22	2.07 ± 5.24
$P^{HF}_{QT/W_{QT}}$	9.45 ± 8.89	1.46 ± 7.10	0.55 ± 4.48	3.82 ± 8.10

Table 1: Errors in ratios over simulated data ($\%, mean \pm std$).

bands. The increased error found in series A is due to the very low power of RR1 and QT1 series in HF band and of RR2 and QT2 in LF band (as illustrated in Fig. 2), resulting in a small absolute error on the estimated PSD measures holding a high percent importance. Eliminating the series A the mean results became lower that 1%, with a relevant decrease on std values (0.41 ± 2.70 for TP, 0.82 ± 4.80 for LF and 1.00 ± 5.94 for HF).

The gain (evaluated as the squared absolute value) and phase (angle) of the complex spectral coherence γ between the model estimated $QT_{W_{RR}}[n]$ and the projection $QT_{RR}[n]$ used as reference are presented in the table below, for each frequency band. The high gains reflect the degree of similarly between the variability distribution shapes and thus $QT_{W_{RR}}[n]$ has frequency contents close to $QT_{RR}[n]$ both in power as in location of peaks. The lower gains relative to $QT_{W_{QT}}[n]$ are partially related with the very low power in some regions of the simulated spectra (case A). Excluding these series the mean gain increase to 0.87 (TP), 0.93 (LF) and 0.86 (HF). The very low phases reflect the no existence of delays in the model.

	$QT_{W_{QT}}[n]$		$QT_{W_{RR}}[n]$	
	γ gain	γ phase (rad)	γ gain	γ phase (rad)
TP	0.98	0.01	0.76	-0.01
LF	0.99	-0.02	0.80	-0.01
HF	0.90	0.01	0.75	-0.01

Table 2: Mean spectral coherence between QTV fractions and the references.

3.2 Real data

We obtained adequate models for 28 segments in POLI/MEDLAV and 132 in Politecnico Ca' Granda database and in 4 cases for which we did not found an adequate order to the AR model part.

In the left panel of figure 5 the orders selected by FPE for each model part are presented. Lower orders are more frequent for the ARARX model than for AR, reflecting different dependence of QT from its own past (memory) and past RR intervals. This can also be seen in simulated signals (left side of figure 4) validating that the QT intervals were realistically simulated.

The fraction uncorrelated with HRV was found to be higher than 40% for 98% of the segments in TP and HF band and for 91% in LF suggesting that other factors rather than RR could drive an important part of QTV. The values found for the ratios (%) between the measures on uncorrelated fraction and total QTV spectrum were very high, as illustrated in Figure 5. It is worthwhile to remark that in this study we aimed to estimate the fraction of QTV that is not correlated with HRV. The uncorrelation between that part of QTV and HRV does not imply that there is not any physiological dependence between them, since non-linear effects are not taken into account.

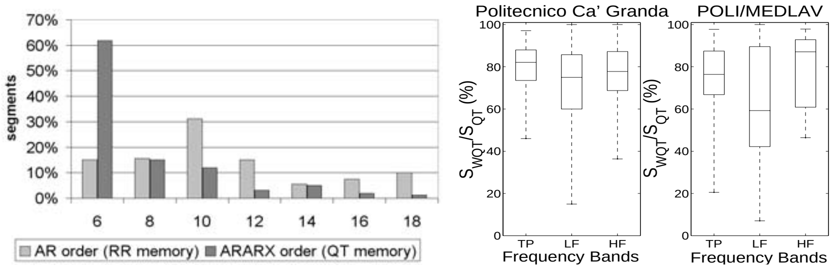


Figure 5: Real data: left) selected orders; right) box and whisker plots of the ratios between $P^{band}_{QT/W_{QT}}$ and P^{band}_{QT}.

4 Concluding remarks

This work discusses the characterization of the short term QT versus RR variabilities by applying a linear open loop model that includes an approach for order selection in each part of the model. The methodology was validated with simulated data and applied to real records. The orders selected for the RR model part are generally higher than for QT what can be associated to differences in the memory of the signals. The results point out that an important part of QTV (more than 40%) is not linearly driven by RR.

The study of the QT versus RR interactions is a complex problem. A deeper characterization requires the incorporation of additional information on the model. Identification and interpretation of the sources non-correlated with RR are the driving force for future studies.

References

[1] Almeida R., Pueyo E., Martínez J.P., Rocha A.P., Olmos S., Laguna P. (2003). *A parametric model approach for quantification of short term QT variability uncorrelated with heart rate variability.* XXX International Conference on Computers in Cardiology, IEEE Computer Society, Thessaloniki, Greece **30**, 165 – 168.

[2] Ljung L. (1999). *System identification theory for the user 2nd edition.* Prentice Hall PTR, Thomas Kailath, Series Editor.

[3] Martínez J.P., Almeida R., Olmos S., Rocha A.P., Laguna P. (2004). *Wavelet-based ECG delineator: evaluation on standard databases.* IEEE Transactions on Biomedical Engineering **51** (4), 570 – 581.

[4] Mateo J., Laguna P. (2000). *Improved heart rate variability signal analysis from the beat occurrence times according to the IPFM model heart timing signal.* IEEE Transactions on Biomedical Engineering **47**, 985 – 996.

[5] Mateo J., Laguna P. (2003). *Analysis of heart rate variability in the presence of ectopic beats using the heart timing signal.* IEEE Transactions on Biomedical Engineering **50**, 334 – 342.

[6] Pinciroli F., Pozzi G., Rossi R., Piovosi M., Capo A., Olivieri R., Della Torre M. (1988). *A respiration-related EKG database.* XV International Conference on Computers in Cardiology, IEEE Computer Society **15**, 477 – 480.

[7] Porta A., Baselli G., Caiani E., Malliani A., Lombardi F., Cerutti S. (1998). *Quantifying electrocardiogram RT-RR variability interactions.* IEEE Transactions on Biomedical Engineering **36**, 27 – 34.

[8] Pueyo E., Smetana P., Malik M., Laguna P. (2003). *Evaluation of QT interval response to marked RR interval changes selected automatically in ambulatory recordings.* XXX International Conference on Computers in Cardiology, IEEE Computer Society, Thessaloniki, Greece. **30**, 157 – 160.

Acknowledgement: The first author acknowledges the grant SFRH/BD/5484/2001 supported by FCT and ESF (III CSF). This work was also supported by the integrated action HP2001-0031/CRUP-E26/02 and projects TIC2001-2167-C02-02 from MCYT/FEDER, P075/2001 from CONSID-DGA (Spain).

Address: R. Almeida, A.P. Rocha, Departamento de Matemática Aplicada, Faculdade de Ciéncias, Universidade do Porto, Rua Campo do Alegre 687, 4169-007 Porto, Portugal,

E. Pueyo, J.P. Martínez, P. Laguna, Comm. Techn. Group, Aragon Institute of Eng. Research, Zaragoza Univ., María de Luna 1, Edificio Ada Byron, 50018 Zaragoza, Spain

E-mail: `rbalmeid@fc.up.pt, aprocha@fc.up.pt, laguna@posta.unizar.es`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

THE THRESHOLD ARMA MODEL AND ITS AUTOCORRELATION FUNCTION

Alessandra Amendola, Marcella Niglio
and Cosimo Damiano Vitale

Key words: SETARMA model, indicator process, autocorrelation.

COMPSTAT 2004 section: Time series analysis.

Abstract: This paper considers the autocorrelation function of a particular family of threshold structure, the Self Exciting AutoRegressive Moving Average (SETARMA) model. Its is initially presented in the original form and some properties of the process which regulates the switching among the regimes are shown. An alternative representation of the SETARMA structure is then proposed and its autocorrelation function is exactly derived.

1 Introduction

The statistical and econometric literature has given an increasing attention to threshold models and different structures have been proposed in the years.

In the present paper the attention has been focused on a particular class of threshold models, the Self Exciting Threshold AutoRegressive Moving Average [4], whose dependence structure has been investigated. In particular the aim of the paper is to derive the analytic expression of the autocorrelation function of the SETARMA models which can be used to compute well known indexes to evaluate how far the process is from the linearity.

In Section 2 the SETARMA model has been shown in its traditional form and the main properties of the process which regulates the switching in a two regimes structure are discussed. An alternative representation of the model is further proposed under well defined conditions on the generating process. In Section 3 the exact form of the autocorrelation function of the model under analysis is presented and its expression for a SETARMA(2; 1,1; 1,1) is shown and discussed. Some concluding remarks are given in the final section.

2 The model under analysis

The Self-Exciting Threshold AutoRegressive Moving Average process [4] of order $(k; p_1, \ldots, p_k; q_1, \ldots, q_k)$ is defined as:

$$X_t = \sum_{i=1}^{k} \left[\phi_0^{(i)} + \sum_{j=1}^{p_i} \phi_j^{(i)} X_{t-j} + \sigma_i \left(e_t - \sum_{w=1}^{q_i} \theta_w^{(i)} e_{t-w} \right) \right] I(X_{t-d} \in R_i) \quad (1)$$

where $\sigma_i e_t \sim WN(0, \sigma_i^2)$, for $i = 1, \ldots, k$, $R_i = [r_{i-1}, r_i)$ forms a partition of the real line such that $-\infty = r_0 < r_1 < r_2 < \ldots < r_k = +\infty$ with r_i the threshold values, d is the threshold delay, p_i and q_i are non negative integers, $\phi_j^{(i)}$ and $\theta_w^{(i)}$ are unknown parameters with $j = 1, 2, \ldots, p_i$ and $w = 1, 2, \ldots, q_i$ and $I(\cdot)$ is a Bernoulli random process.

In the following, in order to easily show the results on the autocorrelation function and to avoid an heavy notation, a SETARMA $(2; p_1, p_2; q_1, q_2)$ model with threshold delay d and threshold value r, such that $R_1 = [r, \infty)$, $R_2 = (-\infty, r)$, is considered. In this case the switching between the two regimes is regulated by the indicator process $I_{t-d} = I(X_{t-d} \in R_1)$ that, given its dichotomy, can be written as:

$$I_{t-d} = \begin{cases} 1 & \text{if} \quad X_{t-d} \geq r \\ 0 & \text{if} \quad X_{t-d} < r \end{cases} \tag{2}$$

for $t = 1, 2, \ldots$, and $d > 0$.

I_{t-d} has a relevant role in the dynamic structure of X_t and it is characterized by some interesting properties shown in the next section.

2.1 The indicator process $\mathrm{I}_{\mathrm{t-d}}$

The process I_{t-d} in (2), which controls the switch between the two regimes of the SETARMA model, satisfy four main properties which are strictly related to its dichotomy:

a) $E[I_{t-d}] = P(I_{t-d} = 1) = P(X_{t-d} \geq r)$
b) $I_{t-d}^v = I_{t-d}$ and $(1 - I_{t-d})^s = 1 - I_{t-d}$
c) $I_{t-d}^v (1 - I_{t-d})^s = 0 \qquad \text{for} \quad v, s = 1, 2, 3, \ldots$
d) $I_{t-d} \cdot I_{t-d-k} = \begin{cases} 1 & \text{if} \;\; X_{t-d} \geq r \;\text{ and }\; X_{t-d-k} \geq r \\ 0 & \text{otherwise} \end{cases} \qquad \text{with} \quad k \in N$

In order to focus the attention on some properties of the process $\{I_{t-d}\}$, $t = d + 1, d + 2, \ldots$, the following assumption is fixed:

[A1.] *The process I_{t-d} is second order stationary and ergodic.*

which implies that:

- $E(I_{t-d}) = p$
- $var(I_{t-d}) = p(1 - p) \qquad t = d + 1, d + 2, \ldots$
- $cov(I_{t-d}, I_{t-d-k}) = \gamma_I(k) = p_k - p^2$

where:

$$p_k = E[I_{t-d} I_{t-d-k}] \qquad \text{for } k = 1, 2, \ldots .$$

It further follows that the double random variable (I_{t-d}, I_{t-d-k}) has the probability distribution shown in Table 1, where p and p_k have to satisfy the inequalities:

I_{t-d}/I_{t-d-k}	**0**	**1**	
0	$1-2p+p_k$	$p-p_k$	$1-p$
1	$p-p_k$	p_k	p
	$1-p$	p	1

Table 1: Probability distribution of the random variable (I_{t-d}, I_{t-d-k}).

$$p_k \leq p \leq \frac{1+p_k}{2} \qquad \text{and} \qquad \max(0, 2p-1) \leq p_k \leq p \tag{3}$$

It implies that:

$$\sum_{k=0}^{\infty} |\gamma_I(k)| = p(1-p) + \sum_{k=1}^{\infty} |p_k - p^2| < \infty \tag{4}$$

where, recalling that $p_0 = E(I_{t-d}^2) = p$, $\gamma_I(0) = p_0 - p^2 = p(1-p)$.

The estimate of the probabilities p and p_k can be obtained taking advantage of the properties of the Bernoulli process I_{t-d}. In particular, given the assumption [A1.], a consistent estimate in probability for p and p_k is given respectively by:

$$\hat{p} = \frac{\#(x_{t-d} \geq r)}{T-d} \qquad\qquad \hat{p}_k = \frac{\#(x_{t-d} \geq r, x_{t-d-k} \geq r)}{T-d} \tag{5}$$

where $x_1, x_2, \ldots, x_T$ is the time series of length T generated from the stochastic process X_t.

2.2 An alternative representation for the SETARMA model

A SETARMA model can be seen as a direct generalization, in the non-linear domain, of the ARMA model of Box and Jenkins [1]. In particular, under well defined conditions, some results reached in the linear context can be properly extended to this class of nonlinear models and alternative representations of them can be derived.

Given these further two assumptions on X_t:

[A2.] *all the roots of the autoregressive polynomials* $\phi_{p_i}(B) = 1 - \sum_{j=1}^{p_i} \phi_j^{(i)} B^j$, *lie outside the unit cycle, (for* $i = 1, 2$*)*;

[A3.] *given the SETARMA model* $X_t = X_t^{(1)} I_{t-d} + X_t^{(2)}(1 - I_{t-d})$, *with* $X_t^{(i)} \sim ARMA(p_i, q_i)$ *for* $i = 1, 2$, *the joint process* $\mathbf{X_t} = (X_t^{(1)}, X_t^{(2)}, I_{t-d})$ *is strict stationary and ergodic*;

model (1), with $k = 2$, can be alternatively written as:

$$X_t = \left[c_0^{(1)} + \sigma_1 \sum_{j=0}^{\infty} \psi_j^{(1)} B^j e_t \right] I_{t-d} + \left[c_0^{(2)} + \sigma_2 \sum_{j=0}^{\infty} \psi_j^{(2)} B^j e_t \right] (1 - I_{t-d}) \quad (6)$$

where

- $c_0^{(i)} = \frac{\phi_0^{(i)}}{1 - \sum_{j=1}^{p_i} \phi_j^{(i)}}$ is the mean value of regime i, for $i = 1, 2$;
- $\sum_{j=1}^{\infty} |\psi_j^{(i)}| < \infty$, with $\psi_0^{(i)} = 1$ and $i = 1, 2$;
- and the weights $\psi_j^{(i)}$ (for $i = 1, 2$ and $j = 0, 1, 2, \ldots$) are computed as:

$$\psi_j^{(i)} = \begin{cases} 1 & \text{when} \quad j = 0 \\ \sum_{s=0}^{j-1} \psi_s^{(i)} \phi_{j-s}^{(i)} - \theta_j^{(i)} & \text{when} \quad j \geq 1 \end{cases} \quad (7)$$

with $\phi_j^{(i)} = 0$, for $j > p_i$, and $\theta_j^{(i)} = 0$, for $j > q_i$.

Model (6) allows to derive, in an easier form, the results in Section 3 and therefore is used as shown in the following pages.

3 The autocorrelation

The autocorrelation coefficient $\rho(k)$, given as:

$$\rho(k) = \frac{cov(X_t, X_{t \pm k})}{[Var(X_t) Var(X_{t \pm k})]^{1/2}} \qquad k = 0, 1, 2, \ldots, N \quad (8)$$

is a remarkable tool for the analysis of linear time series to study the dependence among the random variables and to identify the ARMA models in the Box and Jenkins [1] approach.

When nonlinear models are under analysis, $\rho(k)$ is not sufficient to investigate the relation among the X_t's ($t = 1, 2, \ldots, N$) even if it can give useful practical indications to evaluate how the generating mechanism of $\{X_t\}$ is closed to the linearity. In this context Tong [5] proposes an *index of linearity* based on the use of the square autocorrelations; more recently Nielsen and Madsen [3] generalize some traditional tools, such as the global and partial autocorrelation function, for the identification of nonlinear models. Their use imply the knowledge of the autocorrelation of nonlinear models in order to evaluate their applicability. For example, no advantages can be reached using $\rho(k)$ when the data generating process is related to a purely bilinear model [2] where the autocorrelation is shown to be zero, with the exception of few and well defined representations of the model.

When instead $X_t \sim$ SETARMA$(2; p_1, p_2; q_1, q_2)$, its local linearity naturally lead to investigate on $\rho(k)$.

Starting from the denominator of (8), the stationarity assumed for X_t allows to write $[Var(X_t)Var(X_{t\pm k}]^{1/2} = Var(X_t)$ and using model (6):

$$Var(X_t) = p\sigma_1^2 \sum_{j=0}^{\infty} \left(\psi_j^{(1)}\right)^2 + (1-p)\sigma_2^2 \sum_{j=0}^{\infty} \left(\psi_j^{(2)}\right)^2 + p(1-p)[c_0^{(1)} - c_0^{(2)}]^2 \quad (9)$$

where $c_0^{(1)}$ and $c_0^{(2)}$ are the two regimes constants.

The numerator of (8) instead involves more detailed investigation as shown in the following proposition.

Proposition 3.1. *If* $X_t \sim \mathrm{SETARMA}(2; p_1, p_2; q_1, q_2)$ *and the assumptions [A.2] and [A.3] are satisfied,* [so that the SETARMA model can be written in the alternative form (6)], *the autocovariance of* X_t *at lag* k*, with* $k = 0, 1, \ldots, N$*, is:*

$$\begin{aligned}\gamma(k) \quad &= \quad \sum_{j=0}^{\infty} \Big[p_k \sigma_1^2 \psi_j^{(1)} \psi_{k+j}^{(1)} + (1-2p+p_k)\sigma_2^2 \psi_j^{(2)} \psi_{k+j}^{(2)} + (p-p_k)\sigma_1\sigma_2 \cdot \\ &\quad \cdot \left(\psi_j^{(1)}\psi_{k+j}^{(2)} + \psi_{k+j}^{(1)}\psi_j^{(2)}\right)\Big] + (p_k - p^2)(c_0^{(1)} - c_0^{(2)})^2 \qquad (10)\end{aligned}$$

with $p_k = E[I_{t-d}I_{t-d-k}]$.

Proof. The assumption of stationarity implies the symmetry of the autocovariance of X_t, such that $cov(X_t, X_{t-k}) = cov(X_t, X_{t+k})$ and so the proof can be limited to the case $\gamma(k) = cov(X_t, X_{t-k})$.

Starting from the definition of $\gamma(k)$:

$$\begin{aligned}\gamma(k) \quad &= \quad cov(I_{t-d}X_t^{(1)}, I_{t-d-k}X_{t-k}^{(1)}) + cov(I_{t-d}X_t^{(1)}, (1-I_{t-d-k})X_{t-k}^{(2)}) + \\ &+ \quad cov((1-I_{t-d})X_t^{(2)}, I_{t-d-k}X_{t-k}^{(1)}) + cov((1-I_{t-d})X_t^{(2)}, \qquad (11) \\ &\qquad (1-I_{t-d-k})X_{t-k}^{(2)})\end{aligned}$$

each term of (11) is given as:

a) $cov(I_{t-d}X_t^{(1)}, I_{t-d-k}X_{t-k}^{(1)}) = p_k\gamma_1(k) + (p_k - p^2)(c_0^{(1)})^2$

b) $cov(I_{t-d}X_t^{(1)}, (1-I_{t-d-k})X_{t-k}^{(2)}) = (1-2p+p_k)\gamma_2(k) + (p_k - p^2)(c_0^{(2)})^2$

c) $cov((1-I_{t-d})X_t^{(2)}, I_{t-d-k}X_{t-k}^{(1)}) = (p-p_k)\gamma_{12}(k) - (p_k - p^2)c_0^{(1)}c_0^{(2)}$

d) $cov((1-I_{t-d})X_t^{(2)}, (1-I_{t-d-k})X_{t-k}^{(2)}) = (p-p_k)\gamma_{21}(k) - (p_k - p^2)c_0^{(1)}c_0^{(2)}$

with $E[(1-I_{t-d})(1-I_{t-d-k})] = 1-2p+p_k$ and $E[(1-I_{t-d})I_{t-d-k}] = p-p_k$.

Substituting the results a)-d) in (11), $\gamma(k)$ becomes:

$$\begin{aligned}\gamma(k) \quad &= \quad p_k\gamma_1(k) + (1-2p+p_k)\gamma_2(k) + (p-p_k)[\gamma_{12}(k) + \gamma_{21}(k)] + \\ &\quad +(p_k - p^2)(c_0^{(1)} - c_0^{(2)})^2 \qquad (12)\end{aligned}$$

where

- $\gamma_i(k) = \sigma_i^2 \sum_{j=0}^{\infty} \psi_j^{(i)} \psi_{k+j}^{(i)}$, for $i = 1, 2$;
- $\gamma_{12}(k) = \sigma_1 \sigma_2 \sum_{j=0}^{\infty} \psi_j^{(1)} \psi_{k+j}^{(2)}$;
- $\gamma_{21}(k) = \sigma_1 \sigma_2 \sum_{j=0}^{\infty} \psi_{k+j}^{(1)} \psi_j^{(2)}$;

which lead to the result (10). ◇

The combination of (9) and (10) in (8) allows to obtain $\rho(k)$ for model (6) which assumes different forms when the orders p_i and q_i of the two regimes are selected.

For example when a SETARMA(2; 1,1; 1,1) with no intercepts is chosen, the numerator (12) of the autocorrelation is such that:

$$\begin{aligned}
\gamma_i(k) &= cov(X_t^{(i)} X_{t-k}^{(i)}) = \sigma_i^2 \left(\phi_1^{(i)}\right)^{k-1} \frac{(1-\phi_1^{(i)}\theta_1^{(i)})(\phi_1^{(i)} - \theta_1^{(i)})}{1 - \left(\phi_1^{(i)}\right)^2}, \\
& \quad \text{for} \quad i = 1, 2; \\
\gamma_{12}(k) &= cov(X_t^{(1)} X_{t-k}^{(2)}) = \frac{\sigma_1 \sigma_2}{1 - \phi_1^{(1)} \phi_1^{(2)}} \\
& \left[\left(\phi_1^{(1)}\right)^{k-1} \left(\phi_1^{(1)} - \theta_1^{(1)}\right) \left(1 - \phi_1^{(1)} \theta_1^{(2)}\right) \right] \\
\gamma_{21}(k) &= cov(X_{t-k}^{(1)} X_t^{(2)}) = \frac{\sigma_1 \sigma_2}{1 - \phi_1^{(1)} \phi_1^{(2)}} \\
& \left[\left(\phi_1^{(2)}\right)^{k-1} \left(\phi_1^{(2)} - \theta_1^{(2)}\right) \left(1 - \phi_1^{(2)} \theta_1^{(1)}\right) \right]
\end{aligned}$$

whereas the variance at the denominator is:

$$Var(X_t) = p\sigma_1 \frac{1 + \left(\theta_1^{(1)}\right)^2 - 2\phi_1^{(1)}\theta_1^{(1)}}{1 - \left(\phi_1^{(1)}\right)^2} + (1-p)\sigma_2 \frac{1 + \left(\theta_1^{(2)}\right)^2 - 2\phi_1^{(2)}\theta_1^{(2)}}{1 - \left(\phi_1^{(2)}\right)^2}$$

From the previous results it can be shown that $\rho(k)$ of the SETARMA(2; 1,1; 1,1) model is null as k diverges, that is:

$$\lim_{k \to \infty} \rho(k) = 0$$

and so the autocorrelation between X_t and $X_{t \pm k}$ is zero as the temporal lag $|k|$ grows.

These results can be graphically observed in Figure 1 [frame (b)] where the correlogram of the series generated from a SETARMA(2; 1,1; 1,1) model is considered, with X_t given as:

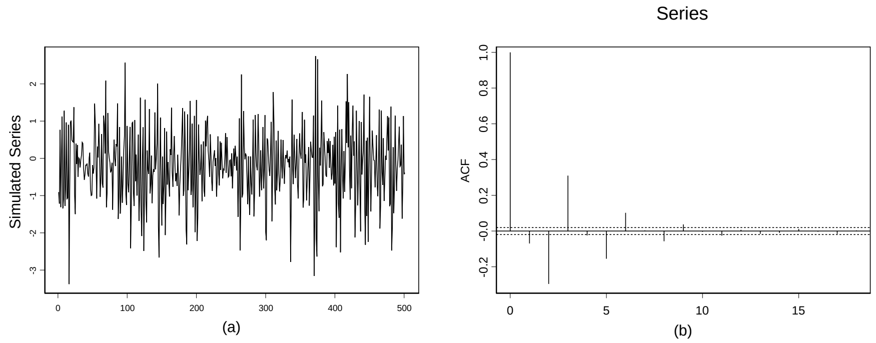


Figure 1: (a): Sample of the generated series of length 500; (b):Correlogram of the complete generated series.

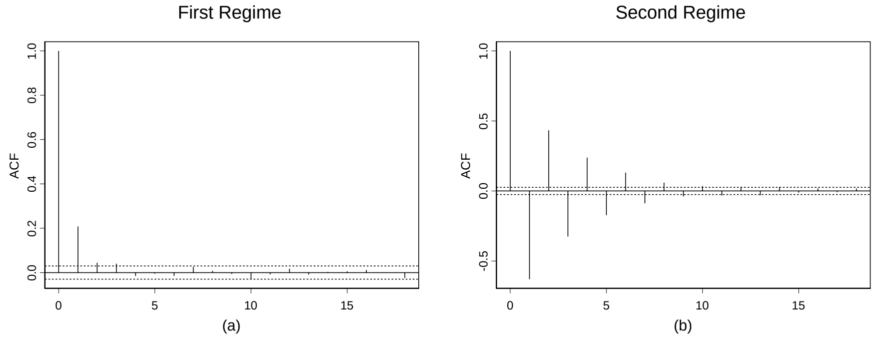


Figure 2: Correlograms of the generated data which belong to the first and second regime [frames (a) and (b) respectively].

$$X_t = \begin{cases} 0.6X_{t-1} + e_t^{(1)} - 0.4e_{t-1}^{(1)} & X_{t-1} \geq 0 \\ -0.6X_{t-1} + e_t^{(2)} + 0.4e_{t-1}^{(2)} & X_{t-1} < 0 \end{cases} \tag{13}$$

where $e_t^{(1)} = e_t$, $e_t^{(2)} = 0.5e_t$ and $\{e_t\}$ is a white noise process with $e_t \sim N(0,1)$, for $t = 1, 2, \dots, 10000$.

In particular the sample autocorrelation decreases exponentially to zero and give useful information about the dependence structure of the series under analysis. As widely discussed in [5], this can be considered a first step to investigate the dependence among the X_t's which needs to be further evaluated with more sophisticated instruments in order to avoid model misspecification. It is even informative for the identification of the two regimes when the threshold delay d and the threshold value r are known and therefore the correlograms of the values which belongs to each regime can be constructed [frames (a) and (b) in Figure 2].

4 Concluding remarks

In time series analysis the autocorrelation function is often used to study the dependence among data and to evaluate which approach to follow to model them. In fact numerous instruments based on its application have been proposed in literature and most of them have been successfully applied.

In this context the exact form of the SETARMA autocorrelation function $\rho(k)$ have been derived in order to investigate some aspects of the dependence structure of this family of process and to show how it changes with respect to the regimes orders and to the lag length between X_t and X_{t-k}.

The results obtained further highlight that some tools, based on $\rho(k)$, which evaluate how the generating process is far from the linearity, can be properly applied in the SETARMA context.

References

[1] Box G.E.P., Jenkins G.M. (1976). *Time series analysis, forecasting and control*. Holden-Day, San Francisco.

[2] Granger C.W.J., Andersen A.P. (1978). *An introduction to bilinear time series models*. Vanderhoeck and Ruprecht, Gottingen.

[3] Nielsen H.A., Madsen H. (2001). *A generalization of some classical time series tools*. Computational Statistics & Data Analysis, **37**, 13 − 31.

[4] Tong H. (1983). *Threshold models in nonlinear time series analysis*. Springer-Verlag, London

[5] Tong H. (1990). *Non-linear time series: A dynamical system approach*. Clarendon Press, Oxford.

Address: A. Amendola, M. Niglio, C. Vitale, Dipartimento di Scienze Economiche e Statistiche, Università degli studi di Salerno, Via Ponte Don Melillo, 84084 Fisciano (Sa), Italy

E-mail: alamendola@unisa.it, mniglio@unisa.it, vitale@unina.it

FUNCTIONAL DISCRIMINANT ANALYSIS FOR MICROARRAY GENE EXPRESSION DATA VIA RADIAL BASIS FUNCTION NETWORKS

Yuko Araki, Sadanori Konishi and Seiya Imoto

Key words: Classification, functional data analysis, maximum penalised likelihood, microarray, radial basis function.

COMPSTAT 2004 section: Functional data analysis.

Abstract: We introduce functional logistic discriminant analysis (FLDA) which is an extension of the classical method of logistic discriminant analysis to data where predictor variables are functions or curves. FLDA approach can effectively classify functions into two distinct classes by imposing smoothness constraint on the predictor functions and coefficient function by radial basis function expansion and regularization. In order to select the value of a smoothing parameter, we derive an information criterion which enables us to evaluate model estimated by regularization. The proposed method is illustrated through the analysis of yeast cell cycle microarray data. It is shown that FLDA performs well especially in terms of prediction ability.

1 Introduction

Classification or discrimination technique is one of the most widely used statistical tools in various fields of natural and social sciences. In recent years, several techniques have been proposed for analyzing multivariate observations with complex structure (see, for example, [3]).

The focus in the present paper will be on the problem of classifying functions, where each observation can be interpreted as a discretized realization of a function evaluated at possibly differing time points. Recently, Cardot et al. [2] used functional approaches for estimating land use based on the temporal evolution of remote sensing data.

Our motivation arises from the analysis of yeast cell cycle gene expression data which provide inference about how gene expression levels evolve in time and how genes are dependent during a given biological process [10] and [5]. Classification of genes enables us to predict functions of unknown genes and to identify the set of co-regulated genes. In the yeast cell cycle data analysis, one wish to classify genes based on the cDNA microarray time series data.

We introduce functional discriminant analysis using Gaussian radial basis function networks with help of regularization. It is designed to construct a decision rule based on data given as a set of functions. We first transfer the vector valued observations to a set of functions. Secondly, functional logistic model is constructed by using Gaussian radial basis functions and then estimation is by regularized maximum likelihood method. In order to select smoothing parameters, we derive model selection criterion within the framework of functional data analysis by developing the generalized information criterion due to [4].

The paper is organized as follows. In Section 2 we describe the radial basis expansion smoothing technique which converts discrete raw data into underlying smooth functional form. In Section 3 the new method, functional logistic discriminant analysis, is set out and the details of its implementation are described. Section 4 presents an application of the proposed method to yeast cell cycle gene expression data collected by [10].

2 Radial basis smoothing techniques

In the context of functional data analysis [9], individual data should be considered to have a functional form in nature even though observed data are usually recorded discretely. In addition, those discrete raw data which are supposed to have functional form may contain observational error. Therefore, converting raw data into underlying smooth functional form requires efficient smoothing techniques.

The typical functional data analysis approach is to fit each curve individually using expansion in basis functions. Common basis functions for smoothing functional data are B-spline basis and Fourier expansions. In our model, we use Gaussian radial basis function with hyperparameter [1]. An advantage of this basis expansion is that it controls the amount of overlapping among basis functions and adopts the information of the desired outputs. For background about radial basis function networks, we refer to [7], [8].

Suppose we have N independent observations $\{(x_i, t_i);\ t_i \in \mathcal{T},\ i = 1, 2, \cdots, N\}$, where x_i are random response variables and t_i are explanatory variables, assuming that they are drawn from the Gaussian nonlinear regression model

$$x_i = u(t_i) + \epsilon_i, \qquad i = 1, \cdots, N, \tag{1}$$

where $u(t)$ is a smooth function to be estimated, and the errors ϵ_i are independently, normally distributed with mean zero and variance σ^2. We consider the function $u(t)$ that can be expanded in the form of the radial basis function network taking the following form;

$$u(t; \boldsymbol{\omega}) = \sum_{k=1}^{m} \omega_k \phi_k(t) + \omega_0, \tag{2}$$

where $\boldsymbol{\omega} = (\omega_0, \omega_1, \cdots, \omega_m)^T$ and $\phi_k(t)$ are a set of Gaussian radial basis functions with hyperparameter ν given as

$$\phi_k(t; \mu_k, \sigma_k^2) = \exp\left\{ -\frac{(t - \mu_k)^2}{2\nu\sigma_k^2} \right\}, \qquad k = 1, \cdots, m, \tag{3}$$

where μ_k is a scalar determining the location of the kth basis function, σ_k is the width, ν is a hyperparameter. The function $\hat{u}(t) \equiv x(t)$ which is estimated from the observed data $\{(x_i, t_i); i = 1, \cdots, N\}$ is called 'functional data', and is proceeded to further analysis.

The nonlinear function $u(t)$ is estimated in two-stage procedure; position the centers and determine the dispersions first, then calculate the weights using an appropriate optimization schemes. This two stage learning is reported to solve the problem of convergence and the identification problem. Among several strategies, k-means clustering method algorithm is used to determine the centers μ_k and the dispersion parameters σ_k^2 of the basis functions. More precisely, observation points $\{t_1, \cdots, t_N\}$ are grouped into m clusters $\{C_1, \cdots, C_m\}$, where m is a given number of radial basis functions. Then the centers and the dispersion parameters are determined by

$$c_k = \frac{1}{n_k} \sum_{t_i \in C_k} t_i, \qquad s_k^2 = \frac{1}{n_k} \sum_{t_i \in C_k} (t_i - c_k)^2,$$

where n_k represents the number of data which belong to the cluster C_k. We define the basis function $\phi_k(t; c_k, s_k^2)$ using those estimates as $\phi_k(t)$. Hence it follows that the nonlinear regression model based on the radial basis function network can be written as

$$f(x_i | t_i; \boldsymbol{\omega}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\left\{x_i - \boldsymbol{\omega}^T \boldsymbol{\phi}(t_i)\right\}^2}{2\sigma^2} \right], \tag{4}$$

where $\boldsymbol{\phi}(t_i) = (1, \phi_1(t_i), \cdots, \phi_m(t_i))^T$.

In fitting data with complex structure, the maximum likelihood method does not yield satisfactory results, since it often occurs overfitting and yields unstable parameter estimates. Moreover, in smoothing functional data, all individual data should be fitted by using the common basis functions in our model. In other words, the number of basis functions is fixed even though the amount of smoothness imposed on a set of discrete data will be differ from each other. Therefore the unknown weights and the error variances are estimated by regularization method. Regularization allows us to adjust individual differences by a smoothing parameter. In addition, implementing the hyperparameter and adjusting the smoothing parameter capture the structure in the data flexibly.

The regularization method maximizes the penalized log-likelihood function

$$l_\gamma(\boldsymbol{\omega}, \sigma^2) = \sum_{i=1}^{N} \log f(x_i | t_i; \boldsymbol{\omega}, \sigma^2) - \frac{N\gamma}{2} \boldsymbol{\omega}^T D_2^T D_2 \boldsymbol{\omega}, \tag{5}$$

where $D_2^T D_2$ is the second order difference matrix and γ is called a smoothing parameter which adjusts the amount of smoothness and also avoids ill-posed problem. The maximum penalized likelihood estimates are

$$\hat{\boldsymbol{\omega}} = (\Phi^T\Phi + N\beta D_2^T D_2)^{-1}\Phi^T \boldsymbol{x}, \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\{x_i - \hat{\boldsymbol{\omega}}^T \boldsymbol{\phi}(t_i)\}^2, \tag{6}$$

where $\Phi = (\boldsymbol{\phi}(t_1), \boldsymbol{\phi}(t_2), \cdots, \boldsymbol{\phi}(t_N))^T$, $\beta = \gamma\sigma^2$ and $\boldsymbol{x} = (x_1, \cdots, x_N)^T$.

The number of basis functions m, the adjusted parameters ν and γ are determined by using an information criterion given by [1]. Thus the observed discrete data $\{(x_i, t_i); t_i \in \mathcal{T}, i = 1, \cdots, N\}$ are smoothed by the method described above and we have a functional data given by $x(t)$;

$$\hat{u}(t) = \sum_{k=1}^{m} \hat{\omega}_k \phi_k(t) + \hat{\omega}_0 \equiv x(t), \qquad t \in \mathcal{T}. \tag{7}$$

Marx and Eilers [6] used B-splines expansion with regularization, called P-splines, and applied the procedure to medical diagnosis and phoneme recognition.

3 Functional logistic discrimination

Suppose we have n independent observations $\{(x_\alpha(t),\ g_\alpha);\ \alpha = 1, \cdots, n\}$, where $x_\alpha(t)$ are functional predictor variables and g_α are indicators of the group membership. For example, we consider two-class classification, i.e. $k = 1$ or 2, $g_\alpha = k$ implies that it belongs to class G_k. A set of functions smoothed by the Gaussian radial basis function smoothing method are given by

$$x_\alpha(t) = \boldsymbol{w}_\alpha^T \boldsymbol{\phi}(t), \qquad \alpha = 1, \cdots, n, \tag{8}$$

where $\boldsymbol{w}_\alpha$ are estimated parameter vectors and $\boldsymbol{\phi}(t)$ is a vector of Gaussian basis functions given in equation (3).

A Bayes rule of allocation is to assign $x_\alpha(t)$ to group $G_k (k = 1, 2)$ with the maximum posterior probability $\Pr(g = k|x_\alpha(t))$. We consider the log-odds of the posterior probability given in the following form;

$$\log\left\{\frac{\Pr(g = 1|x_\alpha(t))}{\Pr(g = 2|x_\alpha(t))}\right\} = \beta_a + \int_{\mathcal{T}} \beta(t) x_\alpha(t) dt. \tag{9}$$

By making use of the same Gaussian radial basis function $\boldsymbol{\phi}(t)$ as in (8), we expand the functional parameter as $\beta(t) = \beta_0 + \sum_{i=1}^{m} \beta_i \phi_i(t) = \boldsymbol{\beta}^T \boldsymbol{\phi}(t) (\in \mathcal{T})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_m)^T$. We denote the posterior probability $\Pr(g = 1|$ $x_\alpha(t)) = \pi(x_\alpha(t))$, so that $\Pr(g = 2|x_\alpha(t)) = 1 - \pi(x_\alpha(t))$. Then the log-odds model (9) can be expressed as

$$\log\left\{\frac{\pi(x_\alpha(t))}{1 - \pi(x_\alpha(t))}\right\} = \boldsymbol{Z}_\alpha^T \boldsymbol{\beta}, \tag{10}$$

where Z is an $n \times (m + 2)$ matrix given by

$$Z^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \Phi^T \boldsymbol{w}_1 & \Phi^T \boldsymbol{w}_2 & \cdots & \Phi^T \boldsymbol{w}_n \end{bmatrix} \tag{11}$$

with $(m+1) \times (m+1)$ matrix Φ having $\phi_{jk} = \int \phi_j(t)\phi_k(t)dt$ as the (j,k)-th element.

We define the binary variable y_α coded as either 0 or 1 to indicate the group membership of a sample, where $y_\alpha = 1$ if $g_\alpha = 1$ and $y_\alpha = 0$ if $g_\alpha = 2$. The log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{\alpha=1}^{n} \left[y_\alpha \log \pi(x_\alpha(t)) + (1 - y_\alpha) \log\{1 - \pi(x_\alpha(t))\} \right], \tag{12}$$

where $\pi(x_\alpha(t)) = \exp(\boldsymbol{Z}_\alpha^T \boldsymbol{\beta})/\{1 + \exp(\boldsymbol{Z}_\alpha^T \boldsymbol{\beta})\}$. We estimate the parameter vector $\boldsymbol{\beta}$ by maximizing the penalized log-likelihood function

$$l(\boldsymbol{\beta}) - \frac{n\lambda}{2} \boldsymbol{\beta}^T D_2^T D_2 \boldsymbol{\beta}. \tag{13}$$

This is because regularization method yields estimates with lower variances, even though they are biased. Therefore we obtain the solution $\hat{\boldsymbol{\beta}}_\lambda$ by the iterative algorithm like Newton-Raphson algorithm.

The crucial issue on regularization method is the choice of the optimal value of smoothing parameter λ. We obtain an information-theoretic criterion within the framework of functional data analysis. An information criterion for evaluating functional logistic discrimination model estimated by regularization is of the form

$$\mathrm{GIC_F} = -2 \log l(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{tr} Q^{-1} R, \tag{14}$$

where Q and R are $(m+2) \times (m+2)$ matrices given by the first and second derivatives of equation (13). We choose the smoothing parameter λ to minimize $\mathrm{GIC_F}$.

4 Real data example

In this section we show the effectiveness of the proposed method through the analysis of the yeast cell cycle gene expression data collected by [10]. Gene expressions for all 6,178 genes in the yeast genome were measured by cDNA microarrays over time during about two cell cycles. These data contain 77 microarrays and consist of two short time-courses (two time points) and four medium time-courses (18, 24, 17 and 14 time points). Spellman et al. [10] identified 800 genes as cell cycle related genes based on the clustering analysis, and also grouped these genes into five classes, G1, S, G2, M, and M/G1, by considering peaks in the expression patterns. Figure 1 shows the expression patterns of the 800 genes in the five classes.

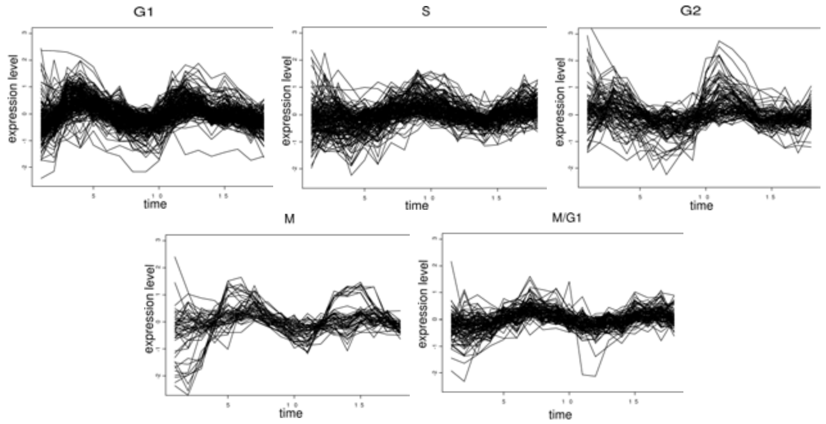


Figure 1: Raw gene expression patterns during the yeast synchronization experiment.

In our analysis, we concentrate on the time-course "α factor-based synchronization experiment data" (18 time points), for simplicity excluding the genes containing missing values. That is, the expression patterns of 612 genes out of 800 cell cycle related genes are used in our analysis, and those expression data are considered as a discretized realization of 612 expression curves evaluated at 18 time points. Note that microarray data usually contain observational noise. Therefore, the smoothing that we will first perform has an important role to remove the observational noise from expression data. In addition, since the gene expression pattern of each cell cycle related gene can be considered as a function of time, the proposed method is appropriate for analyzing time course gene expression data.

We carried out two-class logistic discrimination for all possible combinations. In order to evaluate the effectiveness of our FLDA model, the genes in each class were randomly assigned into training data and test data. That is, the FLDA model is estimated by using the training data, and the predictive ability of the estimated model is evaluated by the test data.

We first performed the Gaussian radial basis smoothing method described in Section 2 to the time-course expression data $\{(t_i, x_{ij});\ i = 1, \ldots, 18;\ j = 1, \ldots, 612\}$, where t_i is the i-th time point and x_{ij} is the expression value of j-th gene at time t_i. In the functional discrimination analysis, each smoothing step has to be carried out by using the same number of basis functions. Hence the differences of the degree of smoothness between different gene expression patterns can be adjusted by the smoothing parameters. However, since there are various gene expression patterns in the same group, adjusting their smoothness by using only the smoothing parameter might not be enough. In such a case, the proposed radial basis function smoothing method with the hyperparameter works efficiently in practice.

Figure 2 shows examples of two G1-grouped genes with the Gaussian ra-

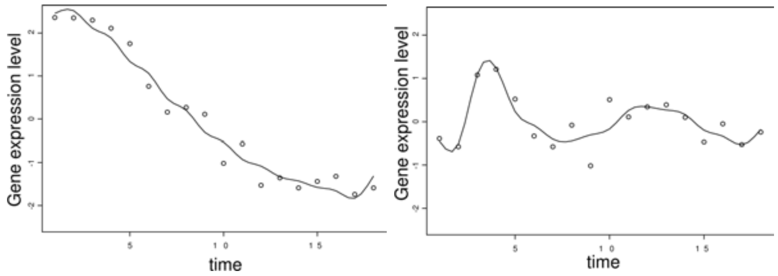


Figure 2: Smoothed gene expression patterns by the Gaussian radial basis function networks with hyperparameter.

dial basis smoothing curves. Although there are various types of expression patterns in the same class, we succeeded in extracting the effective expression curves that are possibly close to the real expression patterns. We observe that the hyperparameter allows flexible curve fitting, and the smoothing parameter adjusts the differences of gene expression patterns effectively.

The linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA) are the most popular classical method for discriminant analysis. We compare FLDA evaluated by the criterion $\mathrm{GIC}_{\mathrm{F}}$ with LDA and QDA which analyze discretized data directly. For almost all combinations of the classes, the proposed method yields a lower test error. We suggest investigating genes that were classified in the opposite group with high posterior probability, since they may have been misclassified by [10].

5 Conclusion

The functional logistic discriminant analysis proposed in this paper appears to be a useful tool for classifying functions or curves. An advantage of our method is that one could treat the samples as a set of functions, hence the problems of the observational point difference and highly correlated data are overcome. Also the model selection criterion enables us to evaluate models subjectively. Potential research would be extending our modeling strategy to the case of sampled surface for multi-group classification.

References

[1] Ando, T. and Konishi, S. (2002). *Nonlinear regression modeling via regularized radial basis function networks.* The Institute of Statistical Mathematics, Research Memorandum **845**.

[2] Cardot, H., Faivre, R. and Goulard, M. (2003). *Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data.* J. of Applied Statist. **30**, 1185 – 1199.

[3] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning.* Springer-Verlag, New York.

[4] Konishi, S. and Kitagawa, G. (1996). *Generalised information criteria in model selection.* Biometrika **83**, 875 – 890.

[5] Luan, Y. and Li, H. (2003). *Clustering of time-course gene expression data using a mixed-effects model with B-splines.* Bioinformatics **19(4)**, 474 – 482.

[6] Marx, B.D. and Eilers, P.H.C. (1999). *Generalized linear regression on sampled signals and curves: A P-spline approach.* Technometrics **41**, 1 – 13.

[7] Moody, J. and Darken, C. J. (1989). *Fast learning in networks of locally-tuned processing units.* Neural Comp. **1**, 281 – 294.

[8] Poggio, T. and Girosi, F. (1990). *Networks for approximation and learning.* Proc. IEEE **78**, 1484 – 1487.

[9] Ramsay, J. O. and Silverman, B. W. (1997). *Functional data analysis.* Springer-Verlag, New York.

[10] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Bostein, D. and Futcher, B. (1998). *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol. Biol. Cell **9**, 3273 – 3297.

Acknowledgement: The authors would like to thank two anonymous reviewers for pointing out related references.

Address: Y. Araki, S. Konishi, Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-Ku, Fukuoka 812-8581, Japan
S. Imoto, Institute of Medical Science, University of Tokyo, 4-6-1
Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

E-mail: `yuko@math.kyushu-u.ac.jp`, `konishi@math.kyushu-u.ac.jp`, `imoto@ims.u-tokyo.ac.jp`

COMPSTAT'2004 Symposium

FRACTAL PECULIARITIES OF BIRTH AND DEATH

Sergei Arhipov

Key words: Fractals, chaos, differential equations.

COMPSTAT 2004 section: Data visualisation.

Abstract: This paper shows how to use differential branches of mathematics (mathematical analysis, theory of differential equations, theory of probability etc.) and informatics (object-oriented analysis and designing, imitative modelling, visual programming etc.) to model the real phenomena. The modelling of the size of population is be demonstrated by the factors of birth and death.

1 Introduction

When exploring real phenomena, the choice of a model is the most significant aspect of applied research. The same real phenomenon may be described from different points of view. This enables to create many abstract conceptions concerning the reality to explore. In order to develop and explore models, different branches of mathematics (mathematical analysis, theory of differential equations, theory of probability etc.) and informatics (object-oriented analysis and designing, imitative modelling, visual programming etc.) are used. The only possible way, how to acquire the modelling, is to demonstrate different phenomena and different models, describing them. Interrelationships among plants and animals and between them and their environment are the significant issue to be explored. The modelling of the number of population is determined by the factors of birth and death.

2 Maltus' model

Maltus' model is one of the first models of the dynamics of population number. It is named after its author and is expressed by $\frac{dx}{dt} = \alpha x$. The integration of equation provides solution $x = x_0 e^{\alpha t}$, where x_0 – the threshold value of x. Maltus' equation models the hypothesis "The change of the number of population is proportional to the number of already existing individuals". Variable x indicates the number of a particular biological population type within at a certain moment. Constant α (Maltus' constant) indicates the coefficient of the increase of population number. Maltus' discreet model we can obtain from the continuous model by means of conceiving the separate interval of time. The birth of new individuals occurs after the interval, equal to one unit. According to the derivative function

$$\frac{x(t+\Delta t) - x(t)}{\Delta t} = \frac{\Delta x}{\Delta t} \overset{\Delta t \to 0}{\to} \frac{dx}{dt} \,.$$

If we introduce symbol: $\Delta t = 1$, $x_n = x(t)$, $x_{n+1} = x(t + \Delta t) = x(t + 1)$, then discreet model will be expressed by:

$$\begin{cases} x_0 = \text{const} \\ x_{n+1} = x_n + \alpha x_n \end{cases} .$$

Maltus' discreet model of the unlimited increase of population means that the number x_{n+1} of every next generation consists of the number x_n of the previous generation and the number αx_n of new-born individuals. If x_0 is the given threshold number of individuals of population, the model provides the answer to the question – how many individuals will be in the biological population after certain discreet period of time.

3 Fractal visualization of Maltus' discreet model

If we formally replace the real symbols by complex numbers, we can view Maltus' discreet model in the complex plane. Maltus' discreet model in the complex plane C is the formal replacement of real variables, constants and coefficients by corresponding complex variables, constants and coefficients.

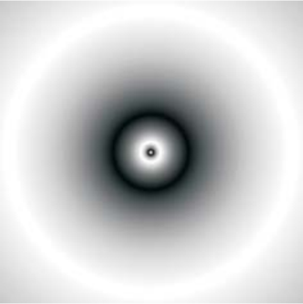


Figure 1:

Then the model

$$\begin{cases} Z_0 = C \\ Z_{n+1} = Z_n + AZ_n \end{cases}$$

may be viewed as iteration in the complex plane, but the results of such process may be described as a fractal portrait. The fractal portrait of iterative process $Z_{n+1} = Z_n + AZ_n$ is made in rectangular discrete area of a complex plane. For each point $Z_0 = C$ from this area process $Z_{n+1} = Z_n + AZ_n$ before performance of one of two conditions repeats. The first condition – whether it the new point Z_{n+1} has left the set area $|Z_{n+1}| < R$ of visibility whether or not. The second condition is a restriction on amount of iterations $n < N$. These two conditions guarantee the ending of iterative process. Or process will leave area of visibility, or process is conditionally infinite and it interrupts. The received number n is accepted for color value in a point

$Z_0 = C$, with which iterative process began. The fractal portrait is a set of integers n which are calculated in each point in rectangular discrete area of a complex plane. Thus, modulating the received digital values n in what or a color palette, the color image of fractal portrait is made. Change of ratio of color components of a palette allows considering features of process $Z_{n+1} = Z_n + AZ_n$.

4 Fractal visualization of Verhulst's model

According to Maltus' law, the number of population should increase exponentially. If we take into consideration that the population lives within the limited space (territory, resources etc.) and among the individuals there exists competition concerning the space, then we should take this factor into account within the model. Verhulst imparts the factor of the population number decrease to Maltus' model. Addend $(-\beta x^2)$ models the hypothesis "the decrease of the population number is proportional to the frequency of cases, when individuals meet each other". Then the equation of population number is expressed by

$$\frac{dx}{dt} = \alpha x - \beta x^2 \,.$$

If we take out of the brackets, we obtain Verhulst's equation

$$\frac{dx}{dt} = x(\alpha - \beta x \,,$$

where β – coefficient of the population decrease. Verhulst's discreet model is described by the recurrent ratio

$$\begin{cases} x_0 = \text{const} \\ x_{n+1} = x_n + x_n(\alpha - \beta x) \end{cases}$$

in the space of real numbers and

$$\begin{cases} Z_0 = C \\ Z_{n+1} = Z_n + Z_n \cdot (A - B \cdot Z_n) \end{cases}$$

in the space of complex numbers. Iteration of the reflection of Verhulst's model, if there are given different complex coefficients A and B, provides the set of fractals, corresponding to Julia's set. The solution of Verhulst's model is called logical function. Many models of birth and death in biology, economics, sociology are described by means of these functions. Perhaps, fundamental principles of increase enabled to reflect the likeness of the logical fractal portrait and the spirals of sun-flower, split into opposite directions, by means of Verhulst's model. Iterative process $Z_{n+1} = Z_n + Z_n \cdot (A - B \cdot Z_n)$ begins in a point $Z_0 = C$.

Amount of such computing processes is equal to amount of complex points on fractal portrait. The purpose of each process is calculation n of amount of

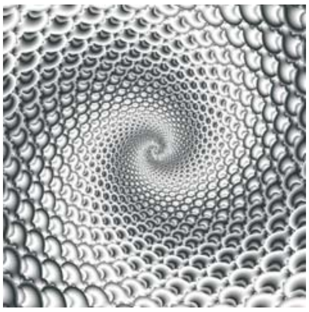


Figure 2:

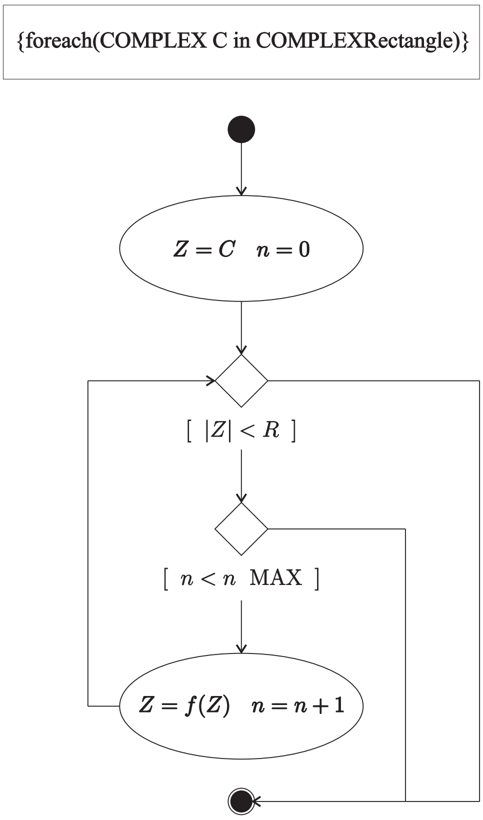

{foreach(COMPLEX C in COMPLEXRectangle)}
$Z = C \quad n = 0$
[$|Z| < R$]
[$n < n$ MAX]
$Z = f(Z) \quad n = n + 1$

Figure 3:

iterations. If iterative process constantly generates points Z_{n+1} in the field of visibility $|Z_n + Z_n \cdot (A - B \cdot Z_n)| < R$, then such process is considered infinite and the variable $n = N$ gives conditionally big number. If iterative process leaves area of visibility, that is at some value $n < N$ the condition $R \leq |Z_n + Z_n \cdot (A - B \cdot Z_n)|$ is true, then the variable n gives number of the executed iterations be satisfied. The integer n received thus sets color value of fractal portrait in a point $Z = C$ of a complex plane.

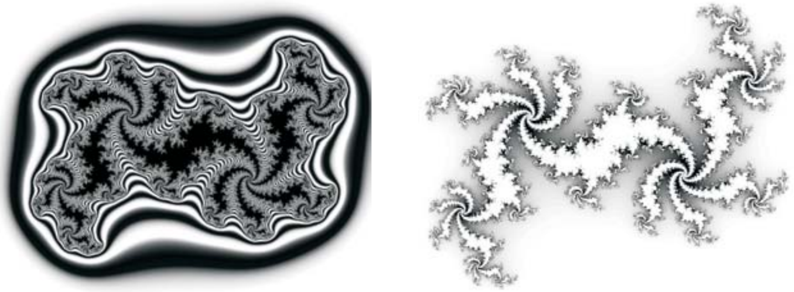


Figure 4: This is a fractal portrait and Julia's set of Verhulst's complex model with the given coefficients A=(0.283;1) and B=(0.5;0.5).

5 Fractal portrait of Volterra model

In Voltera model the problem becomes two-dimensional. There are two populations in the space – predators and their preys. Variable x indicates the number of preys, but variable y indicates the number of predators within the system. Equation of variable x – the number of preys is expressed by:

$$\frac{dx}{dt} = \alpha x - \beta x y\,.$$

The first addend αx, like in Maltus' model, means that the increase of the number of individuals is proportional their amount at a certain moment.

The second addend $(-\beta x y)$ means that the decrease of the population number is proportional to the frequency of cases, when the preys meet predators. If we take x out of the brackets, we obtain the equation for preys:

$$\frac{dx}{dt} = x(\alpha - \beta x)\,.$$

Equation of variable y – the number of dead predators is expressed by:

$$\frac{dy}{dt} = -\gamma y + \delta x y\,.$$

The first addend $(-\gamma y)$, means that the predators die if there is lack of prey. The second addend $\delta x y$ means that the increase of the population of predators is proportional to the number of cases, when the predator meets the prey. If we take out of the brackets, we obtain the equation for predators:

$$\frac{dy}{dt} = -y(\gamma - \delta x)\,.$$

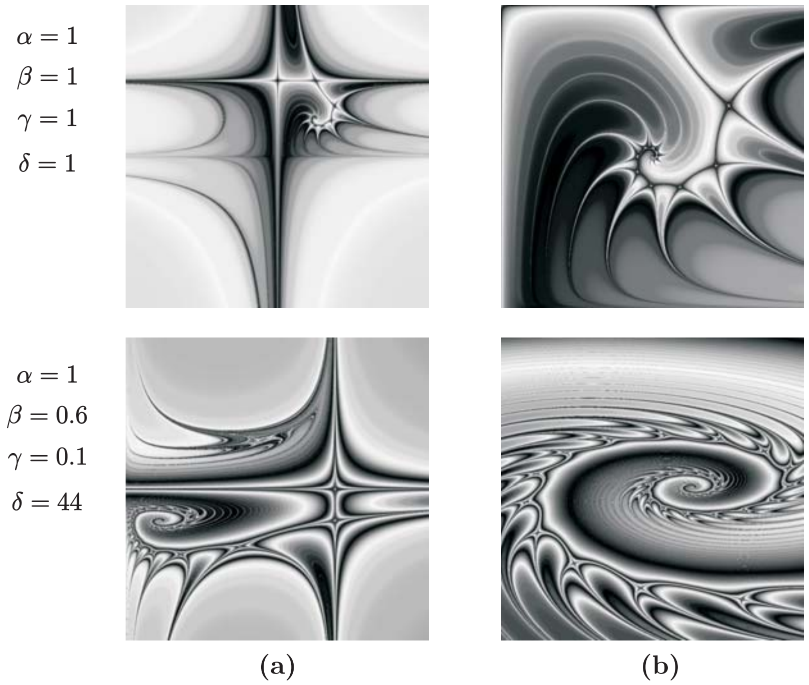


Figure 5: **(a)** Fractal portrait of Volterra model with the given parameters. **(b)** The enlarged centre of the spiral of left portrait.

Thus we obtain the non-linear system of differential equations:

$$\begin{cases} \frac{dx}{dt} = x(\alpha - \beta x) \\ \frac{dy}{dt} = -y(\gamma - \delta x)\,, \end{cases}$$

where α – the increase coefficient of the prey population number, β – the decrease coefficient of the prey population number, when the preys are eaten by the predators, γ – the decrease coefficient of the predators population number if there is a lack of preys. δ – the increase coefficient of the predators population number, when the predators eat preys. We obtain Volterra discreet model by replacing the derivative function by its discreet analogue and assuming that the period of time is a unit:

$$\begin{cases} x_0 = c_x \\ y_0 = c_y \\ x_{n+1} = x_n + x_n(\alpha - \beta x_n) \\ y_n = y_n - y_n(\gamma - \delta y_n) \end{cases}$$

Volterra model may be viewed as the one-dimensional reflection of complex plane in relation to oneself:

$$\begin{cases} Z_0 = C \\ Z_{n+1} = f(Z_n) = f(Z_n)\,, \end{cases}$$

where $Z_n = (x_n; y_n)$ and $C = (c_x; c_y)$ are the points in the complex plane.

Julia's set is the result of iteration of the complex plane reflections and enables to visualise the behaviour of the non-linear dynamic systems, developing a series of fractal portraits.

6 Conclusions

The present paper deals with some examples of the models of the populations dynamics and shows corresponding fractal portraits of systems. Some fractal portraits of systems include Julia's set. Some Julia's sets have a regular boarder, but some sets are structured fractually. The basic peculiarity of the fractal portraits of dynamic systems is the fact that the recursive procedure of the development of fractals naturally arise from the recurrent interrelations of corresponding discreet models. Therefore it is common to analyse the phase's portraits of systems together with their fractal peculiarities.

References

[1] Mandelbrot B.B. (2002). *The fractal geometry of nature.* W. H. Freedman and Company, New York.

[2] Arditi R. (2001). *Directed movement of predators and the emergence of density-dependence in predator_Prey Models.* Theoretical Population Biology **59**, 207−221.

Address: Arhipov S., Faculty of Information Technologies, Latvia University of Agriculture, 2 Liela street, Jelgava, LV-3001, Latvia

E-mail: arx@cs.llu.lv

THE PROBLEM OF CHOOSING STATISTICAL HYPOTHESES IN APPLIED STATISTICS

Irina Arhipova and Signe Balina

Key words: Statistical methods, hypotheses.

COMPSTAT 2004 section: Teaching statistics.

Abstract: Based on our experience at Latvia University of Agriculture and University of Latvia, we would like to present views on how to teach statistics for undergraduate students in economics, management science and related disciplines. The approach of designing statistics' course depends both on the type of the course and on the students. Teaching statistics to graduate students or others, certain balances have to be observed and maintained. One of them is the balance between the theory and practice, because if theoretical materials take up too much time and energy the students will lose interest. Therefore there should be another balance - the balance between mathematical statistics, subject of statistical modelling and software usage. The goal of the statistical course for graduate is to acquaint students with a variety of problems with which they might encounter in their studies during which each problem requires students to collect "real-world" data and to combine several statistical methods. The interrelation of the statistical methods and statistical hypothesis had been considered in the statistics' course.

1 The tasks of teaching statistics at universities

The statisticians have three tasks at universities: teaching statistics, carrying out research in statistics and consulting in statistics. In this paper we discuss the task of teaching. It is very difficult for students to become familiar with a wide range of the statistical methods and statistical hypotheses. The one of the problems in the teaching process is how to show classification of the statistical methods and interrelation of statistical hypotheses. The first step is the identification of the number of variables and their scale. The next step is the definition of dependent and independent variables. The four possibilities according to the type and measurement scale of variables are as follows:

- The number of independent variables – one or more than one.
- The number of dependent variables – one or more than one.
- The type of measurement scale used for the dependent variables (i.e. metric or non-metric).
- The type of measurement scale used for the independent variables (i.e. metric or non-metric).

The third step is the definition of appropriate hypothesis (null and alternative) and the choosing of appropriate statistical method.

Introducing this classification there is no problem to define univariate and multivariate methods in teaching. For example, for one non-metric independent variable and one metric dependent variable the appropriate method is t-test but for more than one non-metric independent variable and one metric dependent variable the appropriate method is analysis of variance (ANOVA). At the same time for one metric independent variable and one metric dependent variable the appropriate method is regression analysis. And for more than one metric independent variable and one metric dependent variable the appropriate method is multivariate regression analysis.

However there are such data sets for which it is impossible to conceptually designate which set of variables is dependent, and which - independent. For these types of data sets the objectives are to identify how and why the variables are related among themselves. Statistical methods for analysing these types of data sets are called *interdependence methods*, for example, cluster analysis, factor analysis for more than two metric variables and multiway contingence tables for more than two non-metric variables.

Only after the regularities of statistical hypotheses have been acquired, the statistical software packages can be used in problem solving, because the unsophisticated use of the applied statistical packages hinders students' deep acquisition of the statistics.

2 Interrelation of statistical hypotheses

Teaching statistics different topics are taught separately without emphasising the fact that all the methods are closely interrelated, and, if the number or the types of the variables change, the method to be applied also changes. The students usually have difficulties to classify the methods accordingly to the information provided by the "real-world" problem and consequently they also have difficulties to choose the hypothesis to be verified. Teaching a new topic it is indispensable to show the link between the already acquired methods and this new topic. The interrelations between statistical methods are easy to show, if one "real-world" problem is used and developed. The first step is to word the null and alternative hypotheses of the problem, the second – to transform the hypotheses into mathematical symbols.

Let us consider the interrelation of statistical hypothesis using the hypotheses about the population mean. The first of them is T-*test* $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. For example, suppose that statistician is interested in determining, whether the mean salary per month for inhabitants of Latvia is equal to EUR 500 or not: $H_0 : \mu = 500$ versus $H_1 : \mu \neq 500$.

The extension of these hypotheses is two-population hypothesis test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. For example, suppose you need to compare the equality of the mean salaries per month for inhabitants at the age below and over 40 years: $H_0 : \mu_{<40} = \mu_{>40}$ versus $H_1 : \mu_{<40} \neq \mu_{>40}$.

Let us suppose you need to compare the mean salary per month for inhabitants in the following groups of age: below 25 years, from 26 to 40 years, from 41 to 55 years, over 55 years. In this case more than two population means must be compared, and the method of one-way analysis of variance (ANOVA) is appropriate: $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_a$ versus H_1: *not all populations have the same mean*, where a is the number of groups, or: $H_0 : \mu_{<25} = \mu_{(25-40)} = \mu_{(41-55)} = \mu_{>55}$ versus H_1: *not all mi are equal*.

This ANOVA test is the same as F-test: $H_0 : F = 0$ versus $H_1 : F > 0$, where $F = \frac{s_a^2}{s_e^2}$ and s_a^2 is between method estimate of σ^2 and s_e^2 – within method estimate of σ^2. The null hypothesis means that the factor (age) is not significant, and the alternative hypothesis – that the factor is significant.

Let us consider the age factor as the continuous factor. In this case the factor significance must be analysed by correlation analysis or one-factor regression analysis. Using the hypotheses: $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ the correlation between depended and independent variables can be defined, in our case, between salary and age. Either the hypotheses $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ or $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ are the same for one-factor correlation and regression analysis, where β_1 is the slope coefficient of the linear regression model. The first hypothesis states that there is no correlation between age and salary. The second hypothesis states that if a regression line is fitted to the population salary – age data, this line will be horizontal, i.e., it will have a slope of zero. In others words, salary and age have no correlation in the population. If one of these hypotheses is true, the other also is. In fact, if these two null hypotheses are tested separately, t statistic and test conclusions will be the same.

Another key statistics in regression analysis, the F statistics, is used to test the null hypothesis that the sample regression equation does not explain a significant percentage of the dependent variable's variance. The null and alternative hypotheses are $H_0 : \rho^2 = 0$ versus $H_1 : \rho^2 > 0$.

Thus, the F test can be used to determine the existence of a linear relationship between independent and dependent variables. In the simple linear regression, this test is equivalent to the t test: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

Therefore, for one factor regression analysis, you may conduct either F test or t test. The results of the two tests will lead to the same conclusions. The statistical methods and hypotheses interrelation is shown in figure 1.

Let us suppose that a new variable – inhabitants' work experience (in years) – has been added to the problem. Thus a hypothesis to be verified is set up – whether the salary depends on the inhabitant's age and work experience. The most appropriate method to verify the chosen hypothesis is the multiple regression. Because there is more than one explanatory variable, the null and alternative hypotheses can be set up as follows: $H_0 : \beta_1 = \beta_2 = 0$ versus H_1: at least one $\beta_j \neq 0$ where null hypothesis means that there is no linear relationship between the salary and explanatory variables (age and

$$\begin{tikzcd}
 & & & \textit{One-factor regression and correlation analysis} & \\
 & & & \begin{matrix} H_0:\beta_1=0 \\ H_1:\beta_1\neq 0 \end{matrix} \arrow[d, Leftrightarrow] & \begin{matrix} H_0:\rho=0 \\ H_1:\rho\neq 0 \end{matrix} \arrow[l, Rightarrow] \arrow[d, Rightarrow] \\
\textit{T-test} & & & & \\
\begin{matrix} H_0:\mu=\mu_0 \\ H_1:\mu\neq\mu_0 \end{matrix} \arrow[r, equal] & \begin{matrix} H_0:\mu_1=\mu_2 \\ H_1:\mu_1\neq\mu_2 \end{matrix} \arrow[r, Rightarrow] & \begin{matrix} H_0:\mu_1=\mu_2=\cdots=\mu_a \\ H_1 \textit{ not all } \mu_i \textit{ are equal} \end{matrix} \arrow[r, Rightarrow] & \begin{matrix} H_0:F=0 \\ H_1:F>0 \end{matrix} & \begin{matrix} H_0:\rho^2=0 \\ H_1:\rho^2>0 \end{matrix} \arrow[l, Rightarrow] \\
 & & \textit{One-way analysis of variance (ANOVA)} & &
\end{tikzcd}$$

Figure 1: The interrelation of statistical hypotheses.

work experience), but alternative hypothesis means that there is a linear relationship between the salary and at least one of the explanatory variables.

Let us add a new variable – inhabitant's gender – to the data base. Thus a new hypothesis to be verified is set up – whether the salary depends on the gender, age and work experience. In the particular case there are both qualitative and quantitative variables are among the independent variables. The chosen hypothesis can be verified by means of CANOVA (covariance analysis of variance), which is a combination of the regression analysis and the analysis of variance. If the goal of the problem is to verify, whether the salary depends on gender and age, the simple linear covariance model is valid: $Y_{ij} = \mu+\alpha_i+\beta X_{ij}+\varepsilon_{ij}$ where μ is the overall mean effect, α_i is the true effect of the gender and β is a slope parameter associated with the independent factor of age. If the goal of the problem is to verify, whether the salary depends on gender, age and work experience, the multiple linear covariance model is valid: $Y_i = \mu + \alpha_i + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \varepsilon_{ij}$ where β_1 is a slope parameter associated with the age and β_2 is a slope parameter associated with the work experience. If the goal of the problem is to verify, whether the salary and work experience depend on gender, then the appropriate method is the multivariate analysis of variance (MANOVA) that combines results from the several ANOVAs.

Besides, if the goal of the problem is to analyse, whether the salary and work experience depend on gender and age, then the appropriate method is the multivariate covariance analysis of variance (MCANOVA). As we see, it is possible to show different methods and their interrelation on the basis of a particular example.

$$\begin{tikzcd}
 & \text{ANOVA} & & \text{MANOVA} \\
T\text{-test} \arrow[ur, Rightarrow] \arrow[dr, Rightarrow] & \arrow[r, Rightarrow] & \text{CANOVA} \arrow[ur, Rightarrow] \arrow[dr, Rightarrow] & \\
 & \text{Regression} & & \text{MCANOVA}
\end{tikzcd}$$

Figure 2: The example of the interrelation of statistical methods.

In order to help the students to classify the methods accordingly to the information provided by the problem, the classification of the statistical methods is provided in the table below.

			Dependent variable			
			One		**More than one**	
			Quantitative	**Qualitative**	**Quantitative**	**Qualitative**
Independent variable	**One**	**Quantitative**	**Univariate methods**			
			Simple regression	Two-group discriminant analysis	Canonical correlation	Multi-group discriminant analysis
		Qualitative	T-test	Discrete discriminant analysis	Multivariate analysis of variance	Discrete multi-group discriminant analysis
	More than one	**Quantitative**	Multiple regression	Two-group discriminant analysis	Canonical correlation	Multi-group discriminant analysis
		Qualitative	Analysis of variance, ANOVA	Discrete discriminant analysis	Multivariate analysis of variance	Discrete multi-group discriminant analysis

Table 1: The classification of the statistical methods.

3 The problem-orientated statistical method choice

For the illustration let us consider the typical problem in statistics' course, namely, let us consider data about average gross wages per month by sex and occupation in Latvia. The problems can be defined as follows:

- Is there a significant difference in the average gross wages for males and females?
- Is there a significant difference in the average gross wages for the various categories of occupation?
- Is there a significant relation between male and female gross wages?

Students should define the independent and dependent variables and their scale. For example, in the first problem factor sex is qualitative independent variable with two categories, and gross wages is quantitative dependent variable. In this case the t-test is the appropriate method for the statistical hypothesis for the proving the significant difference between the average wages for males and females. For the second problem factor occupation is qualitative independent variable with various categories, and gross wages is quantitative dependent variable. So, the analysis for variance (ANOVA) is appropriate method for the statistical hypothesis for proving the difference between the average wages for the various categories of occupation. For the last problem there are two quantitative variables: male's wage and female's wage, where the simple regression analysis allows defining a relation between wages of males and females.

4 The use of statistical packages in teaching statistics

After the acquisition of theoretical course, the students are ready to use the acquired knowledge to solve a problem, using ready-made statistical software, for example, SPSS, Minitab and Microsoft Excel Data Analysis Tools. The general linear model (GLM) is flexible statistical model which incorporates analyses involving normally distributed dependent variables and combinations of categorical and continuous factor variables. The SPSS GLM procedure can accommodate univariate models (one dependent variable) involving:

- Categorical predictors only (ANOVA)
- Continuous predictors only (Regression)
- Combinations of categorical predictors and continuous predictors (CANOVA)

as well as multivariate versions (two or more dependent variables) of all of these models (MANOVA and MCANOVA). Categorical predictors are referred to as factors, while continuous predictors are called covariates. The application of statistical packages allows the students to obtain the result quickly, as well as help them better understand the interrelation between the statistical hypotheses. Unfortunately, the simple data processing without profound understanding and interpretation of obtained results does not allow achieving a real result in problem solving.

Therefore teaching statistics much attention should be paid to the methodology of statistical analysis:

- The definition of a hypothesis.
- The specification of the hypothesis mathematical model.
- The specification of the statistical model.
- Data obtaining.
- The evaluation of the statistical model parameters.
- The verification of hypotheses.
- The interpretation of obtained results.

References

[1] Christensen R. (1996). *Analysis of variance, design and regression.* Chapman & Hall.

[2] Olsson U. (1999). *Teaching and examination of statistics using real-world problems.* Second Nordic-Baltic Agrometric Conference, Karaski, Estonia, 1 – 6.

[3] Ostle B., Mensing R. W. (1997). *Statistics in research.* The Iowa State University Press/AMES.

[4] Sharma S. (1996). *Applied multivariate techniques.* John Wiley & Sons, Inc.

Address: I. Arhipova, Faculty of Information Technologies, Latvia University of Agriculture, 2 Liela street, Jelgava, LV-3001, Latvia
S. Balina, Faculty of Economics and Management, University of Latvia, 5 Aspazijas blvd., Riga, LV-1050, Latvia

E-mail: `irina@cs.llu.lv, signe@lanet.lv`

REDUCING THE BIAS OF THE LOG-PERIODOGRAM REGRESSION IN PERTURBED LONG MEMORY SERIES

Josu Arteche

Key words: Long memory, log periodogram, stochastic volatility.

COMPSTAT 2004 section: Time series analysis.

Abstract: This paper proposes an extension of the log periodogram regression which explicitly accounts for the added noise in Long Memory in Stochastic Volatility (LMSV) and other perturbed long memory time series. Contrary to the non linear log periodogram regression of [8], no linear approximation of the logarithmic term which accounts for the noise is used. This produces a reduction of the bias and increases the asymptotic efficiency in long memory signal plus noise series. Asymptotic and finite sample properties of the estimator are analyzed. Finally an application to the Spanish stock index Ibex35 is included.

1 Introduction

The estimation of the memory parameter in perturbed long memory series has become a subject of increasing interest motivated especially by the strong persistence found in the volatility of many financial and economic series. Alternatively to the different extensions of ARCH and GARCH models, the Long Memory in Stochastic Volatility (LMSV) has proved a useful tool to model such a strong persistent volatility. Estimation of the memory parameter of the volatility component in LMSV corresponds to a problem of estimation in a long memory signal plus noise model.

The perturbed long memory series recently considered in the literature are of the form,

$$z_t = \mu + y_t + u_t \tag{1}$$

where μ is a finite constant, y_t is a long memory (LM) process such that its spectral density satisfies

$$f_y(\lambda) = C\lambda^{-2d}(1 + O(\lambda^{\alpha})) \qquad \text{as } \lambda \to 0 \tag{2}$$

for a positive finite constant C, $\alpha \in [1, 2]$ and $0 < d < 0.5$, and u_t is a weakly dependent process. The condition of positive memory $0 < d < 0.5$ guarantees the asymptotic equivalence between spectral densities of y_t and z_t.

The version of [6] of the log periodogram regression estimator (LPE) in a fully observable LM series is based on the least squares regression $\log I_{yj} = a + d(-2\log\lambda_j) + v_j$, $j = 1, \ldots, m$, where $I_{yj} = I_y(\lambda_j) = |\sum_{t=1}^{n} y_t \exp(-i\lambda_j t)|^2/(2\pi n)$ is the periodogram of the series y_t, $t = 1, \ldots, n$, at

Fourier frequency $\lambda_j = 2\pi j/n$ and m is the bandwidth such that at least $m^{-1} + mn^{-1} \to 0$ as $n \to \infty$. The motivation of this estimator is the log linearization in (2) such that

$$\log I_{yj} = a + d(-2\log\lambda_j) + U_{yj} + O(\lambda_j^{\alpha}), \quad j = 1, 2, \ldots, m, \tag{3}$$

where $a = \log C - c_0$, $c_0 = 0.5772\ldots$ is Euler's constant and $U_{yj} = \log(I_{yj} f_y^{-1}(\lambda_j)) + c_0$. The bias of the least squares estimate of d is dominated by the $O(\lambda_j^{\alpha})$ term which is not explicitly considered in the regression such that a bias of order $O(\lambda_m^{\alpha})$ arises [4].

The main rival semiparametric estimator of the LPE is the local Whittle or Gaussian semiparametric estimator (GSE) of [7] which has the computational disadvantage of requiring nonlinear optimization but it is more efficient than the log periodogram regression. However both share important affinities. In particular the bias is in both cases of order $O(\lambda_m^{\alpha})$.

Both estimators preserve the consistency and asymptotic normality when applied to perturbed long memory series [3] and [1]. In this case

$$f_z(\lambda) = C\lambda^{-2d}\left(1 + \frac{f_u(\lambda)}{C}\lambda^{2d} + O(\lambda^{\alpha})\right) \quad \text{as } \lambda \to 0 \tag{4}$$

with $f_u(\lambda)$ positive and bounded. The leading term of the bias is in both estimators $f_u(\lambda)C^{-1}\lambda^{2d}$ which is the dominant part not considered explicitly in the estimation. Then the bias is of order $O(\lambda_m^{2d})$ which can be quite severe, especially if d is low. Correspondingly the asymptotic normality requires at least $m^{1+4d}n^{-4d} \to 0$ as $n \to \infty$ which limits the size of the bandwidth and consequently the asymptotic efficiency of both estimators.

In order to reduce the bias of the GSE [5] suggested to incorporate explicitly a $\beta\lambda_j^{2d}$ term in the estimation to take into account the effect of the added noise on the spectral density of z_t and proposed a modified Gaussian semiparametric estimator (MGSE) defined as

$$(\tilde{d}_M, \tilde{\beta}_M) = \arg\min_{\Delta\times\Theta}\left[\log\left(\frac{1}{m}\sum_{j=1}^{m}\frac{\lambda_j^{2d} I_{yj}}{1+\beta\lambda_j^{2d}}\right) + \frac{1}{m}\sum_{j=1}^{m}\log\{\lambda_j^{-2d}(1+\beta\lambda_j^{2d})\}\right]$$

where $\Theta = (0, \Theta_1)$, $\Theta_1 < \infty$, $\Delta = [\Delta_1, \Delta_2]$, $0 < \Delta_1 < \Delta_2 < 1/2$. The upper bound in the bandwidth is relaxed now to comply $m^{1+2\alpha}(\log m)^2/n^{2\alpha} \to 0$ as $n \to \infty$ which allows a gain in asymptotic efficiency. In the fractional ARIMA processes the MGSE achieves a rate of convergence arbitrarily close to $n^{2/5}$ which is the upper bound of the rate of convergence of the Gaussian semiparametric estimator in the absence of additive noise. However with an additive noise the best possible rate of convergence achieved by the GSE is $n^{2d/(4d+1)}$. Regarding the bias, the MGSE has a bias of order $O(\lambda_m^{\alpha})$ instead of $O(\lambda_m^{2d})$ which is the bias of the GSE in the presence of an additive noise.

Sun and Phillips [8] extended the log periodogram regression in a similar manner. From (4) with $f'_u(0) = 0$

$$\begin{aligned} \log I_{zj} &= \log C - c_0 + d(-2\log\lambda_j) + \log\left(1 + \frac{f_u(0)}{C}\lambda_j^{2d}\right) + O(\lambda_j^{\alpha}) + U_{zj} \qquad (5) \\ &= \log C - c_0 + d(-2\log\lambda_j) + \frac{f_u(0)}{C}\lambda_j^{2d} + O(\lambda_j^{\alpha^*}) + U_{zj} \end{aligned}$$

where $\alpha^* = \min(4d, \alpha)$. Sun and Phillips [8] proposed a non linear log periodogram regression $\log I_{zj} = a + d(-2\log\lambda_j) + \beta\lambda_j^{2d} + U_{zj}$ for $\beta = f_u(0)/C$, such that the non linear log periodogram regression estimator (NLPE) are

$$(\hat{a}, \hat{d}, \hat{\beta}) = \arg\min \sum_{j=1}^{m} (\log I_{zj} - a + d(2\log\lambda_j) - \beta\lambda_j^{2d})^2 \qquad (6)$$

The bias of $\hat{d}$ is of order $O(\lambda_m^{\alpha^*})$ which is produced by the $O(\lambda_j^{\alpha^*})$ omitted in the regression in (6). Correspondingly the upper bound of m for the asymptotic normality is $O(n^{2\alpha^*/(2\alpha^*+1)})$. Sun and Phillips [8] only consider the case $\alpha = 2$ so that $\alpha^* = 4d$ and the behaviour of m is restricted to be $O(n^{8d/(8d+1)})$ with a bias of order $O(\lambda_m^{4d})$, but the extension to $\alpha < 2$ is straightforward. The asymptotic efficiency of the NLPE is higher than in the standard LPE but lower than the asymptotic efficiency of the MGSE when $\alpha > 4d$. The reason of this behaviour is the approximation of the log expression in (5). After this modification the regression model of [8] is still non linear and does not imply any computational advantage. Instead, noting (5) we propose the following non linear regression model

$$\log I_{zj} = a + d(-2\log\lambda_j) + \log(1 + \beta\lambda_j^{2d}) + U_{zj} \qquad (7)$$

which only leaves an $O(\lambda_j^{\alpha})$ term out of explicit consideration.

2 Asymptotic bias of the periodogram

Consider the following assumptions:

A.1: z_t in (1) is a long memory signal plus noise process with y_t an LM process with spectral density function in (2) with $d < 0.5$ and u_t is stationary with positive and bounded continuous spectral density function $f_u(\lambda)$.

A.2: y_t and u_t are independent.

Theorem 2.1. *Let z_t satisfy assumptions A.1 and A.2 and define $L_j(d) = E\left[\frac{I_{zj}}{C\lambda_j^{-2d}}\right]$. Then, considering j fixed $L_j(d) = A_{1j} + A_{2j} + o(n^{-2d})$ where*

$$\lim_{n\to\infty} A_{1j} = \int_{-\infty}^{\infty} \psi_j(\lambda) \left|\frac{\lambda}{2\pi j}\right|^{-2d} d\lambda\,, \quad \lim_{n\to\infty} n^{2d} A_{2j} = \int_{-\infty}^{\infty} \psi_j(\lambda) \frac{f_u(0)}{C(2\pi j)^{-2d}} d\lambda$$

with $\psi_j(\lambda) = \frac{2}{\pi}\frac{\sin^2\frac{\lambda}{2}}{(2\pi j - \lambda)^2}$.

Remark 1: In the LMSV case $f_u(0) = \sigma_\xi^2/2\pi$. The influence of the noise is clear here, the larger the variance of the noise the higher the relative bias of the periodogram. This explains the high bias of semiparametric estimates in LMSV models under a low signal to noise ratio in [2] and [1].

Remark 2: When $d < 0$ the bias increases without limit as n increases. This justifies the difficulties encountered when estimating a negative d in perturbed long memory series [3] and [1].

Theorem 2.2. *Let z_t satisfy assumptions A.1 and A.2, and consider a sequence of positive integers $j = j(n)$ such that $j/n \to 0$ as $n \to \infty$. Then*

$$L_j(d) = 1 + O\left(\frac{\log j}{j} + \lambda_j^{\min(\alpha,2d)}\right)$$

3 The ALP estimator

The augmented log periodogram estimator (ALPE) is defined as

$$(\tilde{a}, \tilde{d}, \tilde{\beta}) = \arg\min \sum_{j=1}^{m} (\log I_{zj} - a + d(2\log\lambda_j) - \log(1 + \beta\lambda_j^{2d}))^2 \qquad (8)$$

Concentrating the constant a out

$$(\tilde{d}, \tilde{\beta}) = \arg\min_{\Delta\times\Theta} \sum_{j=1}^{m} (\log I_{zj}^* + d(2\log\lambda_j)^* - \log^*(1 + \beta\lambda_j^{2d}))^2 \qquad (9)$$

where for a general ξ_t we use the notation $\xi_t^* = \xi_t - \bar{\xi}$ where $\bar{\xi} = \sum \xi_t/n$.

The first order conditions[1] of this minimization problem are

$$S(\tilde{d}, \tilde{\beta}) = 0 \qquad \text{where} \qquad S(d,\beta) = \sum_{j=1}^{m} \begin{pmatrix} x_{1j}^*(d,\beta) \\ x_{2j}^*(d,\beta) \end{pmatrix} W_j(d\beta)$$

with

$$\begin{aligned} x_{1j}(d,\beta) &= 2\left(1 - \frac{\beta\lambda_j^{2d}}{1+\beta\lambda_j^{2d}}\right)\log\lambda_j \,, \quad x_{2j}(d,\beta) = -\frac{\lambda_j^{2d}}{1+\beta\lambda_j^{2d}} \,, \\ W_j(d,\beta) &= \log I_{zj}^* + d(2\log\lambda_j)^* - \log^*(1+\beta\lambda_j^{2d}) \end{aligned}$$

Let d_0 be the true unknown memory parameter and d any admissible value and consider the same notation for the rest of parameters to estimate. Define the diagonal matrix $D_n = diag(\sqrt{m}, \lambda_m^{2d_0}\sqrt{m})$ and the matrix

$$\Omega = \begin{pmatrix} 4 & -\frac{4d_0}{(2d_0+1)^2} \\ -\frac{4d_0}{(2d_0+1)^2} & \frac{4d_0^2}{(4d_0+1)(2d_0+1)^2} \end{pmatrix}$$

[1] If we allow $\beta = 0$ a similar result (available upon request) to that in [8] is achieved.

Consider the following assumptions:

B.1: y_t and u_t are independent Gaussian processes.

B.2: $f_u(\lambda)$ is continuous on $[-\pi, \pi]$, bounded above and away from zero with bounded second derivative in a neighbourhood of zero.

B.3: The spectral density of y_t satisfy

$$f_y(\lambda) = C\lambda^{-2d}(1 + E\lambda^{\alpha} + o(\lambda^{\alpha}))$$

for some finite E and $\alpha \in (4d_0, 2]$.

B.4: For some positive constant K, as $n \to \infty$,

$$\frac{m^{2\alpha+1}}{n^{2\alpha}} \to K.$$

Assumption B.1 is quite severe and excludes LMSV models where u_t is not Gaussian but a log chi-square. Considering recent results, Gaussianity of signal and noise could be relaxed. However this would significantly complicate the technical details of the proofs and we prefer to keep the technical requirements to a minimum. Assumption B.2 restricts the behaviour of u_t and B.3 imposes a particular spectral behaviour of y_t around zero. This local specification permits to obtain the asymptotic bias of $\tilde{d}$ in terms of E. We restrict our analysis to the case $\alpha > 4d_0$ where the ALPE achieves a lower bias and higher asymptotic efficiency than the NLPE. In the standard fractional ARIMA process as considered in [8] $\alpha = 2$. We consider also $\alpha < 2$ which may be relevant in some situations, and permit a direct extension to the seasonal or cyclical long memory case. Assumption B.4 restricts the behaviour of the bandwidth m in a similar manner as in [8] but allowing a larger m.

Theorem 3.1. *Let z_t in (1) satisfy assumption B.1-B.3 and m satisfy B.4. Then as $n \to \infty$*

$$D_n \begin{pmatrix} \tilde{d} - d_0 \\ \tilde{\beta} - \beta_0 \end{pmatrix} \xrightarrow{d} N\left(\Omega^{-1}b, \frac{\pi^2}{6}\Omega^{-1}\right)$$

where

$$b = (2\pi)^{\alpha} K 2 \begin{pmatrix} -\frac{\alpha}{(1+\alpha)^2} \\ \frac{\alpha d_0}{(2d_0+\alpha+1)(2d_0+1)(1+\alpha)} \end{pmatrix} E.$$

The asymptotic bias of $\tilde{d}$ is then

$$ABias(\tilde{d}) = \left(\frac{m}{n}\right)^{\alpha} K_0 \quad \text{where} \quad K_0 = \frac{(2\pi)^{\alpha}\alpha(2d_0+1)(\alpha - 2d_0)E}{4d_0(1+\alpha)^2(2d_0+\alpha+1)}$$

In contrast to the LPE and NLPE, the ALPE $\tilde{d}$ has an asymptotic positive bias which decreases with d_0. The asymptotic variance and mean square error are

$$
\begin{aligned}
AVar(\tilde{d}) &= \frac{\pi^2}{24m}C_d \quad \text{where} \quad C_d = 1 + \frac{1+4d_0}{4d_0^2} \\
AMSE(\tilde{d}) &= \frac{\pi^2}{24m}C_d + \left(\frac{m}{n}\right)^{2\alpha} K_0^2.
\end{aligned}
$$

The optimal bandwidth, in an asymptotic MSE sense, is

$$
m_{opt} = \left(\frac{\pi^2 C_d}{48\alpha K_0^2}\right)^{\frac{1}{2\alpha+1}} n^{\frac{2\alpha}{2\alpha+1}}.
$$

The optimal bandwidth of the ALPE increases with n faster than the corresponding m_{opt} of the NLPE. Correspondingly the AMSE($\tilde{d}$) converges to zero at a rate $n^{-2\alpha/(2\alpha+1)}$ which is faster that the $n^{-4d_0/(4d_0+1)}$ obtained with the optimal m in the LPE and if $\alpha > 4d_0$ (as in the $\alpha = 2$ case) it is faster than the $n^{-8d_0/(8d_0+1)}$ rate achieved by the NLPE with an optimal m.

4 Finite sample performance

We compare the finite sample performance of the LPE and NLPE with the ALPE in a LMSV

$$
z_t = y_t + u_t
$$

for $(1-L)^{0.45}y_t = w_t$ and $u_t = \log \varepsilon_t^2$, for ε_t and w_t independent, ε_t is standard normal and $w_t \sim N(0, \sigma_w^2)$ for $\sigma_w^2 = 0.5, 0.1$. These values correspond to long run noise to signal ratios $f_u(0)/f_w(0) = \pi^2$, $5\pi^2$. The first value is close to the values considered in [3] and [8]. The second one corresponds more closely to the values found in financial time series. Since ε_t is standard normal, u_t is a $\log \chi_1^2$ and consequently assumption B.1 does not hold. However we consider relevant to show that the ALPE can be applied in LMSV models which are an essential tool in the modelling of financial time series, and justify in that way our conjecture of no necessity of Gaussianity of the added noise.

The Monte Carlo is carried out in SPlus 2000, generating y_t with the option arima.fracdiff.sim and for the non linear optimization we use nlminb for $0.0001 < d < 0.7$ providing the gradient and the hessian. We just consider a sample size of $n = 1024$ which is comparable with the size of many financial series and permits the exact use of the Fast Fourier Transform. The grid of bandwidths analysed is $m = 10(2)\ldots, 500$. The number of replications is 1000. The advantages of the ALPE over both the LPE and NLPE are clear in Figure 1.

5 Long memory in Ibex35 volatility

In this section we analyze the persistence of the volatility of the Spanish stock index Ibex35 composed of the 35 more actively traded stocks for the period 1-10-93 to 22-3-96 half-hourly. The returns are constructed by first

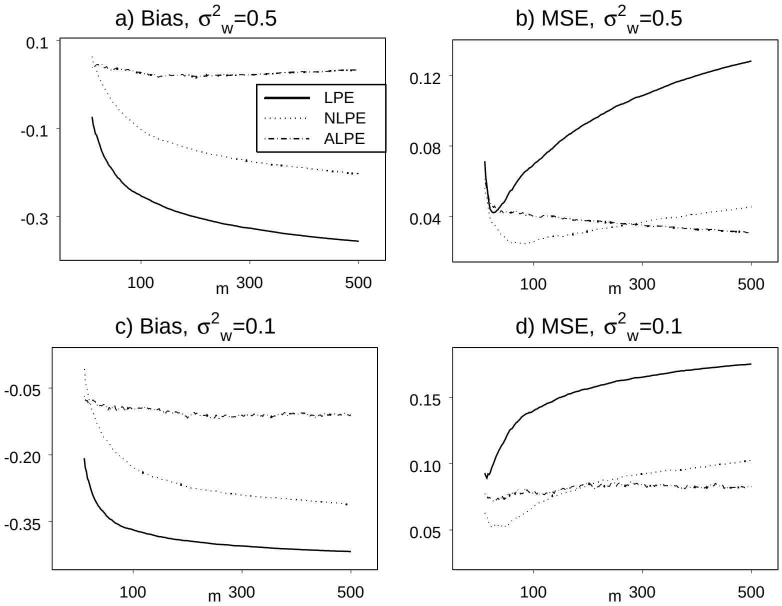


Figure 1: Bias of LPE, NLPE and ALPE, $d = 0.45$.

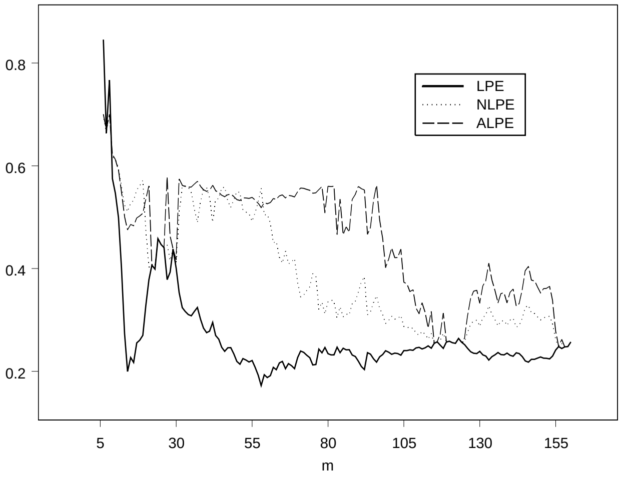


Figure 2: Log periodogram estimates (IBEX35).

differencing the logarithm of the transaction prices of the last transaction every 30 minutes, omitting incomplete days. After this modification we get the series of intra-day returns x_t, $t = 1, \ldots, 7260$. Figure 2 show the LPE, NLPE and ALPE for a grid of bandwidths $m = 6, \ldots, 160$. The decreasing behaviour of the LPE is similar to that of the Gaussian semiparametric estimation in Figure 3. However the NLPE and ALPE are higher and more stable with m sustaining the Monte Carlo results in the previous section. Comparing Figures 2 and 3 the resemblance of the LPE and GSE on one hand and the MGSE and the ALPE on the other, are evident.

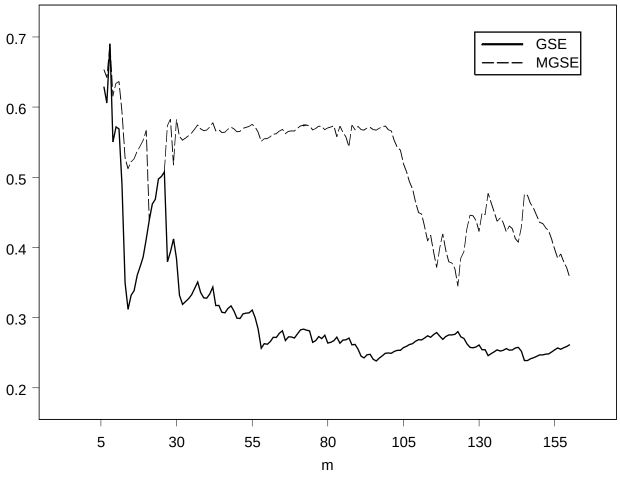


Figure 3: Standard and modified Gaussian semiparametric estimates (IBEX35).

References

[1] Arteche J. (2004). *Gaussian semiparametric estimation in long memory in stochastic volatility and signal plus noise models.* Journal of Econometrics **119**, 131 – 154.

[2] Crato N., Ray B.K. (2002). *Semi-parametric smoothing estimators for long-memory processes with added noise.* Journal of Statistical Planning and Inference **105**, 283 – 297.

[3] Deo R.S., Hurvich C.M. (2001). *On the log periodogram regression estimator of the memory parameter in long memory stochastic volatility models.* Econometric Theory **17**, 686 – 710.

[4] Hurvich C.M., Deo R., Brodsky J. (1998). *The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter in a long-memory time series.* Journal of Time Series Analysis **19**, 19 – 46.

[5] Hurvich C.M., Moulines E., Soulier P. (2003). *Estimating long memory in volatility.* Working paper SOR-2003-5, NYU Stern.

[6] Robinson P.M. (1995a). *Log-periodogram regression of time series with long-range dependence.* The Annals of Statistics **23**, 1048 – 1072.

[7] Robinson P.M. (1995b). *Gaussian semiparametric estimation of long-range dependence.* The Annals of Statistics **23**, 1630 – 1661.

[8] Sun Y., Phillips P.C.B. (2003). *Nonlinear log-periodogram regression for perturbed fractional processes.* Journal of Econometrics **115**, 355 – 389.

Acknowledgement: Research supported by the University of the Basque Country grant 9/UPV 00038.321-13503/2001 and the Spanish *Ministerio de Ciencia y Tecnología* and *FEDER* grant BEC2003-02028.

Address: J. Arteche, Departamento de Econometría y Estadística, Facultad de Ciencias Económicas y Empresariales, University of the Basque Country (UPV-EHU), Avda. Lehendakari Agirre 83, Bilbao 48015, Spain

E-mail: `etpargoj@bs.ehu.es`

DISTAL POINTS VIEWED IN KOHONEN'S SELF-ORGANIZING MAPS

Anna Bartkowiak

Key words: Multivariate data visualization, self-organizing maps, outliers, distal points.

COMPSTAT 2004 section: Data visualization.

Abstract: Kohonen's self-organizing maps are a recognized tool for finding representative data vectors and clustering the data. To what extent is it possible to preserve the topology of the data in the constructed planar map? We address the question looking at distal data points located at the peripherals of the data cloud and their position in the provided map. Several data sets have been investigated; we present the results for two of them: the Glass data (dimension $d = 7$) and the Ionosphere data (dimension $d = 32$). It was found that the distal points are reproduced either at the edges (borders) of the map, or at the peripherals of dark regions visualized in the maps.

1 Problem

Kohonen's self-organizing maps [5], [12] are a recognized tool for finding representative data vectors and clustering the data. The method provides also a 'map' of the data. It is expected that the map will preserve a large part of the topology from the data space. To what extent is it possible to preserve the topology of the data in the constructed planar map? After all, the data are multidimensional, and the map has only two dimensions! Various neighborhood relations have been investigated, see [11], [4] and the references therein. However, little attention has been paid to the outliers and distal points of the data and how are they reproduced in the map. Maruzabál and Muñoz [6] have discussed extensively that problem and concluded that self-organizing maps are not a good tool for identifying outliers. An assessment of Kohonen's method was given by Morlini [7].

Since then, we got a graphical tool, the UMAT technique, which permits to represent distances in a planar map by smoothed color hues. The UMAT technique was already implemented in Kohonen's SOM_PAK package (1995); the technique is also available in the Matlab `SomTB2` package [13]. This gives us a powerful tool to analyze the representation of the data in the map.

The aim of the paper is to investigate, how the data points located at the peripherals of the data cloud are represented in the map. Such points are also called 'distal points', see Wouters et al., Biometrics 2003, p. 1136. The distal points might be outliers, or alternatively, they might be just extreme or peripheral points of the data cloud. Several data sets have been investigated. We present here the results for two of them: the Glass data, which contain possibly four outliers (this was not further pursued), and the Ionosphere data, which seem to be a mixture with no apparent outliers.

For each of the investigated data sets we have found the top 20 distal points; this was done using robust Mahalanobis distances [9]. Next the found 20 points were projected to the map and their position in the map was observed. It was stated that the distal points are reproduced either at the edges (borders) of the map, or at the peripherals of dark regions visualized in the map.

The paper is scheduled as follows: In section 2 we present briefly the methods. Sections 3 and 4 contain the detailed analyzes for the Glass and the Ionosphere data. Section 5 contains some concluding remarks.

2 Methods

We consider data vectors $\mathbf{x}_i,\ i = 1, \ldots, N$, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$. Thus d denotes the number of variables (dimensionality), and N is the number of data vectors. The size of the data set is: $N \times d$.

Determining the distal points

The distal points were determined using robust Mahalanobis distances evaluated using the `fastmcd` algorithm developed by Rousseeuw and van Driessen [9]. For calculations we have used the Matlab function `fastmcd` offered by the cited authors. The function yields (a.o.) an index-plot of the robust Mahalanobis distances. Let $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ i = 1, \ldots, N$. It may be shown that squared Mahalanobis distances, evaluated for the data vectors $\mathbf{x}_i$, are distributed as χ^2_d.

Let $\chi^2_{.975}$ denote the (upper) 0.975 quantile of the χ^2_d distribution. Data points with Mahalanobis distances greater then $\sqrt{chi^2_{.975}}$ are suspected to be outliers.

Constructing the map

The methodology of Kohonen's self-organized maps id described in many sources, see, e.g., [5], [12], [6]. For the calculations we have used the Matlab `SomTB2` package [13]. Before starting the analysis, the data were standardized to mean equal zero and variance equal 1.

The basis for a self-organizing map is a lattice of given size. We have chosen a rectangular lattice composed of square units. The nodes of the grid contain neurons that are able to learn the distribution of the data. Let M denote the number of neurons. We have: $M = m_1 \times m_2$, with m_1 and m_2 denoting the side-lengths of the lattice. Each neuron designates a regular (in our case: square) map unit. The neurons are in one–to–one correspondence with prototypes of the data (called by Kohonen 'code-book vectors') located in the data space R^d.

During the process of learning, the code-book vectors adapt themselves to the distribution of the data. At the end of the learning, the entire data space is subdivided into adjacent regions; each region contains one representative code-book vector. Each data point has assigned its nearest code-book vector.

We may display in the map various information about the data. For instance, we may show, by appropriate shadowing of the map, how distant are the prototypes in the data space. The hues depend on the chosen colormap. Dark hues may, e.g., indicate that the respective prototypes are distant, bright hues may mean that the prototypes are close.

It might be also interesting to know how many data vectors are represented by each prototype. This information may be obtained using the technique 'som_hits', shortly: 'hits'. We pass through the data and find for each data vector $\mathbf{x}_i$, $(i = 1, \ldots, N)$ its closest prototype. Say, for $\mathbf{x}_i$ this is the prototype no. h. It is in correspondence with the neuron no. h, which in turn is contained in map unit no. h. Shortly it is said that the map unit no. h was hit by the vector $\mathbf{x}_i$. After passing through the entire set of data we obtain the *counts*, an array of size [1:M], memorizing how many times a map unit was 'hit' by subsequent data vectors. The same may be done for a subset of the data (in our case: the subset of distal points), or for a different set of data with the same dimension d. The counts (number of hits) may be also displayed in the map: either in the form of properly enlarged markers, or in the form of labels expressing digitally the number of hits. We will display them in a template of the map.

3 Visualization of the Glass data

The Glass data (source: [10]) contains $N = 214$ data vectors, each characterized by $d = 7$ variables (from the original data we have taken for our analysis the variables 2–8). The data exhibit a high multivariate kurtosis (the excess kurtosis $G_2 = 142.67$).

Index-plot for robust Mahalanobis distances calculated by the `fastmcd` procedure [9] is shown in Figure 1. One may notice that there are four outstanding points, which might be eventually considered as outliers. However, there are other data points which have quite large Mahalanobis distances. We have identified the top 20 points with the largest Mahalanobis distances – in the following these 20 points will be referred to as 'distal' points.

The map constructed for the entire glass data is shown in Figure 2. Nodes of the map represent neurons, which are in one-to-one correspondence with data prototypes located in the data space. Color shades indicate for distances of corresponding prototypes in R^d, looking in east-south directions of the map. The 20 distal points are marked by big squares squares. The topological error [13] amounts $t_e = 0.084$ (in scale [0, 1]), which is quite good.

The *counts*, containing the number of hits received by subsequent units (see explanation in previous page) are shown in the templates below.

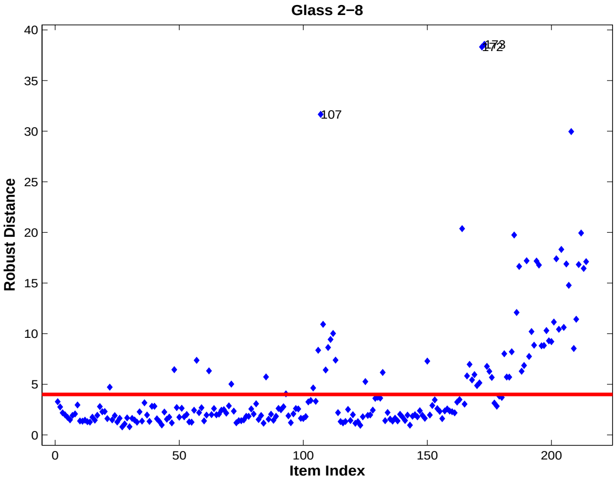


Figure 1: Robust Mahalanobis distances for the Glass data. The top four outlying points are: no.s 173, 172, 107, 208.

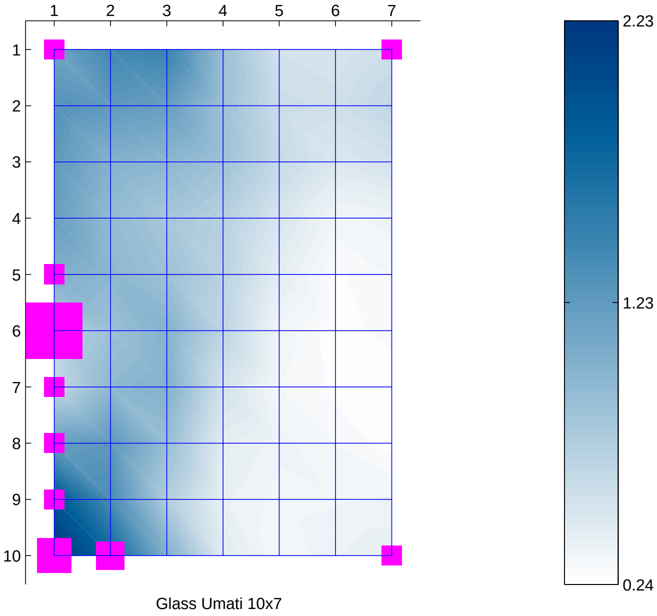


Figure 2: Map designed by a lattice 10 × 7 obtained for the standardized Glass data. The 20 distal points are shown as big squares put over the map nodes. Notice, that all distal points appear at borders of the map.

Hits of 20 distal points

	1	2	3	4	5	6	7
1:	1	0	0	0	0	0	1
2:	0	0	0	0	0	0	0
3:	0	0	0	0	0	0	0
4:	0	0	0	0	0	0	0
5:	1	0	0	0	0	0	0
6:	8	0	0	0	0	0	0
7:	1	0	0	0	0	0	0
8:	1	0	0	0	0	0	0
9:	1	0	0	0	0	0	0
10:	3	2	0	0	0	0	1

Remaining 194 points

	1	2	3	4	5	6	7
1:	5	1	1	3	4	2	7
2:	2	1	0	0	5	1	2
3:	3	2	2	1	1	2	6
4:	3	1	3	0	5	4	6
5:	0	1	4	0	2	3	5
6:	0	0	1	0	5	1	2
7:	6	1	0	3	6	3	9
8:	5	0	1	6	0	4	10
9:	0	1	3	3	7	1	5
10:	0	0	1	6	7	6	3

4 Visualization of the Ionosphere data

The Ionosphere data (source: [10]) contains $N = 351$ data vectors, each characterized by 35 variables; from these the first variable is the no. of the data vector, and the last a string characterizing class membership (binary). From the original data we have taken for our analysis the variables 3–34, thus we have $d = 32$. The data exhibit a high multivariate kurtosis (the excess kurtosis amounts $G_2 = 1684.65$).

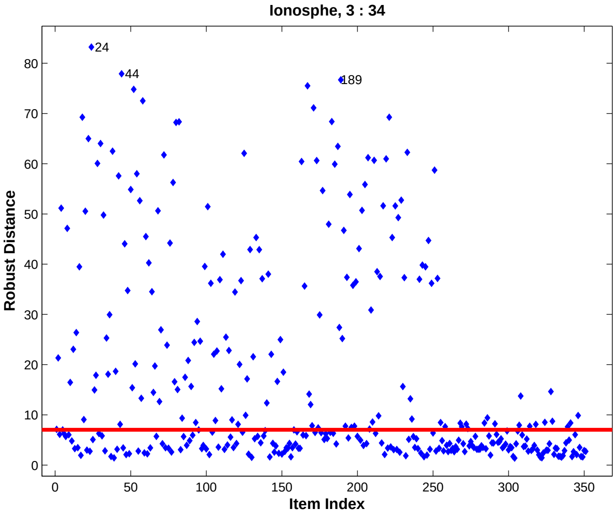


Figure 3: Robust Mahalanobis distances for the Ionosphere data. There are no outliers, only distal points. The distribution seems to be a mixture.

Plot of robust Mahalanobis distances for these data is shown in Figure 3. One may notice, that truly there are no outliers, only a lot of distal (peripheral) points. The composition of the data is quite interesting. It seems that we have here to deal with a mixture.

The top 20 distal points are the following ones (listed in increasing value of their Mahalanobis distances): [207 72 125 233 38 187 30 22 80 82 183 18 7221 171 58 52 167 189 44 24].

The map constructed for the entire (standardized) Ionosphere data is shown in Figure 4. The topological error [13] amounts here $t_e = 0.064$ (in scale [0, 1]), which is very satisfactory and indicates for a good preservation of the topology of the original data. The map units hit by the chosen 20 distal points are shown as big squares put over the nodes of the map. Distances between prototypes in R^d (when looking at east-south directions of the map) are indicated by hues of the colors, as shown in the color-bar at the right of the figure. One may notice that some of the distal points have placed themselves at the borders of the map; other distal points appear at the borders of dark regions appearing in the map.

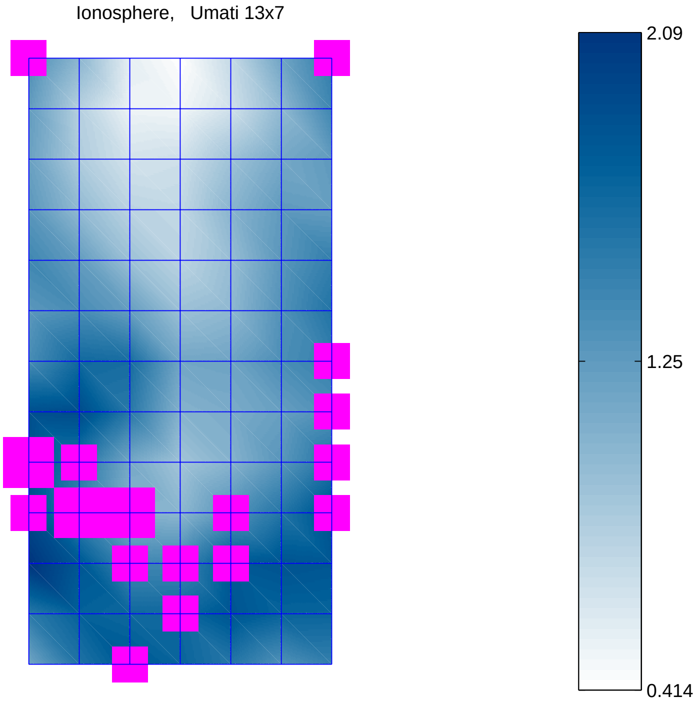


Figure 4: SOM obtained for the standardized Ionosphere data ($d = 32$) with 20 distal points marked by big squares put over the nodes of the map. Increasing darkness indicates for increasing distance between the prototypes. Notice that distal points appear either on the borders of the map, or at the borders of dark areas.

The *counts*, i.e. the number of hits into subsequent map units (a unit is identified by a map node) are shown in the templates below.

```
   Top 20 distal points
      1  2  3  4  5  6  7
   ---------------------
 1:   1  0  0  0  0  0  1
 2:   0  0  0  0  0  0  0
 3:   0  0  0  0  0  0  0
 4:   0  0  0  0  0  0  0
 5:   0  0  0  0  0  0  0
 6:   0  0  0  0  0  0  0
 7:   0  0  0  0  0  0  1
 8:   0  0  0  0  0  0  1
 9:   2  1  0  0  0  0  1
10:   1  2  2  0  1  0  1
11:   0  0  1  1  1  0  0
12:   0  0  0  1  0  0  0
13:   0  0  1  0  0  0  0
```

```
   71 mild distal points
      1  2  3  4  5  6  7
   ---------------------
 1:   5  0  0  0  0  1  4
 2:   0  0  0  0  0  0  0
 3:   1  0  0  0  0  0  0
 4:   0  0  0  0  0  0  0
 5:   2  0  0  0  2  0  0
 6:   0  0  0  1  1  0  1
 7:   1  0  0  0  1  0  3
 8:   3  0  0  3  2  3  2
 9:   1  1  2  1  3  3  6
10:   0  1  4  2  0  1  1
11:   1  0  1  0  4  0  0
12:   0  0  1  0  0  0  0
13:   0  0  0  1  0  1  0
```

```
   Main cloud, 260 points
      1  2  3  4  5  6  7
   ---------------------
 1:   1  2 13 11  3  2  1
 2:   2  1  4  5  7  2  1
 3:   3  2  7  5  2  4  6
 4:   3  3  3  7  5  0  7
 5:   3  1  7  6  2  3  4
 6:   7  3  3  5  6  1  2
 7:   7  0  1  3  2  0  1
 8:   0  0  0  2  0  0  0
 9:   0  0  1 10  1  1  0
10:   0  0  1  0  0  1  0
11:   1  2  0  1  0  2 10
12:   7  1  3  0  0  0  2
13:  11  1  1  2  7  6  7
```

The entire data set was subdivided into 3 categories: 1) Top 20 distal points exhibiting the largest Mahalanobis distances in Figure 3; 2) Mild distal points with Mahalanobis distances greater then 15.0; remind that for normal data the respective 0.975 quantile equals to $\sqrt{\chi^2_{.975;32}} = 7.03$; 3) Remainder, containing 260 data points with moderate and small Mahalanobis distances (≤ 15.0).
One may notice the difference in location of hits caused by data points belonging to the 1st and 2nd categories opposed to the 3rd category.

5 Discussion and final remarks

The presented results agree with conclusions of Morlini [7]. Indeed, Kohonen's maps, when properly used, offer many highly interesting possibilities of data exploration.

Competitive methods might be: The *Neuroscale* mapping and the *Generative Topographic mapping.* The *neuroscale* mapping (see [3], also [8] and the references therein) seems to be very promising. An example of application in analysis of erosion data is shown in [2]. Preliminary comparisons of self-organizing maps (SOMs) and the GTM method are reported in [1].

References

[1] Bartkowiak A. (2003). *SOM and GTM: Comparison in figures.* Report December 2003. `http://www.ii.uni.wroc.pl/~aba/papers.html`

[2] Bartkowiak A., Zdziarek J., Evelpidou N., Vasilopulos A. (2003). *Choosing representative data items: Kohonen, Neural Gas or Mixture Model? A case study of erosion data.* ACS'2003, The Tenth International Multiconference on Advanced Computer Systems, Szczecin-Międzyzdroje, October 22–24, 2003. Printed in disk, pp. 1 – 8. Accepted for Proceedings to be printed by Kluwer Academic Publishers.

[3] Bishop C.M., Svensèn M., Williams C.K.I. (1996). *The generative topographic mapping.* Neural Computation **10** (1), 215 – 235.

[4] Kiviluoto K. (1996). *Topology preservation in self-organizing maps. Proceedings ICNN'96.* V. 1, IEEE Neural Networks Council, June 1996, Piscataway, New Jersey, USA, **1**, 294 – 299.

[5] Kohonen T. (1995). *Self-organizing maps.* Springer Series in Information Sciences, **30**, Berlin.

[6] Maruzabál J., Muñoz A. (1997). *On the visualization of outliers via self–organizing maps.* J. of Computational and Graphical Statistics **6**, 355 – 382.

[7] Morlini I. (1998). *Multivariate outlier detection with Kohonen's networks: an useful tool for routine exploration of large data sets.* NTSS'98, Int. Seminar on New Techniques & Technologies for Statistics, Sorrento, Italy. **2**, Contributed papers. 345 – 350.

[8] Nabney I.T. (2001). *Netlab: Algorithms for pattern recognition.* Springer London, Berlin, Heidelberg. Springer Series: Advances in Pattern Recognition.

[9] Rousseeuw P.J., van Driessen K. (1999). *A fast algorithm for the minimum covariance determinant.* Technometrics **41**, 212 – 223.

[10] University of California at Irvine data repository,
`http://www.ics.uci.edu/pub/machine-learning-databases/`

[11] Venna J., Kaski S. (2001). *Neighborhood preservation in nonlinear projection methods: An experimental study.* In: Proceedings ICANN 2001, Vienna. Edited by G. Dorffner, H. Bischof, and K. Hornik. Springer, Berlin, 485 – 491.

[12] Vesanto J. (1999). *SOM–based data visualizing methods.* Intelligent Data Analysis, **3**(2), 111 – 126.

[13] Vesanto J., Himberg J. et al. (2000). *SOM Toolbox for Matlab 5.* SOM Toolbox Team, HUT, Finland, Libella Oy, Espoo, 1 – 54. `http://www.cis.hut.fi/projects/somtoolbox`, Version 0beta 2.0, Nov. 2001.

Address: A. Bartkowiak, Institute of Computer Science, University of Wrocław, Przesmyckiego 20, Wrocław 51-151, Poland

E-mail: `aba@ii.uni.wroc.pl`

PLS-COX MODEL: APPLICATION TO GENE EXPRESSION

Philippe Bastien

Key words: Cox model, PLS regression, gene profiling.

COMPSTAT 2004 section: Partial least squares.

Abstract: With advances in high-density DNA microarray technology, gene expression profiling is extensively used to discover new markers and new therapeutic targets. This technique supposes to take into account the expression of thousands of genes with respect to a limited number of patients. To predict survival probability on the basis of gene expression signatures can become a very useful diagnostic tool. In the context of highly multidimensional data the classical Cox model does not work. The PLS-Cox model by operating a dimension reduction of the gene expression space directed towards the explanation of the risk function appears particularly useful. It allows the determination of signatures of genomic expressions associated with survival, to predict the survival probability from these profiles, and reduce inter individual variability by changing the level of adjustment from a phenotypical level to a genotypical level.

1 Introduction

The proportional hazard regression model suggested by [5] to study the relationship between the time to event and a set of covariates in the presence of censoring, is the model most commonly used for the analysis of survival data. However, like multivariate regression models, it supposes that there are more observations than variables, complete data, and variables not strongly correlated between them. These constraints are often crippling in practice. In particular the analysis of transcriptomic data supposes to take into account the expression of thousands of genes compared to only a limited number of patients. The solution suggested is to initially operate a dimension reduction of the space of genes directed towards the explanation of the risk function. One then builds a Cox model on the PLS components.

Alizadeh et al. [2] identified from the expression of genes of 40 subjects suffering from diffuse large B-cell lymphomas (DLBCL) two subgroups, each characterized by a distinct gene expression signature. These were associated with very different clinical prognoses. Using information on patients survival allows the determination of genotypic signatures linked to the risk function. Survival probabilities have then been carried out from these expression profiles. We show that these genotypic signatures bring additional informations to an index of existing clinical risk.

2 Methods

The suggested method associate PLS regression [9] with the Cox model. It has already been used on epidemiological data [3]. Its specificity is that it takes into account the censoring information in the construction of PLS components.

PLS-Cox algorithm

Let $X_0 = x_1, \ldots, x_p$ a matrix whose columns are gene expression (log ratio). One seeks successively m orthogonal PLS components T_h which are linear combinations of the x_j. In particular the research of the h-th.PLS component Th is carried out according to the following steps:

Step 1: For $j = 1$ to p, calculate the coefficients of regression a_{hj} of x_j in the Cox model with covariates $T_1, T_2, \ldots, T_{h-1}$ and x_j.
Step 2: Normalize the column vector ah formed by $a_{hj} : w_h = a_h / \|a_h\|$.
Step 3: Calculate the residual X_{h-1} of the linear regression of X_0 on $T_1, \ldots, T_{h-1}$.
Step 4: Calculate the component $T_h = X_{h-1} w_h / w'_h w_h$.
Step 5: Express the component T_h according to $X_0 : T_h = X_0 w^*_h$.

The prediction of the risk function h(t) is then carried out in a natural way with the Cox model adjusted on PLS components. The regression equation can also be written according to the original data with the coefficients confidence intervals estimated by bootstrap resampling.

Cross-validation

The number k of PLS components T_h was chosen by cross-validation. Each patient's score was estimated using a training data set of $N - 1$ samples (leave-one-out CV). Let i be the subscript for sample i and $-i$ the subscript when sample i is leaved out. The score for patient i on h-th PLS component is defined as:

$$t_{h,j} = x_{h-1,i}\, w_{h,-i} = \left(x_{0,i} - \sum_{j=1}^{h-1} t_{j,i}\, p'_{j,-i} \right) w_{h,i}$$

with: $x_{h-1,i}$ the i-th row of the residual matrix X_{h-1}; $w_{h,-i}$ the weights based on $X_{h-1,-i}$; $p'_{j,-i}$ the loadings, defined as the coefficients of $T_{j,-i}$ in the regression of $X_{0,-i}$ on $T_{1,-i}, \ldots, T_{j,-i}$, the j first PLS components carried out on $X_{0,-i}$.

PLS-Cox and PLS-GR

The Cox-PLS algorithm uses the principles of the NIPALS algorithm [10] and can also function in the presence of missing data. The PLS-Cox model is a particular case of PLS generalized linear regression [4].

Estimation of the survivor function

During the prediction phase, a proportional hazard model is fitted with the k PLS scores $T_1, \ldots, T_k$ as covariates. Let $S_0(t) = \exp\left[-\int_0^i h_o(u)du\right]$ the baseline unspecified survivor function, $T_i' = (t_{i1}, \ldots, t_{ik})$, and $\beta' = (\beta_1, \ldots, \beta_k)$. The survivor function given the scores T_i is: $S(t, T_i) = S_0(t, T_i)^{\exp(T_i', \beta)}$. The calculation of the non-parametric maximum likelihood of $S_0(t)$ [6] is based on the product limit estimate with similar argument to that used in obtaining kaplan-Meier estimate.

Let $t_{(1)}, \ldots, t_{(l)}$ be the distinct failures times, the likelihood function is maximized by taking $S_0(t) = S_0(t_{(i)} + 0)$ for $t_{(i)} < t \leq t_{(i+1)}$ and allowing probability mass to fall only at the observed failure time $t_{(i)}$. This leads to the consideration of a discrete model with hazard contribution $1 - \alpha_i$ at $t_{(i)}$. Let

$$\hat{S}_0(t) = \prod_{j/t(j)<t} \hat{\alpha}_j$$

the survival probability at time t

$$\hat{S}_i(t) = \prod_{j/t(j)<t} \hat{\alpha}_j^{\exp T_i \beta}$$

the survival probability at time t for a patient with covariates T_i.

The maximum likelihood estimate $\hat{\alpha}_i$ of α_i is obtained numerically from:

$$\sum_{k \in F_i} \frac{\hat{u}_k}{1 - \hat{\alpha}_i^{\hat{u}_k}} = \sum_{l \in R(t_i)} \hat{u}_l$$

with $\hat{u}_k = \exp(T_k' \hat{\beta})$, F_i the set of individuals failing at time $t_{(i)}$ and $R_{(ti)}$ the risk set at time $t_{(i)}$. In case where there are no ties then the set F_i contains only one individual and the solution to the above equation can be solved analytically and is given by

$$\hat{\alpha}_i = \left[1 - \left(\hat{u} / \sum_{l \in R(t_i)} \hat{u}_l\right)\right]^{\hat{u}_i^{-l}} .$$

One finds the Kaplan-Meier estimator when $T_i = 0$ for all the individuals:

$$\hat{S}(t) = \prod_{j/t(j)<t} \frac{(n_j - d_j)}{n_j} .$$

3 Application

The data set from [2] consists of gene expression level from cDNA experiments involving three prevalent adult lymphoid malignancies: Diffuse large B-cell lymphoma (DLBCL), B-Cell chronic Lymphocytic Leukemia (BCLL), and Follicular Lymphoma (FL). Data are available on the study supplement web site (`http://llmpp.nih.gov/lymphoma/data.shtml`).

CDNA targets were prepared from experimental mRNA samples and were labelled with Cy5-dye during reverse transcription. A reference cDNA sample was prepared from a combination of nine different lymphoma cell lines and was labelled with Cy3-dye. Cy-labelled experimental and reference cDNAs were mixed and hybridised onto the microarray The standardized intensity ratio of fluorescence was then quantified for each gene. It reflects the relative abundance of the gene in each experimental sample of mRNA compared to the reference sample.

By using clustering analysis, Alizadeh and al. [2] identified two DLBCL sub-groups with different transcriptomic profiles. They correspond to distinct levels of lymphocytes B differentiation: Germinal Center B-like (19 patients) and Activated B-like (21 patients). In complement to the transcriptomic data, the duration of patients survival was also collected. Among the 40 patients 22 events (death) were observed and the 18 remaining survival durations being censured. Patients with a DLBCL of the Germinal center B-like have, on average, a significantly better survival than those with Activated B-like type as shown on figure 1. The molecular classification of the tumours on the basis of their genetic expression profile thus allows to highlight sub-types of cancer non identified.

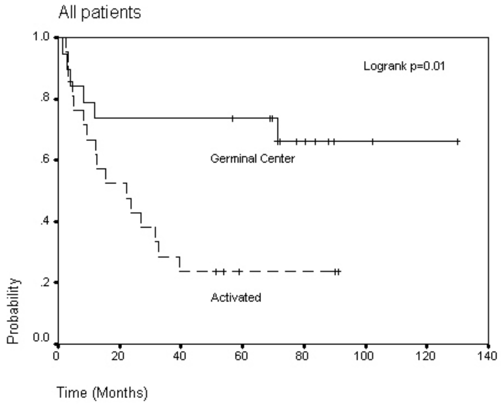


Figure 1: Kaplan-Meier survival curves estimates by molecular sub-groups.

4 Results

One Thousand and height hundred genes were selected from over more than 13000 to having different expressions to the two molecular types (ttest, $p < 0.05$). We retained two PLS components by cross-validation. Once PLS-

Cox model has been estimated, the significance of genes coefficients could be ascertain in a non parametric framework by means of a bootstrap procedure. Bootstrap confidence intervals were computed based on the 2.5 and 97.5 percentiles of the bootstrap empirical distribution (balanced bootstrap, B=500).

Figure 2 presents coefficients confidence intervals of the PLS-Cox model on two components expressed according to their original data (log ratio). The coefficients were sorted by ascending values. In order to simplify PLS components, only genes having a significant contribution at the 5 % threshold were taken into account. It explains the clear separation on both sides of the ordinate axis.

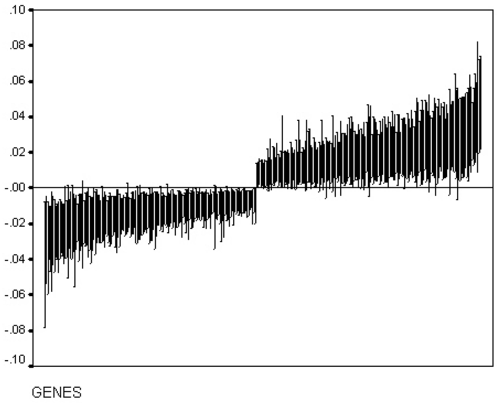


Figure 2: 95% bootstrap confidence intervals for genes coefficients.

The following graph (figure 3) presents the individual distributions for the patients of the Activated B-like type (dotted line) and for those of Germinal center B-like type (continuous line). The letters represent the average distributions by molecular type. The distributions were estimated by cross-validation with two PLS components.

Based on the log-ratio of gene expressions for the mean levels of PLS components, the survival curves demonstrated more marked prognoses between the two molecular types in comparison to the survival estimation using Kaplan-Meier. The genotypic signatures of the two molecular types appear well associated with the different prognoses, with very minor overlaps.

International prognostic indicator (IPI)

A clinical index scored from 0 to 5 is used to define sub-groups of patients suffering from DLBCL. The subjects of the group with the lowest scores IPI (0-2) have a better prognostic than those having highest scores (3-5). Alizadeh et al [2] showed that in the group with the lowest risk factors, the

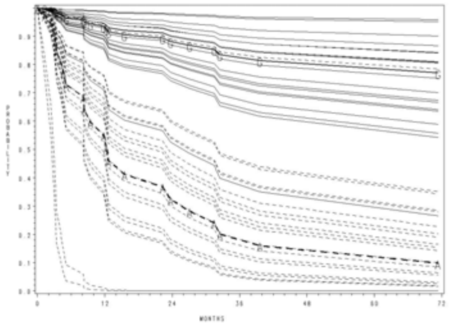


Figure 3: Cross-validated survival curves.

patients presenting a profile of genetic expression of Germinal center B-like type had a significantly better survival (Logrank, $p < 0.05$) that those of Activated B-like type. They did not observe a similar effect in the higher risk factors group ($p = 0.55$) as illustrated in figure 4.

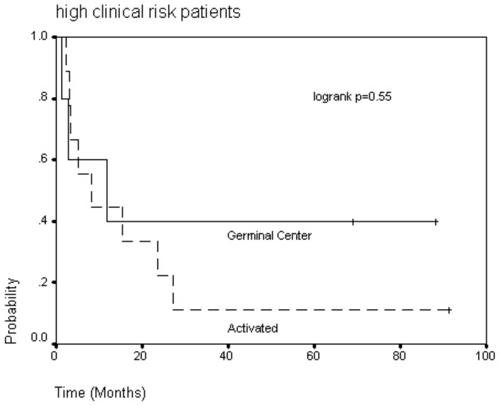


Figure 4: Kaplan-Meier survival curves for the high clinical risk patients.

The Cox-PLS model on the higher risk factors group, taking into account the transcriptomic information is more selective and allows the differentiation of the two molecular types. Figure 5 shows the individual distributions of survival estimated by cross-validation.

The gene expression signatures makes it possible to differentiate the two molecular types.

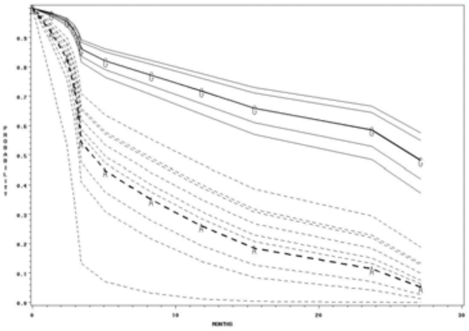


Figure 5: Cross-validated survival curves for the high clinical risk patients.

5 Discussion

With advances in high-density cDNA microarray technology, gene expression profiling is extensively use to discover new markers and new therapeutic targets. This technique supposes to take into account the expression of thousands of genes with respect to only a limited number of patients. To predict survival probability on the basis of gene expression signatures can become a very useful diagnostic tool. In the context of highly multidimensional data the classical Cox model does not work.

Recently Nguyen and Rocke [7] illustrated using the example of [2] the use of PLS components as covariates to predict the probabilities of survival in a Cox model. However their model was not completely satisfactory, since it did not take into account the censoring information in the estimation of PLS component, thus inducing a potential bias in their estimates.

The PLS-Cox model described above shows major improvement with respect to the method proposed by [7]. It takes into account the censoring information in the estimation of PLS components. In case of missing data, PLS components are computed in accordance with the NIPALS algorithm. Moreover statistical significance of gene coefficients is ascertain using bootstrap validation procedure.

The PLS-Cox model by operating a dimension reduction of the genes expression space directed towards the explanation of the risk function appears particularly useful. It allows the determination of signatures of genomic expressions associated with survival, to predict the survival probability from these profiles, and reduce inter individual variability by changing the level of adjustment from a phenotypical level to a genotypical level. In order to assess the efficacy of new drugs, study design will benefit from a better characterisation of patient groups made possible by genomic expression profiling.

References

[1] Allison Paul D. (1995). *Survival analysis using the SAS system.* A Practical Guide, SAS Inc, Cary, NC.

[2] Alizadeh A.A. et al. (2000). *Distinct types of diffuse large B-cell lymphoma identified by gene expression profile.* Nature, **403**, 503 – 511.

[3] Bastien P., Tenenhaus M. (2001). *PLS generalized linear regression. Application to the analysis of life time data.* In PLS and Related Methods, Proceedings of the PLS'01 International Symposium, Esposito Vinzi V., Lauro C., Morineau A. & Tenenhaus M. (Eds). CISIA-CERESTA Editeur, Paris, 131 – 140.

[4] Bastien P., Esposito Vinzi V., Tenenhaus M. (2004). *PLS generalized linear regression.* Computational Statistics & Data Analysis, to appear.

[5] Cox, D.R. (1972). *Regression models and life tables (with discussion).* Journal of the Royal Statistical Society, B, **74**, 187 – 220.

[6] Kalbfleich J.D., Prentice R.L. (1973). *Marginal Likelihoods based on Cox's regression and life model.* Biometrika, **60**, 267 – 278.

[7] Nguyen D.V., Rocke D. (2001). *Partial least squares proportional hazard regression for application to DNA microarray survival data.* Bioinformatics, **18**, 1625 – 1632.

[8] Tenenhaus M. (1998). *La régression PLS.* Technip, Paris.

[9] Wold S., Martens & Wold H. (1983). *The multivariate calibration problem in chemistry solved by the PLS method.* In Proc. Conf. Matrix Pencils, Ruhe A. & Kastrom B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, 286 – 293.

[10] Wold H., (1966). *Estimation of principal components and related models by iterative least squares.* In Multivariate Analysis, Krishnaiah P.R. (Ed.), Academic Press, New York, 391 – 420.

Address: P. Bastien, L'Oréal Recherche, 1 avenue Eugene Schueller - BP 22 - 93601 Aulnay sous Bois Cedex

E-mail: `pbastien@recherche.loreal.com`

TESTING SOLUTION QUALITY IN STOCHASTIC PROGRAMMING: A SINGLE REPLICATION PROCEDURE

Guzin Bayraksan and David P. Morton

Key words: Stochastic programming, Monte Carlo simulation.

COMPSTAT 2004 section: Optimization.

Abstract: We develop a procedure for testing the quality of a candidate solution for a class of stochastic programs. Quality is defined via the optimality gap and the procedure's output is a confidence interval on this gap. We review a multiple-replications procedure and then, we present a result that justifies a computationally simplified single-replication procedure. We compare empirical coverage results of the two procedures for a newsvendor problem.

1 Introduction

We consider a stochastic optimization problem of the form

$$z^* = \min_{x \in X} Ef(x, \tilde{\xi}), \tag{1}$$

where f is a real-valued function that determines the cost of operating with decision x under a realization of the random vector $\tilde{\xi}$, whose distribution is assumed known. $X \subseteq \mathbb{R}^n$ denotes the set of constraints that the decision vector x must obey and E is the expectation operator. As simple as it is to state, (SP) represents a large class of problems that can be found in the statistics and operations research literature. Our motivation comes from a special class of (SP) known as two-stage stochastic programs with recourse. The well-known two-stage stochastic linear program with recourse was introduced independently by [2], [4], in which

$$\begin{aligned} f(x, \tilde{\xi}) = cx \ + \ & \min_{y \geq 0} \quad \tilde{q}y \\ & \text{s.t} \quad \tilde{W}y \geq \tilde{r} - \tilde{T}x, \end{aligned}$$

$X = \{x : Ax = b, x \geq 0\}$ and $\tilde{\xi} = (\tilde{q}, \tilde{W}, \tilde{r}, \tilde{T})$ is a random vector on $(\Xi, \mathcal{B}, P)$. This formulation can be extended to multiple stages, integer restrictions can be imposed in any of the stages and nonlinear constraints and objective function terms can be added. Stochastic programs with recourse have been successfully applied to problems from finance, energy, telecommunications, transportation, logistics and supply chain management (e.g., [10]).

We do not restrict ourselves to the linear recourse model. Instead we make the following assumptions with respect to (SP).

(A1) $f(\cdot, \tilde{\xi})$ is continuous on X, w.p.1,
(A2) $E \sup_{x \in X} f^2(x, \tilde{\xi}) < \infty$,
(A3) $X \neq \emptyset$ and is compact.

As the dimension of the random vector $\tilde{\xi}$ grows, (SP) gets harder and often impossible to solve exactly, unless the cost function f has simple structure, or the number of realizations is small. In such cases, an intuitive approach is to resort to sampling and approximate the problem with

$$z_n^* = \min_{x \in X} \frac{1}{n} \sum_{i=1}^{n} f(x, \tilde{\xi}^i). \tag{2}$$

$\tilde{\xi}^1, \tilde{\xi}^2, \ldots, \tilde{\xi}^n$ may be independent and identically distributed (i.i.d.) as $\tilde{\xi}$ or may be generated according to another sampling scheme. Let x^* denote an optimal solution to (SP) with optimal cost z^*. Similarly, let x_n^* and z_n^* denote an optimal solution and the optimal cost of (SP$_n$). Consistency and other asymptotic properties of estimators x_n^* and z_n^* have been studied extensively in the literature, see e.g., [1], [5], [6], [9].

In this paper, we discuss Monte Carlo sampling-based procedures for testing solution quality in stochastic programs. Determining whether a solution is of high quality (optimal or near optimal) is a fundamental question in optimization. Given a candidate solution $\hat{x}$, we define its quality by its optimality gap, $\mu_{\hat{x}} = Ef(\hat{x}, \tilde{\xi}) - z^*$. In the next section, we review how to construct confidence intervals (CIs) on the optimality gap using a multiple replications procedure [7]. Then, we show how to obtain a valid CI using only a single replication. Finally, we examine a newsvendor problem with uniform demand and compare empirical coverage results for the two procedures.

2 Review of multiple replications procedure

Let $\tilde{\xi}^1, \tilde{\xi}^2, \ldots, \tilde{\xi}^n$ be i.i.d. from the distribution of $\tilde{\xi}$. Then, by interchanging minimization and expectation we obtain a statistical lower bound on z^*,

$$Ez_n^* = E\left[\min_{x \in X} \frac{1}{n} \sum_{i=1}^{n} f(x, \tilde{\xi}^i)\right] \leq \min_{x \in X} E\left[\frac{1}{n} \sum_{i=1}^{n} f(x, \tilde{\xi}^i)\right] = \min_{x \in X} Ef(x, \tilde{\xi}) = z^*.$$

This result establishes that z_n^* has a negative bias, $Ez_n^* - z^* \leq 0$. It can also be shown that $Ez_n^* \leq Ez_{n+1}^*$ for all n. This monotonicity result tells us that on average we get better estimates of the optimal value as the sample size increases.

Given a feasible decision $\hat{x} \in X$ and a sample size n for (SP$_n$), we bound the optimal value of (SP) using the above lower

bound result, $Ez_n^* \leq z^* \leq Ef(\hat{x}, \tilde{\xi})$. The right inequality comes from suboptimality of $\hat{x}$. An upper bound on the optimality gap for $\hat{x}$ is then

$Ef(\hat{x},\tilde{\xi}) - Ez_n^*$. We estimate this quantity by

$$G_n(\hat{x}) = \frac{1}{n}\sum_{i=1}^{n} f(\hat{x},\tilde{\xi}^i) - \min_{x\in X}\frac{1}{n}\sum_{i=1}^{n} f(x,\tilde{\xi}^i). \tag{3}$$

The first term on the right-hand side of (3) is an upper bound estimate and converges to $Ef(\hat{x},\tilde{\xi})$, w.p.1, by the strong law of large numbers. The second quantity, z_n^*, is a lower bound estimate on z^*. In expectation, it provides a lower bound and under (A1)-(A3) converges to z^*, w.p.1 (see subsequent Lemma 3.1). When a common stream of random numbers, $\tilde{\xi}^1, \tilde{\xi}^2, \ldots, \tilde{\xi}^n$, is used in calculating both terms in (3), $G_n(\hat{x}) \geq 0$, w.p.1. This approach also facilitates variance reduction.

Because of the minimization in (3), $G_n(\hat{x})$ (or, its scaled version $\sqrt{n}(G_n(\hat{x}) - \mu_{\hat{x}})$) is, in general, not normally distributed even as n grows large. Therefore, in [7] confidence intervals are constructed by employing replications, an approach frequently used in simulation for estimating the mean of a random variable with an unknown or non-normal distribution. We summarize below the multiple replications procedure (MRP) to construct a CI on the optimality gap. Let $t_{n,\alpha}$ be the $1-\alpha$ quantile of the Student's t distribution with n degrees of freedom.

MRP:

Input: Desired CI level $0 < \alpha < 1$, sample size n, replication size n_g and a candidate solution $\hat{x} \in X$.

Output: $(1-\alpha)$-level confidence interval on $\mu_{\hat{x}}$.

```
1. For $i = 1, 2, \ldots, n_g$,
    1.1. Sample i.i.d. observations $\tilde{\xi}^{i1}, \tilde{\xi}^{i2}, \ldots, \tilde{\xi}^{in}$ from the distribution of $\tilde{\xi}$,
    1.2. Solve (SP$_n$) using $\tilde{\xi}^{i1}, \tilde{\xi}^{i2}, \ldots, \tilde{\xi}^{in}$ to obtain $x_n^{i*}$,
    1.3. Calculate $G_n^i(\hat{x}) = \frac{1}{n}\sum_{j=1}^{n}\left(f(\hat{x},\tilde{\xi}^{ij}) - f(x_n^{i*},\tilde{\xi}^{ij})\right)$.
2. Calculate gap estimate and sample variance by
```

$$\bar{G}(n_g) = \frac{1}{n_g}\sum_{i=1}^{n_g} G_n^i(\hat{x}) \quad \text{and} \quad s_G^2(n_g) = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}\left(G_n^i(\hat{x}) - \bar{G}(n_g)\right)^2.$$

```
3. Output one-sided CI on $\mu_{\hat{x}}$, $\left[0, \bar{G}(n_g) + t_{n_g-1,\alpha}s_G(n_g)/\sqrt{n_g}\right]$.
```

Since $\bar{G}(n_g)$ is a sample mean of i.i.d. random variables, it is possible to use the standard central limit theorem (CLT) to construct a $(1-\alpha)$-level CI for the optimality gap given in step 3. Due to the negative bias of z_n^*, $E\bar{G}(n_g) \geq Ef(\hat{x},\tilde{\xi}) - z^*$. Thus, for sufficiently large n_g, we can infer that

$$P\left(Ef(\hat{x},\tilde{\xi}) - z^* \leq \bar{G}(n_g) + \frac{t_{n_g-1,\alpha}s_G(n_g)}{\sqrt{n_g}}\right) \approx 1-\alpha \tag{4}$$

and hence that the CI formed by MRP will cover the optimality gap of $\hat{x}$ with the desired probability.

3 Single replication procedure

When applying the multiple replication procedure reviewed above, the replication size is typically taken to be $n_g \geq 30$ to have a valid statistical inference. This constitutes a major drawback as one needs to solve at least 30 optimization problems (in step 1.2) in order to determine whether a candidate solution is of high quality. In this section, we show how a single replication, $n_g = 1$, can be used to make a valid statistical inference on the quality of a candidate solution.

As before, we assume that the candidate solution $\hat{x} \in X$ is given, and we use the following additional notation. For a feasible solution, $x \in X$, let $\bar{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} f(x,\tilde{\xi}^i)$, $\sigma^2(x) = \text{var}[f(\hat{x},\tilde{\xi}) - f(x,\tilde{\xi})]$ and $s_n^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}[(f(\hat{x},\tilde{\xi}^i) - f(x,\tilde{\xi}^i)) - (\bar{f}_n(\hat{x}) - \bar{f}_n(x))]^2$. Note that $G_n(\hat{x})$ given in equation (3) can be written as $\bar{f}_n(\hat{x}) - z_n^*$, with the understanding that the same n observations $\tilde{\xi}^1, \tilde{\xi}^2, \ldots, \tilde{\xi}^n$ are used in $\bar{f}_n(\hat{x})$ and z_n^*. Below we state the single replication procedure (SRP).

SRP:

Input: Desired CI level $0 < \alpha < 1/2$, sample size n and a candidate solution $\hat{x} \in X$.

Output: $(1-\alpha)$-level confidence interval on $\mu_{\hat{x}}$.

1. Sample i.i.d. observations $\tilde{\xi}^1, \tilde{\xi}^2, \ldots, \tilde{\xi}^n$ from the distribution of $\tilde{\xi}$.
2. Solve (SP$_n$) to obtain x_n^*.
3. Calculate $G_n(\hat{x})$ as given in (3) and

$$s_n^2(x_n^*) = \frac{1}{n-1}\sum_{i=1}^{n}\Big[(f(\hat{x},\tilde{\xi}^i) - f(x_n^*,\tilde{\xi}^i)) - (\bar{f}_n(\hat{x}) - \bar{f}_n(x_n^*))\Big]^2 .$$

4. Output one-sided CI on $\mu_{\hat{x}}$, $[0,\ G_n(\hat{x}) + t_{n-1,\alpha} s_n(x_n^*)/\sqrt{n}\ ]$.

In the MRP, n_g i.i.d. observations of $G_n(\hat{x})$ are calculated and the sample variance of these gap estimates is used to form the CI. In contrast, only one value of $G_n(\hat{x})$ is calculated in SRP and the individual observations, $f(\hat{x},\tilde{\xi}^i) - f(x_n^*,\tilde{\xi}^i)$ for $i = 1, \ldots, n$, are used to calculate the sample variance. Below, we show how solving a single replication yields enough information to make a valid statistical inference concerning the quality of a candidate solution. Before stating the theorem, we require the following lemma that establishes consistency of the estimators. Let X^* denote the set of optimal solutions to (SP) and let $x_{\min}^* \in \arg\min_{x\in X^*} \text{var}[f(\hat{x},\tilde{\xi}) - f(x,\tilde{\xi})]$. So, $x_{\min}^*$ is a solution with minimum variance of $f(\hat{x},\tilde{\xi}) - f(x,\tilde{\xi})$, $\sigma^2(x_{\min}^*)$, among all the optimal solutions.

Lemma 3.1. *Assume (A1)-(A3) and that $\tilde{\xi}^1, \tilde{\xi}^2, \ldots, \tilde{\xi}^n$ are i.i.d. as $\tilde{\xi}$. Then,*

(i) $z_n^* \to z^*$, *w.p.1,*

(ii) all limit points of $\{x_n^*\}$ *lie in* X^*, *w.p.1,*

(iii) $\liminf_{n\to\infty} s_n^2(x_n^*) \geq \sigma^2(x_{\min}^*)$, *w.p.1.*

Proof: (A2) implies that $E \sup_{x \in X} f(x, \tilde{\xi}) < \infty$. Therefore, (i) follows immediately from Theorem A1 of [8, p.69]. (A1)-(A3) implies $\bar{f}_n(x)$ converges uniformly to $Ef(x, \tilde{\xi})$, w.p.1 on X. This coupled with (i) implies (ii). To prove (iii), we first show that the sequence of continuous functions $s_n^2(x)$ converges to $\sigma^2(x)$ uniformly, w.p.1 on X. Let $g(x, \tilde{\xi}) = f(\hat{x}, \tilde{\xi}) - f(x, \tilde{\xi})$. Then, with $\bar{g}_n(x) = \frac{1}{n} \sum_{i=1}^n g(x, \tilde{\xi}^i)$ we have

$$s_n^2(x) = \frac{n}{n-1} \left\{ \frac{1}{n} \sum_{i=1}^n \Big(g(x, \tilde{\xi}^i) - Eg(x, \tilde{\xi}) \Big)^2 - \Big(\bar{g}_n(x) - Eg(x, \tilde{\xi}) \Big)^2 \right\}.$$

The first term in the curly brackets is a sample mean of i.i.d. random variables and by Lemma A1 of [8, p.67] converges uniformly, w.p.1, to $\sigma^2(x) = \operatorname{var} g(x, \tilde{\xi})$. Also, by the same lemma, $\bar{g}_n(x)$ converges uniformly to $Eg(x, \tilde{\xi})$, w.p.1, i.e., $\sup_{x \in X} \Big| \bar{g}_n(x) - Eg(x, \tilde{\xi}) \Big| \to 0$, w.p.1. This implies

$$\sup_{x \in X} \Big(\bar{g}_n(x) - Eg(x, \tilde{\xi}) \Big)^2 = \left(\sup_{x \in X} \Big| \bar{g}_n(x) - Eg(x, \tilde{\xi}) \Big| \right)^2 \to 0, \text{ w.p.1.}$$

The sum of these two terms, $a_n(x) = \frac{1}{n} \sum_{i=1}^n (g(x, \tilde{\xi}^i) - Eg(x, \tilde{\xi}))^2 - (\bar{g}_n(x) - Eg(x, \tilde{\xi}))^2$, then converges uniformly to $\sigma^2(x)$, w.p.1. To show uniform convergence of $\frac{n}{n-1} a_n(x)$, consider the following inequality

$$\sup_{x \in X} \left| a_n(x) + \frac{a_n(x)}{n-1} - \sigma^2(x) \right| \le \sup_{x \in X} \left| a_n(x) - \sigma^2(x) \right| \quad + \\ \sup_{x \in X} \left| \frac{a_n(x) - \sigma^2(x)}{n-1} \right| \quad + \quad \sup_{x \in X} \left| \frac{\sigma^2(x)}{n-1} \right|.$$

By the above argument the first two terms on the right-hand side converge to 0, w.p.1. By (A2), $\sup_{x \in X} \sigma^2(x) < \infty$. Thus, the last term also converges to 0, establishing uniform convergence.

Since X is compact, there exists a subsequence N along which $\{x_n^*\}_{n \in N}$ converges to a point in X and by (ii) this point is in X^*. So, using the uniform convergence shown above,

$$\lim_{\substack{n \to \infty \\ n \in N}} s_n^2(x_n^*) \ge \inf_{x \in X^*} \sigma^2(x), \text{ w.p.1.}$$

The subsequence N is arbitrary and hence we obtain (iii).

Theorem 3.1. *Assume (A1)-(A3) and that $\tilde{\xi}^1, \tilde{\xi}^2, \ldots, \tilde{\xi}^n$ are i.i.d. as $\tilde{\xi}$. Given $0 < \alpha < 1/2$, for the SRP,*

$$\liminf_{n \to \infty} P \left(\mu_{\hat{x}} \le G_n(\hat{x}) + \frac{t_{n-1,\alpha} s_n(x_n^*)}{\sqrt{n}} \right) \ge 1 - \alpha. \tag{5}$$

Proof: When $\hat{x} \in X^*$, inequality (5) is trivial. Suppose $\hat{x} \notin X^*$. Recall that $z_n^* = \min_{x \in X} \bar{f}_n(x)$. Thus,

$$G_n(\hat{x}) = \bar{f}_n(\hat{x}) - z_n^* \geq \bar{f}_n(\hat{x}) - \bar{f}_n(x), \quad \forall x \in X.$$

Replacing x by $x^*_{\min} \in \arg\min_{x \in X^*} \sigma^2(x)$ we obtain,

$$\begin{aligned} & P\left(G_n(\hat{x}) + \frac{t_{n-1,\alpha} s_n(x_n^*)}{\sqrt{n}} \geq \mu_{\hat{x}} \right) \\ \geq \quad & P\left(\bar{f}_n(\hat{x}) - \bar{f}_n(x^*_{\min}) + \frac{t_{n-1,\alpha} s_n(x_n^*)}{\sqrt{n}} \geq \mu_{\hat{x}} \right) \qquad (6) \\ = \quad & P\left(\frac{(\bar{f}_n(\hat{x}) - \bar{f}_n(x^*_{\min})) - \mu_{\hat{x}}}{\sigma(x^*_{\min})/\sqrt{n}} \geq -t_{n-1,\alpha} \frac{s_n(x_n^*)}{\sigma(x^*_{\min})} \right), \qquad (7) \end{aligned}$$

where in (7) we assume $\sigma^2(x^*_{\min}) > 0$. Note that if $\sigma^2(x^*_{\min}) = 0$ then $\text{var}\,[\bar{f}_n(\hat{x}) - \bar{f}_n(x^*_{\min})] = \frac{1}{n}\sigma^2(x^*_{\min}) = 0$ and it follows from (6) that (5) is again trivial. Since $\bar{f}_n(\hat{x}) - \bar{f}_n(x^*_{\min})$ is a sample mean of i.i.d. random variables, by the CLT we have

$$\lim_{n \to \infty} P\left(\frac{(\bar{f}_n(\hat{x}) - \bar{f}_n(x^*_{\min})) - \mu_{\hat{x}}}{\sigma(x^*_{\min})/\sqrt{n}} \geq -t_{n-1,\alpha} \right) = 1 - \alpha. \qquad (8)$$

By Lemma 3.1,

$$\liminf_{n \to \infty} \frac{s_n(x_n^*)}{\sigma(x^*_{\min})} \geq 1, \quad \text{w.p.1.} \qquad (9)$$

Combining (8) and (9) via a converging-together argument we obtain (5).

Theorem 3.1 justifies construction of the approximate $(1-\alpha)$-level one-sided confidence interval for $\mu_{\hat{x}} = Ef(\hat{x}, \tilde{\xi}) - z^*$, given in step 4 of SRP without requiring $G_n(\hat{x}) = \bar{f}_n(\hat{x}) - z_n^*$ to be asymptotically normal. The intuitive reason for this is that minimization of the sample mean in z_n^*, while making asymptotic analysis of this random variable more difficult, projects the normal distribution so that the resulting confidence interval is conservative.

We reviewed a procedure in which we use $n_g \geq 30$ replications and introduced a procedure with just one replication, $n_g = 1$. However, the theory presented above is asymptotic, giving us the desired coverage only as the sample size grows to infinity. To test the small-sample performance of the described procedures, we apply them in the next section to the newsvendor problem and compare empirical coverage results.

4 An example: newsvendor problem

The newsvendor problem is a classical example of a stochastic program with simple recourse and its properties are well known, e.g., [3, p.15]. We briefly review its formulation. Let r be the selling price of a newspaper, $0 < c < r$

be its cost to the vendor, and $\tilde{\xi}$ denote the nonnegative random demand. The vendor's problem is to find the number of papers to buy, x, so that the expected profit is maximized. So, the problem is formulated as $\max\{-cx + rE\min\{x,\ \tilde{\xi}\} : x \geq 0\}$ and its solution is given by x^* that solves $\inf_{x \geq 0} P(\tilde{\xi} \leq x) \geq (r-c)/r$, which is simply $\int_0^{x^*} dF(\xi) = (r-c)/r$, when the demand distribution is continuous with cumulative distribution function F. Note that the newsvendor problem is of the form (SP) with $f(x, \tilde{\xi}) = cx - r\min\{x,\ \tilde{\xi}\}$ and $X = \{x : x \geq 0\}$.

We assume $\tilde{\xi} \sim U(0,b)$, $b > 0$ and hence modify X to $\{x : 0 \leq x \leq b\}$. Note that (A1)-(A3) hold. To perform the tests, we set $\alpha = 0.10$. For the problem parameters, we use $c = 5$, $r = 15$ and $b = 10$. This problem has optimal solution $x^* = 6\frac{2}{3}$ with expected profit $z^* = 33\frac{1}{3}$. For the candidate solution $\hat{x}$, we pick a solution that has expected profit 10% from the optimum. We use $\hat{x} = 8.775$ with $Ef(\hat{x}, \tilde{\xi}) = 30$ and with an optimality gap of $\mu_{\hat{x}} = 3\frac{1}{3}$. For the SRP, we construct 100,000 confidence intervals and for the MRP, we take $n_g = 30$ and construct 10,000 intervals for each value of the sample size. We take sample sizes, n, between 50 and 1,000. The table below summarizes the results. We report "coverage", i.e., the proportion, $\hat{p}$, of CIs containing the optimality gap and the half width, $1.645(\hat{p}(1-\hat{p})/k)^{1/2}$, of a two-sided 90% CI for the true coverage probability, where k =10,000 for MRP and 100,000 for SRP. For example, when n =1,000 for the MRP the table indicates $\hat{p} = 0.9267$ so that we are confident at level 0.90 that the coverage probability, i.e., the left-hand side of (4), is in $[0.9224, 0.9310]$.

n	MRP	SRP
50	0.9873 ± 0.0018	0.8756 ± 0.0017
100	0.9741 ± 0.0026	0.8895 ± 0.0016
200	0.9594 ± 0.0032	0.8898 ± 0.0016
300	0.9483 ± 0.0036	0.8946 ± 0.0016
400	0.9390 ± 0.0039	0.8944 ± 0.0016
500	0.9359 ± 0.0040	0.8937 ± 0.0016
600	0.9350 ± 0.0041	0.8962 ± 0.0016
700	0.9299 ± 0.0042	0.8960 ± 0.0016
800	0.9287 ± 0.0042	0.8959 ± 0.0016
900	0.9317 ± 0.0041	0.8970 ± 0.0016
1000	0.9267 ± 0.0043	0.8970 ± 0.0016

Empirical coverage results: $\hat{p} \pm 1.645(\hat{p}(1-\hat{p})/k)^{1/2}$ for various values of n, where k =10,000 for MRP and 100,000 for SRP.

The coverage for the MRP exceeds the desired coverage of 90% but shrinks toward 90% as the sample size increases. The bias, $Ez_n^* - z^*$, constitutes a major part of the CI formed by MRP and thus this CI tends to overestimate the optimality gap. As indicated in Section 2, the bias shrinks as n increases and the coverage of MRP falls as n grows. The SRP, on the other hand, has slightly less than the desired coverage of 90%. Even though the bias is larger when the sample size is small, the number of times a single replication CI contains the optimality gap approaches 90% from below. We explain this as follows. $G_n(\hat{x})$ is more variable for small sample sizes, and we have observed from the individual replications that when it is small, $s_n(x_n^*)$ also tends to be small, resulting in a narrow CI width. In particular, this happens when x_n^* is close to $\hat{x}$, even though $\hat{x}$ is not close to x^*. In ongoing research we further examine this effect and other alternatives to the single replication procedure that improve empirical performance.

References

[1] Attouch H., Wets R.-B. (1981). *Approximation and convergence in nonlinear optimization.* In O. Mangasarian, R. Meyer, and S. Robinson (Eds.). Nonlinear Programming 4, Academic Press, New York, 367 – 394.

[2] Beale E. (1955). *On minimizing a convex function subject to linear inequalities.* Journal of the Royal Statistical Society **17B**, 173 – 184.

[3] Birge J., Louveaux F. (1997). *Introduction to Stochastic Programming.* Springer-Verlag, New York.

[4] Dantzig G. (1955). *Linear programming under uncertainty.* Management Science **1**, 197 – 206.

[5] Dupačová J., Wets R. (1988). *Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems.* The Annals of Statistics **16**, 1517 – 1549.

[6] King A., Rockafellar R. (1993). *Asymptotic theory for solutions in statistical estimation and stochastic programming.* Mathematics of Operations Research **18**, 148 – 162.

[7] Mak W., Morton D., Wood R. (1999). *Monte Carlo bounding techniques for determining solution quality in stochastic programs.* Operations Research Letters **24**, 47 – 56.

[8] Rubinstein R., Shapiro A. (1993). *Discrete Event Systems: Sensitivity and Stochastic Optimization by the Score Function Method.* John Wiley & Sons, Chichester.

[9] Shapiro A. (1991). *Asymptotic analysis of stochastic programs.* Annals of Operations Research **30**, 169 – 186.

[10] Wallace S., Ziemba W.(Eds.) (2001). *Applications of Stochastic Programming.* MPS-SIAM Series in Optimization, in progress.

Acknowledgement: This work was partially supported by the National Science Foundation under Grant DMI-0217927.

Address: G. Bayraksan, D.P. Morton, Graduate Program in Operations Research; The University of Texas at Austin; 1 University Station, C2200; Austin, TX 78712-0292; USA

E-mail: guzinb@mail.utexas.edu, morton@mail.utexas.edu

LOW RISK FITS TO DISCRETE INCOMPLETE MULTI-WAY LAYOUTS

Rudolf Beran

Key words: Submodel selection, multiparametric asymptotics.

COMPSTAT 2004 section: Model selection, functional data analysis.

Abstract: The discrete multi-way layout is a widespread data-type associated with regression, experimental designs, gene or protein chips, digital images or videos, and more. A discrete multi-way layout has a finite number of factor level combinations. The layout may be unbalanced or incomplete or both. We consider candidate fits to an incomplete layout that are least squares fits to certain submodels induced by tensor product space ANOVA models for a complete layout. The candidate estimator with smallest estimated risk is selected. Multiparametric asymptotics under a general (saturated) Gaussian model show that the selected estimator achieves smallest asymptotic risk over the candidate class through bias-variance trade-off.

1 Introduction

Each factor that influences the responses in a *discrete* multi-way layout has a finite number of levels, as in classical experimental design, digital images or videos, gene or protein chips, and regression. The factors can be either ordinal or nominal or some of each. Pioneering results on low-risk fits to multi-way layouts include Stein's [8] shrinkage estimators for complete balanced discrete two-way layouts with both factors nominal and Mallow's [7] study of C_p as a criterion for selecting a submodel fit. Tukey's [10] computational experiments in discrete one- or two- or three-way layouts with ordinal factors indicated that smoothing can bring out pattern. Beran [1], [2] studied low risk adaptive penalized least squares fits to a discrete one-way layout whose factor is either ordinal or nominal.

Related in spirit are spline estimators of a mean function on a one-way layout with a *continuous* ordinal factor that is observed at discrete points [11], [5] and smoothing spline tensor product space ANOVA techniques for functional data analysis [12], [6]. The methods in this paper are designed for large, incomplete, *discrete* multi-way layouts with little or no replication. Regression models are a leading case.

As candidate estimators, we consider least squares fits to submodels for the observed incomplete layout that are induced by tensor product space ANOVA submodels for an associated complete multi-way layout of means. We estimate the risk of each candidate estimator under a general model that does *not* assume correctness of any submodel. Finally, we select the candidate fit that minimizes estimated risk. We show that the selected submodel fit has relatively low asymptotic risk under the general model as the number of observed factor-level combinations tends to infinity. This low asymptotic risk is achieved through variance-bias trade-off.

2 Candidate fits to the saturated model

Consider k_0 factors, either nominal or ordinal, in which factor k has p_k distinct levels. Let $\mathcal{I}$ denote the set of all k_0-tuples $i = (i_1, i_2, \ldots, i_{k_0})$ such that $1 \leq i_k \leq p_k$ for $1 \leq k \leq k_0$. The component i_k indexes the levels of factor k. A *complete* k_0-way layout of means consists of real values $\{m_i : i \in \mathcal{I}\}$. We order the $p = \prod_{k=1}^{k_0} p_k$ elements of the index set $\mathcal{I}$ in mirrored dictionary order: i_{k_0} serves as the first "letter" of the word, i_{k_0-1} as the second "letter", and so forth. Hereafter we assume that $\mathcal{I}$ is so ordered. Taken in this order, the indexed means for the complete multi-way layout form the $p \times 1$ vector

$$m = \{\ldots\{\{m_{i_1,i_2,\ldots,i_{k_0}} : 1 \leq i_1 \leq p_1\}, 1 \leq i_2 \leq p_2\}, \ldots, 1 \leq i_{k_0} \leq p_{k_0}\}. \quad (1)$$

Observations are available on the means $\{m_i : i \in \mathcal{I}_0\}$, where $\mathcal{I}_0$ is a subset of $\mathcal{I}$. In general, these observations $y = \{\{y_{ij} : 1 \leq j \leq n_i\}, i \in \mathcal{I}_0\}$ form an *incomplete unbalanced* k_0-way layout. The vector y is $n \times 1$ with $n = \sum_{i \in \mathcal{I}_0} n_i$. Let $q \leq p$ denote the cardinality of $\mathcal{I}_0$. Define the means-incidence matrix D to be the $q \times p$ matrix of zeroes and ones such that $m_D = Dm$ lists, in vector form, the means $\{m_i : i \in \mathcal{I}_0\}$ for the observed incomplete k_0-way layout. Let C be the $n \times q$ data-incidence matrix that suitably replicates components of the vector $m_D = Dm$ into the vector $\eta = \mathrm{E}(y) = Cm_D$. For a complete layout of data, q equals p and D is just the identity matrix. The general Gaussian *saturated model* for the incomplete layout of observations y puts no restrictions on the mean vector m:

$$y \sim N(\eta, \sigma^2 I_n), \text{ where } \eta = Cm_D = CDm, \quad m \in R^p. \quad (2)$$

2.1 Generic candidate submodel fits

Consider a submodel of the saturated model in which m is restricted to a given subspace of R^p. This condition has several mathematical expressions.

Theorem 2.1. *Let V be a $p \times r$ matrix whose columns form an orthonormal basis for the r-dimensional subspace $\mathcal{V}$. Let $Q = VV'$. Then the following assertions are equivalent: (a) $m \in \mathcal{V}$; (b) $m = V\gamma$ for some $\gamma \in R^r$; (c) $m = Q\beta$ for some $\beta \in R^p$; (d) $m = Qm$.*

The symmetric idempotent matrix Q is the unique orthogonal projection of R^p into $\mathcal{V}$. Its eigenvalues are either zero or one. The expression $Q = VV'$ is a spectral decomposition of Q. Because Q has eigenvalue one r times and eigenvalue zero $p - r$ times, there exist many eigenvector matrices V such that $Q = VV'$.

For Q a projection as in the Theorem, define *submodel*(Q) by imposing the constraint $m = Qm$ on the saturated model (2). Taking Q to be the identity matrix I_p recovers the saturated model itself. Let $\eta(Q) = \mathrm{E}(y)$, evaluated under submodel(Q). By the foregoing Theorem, $\eta(Q) = CDQ\beta = CDV\gamma$. Consequently, the least squares estimator of $\eta(Q)$ under submodel(Q) has several equivalent expressions in which the superscript $^+$ denotes the Moore-Penrose inverse.

Theorem 2.2. *Let $M(Q) = CDQ(CDQ)^+ = CD(CDQ)^+ = CDV(CDV)^+$ and let $M = M(I_p) = CD(CD)^+ = CC^+$. The least squares estimator of $\eta(Q)$ under submodel(Q) is $\hat{\eta}(Q) = M(Q)y$. In particular, the least squares estimator of η under the saturated model is $\hat{\eta} = My$.*

This result can be derived thorough properties of the Moore-Penrose inverse. The matrices $M(Q)$ and M are both symmetric idempotent and satisfy $MM(Q) = M(Q) = M(Q)M$.

For each Q in a class of projection matrices that express tentative prior conjectures about m, we will consider $\hat{\eta}(Q)$ as a (usually biased) candidate estimator for η in the saturated model for the unbalanced incomplete k_0-way layout. The associated candidate estimator for the cell means $m_D = Dm$ is then $\hat{m}_D(Q) = C^+\hat{\eta}(Q)$. Although submodel fits generate the candidate estimators, it is not assumed in this paper that any submodel of the saturated model (2) holds.

2.2 Tensor product candidate submodel fits

To generate useful candidate projections Q, we first express the ANOVA decomposition for a complete k_0-way layout of means in projection form. Let $\mathcal{S}$ denote the set of all subsets of $\{1, 2, \ldots, k_0\}$, including the empty set $\emptyset$. The cardinality of $\mathcal{S}$ is 2^{k_0}.

For every k, define the $p_k \times 1$ vector u_k and the $p_k \times p_k$ matrices J_k, H_k by $u_k = p_k^{-1/2}(1, 1, \ldots, 1)'$, $J_k = u_k u_k'$, and $H_k = I_{p_k} - J_k$. For each k, the symmetric idempotent matrices J_k and H_k have rank (or trace) 1 and $p_k - 1$ respectively. Let U_k be any $p_k \times (p_k - 1)$ matrix such that $(u_k|U_k)$ is an orthogonal matrix. Then the foregoing entails that $H_k = U_k U_k'$.

For every set $S \in \mathcal{S}$, define $P_{S,k} = J_k$ if $k \notin S$ and $P_{S,k} = H_k$ if $k \in S$. Define the $p \times p$ Kronecker product matrix $P_S = \bigotimes_{k=1}^{k_0} P_{S,k_0-k+1}$. Evidently P_S is symmetric idempotent for every $S \in \mathcal{S}$; if $S \neq \emptyset$, the rank (or trace) of P_S is $\prod_{k \in S}(p_k - 1)$; the rank (or trace) of $P_\emptyset$ is 1; if S_1 and S_2 are two different sets in $\mathcal{S}$, then $P_{S_1}P_{S_2} = 0 = P_{S_2}P_{S_1}$; and $\sum_{S \in \mathcal{S}} P_S = I_p$.

Consequently, the $\{P_S \colon S \in \mathcal{S}\}$ are orthogonal projections that decompose R^p into 2^{k_0} mutually orthogonal subspaces. The ANOVA decomposition of a complete k_0-way layout of means is the identity, for every $m \in R^p$,

$$m = \sum_{S \in \mathcal{S}} P_S m. \tag{3}$$

Here $P_\emptyset m$ is the overall mean term in the decomposition. If $S \neq \emptyset$, $P_S m$ is the main effect or interaction term defined by the factors $k \in S$. This

ANOVA decomposition suggests a rich variety of choices for the projection matrix Q that determines submodel(Q) of the saturated model (2) for the incomplete k_0-way layout. The central ideas are as follows:

All factors nominal. Let $\{\mathcal{A}_j \colon 1 \leq j \leq j_0\}$ denote a collection of subsets of $\mathcal{S}$ that is partially ordered under the inclusion operation. Let

$$Q_{\mathcal{A}_j} = \sum_{S \in \mathcal{A}_j} P_S \tag{4}$$

and let $\mathcal{Q} = \{Q_{\mathcal{A}_j} \colon 1 \leq j \leq j_0\}$. The candidate estimators $\{\hat{\eta}(Q) \colon Q \in \mathcal{Q}\}$ for η are least squares fits to the designated ANOVA submodels.

All factors ordinal. Without loss of generality, assume that the indexing of the levels of an ordinal factor follows their numerical order. Prior conjecture may then be that adjacent means in the k_0-way layout vary smoothly as a function of the ordinal factor levels. Tensor product space submodels that express this are described through the following general scheme.

Let $W_k(c_k)$ be any $p_k \times c_k$ matrix of rank $c_k \leq p_k$ whose first column is proportional to u_k and whose successive columns are increasingly wiggly. For instance, the columns of $W_k(c_k)$ could be successive powers, from zero to $c_k - 1$, of the vector of levels of factor k. Define the orthogonal projections $G_k(c_k) = W_k(c_k)W_k(c_k)^+$ and $K_k(c_k) = G_k(c_k) - J_k$. Define $P_{S,k} = J_k$ if $k \notin S$ and $P_{S,k} = K_k(c_k)$ if $k \in S$. For $c = (c_1, c_2, \ldots, c_{k_0})$, define $P_S(c) = \bigotimes_{k=1}^{k_0} P_{S,k_0-k+1}$.

In analogy to the preceding paragraph, let

$$Q_{\mathcal{A}_j}(c) = \sum_{S \in \mathcal{A}_j} P_S(c) \tag{5}$$

and let $\mathcal{Q} = \{Q_{\mathcal{A}_j}(c) \colon 1 \leq j \leq j_0, 1 \leq c_1 \leq d_1, 1 \leq c_2 \leq d_2, \ldots 1 \leq c_{k_0} \leq d_{k_0}\}$. The estimators $\{\hat{\eta}(Q) \colon Q \in \mathcal{Q}\}$ are a class of candidate submodel estimators for η.

Some factors nominal, some factors ordinal. We proceed in similar fashion, choosing the multiplicands $P_{S,k}$ in the cross-factor Kronecker product according to the nominal or ordinal nature of each factor, as above.

3 Adaptive low risk submodel fits

We will assess the performance of the candidate estimator $\hat{\eta}(Q)$ of $\eta = CDm$ through its risk under normalized quadratic loss, $L(\hat{\eta}(Q), \eta) = q^{-1}|\hat{\eta}(Q) - \eta|^2$. Let $r(Q) = \text{rank}(M(Q)) = \text{rank}(CDQ)$. The risk of candidate estimator $\hat{\eta}(Q) = M(Q)y$ under the saturated model is then

$$R(\hat{\eta}(Q), \eta, \sigma^2) = \mathrm{E}L(\hat{\eta}(Q), \eta) = q^{-1}[\sigma^2 r(Q) + |M(Q)\eta - \eta|^2]. \tag{6}$$

Recall that $\hat{\eta} = My$ is the least squares estimator of η under the saturated model. Let denote $\hat{\sigma}^2$ denote a consistent estimator of σ^2. Apart from possible bias in $\hat{\sigma}^2$, Mallow's [7] C_p criterion and Stein's [9] unbiased risk estimator both yield the estimator

$$\hat{R}(\hat{\eta}(Q)) = q^{-1}[|\hat{\eta} - \hat{\eta}(Q)|^2 + (2r(Q) - q)\hat{\sigma}^2] \tag{7}$$

for the risk (6) of $\hat{\eta}(Q)$ under the saturated model.

The *pooling* variance estimator $\hat{\sigma}^2_P$, potentially useful when n equals or is not much larger than q, is $\hat{\sigma}^2_P = [n - r(Q_L)]^{-1}|y - \hat{\eta}(Q_L)|^2$, where Q_L is a projection in $\mathcal{Q}$ with rank $r(Q_L)$. This biased estimator is consistent for σ^2 when $n - r(Q_L)$ tends to infinity and $[n - r(Q_L)]^{-1}|\eta - \eta(Q_L)|^2$ tends to zero.

The *first-difference* variance estimator $\hat{\sigma}^2_{FD}$ is potentially useful when $n = q$, in which case the data forms an incomplete multi-way layout with one observation per cell. Form all possible first differences of adjacent $\{y_{i1}\}$ along each coordinate direction. Then $\hat{\sigma}^2_{FD}$ is defined to be one-half of the average of the squared first differences. This biased estimator is consistent for σ^2 when q tends to infinity and the quantity obtained by replacing each y_{i1} in $\hat{\sigma}^2_{FD}$with m_i tends to zero.

Let $\mathcal{Q}$ denote a finite class of candidate projections Q, constructed as in the previous section, whose cardinality may depend on q. The *adaptive estimator* of η is defined to be the candidate estimator with smallest estimated risk:

$$\hat{\eta}_{\mathcal{Q}} = \hat{\eta}(\hat{Q}), \text{ where } \hat{Q} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \hat{R}(\hat{\eta}(Q)). \tag{8}$$

Asymptotic analysis supports the claim that the risk of $\hat{\eta}_{\mathcal{Q}}$ approximately minimizes risk among all candidate estimators $\{\hat{\eta}(Q) \colon Q \in \mathcal{Q}\}$.

Theorem 3.1. *Let $\#\mathcal{Q}$ be the number of projections in the class $\mathcal{Q}$. Suppose that $\lim_{q\to\infty} q^{-1/2}\#\mathcal{Q} = 0$ and $\lim_{q\to\infty} \#\mathcal{Q} \sup_{|\eta|^2 \le qa\sigma^2} \mathrm{E}|\hat{\sigma}^2 - \sigma^2| = 0$ for every $a > 0$ and $\sigma^2 > 0$. Then, for W equal to either $R(\hat{\eta}_{\mathcal{Q}}, \eta, \sigma^2)$ or $L(\hat{\eta}_{\mathcal{Q}}, \eta)$,*

$$\lim_{q\to\infty} \sup_{|\eta|^2 \le q\sigma^2 a} \mathrm{E}|W - \min_{Q\in\mathcal{Q}} R(\hat{\eta}(Q), \eta, \sigma^2)| = 0 \tag{9}$$

and

$$\lim_{q\to\infty} \sup_{|\eta|^2 \le q\sigma^2 a} \mathrm{E}|W - \hat{R}(\hat{\eta}_{\mathcal{Q}})| = 0. \tag{10}$$

Under the multiparametric asymptotics of this theorem, in which the number q of unknown means tends to infinity, the loss and corresponding risk of a candidate estimator converge to a common value. According to (9), the risk and loss of the adaptive estimator $\hat{\eta}_{\mathcal{Q}}$ converge to the minimum risk achievable over the class of candidate estimators $\{\hat{\eta}(Q) \colon Q \in \mathcal{Q}\}$. Equation (10) asserts that the estimated risk of the adaptive estimator $\hat{\eta}_{\mathcal{Q}}$ is a trustworthy asymptotic approximation to both its actual risk and loss.

Proof idea. Matrices $M(Q)$ and M are each symmetric and idempotent and satisfy $MM(Q) = M(Q) = M(Q)M$. Evidently $\operatorname{rank}(M) = \operatorname{tr}[C(C'C)^{-1}C'] = q$. The matrix $M - M(Q)$ is therefore symmetric and idempotent; has rank $q - r(Q)$; and satisfies $M(Q)[M - M(Q)] = 0$. We have the spectral decompositions $M(Q) = U_Q U_Q'$ and $M - M(Q) = \bar{U}_Q \bar{U}_Q'$, where U_Q is $n \times r(Q)$, $\bar{U}_Q$ is $n \times (q - r(Q))$, $U_Q' U_Q = I_{r(Q)}$, $\bar{U}_Q' \bar{U}_Q = I_{q-r(Q)}$, and $U_Q' \bar{U}_Q = 0$. It follows that the $n \times q$ matrix $U = (U_Q | \bar{U}_Q)$ is orthogonal and that $M = U_Q U_Q' + \bar{U}_Q \bar{U}_Q' = UU'$.

Let $z = U'y$ and $\xi = U'\eta$. Under the saturated model, z has a $N(\xi, \sigma^2)$ distribution and $\eta = M\eta = U\xi$. Let f_Q denote the q-dimensional vector whose first $r(Q)$ components equal 1 and whose other components equal 0. Let $F_Q = \operatorname{diag}\{f_Q\}$. By Theorem 1 and the preceding paragraph, $\hat{\eta}(Q) = U_Q U_Q' y = U F_Q z$ and $\hat{\eta} = UU'y = Uz$. For any vector h, let $\operatorname{ave}(h)$ denote the average of the components of h. For any two vectors g and h of the same dimension, let gh denote the vector formed by componentwise multiplication. Equations (6), (7), and the preceding notations yield the canonical forms $L(\hat{\eta}(Q), \eta) = q^{-1} |f_Q z - \xi|^2$, $R(\hat{\eta}(Q), \eta, \sigma^2) = \operatorname{ave}[f_Q^2 \sigma^2 + (1 - f_Q)^2 \xi^2]$, and $\hat{R}(\hat{\eta}(Q)) = \operatorname{ave}[f_Q^2 \hat{\sigma}^2 + (1 - f_Q)^2 (z^2 - \hat{\sigma}^2)]$ for loss, risk, and estimated risk.

Let V denote either loss $L(\hat{\eta}(Q))$ or the estimated risk $\hat{R}(\hat{\eta}(Q))$. Applying Theorem 2.1 in [3] to the preceding canonical forms establishes existence of a finite constant C such that

$$\mathrm{E}|V - R(\hat{\eta}(Q), \eta, \sigma^2)| \le C[q^{-1/2}(\sigma^2 + \sigma\{\operatorname{ave}(\xi)^2\}^{1/2}) + \mathrm{E}|\hat{\sigma}^2 - \sigma^2|]. \quad (11)$$

for every projection $Q \in \mathcal{Q}$. An argument based on this inequality yields the Theorem conclusions.

4 Example: coal ash data

The coal ash data from [4, p. 34], records percentage of coal ash in 208 assay samples. The data forms an incomplete two-way layout with one observation for each coordinate pair at which an assay sample is obtained. The factors row coordinate and column coordinate are both ordinal and range over $p_1 = 23$ and $p_2 = 16$ equally spaced levels respectively. It seems likely, a priori, that the coal ash means vary smoothly with geographical location. We construct candidate projections by the tensor product space method previously described, using the discrete cosine basis: For $1 \le k \le 2$, the first column of $W_k(c_k)$ is the vector u_k already defined and the succeeding columns are

$$w_{kc} = \{(2/p_k)^{1/2} \cos[(2r-1)(c-1)\pi/(2p_k)] \colon 1 \le r \le p_k\}, \quad 2 \le c \le c_k. \quad (12)$$

Let $\mathcal{A}_1 = \{\emptyset, \{1\}, \{2\}\}$ and let $\mathcal{A}_2 = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$. Let $\mathcal{C} = \bigcup_{j=2}^{8} \{(j,j)\}$. For every $c \in \mathcal{C}$, define $Q_{\mathcal{A}_j}(c)$ by equation (5) and its preceding paragraph and let $\mathcal{Q} = \{Q_{\mathcal{A}_j}(c) \colon 1 \le j \le 2, c \in \mathcal{C}\}$. For the coal-ash data, the first difference variance estimate is $\hat{\sigma}_{FD}^2 = 1.038$.

The following table lists estimated risks for the candidate estimators $\{\hat{\eta}(Q) \colon Q \in \mathcal{Q}\}$:

c	(2,2)	(3,3)	(4,4)	(5,5)	(6,6)	(7,7)	(8,8)
$Q_{\mathcal{A}_1}(c)$	.213	.232	.150	.148	.134	.151	.155
$Q_{\mathcal{A}_2}(c)$	.222	.243	.192	.238	.261	.333	.385

The additive submodel fit $\hat{\eta}(Q_{\mathcal{A}_1}(6,6))$ has smallest estimated risk among these competitors. Moreover, its estimated risk .134 is about one-eighth of the estimated risk 1.038 of the least squares fit to the saturated model. The latter fit coincides with the raw data in this example and is clearly not useful. Through variance-bias trade-off, our data-driven algorithm selects a submodel fit that is visually appealing, has the structural simplicity of additivity, and has much lower risk under the saturated model (see the preceding theorem). This adaptive submodel fit compresses the data so as to discard more noise than signal.

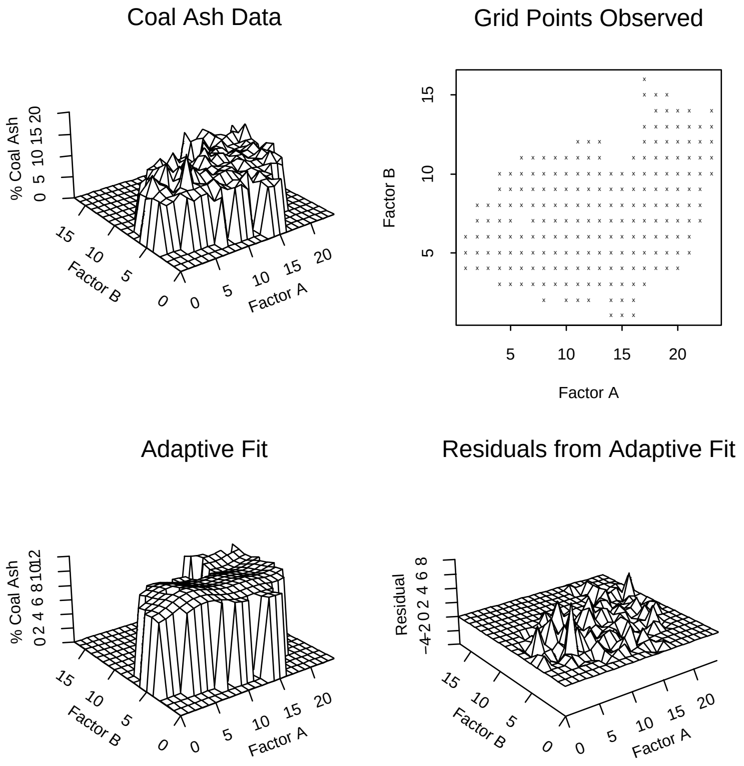


Figure 1: The Coal Ash data, its factor level grid, the low risk adaptive submodel fit $\hat{\eta}(Q_{\mathcal{A}_1}(6,6))$ using the discrete cosine basis, and residuals.

References

[1] Beran R., (2000). *REACT scatterplot smoothers: superefficiency through basis economy.* Journal of the American Statistical Association **63**,155 – 171.

[2] Beran R. (2002). *Improving penalized least squares through adaptive selection of penalty and shrinkage.* Annals of the Institute of Statistical Mathematics **54**, 900 – 917.

[3] Beran R., Dümbgen L. (1998). *Modulation of estimators and confidence sets.* Annals of Statistics **26**, 1826 – 1856.

[4] Cressie N. A. (1993). *Statistics for spatial data* (revised edition). Wiley, New York.

[5] Heckman N.E., Ramsay J.O. (2000). *Penalized regression with model-based penalties.* Canadian Journal of Statistics **28**, 241 – 258.

[6] Lin Yi (2000). *Tensor product space ANOVA fits.* Annals of Statistics **28**, 734 – 755.

[7] Mallows C.L. (1973). *Some comments on C_p.* Technometrics **15**, 661 – 676.

[8] Stein C. (1966). *An approach to the recovery of inter-block information in balanced incomplete block designs.* Festschrift for Jerzy Neyman F.N. David (ed.), Wiley, New York, 351 – 364.

[9] Stein C. (1981). *Estimation of the mean of a multivariate normal distribution.* Annals of Statistics **9**, 1135 – 1151.

[10] Tukey J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading MA.

[11] Wahba G. (1990). *Spline models for observational data.* Society for Industrial and Applied Mathematics, Philadelphia.

[12] Wahba G., Wang Y., Gu C., Klein R., Klein B. (1995). *Smoothing spline ANOVA for exponential families with application to the Wisconsin epidemiological study of diabetic retinopathy.* Annals of Statistics **23**, 1868 – 1895.

Acknowledgement: This research was supported in part by National Science Foundation Grant DMS 0300806.

Address: R. Beran, Department of Statistics, University of California, Davis, CA 95616, USA

E-mail: beran@wald.ucdavis.edu

APPROXIMATE REGENERATIVE BLOCK-BOOTSTRAP FOR MARKOV CHAINS

Patrice Bertail and Stéphan Clémençon

Key words: Bootstrap, Markov chains, Edgeworth expansions, Nummelin splitting, algorithms.

COMPSTAT 2004 section: Resampling method.

Abstract: In this paper we propose a modification of the original ARBB algorithm based on the "2-split" method considered by Schick [10]. We also show how the asymptotic results obtained for the RBB in the regenerative case may be extended to this modified ARBB procedure at the cost of some small loss in the Edgeworth expansions, which is closely linked to the uniform rate for estimating the transition kernel of the chain over a well chosen small set.

1 Introduction

Prolongating ideas introduced in Datta & McCormick [4], Bertail & Clémençon [1], [2] proposed a general resampling method, namely the *Regenerative Block Bootstrap* (RBB in abbreviated form), for bootstrapping statistics based on data $X_1, \ldots, X_n$ drawn from (eventually nonstationary) regenerative Markov chains. When the chain (positive Harris recurrent) possesses a known atom, they proved that this resampling method is second order correct up to $O_P(n^{-1})$ in the case of the studentized sample mean statistic under specific Cramer and "block moment" conditions (less restrictive than the exponential strong mixing rate condition generally assumed when the matter is to deal with dependent data). This is the optimal rate that may be attained by the naive Bootstrap method in the i.i.d. case (see Hall [6]). These results should be put in contrast with the usual rates that may be attained by the *Moving Block Bootstrap* (MBB), which are at best $O_P(n^{-3/4})$ (see Götze & Künsch [5]). We emphasize that the RBB straightforwardly applies to numerous specific regenerative models, widely used in the modeling of queuing and storage systems, and to all countable Markov chains. Resting on the theoretical construction introduced by Nummelin [9], namely the *Nummelin splitting technique*, which is based on the crucial notion of *small set* (cf. [8]), any general Harris Markov chain could be considered as regenerative in the sense of the existence of a regenerative extension. Bertail & Clémençon [2] proposed a resampling procedure, the *Approximate Regenerative Block Bootstrap* (ARBB), that generalizes the RBB method and applies to all Harris Markov chains. The method is based on the prior knowledge of a small set for the chain and a practical approximation of the *Nummelin splitting extension*. It thus consists in using an empirical method to build approximatively a realization drawn from a regenerative extension of the chain and in applying the RBB methodology to the latter.

The outline of the paper is as follows. In section 2 the principles of the ARBB are briefly recalled and a modification of the original method using a variant of the "2-split" trick is presented. In section 3 an asymptotic result claiming the second order asymptotic validity of this ARBB method for studentized sample mean statistics is stated. Finally, in section 4, practical selection rules for the tuning parameters of the algorithm are proposed and some simulation results are presented.

2 Nummelin splitting approximation and ARBB

2.1 Notation and basic notions

Here and throughout we shall use the same notations as in section 2 of Bertail & Clémençon [2]. Consider $X = (X_n)_{n\in\mathbb{N}}$ a positive recurrent Markov chain on a countably generated state space $(E, \mathcal{E})$ with transition probability $\Pi(.,.)$, stationary probability measure μ and initial distribution ν. We denote by P_ν (respectively P_x for x in E, resp. P_A for $A \in \mathcal{E}$) the probability measure on the underlying space such that $X_0 \sim \nu$ (resp. conditionally to $X_0 = x$, resp. conditionally to $X_0 \in A$), by $E_\nu(.)$ the P_ν-expectation (resp. by $E_x(.)$ the P_x-expectation, resp. by $E_A(.)$ the P_A-expectation) and by $I\{\mathcal{A}\}$ the indicator function of the event $\mathcal{A}$.

We recall that a set $S \in \mathcal{E}$ is said to be *small* (see Meyn & Tweedie [8]) if there exist $k \in \mathbb{N}$, a probability measure Φ supported by S, and $\delta > 0$ such that $\forall x \in S, \forall A \in \mathcal{E}, \quad \Pi^k(x, A) \geq \delta\Phi(A)$, denoting by Π^k the k-th iterate of Π (recall that small sets always exist for irreducible chains). When this holds, we shall say that X satisfies the minorization condition $\mathcal{M}(k, S, \delta, \Phi)$. Even if it entails to replace the chain $(X_n)_{n\in\mathbb{N}}$ by $\big((X_{nk}, \dots, X_{n(k+1)-1})\big)_{n\in\mathbb{N}}$, we suppose $k = 1$ in what follows. We assume further that the family of the conditional distributions $\{\Pi(x, dy)\}_{x\in E}$ and the initial distribution ν are dominated by a σ-finite measure λ of reference, so that $\nu(dy) = f(y)\lambda(dy)$ and $\Pi(x, dy) = p(x, y)\lambda(dy)$ for all $x \in E$. In this case, the condition $\mathcal{M}(k, S, \delta, \Phi)$ entails that Φ is also absolutely continuous with respect to λ and $p(x, y) \geq \delta\phi(y)$, $\lambda(dy)$ a.s, for any $x \in S$, with $\Phi(dy) = \phi(y)d\lambda(y)$. We assume

$\mathcal{H}_0$: The chain X satisfies condition $\mathcal{M}(1, S, \delta, \Phi)$ for some known parameters $S \in \mathcal{E}$ such that $\mu(S) > 0$, $\delta > 0$ and probability $\Phi(dy) = \phi(y)d\lambda(y)$ supported by S such that $\inf_{y\in S}\phi(y) > 0$.

The Nummelin splitting technique consists in constructing a bivariate Markov chain $X^{\mathcal{M}} = ((X_n, Y_n))_{n\in\mathbb{N}}$, called the *split chain*, taking its values in the state space $E \times \{0, 1\}$. This construction entails that, conditionally to $X^{(n+1)} = (X_1, \dots, X_{n+1})$, the Y_i's, $1 \leqslant i \leqslant n$, are independent Bernoulli r.v.'s. The Bernoulli parameter is δ, unless X has hit the small set S at time i. And in the case when $X_i \in S$, Y_i is drawn from the Bernoulli distribution with parameter $\delta\phi(X_{i+1})/p(X_i, X_{i+1})$. We denote by $\mathcal{L}^{(n)}(p, S, \delta, \phi, X^{(n+1)})$

the probability distribution of $Y^{(n)} = (Y_1, \ldots, Y_n)$ conditionally to $X^{(n+1)}$, which is simply the tensor product of these Bernoulli distributions. The whole point of the construction consists in the fact that $A_{\mathcal{M}} = S \times \{1\}$ is an atom for the split chain $X^{\mathcal{M}}$, which inherits all the communication and stochastic stability properties from X. In particular, the sample path of $X^{\mathcal{M}}$ can be classically divided into *regeneration blocks* corresponding to the blocks of observations between successive visits of the split chain to $A_{\mathcal{M}}$, which are i.i.d. r.v.'s valued in the torus $T = \cup_{n=1}^{\infty} E^n$, by virtue of the strong Markov property. For a given time $m^* \in \mathbb{N}$ that will be fixed later, we shall here consider the regeneration times

($i.e.$ the times i at which $X_i \in S$ and $Y_i = 1$) posterior to m^*, which are denoted by $\tau_{m^*} \doteq \tau_{m^*}(1) = \inf\{k \geqslant m^* + 1/\ X_k \in S, Y_k = 1\}$, $\tau_{m^*}(j) = \inf\{k > \tau_{m^*}(j-1)/\ X_k \in S, Y_k = 1\}$ for $j \geqslant 2$. We denote by $l_{m^*,n} = \sum_{i=m^*+1}^{n} I\{X_i \in S, Y_i = 1)$ the number of visits to the set $A_{\mathcal{M}} = S \times \{1\}$ between time $m^* + 1$ and time n. The corresponding regeneration blocks are denoted by $\mathcal{B}_{0,m^*} = (X_{m+1}, \ldots, X_{\tau_{m^*}(1)})$, $\mathcal{B}_{1,m*} = (X_{\tau_{m^*}(1)+1}, \ldots, X_{\tau_{m^*}(2)}), \ldots, \mathcal{B}_{l_{m^*,n}-1,m^*} = (X_{\tau_{m^*}(l_{m^*,n}-1)+1}, \ldots, X_{\tau_{m^*}(l_{m^*,n})})$, $\mathcal{B}^{(n)}_{l_{m^*,n},m^*}= (X_{\tau_{m^*}(l_{m^*,n})+1}, \ldots, X_n)$.

2.2 Approximate Nummelin splitting construction

Of course these blocks are practically unknown since their construction explicitly depends on the unknown transition density $p(x, y)$ (see § 2.1). The proposal of Bertail & Clémençon [2] for approximating this construction consists in using an estimate $p_n(x, y)$ of the transition density computed from data $X_1, \ldots, X_n$ to generate a random vector $(\widehat{Y}_1, \ldots, \widehat{Y}_n)$, conditionally to $X^{(n+1)}$, drawn from the distribution $\mathcal{L}^{(n)}(p_n, S, \delta, \phi, X^{(n+1)})$. However this estimation step induces strong dependency problems that make the second order properties of the ARBB procedure very difficult to study, when applied to the data $(X_1, \widehat{Y}_1), \ldots, (X_n, \widehat{Y}_n)$. Here we propose a modification of the method based on the well known semiparametric "splitting trick".

Given the data $X^{(n+1)}$, keep the first m observations $X^{(m)} = (X_1, \ldots, X_m)$ only to compute an estimate $p_m(x, y)$ of $p(x, y)$ such that $p_m(x, y) \geq \delta\phi(y)$, $\lambda(dy)$ a.s. and $p_m(X_i, X_{i+1}) > 0$, $1 \leqslant i \leqslant n$. To ensure that the observations $X^{(m^*,n+1)} = (X_{m^*+1}, \ldots, X_{n+1})$, which shall be used for forming the pseudo-regeneration blocks to resample, are independent from the first m observations ($i.e.$ that a regeneration, or equivalently a visit of $X^{\mathcal{M}}$ to $A_{\mathcal{M}}$, occurs between time $m + 1$ and time m^*) with overwhelming probability, we separate them by a small gap of length p. We will typically choose $m^* = m+p$ with m, p and m^* depending on n such that $p = O(m)$ as $n \to \infty$. This procedure is very similar to the *2-split method* proposed in Schick [10], except that the user is here free to pick the exact number p of deleted observations, within the limits of the previous asymptotic constraint. In the following, we take $m \to \infty$ as $n \to \infty$, so as to get a consistent estimator $p_m(x, y)$, at a rate sufficiently slow (typically such that $\frac{m}{n} \to 0$ as $n \to \infty$) to ensure that the number of pseudo-blocks to resample also tends to infinity as $n \to \infty$.

Conditionally to $X^{(n+1)}$, draw then a vector $(\widehat{Y}_{m^*+1}, \ldots, \widehat{Y}_n)$ from the distribution estimate $\mathcal{L}^{(n-m^*)}(p_m, S, \delta, \phi, X^{(m^*, n+1)})$. From a practical viewpoint, it actually suffices to draw the $\widehat{Y}_i$'s at times i when the chain visits the set S (*i.e.* when $X_i \in S$), which are the only time points at which the split chain may regenerate: at such a time i, draw $\widehat{Y}_i$ according to the Bernoulli law with parameter $\delta\phi(X_{i+1})/p_m(X_i, X_{i+1}))$. Count then the number of visits $\widehat{l}_{m^*,n} = \sum_{i=m^*+1}^{n} I\{X_i \in S, \widehat{Y}_i = 1)$ to $A_{\mathcal{M}} = S \times \{1\}$ between time $m^* + 1$ and time n and divide the truncated sample path $X^{(m^*,n)}$ into $\widehat{l}_n + 1$ blocks, corresponding to the pieces of the data segment between consecutive visits to $A_{\mathcal{M}}$, $\widehat{\mathcal{B}}_{0,m^*} = (X_{m^*+1}, \ldots,\ X_{\widehat{\tau}_{m^*}(1)})$, $\widehat{\mathcal{B}}_{1,m^*} = (X_{\widehat{\tau}_{m^*}(1)+1}, \ldots,\ X_{\widehat{\tau}_{m^*}(2)}), \ldots,$ $\widehat{\mathcal{B}}^{(n)}_{l_{m^*,n},m^*} = (X_{\widehat{\tau}_{m^*}(\widehat{l}_{m^*,n})+1}, \ldots,\ X_n)$ with $\widehat{\tau}_{m^*}(0) = m^*$ and for any $j \geqslant 1$, $\widehat{\tau}_{m^*}(j) = \inf\left\{k > \widehat{\tau}_{m^*}(j-1),\ X_k \in S, \widehat{Y}_k = 1\right\}$. For convenience, denote by $l(\widehat{\mathcal{B}}_{j,m^*}) = \widehat{\tau}_{m^*}(j+1) - \widehat{\tau}_{m^*}(j)$ the length of the block $\widehat{\mathcal{B}}_{j,m^*}$, $j \geqslant 1$.

2.3 Approximate Regenerative Block Bootstrap

Let $T_{n+1} = T_{n+1}(X^{(n+1)})$ be a statistic of interest and $S_{n+1} = S_{n+1}(X^{(n+1)})$ be an adequate standardization of the latter. The modified ARBB algorithm (which we still call ARBB algorithm for the sake of the simplicity) consists then in applying the RBB procedure in the following manner.

1. Draw sequentially bootstrap data blocks $\mathcal{B}_1^*, \ldots,\ \mathcal{B}_k^*$ independently from the empirical distribution $F_{m^*,n} = (\widehat{l}\ _{m^*,n} - 1)^{-1} \sum_{j=1}^{\widehat{l}_{m^*,n}-1} \delta_{\widehat{\mathcal{B}}_{j,m^*}}$ of the blocks $\widehat{\mathcal{B}}_{1,m^*}, \ldots,\ \widehat{\mathcal{B}}_{\widehat{l}_{m^*,n}-1,m^*}$ conditioned on $X^{(n+1)}$, until the length of the bootstrap data series $L^*(k) = \sum_{j=1}^{k} l(\mathcal{B}_j^*)$ is larger than n. Let $l_n^* = \inf\{k \geqslant 1,\ L^*(k) > n\}$.

2. From these bootstrap data blocks, reconstruct a pseudo-trajectory by binding the blocks together, getting the reconstructed *ARBB sample path* $X^{*(n)} = (\mathcal{B}_1^*, \ldots, \mathcal{B}_{l_n^*-1}^*)$. Then compute the *ARBB statistic* $T_n^* = T_{L^*(l_n^*)}(X^{*(n)})$ and the *ARBB standardization* $S_n^* = S_{L^*(l_n^*)}(X^{*(n)})$.

3. The *ARBB distribution* is then given by $H_{ARBB}(x) = P^*(S_n^{*-1}(T_n^* - T_{n+1}) \leqslant x \mid X^{(n+1)})$, which may be approximated by a classical Monte-Carlo resampling scheme.

As shown in Bertail & Clémençon [2], the sequential resampling in step 1 allows to approximatively mimic the renewal property of the split chain and to efficiently reproduce the second order structure.

3 Second order properties for linear functionals

3.1 Basic estimators

Let $f : E \to \Re$ be a μ-integrable function. Our parameter of interest is now the unknown mean $\mu(f) = E_\mu(f(X_1))$. Although the sample mean $\mu_{n+1}(f) = (n+1)^{-1}\sum_{i=1}^{n+1} f(X_i)$ is an asymptotically normal estimator of $\mu(f)$ under simple moment conditions, we shall consider the truncated sample mean based on the data segment $(X_{\widehat{\tau}_{m^*}(1)+1}, \ldots,, X_{\widehat{\tau}_{m^*}(1_{m^*,n})})$ only (or equivalently on the blocks $\widehat{\mathcal{B}}_{1,m^*}, \ldots, \widehat{\mathcal{B}}_{\widehat{l}_{m^*,n}-1,m^*}$), since the matter is here to deal with estimators of which the distribution may be accurately approximated (refer to the discussions in Bertail & Clémençon [1], [2]). Denote by $\widehat{n} = \widehat{\tau}_{_{m^*}}(\widehat{l}_{_{m^*},n}) - \widehat{\tau}_{_{m^*}}(1) = \sum_{j=1}^{\widehat{l}_{m^*,n}-1} l(\widehat{\mathcal{B}}_{j,m^*})$ the length of this segment. Set $f(\widehat{\mathcal{B}}_{j,m^*}) = \sum_{i=1+\widehat{\tau}_{m^*}(j)}^{\widehat{\tau}_{m^*}(j+1)} f(X_i)$, $j \geqslant 1$, $\widehat{\mu}_{m^*,n}(f) = \widehat{n}^{-1}\sum_{j=1}^{\widehat{l}_{m^*,n}-1} f(\widehat{\mathcal{B}}_{j,m^*})$, $\widehat{\sigma}^2_{m^*,n}(f) = \widehat{n}^{-1}\sum_{j=1}^{\widehat{l}_{m^*,n}-1}\{f(\widehat{\mathcal{B}}_{j,m^*}) - \widehat{\mu}_{m^*,n}(f)l(\widehat{\mathcal{B}}_{j,m^*})\}^2$. It can easily be shown by using the argument of Theorems 17.2.2 and 17.3.6 in Meyn & Tweedie [8] that, under suitable block moment conditions, $\widehat{\mu}_{m^*,n}(f)$ is asymptotically normal and $\widehat{\sigma}^2_{m^*,n}(f)$ is a consistent estimator of the asymptotic variance of $\widehat{\mu}_{m^*,n}(f)$ (resp., of $\mu_{n+1}(f)$), namely $\sigma^2(f) = E_{A_\mathcal{M}}(\tau_{A_\mathcal{M}})^{-1}E_{A_\mathcal{M}}((\sum_{i=1}^{\tau_{A_\mathcal{M}}}\{f(X_i) - \mu(f)\})^2)$, where $\tau_{A_\mathcal{M}} = \inf\{k \geqslant 1/\ X_k \in S,\ Y_k = 1\}$ and $E_{A_\mathcal{M}}(.)$ denotes the conditional expectation given $(X_0, Y_0) \in S \times \{1\}$. We then define the *unstudentized mean* $\widehat{\varsigma}_n = \widehat{n}^{1/2}\frac{\widehat{\mu}_{m^*,n}(f)-\mu(f)}{\sigma(f)}$ *and the studentized mean* $\widehat{t}_n = \widehat{n}^{1/2}\frac{\widehat{\mu}_n(f)-\mu(f)}{\widehat{\sigma}_{m^*,n}(f)}$. Bertail & Clémençon [1] have shown how to obtain Edgeworth expansions up to $O(n^{-1})$ for such quantities using the same technique as in Bolthausen [3] and in Malinovskii [7].

3.2 Asymptotic validity of the ARBB

Let $P^*(.)$ denote the conditional probability under the resampling scheme described in step 1 (see § 2.3) for given $X^{(n+1)}$. Consider now the ARBB counterparts of the statistics introduced above: $\mu_n^*(f) = n^{*-1}\sum_{j=1}^{l_n^*-1} f(\mathcal{B}_j^*)$ and $\sigma_n^{*2}(f) = n^{*-1}\sum_{j=1}^{l_n^*-1}\{f(\mathcal{B}_j^*) - \mu_n^*(f)l(\mathcal{B}_j^*)\}^2$ with $n^* = \sum_{j=1}^{l_n^*-1} l(\mathcal{B}_j^*)$. Define also the ARBB version of the pseudo-regenerative unstudentized sample mean by $\widehat{\varsigma}_n^* = n^{*1/2}\sigma_n^*(f)^{-1}(\mu_n^*(f) - \widehat{\mu}_n(f))$ and the one of the pseudo-regenerative studentized mean by $\widehat{t}_n^* = n^{*1/2}\sigma_n^*(f)^{-1}(\mu_n^*(f) - \widehat{\mu}_n(f))$. We shall use the following assumptions. Let $k \geqslant 2$ and set $\tau_S = \inf\{i \geqslant 1/\ X_i \in S\}$.

$\mathcal{H}_1(f, k)$: The small set S is such that $\sup_{x\in S} E_x((\sum_{i=1}^{\tau_S}|f(X_i)|)^k) < \infty$.

$\mathcal{H}_2(k)$: The small set S is such that $\sup_{x\in S} E_x(\tau_S^k) < \infty$.

These conditions may be classically replaced by some Liapounov's drift conditions (see Meyn & Tweedie [8]). For a sequence of nonnegative real numbers $\alpha = (\alpha_n)_{n\in\mathbf{N}}$ converging to 0 as $n \to \infty$, consider

$\mathcal{H}_3 : p(x,y)$ is uniformly estimated by $p_m(x,y)$ based on $X^{(m)}$ at the rate α_m at least for the MSE when error is measured by the L^∞ loss over $S \times S$:

$$\lim_{m\to\infty} \alpha_m^{-1} \left(E \left(\sup_{(x,y)\in S\times S} |p_m(x,y) - p(x,y)|^2 \right) \right)^{1/2} = 0.$$

$\mathcal{H}_4(k)$: The sequences $m = m(n)$ and $p = p(n)$ are chosen such that $n^{1/k} \leq p \leq m$ and $m/n \to 0$ as $n \to \infty$.

$\mathcal{H}_5$: $\overline{lim}_{t\to\infty}$ $\sup_{x\in S}$ $|E_x(\exp(it\sum_{i=1}^{\tau_S}\{f(X_i) - \mu(f)\}))| < 1$ (Cramer type condition).

$\mathcal{H}_6$: There exists $N > 0$ such that the N-fold convolution of the density of $(\sum_{i=1}^{\tau_S}\{f(X_i) - \mu(f)\})^2$ is uniformly bounded over any starting value $X_0 = x$ in S.

We then have the following results :

Theorem 3.1. *Under assumptions $\mathcal{H}_0$, $\mathcal{H}_1(f,k)$, $\mathcal{H}_2(k)$ $\mathcal{H}_3$, $\mathcal{H}_4(k)$ and $\mathcal{H}_5(k)$ with $k > 6$, we have the second order validity of the ARBB distribution both in the standardized and unstandardized case:*

$$\sup_{x\in\mathbb{R}} |P^*(\widehat{\varsigma}_n^* \leq x) - P_\nu(\widehat{\varsigma}_n \leq x)| = O_{P_\nu}(n^{-1/2}\alpha_m \vee n^{-1/2}n^{-1}m\}) \ ,$$

as $n \to \infty$. And if these conditions holds for some $k > 8$ and $\mathcal{H}_6$ hold, we have as $n \to \infty$:

$$\sup_{x\in\mathbb{R}} |P^*(\widehat{t}_n^* \leq x) - P_\nu(\widehat{t}_n \leq x)| = O_{P_\nu}(n^{-1/2}\alpha_m \vee n^{-1/2}n^{-1}m).$$

In particular if $\alpha_m = m^{-1/2}\log(m)$, by choosing $m = n^{2/3}$, the ARBB is second order correct up to $O(n^{-5/6}\log(n))$.

Proof: The proof is based on the same technical ideas as in Bertail & Clémençon [1], [2] (refer to these papers for further details). It relies on establishing the closeness between the conditional distribution of the blocks $\mathcal{B}_{1,m^*}, \ldots, \mathcal{B}_{l_{m^*,n},m^*}$ dividing the segment $X^{(m^*,n)} = (X_{m^*+1}, \ldots, X_{n+1})$ according to the $l_{m^*,n}$ visits of $(X_i, Y_i)_{m^*<i\leqslant n}$ to the atom $A_\mathcal{M}$ between time m^*+1 and time n and the conditional distribution of the blocks $\widehat{\mathcal{B}}_{1,m^*}, \ldots, \widehat{\mathcal{B}}_{\widehat{l}_{m^*,n},m^*}$ dividing $X^{(m^*,n)}$ according to the $\widehat{l}_{m^*,n}$ successive visits of $(X_i, \widehat{Y}_i)_{m^*<i\leqslant n}$ to $A_\mathcal{M}$, for given $X^{(n+1)}$. By coupling arguments one may show that, under $\mathcal{H}_2(2\gamma)$, $\gamma \geqslant 2$ and $\mathcal{H}_3$, there exists a constant C such that for $i \in \{1,2\}$,

$$E_\nu(|\widehat{\tau}_i - \tau_i|^\gamma) \leqslant C\alpha_m, \tag{1}$$

with the further notations $\tau_1 = \tau_{m^*}(1)$, $\widehat{\tau}_1 = \widehat{\tau}_{m^*}(1)$, $\tau_2 = \tau_{m^*}(l_{m^*,n})$ and $\widehat{\tau}_2 = \widehat{\tau}_{m^*}(\widehat{l}_{m^*,n})$. Now set $T_n^{(k)}(f) = n^{-1}\sum_{j=1}^{l_{m^*,n}-1} f(\mathcal{B}_{j,m^*})^k$ and $\widetilde{T}_n^{(k)}(f) =$

$n^{-1}\sum_{j=1}^{\widehat{l}_{m^*,n}-1} f(\widehat{\mathcal{B}}_{j,m^*})^k$ for $1 \leqslant k \leqslant 3$, with by convention $T_n^{(k)}(f) = 0$ (respectively, $\widetilde{T}_n^{(k)}(f) = 0$) when $l_{m^*,n} \leqslant 1$ (resp., when $\widehat{l}_{m^*,n} \leqslant 1$) and set

$$D_n^{(k)}(f) = E_\nu \left| T_n^{(k)}(f) - \widetilde{T}_n^{(k)}(f) \right| .$$

Then, following line by line the argument in Bertail & Clémençon [2], we have as $n \to \infty$

$$D_n^{(1)}(f) = O((n-m-p)^{-1}\alpha_m), \tag{2}$$

$$D_n^{(k)}(f) = O(\alpha_m),\ \text{for } k = 2,\ 3. \tag{3}$$

Observing that, conditioned on $X^{(n+1)}$, the reconstructed ARBB sample path does not keep the markovian structure but still forms a regenerative sequence, the results in Malinovskii [7] (resp. in Bertail & Clémençon [1] allow to derive an explicit Edgeworth expansion (E.E.) up to the second order for the unstudentized ARBB version (resp., for the studentized ARBB version). Given (2) and (3) it is straightforward to check that the conditions of validity of these E.E. hold and that the empirical moments appearing in the empirical E.E. of the ARBB distribution converges to their theoretical counterparts at the rate α_m at least. Moreover the bias induced by the first and last pseudo-regeneration blocks does not perturb the E.E. up to $O_P(n^{-1}\alpha_m)$. The main difficulty actually consists in establishing an E.E. for the original statistic. In the unstudentized case, since the functional is then linear, it simply amounts to control the error induced by a split at the "wrong place" for the first (resp. the last) block (*i.e.* the distance between τ_i and $\widehat{\tau}_i$, $i = 1, 2$): this is typically of the same order as the deviation (2). The unstandardized mean thus admits an E.E. on powers of $(n-m-p)^{-1/2}$, which in turn coincides with the E.E. of the empirical mean up to $O(n^{-1/2}(m/n))$. In the studentized case one must first check that the variance estimate computed from the pseudo-blocks is close to the variance estimate based on the regeneration blocks up to $O_P(n^{-1}\alpha_m)$, conditionally to the first m observations. Combining $\mathcal{H}_4(k)$ with $\mathcal{H}_2(k)$, for $k > 4$, it is straightforward that the probability that the split chain does not visit the regeneration set $S \times \{1\}$ between m and $m+p$ is typically of order $O(n^{-1})$. Subsequently to a regeneration occurring between $m+1$ and $m+p$, the remaining observations may be then decomposed into true regeneration blocks (independent from the first m observations) using the same partitioning arguments as in Malinovskii [7] or Bertail & Clémençon [1], [2]. This yields the validity of the E.E. on powers of $(n-m-p)^{-1/2} = n^{-1/2} + O(n^{-1/2}(m/n))$. A straightforward optimization argument leads to the last statement. ■

4 Tuning parameters and simulation results

The main tuning parameter relies in the choice of the small set. If the transition density $p(x, y)$ is continuous on some neighborhood $V_{x_0}(\varepsilon)^2 =$

$[x_0-\varepsilon, x_0+\varepsilon]^2$ of some fixed point $(x_0,\, x_0)$ such that $p(x_0, x_0) > 0$, then there exists $\delta = \delta(\varepsilon, p) \in]0,\, 1[$ such that $\inf_{(x,y)\in V_{x_0}^2} p(x,y) \geqslant \delta(2\varepsilon)^{-1}$. Such a compact interval $V_{x_0}(\varepsilon)$ is thus a small set for X. It satisfies condition $\mathcal{M}(1, V_{x_0}(\varepsilon), \delta, \mathcal{U}_{V_{x_0}(\varepsilon)})$, where $\mathcal{U}_{V_{x_0}(\varepsilon)}$ denotes the uniform distribution on $V_{x_0}(\varepsilon)$. Hence, in the case when one knows x_0, ε and δ such that (2) holds (this simply amounts to know a uniform lower bound estimate for the probability of returning to $V_{x_0}(\varepsilon)$ in one step), one may effectively apply the ARBB methodology to X. A possible selection rule for ε relies on fixing x_0 and searching for $\varepsilon > 0$ so as to maximize the expected number of regeneration-blocks conditionally to the observed trajectory $X^{(n+1)}$, that is

$$\begin{aligned} N_n(\varepsilon,p) &= E(\textstyle\sum_{i=m^*+1}^{n} I\{X_i \in V_{x_0}(\varepsilon),\, Y_i = 1\}\,|X^{(n+1)}) \\ &= \tfrac{\delta(\varepsilon,p)}{2\varepsilon}\textstyle\sum_{i=m^*+1}^{n} I\{(X_i, X_{i+1}) \in V_{x_0}(\varepsilon)^2\}\tfrac{1}{p(X_i,X_{i+1})}. \end{aligned}$$

Since the transition density p and its minimum over $V_{x_0}(\varepsilon)^2$ are unknown, a practical criterion $\widehat{N}_n(\varepsilon)$ to optimize is obtained by replacing p by p_m and $\delta(\varepsilon,p)/2\varepsilon$ by a sharp lower bound $\widehat{\delta}_m(\varepsilon, p_m)/2\varepsilon$ for p_m over $V_{x_0}(\varepsilon)^2$. The final procedure may be then implemented in 4 steps as follows. Let x_0 be fixed.

1. Compute an estimator $\widehat{p}_m$ of the transition density, for instance of Nadaraya-Watson's type, with $m = Cn^{2/3}, C > 0$.
2. Select the small set $V_{x_0}(\varepsilon)$ by maximizing the empirical criterion $\widehat{N}_n(\varepsilon)$ described above over $\varepsilon > 0$. This yields $\widehat{\varepsilon}_{m,opt}$ and a corresponding minimum value $\widehat{\delta}_{m,opt}$.
3. At each time $i > m^*$ when $(X_i, X_{i+1}) \in [-\widehat{\varepsilon}_{m,opt}, \widehat{\varepsilon}_{m,opt}]^2$, draw independent Bernoulli r.v.'s $\widehat{Y}_i$ with parameter $1-\widehat{\delta}_{m,opt}(2\varepsilon_{m,opt})^{-1}/\widehat{p}_m(X_i, X_{i+1})$. At each time i such that $\widehat{Y}_i = 1$, divide the trajectory, getting data blocks of random size.
4. Apply the ARBB procedure to the sample mean as previously described.

Because the tuning parameters p_m, $\widehat{\varepsilon}_{m,opt}$, $\widehat{\delta}_{m,opt}$ explicitly depends on the first m observations only, the "2-split" technique ensures that the ARBB resampling will not be asymptotically perturbed by the latter.

In the following tables, we compare the quantile of order γ of the true distribution (TD) of the mean respectively. We take $X_0 = 0$, ε_i *i.i.d.* $\sim N(0,1)$ and consider

-an AR(1) model : $X_i = \rho X_{i-1} + \varepsilon_i$, with $\rho = 0.95$ and $n = 200$, $m = 68 = [2 * n^{2/3})]$.

-an AR model with a ARCH(1) structure $X_i = \rho X_{i-1} + (1+\alpha X_{i-1}^2)^{1/2}\varepsilon_{i,}$ $\rho = 0.6$, $\alpha = 0.1$. See Bertail and Clémençon [2] for comparison with the ARBB without the double splitting trick. The performance are quite similar and suggest that the ARBB without the splitting trick enjoy the same second order properties.

	AR		AR-ARCH		
γ	TD	ARBB	TD	ARBB	ASY
1	-3.63	-3.72	-2.53	-2.65	-2.32
2.5	-2.77	-2.81	-2.02	-2.09	-1.96
5	-2.34	-2.36	-1.79	-1.84	-1.65
10	-1.74	-1.73	-1.42	-1.44	-1.28

	AR		AR-ARCH		
γ	TD	RBB	TD	ARBB	ASY
90	1.68	1.61	1.36	1.41	1.28
95	2.16	1.99	1.73	1.82	1.65
97.5	2.73	2.46	2.00	2.14	1.96
99	3.62	3.60	2.53	2.69	2.32

Table 1: Comparison of the tails of the true (TD), modified ARBB and gaussian (ASY) distributions for the two models.

References

[1] Bertail P., Clémençon S. (2003). *Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains.* To appear Probability Theory and Related Fields.

[2] Bertail P., Clémençon S. (2003). *Regenerative block bootstrap for Markov chains.* Submitted to Ann. Statist.

[3] Bolthausen E. (1982). *The Berry-Esseen Theorem for strongly mixing Harris recurrent Markov Chains.* Z. Wahrsch. Verw. Gebiete, **60**, 283 – 289.

[4] Datta S., McCormick W.P. (1993). *Regeneration-based bootstrap for Markov chains.* Canadian J. Statist., **21**, No.2, 181 – 193.

[5] Götze F., Künsch H.R. (1996). *Second order correctness of the blockwise bootstrap for stationary observations.* Ann. Statist., **24**, 1914 – 1933.

[6] Hall P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer.

[7] Malinovskii V. K. (1987). *Limit theorems for Harris Markov chains I.* Theory Prob. Appl., **31**, 269 – 285.

[8] Meyn S.P., Tweedie R.L., (1996). *Markov chains and stochastic stability.* Springer.

[9] Nummelin E. (1978). *A splitting technique for Harris recurrent chains.* Z. Wahrsch. Verw. Gebiete, **43**, 309 – 318.

[10] Schick A (2001). *Sample splitting with Markov Chains.* Bernoulli, **7**, No. 1, 33 – 61.

Address: P. Bertail, CREST, Laboratoire de Statistique, Timbre J340, 1, Bd A. Pinard, 75675, Paris
S. Clémençon, MODAL'X, Université Paris X

E-mail: `Patrice.Bertail@ensae.fr, sclemenc@u-paris10.fr`

TWO MEASURES OF CREDIBILITY OF EVOLUTIONARY TREES

Martin Betinec

Key words: Evolutionary tree, clustering, classification, bootstrap, DNA, Trichomonas.

COMPSTAT 2004 section: Biostatistics, Clustering.

Abstract: Evolutionary tree is a dendrogram that is used to asses genetical similarity of biological objects. The contribution concerns two methods that use bootstrap to verify the credibility of the built tree. The measures are compared on gene sequences of Trichomonadinæ family.

1 Introduction

Evolutionary trees, also known as *Phylogenetic trees*, constitute in biology a way of describing evolution of the observed organisms. From the mathematical point of view, these trees are a special type of classification trees (dendrograms). Not only this method enables a demonstration of relationships among objects, i.e. to answer the question 'which assay is close to this one in evolution', but it can also quantify the distance between clusters of objects. The distance is proportional to the length of the branch connecting the corresponding clusters.

This paper concerns two methods (sections 4 and 5) that use bootstrap for estimation of the confidence mentioned above. We will try to demonstrate why it is possible to claim something reliable about connection between our data set and the real world from the information bootstrap populations can produce. We compare these two methods in section 6 using a real data set that is introduced in section 2.

2 Data

For the demonstration of the results we will use the data that were assembled by a group of biologists of the Faculty of Science of Charles University in Prague lead by prof. Flégr, see `http://prfdec.natur.cuni.cz/~flegr/`. They investigated a part of the genome of a unicellular organisms of the *Trichomonadinæ* family. The data are to be published, but they have not been up to now, so the names of the isolates are replaced with ordinals. For our illustration we will use 22 of them. For $i = 1, \ldots, 22$ we denote the isolates:

$$\boldsymbol{x}_i = \underbrace{(\text{A,T,G, } \ldots, \text{ C,T,A,T})}_{1870 \text{ nucleotides}},$$

where the capital letters are the first letters of the nucleotides' names:
A – adenine, T – thymine, C – cytosine, G – guanine.

3 Tree growing

We compose a data matrix $\boldsymbol{X}$ from the row vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{22}$. So, its dimensions are 22×1870.

There are three important questions for determination of distances between the vectors of $\boldsymbol{X}$. Choice of a *method* of agglomeration, a *distance* and *encoding* the nucleotides which can influence shaping the estimated tree.

3.1 Encoding of nucleotides

For the latter, we need to consider their chemical affinity, probability of their point mutation etc. For more detailed discussion on this topic see [1]. After testing various reasonable encodings (see [2]) a relative independence of the shape of the tree on this factor was realized. Finally, we will use the following in accordance with ([5]):

$$\big(\text{A,G,C,T}\big) = \big(1, 2, 5, 6\big)\,. \tag{1}$$

3.2 Distance

All the metrics that were used (with logical exception of the *supremal*) caused very similar trees. Finally, we will work with the *Euclidean* metric, which produces exactly the same tree as the *Manhattan*, and very similar to *Canberra* metric, see [2].

Obviously, it is possible to leave out the constant column vectors of the matrix $\boldsymbol{X}$. They don't bring any additional information about distance of rows because the corresponding terms are zeros. By that we reduce the dimension of the matrix $\boldsymbol{X}$ from (22×1870) to (22×566).

Various methods of cluster analysis were examined, i.e. *single* and *complete linkage*, *group average*, *Ward's*, *centroid* and *median*. It was realized that the resulting trees were very similar (except *Ward's* method). Finally, we will use *the nearest neighbour principle* to construct dendrograms in this article (see Figure 1). Not only because of the fact that the group of the similar methods was bigger; we also made sure of the proper method by principal component analysis, for details see [2].

4 The method of relative frequencies

The simplest estimate of confidence level of a single branch v of the tree $\widehat{\Psi}$ can be obtained as a relative frequency of occurrence of v at the bootstrapped trees. According to its author, the method is called Felsenstein's.

In each iteration $b = 1, \cdots, B$, we randomly select some of the columns of the matrix $\boldsymbol{X}$ with replacement and from these chosen vectors we create

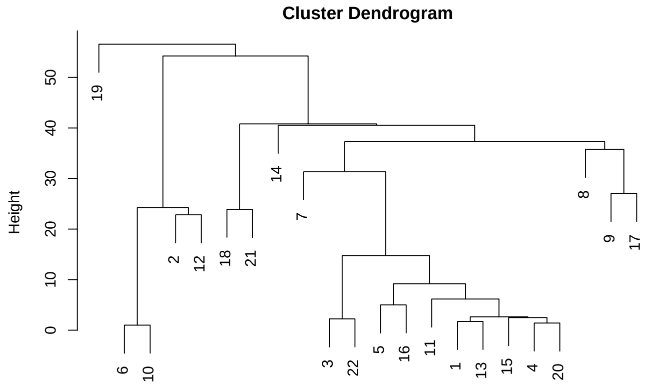


Figure 1: The single linkage tree.

a bootstrap matrix $\boldsymbol{X}_b^*$ of the same size as $\boldsymbol{X}$, i.e. every column vector of the matrix $\boldsymbol{X}$ can be drawn with the same probability. Then we grow a tree $\widehat{\Psi}_b^*$ from the matrix $\boldsymbol{X}_b^*$.

We define the **confidence level** of the branch v as:

$$\alpha_F(v) \overset{df}{=} \frac{1}{B}\#\big\{b : \widehat{\Psi}_b^* \ni v\ \big\}, \tag{2}$$

where # means "a number of elements of" and B is a number of bootstrap iterations. It is not far to seek adaptation of the algorithm for more branches.

The following section will focus in some detail on the relation of the calculated estimate of the confidence level and reality, i.e. in explaining how it is possible to estimate $\alpha = \mathbf{P}\widehat{\Psi} \equiv \Psi$.

4.1 Theoretical background

We suppose that the columns $\boldsymbol{x}^1, \dots, \boldsymbol{x}^{566}$ of the matrix $\boldsymbol{X} = \big(\boldsymbol{x}^1, \dots, \boldsymbol{x}^{566}\big)$, where $x_l^k \in \big\{\text{A,G,T,C}\big\}$, $k = 1, \dots, 566$, $l = 1, \dots, 22$ are randomly selected (with range $n = 566$) from some probability distribution on the space of **all possible non-constant** vectors $\boldsymbol{\xi}$ of dimension 22 over the alphabet $\big\{\text{A,G,T,C}\big\}$. We denote this space as $\mathcal{X}$, i.e. $\mathcal{X} = \big\{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_K\big\}$, where $K = 4^{22} - 4 = 1.7592 \cdot 10^{13}$.

By the diagram

$$\boldsymbol{X} \Longrightarrow \text{cluster an.} \Longrightarrow \widehat{\Psi} \qquad \boldsymbol{X}^* \Longrightarrow \text{cluster an.} \Longrightarrow \widehat{\Psi}^* \quad \text{resp.}, \tag{3}$$

we can describe both areas of reality - the *practical* and the *virtual*, but it cannot be used for the description of the *theoretical* area we are interested in. That is why we move into space $\mathcal{P}$, which is space of **probability vectors**. Their components π_i are probabilities of including the vectors $\boldsymbol{\xi}_i \in \mathcal{X}$ into the sample $(i = 1, \dots, K)$.

$$\mathcal{P} \overset{df}{=} \Big\{\boldsymbol{\pi} = \big(\pi_1, \dots, \pi_K\big)' \quad \pi_i = \mathbf{P}\boldsymbol{\xi}_i \text{ involved into the sample}\Big\} \tag{4}$$

Theoretical probabilities $\boldsymbol{\pi}$ correspond to the **observed frequencies** of appearance of the vectors $\boldsymbol{\xi}_i$ in the columns of the data matrix $\boldsymbol{X}$, i.e.

$$\boldsymbol{\Upsilon} = \left(\Upsilon_1, \ldots, \Upsilon_K\right)', \qquad \text{where} \quad \Upsilon_i \overset{df}{=} \#\left\{j :\ \boldsymbol{x}^j = \boldsymbol{\xi}_i\right\} \quad i = 1, \ldots, K\,,$$

or the **relative frequencies**

$$\widehat{\boldsymbol{\pi}} = \left(\widehat{\pi}_1, \ldots, \widehat{\pi}_K\right)', \qquad \text{where} \quad \widehat{\pi}_i \overset{df}{=} \frac{1}{n} \#\left\{j :\ \boldsymbol{x}^j = \boldsymbol{\xi}_i\right\} = \frac{\Upsilon_i}{n}\,. \tag{5}$$

From the bootstrapped matrices $\boldsymbol{X}_b^* = \left(\boldsymbol{x}^n * 1b, \ldots, \boldsymbol{x}^n * 566b\right)$, $b = 1, \ldots, B$, we could obtain for $i = 1, \ldots, K$ by analogy $\boldsymbol{\Upsilon}_b^* = \left(\Upsilon_{b,1}^*, \ldots, \Upsilon_{b,K}^*\right)'$ and $\widehat{\boldsymbol{\pi}}_b^* = \left(\widehat{\pi}_{b,1}^*, \ldots, \widehat{\pi}_{b,K}^*\right)'$.

Obviously (for $K = 4^{22} - 4$ and $n = 566$):

$$\boldsymbol{\Upsilon} \sim \text{Multi}_K(n, \boldsymbol{\pi}) \qquad\qquad \text{a} \qquad\qquad \boldsymbol{\Upsilon}^* \sim \text{Multi}_K(n, \widehat{\boldsymbol{\pi}})\,. \tag{6}$$

Now, we can express the tree growing, described by diagram (3) as:

$$\widehat{\boldsymbol{\pi}} \Longrightarrow \widehat{\Psi}\,, \qquad \text{resp.} \qquad \widehat{\boldsymbol{\pi}}_b^* \Longrightarrow \widehat{\Psi}_b^*.$$

Similarly, we can imagine that also the theoretical probabilities $\boldsymbol{\pi}$ correspond to some theoretical tree Ψ that describes the true relations among the observed organisms.

It is also worth noticing that one and the same tree can be grown from two different probabilistic vectors $\widehat{\boldsymbol{\pi}}^{(1)}$ and $\widehat{\boldsymbol{\pi}}^{(2)}$. Therefore, the space $\mathcal{P}$ is split into the disjoint classes of equivalency $\mathcal{P} = \dot{\bigcup}_i \mathcal{P}_i$. For these classes holds:

$$\left\{\mathcal{P}_i \ni \widehat{\boldsymbol{\pi}}^{(1)} \quad \& \quad \mathcal{P}_j \ni \widehat{\boldsymbol{\pi}}^{(2)}\right\} \qquad \Longrightarrow \qquad \left\{\,\widehat{\Psi}^{(1)} \equiv \widehat{\Psi}^{(2)} \quad \Leftrightarrow \quad i = j\,\right\}, \tag{7}$$

where $\widehat{\Psi}^{(j)}$ corresponds to the tree grown from $\widehat{\boldsymbol{\pi}}^{(j)}$, $j = 1, 2$.

Let us return to the original problem of estimating confidence of our tree Ψ. What we want to know is the probability of the accordance of $\widehat{\Psi}$ with its theoretical pattern Ψ. So, we are interested in:

$$\alpha = \mathbf{P}\widehat{\Psi} \equiv \Psi = \mathbf{P}\boldsymbol{\pi} \in \mathcal{P}_i\,|\widehat{\boldsymbol{\pi}} \in \mathcal{P}_i \tag{8}$$

We estimated this probability by (2) as an empirical probability (relative frequency) of the fact that the branch in question of the bootstrapped tree $\widehat{\Psi}^*$ grown from $\boldsymbol{X}^*$ corresponds to the given branch of the original tree $\widehat{\Psi}$:

$$\alpha_F = \widehat{\mathbf{P}}_{\widehat{\boldsymbol{\pi}}^*|\widehat{\boldsymbol{\pi}}}(\widehat{\boldsymbol{\pi}}^* \in \mathcal{P}_i\,|\widehat{\boldsymbol{\pi}} \in \mathcal{P}_i)\,. \tag{9}$$

To have an idea of the error we caused by this estimation, it is necessary to know the conditional distribution $\widehat{\boldsymbol{\pi}}^*$ (the aposteriori distribution of $\boldsymbol{\pi}$ respectively) with $\widehat{\boldsymbol{\pi}}$ fixed, so $\mathcal{L}\left(\widehat{\boldsymbol{\pi}}^*\,|\widehat{\boldsymbol{\pi}}\right)$ (resp. $\mathcal{L}\left(\boldsymbol{\pi}\,|\widehat{\boldsymbol{\pi}}\right)$).

We know the first one, because it follows from the formula (6):

$$\mathcal{L}\big(n\widehat{\boldsymbol{\pi}}^* \,|\widehat{\boldsymbol{\pi}}\big) = \mathrm{Multi}_K(n, \widehat{\boldsymbol{\pi}})\,. \tag{10}$$

The aposteriori distribution $\mathcal{L}\big(\boldsymbol{\pi}\,|\widehat{\boldsymbol{\pi}}\big)$ can be expressed by the sample distribution $\widehat{\boldsymbol{\pi}}$ conditional by $\boldsymbol{\pi}$. Considering the model, it holds:

$$\mathcal{L}\big(n\widehat{\boldsymbol{\pi}}\,|\boldsymbol{\pi}\big) = \mathrm{Multi}_K(n, \boldsymbol{\pi})\,. \tag{11}$$

Multinomial distribution is conjugated with Dirichlet's, therefore:

Lemma 4.1. *$\mathcal{L}\big(\widehat{\boldsymbol{\pi}}^*\,|\widehat{\boldsymbol{\pi}}\big)$ and $\mathcal{L}\big(\boldsymbol{\pi}\,|\widehat{\boldsymbol{\pi}}\big)$ have the following properties:*

1. *Let the apriori distribution of $\boldsymbol{\pi}$ be uninformative, then*

$$\mathcal{L}\big(\boldsymbol{\pi}\,|\widehat{\boldsymbol{\pi}}\big) = Dir_K(n\widehat{\boldsymbol{\pi}})\,. \tag{12}$$

2. *They both are concentrated only on those $\boldsymbol{\xi}_i$ for which $\pi_i > 0$, i.e. on columns of the matrix $\boldsymbol{X}$.*
3. *They have identical mean.*
4. *Their covariance matrices equal asymptotically, (i.e. $\Sigma^* = \frac{n+1}{n}\Sigma$)*

The main advantage of this estimate of the confidence level of the evolutionary tree (defined by (2)) is its simplicity. Moreover, it is easy to implement its algorithm. Last but not least benefit lies in rapidity of its calculation. However, the resulting estimate could be sometimes quite misguiding.
A problem appears if we cannot guarantee fulfilling of the assumptions of the lemma 4.1 – its consequences cannot be guaranteed either.

The estimated confidence level α_F could be very distant from the real α. Considering the complexity of the space $\mathcal{P}$, it is not easy to recognize whether that situation comes on. Let us try a more delicate approach.

5 Improved estimator

5.1 Theory

We suppose the model (4) like in section 4. Further, we shall remind ourselves that the space $\mathcal{P}$ is split by the rule (7), so $\mathcal{P} = \dot{\bigcup}_i \mathcal{P}_i$.

Let us reserve a subscript $i = 1$ for that part of the space $\mathcal{P}$ which contains the vector $\widehat{\boldsymbol{\pi}}$, hence $\mathcal{P}_1 \ni \widehat{\boldsymbol{\pi}}$. Let $\overline{\mathcal{P}}_1 = \dot{\bigcup}_{i\neq 1} \mathcal{P}_i$ be a complement of $\mathcal{P}_1$. Not only will we consider the fact that $\widehat{\boldsymbol{\pi}} \in \mathcal{P}_1$ like in previous section, but we will also take into account the distance of $\widehat{\boldsymbol{\pi}}$ from the boundary of the area $\mathcal{P}_1$. Motivation for this effort is following: the vector $\boldsymbol{\pi}$ should lie "near around" the vector $\widehat{\boldsymbol{\pi}}$, because $\widehat{\boldsymbol{\pi}}$ (a vector of the observed relative frequencies) is generated from $\boldsymbol{\pi}$ by assumed model (5). Hence, the confidence level

$$\alpha = \mathbf{P}\boldsymbol{\pi} \in \mathcal{P}_1\,|\widehat{\boldsymbol{\pi}} \in \mathcal{P}_1\,, \tag{13}$$

should be higher, if $\widehat{\boldsymbol{\pi}}$ lied “somewhere in the middle” of $\mathcal{P}_1$ than in the case of $\widehat{\boldsymbol{\pi}}$ situated near the boundary of $\mathcal{P}_1$.

Let $\partial\mathcal{P}_1$ assign a border of $\mathcal{P}_1$. Let $\boldsymbol{\pi}_0$ be the closest point of $\partial\mathcal{P}_1$ to $\widehat{\boldsymbol{\pi}}$, hence

$$\boldsymbol{\pi}_0 \overset{df}{=} \arg \min_{\breve{\boldsymbol{\pi}} \in \partial\mathcal{P}_1} d(\widehat{\boldsymbol{\pi}}, \breve{\boldsymbol{\pi}}) \,, \tag{14}$$

where $d(.\,,.)$ is an arbitrary *metric* on $\mathcal{P}$.

In the previous section we got a bootstrap population by means of frequencies of the occurrence that had been generated from $\widehat{\boldsymbol{\pi}}$, see (6). Now, we will generate new frequencies from $\boldsymbol{\pi}_0$, so for $b = 1, \dots, B_2$

$$\boldsymbol{\Upsilon}_b^{**} = \left(\Upsilon_{1b}^{**}, \dots, \Upsilon_{Kb}^{**}\right)' , \qquad \text{where} \qquad \Upsilon_{ib}^{**} \overset{df}{=} \#\left\{j : \boldsymbol{x}^n * * jb = \boldsymbol{\xi}_i\right\} , \tag{15}$$

$$\boldsymbol{\Upsilon}_b^{**} \sim \text{Multi}_K(n, \boldsymbol{\pi}_0) \qquad \text{and} \qquad \widehat{\boldsymbol{\pi}}_b^{**} = \frac{1}{n} \boldsymbol{\Upsilon}^{**} \,. \tag{16}$$

Our intention is to estimate the confidence value (13). We can proceed in our reflection as follows. The elements of the bootstrap population $\widehat{\boldsymbol{\pi}}_b^{**}$ are generated from the boundary point $\boldsymbol{\pi}_0$. The closer these elements $\widehat{\boldsymbol{\pi}}_b^{**}$ are to $\overline{\mathcal{P}}_1$ (compared to the position of $\widehat{\boldsymbol{\pi}}$, i.e. the smaller the estimated variance of $\widehat{\boldsymbol{\pi}}_b^{**}$ is), the smaller is the probability of the placement of the vector of the true probabilities $\boldsymbol{\pi}$ in $\overline{\mathcal{P}}_1$. The reason is that we suppose that the vector of the observed relative frequencies $\widehat{\boldsymbol{\pi}}$ was obtained by $\boldsymbol{\pi}$ by the same type of generation as in the case of generation of $\widehat{\boldsymbol{\pi}}_b^{**}$ from $\boldsymbol{\pi}_0$, cf. (6), (16). Summary: The closer $\widehat{\boldsymbol{\pi}}_b^{**}$ are to $\overline{\mathcal{P}}_1$ compared to the position of $\widehat{\boldsymbol{\pi}}$, the higher α is.

We define a new *confidence level:*

$$\alpha_{_E} \overset{df}{=} \widehat{\mathbf{P}}_{\widehat{\boldsymbol{\pi}}***|\boldsymbol{\pi}_0}(d(\widehat{\boldsymbol{\pi}}^{**}\!, \overline{\mathcal{P}}_1) \le d(\widehat{\boldsymbol{\pi}}, \overline{\mathcal{P}}_1) \,|\, \widehat{\boldsymbol{\pi}} \in \mathcal{P}_1) \,. \tag{17}$$

There is a consequence of this definition: $1 - \alpha_{_E}$ is an analogy of p value in the classical testing of hypotheses which can be illustrated by the following example.

Let $Z_1, \dots, Z_n \overset{iid}{\sim} N(\mu, 1)$. The investigated hypotheses are:
H: $\mu \le \mu_0$ vs. K: $\mu > \mu_0$.

Vector $\boldsymbol{\pi}_0 \in \partial\mathcal{P}_1$ corresponds to $\mu = \mu_0$ that is the point at the border between H and K. We estimate μ by $\widehat{\mu} = \widehat{\mu}(Z_1, \dots, Z_n)$ after finishing the experiment ($\widehat{\mu}$ is the analogy of $\widehat{\pi}$). Knowing the value of $\widehat{\mu}$, we know $\mathcal{P}_1 \ni \widehat{\mu}$, too. We want to confirm that $\mu \in \mathcal{P}_1$, i.e. we want to reject the negation, hence $H \equiv \overline{\mathcal{P}}_1$.

We denote $Z_1^{**}, \dots, Z_n^{**} \overset{iid}{\sim} N(\mu_0, 1)$ a sample bootstrapped from the null hypothesis H and an appropriate estimate $\widehat{\mu}^{**} = \widehat{\mu}^{**}(Z_1^{**}, \dots, Z_n^{**})$ (which is an analogy of $\widehat{\boldsymbol{\pi}}^{**}$), then

$$p = \mathbf{P}_{\mu_0}(\widehat{\mu}^{**} > \widehat{\mu}) = \mathbf{P}_{\mu_0}(\widehat{\mu}^{**} - \mu_0 > \widehat{\mu} - \mu_0) = \mathbf{P}_{\mu_0}(\mathrm{d}\left(\widehat{\boldsymbol{\mu}}^{**}, \overline{\mathcal{P}}_1\right) > \widehat{\mu} - \mu_0) \,,$$

where $\mathrm{d}\left(\widehat{\boldsymbol{\mu}}^{**}, \overline{\mathcal{P}}_1\right) = \widehat{\mu}^{**} - \mu_0$ for $\widehat{\mu}^{**} > \mu_0$ or 0 otherwise.

The $\alpha_{_E}$ level defined by (17) takes into account the shape of the $\mathcal{P}_1$ area more than the $\alpha_{_F}$ level does, e.g. while $\alpha_{_E} = \alpha_{_F}$ in the case of straight boundary, in the case of strictly convex $\overline{\mathcal{P}}_1$ area it holds $\alpha_{_E} \leq \alpha_{_F}$, as is shown in [5] and also in [1].

There still remains a question: how to compute $\alpha_{_E}$.

5.2 Calculation

The authors of [5] proposed a solution based on bias-corrected percentile confidence intervals, see [3], [4]. Detailed proof can be found in [1].

The technique of calculation is quite complicated, for detailed description see [5] or [1]. It contains two bootstrap resamplings. First one (of size B_1) is the same procedure as estimation of α_F. It serves for the identification of the bootstrap trees $\widehat{\Psi}_b^*$ that do not correspond to $\widehat{\Psi}$. Using those $\widehat{\Psi}_b^*$ the π_0 is estimated by $\widehat{\pi}_0^b$ with precision 2^{-L}. In this case we use $L = 10$ and 14, see tab. 1. The next step is bootstrapping (of size B_2) from $\widehat{\pi}_0^b$ described at (16). Then α_E is estimated:

$$\alpha_E \doteq \Phi\left(\frac{z^{(\alpha_F)} - z_0}{1 + a(z^{(\alpha_F)} - z_0)} - z_0\right), \tag{18}$$

where Φ is the distribution function of N(0,1), $z^{(\alpha_F)} = \Phi^{-1}(\alpha_{_F})$, $z_0 = \Phi^{-1}\Big(\mathbf{P}_{\boldsymbol{\pi}_0}(\widehat{\pi}^{**} \in \mathcal{P}_1)\Big)$ and 'acceleration constant' a is computed during second bootstrap.

6 Results

The branches are numbered increasingly according to the order they are joined by in the Table 1. It can be seen there that α_F is much more stable in changing parametres then α_E. The unstability of α_E could be caused by the small values of B_2 as claims [5]. In [1], there were estimated α_E for dendrogram of 13 objects (of only 73 different components each) using B_1 up to 10000, B_2 up to 200 and the results were stable. With 22 objects (of 566 components each), the computation is very time consuming[1], see Table 1[2].

Besides the description of evolution, there could be another task – to find reasonable clusters of related objects. Figure 1 proposes three clusters: {`19`}, {`6,..., 12`} and {`18, ..., 17`}. If there are more clusters, we continue to divide the largest one. There is a sequence of joining levels which are very close to each other, so the division finally stops on 7 or 8 clusters (the same results come from other hierarchical clustering methods from PCA, too – see [2]). The levels α_F and α_E can help to determine the proper number

[1] calculated on `AMD Duron, 750 MHz, 64 kB Cache, 512 Mb RAM`

[2] the format of time is `h: mm: ss`; the NA value is in fact 6:28:01, but this calculation was not the only process running at that time.

		α_F		α_E					
	B_1	100	1000	100				1000	
			B_2	30		100		30	
Nr.	level		L	10	14	10	14	10	14
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	1.414	0.600	0.622	0.752	0.761	0.739	0.811	0.742	0.735
3	1.732	0.930	0.901	0.803	0.939	0.773	0.860	0.800	0.819
4	2.236	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	2.500	0.550	0.516	0.630	0.661	0.653	0.562	0.657	0.691
6	2.646	0.890	0.896	0.782	0.858	0.712	0.792	0.798	0.825
7	5.000	0.910	0.919	0.908	0.856	0.765	0.735	0.801	0.853
8	6.164	0.850	0.844	0.787	0.883	0.744	0.810	0.759	0.740
9	9.179	0.960	0.971	0.952	0.938	0.779	0.868	0.953	0.883
10	14.765	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11	22.820	0.690	0.690	0.612	0.688	0.675	0.839	0.710	0.699
12	23.916	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
13	24.218	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
14	27.032	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15	31.333	0.980	0.995	1.000	0.841	0.997	0.735	0.976	0.982
16	35.760	0.600	0.638	0.647	0.643	0.600	0.781	0.700	0.697
17	37.289	0.930	0.957	0.938	0.930	0.871	0.651	0.775	0.855
18	40.534	0.550	0.546	0.799	0.695	0.668	0.738	0.670	0.661
19	40.810	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
20	54.242	0.920	0.900	0.893	0.880	0.745	0.772	0.815	0.746
calcul.time		0:00:15	0:02:07	2:05:31	2:19:11	NA	4:42:44	21:47:49	24:15:34

Table 1: Comparison of the levels α_F a α_E.

of clusters. Table 1 shows that joining of the cluster {`3, ..., 20`} with `7` is quite trustworthy, so it recommends rather to stop at 7 clusters, if not at 3 before (proper lines in Table 1 are underlined).

References

[1] `http://betinec.matfyz.cz/doc/Robust02/Robust02.html`

[2] `http://betinec.matfyz.cz/doc/Compstat04/Compstat04.html`

[3] Efron B. (1982). *The Jacknife, the bootstrap, and other resampling plans.* CBMS **38**, SIAM-NSF.

[4] Efron B. (1987). *Better bootstrap confidence intervals.* JASA **82**, 171-185.

[5] Efron B., Halloran E., Holmes S. (1996). *Bootstrap confidence levels for phylogenetic trees.* Proc. Natl. Acad. Sci. USA **93**, 13429 – 13434

Address: M. Betinec, Department of Sociology, Charles University in Prague, Celetná 20, 116 42 Prague, Czech Republic

E-mail: `betinec@matfyz.cz`

A STATISTICAL DATABASE FOR THE TRADE SECTOR: A PROTOTYPE OF A NEW DATA COLLECTION TOOL

Silvia Biffignandi and Stefano Pisani

Key words: Satellite account for trade, fiscal data, interactive database.

COMPSTAT 2004 section: Official statistics.

Abstract: This paper focuses on the problems arising in the construction of an integrated database which uses administrative financial registers, on the advantages of the database obtained and on its characteristics. To be precise, in this study information coming from two different types of taxes is compared. The first tax is VAT (value added tax), which is applied in a standardized way at European level. The second tax is the Regional Tax on Production (in Italian IRAP), which is a tax applied in Italy. Each source of data is set up with reference to the reasons why tax is applied, therefore each gives a partial description of the internal trade. In order to use the above-mentioned sources for economic analysis, statistical work is required.

1 Introduction

New strategies for the data formation process enable us to obtain large amounts of information, regarding both the kind of phenomena being studied and the sectorial and territorial level of detail.

In addition, the use of administrative data conveniently combined with survey data permits us to obtain a wider coverage of the universe. Innovations in progress and the harmonization of the definitions belonging to the statistical system tend to create the basis for the increasing transparency of the data production process and improved comparability of data originating from different sources.

The data base, presented in this paper (called ECOFISCOM[1]) is aimed at the economic analysis of internal trade, carried out by satellite accounting tool [1], [2], therefore, the administrative archives which retain the information relative to the taxable base and not the tax itself. This taxable base, correctly evaluated, allows us to calculate approximately the typical size of macroeconomies which, given the enormity of the data base used, can be employed for the micro analysis of internal trade.

This paper is organized as follows. Paragraph 2 presents the objectives of the proposed data base and the conceptual framework contained within the information as the need arises, paragraph 3 describes the characteristics of the data base regarding both content in terms of data, indicators, data

[1] This acronym show that the data base aims at the economic analysis of the internal trade using fiscal data (ECOnomic, FISCal, COMmerce database).

processing facilities and information technology aspects. Some concluding remarks (par. 4) summarize the innovative results of the paper and future developments.

2 The conceptual framework

Archive integration needs the creation of a series of information which is coherent and harmonized from various points of view. First of all it is necessary to create a connection between the definitions and the classification of aggregates from different sources of integration.

Regarding the IVA and IRAP archives, this has been done by referring to the definitions in national accounting. These definitions are standardized at an internazional level [6]. and, therefore, the Italian administrative data can be compared with that of other nations. Table 1 outlines the first table comparing the fiscal dimensions and between these and the national accounting aggregates.

The table shows that the two revenues being examined supply a representative picture which is almost complete for the main macro-economical quantities which make up the total of the sources and end uses.

IRAP	**IVA**	**National accounting**
Total Positive Components	Turnover	Production
Total Negative Components	Total purchases and import	Interm. Consumption Import
Value of net production (VPN)	Value Added	Value Added
Labour costs	Labour costs	Labour costs
Redemptions	Redemptions	Redemptions
	Export	Export
	Redeemable goods (purchases and transfer)	Investments

Table 1: Table of connections between financial account definitions and national accounting.

While analysing fiscal data, it is important to bear in mind the limits, it is therefore essential to underline that measurements based on such data are affected by the phenomenon of turnover under-declaration (or costs over-declaration), due to fraud or tax evasion. However, it is possible to achieve useful information that relies on relative assumptions rather than absolute considerations. Observing the fiscal reality constitutes a valid starting point for the formulation of hypotheses on compulsory integrations in order to introduce the so-called “underground economy” issue.

3 The ECOFISCOM database

3.1 Content

One of the characteristic and qualifying aspects of the archive has been the effort made to build up metadata, which allows data stored in the databases to be produced in a conceptual framework which is coherent not only between administrative databases but also with national accountability. This archive of metadata, is mainly made up of an essential reference for the transparency of data and the correct use of data for statistical analysis.

For each of the IVA and IRAP aggregates shown in table 1, in actual fact, both the definitions in fiscal terms as well as the existing relationship with the national accounting system are reported in the metadata database.

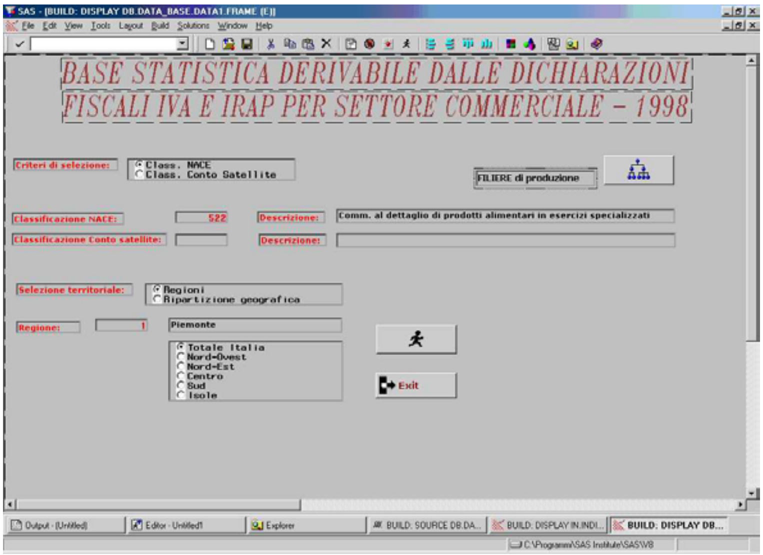
When the archive is consulted the results are presented in a standard table which is compared alongside each aggregate IRAP with the corresponding IVA.

For the aggregates which do not correspond to either of the two sides of the table are empty. The analysis deriving from the information which is obtained from two diverse sources, provided the metadata available, enables each user to carry out quality analysis of the data, aimed at understanding if the numerical differences reflect a differing economic meaning from the aggregate or that there is an error in the way that information is collected.

Going over to the analysis of the structure of data files, the illustration of the initial template (figure 1) from the ECOFISCOM data base shows the key questions proffered by the user. These are: economic activity, territory, value chain production. The interrogative key for a sector of economic activities, is articulated in two sectors: in the first a standard European classification is proposed (NACE rev.1), the second supplies a more aggregated version aimed at economic analysis, which is typical of the satellite accountability of internal trade.

This level of aggregation, defined in a Eurostat [4], offers a synthetic representation which allows a complete analysis for both the distribution channels as well as by market sector. Regarding interrogation in the NACE sector, it is interesting to underline the dynamic character of ECOFISCOM, since it is possible to ask the database with a digit ranging from 5 to 1. The data base, in actual fact, is not based on pre-coded tables, but is interfaced with a program which, each time, makes up the result of the information required. Regarding the key to territorial questions, apart from being generated for the whole of Italy, these can be layered by geographical breakdowns (5) and regions (20).

Examining the territorial differences, the potential of the information clearly emerges, deriving from the use of two different tax source data bases: IVA and IRAP. From the first, the location of the company can be recognized while the second, shows the position of local units in the country. The combined analyses of the two archives, gives us a view of the location of where


SAS - [BUILD: DISPLAY DB.DATA_BASE.DATA1.FRAME (E)]
File Edit View Tools Layout Build Solutions Window Help
Exit
Output - (Untitled)
Editor - Untitled1
Explorer
BUILD: SOURCE DB.DA...
BUILD: DISPLAY IN.INDI...
BUILD: DISPLAY DB...
C:\Programmi\SAS Institute\SAS\V8

BASE STATISTICA DERIVABILE DALLE DICHIARAZIONI
FISCALI IVA E IRAP PER SETTORE COMMERCIALE – 1998

Criteri di selezione: ● Class. NACE ○ Class. Conto Satellite

FILIERE di produzione

Classificazione NACE: 522 Descrizione: Comm. al dettaglio di prodotti alimentari in esercizi specializzati

Classificazione Conto satellite: Descrizione:

Selezione territoriale: ● Regioni ○ Ripartizione geografica

Regione: 1 Piemonte

● Totale Italia
○ Nord-Ovest
○ Nord-Est
○ Centro
○ Sud
○ Isole

Figure 1: Illustration of the initial template form the ECOFISCOM database.

decisions are made for production units and how these decision making centres deploy local workforces.

An important innovative aspect of ECOFISCOM is the reconstruction of the productive value chain [5]. which is provided as pre-packed output from the database. The value chain approach means, among other things, that the process of price formation can be clarified during the various exchanges which take place from the moment something is produced to its final end use. Particularly, a quantitative comparison can be obtained of the gap which is created between the price of goods leaving the production unit (at production) and the consumer price, which means that the inflationary rate, as it is perceived by the end users, can be evaluated[2]. Within this gap the most influential part is represented by profit margins, in as much that the volume is represented by a buffer which allows the inflationistic tensions to be tightened or loosened as the case may be Biffignandi et al. [2].

3.2 The implementation of information technology and the output characteristics

This database has been developed within an SAS framework applying AF modules for the value chains production for the representation and interrogation of data. The data selected from the statistical basis according to the above mentioned criteria can be extrapolated according to the most common formats (Excel, Dbase, etc).

[2] The two price systems differ for transports, trade margins and VAT, since the three items are all excluded from the production prices; they are included in the consumer prices.

Several standard analyses of the data, with their relative graphic representations, have already been drawn up. In Fig. 2, for example, the value chains production, measured by the aggregates IRAP and IVA are shown, according to the conceptual scheme illustrated in table 1, in other words represented by the positive components of income (IRAP) and the turnover (IVA). The example allows us to further illustrate the informative interest of the fiscal data base. The information relative to each variable can be further broken down according to the following characteristics of the company: location in the country, judicial status, size (expressed in terms of extent of turnover).

As mentioned previously, ECOFISCOM offers the possibilty of analysing the territorial disaggregation of the trade sector. Given the relevance of the information, however, even referring to national data, it is considered useful to include a distinction between companies acting on a regional scale to those present on a larger national scale (pluri-regional).

The judicial status of a company, above all, provides us with further information which is useful for economical analysis, as it brings to light how internal trade is articulated on traditional family values (taxable individuals) or on more complex setups like joint stock companies.

The size aspect is highlighted by distinguishing between economic subjects according to the turnover. In this case, Fig. 2, clearly shows that the Italian distribution system is extremely fractionalized and characterized by an extremely high presence of small companies. For this reason, the lower turnover companies stand out, grouped together into a single open class those companies which present turnovers of over 10 billion lira.

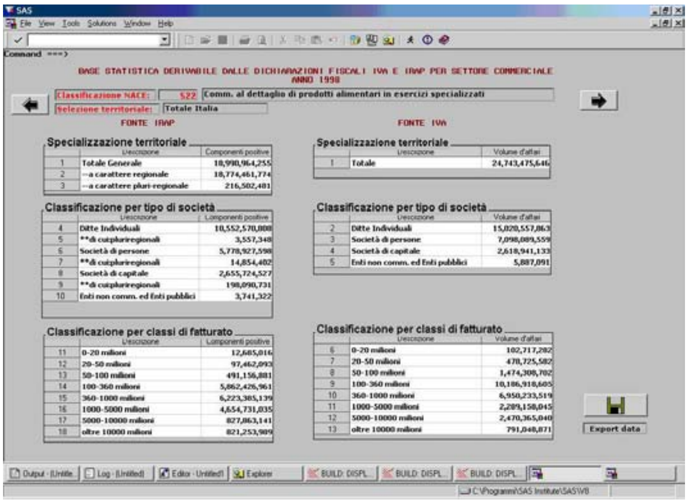
Figure 3 shows the template relative to the value chain production of the product group “Texture and finishing of textiles”. Here, only the IVA database is used since the kind of tax applied lends itself better to the template used. Since IVA is applied at the moment when the sale is carried out, the corresponding base highlights the path goods take from production to when they are sold.

In the template, on the top right hand side, the turnover of the production sector is shown, as distinguished from that part destined for overseas markets and that sold on the Italian market.

The relative data for the economic categories of wholesale exchanges are shown below (code 51561), which are specialized in the sale of the same goods. In this group, purchase data is highlighted for purchase and resale, which represent the goods acquired or the production sector (the upper part of the template) or imported.

These goods can in turn be destined for the domestic market or export markets. The sum of these two components gives us the overall turnover.

On the same line, the information regarding the sales margin is obtained by comparing the turnover with the purchase of goods for resale. This information is of vital importance as it supplies a taxable index offered by a commercial exchange at a wholesale level, influencing prices.



BASE STATISTICA DERIVABILE DALLE DICHIARAZIONI FISCALI IVA E IRAP PER SETTORE COMMERCIALE
ANNO 1998

Classificazione NACE: 522 Comm. al dettaglio di prodotti alimentari in esercizi specializzati
Selezione territoriale: Totale Italia

FONTE IRAP

Specializzazione territoriale

	Descrizione	Componenti positive
1	Totale Generale	18,990,964,255
2	--a carattere regionale	18,774,461,774
3	--a carattere pluri-regionale	216,502,481

Classificazione per tipo di società

	Descrizione	Componenti positive
4	Ditte Individuali	10,552,570,808
5	**di cuipluriregionali	3,557,348
6	Società di persone	5,778,927,598
7	**di cuipluriregionali	14,854,402
8	Società di capitale	2,655,724,527
9	**di cuipluriregionali	198,090,731
10	Enti non comm. ed Enti pubblici	3,741,322

Classificazione per classi di fatturato

	Descrizione	Componenti positive
11	0-20 milioni	12,685,016
12	20-50 milioni	97,462,093
13	50-100 milioni	491,156,881
14	100-360 milioni	5,862,426,961
15	360-1000 milioni	6,223,385,139
16	1000-5000 milioni	4,654,731,035
17	5000-10000 milioni	827,863,141
18	oltre 10000 milioni	821,253,909

FONTE IVA

Specializzazione territoriale

	Descrizione	Volume d'affari
1	Totale	24,743,475,646

Classificazione per tipo di società

	Descrizione	Volume d'affari
2	Ditte Individuali	15,020,557,863
3	Società di persone	7,098,089,559
4	Società di capitale	2,618,941,133
5	Enti non comm. ed Enti pubblici	5,887,091

Classificazione per classi di fatturato

	Descrizione	Volume d'affari
6	0-20 milioni	102,717,282
7	20-50 milioni	478,725,582
8	50-100 milioni	1,474,308,702
9	100-360 milioni	10,186,918,605
10	360-1000 milioni	6,950,233,519
11	1000-5000 milioni	2,289,158,045
12	5000-10000 milioni	2,470,365,040
13	oltre 10000 milioni	791,048,871

Figure 2: Template of the data base ECOFISCOM of production, measured according to the positive components of income (IRAP data source) and according to turnover (IVA data source).

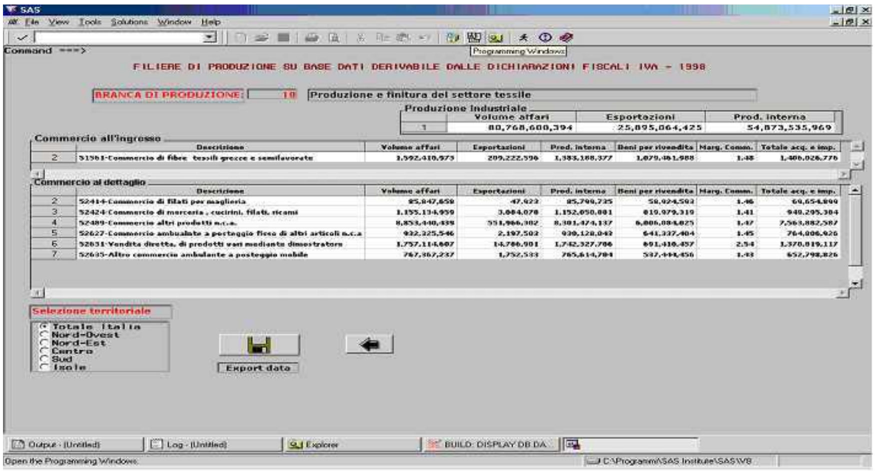
Subsequently, the same information is illustrated regarding wholesale prices against retail prices. In this case, more than one market sector is represented, given the lower level of specialization of the goods in retail sales as opposed wholesale sales.

For any further information, on request, the territorial disaggregation is available (the last part of the template).

The database therefore provides a logical path which follows a group of products from the production stage to final consumption. Although supplying important information, the following path discounts certain levels of simplification which, at the present state of information, do not seem to be able to be solved. In the first place, no information exists which allows the first question to be answered "who sells to who?". In other words, we do not know how much of the production in the textile sector is sold to wholesalers and how much is destined for the retail sector.

Another limitation is due to the presence of un-specialized commercial sales networks. This is not actually shown in the template, since due to the high supply in the sector which characterizes it, it is not easily attributable to any specific group of products. This failing can be partially completed in future by integrating fiscal data with the data from the Italian Statistical Institute through a check on sales[3].

[3] Methodological aspects are presented in Biffignandi et al. [3].



Command ===>

FILIERE DI PRODUZIONE SU BASE DATI DERIVABILE DALLE DICHIARAZIONI FISCALI IVA - 1998

BRANCA DI PRODUZIONE: 18 Produzione e finitura del settore tessile

Produzione Industriale

	Volume affari	Esportazioni	Prod. interna
1	80,768,600,394	25,895,064,425	54,873,535,969

Commercio all'ingrosso

	Descrizione	Volume affari	Esportazioni	Prod. interna	Beni per rivendita	Marg. Comm.	Totale acq. e imp.
2	51561-Commercio di fibre tessili grezze e semilavorate	1,592,410,973	209,222,596	1,383,188,377	1,079,481,988	1.48	1,406,026,776

Commercio al dettaglio

	Descrizione	Volume affari	Esportazioni	Prod. interna	Beni per rivendita	Marg. Comm.	Totale acq. e imp.
2	52414-Commercio di filati per maglieria	85,847,658	47,923	85,799,735	58,924,593	1.46	69,654,899
3	52424-Commercio di merceria , cucirini, filati, ricami	1.155.134.959	3.084.078	1.152.050.881	819.979.319	1.41	940.295.384
4	52489-Commercio altri prodotti n.c.a.	8,853,440,439	551,966,302	8,301,474,137	6,006,084,025	1.47	7,563,882,587
5	52627-Commercio ambulante a posteggio fisso di altri articoli n.c.a	932,325,546	2,197,503	930,128,043	641,337,404	1.45	764,006,926
6	52631-Vendita diretta, di prodotti vari mediante dimostratore	1,757,114,607	14,786,901	1,742,327,706	691,410,457	2.54	1,370,819,117
7	52635-Altro commercio ambulante a posteggio mobile	767,367,237	1,752,533	765,614,704	537,444,456	1.43	652,798,826

Selezione territoriale

● Totale Italia
○ Nord-Ovest
○ Nord-Est
○ Centro
○ Sud
○ Isole

Open the Programming Windows.

C:\Programmi\SAS Institute\SAS\V8

Figure 3: Template of the data base ECOFISCOM for the value chain of the group of products: Texture and finishing of textiles

4 Concluding remarks

The ECOFISCOM data base is presented in this paper, which is based on administrative data (fiscal), which has allowed more specific analysis to be generated.

The information is presented following a satellite accountability approach and grants a micro reading of the structural characteristics of internal trade. Regarding studies carried out on the variable definitions, these microanalyses can be linked to the macro approach of national accounting.

The database structure has been developed in an SAS environment, with the application of AF modules for the construction of visual templates and data interrogation. It is important to underline the development of information technology in a user friendly environment, with the creation of innovative contents (such as value chains) and output containing indicators which are useful from an economic point of view (such as the calculation of sales margins). The data base allows key questions to be asked by sector, by territorial areas and by value chain; it is also possible to obtain information of a judicial nature, the territorial spread and the size of the company. Given the perfect comparability of ECOFISCOM with the common software (such as Excel, Dbase, etc.), this tool can also be used to create personalized databases which can supply data to other software applications.

ECOFISCOM is a dynamic database based on data taken from IVA and IRAP tax returns which are drawn from the economic activities, from the country and production bases.

One aspect of the value added to the product which must not be overlooked is the application of metadata associated with it, which allows the informative content to be unquestionably singled out of each variable and to compare this content with international standardized definitions (given by the national accounts).

The database, in its present form, can be accessed only in an SAS form; in future it will be able to be consulted on the web; a version with selected tables may be available on CD-Rom. A further planned development is the widening of the database through its integration with data made available from surveys.

References

[1] Biffignandi S., Pisani S. (2000). *Satellite accounts for trade: How to build them and how to set up suitable data.* Paper presented at Second International Conference on Establishment Surveys, Buffalo , N. Y. USA, June 17-21 2000.

[2] Biffignandi S. , Gismondi R. (2004). *Analysing the trade and production value chain based on integration between fiscal data and Istat surveys.* Paper submitted to the XLII scientific meeting of the Italian Statistical Society, Bari, Italy.

[3] Biffignandi S., Della Rocca G., Pisani S. (2004). *The role of trade margins in the price determination process.* Paper submitted to the XLII scientific meeting of the Italian Statistical Society, Bari, Italy.

[4] Eurostat. (2000). *Satellite accounts for trade.* Theme.

[5] Morvan Y. (1985). *L'Analyse de filiére.* Economica, Paris.

[6] United Nation, Commission of the European Communities, IMF, OECD, World Bank (1993). *System of National Accounts*, New York.

Acknowledgement: This paper is financially supported by the Cofin Miur 40"The use of new construction strategies of the data and the statistical information for studying and planning economic systems", local unit University of Bergamo, project on "Satellite account for trade and price data: criteria, methods and applications". Database by Giorgio Della Rocca.

Address: S. Biffignandi, Dipartimento di Matematica Statistica, Informatica e Applicazioni, Facolta Economia, Universita di Bergamo, Via Caniana n. 2, 24127 Bergamo (Italy)
S. Pisani, Agenzia Entrate, Roma, Italy

E-mail: silvia.biffignandi@unibg.it;
stefano.pisani@agenziaentrate.it.

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

LOCALIZED LOGISTIC CLASSIFICATION WITH VARIABLE SELECTION

Harald Binder and Gerhard Tutz

Key words: Local logistic regression, discrimination, data adaptive tuning parameters, selection of predictors, localized discrimination.

COMPSTAT 2004 section: Classification.

Abstract: The main problem with localized discriminant techniques is the curse of dimensionality, which seems to restrict their use to the case of few variables. In the present paper logistic discrimination is used locally and combined with local variable selection. A robust localized logistic regression method is developed for which all tuning parameters are chosen data-adaptively. The procedure allows to use higher numbers of variables with discrimination performance that is comparable to the best available classifiers. The performance is evaluated in simulation studies and by using real data sets and compared to various alternative procedures such as linear discriminant analysis, nearest neighbourhood classifiers, trees, random forests and variants of boosting.

1 Introduction

In recent years the local fitting of parametric models has become a powerful tool in nonparametric regression, see e.g. Fan & Gijbels [4], Loader [10]. Local binary regression may reveal how covariates determine the binary response. When used in classification problems the decision boundaries are smooth functions instead of hyperplanes as is the case in simple logistic discrimination. A drawback of the method which carries over to the classification problems is the restriction to the low dimensional case. For example Loader [10] applied localized logistic discrimination techniques to cases with two and four variables. The approach presented here deals with the problem of higher dimensions by combining localization with local dimension reduction by localized variable selection.

In a first localization step parameter estimation for the logistic regression model is obtained by maximizing weighted likelihood where the weights are based on Euclidian distances between the observations from the training set and the target value for which prediction is wanted. In a second step we use a simple simultaneous variable selection procedure. In the reduced space of the selected variables the distance-based weights are recalculated and estimation is repeated to obtain the final model for prediction.

This classification approach is observation specific in the sense that a new classifier is computed for each new observation. The same holds for example for nearest neighbourhood classifiers in contrast to procedures like linear or quadratic discrimination where the classification rule is computed once by estimating parameters. Procedures such as linear discrimination may be seen as global approaches.

Alternative approaches to local dimension reduction have been given by Schaal et al. [13], Hastie et al. [6]. In contrast to these approaches variable selection has the advantage that one obtains information about the relevance of variables even if the selection is performed locally. In statistical applications the user is often interested which variables are relevant and have to be collected in the future.

2 Localized logistic discrimination

Let the observations be given by (x_i, y_i), $i = 1, \ldots, n_L$, for $x_i' = (x_{i1}, \ldots, x_{ip})$ being a p-variate predictor variable and $y_i \in \{0, 1\}$ the class indicator. The objective is to predict y for a new observation for which only x is observed. In the following only the case of two classes is considered although the extension to the k class situation is straightforward.

Localized logistic discrimination is based on the fitting of the model

$$\log\left\{\frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)}\right\} = z_i'\beta$$

where β is a parameter vector of length $m + 1$ and z_i is a design vector built from x_i. For linear logistic discrimination $z_i' = (1, x_i')$ and for quadratic logistic discrimination $z_i' = (1, x_i', x_{i1}^2, \ldots, x_{ip}^2)$ is used.

The model is fitted locally by using weights within the (log-)likelihood. For target value x the weighted log-likelihood is given by

$$l_x(\beta) = \sum_i \left(y_i \log \pi(x_i) + (1 - y_i)\log(1 - \pi(x_i))\right) w_k(z, z_i) \tag{1}$$

where $\pi(x_i) = P(y_i = 1|x_i)$ and z, z_i are the predictor values connected to x, x_i, i.e. $z' = (1, x')$, $z_i = (1, x_i')$ in the linear case and $z' = (1, x', x_1^2, \ldots, x_p^2)$, $z_i' = (1, x_i', x_{i1}^2, \ldots, x_{ip}^2)$ in the quadratic case. By using a loess type localization the locally adaptive weights $w_k(z, z_i)$ are chosen to depend on the (Euclidian) distance between the (transformed) target value z and the (transformed) observation z_i and a kernel window

$$w_k(z, z_i) = K\left(\frac{||z - z_i||}{\lambda(z)}\right) \tag{2}$$

where the kernel width parameter $\lambda(z)$ is chosen as the distance to the kth nearest neighbour $z_{(k)}$ of z, i.e. $\lambda(z) = ||z - z_{(k)}||$. Various kernel functions K can be used. For our investigations we used the tricube kernel which has the computational advantage that in estimation only points in the neighbourhood are included since all other points receive weight zero.

Parameter estimation is performed by solving the local score equation $s_{x,k}(\beta) = 0$ by iterative Fisher scoring of the form

$$\hat{\beta}_x^{(s+1)} = \hat{\beta}_x^{(s)} + F_{x,k}(\hat{\beta}_x^{(s)})^{-1} s_{x,k}(\hat{\beta}_x^{(s)}) \tag{3}$$

where $s_{x,k}(\beta) = \partial l_x/\partial\beta$ is the local score function which for the logistic model has the simple form

$$s_{x,k}(\beta) = \sum_i w_k(z, z_i) z_i (y_i - \pi_i(\beta))$$

with $\pi_i(\beta)$ denoting the response probability evaluated at β and $F_{x,k} = E(-\partial^2 l_x/\partial\beta\partial\beta')$ denoting the weighted Fisher matrix

$$F_{x,k} = \sum_i w_k(z, z_i) z_i z_i' \frac{\partial h(\eta_i)}{\partial\eta}$$

with $\eta_i = z_i'\beta$ and $h(x) = \exp(x)/(1 + \exp(x))$ being the response function of the logistic regression model. The dependence of the parameter estimates on the target value shows in the notation $\hat\beta_x$. For the asymptotic behaviour of local estimates see Fan and Gijbels [4].

3 Local variable selection

In addition to localizing the logistic model the variables which are used for discrimination are locally selected. The underlying assumption is that not all predictors are equally informative on class membership throughout the space spanned by all predictors.

For computational efficiency a one-step selection procedure is used that determines the relevance of predictors by a simple variant of Wald tests. For the local estimates $\hat\beta_x$ at target value x the variance may be approximated by $\hat{\text{cov}}(\hat\beta_x) = F_{x,k}(\hat\beta_x)^{-1}$ (see [8]). Based on the approximation variables are selected by considering the studentized value

$$c_{x,k}(\hat\beta_{x,j}) = \frac{|\hat\beta_{x,j}|}{\sqrt{\hat{\text{var}}(\hat\beta_{x,j})}} \quad , \ j = 1, \ldots, m,$$

where $\beta_x' = (\beta_{x,0}, \beta_{x,1}, \ldots, \beta_{x,m})$. In a single step those predictors are selected for which $c_{x,k}(\hat\beta_j)$ exceeds a threshold c_β. $c_{x,k}(\hat\beta)$ is a localized version of the Wald statistic for testing the null hypothesis $\beta_j = 0$ locally. In the case of linear logistic discrimination where z_i is given by $(1, x_{i1}, \ldots, x_{ip})$ predictor selection refers to the original variables $x_1, \ldots, x_p$ whereas in quadratic logistic discrimination where z_i is given by $(1, x_{i1}, \ldots, x_{ip}, x_{i1}^2, \ldots, x_{ip}^2)$ predictor selection refers to the extended set of variables $x_1, \ldots, x_p, x_1^2, \ldots, x_p^2$.

When the number of predictors has been reduced the weights $w(z, z_i)$ are recalculated for the subspace spanned by the selected predictors and the estimation is performed for the reduced β-vector of the final local model. Prediction for target value x then is based on the reduced and re-estimated model.

In order to avoid instabilities of the estimation procedure it is useful to use some robustified estimator for β_x. We consider a penalized estimator (e.g. Le Cessie & Van Houwelingen [9]) by using the weighted log-likelihood

$$l(\beta) = \sum_i \left(y_i \log \pi(x_i) + (1 - y_i) \log(1 - \pi(x_i))\right) w_k(z, z_i) - \lambda \beta' P \beta. \tag{4}$$

where P is the penalization matrix and λ determines the strength of the penalization. Setting λ to 0 would result in the unpenalized likelihood (1). For the penalty matrix P a simple identity-matrix is used. This leads to a penalization of β_j^2 and is similar to logistic ridge regression [9].

The modifed expression for the penalized weighted local score function and Fisher matrix are

$$s_{x,k}(\beta) = \sum_i w_k(z, z_i) z_i (y_i - \pi_i(\beta)) - 2\lambda P \beta$$

and

$$F_{x,k} = \sum_i w_k(z, z_i) z_i z_i' + 2\lambda P.$$

Special attention has to be given to the intercept parameter β_0. All predictors are centered and standardized in a weighted way and the intercept is kept fixed at the value of the transformed local class membership proportion.

The performance of the algorithm depends on some parameters which have to be chosen. These are: 1. the order k of the nearest neighbourhood used in (2); 2. the threshold c_β which determines how many variables are selected locally; 3. the tuning parameter λ which determines the strength of penalization.

In order to obtain a fully data-adaptive procedure these tuning parameters are chosen by cross-validation. The cross-validation criterion is given by the prediction error rate. Optimization is based on a recently proposed procedure that uses quadratic approximations (see [12]).

4 Simulation study

The simulation study compares the performance of various classification procedures for various types of data structures. The investigated classification procedures are localized logistic regression with variable selection (LLD), linear discriminant analysis (LDA), single-hidden-layer neural networks with five units in the hidden layer and committee voting (NNet), 10-nearest-neighbourhood classification (10-NN), cross-validated classification trees (Tree) and random forests (see [2]). The implementations used are those from the statistical environment R [7] and we used the standard settings for all procedures. The investigated data structures are based on the structure suggested by Friedman [5] and Hastie and Tibshirani [6] denoted by 'Fx' and 'HTx' where 'x' is the number of the example in the respective articles. The size of

	LLD	LDA	NNet	10-NN	Tree	RF
multivariate normal						
2-dim (HT1)	0.071	0.065	0.070	0.074	0.104	0.086
2-dim with noise (HT2)	0.100	0.077	0.089	0.178	0.115	0.135
10-dim, var. inf. (F1)	0.016	0.019	0.019	0.026	0.089	0.019
10-dim, diff. inf. (F2)	0.018	0.022	0.026	0.031	0.124	0.021
no overlap, connected						
linear combination (F5)	0.060	0.053	0.030	0.171	0.329	0.166
quadratic comb. (F4)	0.199	0.453	0.263	0.413	0.309	0.179
weighted quad. (F3)	0.249	0.507	0.240	0.403	0.229	0.164
quad., no noise (HT5)	0.121	0.496	0.128	0.223	0.196	0.126
quad., some noise (HT5)	0.218	0.499	0.241	0.325	0.210	0.145
quad., more noise (HT5)	0.313	0.505	0.397	0.381	0.218	0.180
fractioned class structure						
without noise (HT3)	0.045	0.330	0.045	0.023	0.048	0.031
with noise (HT4)	0.244	0.342	0.164	0.262	0.059	0.096

Table 1: Mean error rates for different simulated data examples and classification procedures.

the learning data was 200 and the number of observations in the test data was 1000 when not stated otherwise. In both data sets the frequencies of the true classes have been chosen to be equal. The results are based on 50 replications for each example. Table 1 gives the mean error rates for all procedures and all examples.

In example HT1 two covariates are drawn from normal distributions with non-zero covariance and different means for the two classes. Linear discriminant analysis (LDA) is seen to have the best performance. This does not come as a surprise because the decision boundary is a straight line and can be matched very well by LDA. Localized logistic regression (LLD) is similar to LDA in performance. This similarity is also releflected in the selection of large values of k. When 14 noise variables are added (HT2) the performance of LLD degrades but not as much as that of 10-nearest-neighbourhood. It still performs better than tree-based methods.

In examples F1 and F2 the amount of information provided by the variables is systematically varied. For both classes the covariates are drawn from a 10-dimensional normal distribution. For one of the two classes the mean and the variance depend on the variable index. In example F1 the covariates with the higher index j are intended to be more relevant. The good performance of procedures using linear combinations of predictors (LLD, LDA and neural networks) indicates that the Bayes decisions boundary can be approximated very well by hyperplanes. In the second example (F2) the mean structure of the second class is changed in that the variables with lower index contain more relevant information on class membership due to the mean and the variables with higher index due to the variance. Again a hyperplane approximation of the Bayes decision boundary seems to be efficient.

In example F5 the class boundary is defined by a linear combination of ten covariates. The procedures that utilize a linear combination of predictors clearly have the best performance (with LLD among them). With a quadratic combination of covariates and training sample size of $n_L = 500$ and $n_T = 1000$ (F4) the performance of the linear procedures LDA and neural networks degrades. LLD shows good performance performance, with only random forests being better. When a weighted contribution of the variables is used (F3) the performance of linear LLD degrades to the level of neural networks. This indicates that the class boundary is too complicated to be approximated very well by local linear models.

In the examples HT5 there are four covariates drawn from standard normal distributions that define a spherical class region augmented with a varying number of noise variables. When no noise is present LLD performs very well here. When adding six noise variables the performance of LLD decreases relative to procedures like random forest and decreases even more with 16 noise variables.

In the two examples with extremely fractioned class regions (HT3 and HT4) the distribution of each of the two classes is defined as a mixture of six spherical bivariate normal subclasses with extremely scattered means. There are 20 observations drawn from the distribution of each subclass and so there are 140 observations per class with a total of $n_L = 240$ observations ($n_T = 960$). With no noise (HT3) all procedures (except LDA) perform very similar on this data. When we augmented the two variables carrying information on class membership with eight noise covariates (HT4) the performance of LLD degrades compared to the partition-based classification tree and random forest procedures. The latter procedures seem to perform very well in seperating informative variables from noise variables.

4.1 Summary of simulation results

It is seen that for different situations different classification methods turn out to be the best choice, but some procedures react more flexibly to varying data structures. Given that it cannot be expected that one method is superior in all data situations LLD performs rather well on a variety of different data structures.

LLD shows better performance than LDA and the 10-nearest neighbourhood approach in almost all examples. Neural networks as used here perform distinctly better only in two examples, despite the fact that in contrast to LLD they can model interactions of covariates directly. The same holds for simple trees. For data with fractioned class structure with noise variables tree-based approaches perform very well in particular if noise variables are included. Although LLD performs better without noise variables it is outperformed if much noise is present. Advanced tree methodology as present in random forests clearly performs best in this case.

	Australian credit	breast cancer	sonar
LLD	0.144	0.029	0.091
LDA	0.146	0.037	0.273
NNet	0.140	0.035	0.165
10-NN	0.313	0.029	0.336
Tree	0.153	0.054	0.271
RF	0.125	0.028	0.164
L2Boost*	0.123	0.037	0.228
L2WCBoost*	0.123	0.040	0.190
LogitBoost*	0.131	0.039	0.158

* from Bühlmann and Yu [3]

Table 2: Error rates for real data and various classification procedures. The numbers are mean error rates for 50 random splits into a 90% training and 10% test set.

5 Application to real data

We use the Australian credit data from the Statlog project [11] and the breast cancer and the sonar data from the UCI machine learning repository [1]. One reason for this selection of data sets is that they have been used in recent work on boosting methods [3] and so information on error rates is available for a class of procedures that is considered to perform very well.

Each data set has been split 50 times randomly into a 90% training and 10% test set and all procedures used in the simulation study have been applied. Table 2 shows the error rates for the three data sets and all procedures used in the simulation study. In addition the error rates for several boosting procedures as given in Bühlmann and Yu [3] are shown.

For the Australian credit data 10-nearest neighbourhood classification rules yield very bad performance while the rest of the procedures are well comparable. For the breast cancer example LLD, 10-nearest neighbourhood classification and random forests distinctly outperform the rest. For the sonar data LLD is distinctly the best performer. As is seen LLD performs well for all three data sets.

6 Concluding remarks

A localized discrimination procedure has been proposed which in combination with local selection of predictors shows promising results, that might be even improved by different parameter selection schemes. Although a method cannot be expected to be best for all potential data structures the performance is surprisingly good over a wide range of data structures. While it outperforms advanced tree methodology for simple structures, the latter dominate at least a linear version for quadratically separated classes with many noise variables. For real data sets, the localizing methodology works very well with the best performance for two of the considered data sets. This shows the high potential in statistical applications.

References

[1] Blake C., Merz C. (1998). *UCI machine learning repository.*

[2] Breiman L. (2001). *Random Forests.* Machine Learning **45**(1), 5 – 32.

[3] Bühlmann P., Yu B. (2003). *Boosting with the L2Loss: regression and classification.* J. Amer. Statist. Assoc. **98**(462), 324 – 339.

[4] Fan J., Gijbels I. (1996). *Local polynomial modelling and its applications.* London: Chapman & Hall.

[5] Friedman J.H. (1994). *Flexible metric nearest neighbor classification.* Technical report, Standford University.

[6] Hastie T., Tibshirani R. (1996). *Discriminant adaptive nearest neighbor classification.* IEEE Trans. Pattern Analysis and Machine Intelligence **18**(6), 607 – 615.

[7] Ihaka R., Gentleman R. (1996) *R: A language for data analysis and graphics.* J. Comp. and Graph. Statistics **51**(3), 299 – 314.

[8] Kauermann G., Tutz G. (2000). *Local likelihood estimates and bias reduction in varying coefficients models.* J. Nonparam. Statistics **12**, 343 – 371.

[9] Le Cessie S., van Houwelingen J.C. (1992). *Ridge estimators in logistic regression.* Applied Statistics **41**(1), 191 – 201.

[10] Loader C. (1999). *Local regression and likelihood*. New York: Springer.

[11] Michie D., Spiegelhalter D.J., Taylor C.C. (1994). *Machine learning, neural and statistical classification*. New York: Ellis Horwood.

[12] Powell M.J.D. (2002) *UOBYQA: unconstrained optimization by quadratic approximation.* Math. Program. **92**, 555 – 582.

[13] Schaal S., Vijayakumar S., Atkeson C.G. (1998). *Local dimensionality reduction.* In: M.I. Jordan, M.J. Kearns, and S.A. Solla (eds.): Advances in Neural Information Processing Systems 10. Cambridge, MA: MIT Press.

Acknowledgement: We thank the German Science Foundation DFG (Sonderforschungsbereich 386) for financial support.

Address: H. Binder, Institut für Statistik, Ludwig-Maximilians-Universität München, Akademiestr. 1, D–80799 München, Germany
G. Tutz, Klinik für Psychiatrie und Psychotherapie, Universität Regensburg, Germany

E-mail: tutz@stat.uni-muenchen.de

NEW STAR MODELS OF TIME SERIES AND THEIR APPLICATION IN FINANCE

Tomáš Bognár, Jozef Komorník and Magda Komorníková

Key words: Time series, regime-switching autoregressive models, transition functions.

COMPSTAT 2004 section: Time series analysis.

Abstract: A new class PSTAR of Smooth Transition Autoregressive models, based on cubic spline type transition functions, has been recently introduced in [1] subjected to comparation with LSTAR models based on the traditional logistic functions. A very high degree of similarity between both classes of models has been demonstrated. PSTAR models can be slightly preferable because of their more simple formal and geometrical structure that may enable users more convenient manipulation in statistical inference procedures.

In this paper, a general approach to construction of STAR models that covers both PSTAR and LSTAR classes is suggested.

An interesting application to modeling exchange rates time series, where the threshold variable has delay 5 (corresponding to the number of business days in 1 week), has been found.

1 Introduction

Smooth transition autoregressive models (STAR) have been extensively analyzed and applied by many authors during the last two decades. They have been introduced in [3] as a smooth alternative to Treshold Autoregressive models (representing nonlinear generalizations of autoregressive models) that assume different autoregressive models describing behaviour of an investigated time series y_t in different regimes. Many interesting applications of STAR and other nonlinear models have been presented in [2].

2 New STAR models

A formal representation of a 2-regimes STAR can be expressed by

$$y_t = \Phi_1(B) y_t [1 - G(y_{t-d}; \gamma, c)] + \Phi_2(B) y_t G(y_{t-d}; \gamma, c) + \epsilon_t \qquad (1)$$

(see [4]), where

ϵ_t is a white noise sequence with variance σ^2,

the autoregressive polynomials

$$\Phi_i(B) = \phi_{i,0} + \phi_{i,1} B + \cdots + \phi_{i,p_i} B^{p_i}, \qquad i = 1, 2$$

in the shift operator B (defined by $By_t = y_{t-1}$) are related to regimes that are determined by values of a threshold variable y_{t-d} and its treshold level value c.

The original LSTAR models proposed in [3], [4] are based on the logistic transition function.

A logistic transition function has the form

$$G(y_{t-d}; \gamma, c) = \frac{1}{1 + \exp(-\gamma[y_{t-d} - c])} \tag{2}$$

where γ is the smoothness parameter.

It is obvious that $G(c; \gamma, c) = \frac{1}{2}$ and $G(y; 0, c) = \frac{1}{2}$ for any $y, c \in R$ and $\gamma \geq 0$.
If we put $q_t = y_{t-d} - c$
and

$$\begin{aligned} G^*(q; \gamma) &= G(c + q; \gamma, c) - \frac{1}{2} = \frac{1}{1 + \exp(-q\gamma)} - \frac{1}{2} = & (3) \\ & & (4) \\ &= \frac{1 - \exp(-q\gamma)}{2(1 + \exp(-q\gamma))} = \frac{1}{2}\text{tgh}(-q\gamma) & (5) \end{aligned}$$

we can rewrite (1) in the form (see [2])

$$y_t = \frac{1}{2}[\Phi_1(B) + \Phi_2(B)]y_t + [\Phi_2(B) - \Phi_1(B)]y_t G^*(q_t; \gamma) + \epsilon_t \tag{6}$$

which can be applied for testing linearity of the model (the hypothesis $\Phi_1(B) = \Phi_2(B)$ which is equivalent to the hypothesis $H_0 : \gamma = 0$ in (3)). For this test, the following third-order Taylor polynomial approximation to $G^*(q; \gamma)$ in the right neighborhood of $\gamma = 0$ was utilized in [2] (where an error in the sign of the third-order term occurred).

For any $\gamma > 0$, G^* as a function of q is convex for $q < 0$ and concave for $q > 0$, thus the partial derivative $\frac{\partial G^*(q;\gamma)}{\partial q}$ attains its maxima $\frac{\gamma}{4}$ for $q = 0$:

$$T_3(q, \gamma) = \gamma \left[\frac{\partial G^*(q; \gamma)}{\partial \gamma}\right]_{\gamma=0} + \frac{1}{6}\gamma^3 \left[\frac{\partial^3 G^*(q; \gamma)}{\partial \gamma^3}\right]_{\gamma=0} = \frac{1}{4}\gamma q - \frac{1}{48}\gamma^3 q^3 \tag{7}$$

Note that $T_3(q, \gamma)$ as well as G^* and G are symmetric in the pair of variables q and γ since they depend only on the product $x = q\gamma$.

For a new STAR model, we can deal with any even non-decreasing smooth surjective function

$$\mathcal{G} : [-\infty, +\infty] \to \left[-\frac{1}{2}, \frac{1}{2}\right]$$

which will be called a *shape function.*

For $\gamma \geq 0$ we will so deal with

$$H^*(q;\gamma) = g(\gamma.q).$$

Then our transition function is given by

$$H(y_{t-d};\gamma;c) = g(\gamma.(y_{t-d} - c)) + \frac{1}{2}$$

Typical examples of shape functions are:

$$g(x) = \frac{1}{\pi} \text{ arctg } \text{ x};$$

$$g(x) = \begin{cases} -\frac{1}{2} & \text{if } x < -\frac{1}{\sqrt{2}} \\ (x + \frac{1}{\sqrt{2}})^2 - \frac{1}{2} & \text{if } -\frac{1}{\sqrt{2}} \leq x < 0 \\ -(x - \frac{1}{\sqrt{2}})^2 + \frac{1}{2} & \text{if } 0 \leq x \leq \frac{1}{\sqrt{2}} \\ \frac{1}{2} & \text{if } x > \frac{1}{\sqrt{2}} \end{cases} \quad \text{(quadratic spline);}$$

$$g(x) = \begin{cases} -\frac{1}{2} & \text{if } x < -\frac{\pi}{2} \\ \frac{1}{2} sin \text{ x} & \text{if } -\frac{\pi}{2} \leq x \leq \frac{\pi}{2} \\ \frac{1}{2} & \text{if } x > \frac{\pi}{2} \end{cases}$$

In [1] a third-order spline shape function P^* has been introduced

$$g(x) = \text{ min } \left(\frac{1}{2}, \text{ max } \left(-\frac{1}{2}, \frac{1}{4}x - \frac{1}{108}x^3 \right) \right).$$

Then

$$P^*(q,\gamma) = \begin{cases} -\frac{1}{2} & q < -\frac{3}{\gamma} \\ \frac{1}{4}q\gamma - \frac{1}{108}q^3\gamma^3 & -\frac{3}{\gamma} \leq q \leq \frac{3}{\gamma} \\ \frac{1}{2} & q > \frac{3}{\gamma} \end{cases} \tag{8}$$

Each shape function H^* has the following properties:

a)

$$H^*(0,\gamma) = 0 \qquad \text{for any } \gamma > 0.$$

b)

$$H^*(q,\gamma) + H^*(-q,\gamma) = 0 \qquad \text{for any } q \in R \text{ and } \gamma > 0.$$

c)

$$\frac{\partial H^*(q;\gamma)}{\partial q} > 0 \qquad \text{for all } q \in R \text{ and } \gamma > 0.$$

d) For any $\gamma > 0$

$$\lim_{q\to\infty} H^*(q;\gamma) = \frac{1}{2} \qquad \text{and} \qquad \lim_{q\to-\infty} H^*(q;\gamma) = -\frac{1}{2}.$$

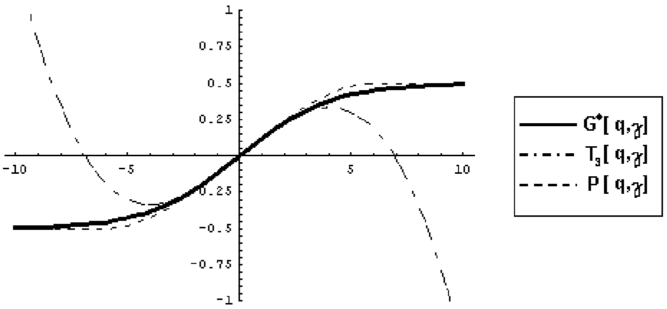


Figure 1: $\gamma = 0.5$

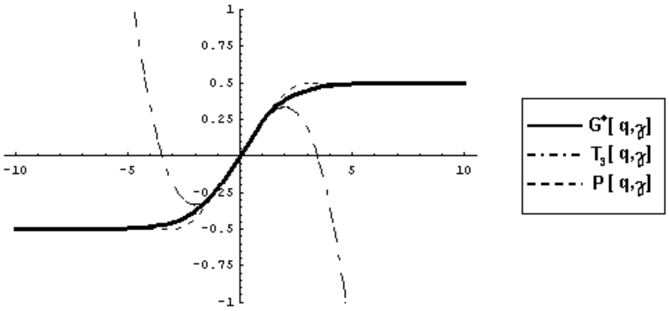


Figure 2: $\gamma = 1$

e) For any $q > 0$, $H^*(q;\gamma)$ is nondecreasing in γ,

$$\lim_{\gamma\to\infty} H^*(q;\gamma) = \frac{1}{2} \qquad \text{and} \qquad \lim_{\gamma\to-\infty} H^*(-q;\gamma) = -\frac{1}{2}.$$

f) For any $q \in R$

$$\lim_{\gamma\to 0} H^*(q;\gamma) = 0.$$

In case of cubic splines, we can see in the Figures 1-3, the polynomials $P^*(q,\gamma)$ provide much better global approximations to $G^*(q,\gamma)$ than $T_3(q,\gamma)$.

Moreover, the function $P^*(q,\gamma)$ is identical with its third-order Taylor polynomial approximation in the right neighborhood of $\gamma > 0$ for $|q| < \gamma^{-1}$ which guarantees higher precision of the same test of linearity that has been applied for logistic models.

Moreover

$$\frac{\partial G^*(0;\gamma)}{\partial\gamma} = \frac{\partial P^*(0;\gamma)}{\partial\gamma} = \frac{\gamma}{4}$$

Inspecting behaviour of the difference $\frac{\partial P^*(q;\gamma)}{\partial\gamma} - \frac{\partial G^*(q;\gamma)}{\partial\gamma}$ we conclude that for any $\gamma > 0$ the difference $P^*(q,\gamma) - G^*(q,\gamma)$ is positive for $q > 0$, increasing on

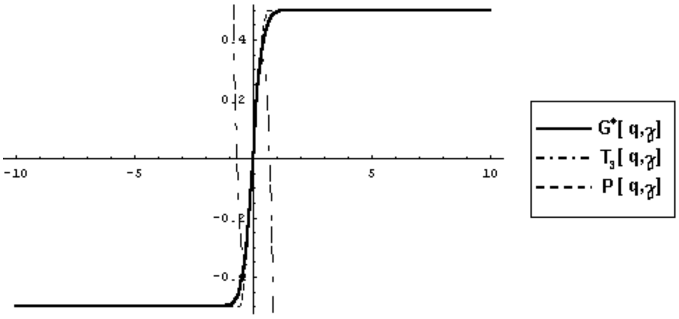


Figure 3: $\gamma = 5$

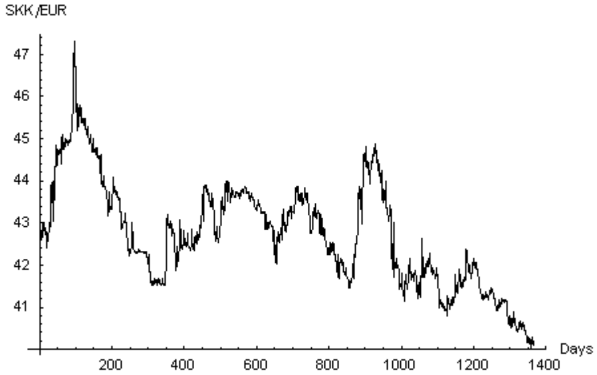


Figure 4: The time series daily exchange rates of Slovak Crown to Euro.

the interval $(0, q_0)$, where $q_0 \approx \frac{2.58}{\gamma}$, and decreasing on (q_0, ∞). The maximal difference $P^*(q_0, \gamma) - G^*(q_0, \gamma) \approx 0.056$ independently on γ.

Similarly we can obtain that the difference $G^*(q, \gamma) - T_3(q, \gamma)$ is positive for any $\gamma > 0, q > 0$.
$T_3(q, \gamma)$ is better approximation to $G^*(q, \gamma)$ than $P^*(q, \gamma)$ on the interval $(0, q_1)$, where $q_1 \approx \frac{1.966}{\gamma}$ and $G^*(q_1, \gamma) - T_3(q_1, \gamma) = P^*(q_1, \gamma) - G^*(q_1, \gamma) \approx 0.044$ independently on γ.

The maximum of differences $[(P^*(q, \gamma) - G^*(q, \gamma)) - (G^*(q, \gamma) - T_3(q, \gamma)]$ is obtained in $q = q_2 \approx \frac{1.482}{\gamma}$ and its value is 0.013 independently on γ.

3 Application

We have investigated the time series daily exchange rates of Slovak Crown to Euro in the period January 1, 1999 - April 22, 2004 (see Figure 4).

The partial autocorrelation function for this time series (see Figure 5)

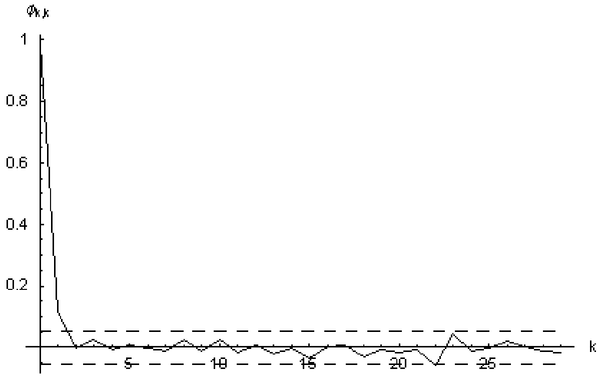
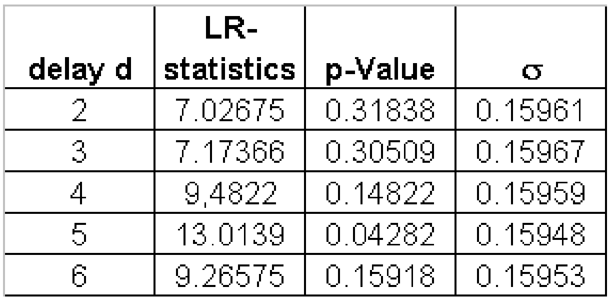


Figure 5: Partial autocorrelation function for time series daily exchange rates of Slovak Crown to Euro.



delay d	LR-statistics	p-Value	σ
2	7.02675	0.31838	0.15961
3	7.17366	0.30509	0.15967
4	9,4822	0.14822	0.15959
5	13.0139	0.04282	0.15948
6	9.26575	0.15918	0.15953

Table 1: Test of nonlinearity for AR(2).

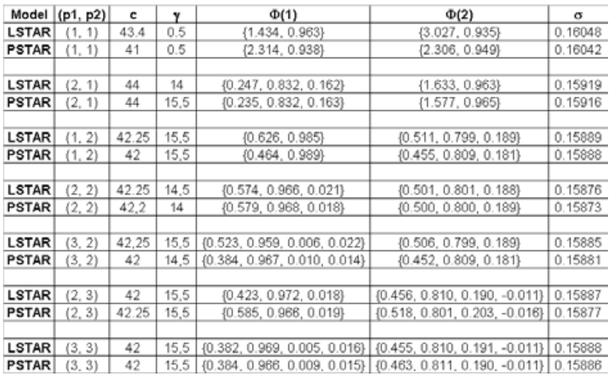
suggests that in the class of linear AR(p) models are preferable low order models. Stepwise tests of AR(p) agains submodels AR(p-1) showed that the optimal AR(p) model is AR(2).

Further we continued by investigating 2-regimes models of the class AR(2) with threshold variables y_{t-d} for d = 1, 2, 3, 4, 5, 6. We applied tests of nonlinearity (see [2] based on the Taylor polynomial approximation mentioned above. From the Table 1 we conclude that only for d = 5 the linear model can be rejected against the general nonlinearity hypothesis.

These results can be related to the fact that there are typically 5 working days in a week and Mondays' and Fridays' trading may exhibit certain specific features.

The following Table 2 sumarizes the results of computations for different LSTAR and PSTAR models with threshold variable y_{t-5}.

We see that for all pairs (p1, p2) PSTAR models provide slightly better fit than their LSTAR alternatives. However the value of coefficients of corresponding PSTAR and LSTAR models are very similar and differences in



Model	(p1, p2)	c	γ	Φ(1)	Φ(2)	σ
LSTAR	(1, 1)	43.4	0.5	{1.434, 0.963}	{3.027, 0.935}	0.16048
PSTAR	(1, 1)	41	0.5	{2.314, 0.938}	{2.306, 0.949}	0.16042
LSTAR	(2, 1)	44	14	{0.247, 0.832, 0.162}	{1.633, 0.963}	0.15919
PSTAR	(2, 1)	44	15,5	{0.235, 0.832, 0.163}	{1.577, 0.965}	0.15916
LSTAR	(1, 2)	42.25	15,5	{0.626, 0.985}	{0.511, 0.799, 0.189}	0.15889
PSTAR	(1, 2)	42	15,5	{0.464, 0.989}	{0.455, 0.809, 0.181}	0.15888
LSTAR	(2, 2)	42.25	14,5	{0.574, 0.966, 0.021}	{0.501, 0.801, 0.188}	0.15876
PSTAR	(2, 2)	42,2	14	{0.579, 0.968, 0.018}	{0.500, 0.800, 0.189}	0.15873
LSTAR	(3, 2)	42,25	15,5	{0.523, 0.959, 0.006, 0.022}	{0.506, 0.799, 0.189}	0.15885
PSTAR	(3, 2)	42	14,5	{0.384, 0.967, 0.010, 0.014}	{0.452, 0.809, 0.181}	0.15881
LSTAR	(2, 3)	42	15,5	{0.423, 0.972, 0.018}	{0.456, 0.810, 0.190, -0.011}	0.15887
PSTAR	(2, 3)	42.25	15,5	{0.585, 0.966, 0.019}	{0.518, 0.801, 0.203, -0.016}	0.15877
LSTAR	(3, 3)	42	15,5	{0.382, 0.969, 0.005, 0.016}	{0.455, 0.810, 0.191, -0.011}	0.15888
PSTAR	(3, 3)	42	15,5	{0.384, 0.966, 0.009, 0.015}	{0.463, 0.811, 0.190, -0.011}	0.15886

Table 2: The results of computations for different LSTAR and PSTAR models with threshold variable y_{t-5}.

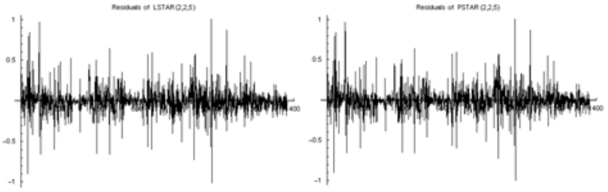


Figure 6: Residuals of LSTAR and PSTAR models.

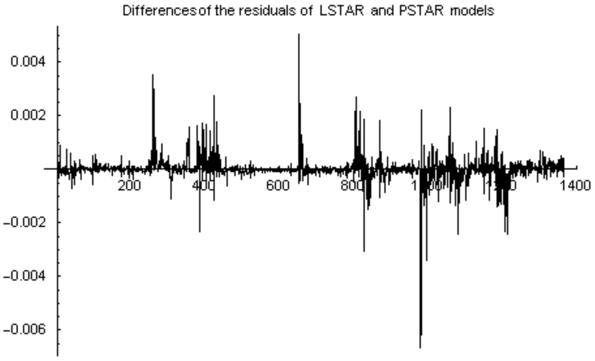


Figure 7: Differences of the residuals of LSTAR and PSTAR models.

their residual variances are small. The best fit is obtained for the model of the class PSTAR(2,2). A surprising phenomenon that higher order models do not provide lower values of estimates of residual variance can be explained by the fact that higher order models deal with smaller number of residulas. Note that the scale for differences of the residulas of optimal PSTAR(2, 2) and LSTAR(2, 2) is smaller by almost three orders than the ones for these residuals (see Figures 6 and 7).

References

[1] Bognar T., Komornik J., Komornikova M. (2004). *Regime-switching models of time series with cubic spline transition function in geodetic application.* Kybernetika **40** (1), 143 – 150.

[2] Franses P.H., Dijk D. (2000). *Non-linear time series models in empirical finance.* Cambridge University Press.

[3] Granger C.W.J., Terasvirta T. (1993). *Modelling nonlinear economic relationships.* Oxford University Press.

[4] Terasvirta T. (1994). *Specification, estimation, and evaluation of smooth transition models.* Journal of American Statistical Association **89**, 208 – 218.

[5] Tong H. (1978). *On a threshold model.* In: C.H. Chen (ed.), Pattern recognition and Signal Processing. Amsterdam, 101 – 141.

[6] Tong H. (1990). *Non-linear time series: A dynamical systems approach.* Oxford University Press, Oxford.

Acknowledgement: The research summarized in this paper was partly supported by the Grants APVT-20-023402 and GACR 402/04/1026.

Address: T. Bognar, Faculty of Civil Engineering, Slovak University of Technology, Bratislava, Slovakia

J. Komornik, Faculty of Management, Comenius University, Bratislava, Slovakia

M. Komornikova, Faculty of Civil Engineering, Slovak University of Technology, Bratislava, Slovakia and UTIA AV CR Prague, Czech Republic

E-mail: `bognar@math.sk,jozef.komornik@fm.uniba.sk,magda@math.sk`

COMPSTAT'2004 Symposium

THE TRADE-OFF BETWEEN GENERATIVE AND DISCRIMINATIVE CLASSIFIERS

Guillaume Bouchard and Bill Triggs

Key words: Statistical computing, numerical algorithms.

COMPSTAT 2004 section: Classification.

Abstract: Given any generative classifier based on an inexact density model, we can define a discriminative counterpart that reduces its asymptotic error rate. We introduce a family of classifiers that interpolate the two approaches, thus providing a new way to compare them and giving an estimation procedure whose classification performance is well balanced between the bias of generative classifiers and the variance of discriminative ones. We show that an intermediate trade-off between the two strategies is often preferable, both theoretically and in experiments on real data.

1 Introduction

In supervised classification, inputs x and their labels y arise from an unknown joint probability $p(x, y)$. If we can approximate $p(x, y)$ using a parametric family of models $\mathcal{G} = \{p_\theta(x, y), \theta \in \Theta\}$, then a natural classifier is obtained by first estimating the class-conditional densities, then classifying each new data point to the class with highest posterior probability. This approach is called *generative* classification.

However, if the overall goal is to find the classification rule with the smallest error rate, this depends only on the conditional density $p(y|x)$. *Discriminative* methods directly model the conditional distribution, without assuming anything about the input distribution $p(x)$. Well known generative-discriminative pairs include Linear Discriminant Analysis (LDA) vs. Linear logistic regression and naive Bayes vs. Generalized Additive Models (GAM). Many authors have already studied these models e.g. [3], [4]. Under the assumption that the underlying distributions are Gaussian with equal covariances, it is known that LDA requires less data than its discriminative counterpart, linear logistic regression [2]. More generally, it is known that generative classifiers have a smaller variance than.

Conversely, the generative approach converges to the best model for the joint distribution $p(x, y)$ but the resulting conditional density is usually a biased classifier unless its $p_\theta(x)$ part is an accurate model for $p(x)$. In real world problems the assumed generative model is rarely exact, and asymptotically, a discriminative classifier should typically be preferred [6], [3]. The key argument is that the discriminative estimator converges to the conditional density that minimizes the negative log-likelihood classification loss against the true density $p(x, y)$ [1]. For finite sample sizes, there is a bias-variance tradeoff and it is less obvious how to choose between generative and discriminative classifiers.

In this paper, we will first consider the parameter estimation problem, focusing on the theoretical distinction between generative and discriminative classifiers. Then we propose a new technique for combining the two classifiers: the Generative-Discriminative Trade-off (GDT) estimate. It is based on a continuous class of cost functions that interpolate smoothly between the generative strategy and the discriminative one. Our method assumes a joint density based parametrization $p_\theta(x, y)$, but uses this to model the conditional density $p(x|y)$. The goal is to find the parameters that maximize classification performance on the underlying population, but we do this by defining a cost function that is intermediate between the joint and the conditional log-likelihoods and optimizing this on training and validation sets.

Given that the generative model based on maximum likelihood (ML) produces minimum variance — but possibly biased — parameter estimates, while the discriminative one gives the best asymptotic classification performance, there are good reasons for thinking that an intermediate method such as the GDT estimate should be preferred. We illustrate this on simulations and on real datasets.

2 Preliminaries

Using independent training samples $\{x_i, y_i\}, i = 1, \ldots, n$, $x_i \in \mathbb{R}^d$,and $y_i \in \{1, \ldots, K\}$ sampled from the unknown distribution $p(x, y)$, we aim to find the rule that gives the lowest error rate on new data. This is closely related to estimating the conditional probability $p(y|x)$.

For each of the K classes, the class-conditional probability $p(x|y = k)$ is modeled by a parametric model f_k with parameters θ_k. The y follows a multinomial distribution with parameters $p_1, \ldots, p_K$. The full parametrization of the joint density is $\theta = (p_1, \ldots, p_K, \theta_1, \ldots, \theta_K)$. Given θ, new data points x are classified to the group k giving the highest posterior probability

$$P_\theta(Y = k|X = x) = \frac{p_k f_k(x; \theta_k)}{\sum_{l=1}^{K} p_l f_l(x_i; \theta_l)}. \tag{1}$$

The generative and the discriminative approaches differ only in the estimation of θ.

Generative classifier. Given data $\{x_i, y_i\}$, $i = 1, \ldots, n$, a standard way to estimate the parameters of densities is the Maximum Likelihood (ML) estimate (we assume that the solution is unique):

$$\hat{\theta}_{GEN} = \arg\max_{\theta \in \Theta} \mathcal{L}_{GEN}(\theta), \qquad \mathcal{L}_{GEN}(\theta) = \textstyle\sum_{i=1}^{n} \log p_{y_i} f_{y_i}(x_i; \theta). \tag{2}$$

Discriminative classifier. Let $\mathcal{D} = \{p_\theta(y|x) = p_\theta(x,y)/\sum_z p_\theta(x,z), \theta \in \Theta\}$ be the set of conditional densities derived from the generative model. Our aim is to find the conditional density in $\mathcal{D}$ that minimizes a classification loss function on the training set. Here, we consider only the negative conditional log-likelihood $-\mathcal{L}_{DISC}$, which can be viewed as a convex approximation to the error rate:

$$\hat{\theta}_{DISC} = \arg\max_{\theta\in\Theta} \mathcal{L}_{DISC}(\theta), \qquad \mathcal{L}_{DISC}(\theta) = \sum_{i=1}^n \log \frac{p_{y_i} f_{y_i}(x_i;\theta)}{\sum_k p_k f_k(x_i;\theta)}. \tag{3}$$

The discriminative approach allows to eliminate parameters that influence only $p(x)$,not $p(y|x)$ (e.g. shared covariance matrix in Gaussian distributions), leading to logistic regression over lower dimensional parameter spaces. However, we will not use this reduction, as we need to maintain a common parametrization for the discriminative and generative cases. thus, the solution (3) of the discriminative classifier may not be unique — there may exist infinitely many parameters that give the same conditional distribution $p_\theta(x|y)$. However, the classification performance is the same for all such solutions.

Relationship. The quantity $\mathcal{L}_{DISC}$ can be expanded as follows:

$$\mathcal{L}_{DISC}(\theta) = \underbrace{\sum_{i=1}^n \log p_{y_i} f_{y_i}(x_i;\theta)}_{\mathcal{L}_{GEN}(\theta)} - \underbrace{\sum_{i=1}^n \log \sum_{k=1}^K p_k f_k(x_i;\theta)}_{\mathcal{L}_x(\theta)} \tag{4}$$

The difference between the generative and discriminative objective functions $\mathcal{L}_{GEN}$ and $\mathcal{L}_{DISC}$ is thus $\sum_{i=1}^n \sum_k \log p_\theta(x_i,k)$, the log-likelihood of the input space probability model $p_\theta(x)$. Equation (4) shows that compared to the discriminative approach, the generative strategy tends to favor parameters that give high likelihood on the training data.

3 Between generative and discriminative classifiers

To get a natural trade-off between the two approaches, we can introduce a new objective function $\mathcal{L}_\lambda$ based on a parameter $\lambda \in [0,1]$ that interpolates continuously between the discriminative and generative objective functions:

$$\begin{aligned} \mathcal{L}_\lambda(\theta;\boldsymbol{x},\boldsymbol{y}) &= \mathcal{L}_{GEN}(\theta;\boldsymbol{x},\boldsymbol{y}) - (1-\lambda)\mathcal{L}_x(\theta;\boldsymbol{x}) \qquad (5)\\ &= \lambda\mathcal{L}_{GEN}(\theta) + (1-\lambda)\mathcal{L}_{DISC}(\theta). \qquad (6)\end{aligned}$$

For $\lambda \in [0,1]$, the GDT estimate is

$$\hat{\theta}_\lambda \quad = \arg\max_{\theta\in\Theta} \mathcal{L}_\lambda(\theta). \tag{7}$$

Taking $\lambda = 0$ leads to the discriminative estimate $\hat{\theta}_{DISC}$, while $\lambda = 1$ leads to the generative one $\hat{\theta}_{GEN}$. We expect that the GDT estimates $\hat{\theta}_\lambda$ $(0 < \lambda < 1)$ will sometimes have better generalization performances than these two extremes. Even if the discriminative estimate (3) is not unique, the maximum of (7) is unique for all $\lambda \in [0, 1)$ if the ML estimate $\hat{\theta}_{GEN}$ is unique.

Computation of $\hat{\theta}_\lambda$. Since we use a differentiable classification loss, the maximization problem (7) can be solved by any gradient ascent method. The Newton algorithm converges rapidly, but requires the computation of the Hessian matrix, The Conjugate Gradient (CG) algorithm may be more suitable for large scale problems: it needs only the first derivative and it is possible to avoid the storage of the quasi-Hessian matrix which can be huge when the number of parameters is large.

For simplicity, we assume that the parameters θ_k of the different class densities are independent. Taking the derivative of (5) with respect to θ_k and π_k, we get

$$\left\{ \begin{array}{l} \frac{\partial}{\partial \theta_k} \mathcal{L}_\lambda(\theta_k) = \sum_{i=1}^n (\mathbf{I}_{\{y_i = k\}} - (1 - \lambda) \tau_{ki}) \frac{\partial \log f_k(x_i; \theta_k)}{\partial \theta_k} \\ \frac{\partial}{\partial \pi_k} \mathcal{L}_\lambda(\theta_k) = \frac{1}{\pi_k} \left(n_k - (1 - \lambda) \sum_{i=1}^n \tau_{ki} \right) \end{array} \right. \tag{8}$$

with $n_k = \sum_{i=1}^n \mathbf{I}_{\{y_i = k\}}$ and $\tau_{ki} = \frac{\pi_k f_k(x_i; \theta_k)}{\sum_{l=1}^K \pi_l f_l(x_i; \theta_l)}$. The optimal parameters are zeros of the equations (8) for $k = 1, \ldots, K$.

For a given class k, these equations are analogous to the ML equations on weighted data, although unlike ML, the weights can be negative here Each point has a weight $\mathbf{I}_{\{y_i = k\}} - (1 - \lambda) \tau_{ki}$. The examples that have most influence on the θ_k-gradient are those that belong to the class k but have a low probability to be in it (τ_{ki} is small), and conversly those that do not belong to the class k but that are assigned to it with a high probability. The influence of the assignment probabilities is controlled by the parameter λ. This remark may ultimately help us to link our approach to boosting, and similar algorithms that iteratively re-weight misclassified data. It also shows that the generative estimator (λ=1) is not affected by the classification rate of the data points.

Choice of λ. The GDT estimate contains a tuning parameter to set, which functions like the smoothing parameter in regularization methods. λ cannot be set on the basis of minimum classification loss on the training set, since by definition, $\lambda = 0$ gives the optimal θ for training set classification. Instead, λ is set to the value $\hat{\lambda}$ that minimizes the cross-validation error rate.

If the optimal $\hat{\lambda}$ is close to one, the generative classifier is preferred. This suggests that the bias in $p_\theta(x, y)$ (if any) does not affect the discrimination of the model too much. Similarly, if $\hat{\lambda}$ is close to 1, it suggests that the model $p_\theta(x, y)$ does not fit the data well, and the bias of the generative classifier is too high to provide good classification results. In this case, a more complex model — i.e. with more parameters, or less constrained — may be needed to reduce the bias. For other $\hat{\lambda}$, there is an equilibrium between the bias and the variance, meaning that the model complexity is well adapted to the amount of training data.

4 Simulations

To illustrate the behavior of the GDT method, we study its performance on two synthetic test problems. We define the true distributions of the data as follows: In the first experiment, the class conditional probabilities are gaussian with identity covariance matrix and means $m_1 = (1.25, 0, 0, 0)$ and $m_0 = (-1.25, 0, 0, 0)$. In the second case, we simulate x according to a uniform density with correlated covariates : $x^{(1)} \sim \mathcal{U}[0; 1]$ and $x^{(d)} \sim \mathcal{U}[x^{(d-1)}; 1 + x^{(d-1)}]$ with $d \in \{2, 3, 4\}$ and $x^{(i)}$ denotes the i^{th} covariate. Then $y|x$ is simulated according to a Bernoulli distribution with parameter $1/\exp(-2.5x^{(1)})$. Note that the linear logistic model is true in the two experiments.

The assumed model is a Gaussian distribution for each class with shared diagonal covariance matrices and prior probabilities equal to $\frac{1}{K}$. Hence, the model does not correspond exactly to the true density in the second experiment, but it can provide a good approximation when the differences between the variances are small.

In each case, we estimated the true error rate of the classifiers learned on training samples of size 50, 100 and 200. The results are plotted in figure 1. We used standard plug-in estimates for $\lambda = 1$ and closed form logistic regression for $\lambda = 0$. For intermediate estimates, the conjugate gradient method was used. The first row illustrates the fact that the generative classifier performs better than the other estimates, but this difference tends to decrease when the sample size increases. In the second row, the best performance is from the BDG estimate for all training set sizes, and the optimal value of λ (the one that minimizes the expected loss) decreases with n since we know that the discriminative approach becomes optimal when n tends to infinity.

5 Experiments

We tried our classification method on some of the publically available Statlog datasets. In our implementation of the GDT estimates, the parameter dimension is limited due to the size of the optimization problem (7). To make the computation feasible, we reduced the dimension of the data by computing the first four Fisher discriminant variables and using them as inputs (when the number of classes was less than 5, so that there were fewer than four discriminant directions, we computed the remaining directions by PCA on the residuals). These directions are computed using the training data and do not involve the test data.

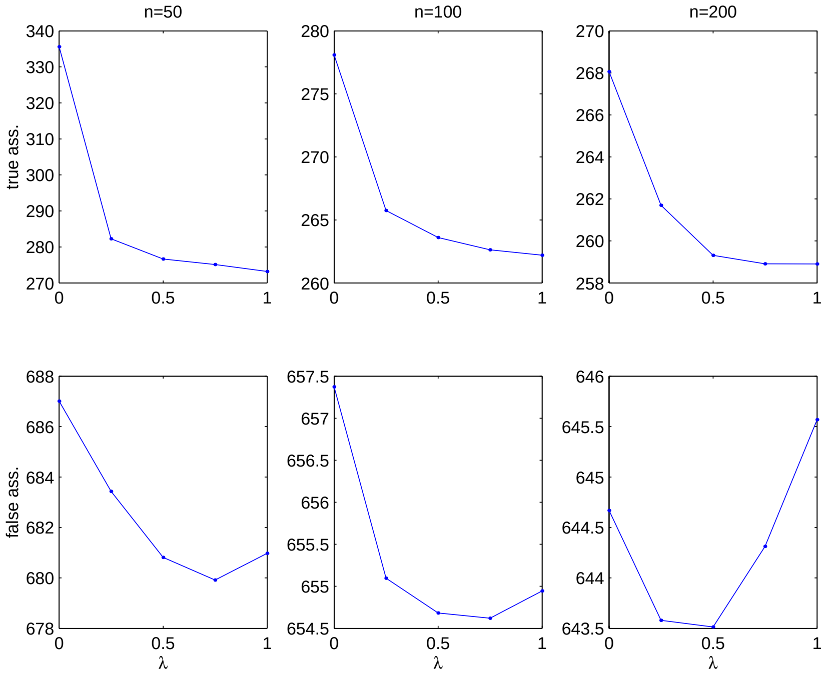


Figure 1: The full lines plot logistic loss computed on test sets of size 10^5 against the tuning parameter λ. Each plotted value is the median of 200 experiments. The rows correspond to the first and second simultations. The columns correspond to different training sample sizes.

We tried four types of density for the class-conditional distributions: 1. Gaussian densities with common covariance matrix (LDA), 2. Gaussian densities with unconstrained covariance (QDA), 3. Gaussian densities with spherical covariance matrix (Balls1), 4. Mixture of two Gaussian densities with spherical covariance matrix (Balls2). These distributions do not exactly fit the data, but they are distributions that are often used to approximate real datasets. Therefore, when the training sample is small, the generative approach may still behave better than the discriminative one. Training sample sizes were set to 50 times the number of classes so the discriminative classifiers should not have reached their asymptotic behavior.

We used a Cholesky-based parametrization of the inverse covariance matrix, so there was no need for a separate positivity constraint on the parameters. Derivatives with respect to this parametrization were obtained for each density, and we used the generative solution — which is explicit for densities 1-3 and obtained by the EM algorithm for the densities 4 — to initialize the CG algorithm.

Table 1 shows the generalization performance for each dataset and each

Dataset	australian	diabetes	heart	satimage	vehicle
Training size	100	100	100	300	200
LDA GEN	**0.143**	0.253	0.178	0.188	0.237
LDA GDT0.75	0.144	0.252	0.178	0.187	0.235
LDA GDT0.5	0.144	**0.249**	0.179	0.186	0.235
LDA GDT0.25	0.144	0.250	0.182	0.185	0.236
LDA DISC	0.145	0.249	0.185	0.191	0.243
QDA GEN	0.149	0.262	0.181	0.181	0.235
QDA GDT0.75	0.151	0.261	0.182	**0.179**	0.234
QDA GDT0.5	0.150	0.262	0.181	0.180	0.235
QDA GDT0.25	0.151	0.262	0.182	0.181	0.234
QDA DISC	0.168	0.270	0.204	0.215	0.267
Balls1 GEN	0.146	0.262	0.168	0.185	0.318
Balls1 GDT0.75	0.145	0.260	0.167	0.183	0.293
Balls1 GDT0.5	0.144	0.259	**0.165**	0.182	0.271
Balls1 GDT0.25	0.144	0.257	0.169	0.181	0.254
Balls1 DISC	0.150	0.253	0.190	0.194	0.242
Balls2 GEN	0.146	0.266	0.181	0.185	0.239
Balls2 GDT0.75	0.145	0.265	0.180	0.185	0.239
Balls2 GDT0.5	0.146	0.265	0.180	0.184	0.236
Balls2 GDT0.25	0.146	0.268	0.181	0.183	**0.232**
Balls2 DISC	0.166	0.279	0.211	0.210	0.250

Table 1: Test error rate on real datasets, averaged over 100 trials. For each trial, training data were randomly chosen and the error rate was computed on the remaining data. In the *heart* dataset, a misclassified heart disease has a cost of 5 instead of 1.

model with different values of λ. These results show that substantial improvements in the classification rate can be obtained for intermediate values of λ. However, they do not directly show the performance of the GDT estimate because we fixed λ rather than selecting it by cross-validation on the training set. The evaluation of the cost as a function of λ could be used as a model selection criterion. For example, on the vehicle dataset, the simple Gaussian model (Balls1) gives an optimal λ equal to 0. This suggests that the bias is dominating the error, and indeed the results are improved by using two Gaussian densities for each class (Balls2).

One can object that the gain in error rate in these experiments is not sufficient to really conclude the usefulness of the GDT estimator.

6 Conclusion

In this study, the relationship between generative and discriminative classifiers has been clarified: they correspond to two different maximizations in the parameter space. By interpolating linearly between the two objective functions, we introduced the GDT estimator. This can be seen either as a less biased variant version of the discriminative solution, or as an improvement of the generative classifier. The regularization is "natural" in the sense that the parameters are encouraged to fit the inputs. Our preliminary results on real data showed that the intermediate model often gives better classification performances than the discriminative and generative classifiers.

The real interest of the GDT estimate resides in its application to generative models. Probabilistic models already exist in many areas: time series models, mixed models and graphical models — including Markov Random Fields and Hidden Markov Models — are examples of widely used generative models. When class-conditional probabilities are modelled generatively, then the GDT estimator should often improve the classification performances.

Currently, the main difficulty with the GDT method is the choice of the tuning parameter, as this requires an expensive cross-validation computation. We believe that more computationally efficient criteria can be developed by analyzing the solutions on the training set, in the spirit of the Bayesian Information Criterion [5].

References

[1] Devroye L., Györfi L., Lugosi L. (1997). *A probabilistic theory of pattern recognition.* New York: Springer-Verlag, 270 – 276.

[2] Efron. B. (1975). *The efficiency of logistic regression compared to normal discriminant analysis.* Journ. of the Amer. Statist. Assoc. **70**, 892 – 898.

[3] Ng A.Y., Jordan M.I. (2002). *On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes.* In T. Dietterich, S. Becker and Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems **14**, Cambridge, MA: MIT Press, 609 – 616.

[4] Rubinstein Y.D., Hastie T. (1997). *Discriminative vs. informative learning.* In Proc. of the Third International Conference on Knowledge and Data Mining, AAAI Press, 49 – 53.

[5] Schwarz G. (1978). *Estimating the dimension of a model.* Annals of Statistics **6**, 461 – 464.

[6] Vapnick V.N. (1998). *Statistical learning theory.* John Wiley & Sons.

Acknowledgement: We thank G. Celeux, for valuable discussions. This work was supported by the European LAVA project.

Address: G. Bouchard, B. Triggs, IS2 project, INRIA, 38334 Saint-Ismier Cedex, France

E-mail: `Guillaume.Bouchard@inria.fr`

COMPSTAT'2004 Symposium

PRINCIPAL COMPONENTS ANALYSIS IN THE FREQUENCY DOMAIN

Alain Boudou, Olivier Caumont and Sylvie Viguier-Pla

Key words: Principal components analysis, time series, stationarity, random measure, spectral analysis, applications.

COMPSTAT 2004 section: Functional data analysis.

Abstract: This paper shows how to overcome some problems arising when summarizing a p-dimensional time series by a q-dimensional one. An application on meteorological data is given.

1 Introduction

David Brillinger [5] proposes a method in order to summarize a $p-$dimensional times series $(X_n)_{n\in\mathbb{Z}}$ by a $q-$dimensional times series $(X'_n)_{n\in\mathbb{Z}}$, $q < p$, where $X'_n = \sum_m C'_m X_{n-m}$.

This method is performed with a principal components analysis (PCA) of each spectral component, so it combines harmonic analysis and PCA However, when the spectrum is continuous, it cannot be put into practice because it would need the diagonalization of an infinity of matrices, and so we cannot compute the coefficients $\ldots, C'_{-1}, C'_0, C'_1, \ldots$.

We get round this difficulty by using a discretization of the spectrum and we substitute for this analysis an analysis which requires the diagonalization of a finite set of matrices. We show that under the assumption that the mesh spacing of the discretization tends to 0 the quality of the approximate solution tends to the quality of the summary $(X'_n)_{n\in\mathbb{Z}}$.

2 The PCA in the frequency domain

A series $(X_n)_{n\in\mathbb{Z}}$ (in the following, a series $(X_n)_{n\in\mathbb{Z}}$ will be noted $(X_n)_n$) composed of $p-$dimensional random vectors X_n is said to be stationary when $\mathbb{E}(X_n\, {}^t\overline{X_m}) = \mathbb{E}(X_{n-m}\, {}^t\overline{X_0})$ for all pair (n, m) of elements of $\mathbb{Z}$.

A second q-dimensional stationary time series $(X'_n)_n$ is said to be stationarily correlated with $(X_n)_n$ when $\mathbb{E}(X_n\, {}^t\overline{X'_m}) = \mathbb{E}(X_{n-m}\, {}^t\overline{X'_0})$ for all pair (n, m) of elements of $\mathbb{Z} \times \mathbb{Z}$.

If $P_{H_{X'}}$ designates the orthogonal projector from $L^2_{\mathbb{C}^p}(\Omega, \mathcal{A}, P)$ onto $H_{X'}=$ $\overline{\text{vect}}\{K\, X'_n;\ n \in \mathbb{Z},\ K\ p \times q \text{ matrix}\}$, we can assert that the $p-$dimensional series $(P_{H_{X'}} X_n)_n$ is a filtered of $(X'_n)_n$. From this fact, it is stationary and stationarily correlated with $(X_n)_n$. The stationary properties allow us to deduce that, for all integer n, $\|X_n - P_{H_{X'}} X_n\| = \|X_0 - P_{H_{X'}} X_0\|$.

We can then consider this last quantity in order to measure, when $q < p$, the quality of the summary $(X'_n)_n$ of $(X_n)_n$. We will note it $r((X'_n)_n)$. Of course, among all the possible $q-$dimensional summaries which are stationary and stationarily correlated with $(X_n)_n$, we will choose the one which is, from this point of view, the most powerful. So it follows the

Definition. *We call principal components analysis of order q in the frequency domain, of the $p-$dimensional stationary series $(X_n)_n$, the research of the $q-$dimensional stationary series $(X'_n)_n$, stationarily correlated with $(X_n)_n$, such that $\|X_0 - P_{H_{X'}}X_0\|$ is the smallest possible.*

3 Mathematical context

We will briefly develop the mathematical context for the resolution of the problem we have just exposed. A more extensive and rigourous mathematical discussion may be found in [2] and [3].

A $p-$dimensional random measure ($p-$r.m.) Z is a vector measure defined on $\mathcal{B}$, Borel $\sigma-$field of $[-\pi,\,\pi[$, with values in $L^2_{\mathbb{C}^p}(\mathcal{A})$, such that $\mathbb{E}(Z\,A^t\overline{Z\,B}) = 0$ for all pair (A,B) of disjoint elements of $\mathcal{B}$.

We easily verify that the applications $M_Z\,:\,A\in\mathcal{B}\mapsto\mathbb{E}(ZA\,{}^t\overline{ZA})\in HS(p,p)$ and $t_Z\,:\,A\in\mathcal{B}\mapsto \text{trace}\,\mathbb{E}(ZA\,{}^t\overline{ZA})\in\mathbb{R}^+$ are measures, $HS(p,q)$ being the vector space of $q\times p$ matrices which, with the inner product $< K_1, K_2 >= \text{trace}\,K_1^t\overline{K_2}$, has a $\mathbb{C}-$Hilbert space structure.

When a $\sigma-$finite measure η dominates t_Z, we can establish the existence of an application $\frac{dM_Z}{d\eta}$ from $[-\pi,\,\pi[$ into $HS(p,p)$, measurable, whose norm is $\eta-$integrable, and for which $M_Z(A) = \int 1_A\frac{dM_Z}{d\eta}\,d\eta$, for all A of $\mathcal{B}$.

A relation of equivalence, related to M_Z, may be defined on a subspace of $HS(p,p)^{[-\pi,\,\pi[}$ and we will denote by $(p,q)-L^2(M_Z)$ the set of classes of equivalence. We will use the same notation for an element of $(p,q)-L^2(M_Z)$, that is a class of equivalence, and one of its representatives, that is an application from $[-\pi,\,\pi[$ into $HS(p,q)$. The application $(\varphi,\psi)\in((p,q)-L^2(M_Z))^2\mapsto\int\text{trace}\,\{\varphi(\lambda)\frac{dM_Z}{d\eta}(\lambda)\,{}^t\overline{\psi(\lambda)}\}\,d\eta(\lambda)\in\mathbb{C}$ is an inner product which gives to $(p,q)-L^2(M_Z)$ a $\mathbb{C}-$Hilbert space structure.

The stochastic integral, relatively to Z, may be defined as the unique isometry from $(p,q)-L^2(M_Z)$ onto $\overline{\text{vect}}\,\{K\,Z(A)\,;\,A\in\mathcal{B},\ K\,q\times p\text{ matrix}\}$ which with 1_AK associates $K\,Z(A) = \int 1_AK\,dZ$ for all A of $\mathcal{B}$ and for all $q\times p$ matrix K.

The $p-$dimensional series $(\int e^{i.n}I_{\mathbb{C}^p}\,dZ)_n$ is stationary. Conversely, with each $p-$dimensional stationary series $(X_n)_n$ we can associate a $p-$r.m., and only one, Z, such that $X_n = \int e^{i.n}I_{\mathbb{C}^p}\,dZ$ for all integer n.

The $q-$dimensional series $(\int e^{i.n}\varphi(.)\,dZ)_n$ is the image of $(X_n)_n$ by the filter φ, element of $(p,q)-L^2(M_Z)$. It is stationary and stationarily correlated with $(X_n)_n$, its associated q-r.m. is $Z_\varphi\,:\,A\in\mathcal{B}\mapsto\int 1_A\varphi\,dZ\in L^2_{\mathbb{C}^q}(\mathcal{A})$.

Under certain conditions, the image by a filter is a moving average :
PROPRIETY 1. *Let Z be a $p-$r.m. for which μ, Lebesgue measure, dominates t_Z and such that $\frac{dM_Z}{d\mu}(.)$ is bounded. If φ is an application from $[-\pi,\,\pi[$ into $HS(p,q)$ measurable and bounded, then the family $\{C_mX_{n-m}\,;\,m\in\mathbb{Z}\}$, where $C_m = (2\pi)^{-1}(\int e^{i\,\cdot\,m}\varphi(.)\,d\mu(.))$, is summable of sum $\int e^{i.n}\varphi\,dZ$.*

4 Solution of the PCA problem

If we denote by $\sum_{j=1}^{p} \mu_j(.)A_j(.)^t\overline{A_j(.)}$ a measurable Schmidt decomposition of $\frac{dM_Z}{d\eta}$ and by F_j the column matrix associated with the j^{th} element of the canonical basis of $\mathbb{C}^q$:

PROPRIETY 2. *The image of $(X_n)_n$ by the filter $\alpha(.) = \sum_{j=1}^{q} F_j^t\overline{A_j(.)}$ is the stationary $q-$dimensional series, stationarily correlated with $(X_n)_n$, solution of the PCA problem of order q of $(X_n)_n$ in the frequency domain.*

We can prove that the series $(X''_n)_n$, where $(X''_n)_n = (P_{H_{X'}}X_n)_n$, is the image of $(X'_n)_n$ by the filter ${}^t\overline{\alpha(.)}$. It is also the image of $(X_n)_n$ by the filter ${}^t\overline{\alpha(.)}\alpha(.)$. This series may be named, a “series reconstitution of data”.

In the particular case where Z is concentrated on a finite set $\{\lambda_1, \ldots, \lambda_k\}$ of elements of $[-\pi, \pi[$, it may be written $Z = \sum_{l=1}^{k} \delta_{\lambda_l}(.)Z_l$, where δ_l is the Dirac measure concentrated on λ_l and $\{Z_1, \ldots, Z_k\}$ a family of elements of $L^2_{\mathbb{C}^p}(\mathcal{A})$ such that $\mathbb{E}(Z_j\,{}^t\overline{Z_{j'}}) = 0$ when $j \neq j'$.

For all φ of $(p,q) - L^2(M_Z)$, it comes $\int \varphi\, dZ = \sum_{l=1}^{k} \varphi(\lambda_l)Z_l$. So the series $(\int e^{i.n} I_{\mathbb{C}^p}\, dZ)_n$ is the sum of stationary uncorrelated series $(e^{i\lambda_l n}Z_l)_n$.

If we denote by λ_{jl} (resp. A_{jl}) the j^{th} eigenvalue (resp. the j^{th} unit eigenvector) of $\|Z_l\|^2\mathbb{E}(Z_l^t\overline{Z}_l)$, a measurable Schmidt decomposition of $\frac{dM_Z}{dt_Z}$ is $\sum_{l=1}^{k} 1_{\{\lambda_l\}} \sum_{j=1}^{p} \lambda_{jl}A_{jl}{}^t\overline{A}_{jl}$.

The series $(X'_n)_n$ and $(X''_n)_n$ corresponding to the PCA of order q in the frequency domain of $(\sum_{l=1}^{k} e^{i\lambda_l n}Z_l)_n$ are respectively
$(\sum_{l=1}^{k} e^{i\lambda_l n} \sum_{j=1}^{q} F_j^t\overline{A}_{jl}Z_l)_n$, image of $(X_n)_n$ by the filter
$\alpha(.) = \sum_{l=1}^{k} 1_{\{\lambda_l\}}(.) \sum_{j=1}^{q} F_j^t\overline{A_{jl}}$, and $(\sum_{l=1}^{k} e^{i\lambda_l n} \sum_{j=1}^{q} A_{jl}\,{}^t\overline{A}_{jl}Z_l)_n$, image of $(X'_n)_n$ by ${}^t\overline{\alpha(.)}$ or equivalently of $(X_n)_n$ by ${}^t\overline{\alpha(.)}\alpha(.)$. This analysis is equivalent to the PCA of each random vector Z_l, that is why we call it PCA in the frequency domain.

5 The particular case of absolutely summable autocovariance function

We propose to perform the PCA in the frequency domain of a p-dimensional stationary time series $(X_n)_n$ of associated p-r.m. Z and of autocovariance function absolutely summable : $\sum_n \|\mathbb{E}(X_n\,{}^t\overline{X_0})\| < +\infty$. This assumption implies the fact that the Lebesgue measure μ dominates t_Z, and consequently that a spectral density $\frac{dM_Z}{d\mu}$ exists (and is equal to $(2\,\pi)^{-1}\sum_n e^{-i.n}$ $\mathbb{E}\,X_n^t\overline{X_0}$). From this, the PCA in the frequency domain, which might be performed by the diagonalization of $\frac{dM_Z}{d\mu}(\lambda)$, for all λ of $[-\pi,\, \pi[$, cannot be obtained in practice. We get round this difficulty by using a discretization of the spectrum.

More precisely, k designating an integer parameter which will tend to infinity, we consider the measurable application from $[-\pi,\, \pi[$ into itself :

$f_k = \sum_{l=-k}^{k-1} \frac{\pi\, l}{k}\, 1_{B_{lk}}$, where $B_{-k,k} = \{-\pi\}$, $B_{lk} = [\frac{\pi\, l}{k} - \frac{\pi}{k}, \frac{\pi\, l}{k}[$ for $l = -k+1, \ldots, -1$, $B_{0k} =]-\frac{\pi}{k}, \frac{\pi}{k}[$, and $B_{lk} = [\frac{\pi\, l}{k}, \frac{\pi\, l}{k} + \frac{\pi}{k}[$ for $l = 1, \ldots, k-1$.

The application $Z_k \ : A \in \mathcal{B} \mapsto Z f_k^{-1} A \in L^2_{\mathbb{C}^p}(\mathcal{A})$ is the $p-$m.a. image of Z by f_k and is equal to $\sum_{l=-k+1}^{k-1} \delta_{\frac{\pi\, l}{k}}(.)\, Z\, B_{lk}$. The PCA in the frequency domain of this last series needs only the diagonalization of the matrices $\mathbb{E}\,(Z\, B_{lk}\, {}^t\overline{B_{lk}})$, $l = -k+1, \ldots, k-1$.

If A_{jlk} denotes the j^{th} unit eigenvector of $\mathbb{E}\,(Z\, B_{lk}\, {}^t\overline{B_{lk}})$, from § 4, the filter, element of $(p,q) - L^2(M_{Z_k})$ that corresponds to the PCA of order q in the frequency domain of $\left(\sum_{l=-k+1}^{k-1} e^{i\frac{\pi\, l}{k} n}\, Z\, B_{lk}\right)_n$ is

$\alpha_k(.) = \sum_{l=-k+1}^{k-1} 1_{\{\frac{\pi\, l}{k}\}}(.) \sum_{j=1}^{q} F_j\, {}^t\overline{A_{jlk}}$.

Let $(X'_n)_n$ denote the $q-$dimensional solution of the problem of PCA of order q in the frequency domain of $(X_n)_n$. This series is of the form $X'_n = \sum_m C'_m X_{n-m}$. Let $(X'^k_n)_n$ be the image of $(X_n)_n$ by the filter $\alpha_k(f_k(.))$. This last series, also stationary and stationarily correlated with $(X_n)_n$, may be considered as a $q-$dimensional summary. If we suppose that, for all λ of $[-\pi, \pi[$, the p eigenvalues of $\frac{d\, M_Z}{d\, \mu}(\lambda)$ are distinct, then we can (cf. [1]) establish that $\lim_{k\to\infty} r((X'^k_n)_n) = r((X'_n)_n)$, what legitimates the discretization.

Of course, as a "series reconstitution of data", one can choose $(X''^k_n)_n$ which is either the image of $(X'^k_n)_n$ by ${}^t\overline{\alpha_k(f_k(.))}$ or the image of $(X_n)_n$ by ${}^t\overline{\alpha_k(f_k(.))}\alpha_k(f_k(.))$.

For all m of $\mathbb{Z}$, we define $C'_{m,k}$, $C''_{m,k}$ and $D_{m,k}$ such that $2\pi\, C'_{m,k} = \int e^{i\,.m} \alpha_k \circ f_k\, d\,\mu = \sum_{l=-k+1}^{k-1} (\int_{B_{lk}} e^{i\,.m}\, d\,\mu) \sum_{j=1}^{q} F_j\, {}^t\overline{A_{jlk}} = 2\pi\, {}^t\overline{C''_{m,k}}$ and $D_{m,k} = (2\pi)^{-1} \sum_{l=-k+1}^{k-1} (\int_{B_{lk}} e^{i\,.m}\, d\mu)\ \sum_{j=1}^{q} A_{jlk}\, {}^t\overline{A_{jlk}}$.

From § 3, we can affirm that
$X'^k_n = \sum_m C'_{m,k} X_{n-m}$ and $X''^k_n = \sum_m C''_{m,k} X'^k_{n-m} = \sum_m D_{m,k} X_{n-m}$.

Then it is obvious that the various "coefficients" $C'_{m,k}$, which can be computed as soon as we know the vectors A_{jlk}, i.e. the matrices $\mathbb{E}\,(Z\, B_{jl}\, {}^t\overline{Z\, B_{jl}})$, allow us to obtain the series (X'^k_n) which is a $q-$dimensional summary all the more powerful as k is large.

If the series $(X_n)_n$ is strictly stationary, in [1] we prove that, almost surely:

$$\lim_m \frac{1}{2\,\pi\, m} \sum_{u=1}^{m} \sum_{v=1}^{m} (\int_{B_{lk}} e^{i\,.(u-v)}\, d\mu) X_v\, {}^t\overline{X_u} = \mathbb{E}(Z\, B_{jl}\, {}^t\overline{Z\, B_{jl}}).$$

Thus, the matrices $\mathbb{E}(Z\, B_{jl}\, {}^t\overline{Z\, B_{jl}})$ can be estimated by
$\frac{1}{2\,\pi\, m} \sum_{u=1}^{m} \sum_{v=1}^{m} (\int_{B_{lk}} e^{i\,.(u-v)}\, d\mu) X_v\, {}^t\overline{X_u}$.
The data being real, we establish $\mathbb{E}(Z\, B_{j,-l}\, {}^t\overline{Z\, B_{j,-l}}) = {}^t\overline{\mathbb{E}(Z\, B_{jl}\, {}^t\overline{Z\, B_{jl}})}$, what allows us to save computing time.

6 A worked example

6.1 The meteorological data of the application

We consider the average annual temperatures in 6 French towns, that is Aix, Biarritz, Toulouse, Bourges, Lille and Lyon, from 1950 to 2000. Then the random vector X_n has values in $\mathbb{C}^6$, and we have $m = 51$ observations. The data are preliminarily centered.

6.2 Convergence when k grows

The results of the PCA in the frequency domain depend on the discretization parameter k of the spectrum $[-\pi, \pi[$. The larger k is, the more the error of reconstitution (square root of the mean square errors) is low (figure 1).

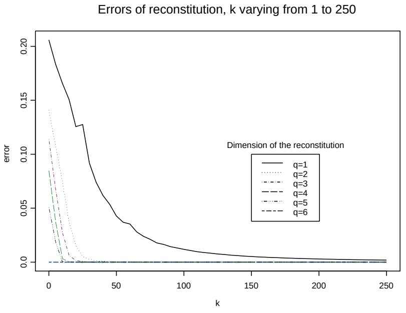


Figure 1. Errors of reconstitution

Let us notice that the usual PCA corresponds to $k = 1$. Hence, it is clear that the PCA in the frequency domain gives best results.

The coefficients $C'_{r,k}$, $C''_{r,k}$ and $D_{r,k}$ seem to converge when k grows (to our knowledge, that is not proved mathematically). For example, for $q = 1$, we obtain

$$C'_{0,5} = \begin{pmatrix} 0.3577 \\ 0.3787 \\ 0.3986 \\ 0.4578 \\ 0.3657 \\ 0.4371 \end{pmatrix} \quad C'_{0,100} = \begin{pmatrix} 0.2809 \\ 0.2986 \\ 0.3001 \\ 0.3888 \\ 0.3171 \\ 0.3956 \end{pmatrix} \quad C'_{0,200} = \begin{pmatrix} 0.2774 \\ 0.2990 \\ 0.2983 \\ 0.3849 \\ 0.3080 \\ 0.3935 \end{pmatrix} \quad C'_{0,300} = \begin{pmatrix} 0.2773 \\ 0.2963 \\ 0.2954 \\ 0.3833 \\ 0.3066 \\ 0.3932 \end{pmatrix} \quad C'_{0,400} = \begin{pmatrix} 0.2767 \\ 0.2955 \\ 0.2947 \\ 0.3828 \\ 0.3064 \\ 0.3931 \end{pmatrix}$$

In figure 2, we see the values of the $C'_{r,k}$ norms, at a stage where convergence can be considered as reached (k=400), whereas figure 3 shows how, for a reconstitution in dimension 1, these norms vary with k.

In [4], we establish that $q = \sum_{n \in \mathbb{Z}^*} \|C'_n\|^2 + \|C_0\|^2$. Hence, from the figure 2, the quantity $\sum_{n \in \mathbb{Z}^*} \|C'_n\|^2$ is not negligible. We also prove that $\sum_{n \in \mathbb{Z}^*} \|C'_n\|^2 = (2\,\pi)^{-1} \int \|\alpha(\lambda) - C'_0\|^2 d\,\mu(\lambda)$. Hence, when $\sum_{n \in \mathbb{Z}^*} \|C'_n\|^2$

is small, $\alpha(\lambda)$ is near C'_0 (for all λ of $[-\pi,\, \pi[$). That means that the various spectral components have very close systems of principal axes. We also prove that, when $(X_n)_n$ is a white noise, $\int \|\alpha(\lambda) - C'_0\|^2 d\,\mu(\lambda) = 0$.

We do not need all of the coefficients $C'_{n,k}$, $C''_{n,k}$ and $D_{n,k}$ for the reconstitution, because they are small when $|n| > 2$ (see figures 2 and 3). Note that they all can be computed. We can prove the equality $\lim_n C'_n = 0$.

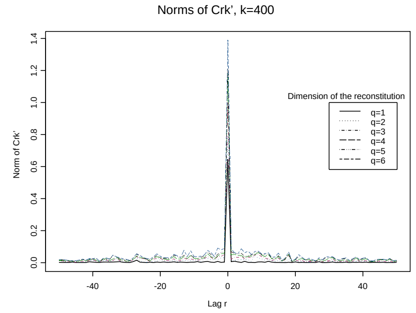


Figure 2. Norms of the $C'_{r,k}$'s for $k = 400$

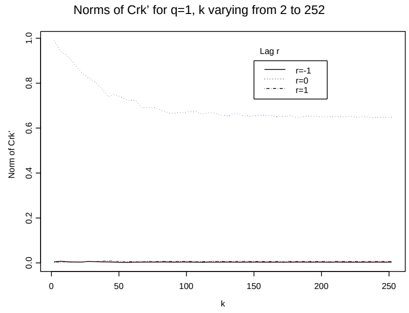


Figure 3. Norms of $C'_{r,k}$'s, k varying

6.3 Reconstitution of data from principal components

When the PCA is used as method of voluminous data storage, one uses, for reconstituting data, the principal components and the reconstitution formula, which is a filter from $\mathbb{C}^q$ into $\mathbb{C}^p$.

From the figure 1, when $k > 50$, the PCA of order 1 in the frequency domain is reaching as good results, from the quality of the reconstitution point of view, as the 5 first steps of the usual PCA. It is the reason why we will compare the quantities of data it is necessary to store in order to "reconstitute" a time series of m observations, according to whether one or the other of the two methods is used.

In the first case, the transformation of the unidimensional series $(X'_{n,k})_n$ into a series reconstituted of dimension 6 $(X''_{n,k})_n$ is made, formally, by the way of a moving average $X''_{n,k} = \sum_l C''_{n-l,k} X'_{l,k}$.

Insofar as we wish to obtain $(X''_{n,k})_{n=1,2,\ldots,m}$, only from the time series $(X'_{n,k})_{n=1,2,\ldots,m}$, the previous formula is approximated by
$X''_{n,k} = \sum_{l=1}^m C''_{n-l,k} X'_{l,k}$. That requires to know $C''_{l,k}$ for $l = 1-m$ to $m-1$.

Taking into account that, for an integer R large enough, when $|l| > R$, $C'_{l,k}$ is very small, the interesting matrices are only $2R+1$: $C''_{-R,k}, \ldots, C''_{R,k}$ (of course, if $m-1 > R$). So this requires the storage of $6(2R+1)$ data. Broadly speaking, adding the 6 data of the mean vector and the m data of $X'_{n,k}$, it is necessary to store $m + 12(R+1)$.

In the usual PCA case, it is necessary to know m values of a 5−dimensional vector, that is $5m$ data, plus the matrix that allows the reconstitution in the dimension 6, that is a matrix of 5×6 data, and the mean vector of

6 data. That adds up $5m + 36$ data. The ratio between these two quantities is $\frac{5m+36}{m+12(R+1)}$, which is all the larger as m is large.

This means that the PCA in the frequency domain is more powerful than the usual PCA, from the data storage point of view.

Let us note that, with the same way of computation, for any number of variables p, q denoting the number of steps selected in the PCA, m still denoting the number of time observations, the compression ratio is $\frac{2Rpq+p+mq}{mp}$. For $q = 1$ (it is often enough to restitute more than 97% of the inertia), this ratio becomes $\frac{(2R+1)p+m}{mp}$. For $R = 5$, this ratio in our example is 38.2%.

Figure 4 indicates the mean square error induced by the reconstitution from the principal components $(X'_{n,k})_{n=1,\dots,m}$. The minimal error is 0.038, for R to its maximum, that is 50. We notice that the error seems to be independent of q. That is because the error of reconstitution with principal components is always larger than the error due to the projection.

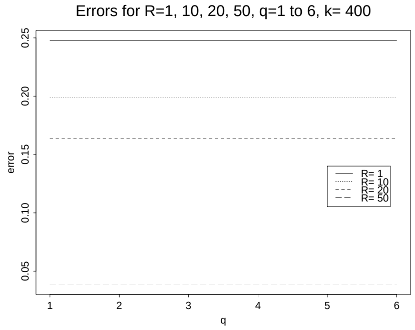


Figure 4. Errors of reconstitution for $R = 1$, 10, 20 et 50 and $k = 400$

We must notice that the more m is large, better the compression ratio is, for a constant error of reconstitution.

6.4 Elements of interpretation

From the fact that, for the PCA of order 1, ${}^tC'_0 = (2\pi)^{-1} \int A_1(\lambda) d\mu(\lambda)$, this matrix $p \times 1$ may be termed "mean of the first latent vectors $A_1(\lambda)$", and can be interpreted as a first latent vector. The other steps of the PCA may be interpreted in a same way.

For this example, the reconstitution ratio of the first principal axis is 97.8% for a total inertia of 2.611. When considering $R = 10$, that is 21 coefficients $C'_{r,k}$ instead of 101, the reconstitution ratio becomes 82.3%.

This allows us to focus on the first step of the PCA for the interpretation.

We can see on figure 5 that this first axis is linked to the mean annual temperature in all the 6 towns. On the left hand of the scatter plot are the colder years, as on the right the warmer ones.

This interpretation is copied from a usual interpretation from a usual PCA. Let us recall that the usual PCA is a particular case ($k = 1$) of the PCA in the frequency domain.

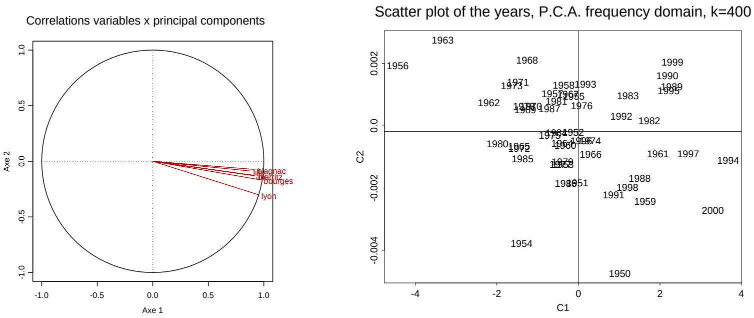


Figure 5. Correlations variables x principal components and scatter plot for principal components 1 and 2

For $k > 1$, instead of reconstituting a data at time t by a linear combination of data at the same t, this reconstitution is obtained by a moving average of linear combinations of data at times $t + n$, $n \in \mathbb{Z}$. This is logically more adequate for times series not necessarily independent in time. When $k = 2$, the reconstitution ratio is slightly higher than $k = 1$ one (90.73% instead of 90.26%), and when $k = 200$, this ratio reaches 99.70%.

References

[1] Boudou A. (1995). *Mise en œuvre de l'analyse en composantes principales d'une série stationnaire multidimensionnelle.* Pub. Inst. Stat. Univ. Paris, XXXIX, fasc.1, 89 − 104.

[2] Boudou A., Dauxois J. (1988). *Analyses de séries multidimensionnelles stationnaires.* Pub. du Labo. de Stat. et Proba., n^0 02 − 88.

[3] Boudou A., Dauxois J. (1994). *Principal Component Analysis for a Stationary Random Function Defined on a Locally Compact Group.* J. Multivariate Anal., **51**, n° 1, 1 − 16.

[4] Boudou A., Viguier-Pla S. (2003). *Sur la proximité entre l'A.C.P. dans le domaine des fréquences et l'A.C.P. classique.* Pub. du Labo. de Stat. et Proba., n^0 12 − 03.

[5] Brillinger D.M. (2001). *Time Series Data Analysis and Theory.* Society for Industrial Applied Mathematics, Philadelphia

Address: A. Boudou, S. Viguier-Pla, Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 31062 Toulouse Cedex 4, France
O. Caumont, Centre national de recherches météorologiques, 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex 1, France

E-mail: boudou@cict.fr,olivier.caumont@meteo.fr,viguier@cict.fr

COMPSTAT'2004 Symposium

FINITE SPATIAL SAMPLING DESIGN AND "QUANTIZATION"

Kamal Boukhetala and Sadoun Ait-Kaci

Key words: Spatial sampling, genetic algorithm, "quantization", localization.

COMPSTAT 2004 section: Spatial statistics.

Abstract: A Spatial Genetic Algorithm (SGA) is proposed to design a finite optimal spatial sampling. A geostatistical application is studied for a problem of optimal localization of meteorological stations. The "quantization" is used to improve the performance of the SGA.

1 Introduction

Optimal sampling of estimation of random quantities is an actual problem for practical and theoretical statistical considerations. For example, real situations where the observed quantities are accompanied by measurement errors.

Whenever errors are correlated, both of sampling problem and methodologies of global search have been studied. In scalar case, this problem has been discussed by [1], using the heuristics methods like genetic algorithms and annealing search algorithms. For a scalar finite sampling, we have confirmed in [4] the performance of the "quantization" (adequate transformation for compression of data), for the used algorithms. For a large sample size, [2] has proved that one may reduce the rate of convergence of mean quadratic error, from n^{-2} to $n^{-1,5}$.

In [3], we have generalized the previous studies in the spatial case by developing an adapted search algorithm. In our work, we propose a generalization of studies of [4], that concern spatial optimal sampling with "quantization", by developing for the spatial case, an adequate global search algorithm. The proposed algorithm is a *Spatial Genetic Algorithm*, that is a powerful heuristic to solve hard combinative problems. We describe two new crossover operators, called: global and local crossover operators. This type of crossover operators explores the spatial aspect and gives better results than [3]. On the other hand, we show that the "quantization" of spatially correlated errors, improve clearly the quality of results, for a small number of sampling points (k), but gives bad results when $k \to \infty$ which is theoretically proved by [2].

We may conclude that the "quantization" has a positive effect with the proposed *Spatial Genetic Algorithm* for a small number of sampling points, that presents a practice interest for different problems encounter in hydrological and environmental sciences.

2 Spatial genetic algorithm

Generally, Genetic Algorithms (GA) represents an enough rich and very interesting family of stochastic algorithms of optimization, that are founded on mechanisms of the natural selection and genetics. Fields of application are very varied; theory of the picture compression, automatic and others. The principle of these algorithms is to proceed by a stochastic research on an important space through "a population" of Pseudo- Solutions. These algorithms are simple and very efficient, because they are not like at hypotheses on the function to optimize, as the continuity and the differentiability. They operated directly one the explored space, after a coding of feasible solutions of the function. They processes on a population of points, instead of an unique point and use values of the studied function only without another auxiliary knowledge and rules of transition probabilistic.

In order to well to fear the genetic algorithm dynamics, let's introduce the main elements of the jargon, used in the literature.

- *Individual or Chromosome: a potential solution.*
- *Population: a set of chromosomes or of points of the space of research.*
- *Environment: the space of research;*
- *Fitness or Function of assessment: the function that one seek to maximize (minimize).*

The working of the GA is based on different basis operators. These operators are inspired directly from the natural selection mechanism and the genetic phenomenon. It consists to evolve the population in order to adapt individuals to the environment. Technically, a new generation gets himself at the end of a cycle by using of three main standard operators, namely, operators of reproduction, crossover, and mutation. For a reason of implementation adapted to computers, a chromosomal representation or coding of individuals (feasible solutions) is necessary. We propose in the case of our problem:

- A spatial coding of the solution,
- Two operators of crossover and mutations appropriated.

What permits us to propose a new Genetic Algorithm adapted to the Spatial case, denoted *SGA*.

2.1 Coding of the solution (CS)

One considers a spatial n-sample, in a space of research, of dimension $L < n$, that is a matrix of real, of the following form:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1L} \\ \vdots & & \dots & \vdots \\ \vdots & & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nL} \end{bmatrix}$$

Thus, a population of solutions in this space is a set of real matrixes $M_{(n\times L)}$. We adept coding relatively to a matrix; it is given for individual i, by a vector of real of size $n * L$:

$$X_i = ((x_{1l}, \dots x_{1L})_i, (x_{2l}, \dots x_{2L})_i, \dots, (x_{nl}, \dots x_{nL})_i) \,.$$

2.1.1 Operator of crossover (OC) One consider two coded real configurations:

$$X = (x_1, x_2, \dots, x_L) \quad \text{and} \quad Y = (y_1, y_2, \dots, y_L)$$

A classic crossover, that we call global crossover produced of the new configurations thereafter. The uniform crossover corresponds here to interchange, at random, a subset of points of IR^L. Without losing generality, we consider an example for the case of the plan in order to illustrate the process of crossover. For two parents X_p and Y_p, such that:

$$\begin{aligned} X_p : &= (x_1, x_2, \dots, x_7) = ((x_1^1, x_1^2), (x_2^1, x_2^2), \dots, (x_7^1, x_7^2)) \\ Y_p : &= (y_1, y_2, \dots, y_7) = ((y_1^1, y_1^2), (y_2^1, y_2^2), \dots, (y_7^1, y_7^2)) \end{aligned}$$

an uniform crossover products two individuals children, X_e and Y_e, with a appropriate Jong schema. We has a difficulty, when we applied this operators with a low probability for the mutation, in accordance with the theory of the AG. What gives a little changes in the diversity of the individuals population relatively to the initial population. On the other hand, the defined crossover operator has for simple role to change indication of components. A first idea was to increase the probability of mutation to maintain a diversity being sufficient, and so to preserve the improvement. The works of Jong clearly demonstrate that this method is not efficient. An uncontrolled increase of the rate of mutation can damage the performance of the genetic algorithm. Therefore, we introduced a new operator, says local crossover, that acts locally on couples of two individual components to cross.

2.1.2 Operator of local crossover (OLC) From a couple $X_1, X_2)$ of individuals, we choose a sample (x_i^1, x_i^2) . We proceed as following to define the OLC operator, by generating a new sample $(x_i^{1'}, x_i^{2'} \in [a,b] \times [a,b]$, with:

$$\begin{cases} x_i^{1'} = P_E\left(\frac{(b-a)}{\text{pas}} \times U\right) \times \text{pas} + \text{pas}/2\,; \\ x_i^{2'} = x_i^1 + x_i^2 - x_i^{1'} \end{cases}$$

for

$$\begin{aligned} a &= \max(x_i^1, x_i^2) - \min(x_i^1, x_i^2) \\ b &= \min(x_i^1, x_i^2, \mathrm{Sup}) \end{aligned}$$

The quantities, denoted by pas, P_E et U, are respectively the increment of sampling space, integer part and a random value, taken from $[0,1]$.

At the time of the research of optimal sample, it can contain two identical samples, because the combinative nature of the problem. For it, one introduces an operator of diversity that is defined as following by increasing or creasing of the gene:
we set

$$\begin{cases} x_i' = x_i & \mathrm{Sgn} \quad \mathrm{pas} \\ y_i' = y_i & \mathrm{Sgn} \quad \mathrm{pas} \end{cases} \qquad (x_i, y_i) \in [a_1, b_1] \times [c_1, d_1] \, .$$

a_1, b_1, c_1, d_1 are some constants, and Sgn is defined by:

$$\mathrm{Sgn} = \begin{cases} + \text{ if} & (x_i = a_1 \text{ or } y_i = c_1) \\ - \text{ if} & (x_i = b_1 \text{ or } y_i = d_1) \\ (+ \; rmor \; -) & \text{with probability } \frac{1}{2} \end{cases}$$

Finally, local crossover is for our problem, an adaptation very interesting of the crossover classic multi-site on chromosomes, produced by the concatenation of binary code.

2.1.3 Operator of Mutation (OM) The operator of mutation that we propose to our spatial sampling problem proceeds as following:
We operate on a gene $S_k = (x_k, y_k)$, to mute it to a new gene $S_k' = (x_k', y_k')$, as following:

$$\begin{cases} x_k' = x_k & \mathrm{Sgn} \quad U \\ y_k' = y_k & \mathrm{Sgn} \quad V \end{cases}$$

such that:

$$\begin{aligned} U &\in \begin{cases} P_E\left(\frac{U_{[x_k, b_1]}}{\mathrm{pas}}\right) \times \mathrm{pas} & \text{if} \quad x_k \geq \frac{a_1+b_1}{2} \\ P_E\left(\frac{U_{[a_1, x_k]}}{\mathrm{pas}}\right) \times \mathrm{pas} & \text{if} \quad x_k < \frac{a_1+b_1}{2} \end{cases} \\ V &\in \begin{cases} U_{[y_k, d_1]} & \text{si} \quad y_k \geq \frac{c_1+d_1}{2} \\ U_{[c_1, y_k]} & \text{si} \quad y_k < \frac{c_1+d_1}{2} \end{cases} \end{aligned}$$

$U_{[a_1,b_1]}$ is a random value from $[a_1, b_1]$.

2.2 Stopping criteria (SC)

The number of iterations is fixed, after some tests on the fitness function. It satisfies a compromise between constraints of convergence of the population, the CPU time and the precision.

Therefore we are now able to propose the following Spatial Genetic Algorithm:

```
SGA algorithm:
Begin
Stage 1 - Function of fitness
Stage 2 - Coding: CS
Stage 3 - Generation of the feasible solutions set
Stage 4 -
        Do
            $OC$, $OLC$ and $OM$
        until SC.
End
```

3 Application

The goal is the installation of a rainfall optimal network, that permits to observe a quantity M of rain data taken on a region D with a number reduces of stations k.

We have:

$$D = \{(x, y, z) \,|\, x \in [540, 800],\ y \in [150, 400],\ z \in [456, 1250]\}$$

where $s = (x, y, z)$ is a point of the rainfall site in the region D, with z is the altitude. The rainfall annual data, obtained from the domain D, permits to give the following regression model, relatively to the observed data of a site $S_i = (x_i, y_i, z_i)$ $Q_{S_i} = \langle \tilde{\alpha\alpha}, S_i \rangle + d + m \cdot \varepsilon(S_i)$ (/millimeter). $\tilde{\alpha} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$, d and m are some real parameters. The parameters a, b, c and d are estimated by:

$$\hat{a} = 60.05.10^{-3} \quad \hat{b} = 0.701 \quad \hat{c} = 0.29 \quad \hat{d} = 94.44$$

with a value of $R62 = 0.89$.

The observations errors are modeled by spatial random process ε_S, by taking in consideration the characteristic of variability, uncertainty and heterogeneity of the real rainfall process. We suppose that the spatial random process of errors associated to each site, is a spatial homogenous Gaussian process, with mean μ_ε and covariance $C(\varepsilon_i, \varepsilon_j)$. Therefore, one determine the optimal sites by the research of the optimal sample of finite size k, relatively to a optimality criteria: $E(\overline{Q} - M)^2$, with $\overline{Q}$ is the mean quantity of rainfall of the k sites $\left(\overline{Q} = \frac{1}{k} \sum_{i=1}^{k} Q_{S_i}\right)$.

3.1 Mathematical formulation of the problem

Let $S = \{s_i \in D \subset IR^3, i = 1, \ldots, k\}$, be a set of sample of sites of size $k = |S| \leq n$, where n is the size of the domain D. The optimality criteria is given by:

$$\sigma^2(S^*/\varepsilon) = \min_{S\in D}\left\{E\left(\sum_{i=1}^{k}\frac{Q(s_i)}{k} - M\right)^2\right\}$$

k: Size of S.
S: Set of sample of size k.
$Q_{S_{ii}}$: Quantity of rainfall observed on site S_i.
$s_i^* = (x_i^*, y_i^*, z_i^*)$ $i = 1, 2, \ldots, k$: coordinates of optimal spatial sample S^*,
$\varepsilon = (\varepsilon(s_i))_{i=1,\ldots,k}$: Vector of correlated observations errors in each site, with mean zero and covariance function $\mathrm{Cov}(s_i, s_j)$ of a Gaussian spatial process, of analytical type. $E(\cdot)$ is mathematical mean, defined relatively to the law of ε.

3.2 Function of fitness

We have:

$$\sigma^2(S^*;\varepsilon) = \min\left\{\frac{1}{k^2}\left[(\mathrm{Cst})^2 + m^2\cdot\sum_{i=1}^{k}\sum_{j=1}^{k}\mathrm{Cov}\left(\varepsilon(s_i)\varepsilon(s_j)\right)\right]\right\}\ldots(I)$$

such that: $\{s_i \in S : i = 1, \ldots, k\}$ and $\{s_j \in S : j = 1, \ldots, k\}$ and Cst is an appropriate determined constant.

We applied our SGA algorithm for following choice of parameters, that are fixed by the expert:

- M: mean quantity of rainfall: 337.9 mm
- k: number of rainfall stations: 5
- Selection by ranging (Pressure of selection: 1.4)
- Crossover: OC and OLC are used, with adequate probabilities P_{OC} and P_{OLC}, receptively
- Mutation: OM is used with an adequate probability P_{OM}
- Spatial Covariance $\mathrm{Cov}(s_i, s_j) = \exp(-\theta\|s_i - s_j\|^2$, with $\theta = 0.05$

The initial choice of individual has an effect on the precision of solution. A positive effect, when this choice is a bad individual. For $\sigma^2(S^*/\varepsilon) = 0,62$, the corresponding optimal spatial configuration is given by the following figure.

The following table give the obtained results of the SGA, for optimum parameters and relatively to a regular sampling, by a repetition of 80 times of the SGA.

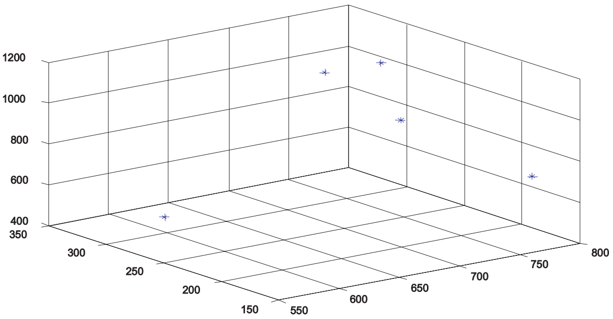


Figure 1: Optimal spatial Localization, obtained by the SGA.

P_{CG}	P_{CL}	**Absolute Error (AE):** $\|\overline{Q} - M\|$	**Confi. Interval of (AE)**
0.9	0.01	4.43, 3.98, 0.52, 1.81, 2.02, 1.4, 0.3, 1.26, 0.76, 0.11, 5.10, 2.41, 3.46, 9.07, 1.81, 0.651, 0.34, 0.41, 2.19, 0.00, 0.94, 0.75, 5.17, 2.80, 7.14, 2.79, 0.19, 1.78, 0.03, 0.19, 8.28, 0.63, 1.44, 7.59, 8.84, 1.42, 2.14, 0.58, 4.09, 2.46, 7.54, 1.42, 4.62, 2.02, 4.14, 0.39, 11.85, 0.11, 0.26, 1.11, 0.56, 1.81, 2.39, 0.84, 2.01, 0.01, 0.98, 1.58, 7.71, 3.28, 1.07, 2.16, 0.07, 0.21,0.95,4.56, 0.92, 0.03, 0.43, 1.22,1.46, 13.74, 7.81, 1.27, 5.51, 1.64, 0.90, 5.31, 1.18, 0.99.	[4.02, 9.96]

4 "Quantization"

The use of a "quantization" operator reduce the variability of the estimator of $\overline{Q}$ and improve the performance of the SGA, as it is shown by the following table:

P_{CG}	P_{CL}	**Absolute Error (AE):** $\|\overline{Q} - M\|$	**Confi. Interval of (AE)**
0.9	0.01	0.35 2.09 1.31 0.22 2.57 0.7 0.56 0.26 0.43 0.61 0.56 2.38 5.58 0.93 5.53 2.70 0.32 0.80 0.441 0.47 6.63 1.95 1.03 1.17 1.67	[2.43, 5.79]

5 Implementation

The SGA is implemented by the MATLAB language. The object procedure of this language are used in generation of Gaussian spatial measure errors and by simulation operators OC, OLC and OM.

6 Conclusion and comments

A Spatial Genetic Algorithm is proposed. The obtained results of sampling are improved by "quantization" of errors, for small size of sample. It justifies itself by the fact that the "quantization" has an effect on the regularity of function fitness, by creating a configuration of picks, desirable for an algorithm of type genetic. The perspective works will be devoted to improve parameters and operators of Spatial Genetic Algorithm, in order to decrease the size of errors and to obtain realistic confidence intervals more adapted to practical applications.

References

[1] Boukhetala K., Benhenni K., Benamara S. (1996). *Optimal sampling for estimating integral of function from observations with correlated measurement errors.* Compstat'96, Edition UPC, A. Prat and E. Ripoll, 161 − 162.

[2] Benhenni K., Stamatis Cambanis (1998). *The effect of quantization on the performance of sampling design.* IEEE transactions on information theory **44**, (5), 109 − 121.

[3] Boukhetala K., Mehassouel Nadjib (2000). *Optimal spatial sampling for an estimating problem, based on correlated observations.* Compstat'2000, edition Statistics Netherlands **1**, 157 − 158.

[4] Boukhetala K., Habib Fatma Zohra (2002). *Finite sampling design with "quantization" for a pharmacokinetic problem.* The 6th WMSCI, Computer Science **XI**, edition IIIS, USA, 167 − 170.

Address: K. Boukhetala, S. Ait-Kaci, Departement de Probabilités et Statistiques Bp, 32, El-Alia, USTHB, Bab-Ezzouar, Alger

E-mail: kboukhetala@usthb.dz

USING PRINCIPAL COMPONENTS ANALYSIS FOR DIMENSION REDUCTION PRIOR TO COMPOSITIONAL ANALYSIS

Mark John Brewer, David A. Elston, Lorna A. Dawson and Robert W. Mayes

Key words: Compositional analysis, principal components.

COMPSTAT 2004 section: Dimensional reduction.

Abstract: We investigate the use of principal components analysis (PCA) as an aid to compositional analysis. We study a data set concerning measurements of chemical markers (*n-alkanes* and *alcohols*) on artificially-created mixtures of grass and clover shoots. We go on to apply our procedure to mixture data sets based on actual Fourier Transform Infrared (FTIR) spectroscopy measurements on soil, where the very large number of markers recorded means that some form of dimension reduction is a necessity.

1 Introduction

We are concerned with data arising from research into the use of chemical markers for apportioning samples of mixtures of plant materials, faeces, soils, etc. into constituent components—i.e. *compositional analysis*. We assume that samples of the original components, here termed (*pure*) *sources* are also available, and that measurements on the chemical markers are made on both mixtures and sources. Note that it is the *problem*, not the *data set*, which is compositional, and that [2] and [3] introduce a Bayesian model for compositional analysis in these contexts.

The numbers of markers involved may be large. For example, spectroscopic methods return traces often comprising over a thousand variables. Consequently, we cannot conduct a compositional analysis without employing some form of dimension reduction. As the number of samples is likely to be relatively low, canonical variates analysis (CVA), an obvious choice where different groups occur in the data, will not be feasible (at least directly).

We first investigate the effectiveness of using PCA prior to compositional analysis on data sets with 15 markers measured on plant shoots, and compare the results with analyses on the full (non-rotated) sets of markers. We then study simulated mixture data sets based on real FTIR source measurements on three different soil types.

2 Compositional analysis

As our model has been introduced elsewhere ([2] and [3]), we provide only a brief description here. We assume there are N sources, for which there are A markers. We have direct measurements of markers for each source, denoted x_j^i for $j = 1, \dots, N$ with $i = 1, \dots, n_j$ samples from each source, and where x_j is a vector of length A. We assume multivariate normality of x_j with mean vectors μ_j and covariance matrices Σ_j, $j = 1, \dots, N$, hence

$$x_j^i \sim \mathrm{MVN}_A(\mu_j, \Sigma_j), \qquad j = 1, \dots, N, \quad i = 1, \dots, n_j. \tag{1}$$

Also, we have measurements on mixture samples, denoted similarly by z^k for $k = 1, \dots, n_z$ samples (z^k being another A-vector).

Further, we assume that each mixture sample is comprised of a weighted combination of *latent* draws from the source distributions at (1); we denote these latent quantities by y_j^k, and hence

$$y_j^k \sim \mathrm{MVN}_A(\mu_j, \Sigma_j), \qquad j = 1, \dots, N, \quad k = 1, \dots, n_z;$$

in addition, the weights are the source proportions, and hence we can model the mixture samples via

$$z^k = \sum_{j=1}^{N} p_j^k y_j^k + \epsilon^k, \quad k = 1, \dots, n_z;$$

where ϵ^k is an A-vector of measurement errors having zero mean (vector) and covariance Σ_ϵ. To model the composition proportions p_j, we use *log-ratios* [1] $q_j^k \equiv \log(p_j^k / p_N^k)$ for $j = 1, \dots, N-1$, and the overall log-ratio compositional $(N-1)$-vector q^k is also assumed multivariate normal, with

$$q^k \sim \mathrm{MVN}_{N-1}\left(\mu_q, \Sigma_q\right), \qquad k = 1, \dots, n_z,$$

for the mean vector μ_q on the log-ratio scale, and covariance Σ_q.

We complete the model definition by defining appropriate diffuse priors: multivariate normal for the mean vectors μ; and Wishart priors on the reciprocals of the covariance matrices Σ. We make inferences on the model via Markov chain Monte Carlo methods using the WinBUGS [5] package.

3 Analysis of plant shoot mixtures

3.1 Background

Chemical markers known as *n-alkanes* have been used in compositional analysis for some time now; [6] considered their use in determining diet composition using matrix methods. Another application using n-alkanes appeared in [3], where they were used to recover the proportions of artificially-mixed samples of grass and clover shoots ($N = 2$). Twelve sets of mixture samples were created at different mean/standard deviation (SD) combinations of clover proportions, shown in columns 1 and 2 of Table 1, expressed as percentages. For each of the twelve sets, 40 “target” clover proportions were generated

Percentages		Non-PCA Analysis		PCA, Mean-only Scaling			
Mean	SD	All 15	Best 3	1 PC	2 PCs	3 PCs	4 PCs
10	1	0.0200	0.0132	0.0324	0.0116	0.0125	0.0126
10	2	0.0196	0.0182	0.0292	0.0172	0.0172	0.0168
10	4	0.0206	0.0192	0.0410	0.0201	0.0201	0.0198
20	2	0.0353	0.0264	0.0335	0.0221	0.0224	0.0242
20	4	0.0280	0.0277	0.0363	0.0251	0.0203	0.0209
20	8	0.0399	0.0374	0.0557	0.0352	0.0374	0.0355
30	3	0.0364	0.0368	0.0492	0.0433	0.0351	0.0362
30	6	0.0515	0.0459	0.0477	0.0394	0.0369	0.0417
30	12	0.0693	0.0594	0.0924	0.0825	0.0746	0.0730
40	4	0.0424	0.0379	0.0448	0.0434	0.0366	0.0398
40	8	0.0411	0.0383	0.0489	0.0440	0.0379	0.0433
40	16	0.0468	0.0555	0.0810	0.0765	0.0575	0.0539

Table 1: Root mean square errors for estimated plant compositions.

to have the stated mean percentage and SD. There were $n_j = 140$ samples each of grass and clover. There were 9 n-alkanes considered in [3], and some attempt was made to determine important subsets of n-alkanes in terms of prediction accuracy measured by root mean square errors (RMSEs, recorded throughout on the *proportion* scale) of the estimates of the individual $n_z = 40$ target proportions for each mixture set.

This paper considers, in addition to the data of [3], measurements of 6 other chemical markers (*alcohols*), giving $A = 15$ markers in total. Columns 3 and 4 of Table 1 give the RMSEs for analyses using all 15 markers, and for the best subset of size 3 respectively. Note the overall superior performance of the subset relative to the full set; several other subsets have been studied (the selection informed by a discriminant analysis) which gave inferior RMSE performance, and which also included larger subsets of markers.

3.2 Dimension reduction

We consider here the use of PCA to reduce the dimensionality of the problem. We have noted how better results are obtained using only a subset of the markers, but that selection of this subset required considerable effort via discriminant analysis and compositional analyses on subjectively-chosen candidate subsets. We would hope that PCA might simplify this process, in a sense, by drawing out the important projections for apportionment. For this reason, we choose a PCA on the source samples *only*, and hence use the same projection for each set of mixture data.

A PCA was conducted on the 280 pure grass and clover samples; we used a mean-centred transform of the data, since transforming to unit variance

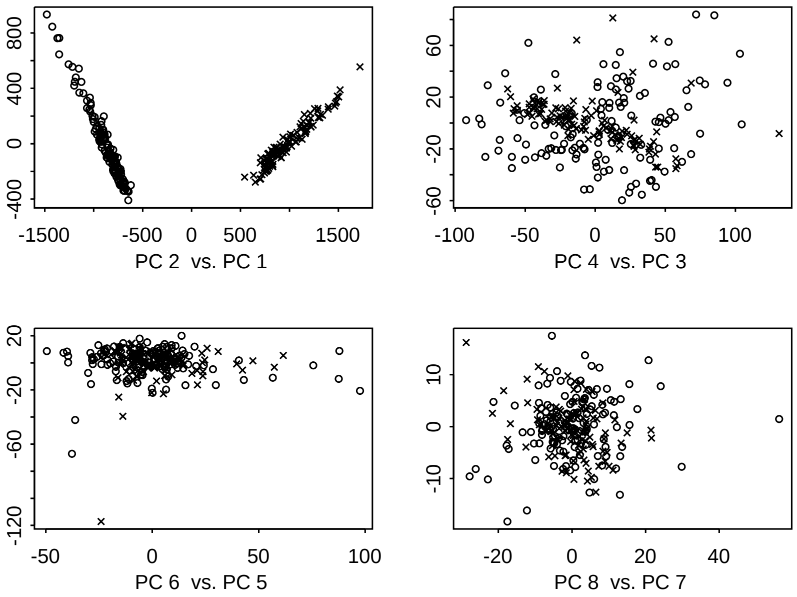


Figure 1: Plots of the first 8 PCs for the plant shoots data: grass (circles) and clover (crosses)

in addition proved less successful in terms of RMSE. The PCA suggested 94.5% of the variation was explained by the first PC, 99.7% by the first two, reaching 99.9% by the first four principal components (PCs). Figure 1 shows pairwise plots of the first 8 PCs. It is clear that the first PC separates the sources very well, the 2nd less well. The plot of PC4 vs. PC3 suggests a difference of scale in a diagonal direction, while the remaining plots show little distinction between grass and clover shoots. It is possible, therefore, that compositional analysis might benefit from the use of the first 4 PCs, while any more than this would be unnecessary. However, we evaluate this objectively below.

3.3 Compositional analysis

Following the PCA, we apply the rotation obtained from the source analysis to the mixed samples, and perform the compositional analysis on any chosen number of PCs. The final 4 columns of Table 1 contain the RMSEs for analyses using from (the first) 1 to 4 PCs. The results for 1 PC are poor; note that the first PC was almost identical to the first (and only) canonical variate (CV) here, and hence using 1 CV gave no improvement. Broadly speaking, it seems that for the mixture sets with the lowest SD, using 2 PCs provided the greatest accuracy; for the 2 sets with the largest SDs, 4 PCs appear to perform best; while for the rest, 3 PCs may be optimal. Results were not improved by using more than 4 PCs; beyond 4, the within-source variance becomes large relative to the between-source variance, and hence we are merely adding more uncertainty into our analysis.

Here, it seems that the larger the variance of the individual sample proportions, the more PCs are required to estimate the composition as best we can. Comparing the analysis of the PCs with the non-PCA analysis, we see that apart from the 30% mean, 12 SD case, use of the PCA-rotated data gives better or equivalent results for one or more numbers of PCs, and on the whole, might be described as doing better.

4 Analysis of FTIR soil data

4.1 Background

Our second example concerns the use of FTIR spectroscopy, where some form of dimension reduction is a *necessity* prior to compositional analysis. FTIR gives information on specific functional chemical groups present in mineral or organic matter by measuring the absorption pattern obtained in the infrared spectrum. The whole spectrum (fingerprint) may also contain information that relates to more general chemical or biological properties related to the influence of vegetation on soil, for example ([4]). FTIR is a rapid procedure, and as such has advantages where chemical or biological analyses are time-consuming or require fresh material.

We have FTIR readings on $N = 3$ sets of source samples of soils—these soils are broadly classified as Brown Earth, Alluvial and Cultivated Podzols. The FTIR spectroscopy returns trace plots which have been summarised by 1894 readings at particular wavelengths. For some regions of the traces, there is no meaningful information, and once these have been removed we are left with $A = 1264$ variables for analysis, with $n_j = 4$ samples of each source. We consider the situation where we might have mixtures of the three soil types, and wish to determine the relative proportions of each.

4.2 Dimension reduction

A PCA analysis of the FTIR source data suggests 61.9% of the variability is explained by the first PC, 92.6% by the first two and 96.6%, 98.4%, 99.3%, 99.8% by the first 3,4,5,6 PCs respectively. (We use the mean-centred transform, since again this provided better results.) We also consider obtaining the two CVs—on the 12 PCs, since the number of original variables is too large for sensible direct application of CVA. Figure 2 shows scatterplots of the first 8 PCs; the sources are well-separated for lower PCs, but the separation becomes less clear beyond the 6th. Thus, 6 PCs might represent the most information to be gained from the data for a compositional analysis; again, we investigate this by attempting to recover known proportions.

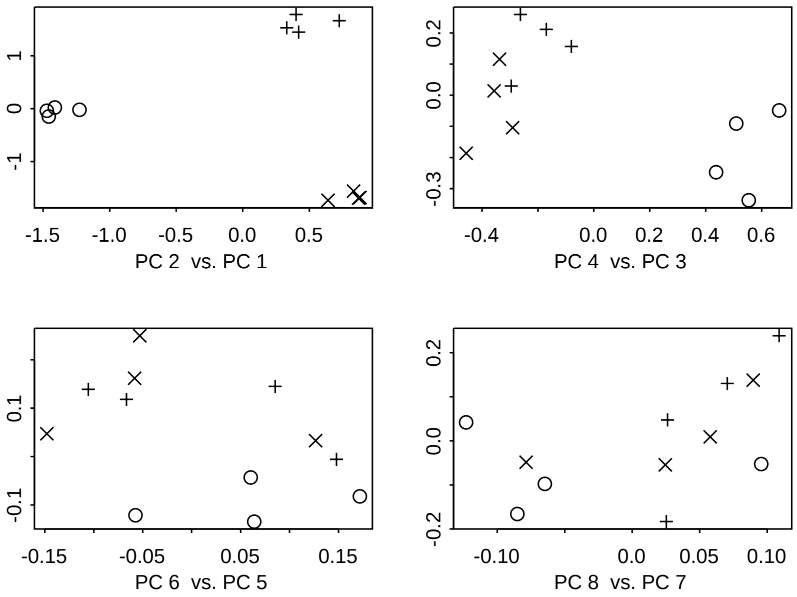


Figure 2: Plots of first 8 PCs for the FTIR soils data: Brown Earth (circles), Alluvial (addition signs), Cultivated Podzols (crosses).

Since no physical mixtures of the soils are as yet available, we obtained numerical mixtures of the source soils stochastically. We achieved this by repeatedly sampling randomly one of the 4 samples from each source, and then combining the FTIR values in the proportions indicated by the first three columns of Table 2. In fact, to obtain $n_z = 40$ mixture samples for each row of the table, we generated 40 uniform random numbers on prop $\pm\, 0.05$ (prop± 0.025 for the bottom six rows, which all include 0.05 as a proportion), and rescaled to ensure summation to 1. We also allowed for slight randomness on the FTIR values, but experiments without this produced similar results. Thus, for each set of mixture proportions, we have a sample of size 40 for analysis, and we can compare the estimated individual p^k with the proportions used in generating the mixture data via RMSEs as before.

4.3 Compositional analysis

The results from use of the first 3, 4, 5 and 6 PCs are shown in Table 2, along with those from use of the 2 CVs. Note that with 3 sources, we require at least 2 variables.

The results for 2 PCs are not shown as they were comparatively poor. For the top 7 proportion combinations in the table the best accuracy appears to occur when using 5 or 6 PCs, although the estimates themselves seem fairly stable with regard to the precise number. For the bottom 6 combinations, where the variability is less, this stability is even more evident, with good

Source Proportions							
Brown		Cultiv.					
Earth	Alluvial	Podzols	3 PCs	4 PCs	5 PCs	6 PCs	2 CVs
0.20	0.40	0.40	0.0041	0.0038	0.0034	0.0033	0.0099
0.40	0.20	0.40	0.0032	0.0029	0.0030	0.0029	0.0075
0.40	0.40	0.20	0.0044	0.0043	0.0037	0.0039	0.0114
0.10	0.10	0.80	0.0082	0.0074	0.0070	0.0071	0.0142
0.10	0.80	0.10	0.0083	0.0079	0.0075	0.0076	0.0152
0.80	0.10	0.10	0.0081	0.0076	0.0068	0.0071	0.0179
0.33	0.33	0.33	0.0036	0.0031	0.0031	0.0029	0.0093
0.05	0.30	0.65	0.0050	0.0049	0.0048	0.0047	0.0086
0.05	0.65	0.30	0.0055	0.0052	0.0052	0.0058	0.0092
0.30	0.05	0.65	0.0043	0.0041	0.0042	0.0042	0.0082
0.30	0.65	0.05	0.0053	0.0052	0.0052	0.0058	0.0103
0.65	0.05	0.30	0.0045	0.0047	0.0043	0.0044	0.0100
0.65	0.30	0.05	0.0045	0.0045	0.0043	0.0042	0.0101

Table 2: Root mean square errors for estimated soil compositions.

results being found using only 3 or 4 PCs. Comparing the RMSE values between the top 7 rows of the table, the accuracy of the estimates is lowest for the 10/10/80-split mixture sets, suggesting that apportionment is harder for cases where the proportions are more uneven.

Note that including more than 6 PCs did not produce further improvements; again, results worsened since variability along projections came to be dominated by within-source variance. Finally, the last column of Table 2 shows the RMSEs from use of the two CVs; perhaps surprisingly, the results are poor—even more so than the 2 PCs case, in fact.

5 Discussion

We have demonstrated the application of PCA as a tool for reducing the dimensionality of data sets in a compositional analysis context. For a relatively small example in terms of number of markers, we found results from analysing a PCA-rotated data set to be as good as or better than those from analysing a non-rotated data set in most cases. For a simulated mixture (from real source components) having well over a thousand markers, we found compositions accurate in all cases to less than 1%.

It was by no means a certainty that PCA would be useful in this context; PCA is concerned with projections maximising variation of a data set taken as whole, not making any distinctions between subsets of the data. Indeed, the procedure relies on source separation occurring in low-numbered PCs; if this does not happen, then the possibility exists of using a different selection of PCs in the analysis than simply the first n. PCA would not work for data where separation of sources was dwarfed by within-source variance, but then it is doubtful whether compositional analysis would be at all sensible.

The decision as to the number of PCs to use is aided by inspection of scatterplots of PCs, using different symbols or colours for each source. We try to find the point at which the separate sources appear indistinguishable. A scree plot is unlikely to be informative; in our examples, we found gains in compositional RMSE accuracy well beyond the point at which the "elbow" would have occurred.

Finally, we note the poor performance of CVA-derived data in this paper. This seems to be due to limitations on the number of CVs which can be derived, whereas the greater number of PCs available allow more subtle features of the compositions to be uncovered.

References

[1] Aitchison J. (1986). *The statistical analysis of compositional data.* Chapman and Hall, London.

[2] Brewer M.J., Dunn S.M. and Soulsby C. (2002). *A Bayesian model for compositional data analysis.* Compstat 2002, Physica-Verlag, Heidelberg, 105–110.

[3] Brewer M.J., Filipe J.A.N., Elston D.A., Dawson L.A., Mayes R.W., Soulsby C. and Dunn S.M. (2004). *A hierarchical model for compositional data analysis.* Journal of Agricultural, Biological and Environmental Statistics. (Submitted)

[4] Chapman, S.J., Campbell, C.D., Fraser, A.R., and Puri, G. (2001). *FTIR spectroscopy of peat in and bordering Scots pine woodland: relationship with chemical and biological properties.* Soil Biology and Biochemistry **33**, 1193-1200.

[5] Lunn D.J., Thomas A., Best N. and Spiegelhalter D.J. (2000). *WinBUGS—a Bayesian modelling framework: concepts, structure and extensibility.* Statistics and Computing **10**, 325–337.

[6] Newman J.A., Thomson W.A., Penning P.D. and Mayes R.W. (1995). *Least-squares estimation of diet composition from n-alkanes in herbage faeces using matrix mathematics.* Australian Journal of Agricultural Research **46**, 793–805.

Acknowledgement: We thank the Scottish Executive Environment and Rural Affairs Department for funding this work. We are grateful to: Colin Campbell and Tony Fraser for the FTIR sample analysis; and Willie Towers and John Bell for work on selecting and sampling of the soils (all Macaulay Institute).

Address: M.J. Brewer, D.A. Elston, Biomathematics and Statistics Scotland, Macaulay Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK
L.A. Dawson, R.W. Mayes, Macaulay Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK

E-mail: M.Brewer@bioss.ac.uk

A ROBUSTIFICATION OF THE JARQUE-BERA TEST OF NORMALITY

Guy Brys, Mia Hubert and Anja Struyf

Key words: Robustness, skewness, tail weight, goodness-of-fit tests.

COMPSTAT 2004 section: Robustness.

Abstract: Many statistical tests have been proposed to find out whether a sample is drawn from a normal distribution or not. Here we discuss the Jarque-Bera test [1] which is based on the classical measures of skewness and kurtosis. As these measures are based on moments of the data, this test has a zero breakdown value. In other words, a single outlier can make the test worthless. In this paper we propose normality tests based on robust measures of skewness and tail weight. We investigate their power and their robustness by means of simulations and examples. We also outline how this approach can be extended to test for other distributions than the normal.

1 Introduction

The third and fourth moments of a distribution are called the skewness and kurtosis. For any distribution F with finite central moments μ_k $(k \leq 3)$, the *skewness* is defined as

$$\gamma_1(F) = \frac{\mu_3(F)}{\mu_2(F)^{3/2}}.$$

Skewness describes the asymmetry of a distribution. A symmetric distribution has zero skewness, an asymmetric distribution with the largest tail to the right has positive skewness, and a distribution with a longer left tail has negative skewness.

For any distribution F with finite central moments μ_k $(k \leq 4)$, the *kurtosis* is defined as

$$\gamma_2(F) = \frac{\mu_4(F)}{\mu_2(F)^2}.$$

There is no agreement on what it really measures. Strictly speaking, kurtosis measures both peakedness and tail heaviness of a distribution relative to that of the normal distribution. Consequently, its use is restricted to symmetric distributions. Finite-sample versions of γ_1 and γ_2 will be denoted by b_1 and b_2.

The classical skewness and kurtosis coefficient have some common disadvantages. They both have a zero breakdown value, and so they are very sensitive to outlying values. One single outlier can make the estimate become very large or small, making it hard to interpret. Another disadvantage is that they are only defined on distributions having finite moments.

In Section 2 we propose several measures of skewness and of left and right tail weight for univariate continuous distributions. Their interpretation is clear and they are robust against outlying values. Contrary to the kurtosis coefficient, the tail weight measures can be applied to symmetric as well as asymmetric distributions. In Section 3 we describe the Jarque-Bera test and introduce some robust normality tests. Section 4 includes simulation results while Section 5 applies the tests on real data. Finally, Section 6 concludes and gives outlines for future research.

2 Robust measures of skewness and tail weight

Assume we have independently sampled n observations $X_n = \{x_1, x_2, \ldots, x_n\}$ from a continuous univariate distribution F. We will consider the *medcouple* (MC), a robust skewness measure, proposed in Brys et al. [2] and extensively discussed in Brys et al. [3]. It is defined as

$$MC(F) = \operatorname*{med}_{x_1 \leq m_F \leq x_2} h(x_1, x_2)$$

with x_1 and x_2 sampled from F, $m_F = F^{-1}(0.5)$ and the kernel function h given by

$$h(x_i, x_j) = \frac{(x_j - m_F) - (m_F - x_i)}{x_j - x_i}.$$

Furthermore, we consider the *left medcouple* (LMC) and *right medcouple* (RMC), respectively the left and right tail weight measure, as defined in Brys et al. [4]. To construct these measures we have applied the medcouple to respectively the left and right half of the samples:

$$\mathrm{LMC}(F) = -\mathrm{MC}(x < m_F) \quad \text{and} \quad \mathrm{RMC}(F) = \mathrm{MC}(x > m_F).$$

Finite sample versions will be denoted by MC_n, LMC_n and RMC_n. These measures can be computed at any distribution, even when finite moments do not exist. Their computation can be performed in $\mathrm{O}(n \log n)$ time due to the fast algorithm described in Brys et al. [3]. They satisfy all natural requirements of skewness or tail weight measures including location and scale invariance. Moreover, they have good robustness properties. More details can be found in the cited references.

3 Normality tests

In this section we discuss four normality tests for the following null and alternative hypothesis:

$$\begin{cases} H_0 : \text{The data are sampled from a normal distribution} \\ H_1 : \text{The data are not sampled from a normal distribution} \end{cases}$$

First, the Jarque-Bera normality test (JB) uses the classical skewness and kurtosis coefficient. As been stated in Moors [7], under the normality assumption ($\gamma_1 = 0$ and $\gamma_2 = 3$) we can write:

	JB	MC1	MC2	MC3
ω	$\begin{pmatrix} 0 & 3 \end{pmatrix}$	$\begin{pmatrix} 0 \end{pmatrix}$	$\begin{pmatrix} 0.199 & 0.199 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.199 & 0.199 \end{pmatrix}$
Σ_k	$\begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix}$	$\begin{pmatrix} 1.25 \end{pmatrix}$	$\begin{pmatrix} 2.62 & -0.0123 \\ -0.0123 & 2.62 \end{pmatrix}$	$\begin{pmatrix} 1.25 & 0.323 & -0.323 \\ 0.323 & 2.62 & -0.0123 \\ -0.323 & -0.0123 & 2.62 \end{pmatrix}$

Table 1: Asymptotic mean ω and covariance matrix Σ_k of the (joint) distribution of several measures of skewness and tail weight.

$$\sqrt{n}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \rightarrow_{\mathcal{D}} N_2\left(\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix}\right)$$

which leads to the Jarque-Bera test statistic:

$$T = n\left(\frac{b_1^2}{6} + \frac{(b_2-3)^2}{24}\right) \approx \chi_2^2.$$

This test can be viewed as a special case of the following generalization. Let $w = (w_1, w_2, \ldots, w_k)$ be estimators of $\omega = (\omega_1, \omega_2, \ldots, \omega_k)$, such that

$$\sqrt{n}\begin{pmatrix} w_1 & \ldots & w_k \end{pmatrix} \rightarrow_{\mathcal{D}} N_k\left(\omega, \Sigma_k\right)$$

then the generalized test statistic T

$$T = n(w-\omega)^t \Sigma_k^{-1}(w-\omega) \approx \chi_k^2.$$

We can thus easily expand the number of goodness-of-fit tests. Taking $k = 2$, $w_1 = b_1$ and $w_2 = b_2$ leads to the JB test with $\omega_1 = \gamma_1 = 0$ and $\omega_2 = \gamma_2 = 3$. A test based on the medcouple (MC1) given in Brys et al. [3] has $k = 1$ and $w_1 = \text{MC}$. Tests solely based on one robust tail weight are proposed in Brys et al. [4]. In the latter paper also a comparison is included with a robust test proposed in Schmid and Trede [8] based solely on quantiles of the data. However as the power of these tests appeared to be rather low, we here construct tests that are based on the joint distribution of several robust measures of skewness and tail weight. First, we define a test based on the left and right medcouple (MC2) with $k = 2$, $w_1 = \text{LMC}$ and $w_2 = \text{RMC}$. Next, we propose a normality test based on MC, LMC and RMC (MC3) where $k = 3$, $w_1 = \text{MC}$, $w_2 = \text{LMC}$ and $w_3 = \text{RMC}$. Table 1 shows the values of ω and Σ_k under the null hypothesis of normality for the proposed normality tests. They are derived from the influence function of the estimators, as described in Brys et al. [3] and in Brys et al. [4].

Note that a robust test of normality could also be obtained by removing the outliers from the data, using an outlier detection rule such as provided by the boxplot or a rule based on robust estimators of location and scale. When the majority of the data are indeed normally distributed, this is a valuable alternative to our robust tests as both the boxplot and the most popular robust estimators of location and scale (such as M-estimators) are based on this normal assumption and thus will indicate the correct set of outliers. However it becomes more complicated when even the majority of the data points do not come from a normal distribution. A boxplot for example then typically shows too many outliers, and to construct an outlier rule one should have knowledge about the underlying distribution of the regular observations. Robust tests, such as the ones presented in Schmid and Trede [8], Moors [7] and in this paper, are based on characteristics of the majority of the data points, and hence they do not require an outlier detection procedure. Consequently, these tests are less powerful to detect non-normality than classical tests, but they are less sensitive to outlying values. This will be demonstrated in the next sections.

4 Simulation study

We investigate the performance of the four normality tests at Tukey's class of gh-distributions [6]. When a random variable Z is standard gaussian distributed, then

$$Y_{g,h} = \begin{cases} \frac{(e^{gZ}-1)}{g} e^{\frac{hZ^2}{2}} & g \neq 0 \\ Ze^{\frac{hZ^2}{2}} & g = 0 \end{cases}$$

is said to follow a gh-distribution $G_{g,h}$ with parameters $g \in \mathbb{R}$ and $h \geq 0$. The parameter g controls the skewness of the distribution, whereas h effects the tail weight. We generated 1000 samples of size 1000 from some $G_{g,h}$ distributions, and summarized the results in Table 2 by calculating the fraction of 1000 samples on which the null hypothesis of normality was rejected at the 5% significance level. In this way, the first column depicts the actual size of the tests, while the other columns represent their power.

It is straightforward to see that the JB test outperforms the other normality tests, followed by MC3 which is much more conservative. The poor performance of MC1 at fat tailed but symmetric distributions $G_{0,h}$, is due to the fact this test is based solely on the skewness of the distribution. Although MC2 is based on tail weight only, it also detects deviations from symmetry, which is a consequence of considering both the left and the right tail weight. Nevertheless, the power values of MC1 and MC2 are mostly lower than those of their combined test MC3.

We now compare the robustness of the normality tests using contaminated normal samples. Contaminated samples were created by taking normal samples of size $1000(1-\varepsilon)$, and adding a normal sample of outliers N(a,σ^2=1) of size 1000ε with $a = 7$ (right contamination, RC) and $a = -7$ (left contamination, LC). With central contamination (CC) a normal sample N(0,σ^2=0.05) of size 1000ε was added. We have also studied a more dispersed symmetric contamination (SC) by adding a normal sample N(0,σ^2=5) of size 1000ε. Here, we take ε equal to 1% and 5%. In Table 3 the fraction of 1000 samples on which the null hypothesis of normality was rejected is given. These values should remain close to the prescribed significance level of 5%.

	$G_{0,0.0}$	$G_{0,0.1}$	$G_{0,0.2}$	$G_{0,0.3}$	$G_{0.1,0.0}$	$G_{0.1,0.1}$	$G_{0.3,0.0}$	$G_{0.3,0.1}$
JB	0.038	0.999	1.000	1.000	1.000	1.000	1.000	1.000
MC1	0.038	0.058	0.063	0.076	0.225	0.235	0.965	0.941
MC2	0.036	0.218	0.688	0.939	0.064	0.254	0.425	0.560
MC3	0.030	0.196	0.617	0.914	0.223	0.383	0.986	0.991

Table 2: Fraction of 1000 samples of data size 1000 from several distributions $G_{g,h}$ on which the null hypothesis of normality was rejected at the 5% significance level.

	RC(1%)	RC(5%)	LC(1%)	LC(5%)	SC(1%)	SC(5%)	CC(1%)	CC(5%)
JB	1.000	1.000	1.000	1.000	0.991	1.000	0.058	0.162
MC1	0.061	0.319	0.067	0.338	0.039	0.072	0.028	0.033
MC2	0.056	0.359	0.044	0.362	0.058	0.094	0.041	0.046
MC3	0.050	0.588	0.052	0.598	0.058	0.092	0.031	0.047

Table 3: Fraction of 1000 samples of data size 1000 from several contaminated normal samples on which the null hypothesis of normality was rejected.

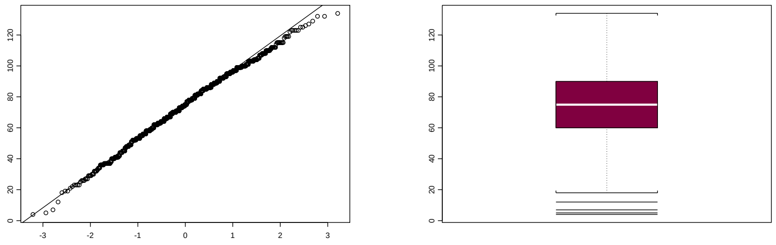
We notice that the JB test is very sensitive to outlying values, especially at right, left and symmetric contamination. The robust normality tests perform better in all cases. Their performance is very comparable except at 5% left and right contamination where MC1 and MC2 are more robust than MC3.

5 Examples

In this section we analyse two data sets which illustrate the robustness of our tests compared to the JB test.

The first data set comes from the Associated Examining Board in Guilford [5] and contains a sample of 1000 scores of students on the writing of a paper. From the normal QQ-plot of Figure 1(a) and the boxplot in Figure 1(b) the assumption of normality seems appropriate. Only a few minor outliers are visible on the boxplot. In Table 4 the non-robustness of the JB test is illustrated. Normality is rejected at the 5% significance level when the outliers from the boxplot are included, but is accepted when they are excluded. On the contrary, the robust normality tests are based on the majority of the data and so they behave the same in both situations. As could be expected, they all detect normality in this data set. Note also the higher significance values of the robust tests compared to the JB test.

Our second example is the speed of light data set [9] which measures the time required for light to travel from a laboratory to a mirror and back, over a total distance of 7400m. This data set contains 66 observations, which is low compared with our simulation study, but we will see that also in this situation the Jarque-Bera test may fail. From the normal QQ-plot and the boxplot in Figure 2 we have the impression that these data come from a normal



(a) (b)

Figure 1: The Guilford data: (a) normal QQ-plot; (b) boxplot.

	JB	MC1	MC2	MC3
Guilford, outliers included	0.039	1.000	0.901	0.975
Guilford, outliers excluded	0.087	1.000	0.967	0.995
Speed of light, outliers included	0.000	1.000	0.308	0.492
Speed of light, outliers excluded	0.855	1.000	0.957	0.992

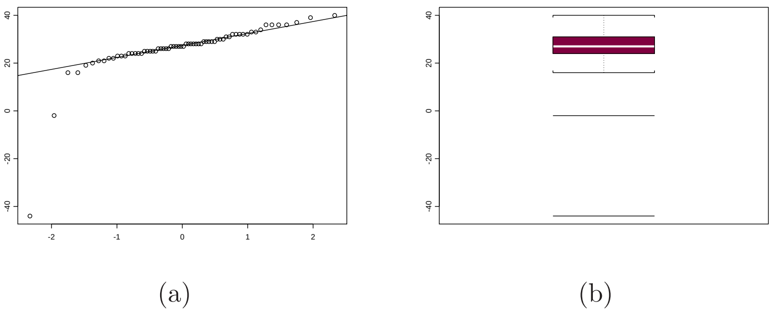
Table 4: Significance of the tests JB, MC1, MC2 and MC3 at the Guilford and speed of light data set, both with outliers included or excluded.

distribution, apart from two clear outliers. As can be seen from Table 4, the JB test applied on the full data rejects the null hypothesis at any significance level, a result which is due to the two outlying values. Excluding those outliers again leads to the opposite conclusion. The robust normality tests detect in both situations the normal behavior which is present in the large majority of the observations.

Of course we do not know whether the so-called outliers in these two data sets are contaminated observations, or whether the true distribution has a long left tail. We therefore recommend to apply both the robust and the classical test. If they contradict each other, the researcher is stimulated to study the nature of these ‘outliers’ in detail.

6 Conclusions

In this paper we have discussed the Jarque-Bera test of normality which is not able to detect normality in the presence of outlying values. Therefore, three robust normality tests have been proposed, which are all based on the medcouple, a robust measure of skewness. The test MC3 which combines the medcouple with a left (LMC) and a right (RMC) measure of tail weight is seen to give the best overall result.



(a) (b)

Figure 2: The speed of light data: (a) normal QQ-plot; (b) boxplot.

In our future research, we will investigate the power of these tests at smaller sample sizes and we will generalize our approach to test whether data are sampled from another distribution than the normal (where again, outlier detection rules become complicated). We then only need to compute the asymptotic mean and covariance matrix as in Table 1. This can be done quickly using a Mathematica program which is available from our website (`www.agoras.ua.ac.be`). Note that the Jarque-Bera test can only be generalized for distributions where the third and fourth moment exist, an assumption which often doesn't hold.

These generalized robust tests will then be used to test distributional assumptions of robust procedures. If, for example, a robust covariance matrix is computed, the underlying assumption is multivariate normality for the majority of the data. A robust test can then be applied to the Mahalanobis distances from the robust fit. The same idea applies to check normality of the residuals from a robust regression procedure.

References

[1] Bera A., Jarque C. (1981). *Efficient tests for normality, heteroskedasticity and serial independence of regression residuals: Monte Carlo evidence.* Economics Letter **7**, 313 − 318.

[2] Brys G., Hubert M., Struyf A. (2003a). *A comparison of some new measures of skewness.* In: Developments in Robust Statistics, ICORS 2001, R. Dutter, P. Filzmoser, U. Gather and P.J. Rousseeuw (eds.), Springer-Verlag, Heidelberg, 98 − 113.

[3] Brys G., Hubert M., Struyf A. (2003b). *A robust measure of skewness.* Journal of Computational and Graphical Statistics, to appear.

[4] Brys G., Hubert M., Struyf A. (2004). *Robust measures of tail weight.* Submitted, available at `http://www.agoras.ua.ac.be`.

[5] Cresswell M.J. (1990). *Gendar effects in GCSE, some initial analyses.* Research Report, Associated Examining Board, Guilford, **517**.

[6] Hoaglin D.C., Mosteller F., Tukey J.W. (1985). *Exploring data tables, trends and shapes.* John Wiley and Sons, New York, 1985.

[7] Moors J.J.A., Wagemakers R.T.A., Coenen V.M.J., Heuts R.M.J., Janssens M.J.B.T. (1996). *Characterizing systems of distributions by quantile measures.* Statistica Neerlandica, **50**, 417 – 430.

[8] Schmid F., Trede M. (2002). *Simple tests for peakedness, fat tails and leptokurtosis based on quantiles.* Computational Statistics and Data Analysis **43**, 1 – 12.

[9] Stigler S.M. (1977). *Do robust estimators work with real data.* The Annals of Statistics, **5**, 1055 – 1098.

Address: G. Brys, Faculty of Applied Economics, Universiteit Antwerpen, Prins-straat 13, B-2000, Antwerpen, Belgium
M. Hubert, Department of Mathematics, Katholieke Universiteit Leuven, W. de Croylaan 54, B-3001 Leuven, Belgium
A. Struyf, Postdoctoral Fellow of the Fund for Scientific Research - Flanders (Belgium), Department of Mathematics and Computer Science, Universiteit Antwerpen, Middelheimlaan 1, B-2020 Antwerpen, Belgium

E-mail: guy.brys@ua.ac.be, mia.hubert@wis.kuleuven.ac.be, anja.struyf@ua.ac.be

COMPSTAT'2004 Symposium

A GENERALISED PAV ALGORITHM FOR MONOTONIC REGRESSION IN SEVERAL VARIABLES

Oleg Burdakow, Anders Grimwall and M. Hussian

Key words: Statistical computing, numerical algorithms, monotonic regression, nonparametric regression, pool-adjacent-violators algorithm.

COMPSTAT 2004 section: Nonparametrical statistics.

Abstract: We present a new algorithm for monotonic regression in one or more explanatory variables. Formally, our method generalises the well-known PAV (pool-adjacent-violators) algorithm from fully to partially ordered data. The computational complexity of our algorithm is $O(n^2)$. The goodness-of-fit to observed data is much closer to optimal than for simple averaging techniques.

1 Introduction

Monotonic regression is a nonparametric method that is appropriate to use when a response variable (y) is increasing or decreasing in one or more explanatory variables ($x_1, \ldots, x_p$). Over the past decades, several numerical algorithms have been developed to facilitate practical application of this method. However, all the algorithms that are currently in use have considerable drawbacks.

The most widespread computational method for monotonic regression is the so-called pool-adjacent-violators (PAV) algorithm [1], [2], [8]. When $p = 1$, this algorithm is computationally efficient and provides solutions that are optimal in the sense that the mean square error is minimised. However, if $p > 1$, the PAV algorithm is less useful. Special cases in which the values of the explanatory variables can be grouped into a moderate number of classes can be handled by repeatedly applying the PAV algorithm to different subsets of data [4], [6], [15], [16]. Other approaches are required for typical multiple regression data in which at least one of the explanatory variables is continuous.

Simple averaging techniques constitute another widespread group of methods. The main idea is to form a weighted mean of two monotonic functions that embrace all observed values of the response variable [11], [12], [16]. In contrast to the PAV algorithm, simple averaging techniques can easily accommodate several explanatory variables. On the other hand, the latter techniques are sensitive to outliers, and the goodness-of-fit can be far from optimal.

Quadratic programming provides yet another approach to monotonic regression [3], [7]. Given a set of observations $\{(x_{1i}, \ldots, x_{pi}, y_i), i = 1, \ldots, n\}$, we shall find a set of fitted values $\{z_i, i = 1, \ldots, n\}$ such that

$$S = \sum_{i=1}^{n} (z_i - y_i)^2$$

is minimised under the constraints induced by the partially ordered data, namely

$$z_i \leq z_j, \text{ if } x_{ki} \leq x_{kj} \text{ for } k = 1, \dots, p.$$

All available algorithms for such solutions entail a considerable computational burden, even for moderately large data sets. The best known computational complexity is $O(n^4)$, and it refers to an algorithm introduced in [10]. Development of more efficient algorithms remains an open problem.

Here, we generalise the PAV algorithm from fully to partially ordered data, and we show how this algorithm can be used for monotonic regression in one or more explanatory variables. In addition, we examine the performance of this algorithm with respect to computational burden and goodness-of-fit to observed data.

2 Main characteristics of the PAV algorithm

Let $M_n = \{(x_i, y_i), i = 1, \dots, n\}$ denote a set of n observations of one explanatory variable (x) and one response variable (y), and assume that the x-values are sorted in increasing order. Then the PAV algorithm computes a non-decreasing sequence of values $\{z_i, i = 1, \dots, n\}$ such that

$$S = \sum_{i=1}^{n} (z_i - y_i)^2$$

is minimised. The cited algorithm is recursive in the sense that the optimal solution for the data set M_n is constructed by starting from the solution for M_1, which is subsequently modified into the solution for M_2, and so on. Moreover, it has the following characteristics:

(i) M_{r+1} is formed by extending M_r with a data point (x_{r+1}, y_{r+1}), such that $x_{r+1} \leq x_i$ for all $i > r + 1$.

(ii) If the values $z_1, \dots, z_r$ denote the solution obtained for M_r, then a preliminary solution for M_{r+1} is formed by setting $z_{r+1} = y_{r+1}$. Thereafter, the final solution for M_{r+1} is derived by pooling adjacent z-values that violate the monotonicity constraints. To be more precise, $z_{r+1}, \dots, z_{r+1-k}$ are assigned the common value $(z_{r+1} + \dots + z_{r+1-k})/(k + 1)$, where k is the smallest non-negative integer such that the new values of $z_1, \dots, z_{r+1}$ form a non-decreasing sequence.

(iii) The optimal solution $\{z_i, i = 1, \dots, n\}$ for M_n is composed of clusters of identical z-values.

3 An alternative formulation of the PAV algorithm

The removal of monotonicity violators implies that adjacent clusters of identical z-values are joined to form new and larger clusters. To achieve a more precise reformulation of the PAV algorithm, we introduce the notation $I = \{i_1, \ldots, i_m\}$ for a cluster consisting of a set of adjacent indices $i_1, \ldots, i_m$. Furthermore, we use $|I|$ for the number of elements in I, and the symbol $z(I)$ for the common value of all $z_i, i \in I$. When two adjacent clusters I_1 and I_2 are joined to form a new cluster $I_1 \cup I_2$, the associated z-value is given by the expression

$$z(I_1 \cup I_2) = \frac{|I_1|z(I_1) + |I_2|z(I_2)}{|I_1| + |I_2|}.$$

If the clusters $I_1, \ldots, I_q$ and their associated values $z(I_1), \ldots, z(I_q)$ compose the optimal solution for M_r, then a preliminary solution for M_{r+1} is formed by introducing the cluster I_{q+1} consisting of the integer $r + 1$, and setting $z(I_{q+1}) = y_{r+1}$. Thereafter, the final solution for M_{r+1} is obtained by joining I_{q+1} with adjacent left-neighbour clusters, one by one, until the z-values violating the monotonicity constraints have been removed.

4 A generalised PAV algorithm for partially ordered data

In our generalisation of the PAV algorithm to partially ordered data, we use the notions introduced in the previous sections. Let $M_n = \{(x_{1i}, \ldots, x_{pi}, y_i),$ $i = 1, \ldots, n\}$ denote a set of n observations of p explanatory variables and one response variable. Also, let $x_i = (x_{1i}, \ldots, x_{pi}), i = 1, \ldots, n$, denote the elements in R^p that are defined by the explanatory variables. Then we can define a partial order on the set $U_n = \{x_i, i = 1, \ldots, n\}$ by setting

$$x_i \preceq x_j \text{ if } x_{ki} \leq x_{kj}, k = 1, \ldots, p,$$

and subsequently sort the elements of U_n (and M_n) in such a way that, for each i, x_i is a minimal element of the set $V_i = \{x_j, j = i, \ldots, n\}$. Furthermore, we can compute the lower cover L_i of each x_i. The latter set consists of all elements x_j, such that

$$x_j \preceq x_i$$

and the inequalites

$$x_j \preceq x_k \preceq x_i$$

are satisfied only if $x_j = x_k$ or $x_k = x_i$.

Like the original PAV algorithm, our generalisation is recursive. Furthermore, it has the following features:

(i) M_{r+1} is formed by extending M_r with a data point (x_{r+1}, y_{r+1}), such that, for all $i > r+1$, either $x_{r+1} \preceq x_i$, or x_{r+1} is incomparable with x_i.

(ii) If the clusters $I_1, \ldots, I_q$ and their associated values $z(I_1), \ldots, z(I_q)$ denote a solution for M_r, then a preliminary solution for M_{r+1} is formed by introducing the cluster I_{q+1} consisting of the integer $r + 1$, and setting $z(I_{q+1}) = y_{r+1}$. Thereafter, the final solution for M_{r+1} is obtained by joining I_{q+1} with left-neighbour clusters, one by one, until the z-values violating the monotonicity constraints have been removed. A cluster I_j is called a left neighbour of I_l if there exists an $i \in I_j$ and a $k \in I_l$ such that x_i belongs to the lower cover of x_k.

(iii) The solution $\{z_i, i = 1, \ldots, n\}$ obtained for M_n is composed of clusters of identical z-values.

Due to their construction, the solutions obtained by using the generalised PAV algorithm are monotonic in the explanatory variables. However, two ambiguities should be noted:

(i) a cluster may have several different left neighbours;

(ii) the pre-sorting which ensures that, for each i, x_i is a minimal element of the set $V_i = \{x_j, j = i, \ldots, n\}$ may be done in different ways.

The first ambiguity is easy to remove. The goodness-of-fit will be improved, if the largest violator of monotonicity is removed first, whenever a cluster has several left neighbours. The second ambiguity is more intricate, and our generalised PAV algorithm will not necessarily attain the minimal value of the mean square error.

5 Computational burden

The computations can be divided into pre-calculations and a recursive establishment of solutions for the data sets $M_r, r = 1, \ldots, n$. The pre-calculations have three major components: (i) establishment of a partial order on the set $\{x_i, i = 1, \ldots, n\}$; (ii) sorting of the observations to ensure that, for each i, x_i is a minimal element of the set $V_i = \{x_j, j = i, \ldots, n\}$; (iii) calculation of the lower cover of each x_i.

Test runs of a VisualBasic implementation of the algorithm showed that data sets consisting of several hundred observations can be processed in less than a second using an ordinary PC. Most of the computer time was usually spent on the calculation of lower covers, followed by the establishment of a partial order and the removal of monotonicity violators. It is also noteworthy that with our generalised algorithm, after the partial order has been established, the number of explanatory variables does not influence the computational burden.

A theoretical analysis of the proposed algorithm showed that its complexity is $O(n^2)$. This result will be proved in a separate paper. The proof is based on the following observations. The pre-calculations can be carried out in $O(n^2)$ elementary arithmetic operations. Each operation of joining clusters is preceded by the search of the largest violator among the left neighbours.

Correlation	Mean square error			
	Normally distr. errors		Exponentially distr. errors	
	GPAV	SA	GPAV	SA
0	0.72	0.80	0.77	1.05
0.9	0.75	0.90	0.75	1.15
-0.9	0.75	0.88	0.75	1.12

Table 1: Mean square error for monotonic regression in two explanatory variables using the generalised PAV algorithm (GPAV) and a simple averaging technique (SA). The table shows mean values for 100 data sets, each consisting of 400 observations.

Such search requires at most n comparisons. The complexity of joining clusters is also $O(n)$. Since the total number of joinings cannot exceed n, the overall complexity related to joining is $O(n^2)$. The number of cases, in which the search of the largest among the left neighbours does not result in joining of clusters, is below n. Thus, the contribution of these cases to the complexity does not exceed $O(n^2)$.

When the structure of partially ordered data is a tree, the algorithm is guaranteed to produce the optimal solution, and it has the same complexity as the algorithm introduced in [13], namely $O(n \log n)$.

6 Goodness-of-fit

A simulation study was undertaken to compare the goodness-of-fit that could be achieved by applying (i) the generalised PAV algorithm and (ii) the simple averaging technique described by Mukarjee in [11]. All of the analysed data sets were generated according to the equation

$$y = x_1 + x_2 + \epsilon$$

where $(x_1, x_2)^*$ was normally distributed with mean zero, variance one, and correlation ρ. The error terms ϵ were either normally or exponentially distributed with variance 1. Table 1 shows that, regardless of the distribution of the error terms or the correlation between the two explanatory variables, the generalised PAV algorithm performed better than the simple averaging technique. Furthermore, the difference in goodness-of-fit was particularly large for heavy-tailed (exponentially distributed) error terms.

7 Discussion

Models of monotonic responses in two or more explanatory variables have a large number of applications in many areas. Thus far, use of monotonic regression has been hampered by the lack of algorithms suitable for typical multiple regression data. With the generalised PAV algorithm, it is feasible to handle data sets that include hundreds or even thousands of observations of one response variable and an arbitrary number of explanatory variables. For moderately large data sets, it is possible to combine monotonic regression with general model selection techniques, such as cross-validation.

The test runs described in this article show that our generalisation of the PAV algorithm has high efficiency from the viewpoint of both its accuracy (goodness-of-fit) and computational time. Two recent conference contributions [5], [9] involving further numerical experiments and applications in environmental science confirm the main results. The present version of the generalised PAV algorithm is superior to simple averaging techniques, but it may not always provide least squares solutions. Further work is needed to determine the extent to which the goodness-of-fit can be improved by removing or reducing the non-optimal performance.

References

[1] Ayer M., Brunk H.D., Ewing G.M., Reid W.T., Silverman E. (1955). *An empirical distribution function for sampling with incomplete information.* The Annals of Mathematical Statistics **26**, 641 – 647.

[2] Barlow R.E., Bartholomew D.J., Bremner J.M., Brunk H.D. (1972). *Statistical inference under order restrictions.* Wiley, New York.

[3] Best M.J., Chakravarti N. (1990). *Active set algorithms for isotonic regression: a unifying framework.* Mathematical Programming **47**, 425 – 439.

[4] Bril G., Dykstra R., Pillers C., Robertson T. (1984). *Algorithm AS 206, isotonic regression in two independent variables.* Applied Statistics **33**, 352 – 357.

[5] Burdakov O., Sysoev O., Grimvall A., Hussian M. (2004). *An algorithm for isotonic regression problems.* To appear in the Proceedings of the 4th European Congress of Computational Methods in Applied Science and Engineering 'ECCOMAS 2004'.

[6] Dykstra R., Robertson T. (1982). *An algorithm for isotonic regression for two or more independent variables.* The Annals of Statistics **10**, 708 – 716.

[7] Gamarnik D. (1998). *Efficient learning of monotone concepts via quadratic optimisation.* Proceedings of the Eleventh Annual Conference on Computational Learning Theory, July 24-26, 1998, USA, Wisconsin, Madison 134 – 143.

[8] Hanson D.L., Pledger G., Wright F.T. (1973). *On consistency in monotonic regression.* Annals of Statistics **1**, 401 – 421.

[9] Hussian M., Grimvall A., Burdakov O., Sysoev O. (2004). *Monotonic regression for trend assessment of environmental quality data.* To appear in the Proceedings of the 4th European Congress of Computational Methods in Applied Science and Engineering 'ECCOMAS 2004'.

[10] Maxwell W.L., Muchstadt J.A. (1983). *Establishing consistent and realistic reorder intervals in production-distribution systems.* Operations Research **33**, 1316 – 1341.

[11] Mukarjee H. (1988). *Monotone nonparametric regression.* The Annals of Statistics **16**, 741 – 750.

[12] Mukarjee H., Stem H. (1994). *Feasible nonparametric estimation of multiargument monotone functions.* Journal of the American Statistical Association **425**, 77 – 80.

[13] Pardalos P.M., Xue G. (1999). *Algorithms for a class of isotonic regression problems.* Algorithmica **23**, 211 – 222.

[14] Salanti G., Ulm K. (2001). *The multidimensional isotonic regression.* Proceedings Book, International Society of Clinical Biostatistics, 19-23 August, Sweden, Stockholm, 162.

[15] Schell M.J., Singh B. (1997). *The reduced monotonic regression method.* Journal of the American Statistical Association **92**, 128 – 135.

[16] Strand M. (2003). *Comparison of methods for monotone nonparametric multiple regression.* Communications in Statistics - Simulation and Computation **32**, 165 – 178.

Acknowledgement: The authors are grateful for financial support from the Swedish Environmental Protection Board and the Swedish Research Council.

Address: O. Burdakov, A. Grimvall, M. Hussian, Department of Mathematics, Linköping University, Sweden

E-mail: `angri@mai.liu.se`

CONDITIONAL QUANTILES WITH FUNCTIONAL COVARIATES: AN APPLICATION TO OZONE POLLUTION FORECASTING

Hervé Cardot, Christophe Crambes and Pascal Sarda

Key words: Functional data analysis, conditional quantiles, robustness, B-spline functions, weighted least squares, backfitting algorithm, Ozone forecasting.

COMPSTAT 2004 section: Functional data analysis.

Abstract: We are interested in estimating conditional quantiles when the covariariates are functions. We modelize the conditional quantiles as a continuous linear functional of the covariates and we propose a spline estimator of the coefficient which minimizes a L^1-type penalized criterion. Then we give some insights on the asymptotic behaviour of this estimator. This approach is illustrated on pollution data in the area of Toulouse.

1 Presentation of the pollution data

Pollution forecasting is nowadays of primary importance, particularly for prevention. In the city of Toulouse (France), the ORAMIP [1] gets measures, in six different places, of specific pollutants. These measures, as well as meteorological measures, are made each hour. So we have functional variables (see Ramsay and Silverman [13], [14]) known in some discretisation points.

More precisely, the data consist in hourly measurements during the period going from the 15th May to the 15th September for the years 1997, 1998, 1999 and 2000. Measurements are obtained for the following variables: Nitrogen Monoxide, Nitrogen Dioxide, Dusts, Ozone, Wind Direction, Wind Speed, Temperature, Humidity and Sun Radiance. Some data are missing, mainly because of breakdowns or missing measurement apparatus.

A PCA of these data has shown that the behaviour of these variables does not vary much from one station to another. This allows us to handle the problem of missing data by considering a "mean" station, taking the functional mean over the stations for each variable.

The aim of the study is to predict the maximum of Ozone for a day knowing the (functional) values of the above five variables the day before. We present in section 2 a generalization of the model proposed by Koenker and Bassett [9] which allows to estimate a conditional quantile for several functional predictor. Estimates are built by a B-splines expansion and the minimization of a L^1-type penalized criterion. Our methodology is illustrated in section 3 with the prediction of Ozone in the area of Toulouse.

[1] "Observatoire Régional de l'Air en Midi-Pyrénées"

2 Quantile regression for functional covariates

2.1 Conditional quantiles for functional covariates

Let us consider a sample $(X_i, Y_i)_{i=1,\ldots,n}$ of pairs of random variables, independent and identically distributed, with the same distribution as (X, Y), with X belonging to the functional space $L^2([0,1])$ of square integrable functions defined on $[0,1]$, and Y belonging to $\mathbb{R}$. Without loss of generality, we suppose that X is a centered variable, that is to say $\mathbb{E}(X) = 0$. Assume that H, the range of X, is a closed subspace of $L^2([0,1])$. Let α be a real number in $]0,1[$ and x a function in H. The *conditional α-quantile* of Y given $[X = x]$ is defined as the scalar $g_\alpha(x)$ such that

$$P(Y \leq g_\alpha(x)|X = x) = \alpha, \tag{1}$$

where $P(.|X = x)$ is the conditional probability given $[X = x]$.

Provided that $\mathbb{E}|Y| < \infty$, $g_\alpha(x)$ can be defined in an equivalent way as the solution of the minimization problem

$$\min_{a \in \mathbb{R}} \mathbb{E}(l_\alpha(Y - a)|X = x), \tag{2}$$

where l_α is the function defined by $l_\alpha(u) = |u| + (2\alpha - 1)u$ (see [9]).

We assume now that g_α is a linear and continuous functional. This condition can be seen as the direct generalization of the model introduced by Koenker and Bassett, the difference being that here, the covariates are functions. Then, there exists a unique function $\Psi_\alpha \in L^2([0,1])$ such that

$$g_\alpha(X) = \langle \Psi_\alpha, X \rangle = c + \int_0^1 \Psi_\alpha(t) X(t)\, dt, \tag{3}$$

where the notation $\langle ., . \rangle$ refers to the usual inner product of $L^2([0,1])$. The norm of $L^2([0,1])$ induced by this inner product is denoted by $\|.\|$.

2.2 Spline estimator of Ψ_α

Our goal is to introduce now an estimator of the function Ψ_α. When the covariate X is real, Koenker and Bassett [9] have proposed an estimator based on the minimization of the empirical version of (2); for nonparametric modelling, estimators have already been proposed: see for example Bhattacharya and Gangopadhyay [2], Fan, Hu and Truong [6] or Lejeune and Sarda [11]. He and Shi [8] proposed a spline estimator and although our setting is quite different, the estimator of Ψ_α defined below is of the same type as the one introduced by He and Shi (based on regression splines). However in our (functional) case there is a need to introduce a penalization in the criterion to be minimized.

We consider a space of spline functions: for this we choose a degree (fixed) and knots in the interval $[0,1]$. For given integers q and $k = k_n$, with $k \neq 0$, we consider splines of degree q and $k-1$ equispaced knots on $[0,1]$ (see [5]). This space of spline functions is a vectorial space of dimension $k+q$. A basis of this vectorial space is the set of the so-called B-spline functions, that we note $\mathbf{B_{k,q}} = {}^{t}(B_1, \ldots, B_{k+q})$.

We estimate Ψ_α by a linear combination of the B_l functions. This leads us to find a vector $\widehat{\boldsymbol{\theta}} = {}^{t}(\widehat{\theta}_1, \ldots, \widehat{\theta}_{k+q})$ in $\mathbb{R}^{k+q}$ such that

$$\widehat{\Psi}_\alpha = \sum_{l=1}^{k+q} \widehat{\theta}_l B_l = {}^{t}\mathbf{B_{k,q}}\widehat{\boldsymbol{\theta}}, \tag{4}$$

where $\widehat{\boldsymbol{\theta}}$ is the solution of the following minimization problem, which is the penalized empirical version of (2),

$$\min_{\theta \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^{n} l_\alpha(Y_i - c - \langle {}^{t}\mathbf{B_{k,q}}\boldsymbol{\theta}, X_i \rangle) + \rho \parallel ({}^{t}\mathbf{B_{k,q}}\boldsymbol{\theta})^{(m)} \parallel^2 \right\}, \tag{5}$$

where $({}^{t}\mathbf{B_{k,q}}\boldsymbol{\theta})^{(m)}$ is the m-th derivative of the spline function ${}^{t}\mathbf{B_{k,q}}\boldsymbol{\theta}$ and ρ is a penalization parameter which role is to control the smoothness of the estimator. This criterion is similar to the one introduced by Cardot et al. [4] for the estimation of the conditional mean in the functional linear model, the quadratic function being replaced here by the loss function l_α. In this case, we have to deal with an optimization problem that does not have an explicit solution, contrary to the case of the functional linear model. That is why we adopted the strategy proposed by Lejeune and Sarda [11]. At first, let us note that the minimization problem (5) is equivalent to

$$\min_{c \in \mathbb{R}, \boldsymbol{\theta} \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \delta_i(\alpha) \mid Y_i - c - \langle {}^{t}\mathbf{B}_{k,q}\boldsymbol{\theta}, X_i \rangle \mid + \rho \parallel ({}^{t}\mathbf{B}_{k,q}\boldsymbol{\theta})^{(m)} \parallel^2 \right\}. \tag{6}$$

where the function δ_i is defined by $\delta_i(\alpha) = 2\alpha \mathbb{1}_{\{Y_i - c - \langle {}^{t}\mathbf{B}_{k,q}\theta, X_i \rangle \geq 0\}} + 2(1-\alpha)\mathbb{1}_{\{Y_i - c - \langle {}^{t}\mathbf{B}_{k,q}\theta, X_i \rangle < 0\}}$. The algorithm for solving (6) described below consists in performing iterative reweighted least squares (see [15]: at step $j+1$, $\delta_i(\alpha)$ is replaced by the value $\delta_i^j(\alpha)$ evaluated at step j and the absolute value is replaced by the ratio of square residuals (at step $j+1$) on the root of residuals computed from step j.

- **Initialization**

  We determine $\boldsymbol{\beta}^1 = {}^{t}(c^1, \boldsymbol{\theta}^1)$ solution of the minimization problem

$$\min_{c \in \mathbb{R}, \boldsymbol{\theta} \in \mathbb{R}^{k+q}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - c - \langle {}^{t}\mathbf{B}_{k,q}\boldsymbol{\theta}, X_i \rangle)^2 + \rho \parallel ({}^{t}\mathbf{B}_{k,q}\boldsymbol{\theta})^{(m)} \parallel^2 \right\},$$

  which solution $\boldsymbol{\beta}^1$ is given by $\boldsymbol{\beta}^1 = \frac{1}{n}(\frac{1}{n}\,{}^{t}\mathbf{D}\mathbf{D} + \rho\mathbf{K})^{-1}\,{}^{t}\mathbf{D}\mathbf{Y}$, with

$$\mathbf{D} = \begin{pmatrix} 1 & \langle B_1, X_1\rangle & \dots & \langle B_{k+q}, X_1\rangle \\ \vdots & \vdots & & \vdots \\ 1 & \langle B_1, X_n\rangle & \dots & \langle B_{k+q}, X_n\rangle \end{pmatrix} \quad \text{and} \quad \mathbf{K} = \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{G} \end{pmatrix},$$

where $\mathbf{G}$ is the $(k+q)\times(k+q)$ matrix such that $\mathbf{G}_{jl} =< B_j^{(m)}, B_l^{(m)} >$.

- **Step j+1**

  Knowing $\boldsymbol{\beta}^j = {}^t(c^j, \boldsymbol{\theta}^j)$, we determine $\boldsymbol{\beta}^{j+1} = {}^t(c^{j+1}, \boldsymbol{\theta}^{j+1})$ solution of the minimization problem

$$\min_{c\in\mathbb{R}, \boldsymbol{\theta}\in\mathbb{R}^{k+q}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i^j(\alpha)(Y_i - c - \langle {}^t\mathbf{B}_{k,q}\boldsymbol{\theta}, X_i\rangle)^2}{[(Y_i - c - \langle {}^t\mathbf{B}_{k,q}\boldsymbol{\theta}^j, X_i\rangle)^2 + \eta^2]^{1/2}} + \rho \parallel ({}^t\mathbf{B}_{k,q}\boldsymbol{\theta})^{(m)} \parallel^2 \right\},$$

  where η is a strictly positive constant that allows us to avoid a denominator equal to zero. Let us define the $n \times n$ diagonal matrix $\mathbf{W}_j$ with diagonal elements given by

$$[\mathbf{W}_j]_{ll} = \frac{\delta_1^j(\alpha)}{n[(Y_l - c - \langle {}^t\mathbf{B}_{k,q}\boldsymbol{\theta}, X_l\rangle)^2 + \eta^2]^{1/2}}.$$

  Then, $\beta^{j+1} = ({}^t\mathbf{D}\mathbf{W}_j\mathbf{D} + \rho\mathbf{K})^{-1}\ {}^t\mathbf{D}\mathbf{W}_j\mathbf{Y}$.

Let us notice that, at each step of the algorithm, the estimator depends on many parameters: the number k of knots, the degree q of the spline functions, the order m of derivation, and the parameter ρ of the penalization. In our experience, we noticed that the penalization parameter ρ is of primary importance at least when the number of knots is large enough (see also Marks and Eilers [12], Besse et. al. [1]). For this reason, we choose in our study (see section 3) to fix $q = 3$ and $m = 2$, which are values commonly used to reach a sufficient degree of regularity. After several attempts, we fix k to be 8 *i.e.* a number of knots which is not too small. Now, the parameter ρ is chosen by minimizing the generalized cross validation criterion (see [16])

$$GCV(\rho) = \frac{\frac{1}{n}\sum_{l=1}^{n}(Y_l - \widehat{Y}_l)^2}{\left(1 - \frac{1}{n}Tr(\mathbf{H}(\rho))\right)^2}, \tag{7}$$

where $\widehat{\mathbf{Y}} = \mathbf{H}(\rho)\mathbf{Y}$ and $\mathbf{H}(\rho) = \frac{1}{n}\mathbf{D}(\frac{1}{n}\ {}^t\mathbf{D}\mathbf{D} + \rho\mathbf{K})^{-1}\ {}^t\mathbf{D}$.

2.3 Multiple conditional quantiles

Let us notice that this estimation procedure can be easily extended to the case where there is more than one covariate. Considering now v functional covariates $X^1, \ldots, X^v$, we have the following model

$$P(Y_i \leq c + g_\alpha^1(X_i^1) + \ldots + g_\alpha^v(X_i^v)/X_i^1 = x_i^1, \ldots, X_i^v = x_i^v) = \alpha. \quad (8)$$

If we assume as before that $g_\alpha^1, \ldots, g_\alpha^v$ are linear and continuous functionals from a clsoed subspace of $L^2([0,1])$, we can write $g_\alpha^r(X_i^r) = \langle \Psi_\alpha^r, X_i^r \rangle$ for $r = 1, \ldots, v$ with $\Psi_\alpha^1, \ldots, \Psi_\alpha^v$ in $L^2([0,1])$. The estimation of each function Ψ_α^r is obtained using the iterative backfitting algorithm (see [7]) combined with the algorithm described previously.

2.4 An asymptotic result

In this section we give some insights on the asymptotic behavior of $\widehat{\Psi}_\alpha$. As a matter of fact, we give an upper bound for the L^2 convergence: the proof of this result can be found in Cardot et al. [3].

Let us introduce now some notations. If we assume that $\mathbb{E}\|X\|^2 < \infty$, the covariance operator of X, denoted by Γ_X, is defined for all u in $L^2([0,1])$ by $\Gamma_X u = \mathbb{E}\left(\langle X, u\rangle X\right)$. This operator is non-negative, so we can associate a semi-norm noted $\|.\|_2$ and defined by $\|u\|_2^2 = \langle \Gamma_X u, u\rangle$. Using notations from Cardot et al. [4], we consider the $(k+q)\times(k+q)$ matrix $\widehat{\mathbf{C}}$ with general term $\langle \Gamma_n(B_j), B_l\rangle$ where Γ_n is the empirical version of Γ. Moreover, we set $\widehat{\mathbf{C}}_{\boldsymbol{\rho}} = \widehat{\mathbf{C}} + \rho\mathbf{G}$ where the matrix $\mathbf{G}$ is defined in section 2.2. It is possible to find a sequence $(\eta_n)_{n\in\mathbb{N}}$ of non negative reals such that $\Omega_n = \left\{\omega/\lambda_{\min}(\widehat{\mathbf{C}}_{\boldsymbol{\rho}}) > c\eta_n\right\}$, where $\lambda_{\min}(\widehat{\mathbf{C}}_{\boldsymbol{\rho}})$ is the smallest eigenvalue of $\widehat{\mathbf{C}}_{\boldsymbol{\rho}}$, has probability going to 1 when n goes to infinity (see [4])

To prove the convergence result of the estimator $\widehat{\Psi}_\alpha$, we assume that the following hypotheses are satisfied.

$(H.1)$ $\| X_i \| \leq C_0 < +\infty, \quad as.$

$(H.2)$ The function Ψ_α has a continuous p'-th derivative $\Psi_\alpha^{(p')}$ which satisfies a Lipschitz condition of order $\nu \in [0,1]$.

In the following, we set $p = p' + \nu$ and we suppose that $q \geq p \geq m$.

$(H.3)$ The eigenvalues of Γ_X are strictly positive.

$(H.4)$ For $x \in H$, Y has a conditional density function f_Y^x given $[X = x]$ continuous and strictly positive at the α-quantile.

Then, we have the following result.

Theorem 2.1. *Under hypotheses* $(H.1) - (H.4)$*, if we also suppose there exists* β, γ *in* $]0,1[$ *such that* $k_n \sim n^\beta$, $\rho \sim n^{-\gamma}$ *and* $\eta_n \sim n^{-\beta-(1-\delta)/2}$*, then*

(i) $\widehat{\Psi}_\alpha$ *exists except on a set whose probability goes to zero as n goes to infinity,*

(ii) $\mathbb{E}\big(\| \widehat{\Psi}_\alpha - \Psi_\alpha \|_2^2 \,|X_1, \ldots, X_n\big) = O_P\left(\dfrac{1}{k_n^{2p}} + \dfrac{1}{n\eta_n} + \dfrac{\rho^2}{k_n\eta_n} + \rho k_n^{2(m-p)}\right).$

3 Application to Ozone prediction

We want to predict the maximum of the Ozone variable the day i, denoted by Y_i, using the functional covariates observed the day before until 5:00 pm. We consider covariates with length of 24 hours. We can assume that, beyond 24 hours, the effects of the covariate are negligible knowing the last 24 hours, so that each curve X_i begins at 6:00 pm the day $i-2$.

We ramdomly splitted the initial sample $(X_i, Y_i)_{i=1,\dots,n}$ into two subsamples:

- a learning sample $(X_{a_i}, Y_{a_i})_{i=1,\dots,n_l}$ with $n_l = 332$, used to determine the estimators $\widehat{c}$ and $\widehat{\Psi}_\alpha$,
- a test sample $(X_{t_i}, Y_{t_i})_{i=1,\dots,n_t}$ with $n_t = 142$, used to evaluate the quality of the models and to make a comparison.

We use the conditional median to predict the value of Y_i, *i.e.* $\alpha = 0.5$. To judge the quality of the models, we give a prediction of the maximum of Ozone for each element of the test sample,

$$\widehat{Y_{t_i}} = \widehat{c} + \int_D \widehat{\Psi}_\alpha(t) X_{t_i}(t)\, dt.$$

Then, we consider three criteria given by

$$C_1 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \widehat{Y_{t_i}})^2}{\frac{1}{n_t}\sum_{i=1}^{n_t}(Y_{t_i} - \overline{Y}_l)^2}, \quad C_2 = \frac{1}{n_t}\sum_{i=1}^{n_t} \mid Y_{t_i} - \widehat{Y_{t_i}} \mid,$$

$$C_3 = \frac{\frac{1}{n_t}\sum_{i=1}^{n_t} l_\alpha(Y_{t_i} - \widehat{Y_{t_i}})}{\frac{1}{n_t}\sum_{i=1}^{n_t} l_\alpha(Y_{t_i} - q_\alpha(Y_l))},$$

where $\overline{Y}_l$ is the empirical mean of the learning sample $(Y_{a_i})_{i=1,\dots,n_l}$ and $q_\alpha(Y_l)$ is the empirical α-quantile of the learning sample $(Y_{a_i})_{i=1,\dots,n_l}$. This last criterion C_3 is similar to the one proposed by Koenker and Machado [10]. We remark that, the more these criteria take low values (close to 0), the better is the prediction. After testing all the possible models with one to five covariates, we finally kept the model using the four covariates Ozone, Nitrogen Monoxide, Nitrogen Dioxide and Wind Speed (we have put some of the results obtained in table 1). For this model, figure 1 represents predicted Ozone vs measured Ozone. Except for one outlier, the prediction seems rather good. The most efficient covariate to estimate the maximum of Ozone is the Ozone curve the day before; however, we noticed an improvement adding other covariates.

We can also think that these results could be improved by considering other covariates that were not available, such as for example the curve of temperature. Finally, let us remark that we can similarly estimate conditional quantiles of Y_i or order 0.9 and 0.1 to derive some kind of prediction intervals.

Models	Variables	C_1	C_2	C_3
1 covariate	N2	$0,814$	$16,916$	$0,906$
	O3	**0,414**	**12,246**	**0,656**
	WS	$0,802$	$16,836$	$0,902$
2 covariates	O3, NO	$0,413$	$11,997$	$0,643$
	O3, N2	$0,413$	$11,880$	$0,637$
	O3, WS	$0,414$	$12,004$	$0,635$
3 covariates	O3, NO, N2	$0,412$	$12,127$	$0,644$
	O3, N2, WD	$0,409$	$12,004$	$0,645$
	O3, N2, WS	$0,410$	$11,997$	$0,642$
4 covariates	**O3, NO, N2, WS**	**0,400**	**11,718**	**0,634**
5 covariates	O3, NO, N2, WD, WS	$0,401$	$11,750$	$0,639$

Table 1: Forecast quality for some models of mediane regression.

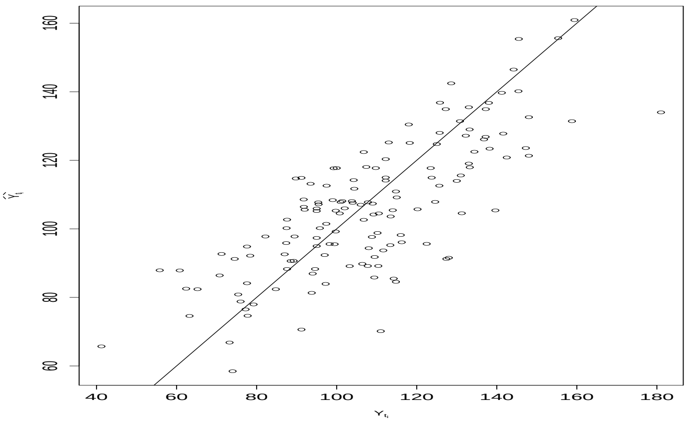


Figure 1: Predicted Ozone vs measured Ozone (prediction with the variables O3, NO, N2 and WS).

References

[1] Besse P.C., Cardot H., Ferraty F. (1997). *Simultaneous nonparametric regression of unbalanced longitudinal data.* Comput. Statist. and Data Anal. **24**, 255 – 270.

[2] Bhattacharya P.K., Gangopadhyay A.K. (1990). *Kernel and nearest-neighbor estimation of a conditional quantile.* Ann. Statist. **18**, 1400 – 1415.

[3] Cardot H., Crambes C., Sarda P. (2004). *Spline estimators of conditional quantiles for functional covariates.* Preprint.

[4] Cardot H., Ferraty F., Sarda P. (2003). *Spline estimators for the functional linear model.* Statistica Sinica **13**, 571 – 591.

[5] de Boor C. (1978). *A practical quide to splines.* Springer, New-York.

[6] Fan J., Hu T.C., Truong Y.K. (1994). *Robust nonparametric function estimation.* Scand. J. Statist **21**, 433 – 446.

[7] Hastie T.J., Tibshirani R.J. (1990). *Generalized additive models.* Chapman and Hall, New-York.

[8] He X., Shi P. (1994). *Convergence rate of B-spline estimators of nonparametric conditional quantile functions.* Nonparametric Statistics **3**, 299 – 308.

[9] Koenker R., Bassett G. (1978). *Regression quantiles.* Econometrica **46**, 33 – 50.

[10] Koenker R., Machado J. (1999). *Goodness of fit and related inference processes for quantile regression.* Journal of the American Statistical Association **94**, 1296 – 1310.

[11] Lejeune M., Sarda P. (1988). *Quantile regression: A nonparametric approch.* Computational Statistics and Data Analysis **6**, 229 – 239.

[12] Marx B.D., Eilers P.H. (1999). *Generalized linear regression on sampled signals and curves: A P-spline approach.* Technometrics **41**, 1 – 13.

[13] Ramsay J.O., Silverman B.W. (1997). *Functional data analysis.* Springer-Verlag.

[14] Ramsay J.O., Silverman B.W. (2002). *Applied functional data analysis.* Springer-Verlag.

[15] Ruppert D., Caroll J. (1988). *Transformation and weighting in regression.* Chapman and Hall.

[16] Wahba G. (1990). *Spline models for observational data.* Society for Industrial and Applied Mathematics, Philadelphia.

Acknowledgement: We would like to thank ORAMIP that provided the pollution data.

Address: H. Cardot, PINRA Toulouse, Biométrie et Intelligence Artificielle,
Chemin de Borde-Rouge, BP 27, 31326 Castanet-Tolosan Cedex, France
C. Crambes, P. Sarda, Université Paul Sabatier, Laboratoire de Statistique et Probabilités, UMR C5583, 118 route de Narbonne, 31062 Toulouse Cedex, France
P. Sarda, Université Toulouse-le-Mirail, GRIMM, EA 2254, 5 allées Antonio Machado, 31058 Toulouse Cedex 9, France

E-mail: `cardot@toulouse.inra.fr`, `crambes@cict.fr`,
`Pascal.Sarda@math.ups-tlse.fr`

COMPSTAT'2004 Symposium

RANDOM EFFECTS VARYING TIME REGRESSION MODELS WITH APPLICATION TO REMOTE SENSING DATA

Hervé Cardot, Robert Faivre and Philippe Maisongrande

Key words: Backfitting, BLUP, covariance operator, ECME, functional data, land use estimation, mixed effects, mixed pixels, splines, unmixing.

COMPSTAT 2004 section: Functional data analysis.

Abstract: The sensor SPOT4/Végétation gives every day satellite images of Europe with medium spatial resolution, each pixel corresponding to an area of 1km×1km. Such data are useful to characterize the development of the vegetation at a large scale. The basic information support units, named *mixed* pixels, aggregate informations of different crops and thus different themes of interest (wheat, corn, forest, ...).

We aim at estimating the local variations of the responses of the different cultures knowing the land use and the temporal evolution of the reflectance of the mixed pixels. We propose an unmixing procedure relying on a random effects varying time regression model. Estimators are based on penalized splines and maximum likelihood. A combination of the backfitting and the ECME algorithms allows us to get a fast estimation procedure. This approach is illustrated in the South-East of France with SPOT4/VGT data obtained during the year 2002.

1 Mixed pixels and characteristic curves expansion

The Végétation sensor of the SPOT4 satellite gives daily images of Europe (high temporal resolution) at a coarse spatial resolution, each pixel corresponding to a ground area of 1 km^2. The information given by this sensor are the reflectances, i.e the proportion of reflected radiation, in a few spectral bands. This information allows to characterize the development of vegetation and crops at the scale of a small country [11]. Nevertheless in Europe, and particularly in France, the size of plots is much less than 1 km^2. Thus the observed reflectances are a mixture of different informations since they contain different agricultural plots (maize, wheat, forest, ...); such pixels are named *mixed pixels*.

When observing mixed pixels, the reflectance is a function of the characteristic reflectances of each particular category. These curves are also called phenological curves when the category is associated to a crop. We can assume the combination is linear so that the reflectance curve $Y_i(t)$ of a pixel i is decomposed as

$$Y_i(t) = \sum_{j=1}^{J} \pi_{ij}\ \rho_{ij}(t) + \varepsilon_{i,t}, \quad t \in [0,T], \tag{1}$$

where π_{ij} is the proportion of land use of crop j, $\rho_{ij}(t)$ is the response curve associated to the category j for the ith pixel and $\varepsilon_{i,t}$ are independent Gaussian random variables with mean zero and variance σ^2. We assume that the land use is fixed during the observation period $[0,T]$, p images are observed at the instants $0 \leq t_1 < \ldots < t_\nu < \ldots < t_p \leq T$. Each image is composed of n pixels so that we get vectors of discretized pixel trajectories $\mathbf{Y}_i = (Y_i(t_1), \ldots, Y_i(t_p))'$, $i = 1, \ldots, n$. In the area under study the land use is supposed to be known and is composed of J different themes (wheat, corn, river, urban, $\ldots$) and we have for each pixel i, the vector of proportions of the surfaces $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ})$ dedicated to each category.

Since agricultural practices may differ from one farmer to another, that the soil properties, the climate, $\ldots$, may also vary with the location, phenological curves which reflect more or less the growth of a culture may be different from one pixel to another. We do not assume anymore that $\rho_{ij}(t) = \rho_j(t)$, as it was done in Cardot et al. [1] to predict the proportion of each crop in a mixed pixel.

Our work deals with the problem of estimating the temporal evolution of the local responses of each crop from the temporal evolution of mixed pixels. Then, it is possible to assimilate the predicted local reflectances of a culture, such as wheat, into a crop growth model in order to predict the total production at a regional scale.

Faivre and Fischer [3] proposed a statistical approach for estimating the local reponses of each crop for a fixed instant t_ν. A random coefficient linear model was built to estimate the distribution of the characteristic reflectances assuming the land use was known; unmixing or disaggregation being done independently on each image, *i.e* at each date. We propose, in section 2, a nonparametric model which can take into account the temporal variations of the responses. It is based on spline decomposition with random coefficients estimated via a penalized maximum likelihood criterion, see section 3. The fixed effects are estimated with a backfitting procedure whereas the ECME algorithm is proposed for the estimation of the random components. A remote sensing application is presented in section 4 and some extensions are proposed in section 5. The method is implemented in the R language and is avalaible on request.

2 A functional random effects model

Assuming the characteristic curves have a Gaussian distribution, we get now, combined with (1), the following model:

$$\begin{cases} Y_i(t) & = & \sum_{j=1}^{J} \pi_{ij}\ \rho_{ij}(t) + \varepsilon_t, \quad t \in [0,T], \\ \rho_{ij} & \sim & \mathcal{N}\left(\rho_j^0, \Gamma_j\right), \quad j = 1, \dots, J, \end{cases} \tag{2}$$

where ρ_j^0 is the mean fonctional response $\rho_j^0(t) = E(\rho_{ij}(t))$ and Γ_j its covariance operator. It is an integral operator whose kernel is $\gamma^j(s,t) = \mathrm{Cov}(\rho_{ij}(s), \rho_{ij}(t))$. We suppose that the noise is not spatially correlated and the response of the different crops are independent, $\mathrm{cov}(\rho_{ij}, \rho_{i'j'}) = 0$ if $i \neq i'$ or $j \neq j'$, so that we do not include cross covariance operators for the conditional variance of the trajectory Y_i.

Model (2) can also be seen as a varying time regression model [4]

$$E\left(Y_i(t)\right) \quad = \quad \sum_{j=1}^{J} \pi_{ij}\ \rho_j^0(t), \tag{3}$$

in which the conditional variance (conditionned to $\boldsymbol{\pi}$) is modelled as follows:

$$\mathrm{Cov}\left(Y_i(s), Y_i(t)\right) \quad = \quad \sigma^2 \delta_{\{s=t\}} + \sum_{j=1}^{J} \pi_{ij}^2\ \gamma^j(s,t)\ . \tag{4}$$

Remark. Models 2 and 4 are extensions of some previous models proposed by Rice and Wu [9], Cardot et al. [1] and Kneip et al. [6]. Rice and Wu [9] have studied nonparametric varying time regression models without explanatory variables whereas Cardot et al. [1] did not include conditional variance terms in a similar model. Kneip et al. [6] studied a semi-parametric model with parametric mean effects and nonparametric random effects but without any covariate.

Spline expansion

Consider two B-splines bases $\{B_1(t), \dots, B_{q+k_1}(t)\}$ and $\{\mathcal{B}_1(t), \dots, \mathcal{B}_{q+k_2}(t)\}$, of order q, with respectively k_1 and k_2 equispaced interior knots [2]. Let us separate the fixed effects from the random ones,

$$\rho_{ij}(t) \quad \approx \quad \sum_{r=1}^{q+k_1} \theta_{r,j}^0 B_r(t) + \sum_{s=1}^{q+k_2} \delta_{s,i}^j \mathcal{B}_s(t) \tag{5}$$

where the coordinates $\theta_{r,j}^0, k = 1, \dots, q+k_1$ are associated to the spline approximation of the mean of the jth phenological curve and $\delta_{s,i}^j$ are the variations from the mean of the trajectory of the pixel i. By hypothesis, $E(\delta_{s,i}^j) = 0$ and we can approximate the mean and the variance of a discretized trajectory $\mathbf{Y}_i$ as follows:

$$\mathrm{E}\left(\mathbf{Y}_i | \boldsymbol{\pi}_i\right) \quad = \quad \mathbf{B}\boldsymbol{\theta}^0 \boldsymbol{\pi}_i' \tag{6}$$

$$\mathrm{Var}\left(\mathbf{Y}_i | \boldsymbol{\pi}_i\right) \quad = \quad \sigma^2 \mathbf{I}_p + \sum_{j=1}^{J} \pi_{ij}^2 \sum_{s,\ell=1}^{q+k_2} \mathrm{Cov}(\delta_{s,i}^j, \delta_{\ell,i}^j) \boldsymbol{\mathcal{B}}_s \boldsymbol{\mathcal{B}}_\ell' \tag{7}$$

where $\mathbf{B}$ is the $p \times (q+k_1)$ matrix whose elements are $[\mathbf{B}]_{\nu,r} = B_r(t_\nu)$, $\boldsymbol{\theta}^0$ is the $(q+k_1) \times J$ matrix whose elements are $[\boldsymbol{\theta}^0]_{r,j} = \theta^0_{r,j}$ and $\boldsymbol{\mathcal{B}}_s = (\mathcal{B}_s(t_1), \ldots, \mathcal{B}_s(t_p))' \in \mathbb{R}^p$.

3 Maximum likelihood estimates

Denote by $\mathbf{V}_i$ the variance of the discretely sampled trajectory $\mathbf{Y}_i$, it is defined in equation (7).

The log-likelihood, equals, up to a constant

$$\mathcal{L} = -\frac{1}{2}\left(\sum_{i=1}^n \log|\mathbf{V}_i| + \sum_{i=1}^n \left(\mathbf{Y}_i - \mathbf{B}\boldsymbol{\theta}^0\boldsymbol{\pi}_i\right)' \mathbf{V}_i^{-1}\left(\mathbf{Y}_i - \mathbf{B}\boldsymbol{\theta}^0\boldsymbol{\pi}_i\right)\right) \quad (8)$$

and the parameters to be estimated are σ^2, the $(q+k_1) \times J$ matrix $\boldsymbol{\theta}^0$ and the J symmetric matrices $\boldsymbol{\gamma}^j$ of size $(q+k_2) \times (q+k_2)$ with $[\gamma^j]_{s,\ell} = \mathrm{Cov}(\delta^j_{s,i}, \delta^j_{\ell,i})$. This latter matrix is independent of i by assumption and has $(q+k_2)(q+k_2+1)/2$ non redundant parameters.

The estimation procedure proposed here is based on combination of the backfitting algorithm and a kind of EM algorithm [7], called ECME which is known to converge faster than the classical EM algorithm [8]. Let us also notice that one major advantage of this approach compared to direct optimization procedures is that the estimated covariance matrices are automatically non negative.

The variance parameters are known

Suppose first that the variance components $\boldsymbol{\gamma}^j, j = 1, \ldots, J$ and σ^2 of the model are known and let us give estimators of the fixed effects.

As in Hoover et al. [4] and Cardot et al. [1], a penalty, tuned by a smoothing parameter λ_j, is added in order to get identifiability and smooth estimators of the mean characteristic curves,

$$J(\lambda, \rho_j) = \lambda_j \times \int_T \left(\frac{d^2\rho_j(u)}{dt^2}\right)^2 du = \lambda_j \; \boldsymbol{\theta}_j'\mathbf{P}_2\boldsymbol{\theta}_j \;, \quad (9)$$

where $[\mathbf{P}_2]_{\ell,r} = \int_T B_\ell^{(2)}(t)B_r^{(2)}(t)dt$ and $B_r^{(2)}(t)$ is the second order derivative of $B_r(t)$. Then maximizing the log likelihood (8) with penalty terms defined in (9) is equivalent to minimize the following criterion:

$$\sum_{i=1}^n \left\|\mathbf{Y}_i - \sum_{j=1}^J \pi_{ij}\mathbf{B}\boldsymbol{\theta}_j\right\|^2_{\mathbf{V}_i^{-1}} + \frac{1}{2}\sum_{j=1}^J \lambda_j \boldsymbol{\theta}_j'\mathbf{P}_2\boldsymbol{\theta}_j, \quad (10)$$

where $\|\mathbf{Y}_i\|^2_{\mathbf{V}_i^{-1}} = \mathbf{Y}_i'\mathbf{V}_i^{-1}\mathbf{Y}_i$. Noticing that minimizing 10) is equivalent to solve the set of equations in $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J$

$$\sum_{j'=1}^J \left(\sum_{i=1}^n \pi_{ij}\pi_{ij'}\mathbf{B}'\mathbf{V}_i^{-1}\mathbf{B}\right)\boldsymbol{\theta}_{j'} + \lambda_j\mathbf{P}_2\boldsymbol{\theta}_j \;=\; \sum_{i=1}^n \pi_{ij}\mathbf{B}'\mathbf{V}_i^{-1}\mathbf{Y}_i, \quad (11)$$

for $j = 1, \dots, J$, allows to save computation time by using the backfitting algorithm. It avoids inverting a large $J(q + k_1) \times J(q + k_1)$ matrix. Let us remark that computation time can be saved again by considering the following transformed data, $\widetilde{\mathbf{Y}}_i = \mathbf{B}'\mathbf{Y}_i$. Then we have to deal with $(q+k_1) \times (q+k_1)$ individual variance matrices $\mathbf{W}_i = \mathbf{B}'\mathbf{V}_i\mathbf{B}$ with $q + k_1 < p$.

Variance parameters estimation

Define $\boldsymbol{\delta}_i^j = (\delta_{1,i}^j, \dots, \delta_{q+k_2,i}^j)'$. The individual components $\widehat{\boldsymbol{\delta}}_i^j = E\left(\boldsymbol{\delta}_i^j | \mathbf{Y}_i\right)$ and $\widehat{\boldsymbol{\varepsilon}}_i = E\left(\boldsymbol{\varepsilon}_i \mid \mathbf{Y}_i\right)$ can be deduced using the BLUP formula

$$\widehat{\boldsymbol{\delta}}_i^j = \pi_{ij}\boldsymbol{\gamma}^j\boldsymbol{\mathcal{B}}'\mathbf{V}_i^{-1}\left(\mathbf{Y}_i - \sum_{j=1}^{J}\pi_{ij}\mathbf{B}\widehat{\boldsymbol{\theta}}_j^0\right), \tag{12}$$

$$\widehat{\boldsymbol{\varepsilon}}_i = \mathbf{Y}_i - \sum_{j=1}^{J}\pi_{ij}\left(\mathbf{B}\widehat{\boldsymbol{\theta}}_j^0 + \boldsymbol{\mathcal{B}}\widehat{\boldsymbol{\delta}}_i^j\right). \tag{13}$$

Using these expressions we can get estimates of the variance components

$$\begin{aligned} n\,\widehat{\gamma}^j &= E\left(\sum_{i=1}^{n}\widehat{\boldsymbol{\delta}}_i^j(\widehat{\boldsymbol{\delta}}_i^j)' \mid \mathbf{Y}_i\right) \\ &= \sum_{i=1}^{n}\left\{\widehat{\boldsymbol{\delta}}_i^j(\widehat{\boldsymbol{\delta}}_i^j)' + \mathrm{Var}\left(\boldsymbol{\delta}_i^j \mid \mathbf{Y}_i\right)\right\}. \end{aligned} \tag{14}$$

The variance σ^2 of the noise is estimated by

$$\begin{aligned} np\,\widehat{\sigma}^2 &= E\left(\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_i \mid \mathbf{Y}_i, \widehat{\gamma}\right) \\ &= \sum_{i=1}^{n}\left\{\widehat{\boldsymbol{\varepsilon}}_i'\widehat{\boldsymbol{\varepsilon}}_i + \mathrm{tr}\ \mathrm{Var}\left(\boldsymbol{\varepsilon}_i \mid \mathbf{Y}_i\right)\right\}. \end{aligned} \tag{15}$$

Formulas for the expected conditional variances are given in McLachlan and Krishnan [8].

The final algorithm consists in repeating estimation of the fixed effects (11), individual effects (12), (13) and estimations of the variance terms (14), (15) until convergence.

Once the parameters of the model have been estimated, it is easy to get, for every pixel i, the local responses of the different crops applying the BLUP formula (12).

4 An application in remote sensing

We have remote sensing data obtained with the Végétation sensor in the South-East of France. Each pixel correspond to an area of 1 km^2 and we

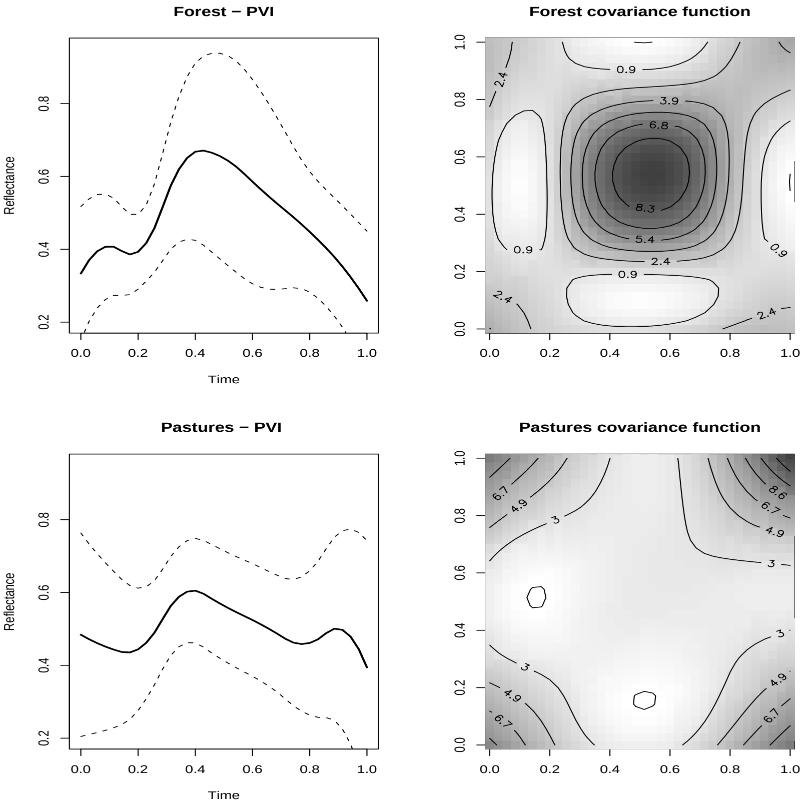


Figure 1: Estimated phenological curves and "normalized" covariance functions for the themes "Forests" and "Pastures". Dotted lines correspond to the mean function ± 1.96 times the standard deviation functions, defined as $\sqrt{\widehat{\gamma}^j(t,t)}$.

observe this region during the year 2002. We have $p = 36$ distinct time instants and $n = 1209$ pixels, corresponding to total area of 1209 km^2.

In this region, the land use is also known thanks the Corine land cover map. We selected the 7th most important themes (forest, pastures, urban, ...) so that we have $J = 8$ classes. We choose B-splines bases of order $q = 3$, with $k_1 = 6$ and $k_2 = 3$ interior knots. The smoothing parameter value is the same for all the phenological curves, it is $\lambda = 10^{-4}$. We deal with the PVI index [10], it is a linear combination of the reflectances in the Red and Near-Infra-Red channels.

It took less than one minute on a PC for the algorithm to converge. The estimated variance of the noise is $\widehat{\sigma}^2 = 0.0022$ whereas the variance of the functions Y_i is about 0.014.

We have drawn in Figure 1 the estimated phenological curves of the themes “Forest” and “Pastures” with their estimated variability as well as discretized estimates of their “normalized” covariance operator defined by $\widehat{\gamma}^j(s,t) = \boldsymbol{\mathcal{B}}'(t)\widehat{\boldsymbol{\gamma}}^j\boldsymbol{\mathcal{B}}(s)$ with $\boldsymbol{\mathcal{B}}'(t) = (\mathcal{B}_1(t), \ldots, \mathcal{B}_{q+k_2}(t))$. Each cell is an estimation of the “normalized” covariance kernel $\widehat{\gamma}^j(s,t)/\widehat{\sigma}^2$ for $s,t \in [0,1]$ and $j \in \{$“forest”,”pastures”$\}$.

Estimated covariance operators can be really different from one class to another as shown in Figure 1. This confirms that different crops are characterized not only by their mean reflectance along time but also by their variations. The largest variations for the theme “Forest” occur during the middle of the year, when the response is close to its maximum, with a strong temporal correlation whereas this is not the case at all for the theme “Pastures”. Furthermore, an eigenanalysis of the estimated covariance operators allows us to exhibit the main modes of variations of the individual responses.

5 Discussion

First notice that we can extend without real difficulties, as in James [5], our estimation procedure with measurements points (sampled time instants) that differ from one pixel to another.

In practice one has to choose the number of knots for the spline basis and values of the smoothing parameters. We prefer a criterion like the AIC criterion compared to cross validation for computation time reasons. James et al. [5] noticed that results were nearly equivalent in a similar study.

Another potential application of this model is land use prediction [1], taking into account the temporal structure of the variability of the crop responses. Indeed, once the parameters have been estimated one can predict the land use in a similar area (assuming the crops have the same modes of variations), that is to say the vector $\boldsymbol{\pi}_{i'}$ of a new pixel i' when observing the trajectory $\mathbf{Y}_{i'} = (Y_{i'}(t_1), \ldots, Y_{i'}(t_p))$. This can be done by maximizing the likelihood criterion, which is equivalent to minimize:

$$\min_{\boldsymbol{\pi}} \ \log|\widehat{\mathbf{V}}_{i'}| + \left(\mathbf{Y}_{i'} - \mathbf{B}\widehat{\boldsymbol{\theta}}^0\boldsymbol{\pi}\right)' \widehat{\mathbf{V}}_{i'}^{-1} \left(\mathbf{Y}_{i'} - \mathbf{B}\widehat{\boldsymbol{\theta}}^0\boldsymbol{\pi}\right) \tag{16}$$

under the constraints that $\pi_j \geq 0$ and $\sum_j \pi_j = 1$, where $\widehat{\boldsymbol{\theta}}^0$ is deduced from the previous estimation procedures and $\widehat{\mathbf{V}}_{i'} = \widehat{\sigma}^2\mathbf{I}_p + \sum_{j=1}^J \pi_j^2 \sum_{s,\ell} \widehat{\gamma}_{s,\ell}^j \boldsymbol{\mathcal{B}}_s\boldsymbol{\mathcal{B}}_\ell'$.

Finally, as it was pointed out by an anonymous referee, we can reasonably think that the spatio-temporal structure of the data is not fully reflected with our model. Incorporating spatial structures would certainly improves the quality of our estimates. It deserves further attention but is beyond the scope of this paper.

References

[1] Cardot H., Faivre R., Goulard M. (2003). *Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data.* J. of Applied Statistics **30**, 1185 – 1999.

[2] Dierckx, P. (1993). *Curve and surface fitting with splines.* Clarendon Press, Oxford.

[3] Faivre R., Fischer A. (1997). *Predicting crop reflectances using satellite data observing mixed pixels.* J. of Agricultural, Biological and Environmental Statistics **2**, 87 – 107.

[4] Hoover D.R., Rice J.A., Wu C.O., Yang L.P. (1998). *Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data.* Biometrika **85**, 809 – 822.

[5] James, G., Hastie, T., and Sugar, C. (2000). *Principal Component Models for Sparse Functional Data.* Biometrika, **87** , 587 – 602.

[6] Kneip A., Sickles R., Song W. (2003). *On estimating the mixed effects model.* Preprint.

[7] Laird, N. and Ware, J. (1982). *Random-effects models for longitudinal data.* Biometrics, **38**, 963 – 974.

[8] McLachlan G., Krishnan T. (1997). *The EM algorithm and extensions.* John Wiley & Sons.

[9] Rice J., Wu C. (2001). *Nonparametric mixed effects models for unequally sampled noisy curves.* Biometrics **57**, 253 – 259.

[10] Richardson A.J., Wiegang C.L. (1977). *Distinguishing vegetation from soil background information.* Photogrammetric Engineering and Remote Sensing **43**, 1541 – 1552.

[11] Tucker C.J. (1979). *Red and photographic infrared linear combinations for monitoring vegetation.* Remote Sensing of Environment **8**, 127 – 150.

Acknowledgement: We would like to thank J.L. Champeaux (Météo-France) for providing us the data and the Staph working group in Toulouse for fruitful discussions. We also thank an anonymous referee for his very constructive remarks.

Address: H. Cardot, R. Faivre, INRA Toulouse, Unité Biométrie et Intelligence Artificielle, 31326 Castanet-Tolosan, France
P. Maisongrande, Centre d'Etudes Spatiales de la Biosphère, UMR CNES-CNRS, 18 Av. Edouard Belin, 31401 Toulouse, France

E-mail: `cardot@toulouse.inra.fr`

CHEMICAL BALANCE WEIGHING DESIGNS FOR $V+1$ OBJECTS WITH DIFFERENT VARIANCES BASED ON TERNARY BALANCED BLOCK DESIGNS

Bronisław Ceranka and Małgorzata Graczyk

Key words: Optimum chemical balance weighing design, ternary balanced block design.

COMPSTAT 2004 section: Statistical software.

Abstract: The paper is studying the estimation problem of individual weights of objects using the chemical balance weighing design under the restriction on the number of times in which each object is weighed. We assume that the errors are uncorrelated and they have different variances. The conditions under which the existence of the optimum chemical balance weighing design for $p=v$ objects implies the existence of the optimum chemical balance weighing design for $p=v+1$ objects are given. The new construction methods for the optimum chemical balance weighing design for $p=v+1$ objects are given. We use the incidence matrices of the ternary balanced block designs for v treatments to construct the design matrix of the optimum chemical balance weighing designs for $p=v+1$ objects.

1 Introduction

The results of n weighing operations to determine the individual weights of p objects with a balance that is corrected for bias will fit into the linear model

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is an $n\times 1$ random observed vector of the recorded results of weights, $\mathbf{X} \in \Phi_{n\times p,m}(-1,0,1)$, where $\Phi_{n\times p,m}(-1,0,1)$ denotes the class of the $n\times p$ matrices with elements $x_{ij}=-1$, 1 or 0 if the jth object is kept on the left pan, right pan or is not included in the particularly weighing, respectively, $i=1,2,\ldots,n,\quad j=1,2,\ldots,p$. Let m be the maximum number of elements equal to -1 and 1 in the columns of the matrix $\mathbf{X}$, $\mathbf{w}$ is a $p\times 1$ column vector representing unknown weights of objects and $\mathbf{e}$ is an $n\times 1$ random vector of errors. We assume that $E(\mathbf{e})=\mathbf{0}_n$ and $Var(\mathbf{e})=\sigma^2\mathbf{G}$, where $\mathbf{0}_n$ is the $n\times 1$ column vector of zeros, $\mathbf{G}$ is the $n\times n$ positive definite diagonal matrix of known elements, $E(\cdot)$ stands for the expectation and $\mathbf{e}^{'}$ is used for transpose of $\mathbf{e}$.

The normal equations estimating $\mathbf{w}$ are of the form

$$\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{y}, \tag{2}$$

where $\hat{\mathbf{w}}$ is the vector of the weights estimated by the least squares method.

The chemical balance weighing design is singular or nonsingular depending on whether the matrix $\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X}$ is singular or nonsingular, respectively. It is obvious that because of the assumption connected with the matrix $\mathbf{G}$ the matrix $\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X}$ is nonsingular if and only if the matrix $\mathbf{X}^{'}\mathbf{X}$ is nonsingular, i.e. if and only if $\mathbf{X}$ is of full column rank ($= p$). If $\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X}$ is nonsingular, the least squares estimator of $\mathbf{w}$ is given in the form

$$\hat{\mathbf{w}} = \left(\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{y} \tag{3}$$

and the variance - covariance matrix of $\hat{\mathbf{w}}$ is given by formula

$$Var(\hat{\mathbf{w}}) = \sigma^2\left(\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X}\right)^{-1}. \tag{4}$$

Various aspects of the chemical balance weighing designs have been studied by Raghavarao [8] and Banerjee [1]. When $\mathbf{G}$ is the positive definite diagonal matrix of known elements Katulska [7] have showed that the minimum attainable variance of each of the estimated weights for a chemical balance weighing design is $\sigma^2/tr(\mathbf{G}^{-1})$. She proved the theorem that each of the variances of the estimated weights attaines the minimum if and only if $\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X} = tr(\mathbf{G}^{-1})\mathbf{I}_p$. This design is said to be the optimum chemical balance weighing design. In this case several methods of construction of the optimum chemical balance weighing designs are given in the literature. But the optimality condition implies that the elements of the design matrix $\mathbf{X}$ are equal to -1 and 1, only.

In present paper we consider the generalisation of the problem of choosing the matrix $\mathbf{X}$ of the optimum chemical balance weighing design with nonhomogeneity of variances of errors. We assume that in each column of the design matrix $\mathbf{X}$ are elements equal to 0, i.e. in each measurement operation not all objects are included. We investigate the necessary and sufficient conditions under which the minimum variance is attained for estimated weights. We give new methods of construction of the optimum chemical balance weighing designs for $p = v + 1$ objects. They are based on the incidence matrices of the ternary balanced block designs for $p = v$ treatments.

In a special case when $\mathbf{G} = \mathbf{I}_n$, the methods of construction of the optimum chemical balance weghing designs for $p = v$ objects with homogeneity of variances of errors based on the same set of the incidence matrices were given in Ceranka and Graczyk [4].

2 Variance limit of estimated weights

Let $\mathbf{X}_h \in \Phi_{n_h \times p, m_h}(-1, 0, 1)$ be the $n_h \times p$ matrix of the rank p of the chemical balance weighing design, where $m_h = max(m_{h1}, m_{h2}, \ldots, m_{hp})$, m_{hj} is the number of elements equal to -1 and 1 in the jth column of the matrix $\mathbf{X}_h$, $\quad h = 1, 2$.

Now, we define the matrix $\mathbf{X} \in \Phi_{n\times p,m}(-1,0,1)$ of the chemical balance weighing design as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \tag{5}$$

where $n = n_1 + n_2$ and $m = m_1 + m_2$.
Let

$$\mathbf{G} = \begin{bmatrix} \frac{1}{a}\mathbf{I}_{n_1} & \mathbf{0}_{n_1}\mathbf{0}'_{n_2} \\ \mathbf{0}'_{n_2}\mathbf{0}_{n_1} & \mathbf{I}_{n_2} \end{bmatrix}. \tag{6}$$

Ceranka and Graczyk [5] showed that the minimum attainable variance for each of the estimated weights for the chemical balance weighing design with the design matrix $\mathbf{X}$ in the form (5) with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given in (6), is $\sigma^2/(am_1 + m_2)$, i.e. $Var(\hat{w}_j) \geq \sigma^2/(am_1 + m_2)$, $j = 1, 2, \ldots, p$.

Definition 1. *Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X}$ in the form (5) with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given in (6), is called the optimal design for the estimation of individual weights of the objects if $Var(\hat{w}_j) = \sigma^2/(am_1+m_2)$, $j = 1, 2, \ldots, p$.*

Ceranka and Graczyk [5] proved the following theorem

Theorem 1. *Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X}$ in the form (5) with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given in (6), is optimal if and only if*

$$\mathbf{X}'\mathbf{G}^{-1}\mathbf{X} = (am_1 + m_2)\mathbf{I}_p. \tag{7}$$

In the particular case when $m_1 = n_1$ and $m_2 = n_2$ the theorem was given in Katulska [7] and when additionally $\mathbf{G} = \mathbf{I}_n$ the theorem was given in Hotelling [6].

3 Optimum chemical balance weghing designs for $p + 1$ objects

Let $\mathbf{X} \in \Phi_{n\times p,m}(-1,0,1)$ be the matrix of the chemical balance weighing design in the form (5). Based on that matrix we form the design matrix $\mathbf{X} \in \Phi_{n\times(p+1),m}(-1,0,1)$ of the chemical balance weighing design for $p + 1$ objects and we receive

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{n_1} \\ \mathbf{X}_2 & \mathbf{0}_{n_2} \end{bmatrix}, \tag{8}$$

where $\mathbf{1}_{n_1}$ is the $n_1 \times 1$ column vector of units.

Theorem 2. *If $\mathbf{X}_h$ is the matrix of the $n_h \times p$ chemical balance weighing design, $h = 1, 2$, then the $n \times (p + 1)$ matrix $\mathbf{X}$ given in the form (8) is the matrix of the optimum chemical balance weighing design with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given in (6), if and only if*

$$\mathbf{X}_1^{'}\mathbf{1}_{n_1} = \mathbf{0}_p$$

and

$$am_1 + m_2 = an_1.$$

Proof. For the design matrix $\mathbf{X}$ given in (8) and the matrix $\mathbf{G}$ in the form (6) we have

$$\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X} = \begin{bmatrix} a\mathbf{X}_1^{'}\mathbf{X}_1 + \mathbf{X}_2^{'}\mathbf{X}_2 & a\mathbf{X}_1^{'}\mathbf{1}_{n_1} \\ a\mathbf{1}_{n_1}^{'}\mathbf{X}_1 & an_1 \end{bmatrix}. \tag{9}$$

Then from (7) and (9) we have $\mathbf{X}_1^{'}\mathbf{1}_{n_1} = \mathbf{0}_p$ and $a\mathbf{X}_1^{'}\mathbf{X}_1 + \mathbf{X}_2^{'}\mathbf{X}_2 = an_1\mathbf{I}_p$ which complet the proof.

4 Ternary balanced block designs

A ternary balanced block design is defined as the design consisting of b blocks, each of size k, chosen from a set of objects of size v, in such a way that each of the v treatments occurs r times altogether and 0, 1 or 2 times in each block, (2 appears at least once) and each of the distinct pairs appears λ times. Any ternary balanced block design is regular, that is, each treatment occurs alone in ρ_1 blocks and is repeated two times in ρ_2 blocks, where ρ_1 and ρ_2 are constant for the design. Let $\mathbf{N}$ be the incidence matrix of the ternary balanced block design. It is straightforward to verify that

$$\begin{aligned}
&vr = bk,\\
&r = \rho_1 + 2\rho_2,\\
&\lambda(v-1) = \rho_1(k-1) + 2\rho_2(k-2) = r(k-1) - 2\rho_2,\\
&\mathbf{NN'} = (\rho_1 + 4\rho_2 - \lambda)\mathbf{I}_v + \lambda\mathbf{1}_v\mathbf{1}_v' = (r + 2\rho_2 - \lambda)\mathbf{I}_v + \lambda\mathbf{1}_v\mathbf{1}_v'.
\end{aligned}$$

5 Construction of the design matrix

Let $\mathbf{N}_h$ be the incidence matrix of the ternary balanced block design with the parameters $v,\quad b_h,\quad r_h,\quad k_h,\quad \lambda_h,\quad \rho_{1h},\quad \rho_{2h},\quad h = 1, 2$. Now, we define the matrix $\mathbf{X}_h$ as $\mathbf{X}_h = \mathbf{N}_h^{'} - \mathbf{1}_{b_h}\mathbf{1}_v^{'}$. Then the matrix $\mathbf{X} \in \Phi_{n\times(p+1),m}(-1,0,1)$ is of the form

$$\mathbf{X} = \begin{bmatrix} \mathbf{N}_1^{'} - \mathbf{1}_{b_1}\mathbf{1}_v^{'} & \mathbf{1}_{b_1} \\ \mathbf{N}_2^{'} - \mathbf{1}_{b_2}\mathbf{1}_v^{'} & \mathbf{0}_{b_2} \end{bmatrix}. \tag{10}$$

In such a design we determine unknown measurements of $p = v + 1$ objects. Thus, each of v first columns of the matrix $\mathbf{X}$ will contain $\rho_{21} + \rho_{22}$ elements equal to 1, $b_1 + b_2 - \rho_{11} - \rho_{12} - \rho_{21} - \rho_{22}$ elements eqaul to -1 and $\rho_{11} + \rho_{12}$ elements eqaul to 0. The last column of $\mathbf{X}$ will contain b_1 elements equal to 1 and b_2 elements equal to zero. Clearly, such a design implies that in $n = b_1 + b_2$ measurement operations the jth object is weighed $b_1 + b_2 - \rho_{11} - \rho_{12}$ times, $j = 1, 2, \ldots, v$, and the $(v + 1)$th object is weighed b_1 times.

Lemma 1. *The chemical balance weighing design* $\mathbf{X}$ *in the form (9) is nonsingular if and only if* $v \neq k_2$.
Proof. Since $\mathbf{G}$ is the positive definite diagonal matrix of known elements then $\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X}$ is nonsingular if and only if $\mathbf{X}^{'}\mathbf{X}$ is nonsingular. For the design matrix $\mathbf{X}$ given in (9) we have

$$\mathbf{X}^{'}\mathbf{X} = \begin{bmatrix} \theta\mathbf{I}_v + \gamma\mathbf{1}_v\mathbf{1}_v^{'} & (r_1 - b_1)\mathbf{1}_v \\ (r_1 - b_1)\mathbf{1}_v^{'} & b_1 \end{bmatrix},$$

where $\theta = r_1 + 2\rho_{21} - \lambda_1 + r_2 + 2\rho_{22} - \lambda_2$, $\quad \gamma = b_1 + \lambda_1 - 2r_1 + b_2 + \lambda_2 - 2r_2$. It is easy to see that

$$det(\mathbf{X}^{'}\mathbf{X}) = \frac{b_1 b_2}{v}(v - k_2)^2 \left(r_1 + 2\rho_{21} - \lambda_1 + r_2 + 2\rho_{22} - \lambda_2\right)^{v-1}.$$

Evidently $r_1 + 2\rho_{21} - \lambda_1 + r_2 + 2\rho_{22} - \lambda_2$ is positive, hence $det(\mathbf{X}^{'}\mathbf{X}) \neq 0$ if and only if $v \neq k_2$. So, the lemma is proved.

From the theorems (1) and (2) we have

Theorem 3. *Any nonsingular chemical balance weighing design with the design matrix* $\mathbf{X}$ *given in the form (10) with the variance-covariance matrix of errors* $\sigma^2\mathbf{G}$, *where* $\mathbf{G}$ *is of the form (6), is optimal if and only if*

$$a\,(b_1 + \lambda_1 - 2r_1) + (b_2 + \lambda_2 - 2r_2) = 0, \tag{11}$$

$$b_1 = r_1 \tag{12}$$

and

$$b_2 = a\rho_{11} + \rho_{12}. \tag{13}$$

Proof. For the design matrix $\mathbf{X}$ given in (10) and the matrix $\mathbf{G}$ in the form (6) we have

$$\mathbf{X}^{'}\mathbf{G}^{-1}\mathbf{X} = \begin{bmatrix} \delta\mathbf{I}_v + \eta\mathbf{1}_v\mathbf{1}_v^{'} & a(r_1 - b_1)\mathbf{1}_v \\ a(r_1 - b_1)\mathbf{1}_v^{'} & ab_1 \end{bmatrix},$$

where $\delta = a(r_1 + 2\rho_{21} - \lambda_1) + (r_2 + 2\rho_{22} - \lambda_2)$, $\quad \eta = a(b_1 + \lambda_1 - 2r_1) + (b_2 + \lambda_2 - 2r_2)$. The design matrix $\mathbf{X}$ in the form (10) is optimal if and only if the conditions given in the theorems (1) and (2) are fulfieled. From these conditions and relations between parameters we receive the thesis of theorem.

If the chemical balance weighing design given by the matrix $\mathbf{X}$ in the form (10) with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is of the form (6), is optimal then

$$Var(\hat{w_j}) = \frac{\sigma^2}{ab_1}, \qquad j = 1, 2, \ldots, v + 1.$$

We have seen in theorem (3) that if parameters of two ternary balanced block designs satisfy the conditions (11), (12) and (13) then the chemical balance weighing design with the design matrix $\mathbf{X}$ given in the form (10) with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where the matrix $\mathbf{G}$ is of the form (6), is optimal. Under these conditions we have formulated the theorems following from the papers of Billington [2] and Billington and Robinson [3].

Theorem 4. *Let $a = \frac{1}{2}$. If the parameters of two ternary balanced block designs are equal to*

(i) $v = 5$, $b_1 = 8(s+2)$, $r_1 = 8(s+2)$, $k_1 = 5$, $\lambda_1 = 2(4s+7)$, $\rho_{11} = 8(s+1)$, $\rho_{21} = 4$ *and* $v = 5$, $b_2 = 5(s+2)$, $r_2 = 3(s+2)$, $k_2 = 3$, $\lambda_2 = s+3$, $\rho_{12} = s+6$, $\rho_{22} = s$, $s = 1, 2, \ldots,$

(ii) $v = 9$, $b_1 = 6(s+4)$, $r_1 = 6(s+4)$, $k_1 = 9$, $\lambda_1 = 2(3s-4)$, $\rho_{11} = 2(3s+4)$, $\rho_{21} = 8$ *and* $v = 9$, $b_2 = 3(s+4)$, $r_2 = 2(s+4)$, $k_2 = 6$, $\lambda_2 = s+5$, $\rho_{12} = 8$, $\rho_{22} = s$, $s = 2, 3, \ldots,$

(iii) $v = 11$, $b_1 = 32$, $r_1 = 32$, $k_1 = 11$, $\lambda_1 = 30$, $\rho_{11} = 12$, $\rho_{21} = 10$ *and* $v = 11$, $b_2 = 11$, $r_2 = 7$, $k_2 = 7$, $\lambda_2 = 4$, $\rho_{12} = 5$, $\rho_{22} = 1$,

(iv) $v = 12$, $b_1 = 8(2s+5)$, $r_1 = 8(2s+5)$, $k_1 = 12$, $\lambda_1 = 2(8s+19)$, $\rho_{11} = 2(8s+9)$, $\rho_{21} = 11$ *and* $v = 12$, $b_2 = 3(2s+5)$, $r_2 = 2(2s+5)$, $k_2 = 8$, $\lambda_2 = 2(s+3)$, $\rho_{12} = 6-2s$, $\rho_{22} = 3s+2$, $s = 0, 1, 2,$

(v) $v = 15$, $b_1 = 10(s+4)$, $r_1 = 10(s+4)$, $k_1 = 15$, $\lambda_1 = 2(5s+19)$, $\rho_{11} = 2(5s{+}6)$, $\rho_{21} = 14$ *and* $v = 15$, $b_2 = 3(s{+}4)$, $r_2 = 2(s+4)$, $k_2 = 10$, $\lambda_2 = s+5$, $\rho_{12} = 6-2s$, $\rho_{22} = 2s+1$, $s = 1, 2,$

then the chemical balance weighing design with the design matrix $\mathbf{X}$ given in the form (10) with the variance-covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given in (6), is optimal.

Proof. It is easy to see that the parameters of the ternary balanced block designs satisfied the equalities (11)-(13).

Theorem 5. *Let $a = \frac{3}{2}$. If the parameters of two ternary balanced block designs are equal to*

(i) $v = 9$, $b_1 = 32$, $r_1 = 32$, $k_1 = 9$, $\lambda_1 = 30$, $\rho_{11} = 16$, $\rho_{21} = 8$ *and* $v = 9$, $b_2 = 27$, $r_2 = 15$, $k_2 = 5$, $\lambda_2 = 6$, $\rho_{12} = 3$, $\rho_{22} = 6$,

(ii) $v = 12$, $b_1 = 28$, $r_1 = 28$, $k_1 = 12$, $\lambda_1 = 26$, $\rho_{11} = 6$, $\rho_{21} = 11$ *and* $v = 12$, $b_2 = 14$, $r_2 = 7$, $k_2 = 6$, $\lambda_2 = 3$, $\rho_{12} = 5$, $\rho_{22} = 1$,

(iii) $v = 15$, $b_1 = 40$, $r_1 = 40$, $k_1 = 15$, $\lambda_1 = 38$, $\rho_{11} = 12$, $\rho_{21} = 14$ *and* $v = 15$, $b_2 = 25$, $r_2 = 15$, $k_2 = 9$, $\lambda_2 = 8$, $\rho_{12} = 7$, $\rho_{22} = 4$,

then the chemical balance weighing design with the design matrix $\mathbf{X}$ *given in the form (10) with the variance-covariance matrix of errors* $\sigma^2\mathbf{G}$, *where* $\mathbf{G}$ *is given in (6), is optimal.*

Theorem 6. *Let* $a = 2$. *If the parameters of the ternary balanced block designs are equal to* $v = 5$, $b_1 = 2(s+4)$, $r_1 = 2(s+4)$, $k_1 = 5$, $\lambda_1 = 2s+7$, $\rho_{11} = 2(s+2)$, $\rho_{21} = 2$ *and* $v = 5$, $b_2 = 5(s+4)$, $r_2 = 3(s+4)$, $k_2 = 3$, $\lambda_2 = s+6$, $\rho_{12} = s+12$, $\rho_{22} = s$, $s = 1, 2, \ldots$, *then the chemical balance weighing design with the design matrix* $\mathbf{X}$ *given in the form (10) with the variance-covariance matrix of errors* $\sigma^2\mathbf{G}$, *where* $\mathbf{G}$ *is given in (6), is optimal.*

Theorem 7. *Let* $a = 3$. *If the parameters of two ternary balanced block designs are equal to*

(i) $v = 9$, $b_1 = 16$, $r_1 = 16$, $k_1 = 9$, $\lambda_1 = 15$ $\rho_{11} = 8$, $\rho_{21} = 4$ *and* $v = 9$, $b_2 = 27$, $r_2 = 15$, $k_2 = 5$, $\lambda_2 = 6$, $\rho_{12} = 3$, $\rho_{22} = 6$,

(ii) $v = 15$, $b_1 = 20$, $r_1 = 20$, $k_1 = 15$, $\lambda_1 = 19$ $\rho_{11} = 6$, $\rho_{21} = 7$ *and* $v = 15$, $b_2 = 25$, $r_2 = 15$, $k_2 = 9$, $\lambda_2 = 8$, $\rho_{12} = 7$, $\rho_{22} = 4$,

then the chemical balance weighing design with the design matrix $\mathbf{X}$ *given in the form (10) with the variance-covariance matrix of errors* $\sigma^2\mathbf{G}$, *where* $\mathbf{G}$ *is given in (6), is optimal.*

References

[1] Banerjee K.S. (1975). *Weighing designs for chemistry, medicine, economics, operations research, statistics.* Marcel Dekker Inc., New York, 1975.

[2] Billington E. J. (1984). *Balanced n-array designs: a combinatorial survey and some new results.* Ars Combin. **17**A, 37 – 72.

[3] Billington E.J., Robinson P.J. (1983). *A list of balanced ternary block designs with* $r \leq 15$ *and some necessary existence conditions.* Ars Combin. **16**, 235 – 258.

[4] Ceranka B., Graczyk M. (2003a). *Optimum chemical balance weighing design for* $v + 1$ *objects.* Kybernetika **39**, 333 – 340.

[5] Ceranka B. and Graczyk M. (2003b). *Optimum chemical balance weighing designs.* Tatra Mountains Mathematical Publications **26**, 49 – 57.

[6] Hotelling H. (1944). *Some improvements in weighing designs and other experimental techniques*. Ann. Math. Stat. **15**, 297 − 305.

[7] Katulska K. (1989). *Optimum chemical balance weighing designs with non-homogeneity of the variances of errors*. J. Japan Statist. Soc. **19**, 95 − 101.

[8] Raghavarao D. (1971). *Constructions and combinatorial problems in designs of experiments.* John Wiley Inc., New York.

Address: B. Ceranka, M. Graczyk, Department of Mathematical and Statistical Methods, Agricultural University of Poznań, ul. Wojska Polskiego 28, 60-637 Poznań, Poland

E-mail: `bronicer@owl.au.poznan.pl, magra@owl.au.poznan.pl`

A COMPARISON OF TWO METHODS OF PRINCIPAL COMPONENT ANALYSIS

Vartan Choulakian

Key words: Transposition invariance; PCA in L$_1$;PCA; Banach spaces; robustness.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: An invariance property of the classical principal component analysis (PCA) is used to develop a new method of PCA in L$_1$. Using the theory of finite dimensional Banach spaces, we show that the mathematical framework of both methods is similar. The new method is robust compared to the classical PCA, and produces simultaneous dichotomies of the variables and the observations. We compare both methods on a real data set: The new method explains much better the underlying structure of the well known Hearing Loss Data.

1 Introduction

We start with some notation. Let $\mathbf{X}$ be a data set of dimension $n \times m$, where n observations are described by the m variables. Let $\mathbf{Y}$ represent the standardized data set, and $\mathbf{V} = \mathbf{Y}'\mathbf{Y}$ is the correlation or covariance matrix. The p-th vector norm of a vector $\mathbf{v} = (v_1, \ldots, v_n)'$ is defined to be $\|\mathbf{v}\|_p = (\sum_{i=1}^n \|v_i\|^p)^{1/p}$ for $p \geq 1$ and $\|\mathbf{v}\|_\infty = \max_i |v_i|$. Let l_p^n represent the finite n-dimensional Banach space equipped with the norm $\|.\|_p$, similarly we denote l_q^m. The matrix $\mathbf{Y}$ is considered an application from l_q^m to l_p^n. For $p \geq 1$ and $q \geq 1$, we define the matrix norm of $\mathbf{Y}$ to be

$$\|\mathbf{Y}\|_{pq} = \max_{\mathbf{v}} \frac{\|\mathbf{Y}\mathbf{v}\|_p}{\|\mathbf{v}\|_q}$$

In statistics, Galpin and Hawkins [7] proposed the use of $\|\mathbf{Y}\|_{11}$ and $\|\mathbf{Y}\|_{12}$ for L$_1$ estimation of a covariance matrix. Choulakian [4] developed the robust Q-mode principal component analysis (PCA) in L$_1$ based on $\|\mathbf{Y}'\|_{12}$, and Choulakian [5] showed that the centroid method is based on $\|\mathbf{Y}\|_{2\infty} = \|\mathbf{Y}'\|_{12}$. Also, Heiser [8], Benayade and Fichet [3], and, Baccini, Besse and de Falguerolles [2], proposed other formulations of PCA in L$_1$. The method developed in this paper differs from the other methods of PCA in L$_1$ found in the statistical literature by an invariance property: the objective function is transposition invariant with respect to the operator norm $\|\cdot\|_{1\infty}$, that is, $\|\mathbf{Y}\|_{1\infty} = \|\mathbf{Y}'\|_{1\infty}$. For this reason the method is named transposition invariant (TI) PCA in L$_1$. We note that the classical PCA is also transposition invariant and it is based on the norm $\|\mathbf{Y}\|_{22} = \|\mathbf{Y}'\|_{22}$.

The classical PCA is discussed quite in detail in two monographs written by Jackson [9] and Jolliffe [10].

This paper is organized as follows: Section 2 presents the main results, where the classical PCA and the new method are compared. In section 3, we analyze the Hearing Loss Data set, where we show that the TI-PCA in L_1 produces much clearer structure of the natural physical process of the hearing loss in adult males than the classical PCA. Finally, we conclude in section 4.

Here, we stress the fact that the exact principal component weights of TI-PCA in L_1 are calculated by an algorithm based on combinatorial optimization, equations 1 and 2 provided below, whose computational complexity is of the order $\mathrm{O}(\max(m,n)2^{\min(m,n)})$.

2 Main results

In the following we shall enumerate some similar mathematical properties of both methods: Classical PCA and TI-PCA in L_1. Proofs of the first five properties can be found in Choulakian [6].

- Variational definition:

  PCA

  $$\max_{\mathbf{v}} \|\mathbf{Yv}\|_2 \text{ subject to } \|\mathbf{v}\|_2 = 1.$$

  TI-PCA in L_1

  $$\max_{\mathbf{v}} \|\mathbf{Yv}\|_{\mathbf{1}} \text{ subject to } \|\mathbf{v}\|_\infty = 1.$$

- Duality:

  PCA

  $$\max_{\mathbf{u}} \|\mathbf{Y'u}\|_2 \text{ subject to } \|\mathbf{u}\|_{\mathbf{2}} = 1.$$

  TI-PCA in L_1

  $$\max_{\mathbf{u}} \|\mathbf{Y'u}\|_1 \text{ subject to } \|\mathbf{u}\|_\infty = 1.$$

- Bilinear application:

  PCA

  $$\max_{\mathbf{v},\mathbf{u}} \mathbf{u'Yv} \text{ subject to } \|\mathbf{u}\|_{\mathbf{2}} = 1 \text{ and } \|\mathbf{v}\|_2 = 1.$$

  TI-PCA in L_1

  $$\max_{\mathbf{v},\mathbf{u}} \mathbf{u'Yv} \text{ subject to } \|\mathbf{u}\|_\infty = 1 \text{ and } \|\mathbf{v}\|_\infty = 1.$$

- Computation of the first principal component weights $\mathbf{v}_1$ and $\mathbf{u}_1$, and the dispersion measure λ_1 :

PCA: Eigen value method

$$\mathbf{Y}'\mathbf{Y}\mathbf{v}_1 = \lambda_1^2\mathbf{v}_1 \quad \text{and} \quad \|\mathbf{v}\|_2 = 1.$$

$$\mathbf{Y}\mathbf{Y}'\mathbf{u}_1 = \lambda_1^2\mathbf{u}_1 \quad \text{and} \quad \|\mathbf{u}\|_2 = 1.$$

$$\lambda_1 = \mathbf{u}_1'\mathbf{Y}\mathbf{v}_1.$$

TI-PCA in L$_1$: Combinatorial optimization

$$\mathbf{v}_1 = \arg\max_{v_i=\pm 1} \|\mathbf{Y}\mathbf{v}\|_{\mathbf{1}} \quad \text{and} \quad \lambda_1 = \|\mathbf{Y}\mathbf{v}_1\|_{\mathbf{1}}. \tag{1}$$

$$\mathbf{u}_1 = \arg\max_{u_i=\pm 1} \|\mathbf{Y}'\mathbf{u}\|_{\mathbf{1}} \quad \text{and} \quad \lambda_1 = \|\mathbf{Y}'\mathbf{u}_1\|_{\mathbf{1}}. \tag{2}$$

$$\lambda_1 = \mathbf{u}_1'\mathbf{Y}\mathbf{v}_1.$$

- Transitional formulae:

PCA

$$\text{1st pc row scores vector} : \mathbf{s}_1 = \mathbf{Y}\mathbf{v}_1, \tag{3}$$

$$\text{1st pc factor loadings vector} : \mathbf{c}_1 = \mathbf{Y}'\mathbf{u}_1, \tag{4}$$

$$\mathbf{c}_1 = \boldsymbol{\lambda}_\mathbf{1}\mathbf{v}_1 \quad \text{and} \quad \mathbf{s}_1 = \boldsymbol{\lambda}_\mathbf{1}\mathbf{u}_1. \tag{5}$$

TI-PCA in L$_1$

$$\text{1st pc row scores vector} : \mathbf{s}_1 = \mathbf{Y}\mathbf{v}_1, \tag{6}$$

$$\text{1st pc factor loadings vector} : \mathbf{c}_1 = \mathbf{Y}'\mathbf{u}_1, \tag{7}$$

$$sgn(\mathbf{c}_1) = \mathbf{v}_1 \quad \text{and} \quad sgn(\mathbf{s}_1) = \mathbf{u}_1. \tag{8}$$

Where $sgn(.)$ is the coordinatewise sign function, $sgn(x) = 1$ if $x > 0$, and $sgn(x) = -1$ if $x \leq 0$.

- Residual data matrix:

PCA

$$\mathbf{Y}^{(1)} = \mathbf{Y} - \mathbf{s}_1\mathbf{c}_1'/\lambda_1. \tag{9}$$

TI-PCA in L$_1$

$$\mathbf{Y}^{(1)} = \mathbf{Y} - \mathbf{s}_1\mathbf{c}_1'/\lambda_1. \tag{10}$$

Note that in both cases, the rows of $\mathbf{Y}^{(1)}$ are orthogonal to $\mathbf{v}_1$, and the columns of $\mathbf{Y}^{(1)}$ are orthogonal to $\mathbf{u}_1$. From which, we deduce the next property.

- Relationships between the consecutive principal components:

  Let $\mathbf{v}_i, \mathbf{c}_i$, $\mathbf{u}_i$ and $\mathbf{s}_i$ be the ith principal component of weights, loadings and scores, respectively, calculated from $\mathbf{Y}^{(i-1)}$ for $i \geq 2$. Then:

  PCA

$$\mathbf{v}_i'\mathbf{v}_j{=}0 \quad \text{for} \quad i \neq j,$$

$$\mathbf{s}_i'\mathbf{s}_j = 0 \quad \text{for} \quad i \neq j.$$

  TI-PCA in L_1

$$sgn(\mathbf{c}_i')\mathbf{c}_j = \mathbf{v}_j'\mathbf{c}_i = 0 \quad \text{for} \quad i < j, \tag{11}$$

$$sgn(\mathbf{s}_j')\mathbf{s}_i = \mathbf{u}_j'\mathbf{s}_i = 0 \quad \text{for} \quad i < j. \tag{12}$$

Remarks

In the case of TI-PCA in L_1, first, (8) shows that there is a clear distinction between pc weights (± 1) and pc loadings; second, (11) shows that the j-th pc loadings vector is orthogonal to the i-th pc weights vector for $i < j$.

Similar to the singular value decomposition, the matrix $\sum_{i=1}^{k} \mathbf{s}_i \mathbf{c}_i^{'}/\lambda_i$ provides a k rank approximation of the data for $k \leq \min(m, n)$.

The classical PCA presupposes the existence of the variances of the variables in a data set, while the TI-PCA-L_1 presupposes the existence of the means of the variables in a data set. This shows that it is specifically useful for long-tailed data, where the existence of the variances of some of the variables is dubious.

Transition formulae, (3) through (8), between the principal scores and the principal component weights are important for the interpretation of graphical displays. A comparison of the transition formulae sheds further insight into the differences between the two methods: TI-PCA-L_1 is more robust than the classical PCA. In the TI-PCA-L_1, all the variables are included *uniformly* in the construction of a score of an observation, and similarly the component loading of a variable is the sum of *signed uniform* contributions of all the observations.

The calculation of the principal scores and the principal component weights of TI-PCA in L_1 can be accomplished by two algorithms. The first one is based on complete enumeration. The second one is based on iterating the transition formulae (6) and (7), similar to Wold's [11] NILES (nonlinear estimation by iterative least squares) algorithm, which is based on (3) and (4). The rows and the columns of the data can be used as initial values for the iterative algorithm.

Finally, we mention that the same citeria used to select the number of principal components in the classical PCA, such as the scree plot or the percentage of the total dispersion, can be used to select the number of principal components in the TI-PCA in L_1.

1a) **Ordinary PCA of** y_{ij}.								
	pc1	pc2	pc3	pc4	pc5	pc6	pc7	pc8
500L	0.39	0.33	0.21	-0.38	0.11	-0.39	-0.22	-0.59
1000L	0.42	0.24	-0.02	-0.45	-0.41	-0.07	0.08	0.63
2000L	0.39	-0.21	-0.46	-0.22	0.35	0.22	0.59	-0.18
4000L	0.28	-0.49	0.41	-0.24	0.18	0.51	-0.41	0.06
500R	0.33	0.39	0.28	0.46	0.58	0.00	0.08	0.33
1000R	0.40	0.23	-0.05	0.44	-0.52	0.48	-0.02	-0.31
2000R	0.34	-0.28	-0.56	0.27	0.06	-0.34	-0.53	0.13
4000R	0.26	-0.51	0.42	0.27	-0.25	-0.45	0.39	-0.03
L_2-norms	19.23	12.53	9.67	8.30	5.96	5.49	4.65	3.96
1b) **TI-PCA-L_1 of** y_{ij}.								
500L	0.44	0.41	0.19	-0.35	-0.31	-0.33	-0.26	0.35
1000L	0.43	0.31	-0.14	-0.42	0.22	-0.37	0.26	-0.35
2000L	0.36	-0.29	-0.42	-0.30	-0.32	0.28	0.43	0.35
4000L	0.28	-0.39	0.52	-0.34	0.40	0.42	-0.42	-0.35
500R	0.32	0.37	0.23	0.39	-0.39	0.33	0.26	-0.35
1000R	0.37	0.32	-0.28	0.37	0.47	0.37	-0.26	0.35
2000R	0.31	-0.32	-0.48	0.35	-0.38	-0.28	-0.43	-0.35
4000R	0.28	-0.41	0.38	0.29	0.29	-0.42	0.43	0.35
L_1-norms	442.89	264.92	200.56	167.25	140.03	126.77	98.96	92.61

Table 1: PCA of hearing loss data.

3 Example: Hearing loss data

The data set of dimension 100 × 8 is found in Jackson [9]. The first four columns represent the hearing measurements on the left ear of an individual, and the last four columns represent the hearing measurements on his right ear. The measuring of hearing is done with an audiometer at 4 different frequencies for each ear: 500Hz, 1000Hz, 2000Hz and 4000Hz. Jackson [9] applied the classical PCA, and he found the first four dimensions to be interpretable. Table 1 displays the 8 principal components obtained by classical PCA and TI-PCA in L_1 applied to the standardized data, $y_{ij} = (x_{ij} - \bar{x}_{.j})/s_j$, where $\bar{x}_{.j}$ and s_j represent the mean and the standard deviation of the jth variable, respectively. We note that the first four principal components are very similar in both approaches. The first principal component is a size factor. The second principal component opposes the low frequencies (500Hz and 1000Hz) to the higher frequencies (2000Hz and 4000Hz). The third principal component contrasts the two extreme frequencies (500Hz and 4000Hz) to the middle frequencies (1000Hz and 2000Hz), and the fourth principal component differentiates the left ear from the right. The remaining principal components obtained by the new method are also interpretable: the fifth principal component shows oppositions between (500Hz and 2000Hz) and (1000Hz and 4000Hz). The last three principal axes reproduce the second, third and fifth principal components but opposing the right ear to the left ear. We conclude that this data set is very well structured, and the complete structure is revealed by the TI-PCA in L_1 and not by the classical PCA. This example provides an empirical substantiation of Arabie [1], where it is argued that the L_1 metric models psychological processes (in this example physiological) better than the L_2 metric.

4 Conclusion

An invariance property of the classical PCA is used to develop a new method named TI-PCA in L_1. The mathematical framework of both methods is similar. The new method is robust compared to the classical PCA, and produces simultaneous dichotomies of the variables and the observations. We compared both methods on a real data set: The new method explained much better the underlying structure of the Hearing Loss Data. In the future, we intend to compare the TI-PCA in L_1 with the other L_1 methods proposed in the statistical literature.

References

[1] Arabie P. (1991). *Was Euclid an unnecessarily sophisticated psychologist?* Psychometrika **56**, 567 – 587.

[2] Baccini A., Besse Ph., de Falguerolles A. (1996). *A L_1-norm PCA and a heuristic approach.* In Ordinal and Symbolic Data Analysis, eds. Diday, E., Lechevalier, Y. and Opitz, O., Springer, 359 – 368.

[3] Benayade M., Fichet B. (1994). *Algorithms for a geometrical PCA in L_1-norm.* In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy (Eds.), New Approaches in Classification and Data Analysis, Berlin: Springer, 75 – 84.

[4] Choulakian V. (2001). *Robust Q-mode principal component analysis in L_1.* Computational Statistics and Data Analysis **37**, 135 – 150.

[5] Choulakian V. (2003). *The optimality of the centroid method.* Psychometrika **68**, 473 – 475.

[6] Choulakian V. (2004). *Matrix norms and principal component analysis.* Technical report in preparation.

[7] Galpin J.S., Hawkins D.M. (1987). *L_1 estimation of a covariance matrix.* Computational Statistics and Data Analysis **5**, 305 – 319.

[8] Heiser W.J. (1987). *Correspondence analysis with least absolute residuals.* Computational Statistics and Data Analysis **5**, 337 – 356.

[9] Jackson J.E. (1991). *A user's guide to principal components.* Wiley: N.Y.

[10] Jolliffe I.T. (2002). *Principal component analysis.* 2nd edition, Springer Verlag: N.Y.

[11] Wold H. (1966). *Estimation of principal components and related models by iterative least squares.* In Multivariate Analysis, Krishnaiah, P.R. (ed.), Academic Press, New York, 391- 420.

Acknowledgement: This research was financed by the Natural Sciences and Engineering Research Council of Canada.

Address: V. Choulakian, Dept. de Math/Statistique, Universite de Moncton, Moncton, N.B., E1A 3E9, CANADA

E-mail: `choulav@umoncton.ca`

COMPSTAT'2004 Symposium

A LOWER BOUND ON INSPECTION TIME FOR COMPLEX SYSTEMS WITH WEIBULL TRANSITIONS

Stephane Chrétien and F. Corset

Key words: Random networks, random linear programming, reliability, maintenance, Dyer-Frieze-McDiarmid's inequality.

COMPSTAT 2004 section: Reliability.

Abstract: The paper studies the expectation of the inspection time in complex aging systems. Under reasonable assumptions, this problem is equivalent to studying the expectation of the length of the shortest path in the directed degradation graph of the systems where the parameters are obtained from experts. The expectation itself being sometimes out of reach, in closed form or even through Monte Carlo simulations in the case of large systems, we study the bound of Dyer, Frieze and McDiarmid which provides an interesting upper bound in the case of exponential transition times between degradation states. On the other hand, we show that this bound does not hold for Weibull distributions. Another problem is that lower bounds are much more useful in the context of estimating inspection times before failure. Such a rigourous lower bound is presented for the case of Weibull distribution with reasonable values of the shape parameter.

1 Introduction

1.1 Problem statement

Consider a complex system whose n degradation states have been identified by experts. Let node 1 represent the state where the system is considered as new and let node n be the state of unacceptable degradation. All maximum paths from any node of the graph end at node n as in the figure below. The system is supposed to possibly evolve from a degradation state to any neighbor in the corresponding directed graph. The transition time between any two given states is assumed to follow a Weibull distribution whose parameters are given by experts or are estimated if the number of observations is sufficiently large. Using Bayesian statistics both informations can also be merged.

Assume we start with a brand new system. Then, evolution of the system starts in state 1. Maintenance policies require that the system be inspected before reaching state n, i.e. unacceptable degradation. Such examples of complex systems have been studied in [1]. The problem posed in this paper is to provide a lower bound on acceptable inspection times.

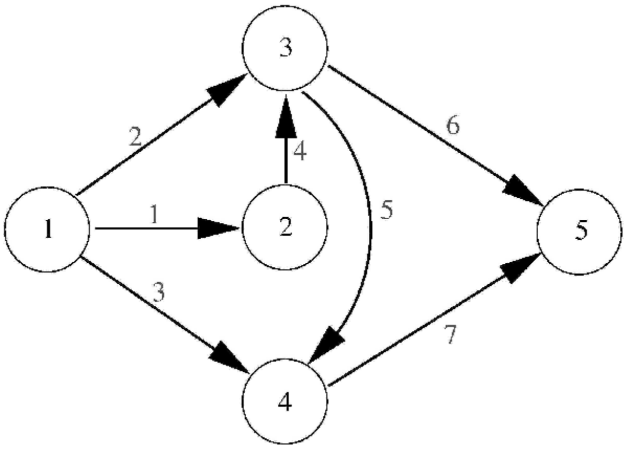


Figure 1: A simple graph.

1.2 Inspection times and shortest paths

In order to simplify the analysis, we assume that evolution inside the degradation graph proceeds following the rule that starting from one node i, the system goes to state j minimizing the transition time among neighbors of state i. Therefore, acceptable inspection times will be the times lower than *the shortest path* from state 1 to state n where each edge is weighted by its transition time. In general situations, we thus may ask for

- an estimator of the expected length of the shortest path from 1 to n,
- a confidence interval for the expected time. path.

This task is in general impossible to achieve because of the huge number of observations this should require in practice. The goal of this paper is to propose a lower bound on the expected length of the shortest path. On the other hand, approximate confidence intervals seem very difficult to obtain. A possible way of doing this may be the use of Talagrand's inequalities but this issue will not be discussed here.

2 The Dyer-Frieze-McDiarmid inequality for exponential transitions

An important step in the search of good bounds for expectations in combinatorial problems was achieved by Dyer, Frieze and Mc Diarmid in [2]; see also [3]. Their bound is an upper bound to the expectation of the optimal value. In comparison, the main objective of our work is to obtain a lower bound but using Dyer-Frieze-McDiarmid's bound gives a first understanding of the problem.

2.1 Linear programming formulation

The main idea is to convert the problem into an equivalent linear programming problem, when possible. Many combinatorial optimization problems cannot be transformed in this manner but it is well known that this is the case for the shortest path problem. Consider the following *extended incidence matrix* A of the oriented degradation graph. Its rows are indexed by the nodes of the graph while its columns are indexed by its edges with an extra column of all ones. In each column indexed by edge (i, j), set the i^{th} component to -1, the j^{th} component to 1 and set all other entries to zero. For instance, the extended incidence matrix for the graph of figure 1 is given by

$$A = \begin{bmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Now, the fact that we start at node 1 and end at node n is encoded in the vector $b = [-1, 0, 0, 0, 1]^t$. Then, the shortest path problem is equivalent to the linear program

$$\begin{aligned} z^\star = \quad & \min \quad c^T x \\ & \text{s.t.} \quad Ax = b \\ & \qquad\quad x \geq 0 \end{aligned} \tag{1}$$

where the random vector c contains the transition times for each edge of the graph.

2.2 The case of exponential transition times

We now apply the Dyer-Frieze-McDiarmid bound to our shortest path problem in the case where the transition times are independent and exponential. In this case, the mean residual times, i.e. the variables $E[c_i \mid c_i \geq h]$ satisfy the equality

$$E[c_i \mid c_i \geq h] = E[c_i] + h.$$

Then, an *upper bound* on the expected length of the shortest path is given by the following theorem.

Theorem 2.1. *(Dyer-Frieze-McDiarmid Inequality [2]) Assume that the random costs c_i are independent and satisfy*

$$E[c_i \mid c_i \geq h] \geq E[c_i] + \alpha h$$

for some $\alpha \in (0, 1]$. Then for any matrix $A \in \mathbb{R}^{n \times m}$ and any vector $b \in \mathbb{R}^n$, the optimal value z^* of the general linear program (1) satisfies

$$E[z^*] \leq \max_{S\,:\,\#S=n} \sum_{i \in S} E[c_i] x_i \tag{2}$$

for any feasible solution x, i.e. any x satisfying $Ax = b$.

A direct application of this theorem gives the following interesting corollary.

Corollary 2.1. *Consider problem (1) where the random costs are assumed to be independent and exponentially distributed and consider the associated deterministic linear program*

$$\begin{aligned} \zeta^\star = \quad & \min \quad E[c]^T x \\ & s.t. \quad Ax = b \\ & \qquad\quad x \geq 0 \end{aligned} \tag{3}$$

where the random costs are replaced by their expected values. Then the shortest path problem with optimal value denoted by z^* satisfies

$$E[z^*] \leq \zeta^*.$$

Proof. Take x equal to any binary vector minimizing (3). It is clear that the number of ones in this vector is less than the number of nodes in the graph. Then, the maximum value over all sets S of cardinality n in the right hand term in (2) is obtained when S is taken to be the set of indices i for which $x_i = 1$. Thus $\sum_{i \in S} E[c_i] x_i$ is exactly the cost of x , i.e. ζ^*. □

An important conclusion of this corollary is that, contrarily to the intuitive idea prescribed by common practice, replacing the random costs by their expected values cannot help for the problem of finding a lower bound to the inspection time of the system. A rigourous lower bound will be presented below.

3 A lower bound on the shortest path and inspection times

3.1 The mean residual time to failure for Weibull distributions

Analyzing the proof of the Dyer-Frieze-McDiarmid bound reveals the importance of studying the mean residual transition times. Here, we recall some results for mean residual times in the case where the variables have a Weibull distribution. Let X be a random variable with Weibull distribution

$$f_X(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^\beta e^{-(\frac{t}{\eta})^\beta}.$$

Then, the mean residual time to failure (MRTF) is given by

$$G_X(h) \triangleq E[X \mid X \geq h] = \eta e^{(\frac{h}{\eta})^\beta} \Gamma\left(1 + \frac{1}{\beta}, \left(\frac{h}{\eta}\right)^\beta\right),$$

where $\Gamma(a, h)$ is the incomplete gamma function defined by

$$\Gamma(a,h) = \int_h^{+\infty} t^{a-1} e^{-t} dt.$$

Lemma 3.1. *The first two derivatives of the MRTF for a Weibull distributed variable X are given by*

$$G'_X(h) = \frac{\beta}{\eta^\beta} h^{(\beta-1)} \left\{ \eta e^{\left(\frac{h}{\eta}\right)^\beta} \Gamma\left(1+\frac{1}{\beta}, \left(\frac{h}{\eta}\right)^\beta\right) - h \right\}$$

and

$$\begin{aligned} G''_X(h) &= -\frac{\beta^2}{\eta^\beta} h^{(\beta-1)} \left(1+\beta\left(\frac{h}{\eta}\right)^\beta\right) \\ &+ \frac{\beta}{\eta^{(\beta-1)}} h^{(\beta-2)} e^{\left(\frac{h}{\beta}\right)^\beta} \Gamma\left(1+\frac{1}{\beta}, \left(\frac{h}{\eta}\right)^\beta\right) \left(\beta\left(\frac{h}{\eta}\right)^\beta + \beta - 1\right). \end{aligned}$$

Moreover, when $\beta \geq 1$ we have

$$\lim_{h\to 0} G'_X(h) = 0 \text{ and } \lim_{h\to+\infty} G'_X(h) = 1$$

and if moreover $\beta < 2$ we have

$$G''_X(h) > 0, \quad \lim_{h\to 0} G''_X(h) = +\infty \text{ and } \lim_{h\to+\infty} G'_X(h) = 0.$$

Using this lemma, we can draw the following simple conclusion.

Proposition 3.1. *Assume that X is Weibull distributed with $\beta \in (1,2)$. Then, for all $h \geq 0$, we have*

$$E[X \mid X \geq h] \leq E[X] + h.$$

This result will be instrumental in the derivation of the bound.

3.2 A lower bound to the inspection time

In order to derive the lower bound, we need to clarify the behavior of the linear program (1) when subject to random costs. The presentation follows [2] and [3, Chapter 4]. For any family of n different columns of A indexed by B which are linearly independent and such that $A_B^{-1}b \geq 0$, we can find a feasible solution x satisfying $Ax = b$ and $x \geq 0$ in the following way:

$$\begin{cases} x_i = 0 \text{ forall } i \notin B \text{ and} \\ x_B = A_B^{-1} b \end{cases} \tag{4}$$

where x_B is the vector whose components are those of x with indices in B and A_B is the matrix whose columns are those of A whose indices are in B also. We will also sometimes use the notation $N = \{1, \ldots, n\} \setminus B$.

We then have the following standard result in linear programming theory.

Lemma 3.2. *([2, 3]) Let B be a subset of n indices such that $A_B^{-1}b \geq 0$. A vector x defined by (4) with respect to B will be a solution to program (1) if and only if $c_i \geq c_B^t A_B^{-1} a_i$ for all $i \notin B$ where a_i is the i^{th} column of A.*

A subset of indices B such that the columns of A_B are linearly independent and $A_B^{-1}b \geq 0$ is called a basis of the linear program. Solving the linear program thus consists of finding an optimal basis characterized by the property

$$c_i \geq c_B A_B^{-1} a_i \text{ for all } i \notin B. \tag{5}$$

The reason for appending the extra column of all ones to the incidence matrix is that the standard incidence matrix has rank equals to $n-1$ when the graph is connected. Thus, any basis for the shortest path problem must contain this extra column of ones which is easily seen to be orthogonal to any of the other columns.

Using these definitions, we now present the lower bound to the inspection time in the following theorem.

Theorem 3.1. *Consider the shortest path problem whose linear programming formulation is given by (1) where the components c_i of the cost vector c are independent and follow a Weibull distribution with parameters η_i and β_i, $i = 1, \ldots, n$. Let x be any feasible solution of (1), i.e. satisfying $Ax = b$ and $x \geq 0$. Then, the expectation of the random optimal value z^* admits the following lower bound*

$$\sum_r p_r \sum_{i \in B_r} E[c_i] x_i \leq E[z^*],$$

where (B_r) is the family of all bases for program (1) and for all r, p_r is the probability that B_r be optimal.

Sketch of the proof. The proof is adapted from the one of Dyer, Frieze and McDiarmid's inequality. For any feasible x, we can write

$$\begin{aligned} E[c^t x \mid B_r \text{ optimal }, c_{B_r}] \quad &= c_{B_r}^t x_{B_r} + \textstyle\sum_{i \notin B_r} E[c_i \mid \ B_r \text{ optimal }, c_{B_r}]^t x_i \\ &= c_{B_r}^t x_{B_r} + \textstyle\sum_{i \notin B_r} (E[c_i \mid c_i \geq c_{B_r} A_{B_r}^{-1} a_i \geq 0, c_{B_r}]^t x_i \\ &\geq c_{B_r}^t x_{B_r} + \textstyle\sum_{i \notin B_r} E[c_i]^t x_i + \textstyle\sum_{i \notin B_r} c_{B_r} A_{B_r}^{-1} a_i x_i \\ &= c_{B_r}^t x_{B_r} + E[c_{N_r}]^t x_{N_r} + c_{B_r} A_{B_r}^{-1} A_{N_r} x_{N_r}. \end{aligned}$$

The second equality is where both the underlying mechanism of linear programming and Proposition 3.1

play a crucial rôle. Now since

$$A_{B_r} x_{B_r} + A_{N_r} x_{N_r} = Ax = b$$

and $E[z^* \mid \ B_r \text{ optimal }, c_{B_r}] = c_{B_r} A_{B_r}^{-1} b$, we get

$$E[c^t x \mid \ B_r \text{ optimal }, c_{B_r}] \geq E[c_{N_r}]^t x_{N_r} + E[z^* \mid \ B_r \text{ optimal }, c_{B_r}].$$

Now, taking the expectation relative to c_{B_r} and using the formula of total probabilities with respect to the events "B_r is optimal" for all r, we easily get the announced result. □

This theorem cannot be used in crude manner. It explicitely requires that all probabilities p_r be estimated which amounts to a usually hudge number in complex cases. The main idea for using this result is to restrict to the bases which appear most often in the experiments. The following bound appears more tractable in practice.

Corollary 3.1. *Let x be the solution of the linear program (3) where the costs are replaced by their expected values, let B be the optimal basis associated with x and let $\hat{p}_B$ be a lower bound to the probability that B be optimal basis in (1) with confidence level $1-\alpha$. Then with probability $1-\alpha$ we have*

$$\hat{p}_B E[c]^t x \leq E[z^*].$$

Remark 3.1. *Of course, one can add to $\hat{p}_B E[c]^t x$ several more terms of the form $\hat{p}_r \sum_{i \in B_r} E[c_i] x_i$ for as many other bases occuring with nonneglectable probability as we want, using underestimations $\hat{p}_r$ of p_r for each new r. This results in sharper lower bounds to the mean inspection time.*

4 Simulations results

In this section we present our simulation results. We generate a hundred different shortest path problems on the graph of Figure 1. Each of these problems differs by the choice of the shape and scale parameters controlling the distribution of the cost on each edges. More precisely, for problem i, the cost c_j of edge j follows a Weibull distribution $W(\eta_j(i), \beta_j(i))$, where $\eta_j(i)$ was drawn at random from the normal distribution $\mathcal{N}(50, 100)$ and $\beta_j(i)$ was drawn from a uniform distribution over $[1, 2]$. For each problem, the expectation of the optimal cost is computed via Monte Carlo simulations over 1000 samples. Note that this is possible because of the simplicity of the chosen example. In order to compute our lower bound and following Remark 3.1, we use all the basis which occured to be optimal for at least one sample among the thousand samples. The bases are obtained as output of the simplex algorithm. Note that the only burden is to store these bases which is of course costless: we only take into account the small number of bases which frequently occur as optimal in the hundred tested samples of the shortest path problem. Finally, the underestimations $\hat{p}_r$ have been chosen as the lower bounds of the respective confidence intervals for the probabilities p_r with risk value $\alpha = 5\%$. The results are given in Figure 2 below.

In most cases, our lower bound is closer than the Dyer-Frieze-McDiarmid (DFM) upper bound to the expected cost in absolute value. Moreover, in general, the mean relative error between our bound and the expected value is less than 20% over the hundred generated problems (19.86 % in the displayed example).

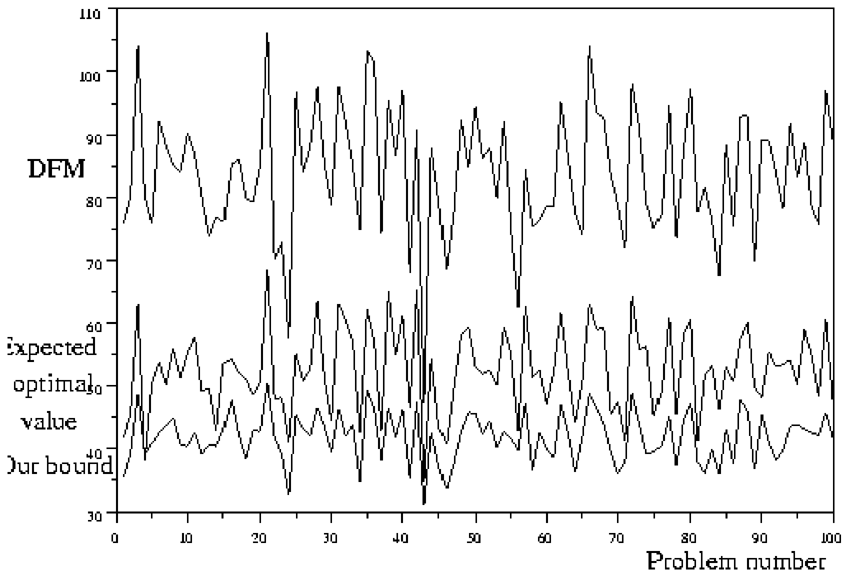


Figure 2: Comparison of the bounds for one hundred problems.

This very preliminary experimental results seem to be quite promising, and we plan to try our bound on real life problems in a short future.

References

[1] Corset F. (2003). *Maintenance optimization from Bayesian networks and reliability with doubly censored data.* Doctoral Thesis, Université Joseph Fourier, Grenoble, France.

[2] Dyer M.E., Frieze A.M. and McDiarmid C.J.H. (1986). *On linear programs with random costs.* Math. Programming **35**, 3 – 16.

[3] Steele J.M. (1997). *Probability Theory and Combinatorial Optimization.* CBMS-NSF Conference Series in Applied Mathematics 69, SIAM.

Address: S. Chrétien, Université de Franche Comté, UMR6623, Dpt de mathématiques, 16 route de Gray 25000 Besançon, France
F. Corset, LabSAD, EA 3698, Université Pierre Mendes France, BSHM, 1251 avenue centrale BP47 38040 Grenoble Cedex 09, France

E-mail: `chretien@math.univ-fcomte.fr,`
`Franck.Corset@upmf-grenoble.fr`

AN INFERENCE CURVE-BASED RANKING TECHNIQUE FOR UNIVARIATE OBSERVATIONS WITH BIOMEDICAL APPLICATIONS

Christos Christodoulou, A. Karagrigoriou and F. Vonta

Key words: Correlation, influence curve, outliers.

COMPSTAT 2004 section: Model selection.

Abstract: In this paper we introduce a correlation ranking method which is established with the use of the influence curve. The method proposed has numerous applications among which one could mention outlier detection. For illustrative purposes, a number of real medical data sets are studied and the so called reference range is obtained.

1 Introduction

A way of "ranking" bivariate observations following a linear regression model in terms of their contribution to the fitting of the model has been introduced by Atkinson and Riani [1]. Their technique focuses exclusively on outlier detection for regression models. In this paper we propose a different technique for ranking univariate observations by means of the correlation between the sample data and their respective ranks. The proposed method is then used in various examples with both simulated and real data. In all the examples we concentrate on the topic of reference range which has been given considerable attention in various scientific areas among which is the medical research.

The paper is presented as follows: In Section 2, the main inference curve-based ranking technique is introduced while in Section 3 a number of examples and applications are fully discussed. A discussion section has been included at the end of the paper.

2 The proposed ranking technique

In this section we introduce the use of a technique for ranking univariate observations by means of the correlation between the sample data x_i and their respective ranks. The proposed method is based on a very clear and simple idea according to which the issue of ordering of the observations should not be based on their size but rather on their effect or contribution on the underlying model or process the observations suppose to satisfy. Such a ranking may be called "model-ranking". For example, for bivariate data that fit a p-dimensional regression model the observations could be ranked according to their contribution on the residuals or on their contribution on the coefficient of determination. A similar approach could be adopted for time series data.

It is of course understood that in many instances the underlying model is not available and therefore a model-ranking is not possible. In medical research and especially in laboratory research where biochemical characteristics behave independently no model may be appropriate or available. In such cases an artificial model may be implemented, like the one describing the (linear) relation between the observations and their corresponding traditional ranks. The alternative approach we propose here in such cases is based therefore on "(pseudo-)model-ranking".

For the pseudo-model-ranking we make use of the definition of the influence curve for bivariate data using as the quantity of interest the correlation coefficient and we rank the data according to their influence on the correlation coefficient. The influence value of the point (x, y) is

$$IC(x,y) = \frac{x-\mu_1}{\sigma_1} \cdot \frac{y-\mu_2}{\sigma_2} - \frac{\rho}{2}\left[(\frac{x-\mu_1}{\sigma_1})^2 + (\frac{y-\mu_2}{\sigma_2})^2 \right],$$

where μ_1, σ_1 are the sample mean and standard deviation of the x's, μ_2, σ_2 the corresponding values of the y's and ρ the sample correlation coeffitient between x and y. The three steps of the procedure are as follows:

Step 1: We construct a single regression model between the sample data and their respective ranks, $y_i = rank\,[x_i]\,,\ i = 1, 2, \ldots, n$. Note that here we refer to the traditional ranking (order statistics).
Step 2: Start with $j = n$ (the initial sample size).
Step 3: Calculate the influence curve of each pair of observations (x_i, y_i), $i = 1, 2, \ldots, j$. The observation with the smallest (worst) influence is assigned the higher ranking, namely $x^*_{(j)}$.
Step 4: Repeat Step 3, excluding all observations that have been assigned a rank such that $j = n - (\#$ of obs. already ranked). Let the ranked data be $x^*_{(1)}, \ldots, x^*_{(n)}$ with $x^*_{(1)}$ and $x^*_{(n)}$ the highest and lowest influence observations respectively.

The above ranking method could be used in various ways by the statistician. One way is by applying the method to identify possible outliers in a given set of data. Various outlier detection techniques can be found in the literature for both univariate and multivariate regression and time series models. Barnett and Lewis [2] provide an extensive literature review, while Davies and Gather [3] introduce the outlier region with the use of finite breakdown points [4]. Peña and Yohai [8] have introduced a procedure of high-breakdown robust point estimates for regression through which one succeeds in detecting groups of outliers. A large collection of graphical and other techniques for outlier diagnostics can be found in Atkinson and Riani [1] which have been applied to different types of regression models. In multivariate samples, a fast outlier detection procedure based on the analysis of the projections of the sample points onto a set of $2p$ directions was recently developed [7].

It is obvious that there are several statistical tools that could be used for outlier detection. For the purpose of this work, we have decided to select for this role, the correlation coefficient. As a result, the following step could be added to the above algorithm for the identification of possible outliers.

Step 5: Start with $x^*_{(1)}$ and $x^*_{(2)}$ and calculate R^2 repeatedly by adding each time the next ordered observation. If the inclusion of the kth ordered observation, $k = 3, 4, \dots, n$ results in $R^2 < \delta$ for some pre-assigned $0 < \delta < 1$ chosen by the statistician, then the observations $x^*_{(k)}, \dots, x^*_{(n)}$ are (possible) outliers.

Intuitively, if no outlying observations are present then the correlation coefficient between the observations and their ranks should be very close to 1 and their linear relation extremely strong. Naturally the value of δ is expected to be chosen very close to 1. Simulated data for various continuous and discrete distributions show that a threshold close to 0.95 is adequate for most applications.

For the sake of completeness we need to mention briefly here the traditional approach of Hoaglin, Mosteller, and Tukey [6] for identifying outliers. According to their approach which is based on the fourth spread, we lay off a multiple of $\frac{3}{2}$, upward from the upper fourth and downward from the lower fourth. The lower fourth, which is denoted by F_L is defined as the median of the observations lying between the minimum and the median of a given data set. The upper fourth, F_U is defined symmetrically with respect to the median. The resulting interval is

$$\Big(F_L - 1.5d_F, F_U + 1.5d_F\Big)$$

where $d_F = F_U - F_L$ is the fourth spread. Observations outside these cut-offs are regarded as outliers. A generalization of the above cut-offs is given by:

$$\Big(F_L - ad_F, F_U + ad_F\Big). \tag{2.1}$$

where a is a real number to be chosen by the statistician according to the definition of outliers she/he is willing to adopt.

3 Examples and applications

The issues raised and discussed in this article have numerous applications. Here we briefly cover the topic of reference range which is a main issue in medical research used for identifying abnormalities in clinical investigations. In most diagnostic tests, the reference range for the variables tested play the key role in determining the type and the extent of the therapeutic or pharmaceutical action to be taken. The reference range is defined as an interval inside which the normal (non abnormal) values of a certain characteristic lie.

Values outside the reference range are indicative of a malfunction and the patients should be properly treated. The most usual reference ranges are the inter-centile intervals based on the upper and lower 2.5th and 5th percentiles. It is natural that for the estimation of these percentiles caution is required in identifying any outliers. Note that the inclusion of outliers in the process of calculating the boundaries of the range may result in an unacceptable range that may be inadequate for preventive, diagnostic or therapeutic purposes.

(a) Simulated data

In order to test the effectiveness of the proposed ranking method in outlier detection, two short scale simulation studies were conducted. It should be pointed out that the objective of the simulation studies is the investigation of the effectiveness of the proposed method for cases with different degree of distinction between the two distributions involved. As a result the studies were decided to be based on the contamination of a basic distribution by distributions located very near as well as far enough from the basic distribution. In both studies, the standard Normal distribution was selected to play the role of the basic distribution. Furthermore, the Normal distribution N(3,1) has been selected as representative of the distributions located relatively close to the standard Normal so that the distinction between them is relatively difficult. Finally, the opposite role has been given to the Normal distribution N(6,1). In the first study, one hundred samples with 100 observations each were selected. Each sample consists of 95 observations randomly selected from the standard Normal distribution and 5 observations randomly selected from the Normal distribution with mean 6 and variance 1. In other words, each sample has been contaminated with a random sample of size 5 from the $N(6,1)$ distribution. The influence curve was calculated and the (pseudo)-model ranks were assigned (the higher the rank the lower the influence). The method has successfully identified all contaminated data as the least influential data in each sample. Indeed, for each sample, the method selects first as the most influential observations, the 95 observations from the $N(0,1)$ (ranks 1 through 95) and then selects the contaminated observations (ranks 96 through 100).

The correlation coefficient remains above the 0.95 threshold for all pure (non-contaminated) observations (a total of 9500 observations in all 100 samples). The 0.95 threshold is crossed for 476 out of 500 contaminated observations while for the remaining 24 contaminated cases the correlation coefficient remained incorrectly above the threshold. As a result, the first part of the method based on the influence curve has a 100% rate of success while the second part based on the correlation coefficient has only a 95.2% rate of success. Table 2 provides for all 100 samples combined the proportion of observations for which the correlation coefficient stays above or below the $\delta = 0.95$ threshold. The results are presented separately for the 9500 pure $N(0,1)$ observations and the 500 contaminated $N(6,1)$ observations.

	9500 $N(0,1)$ obs & 500 $N(6,1)$ obs.	
$\delta = 0.95$	**% of $N(0,1)$ obs.**	**% of $N(6,1)$ obs.**
ABOVE	100%	4.8%
BELOW	0%	95.2%

Table 1: First simulation study.

A second simulation study was contacted identical to the first one except that the $N(3,1)$ distribution is now responsible for the contamination. Note that due to the overlapping of the lower half of $N(3,1)$ with the upper half of the standard Normal, approximately 50% of the contaminated observations are expected never to be identified. The influence curve method though was successfully applied to the data with an overall (with all 100 samples combined) rate of success equal to 61.6%, which means that 61.6% of the contaminated observations where correctly identified as the least influential ones. Furthermore, in 65% of the samples the method correctly identified at least 3 (out of 5) contaminated observations as the least influential ones. Table 2 provides the proportion of samples where the proposed method recovered successfully 0, 1, 2, 3, 4, and 5 contaminated observations per sample.

95 $N(0,1)$ obs & 5 $N(3,1)$ obs per sample						
# of contaminated obs. identified	**0**	**1**	**2**	**3**	**4**	**5**
Proportion of samples (%)	2	10	23	22	29	14

Table 2: Second simulation study.

The above studies clearly indicate that the proposed ranking technique based on the influence curve combined with the correlation coefficient produce an extremely successful technique for identifying contaminated data.

(b) Real data

For illustrating the methodology proposed in this article, we make use of a set of real data on cholesterol and the transition mutation C677T of 5, 10-methylenetetrahydrofolate reductase (MTHFR) provided by the Molecular Genetics Department B and Laboratory of Forensic Genetics of the Cyprus Institute of Neurology and Genetics. A total of 515 individuals, 319 males and 196 females were examined and among other variables, the cholesterol level was measured in mMol/L (min=3.4, max=10.8, mean=6.28, std. dev.= 1.17, and median=6.2). Furthermore for 87 of these individuals the genotype frequencies for the MTHFR C677T polymorphism were established. 41 individuals (47%) were homozygous for the wild type allele (C/C), 34 (39%) were heterozygous (C/T), and 12 (14%) were homozygous for the mutant allele (T/T). The purpose of the study is the estimation of a 95% reference range for the cholesterol level as well as reference ranges for the cholesterol level for the C/C and C/T genotype for MTHFR C677T polymorphism. No range will be provided for the mutant homozygous genotype due to the small number of cases. Since the data do not follow the normal distribution, a series of transformations on the data to achieve normality is considered necessary. Making use of the Kolmogorov test, it is realized that the cholesterol level data do not come from a normal distribution (test statistic=0.119, p-value=0). After applying the log-transformation, we achieve normality (test statistic reduces to 0.0397, p-value=0.3921).

Applying now the 5-step-(pseudo) model-ranking method with $\delta = 0.95$ we observe that no outliers are present. Note that the Hoaglin, Mosteller, and Tukey (HMT) method arrives at the same results. Indeed, consider the generalized formula of HMT we introduced in equation (2.1) and assume that a is defined so that if the distribution is indeed normal then all observations of the sample that depart more than three standard deviations from the mean (or the median, since they coincide) are characterized as outliers and consequently they are discarded from the data set. Under this definition it is easily found that $a = 1.7$. Using this general formula for our example, we identify no outliers. Calculating the upper and lower 2.5th sample percentiles we arrive at the 95% reference range: $(4.25, 9.11)$.

Following the same procedure for the 41 cases for the wild type genotype for MTHFR C677T (MTHFR=1) and the 34 cases of the heterozygous genotype for MTHFR C677T (MTHFR=2) we identify one outlier in the former case and none in the latter. It should be pointed out that both the proposed method and the HMT method arrive at the same result. Figure 1 provides the corresponding correlation coefficients. The single outlier found is easily identified because of the significant drop in the value of the correlation coefficient (for MTHFR=1). The resulting 95% reference ranges are $(4.45, 8.37)$ for the wild type genotype and $(4.65, 8.82)$ for the heterozygous genotype.

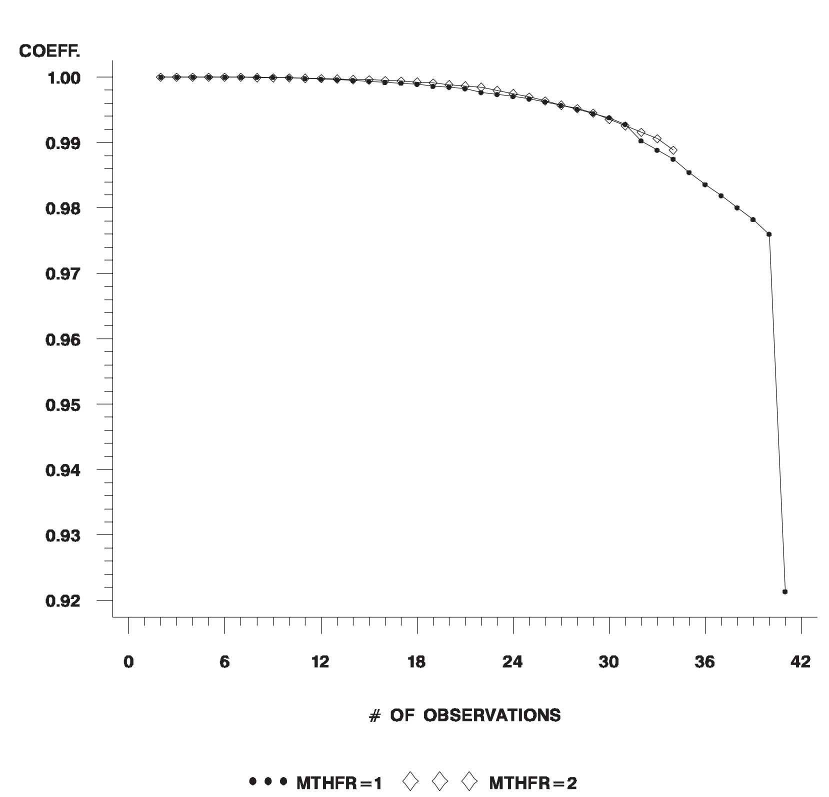
Note that the above results indicate that the MTHFR C677T mutation has no effect on the cholesterol level. Such a conclusion is indeed expected since the MTHFR C677T mutation has not been shown in any study to affect cholesterol levels and further, MTHFR does not interact in anyway with cholesterol metabolism.

4 Discussion

In this work we introduce the notion of pseudo-model-ranking for which the observations are ranked according to their influence on an artificial regression model between the observations and their traditional ranks.

As it is well known the outlier detection plays a very important role not only in the analysis of biomedical data but in general in the analysis of any type of data. Indeed, the modelling, the statistical inference as well as the prediction inference may be heavily affected by the presence of outliers. The simulation studies undertaken as well as the biomedical applications considered in this work clearly indicate both the appropriateness and the effectiveness of the proposed method.

Figure 1: Coefficient of Determination – Cholesterol for MTHFR



References

[1] Atkinson A., Riani M. (2000). *Robust diagnostic regression analysis.* Springer Verlag.

[2] Barnett V., Lewis T. (1984). *Outliers in statistical data*. Wiley.

[3] Davies L., Gather U. (1993). *The identification of multiple outliers.* JASA **88**, 782–792.

[4] Donoho D.L., Huber P.J. (1983). *The notion of breakdown point.* In A Festschrift for E.L. Lehmann, eds. P.J. Bickel, K.A. Doksum, and J.L. Hodges, Jr., Belmont, CA: Wadsworth, 157–184.

[5] Hoaglin D.C., Inglewicz B., Tukey J.W. (1980). *Small-sample performance of a resistant rule for outl ier detection.* Proc. of the Statist. Comput. Sec. of the Amer. Statist. Soc., 148–152.

[6] Hoaglin D.C., Mosteller F., Tukey J.W. (1983). *Understanding robust and exploratory data analysis*. Wiley.

[7] Pena D., Prieto F.J. (2001). *Multivariate outlier detection and robust covariance matrix estimation*. Technometrics **43**.

[8] Pena D. and Yohai V. (1999). *A fast procedure for outlier diagnostics in large regression problems*. JASA **94**, 434−445.

Acknowledgement: The authors wish to express their sincere appreciation to Dr. M. A. Cariolou of the Cyprus Institute of Neurology and Genetics not only for the use of the data sets but also for his guidance and support during the entire course of this work. We also thank Dr. S. L. Xenophontos for numerous helpful comments and suggestions.

Address: C. Christodoulou, A. Karagrigoriou, F. Vonta, Department of Mathematics and Statistics, University of Cyprus, CY-1678 Nicosia, Cyprus

E-mail: `alex@ucy.ac.cy`

COMPSTAT'2004 Symposium

MODEL BASED VISUALIZATION OF PORTFOLIO STYLE ANALYSIS

Claudio Conversano and Domenico Vistocco

Key words: Constrained linear regression, mean integrated squared error, parallel coordinates, mutual funds benchmarking, active vs. passive management.

COMPSTAT 2004 section: Financial data analysis.

Abstract: The paper concerns multimanager management of Mutual Funds portfolios, a kind of investment allocating the financial resources to Mutual Funds instead of standard financial stocks. In particular, Mutual Funds portfolios investing on a single asset category (equity, bond, corporate, derivative) are considered. The goal is to provide a method to rank Mutual Funds investment style with respect to the returns of a target index named benchmark. To this aim, a rolling constrained multiple linear regression model is considered in order to estimate each Mutual Fund portfolio composition as well as the benchmark portfolio composition. Starting from such compositions, the Mean Integrated Squared Error (MISE) is computed to measure the proximity of each Mutual Fund portfolio returns to the benchmark portfolio returns. A visual inspection of this proximities is provided using parallel coordinates plot. The method allows to identify a specific management style for each Mutual Fund, discriminating active management Funds from passive management ones. To evaluate the effectiveness of the proposed method, an application on a set of Italian Mutual Funds operating in the European equity market is presented.

1 The framework

Nowadays, among the range of new financial products offered by investment banks, portfolios composed by Mutual Funds are an interesting opportunity for investors. Their aim is to offer the investor an "optimal" portfolio (in the Markovitz sense), i.e. a portfolio providing the maximum return with the minimum risk. Two main categories can be considered:

1) *Mutual Funds Individual Portfolios* (MFIP). For each client, the investment bank allocates resources in a basket of Mutual Funds instead of a set of different financial stocks.

2) *Mutual Funds Mixed Portfolios* (MFMP). The bank is able to offer the same product obtained by combining different quotas of Mutual Funds to a set of clients. Here, the product is not offered to a single investor but to many of them.

Hereinafter, we refer to MFMP investing on a single asset category (*mono-market management*). This choice allows to start from a set of quite homogeneous products to be evaluated on the basis of common criteria. In many countries (including Italy), investment banks are legally confined to provide general information to the official institutions governing the financial markets about their Mutual Fund asset allocation principles and the *benchmarks*. The latter are reference portfolios composed by one or more financial market indexes. The role of benchmark is threefold:

a) for the portfolio manager, it is a reference parameter in the asset allocation process;

b) for the investment bank internal auditing, it is a tool to evaluate the management style and the results of the management team;

c) for the client, it is a reference index to evaluate the portfolio management activity, since the returns of each Fund should not be too far from the benchmark ones, particularly if negatives.

Often, the official benchmark (also known as *strategic benchmark*) differs from the *operative benchmark*. The first is usually selected to minimize the risk of future negative deviations. In fact, to avoid the risk of losing clients, portfolio managers select benchmarks that can be easily replicated. Instead, operative benchmark is selected according to internal auditing indications, since it is commonly used to monitor the portfolio managing activity. As a consequence, a financial analyst wishing to evaluate the management style of a set of Mutual Funds managed by various investment banks should refer to a benchmark that it is meant to be representative of the Mutual Funds specific sector, since information about each specific portfolio composition can not be retrieved easily.

2 Data, method and notation

We consider Italian Mutual Funds operating to the European equities sector. We collect daily time series from January 4, 1999 to September 30, 2002 (976 observations) concerning:

- a set of 39 Italian Mutual Funds;
- the Morgan Stanley Europe Equity Index (benchmark);
- the benchmark constituents indexes concerning 10 activity sectors, namely: *Energy (ENR), Materials (MAT), Industrials (IND), Consumer Discretionary (CDIS), Consumer Staples (CSTA), Health Care (HC),Financial (FNCL), Information Technology (IT), Telecommunication Services (TEL), Utilities (UTI))*[1].

Each series is quoted in Euro and is rescaled on weekly basis (the Wednesday observation is considered as representative of the whole week, as common practice in financial data analysis). As a result, the final weekly time series are composed of 195 observations.

The proposed method, to be described formally in the following section, is in three steps. The first step is the estimation of the benchmark port-

[1] A more analytical decomposition resulting in more sectors can be used depending on the accuracy level of the analysis. Of course, when increasing the number of sectors the interpretability of final results become more difficult.

folio composition. It takes into account the benchmark constituents using a constrained rolling regression model ([1],[2]), aimed to evaluate the impact of each constituent in the generation of the benchmark portfolio returns (step 1A). The same approach is used for each Mutual Fund portfolio to identify the main differences in the portfolio management style, through the estimation of the portfolio weights of each activity class (step 1B). Once the portfolio compositions of both Mutual Fund portfolios and benchmark portfolio have been estimated, a ranking of the Mutual Fund investment styles is made with respect to both the whole benchmark portfolio composition (global ranking) and each benchmark constituent (partial ranking). The Mean Integrated Squared Error (MISE) allows to obtain these rankings (Step 2). Finally, using interactive parallel coordinates plots, a visual inspection of the results of previous steps is provided (Step 3).

Since we apply a rolling regression model, we denote with h the number of weeks composing the temporal window. Furthermore:

- $t = 1, \ldots, T$ indicates time occasions,
- $f = 1, \ldots, F$ indicates the different Mutual Funds,
- $s = 1, \ldots, S$ indicates the activity sectors of the benchmark constituents.

In our application we use $h = 52$ (corresponding to the number of weeks in a year), $T = 195$, $F = 39$, $S = 10$ (since the sectorial composition of the benchmark portfolio refers to the 10 above-mentioned activity sectors). We apply a constrained rolling regression model $T - h$ times. The model is applied separately for the benchmark and the Mutual Funds, namely:

1.) using as covariates the benchmark constituents indexes returns and as response the benchmark index returns (*step 1A*);

2.) using as covariates the benchmark constituents indexes returns and as response the Mutual Funds returns (*step 1B*).

3 Formal description of the method

STEP 1: Style Analysis (Sharpe-like)[3]. We consider $T - h = 143$ occasions and define:

- $\mathbf{r}^{(j)}_{BMK}$: the $h \times 1$ vector of benchmark returns ($j = 1, \ldots, T - h$)
- $\mathbf{R}^{(j)}_{CONST}$: the $h \times S$ matrix of the benchmark constituents returns ($j = 1, \ldots, T - h$ and $s = 1, \ldots, S$).
- $\mathbf{r}^{(f,j)}_{FUND}$: the $h \times 1$ vector of the returns of the Mutual Fund f ($f = 1, \ldots, F$, and $j = 1, \ldots, T - h$).

STEP 1A: *Estimation of the benchmark sector composition.* To estimate the benchmark sector composition we define the following multiple regression model [2]:

$$\mathbf{r}^{(j)}_{BMK} = \mathbf{R}^{(j)}_{CONST}\beta^{(j)}_{BMK} + \epsilon \qquad (j = 1, \ldots, T - h)$$

with the constraints:

$$\begin{cases} \sum_{s=1}^{S} \beta_{BMK}^{(j)} = 1 \\ \\ 0 \leq \beta_{BMK}^{(j)} \leq 1 \qquad (s = 1, \ldots, S) \end{cases}$$

The model coefficients indicate the benchmark composition with respect to the S different constituents. Since the rolling estimation process is repeated $j = 1, \ldots, T-h$ times, we obtain $T-h$ model coefficient vectors. They form the following matrix:

$$\hat{\mathbf{B}}_{BMK} = \begin{bmatrix} \hat{\beta}_{BMK_1}^{(1)} & \cdots & \hat{\beta}_{BMK_1}^{(j)} & \cdots & \hat{\beta}_{BMK_1}^{(T-h)} \\ \vdots & & \vdots & \vdots & \\ \hat{\beta}_{BMK_S}^{(1)} & \cdots & \hat{\beta}_{BMK_S}^{(j)} & \cdots & \hat{\beta}_{BMK_S}^{(T-h)} \end{bmatrix}$$

$\hat{\mathbf{B}}_{BMK}$ is a $S \times (T-h)$ matrix such that, for each j, $0 \leq \hat{\beta}_{BMK}^{(j)} \leq 1$ and $\sum_{s=1}^{S} \hat{\beta}_{BMK}^{(j)} = 1$.

STEP 1B: *Estimation of Mutual Funds portfolio sector composition.* Likewise step 1, to estimate each Mutual Fund sector composition we define the following multiple regression model [2]:

$$\mathbf{r}_{FUND}^{(f,j)} = \mathbf{R}_{CONST}^{(f,j)} \beta_{FUND}^{(f,j)} + \epsilon \qquad (f = 1, \ldots, F) \quad (j = 1, \ldots, T-h)$$

with the constraints:

$$\begin{cases} \sum_{s=1}^{S} \beta_{FUND}^{(f,j)} = 1 \\ \\ 0 \leq \beta_{FUND}^{(f,j)} \leq 1 \qquad (s = 1, \ldots, S) \end{cases}$$

In this case, for the Mutual Fund f the model coefficients express its portfolio composition with respect to the sector s. Since the rolling estimation process is repeated $j = 1, \ldots, T-h$ times, we obtain $T-h$ vectors, that form the matrix:

$$\hat{\mathbf{B}}_{FUND} = \begin{bmatrix} \hat{\beta}_{FUND_1}^{(f,1)} & \cdots & \hat{\beta}_{FUND_1}^{(f,j)} & \cdots & \hat{\beta}_{FUND_1}^{(f,T-h)} \\ \vdots & & \vdots & \vdots & \\ \hat{\beta}_{FUND_S}^{(f,1)} & \cdots & \hat{\beta}_{FUND_S}^{(f,j)} & \cdots & \hat{\beta}_{FUND_S}^{(f,T-h)} \end{bmatrix}$$

$\hat{\mathbf{B}}_{FUND}$ is a $S \times (T-h)$ matrix such that, for each j, $0 \leq \hat{\beta}_{FUND}^{(f,j)} \leq 1$ and $\sum_{s=1}^{S} \hat{\beta}_{FUND}^{(f,j)} = 1$. A matrix $\hat{\mathbf{B}}_{FUND}$ is obtained for each Mutual Fund f $(f = 1, \ldots, F)$.

Information obtained in Step 1 are the starting point to rank the asset allocation process of each Fund in steps 2 and 3.

STEP 2: Mutual Funds Management Style Ranking. The estimation of the benchmark portfolio composition (step 1A) and that of each Mutual Fund portfolio composition (step 1B) allow to rebuild the F Mutual Funds portfolios and to compare them with the benchmark portfolio composition.

To this purpose, the *Mean Integrated Squared Error* (MISE) is cnsidered. It measures the proximity between an estimated curve $\hat{m}$ and the true function m and it is used in the framework of regression smoothing through kernel functions[4]. Here, f is a marginal density and w is a weight function. The MISE, denoted with d_M, can be meant as the expectation of the *Asymptotic Squared Error* (ASE), denoted with d_I, namely:

$$d_M(\hat{m}, m) = E\{d_I(\hat{m}, m)\}$$

with $d_I(\hat{m}, m) = \int(\hat{m} - m(x))^2 f(x)w(x)dx$.

Hereinafter, we assume the benchmark portfolio time series returns is the true function m to be estimated using the returns of the Mutual Fund recomputed according to the portfolio weights estimated in step 1B. This assumption allows to discriminate between two different portfolio management styles:

a)*passive management*: the asset manager investment style is driven by the benchmark portfolio composition and the corresponding MISE is small;

b)*active management*: the asset manager investment style is not necessarily driven by the benchmark portfolio composition. In this case, the corresponding MISE is high.

STEP 3: *Visualisation of the Management Style*. The whole information obtained in the previous steps can be represented using interactively parallel coordinates plots [5].

These can be considered as a generalisation of a two-dimensional Cartesian Plot, since they maps the k-dimensional space onto two display dimensions using k equidistant axes, which are parallel to one of the display axes. The axes correspond to the dimensions and are linearly scaled from the minimum to the maximum value of the corresponding dimension. Each data item is presented as a polygonal line, intersecting each axes at the point corresponding to the value of the considered dimension. Although the basic idea of parallel coordinates is quite simple, this tool is powerful in revealing a wide range of data characteristics, including the ability to diagnose one-dimensional features (such as marginal densities), two-dimensional features (such as correlations) and nonlinear structures, as well as multidimensional features (such as clustering, hyperplanes and the modes).

In our method, parallel coordinates plots are used to visualize together different portfolio management styles. For each observation, we plot the global MISE (calculated w.r.t. the benchmark portfolio returns) and the sectorial MISE (calculated w.r.t. the returns of sector indexes). The information deriving from this kind of visualization can be combined with return and risk measures in order to evaluate the way the management style affect the Mutual Fund performance.

Mutual Fund	**Global MISE**	**Average Rank**	**Average Return**	**Perfor- mance**	**Historical Volatility**
1. Gestnord	**0.117**	**10.45**	-0.038	-10.391	2.270
2. Arca	0.119	14.73	-0.017	-7.043	2.256
3. Ras	0.148	10.64	**-0.003**	-5.654	2.476
4. Investire	0.157	17.36	-0.044	-11.303	2.242
5. Nextra	0.165	15.91	-0.057	**-13.652**	2.360
...	...	...	...	...	...
35. Gestielle	0.948	26.27	0.004	-3.871	2.328
36. Azimut	1.075	25.64	**0.207**	**33.318**	2.796
37. Zetaswiss	1.628	33.55	0.077	11.365	1.690
38. F&F Pot.	2.041	**30.91**	0.095	7.538	**3.231**
39. Prime	**2.618**	26.82	0.113	18.825	**1.603**

Table 1: The 5 most passive vs. the 5 most active Mutual Funds according to the global MISE (first column) and other risk/return measures.

4 Main results

The proposed method has been applied for the 39 Italian Mutual Funds described in section 2. For sake of brevity we summarize just the main results. Table 1 reports the most 5 *passive Funds* (i.e. those reporting the lowest value of the global MISE) compared to the most 5 *active Funds* (i.e. those reporting the highest value of the global MISE). For comparison purposes, also the most common risk/return measures are reported, namely: a) the average rank (second column), i.e. the average rank of the Mutual Fund, obtained when ordering observations increasingly with respect to the value of the sectorial MISE; b) the avarage return (third column), i.e. the average of the weekly returns each Fund reported in the whole period; c) the performance (fouth column), i.e. the total return of the Fund in the whole period; d) the historical volatility (last column), i.e. the standard deviation of the weekly returns in the whole period. The table shows that the global MISE reflects the overall management style, because the average rank of the most passive Funds is lower than the global MISE of the most active Funds. Furthermore, considering together the performance, the average return and the volatility of the two groups of Mutual Funds, it is evident that an active management style usually improves the Mutual Fund performance and does not increase the risk (volatility).
Finally, the investment style of the two groups of Funds is compared though parallel coordinates plots using the *Mondrian*[6] software.

Figure 1 shows the global and sectiorial MISE for the 39 Funds. In the top panel the polygons corresponding to the 5 most passive Mutual Funds have been highlighted and similar visualization is available for the 5 most active

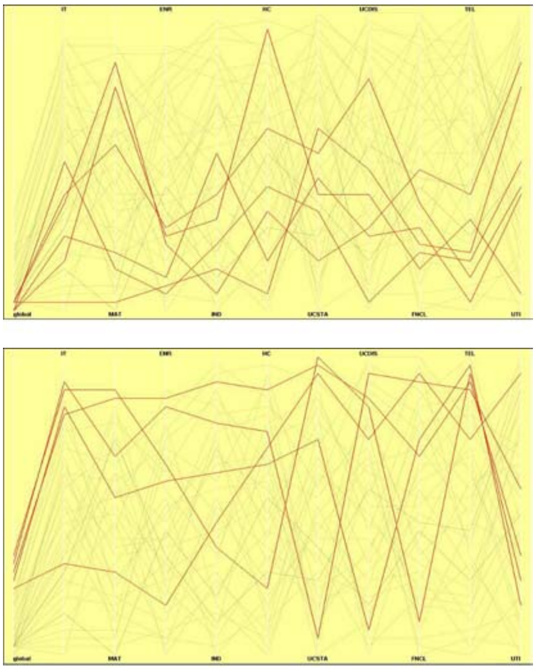


Figure 1: Parallel coordinates plot of the Global and sectorial MISE. Highlited are the polygons of both the 5 most passive Mutual Funds (top panel) and the 5 most active Mutual Funds (bottom panel).

Mutual Funds (bottom panel). Comparing the two panels of the figure, it is possible to note that passive Funds tend to show a relatively similar investment style, since the poligons of the corresponding parallel coordinates are quite homogeneous. Instead, investment style of active Funds is quite heterogeneous, showing that the asset allocation strategy for sectors like Consumer Staples (CSTA) and Financial (FNCL) tends to diverge from the relative benchmark constituents indexes, since the partial ranking (MISE) is high.

5 Final remarks

The analysis presented in the previous section revealed the effectiveness of the proposed method. First, Mutual Funds presenting a small MISE refers to an investment style which is coherent with the benchmark composition (and then with the financial market variations), and they should at least provide the same risk and return of the benchmark. Second, information about portfolio compositions can be considered in a perspective way, since the sector composition of both the benchmark and the Mutual Fund portfolios can be

used to compose a portfolio of Mutual Funds (MFMP) on the basis of a prespecified risk/return target.
The proposed method can be also applied for more analytical sectorial decompositions of the benchmark portfolio in order to rank in different ways Mutual Funds investment styles. With some minor changes the method can be extended to mixed Mutual Funds, in the case two or more benchmarks are used as representatives of the target market. Other results to be drawn are: a) the comparison of our rankings with other commonly used ranking criteria (like those provided by the rating agency Morningstar); b) the interactive visual analysis of the results of the method combined with risk/return measures; c) the different consideration of positive and negative deviations among the returns of the Mutual Funds and those of the benchmark in the computation of MISE; d) the investigation of the relations between the sectorial ranks and the global ranks.

References

[1] Sharpe W., Alexander G.J., Bailey V. (2000). *Fundamentals of investments.* Prentice-Hall.

[2] Gill P.E., Murray W., Wright M.H. (1981). *Practical optimization.* Academic Press, London, UK.

[3] Sharpe W. (1992). *Asset allocation: management style and performance measurement.* The Journal of Portfolio Management.

[4] Hardle W. (1992). *Applied nonparametric regression.* Cambridge University Press.

[5] Wegman E.J. (1990). *Hyperdimensional data analysis using parallel coordinates.* Journal of the American Statistical Association, **85**, 664 – 675.

[6] Theus M. (2004). *Interactive data visualization using mondrian.* Technical Report, University of Augsburg.

Acknowledgement: C. Conversano was supported by Research Funds of the University of Casssino, Dipartimento di Economia e Territorio. D. Vistocco was supported by MIUR Research Funds 2004 (P. I.: Prof. M.R. D'Esposito)

Address: C. Conversano, D. Vistocco, Dipartimento di Economia e Territorio,
Universitá di Cassino, Via Mazzaroppi, I-03043 Cassino (FR), Italy

E-mail: `c.conversano@unicas.it, vistocco@unicas.it`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

VISUALIZATION IN CLASSIFICATION PROBLEMS

D. Cook, D. Caragea and V. Honavar

Key words: Tour methods, classification problems, support vector machines.

COMPSTAT 2004 section: Data visualization.

Abstract: In the simplest form support vector machines (SVM) define a separating hyperplane between classes generated from a subset of cases, called support vectors. The support vectors "mark" the boundary between two classes. The result is an interpretable classifier, where the importance of the variables to the classification, is identified by the coefficients of the variables defining the hyperplane. This paper describes visual methods that can be used with classifiers to understand cluster structure in relation to known classes in data.

1 Introduction

The classification community has been overly focused on predictive accuracy. For many data mining tasks understanding a classification rule is as important as the accuracy of the rule itself. Going beyond the predictive accuracy to gain an understanding of the role different variables play in building the classifier provides an analyst with a deeper understanding of the processes leading to cluster structure. Ultimately this is our scientific goal, to solve a problem and understand the solution. With a deeper level of understanding researchers can more effectively pursue screening, preventive treatments and solutions to problems.

In the machine learning community, which is driving much of the current research into classification, the goal is that the computer operates independently to obtain the best solution. In data analysis, we're still a long way off this goal. Most algorithms will require a human user to twiddle with many parameters in order to arrive at a satisfactory solution. The human analyst is invaluable at the training phase of building a classifier.

This paper describes the use of graphics to build a better classifier based on support vector machines (SVM). We will plot classification boundaries in high-dimensional space and other key aspects of the SVM solution. The visual tools are based on manipulating projections of the data, and are generally described as tour methods. Our analysis is conducted on a particular data problem, the Italian olive oils data where the task is to classify oils into their geographic area of production based on the fatty acid composition. We focus on SVM because they operate by finding a hyperplane which maximizes the margin of separation between the two classes. This is similar to how we believe the eye perceives class boundaries. As a result of visualizing class structure in relation to SVM we have suggestions about how to find simpler but accurate classifiers.

2 Support vector machines

SVM is a binary classification method that takes as input labeled data from two classes and outputs a model for classifying new unlabeled data into one of those two classes. SVM can generate linear and non-linear models. In the linear case, the algorithm finds a separating hyperplane that maximizes the margin of separation between the two classes. The algorithm assigns a weight to each input point, but most of these weights are equal to zero. The points having non-zero weight are called *support vectors*. The separating hyperplane is defined as a weighted sum of support vectors. It can be written as, $\{\mathbf{x} : \ \mathbf{x}'\mathbf{w} + b = 0\}$, where $\mathbf{x}$ is the p-dimensional data vector, and $\mathbf{w} = \sum_{i=1}^{s} (\alpha_i \cdot y_i)\mathbf{x_i}$, where s is the number of support vectors, y_i is the known class for case $\mathbf{x}_i$, and α_i are the support vector coefficients that maximize the margin of separation between the two classes. SVM selects among the hyperplanes that correctly classify the training set, the one that minimizes $\|\mathbf{w}\|^2$, which is the same as the hyperplane for which the *margin* of separation between the two classes, measured along a line perpendicular to the hyperplane, is maximized. The classification for a new unlabeled point can be obtained from $f_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$.

We will be using the software SVM Light 3.50 [5] for the analysis. It is currently one of the most widely used implementations of SVM algorithm.

3 Tours methods for visualization

The classifier resulting from SVM, using linear kernels, is the normal to the separating hyperplane which is itself a 1-dimensional projection of the data space. Thus the natural way to examine the result is to look at the data projected into this direction. It may also be interesting to explore the neighborhood of this projection by changing the coefficients to the projection. This is available in a visualization technique called a manually-controlled tour.

Generally, tours display linear combinations (projections) of variables, $\mathbf{x}'\mathbf{A}$ where $\mathbf{A}$ is a $p \times d(< p)$-dimensional projection matrix. The columns of A are orthonormal. Often $d = 2$ because the display space on a computer screen is 2, but it can be 1 or 3, or any value between 1 and p. The earliest form of the tour presented the data in a continuous movie-like manner [1], but recent developments have provided guided tours [3] and manually controlled tours [2]. Here we are going to use a $d = 2$-dimensional manually-controlled tour to recreate the separating boundary between two groups in the data space.

We will be using the tour methods available in the data visualization package GGobi (`www.ggobi.org`).

4 The analysis of the Italian olive oils

To explain the visualization techniques we will use a data set on Italian olive oils. The olive oil data [4] contains 572 instances (cases) that have been chemically assayed for their fatty acid composition. There are 8 attributes (variables), corresponding to the percentage composition of 8 different fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic), 3 major classes corresponding to 3 major growing regions (North, South, Sardinia), and 9 sub-classes corresponding to areas in the major regions.

The data was collected for studying quality control of olive oil production. It is claimed that olive oils from different geographic regions can be distinguished by their fatty acid signature, so that forgeries can be recognized. The multiple classes create a challenge for classification, and the class structure is interesting: the classes have dramatically different variance structure, there are both linear and nonlinear boundaries between classes and some classes are difficult to separate. There are some surprises in the data.

4.1 Approach

Because there are 9 classes and SVM works only with 2 classes, we need to define an approach to building a classifier working in pairwise fashion. The most obvious start is to develop a classifier for separating the 3 regions, and then hierarchically work within region to build classifiers for separating areas.

In each classification we will run SVM. In the visualization we will highlight the cases that are chosen as support vectors and use the weights of the support vectors to construct the best separating projection. We will also examine the correlations between predicted values and variables to determine the importance of the variables for the classification.

4.2 Analysis

South vs Sardinia/North: This separation is too easy so it warrants only a short treatment. If the analyst blindly runs SVM with the oils from southern Italy in one class and the remaining oils in another class, then the result is a perfect classification, as shown in the plot of the predicted values on the horizontal axis of Figure 1. However if the analyst plots eicosenoic acid alone (vertical axis in Figure 1) she would notice that this also gives a good separation between the two classes. Eicosenoic acid alone is sufficient to separate these regions. When we examine the correlations between the predicted values and the variables we can see that although eicosenoic acid is the variable that is most correlated with predicted values that several other variables, palmitic, palmitoleic and oleic also contribute strongly to the prediction. Clearly, this classifier is too complicated. The simplest, most accurate rule would be to use only eicosenoic acid, and split the two classes by:

```
Assign to south if eicosenoic acid composition is more than 5%.
```

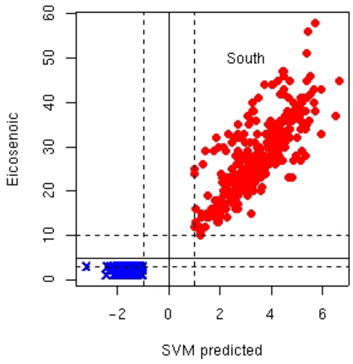
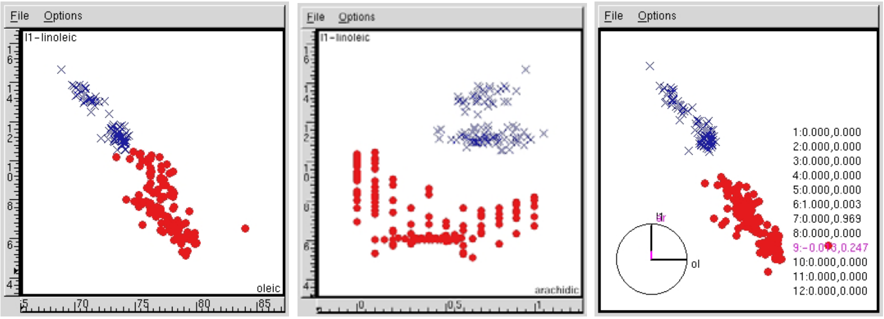


Figure 1: Examining the relationship between the results of the SVM classifier and one variable, eicosenoic acid. The horizontal axis displays the predicted values from the SVM classifier. The vertical axis displays the values of eicosenoic acid.

The minimum eicosenoic acid value for southern oils is 10%, and the maximum eicosenoic acid value for Sardinian and northern oils is 3%. Thus given the variance difference between the two groups a boundary at 5% makes sense. To obtain a solution numerically, the analyst could use eicosenoic acid alone in the SVM, quadratic discriminant analysis or logistic regression. Linear discriminant analysis does not give a good boundary because the two groups have very different variance.

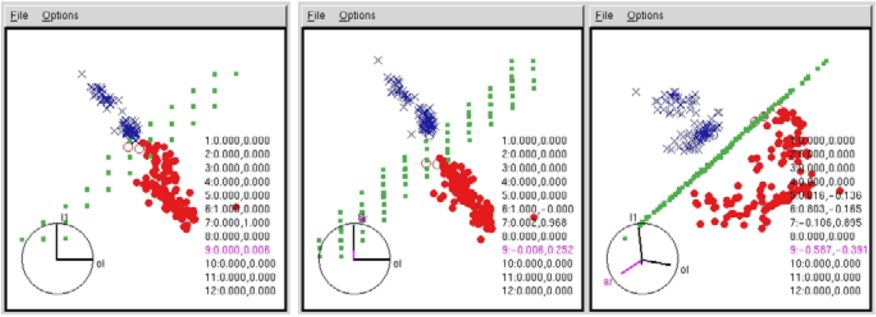
Sardinia vs North: This is an interesting classification problem, so it warrants an in-depth discussion. Plotting the data (Figure 2) two variables at a time reveals both a fuzzy linear classification (left) and a clear non-linear separation (middle) which would be difficult to model. A clean linear separation can be found by a linear combination of linoleic and arachidic acids (right). The circle at lower left in the plot is an axis that represents the combination of variables that are shown, and the numbers at right are the numerical values of the projection. Variable 7, linoleic, has a projection coefficient equal to 0.969, and variable 9, arachidic, has a projection coefficient equal to 0.247. This class boundary warrants some investigation. Very often explanations of SVM are accompanied by an idealized cartoon of two well-separated classes, with support vectors highlighted, and the separating hyperplane drawn as a line. We would like to reconstruct this picture in high-dimensional data space using tours to examine the classification results. To do this we need to turn the SVM output into visual elements. First, we generate a grid of points over the data space, select the grid points within a tolerance of the separating hyperplane. Then we will use the manually



```
1:0.000,0.000
2:0.000,0.000
3:0.000,0.000
4:0.000,0.000
5:0.000,0.000
6:1.000,0.003
7:0.000,0.969
8:0.000,0.000
9:-0.013,0.247
10:0.000,0.000
11:0.000,0.000
12:0.000,0.000
```

Figure 2: (Left) Close to linear separation in two variables, linoleic and oleic, (Middle) Clear nonlinear separation usng linoleic and arachidic acid, (Right) Good linear separation in combination of linoleic and arachidic acid in the vertical direction.

controlled tour to rotate the view until we find the projection of the hyperplane through the normal vector, where the hyperplane reduces to a straight line. Figure 3 shows the sequence of manual rotations made to find this view. Grid points on the hyperplane are small green squares. The large open circles are the support vectors. Samples from Sardinia are represented as blue crosses, and from northern Italy as solid red circles. The large circle at lower left is an axis where the radii represent the projection coefficient of the variables in the current plot. The numbers at the right are the numeric values of the projection coefficients for each variable. Purple text indicates the values for the variable just manipulated, rotated, into the projection. The rotation sequence follows the values provided by the weights, **w** and the



```
1:0.000,0.000
2:0.000,0.000
3:0.000,0.000
4:0.000,0.000
5:0.000,0.000
6:1.000,0.000
7:0.000,1.000
8:0.000,0.000
9:0.000,0.006
10:0.000,0.000
11:0.000,0.000
12:0.000,0.000
```

```
1:0.000,0.000
2:0.000,0.000
3:0.000,0.000
4:0.000,0.000
5:0.000,0.000
6:1.000,-0.000
7:0.002,0.968
8:0.000,0.000
9:-0.006,0.252
10:0.000,0.000
11:0.000,0.000
12:0.000,0.000
```

```
1:0.000,0.000
2:0.000,0.000
3:0.000,0.000
4:0.000,0.000
5:0.016,-0.136
6:0.803,-0.165
7:-0.106,0.895
8:0.000,0.000
9:-0.587,-0.391
10:0.000,0.000
11:0.000,0.000
12:0.000,0.000
```

Figure 3: Sequence of rotations that reveals the bounding hyperplane between oils from Sardinia and northern Italy: (Left) Starting from oleic vs linoleic, (Middle) Arachidic acid is rotated into the plot vertically, giving a combination of linoleic and arachidic acid, (Right) The separating hyperplane in the direction of the nonlinear separation.

correlations between predicted values and variables. According the the correlations linoleic is the most important variable followed by oleic, arachidic, palmitoleic, but stearic acid has a large weight value so we are curious about the contribution of this variable. We begin with the projection of the data into linoleic vs oleic acids because these two variables provided a simple linear boundary. Then other variables will be rotated into the plot in combination with linoleic to match the order of importance as provided by the weights. Since oleic acid has a low weight value we believed that its contribution to the separating hyperplane is mostly due to its strong correlation with linoleic acid. We were wrong. The separating hyperplane runs orthogonally to the linear relationship between oleic and linoleic, lower left to upper right rather than horizontally. Thus, it is clear that oleic acid is used even though its weight is small relative to linoleic acid. Next, arachidic acid is rotated into the plot, in combination with linoleic acid. Here the clear linear separation between the two classes can be seen, but it is not the projection corresponding to the SVM separating hyperplane. So, stearic acid is rotated into the plot in combination with linoleic and arachidic acid. Finally, the separating hyperplane is clearly visible, and the support vectors defining the hyperplane lie on opposing edges of the groups. Its clear that stearic acid is used in building the classifier. The boundary is too close to the northern oils. This surprised us, and we re-checked our planar computations several times, but this is the boundary. A quick fix could be obtained by adjusting the shift parameter value to shift the plane to closer to the middle of the separation between the two classes.

In general, the simplest but as accurate solution would be obtained by entering only two variables, linoleic and arachidic acids, into a classifier which should roughly give a rule as follows:

```
Assign to Sardinia if 0.97 x linoleic+2.5 x arachidic > 11%
```

This was obtained by using the projection coefficients provided by the tour, but it could just have easily been obtained by fitting the SVM model or some other classification model using only linoleic and arachidic acids.

North: The oils from the 3 areas in the north (Umbria, East/West Liguria) are difficult to separate. Working purely from graphical displays we might conclude the areas are not separable. But SVM tells us otherwise. The results are that the 3 areas can be perfectly separated using a linear kernel. So we attempt to find the projection of the data which separates the 3 areas. Working from the correlations between the variables and the predicted values, and from the weights for each variable in the separating hyperplane we rotate variables into view. Figure 4 displays the results. Sure enough, with a combination of most of the variables, a projection of the data where the 3 classes are separated can be found. The combination uses stearic, linoleic, linolenic and arachidic horizontally, and palmitoleic, stearic, linoleic and linolenic vertically. What we learn is that the SVM solution is about as

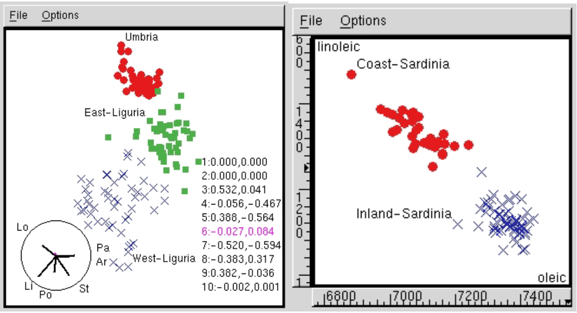


Figure 4: (Left) Projection of northern Italy oils. (Right) Projection of Sardinian oils.

good as possible, and although not easy to simplify using 5 of the 8 variables may provide an adequate classification.

Sardinia: The oils from coastal Sardinia are easily distinguishable from the inland oils using oleic and linoleic acids (Figure 4). A simple rule would be provided by: Assign to Inland Sardinia if $0.5 oleic - 0.75 linoleic > 27$.

South: Now this is the interesting region! There are 4 areas. There is no natural way to break down the 4 areas into a hierarchical pairwise grouping to build a classification scheme using SVM. The best but still poor results are obtained if the analyst first classifies Sicily against the rest. The remaining 3 areas can be almost perfectly separated using linear kernels. Why is this?

Looking at the data, using tours, with the oils from Sicily excluded, it is clear that the oils from the 3 other areas are readily separable (Figure 5, left). But when Sicily (orange open squares) is included it can be seen that the variation on the Sicilian oils is quite large (Figure 5, right). These points almost always overlap with the other oils in the tour projections. These pictures raised suspicions about Sicilian oils. It looks like they are a mix of the oils from the other 3 areas. Indeed, from informal enquiries, it seems that this is the case, that Sicilian olive oils are made from olives that are imported from neighboring areas. The solution is to exclude Sicilian oils from any analysis.

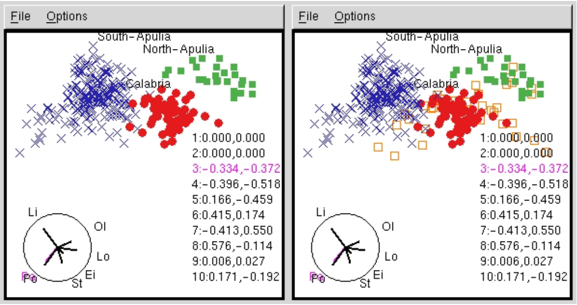


Figure 5: Projection of southern Italy oils.

Summary of analysis: In summary the concluding remarks about this data set are to beware of the oils designated from Sicily, as they do not appear to be pure oils. The remaining geographic production areas do seem to produce distinct fatty acid signatures, and its possible to quantify the differences with linear classifiers.

5 Summary and conclusion

SVMs provide a good complement to visual methods. The results from SVM are visually intuitive, unlike results from methods such as linear discriminant analysis where variance differences can produce a boundary that is too close to the group with the larger variance. Methods such as trees provide boundaries that are restricted to separations in individual variables. Logistic discriminant analysis, though, should compete with SVM and provide similarly accurate and interpretable linear boundaries.

As we raised in the introduction it is reasonable to be laborious in a training phase of building a classifier. Human input to the machine learning process can provide valuable insight into the scientific problem. With the visual tools described in this paper it is possible to visualize class structure in high-dimensional space and use this information to tailor better classifiers for a particular problem. We used just one example data set and one classification technique, but the approach works generally on other real-valued multivariate data, and for understanding other classification techniques.

References

[1] Asimov D. (1985). *The grand tour: a tool for viewing multidimensional data.* SIAM Journal of Scientific and Statistical Computing **6** (1), 128 – 11.

[2] Cook D., Buja A. (1997). *Manual controls for high-dimensional data projections.* Journal of Computational and Graphical Statistics **6** (4), 464 – 480.

[3] Cook D., Buja A., Cabrera J., Hurley C. (1995). *Grand tour and projection pursuit.* Journal of Computational and Graphical Statistics **4** (3), 155 – 172.

[4] Forina M., Armanino C., Lanteri S., Tiscornia E. (1983). *Classification of olive oils from their fatty acid composition.* 189 – 214.

[5] Joachims T. (1999). *Making large-scale SVM learning practical.*

Acknowledgement: This work has been supported in part by grants from the National Science Foundation (#9982341, #9972653), and the Iowa State University Graduate College.

Address: D. Cook, D. Caragea, V. Honavar, Iowa State University, Ames,
IA 50011

E-mail: {dicook,dcaragea,honavar}@iastate.edu

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

ANALYSIS OF THE MIB30 BASKET IN THE PERIOD 2000-2002 BY FUNCTIONAL PC'S

Damiana G. Costanzo and Salvatore Ingrassia

Key words: Functional data, principal components, financial data.

COMPSTAT 2004 section: Functional data analysis.

Abstract: The MIB30 basket refers to the 30 most capitalised and traded companies on the Italian Stock Exchange. Related share prices and related quantities are updated during the phase of continuous trading at a frequency of one a minute on the basis of the prices of the latest contracts concluded on each share. The daily traded volumes of these 30 shares in the period from January 3rd, 2000 to December 30th, 2002 are here investigated from an explorative point of view using functional principal component techniques.

1 Introduction

Functional data are essentially curves and trajectories, the basic rationale is that we should think of observed data functions as single entities rather than merely a sequence of individual observations. Even though functional data analysis often deals with temporal data, its scope and objectives are quite different from time series analysis: while time series analysis mainly focuses on modeling data, or in predicting future observations, the techniques developed in FDA are essentially exploratory in nature: the emphasis is on trajectories and shapes; moreover unequally-spaced and/or different number of observations can be taken into account as well as series of observations with missing values, see Ramsay & Silverman [5], [6].

In this paper statistical properties of the daily series of the traded volumes of the shares composing the MIB30 basket in the period from January 3rd, 2000 to December 30th, 2002 are investigated from a functional data analysis perspective. We remark that the MIB30 basket synthesizes the performance of the Italian Stock Exchange within the Telematic Share Market; at the beginning, in 1975, the term MIB was the acronym of "Milano Indice Borsa"; subsequently it took on the new meaning of "Mercato Italiano di Borsa" since the market had become effectively national.

The rest of the paper is organised as follows: in the next section we outline functional data modeling and give some details about functional principal component analysis; in Section 3 we introduce the MIB30 basket dataset and present the results of our analysis; finally in Section 4 we compare the dynamics of the first functional PC with the dynamics of the MIB30 index in order to explore the possibility of the construction of stock market indices based on functional indicators.

2 Functional PCA

Let $\{\omega_1, \ldots, \omega_n\}$ be a set of n units and let $\mathbf{y}_i = (y_i(t_1), \ldots, y_i(t_p))$ be a sample of measurements of a variable Y taken at p times $t_1, \ldots, t_p \in \mathcal{T} = [a, b]$ in the i-th unit ω_i, $(i = 1, \ldots, n)$. Such data $\mathbf{y}_i$ $(i = 1, \ldots, n)$ are regarded as *functional* because they are considered as single entities rather than merely sequences of individual observations, so they are called *raw functional data*; indeed the term functional refers to the intrinsic structure of the data rather than their explicit form. In order to convert raw functional data into a suitable functional form, a smooth function $x_i(t)$ is assumed to lie behind $\mathbf{y}_i$ which is referred to as the *true functional form*; this implies, in principle, that we can evaluate x at any point $t \in \mathcal{T}$. The set $\mathcal{X}_\mathcal{T} = \{x_1(t), \ldots, x_n(t)\}_{t\in\mathcal{T}}$ is the *functional dataset.*

In functional data analysis the statistical techniques posit a vector space of real-valued functions defined on a closed interval for which the integral of their squares is finite. If attention is confined to functions having finite norms, then the resulting space is a Hilbert space; however we often require a stronger assumption so we assume $\mathcal{H}$ be a *reproducing kernel Hilbert space* (r.k.h.s.), which is a Hilbert space of real-valued functions on $\mathcal{T}$ with the property that, for each $t \in \mathcal{T}$, the evaluation functional L_t, which associates f with $f(t)$, $L_t f \to f(t)$, is a bounded linear functional.

In such spaces the objective in principal component analysis of functional data is the orthogonal decomposition of the empirical variance function:

$$v(t,u) := \frac{1}{n-1}\sum_{i=1}^{n}\{x_i(t) - \overline{x}(t)\}\{x_i(u) - \overline{x}(u)\} \tag{1}$$

(which is the counterpart of the covariance matrix of a multidimensional dataset) in order to isolate the dominant components of functional variation.

In analogy with the multivariate case, the functional PCA problem is characterized by the decomposition of the variance function:

$$v(t,u) = \sum_j \lambda_j \xi_j(t)\xi_j(u) \tag{2}$$

where $\lambda_j, \xi_j(t)$ satisfy the eigenequation: $\langle v(s,), \xi_j\rangle_h = \lambda_j \xi_j(u)$, where the eigenvalues:

$$\lambda_j := \int_\mathcal{T} \xi_j(t)v(t,u)\xi_j(u)dt\,du$$

are positive and non decreasing while the eigenfunctions must satisfy the constraints:

$$\int_\mathcal{T} \xi_j^2(t)dt = 1 \qquad \text{and} \qquad \int_\mathcal{T} \xi_j\xi_i(t)dt = 0 \qquad (i < j).$$

The ξ_j*'s* are usually called *principal component weight functions.* Finally the principal component scores (of $\xi(t)$) of the units in the dataset are the values w_i given by:

$$w_i^{(j)} := \langle x_i, \xi_j \rangle = \int_{\mathcal{T}} \xi(t) x_i(t) dt \, . \tag{3}$$

The decomposition (2) defined by the eigenequation (2) permits a reduced rank least squares approximation to the empirical covariance function v. Thus, the leading eigenfunctions ξ define the principal components of variation among the sample functions x_i.

3 A functional PC analysis of the MIB30 basket dataset

Raw data considered here consist of the total value of the traded volumes of the shares composing the MIB30 basket in the period January 3rd, 2000 - December 30th, 2002. They have been collected in a 30 × 758 matrix. We remark that an important characteristic of this basket is that it is "open" in that the composition of the index is normally updated twice a year, in the months of March and September (ordinary revisions). Moreover, in response to extraordinary events, or for technical reasons ordinary revisions may be brought forward or postponed with respect to the scheduled date.

In particular in our data set there are 21 companies which have remained in the basket for the three years while the other 9 places in the basket have been shared by a set of other companies which have been remaining in the basket for shorter periods. Due to the connection among the international financial markets, data concerning the closing days (as week-ends and holidays) are regarded here as missing data.

In literature functional PCA is usually performed from original data (x_{ij}); here we prefer to work on the daily standardized raw functional data:

$$z_{ij} := \frac{x_{ij} - \bar{x_j}}{s_j} \qquad (i = 1, \ldots, 30, \quad j = 1, \ldots, 758) \, , \tag{4}$$

where $\bar{x_j}$ and s_j are respectively the daily mean and standard deviation of the e.e.v of the shares in the basket. We shall exhibit later how such transformation can gain an insight into the PC trajectories understanding.

The first PC alone accounts for the 89.4% and the second PC accounts for the 6.9% of the whole variability. In Figure 1 we give the trajectories of the first two functional principal components which show the way in which such set of functional data varies from its mean, and, in terms of these modes of variability, quantifies the discrepancy from the mean of each individual functional datum.

The meaning of functional principal component analysis is a more complicated task than the usual multidimensional analysis, however here it emerges the following interpretation:

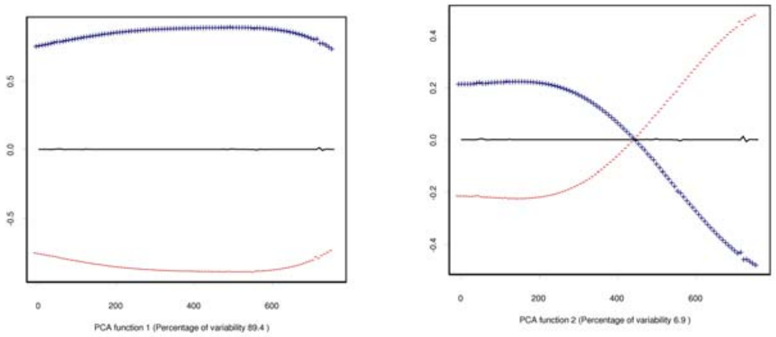


Figure 1: Plot of the first 2 functional principal components.

i. The first functional PC is always positive, then shares with large scores of this component during the considered period have a large traded volume as compared to the mean value on the basket; it can be interpreted as a *long term trend component.*

ii. The second functional PC changes sign at $t = 431$ which corresponds to September 11th, 2001 and the final values, in absolute value, are greater than the initial values: this means that shares having good (bad) performances before September 11th, 2001 have been going down(rising) after this date; it can be interpreted as a *shock component.*

This interpretation is confirmed by the following analysis of the raw data. As it concerns the first PC, for each company we considered its minimum standardized value over the three years $z_i^{(\min)} = \min_{j=1,\dots,758} z_{ij}$ $(i = 1, \dots, 30)$. In particular $z_i^{(\min)}$ is positive (negative) when the traded volumes of the i-th share are always greater (less) than the mean value of the MIB30 basket during the three years.

As for the second PC, let $\bar{x}_{Bi}$ be the average of the traded volumes of the ith company over the days: 1,...,431 (i.e. before September 11th, 2001) and $\bar{x}_{Ai}$ the corresponding mean value after September 11th, 2001. Let us consider the variation per cent:

$$\delta_i := \frac{\bar{x}_{Ai} - \bar{x}_{Bi}}{\bar{x}_{Bi}} 100\% \qquad i = 1, \dots, 30 .$$

If δ_i is positive (negative) then the ith company increased (decreased) its mean e.e.v. after the September 11, 2001.

Consider the scores on the two first PCs given in (3), respectively $\mathbf{w}_i^{(1)}$ and $\mathbf{w}_i^{(2)}$. We observed that the correlation coefficient between the $z_i^{(\min)}$ and $\mathbf{w}_i^{(1)}$ is equal to 0.96 and the correlation coefficient between the δ_i and $\mathbf{w}_i^{(2)}$ is equal to -0.84, see Ingrassia and Costanzo [3] for details.

4 A comparison between the dynamics of the first FPC and the MIB30 index

In our opinion, the obtained results open methodological perspectives for the construction of new financial indices having some suitable statistical properties. As a matter of fact, the construction of some existing stock market indices has been criticized by several authors, see e.g. Elton and Gruber [2].

In Costanzo [1] the dynamics of the MIB30 index in the three years 2000, 2001 and 2002 have been investigated using the *phase-plane plot*; here we compare such dynamics with the ones coming from the first functional. This technique may provide information about pure dynamics of the event and compare it within time, since it focuses as on the rate of change of the index rather than its actual size. The functional index is regarded as a harmonic process in which energy is exchanged between the potential and the kinetic states; thus, as mentioned above, the phase-plane plot draws the acceleration against the velocity. As a matter of fact, the kinetic energy is proportional to the square of the velocity (i.e. the first derivative of the function) and the potential energy is proportional to the acceleration (i.e. the second derivative). In economics, potential energy corresponds to available capital, human resources, raw material and other resources that are at hand to be used in economic activity; kinetic energy corresponds to the manufacturing process in full swing, when these resources are moving along the assembly line (see [6]). For a general dynamic process, velocity represents its rate of change, while acceleration indicates the input or whatever resources or forces produce this change. Thus, in financial markets, the potential energy may correspond to "strength" of the economic situation, both the realistic one and the one perceived by financial operators. The kinetic energy corresponds to the confidence of the financial operators in economy in such "strength". In fact, a financial index reflects the confidence of the financial operators in the economic situation: when the markets are confident, the value of the index tends to increase and exhibits stable patterns over time; shocks, as wars or dramatic events, produce both short transitory effects and longer-lasting readjustments in market behaviour. In the phase-plane we plot acceleration on the vertical axis, versus the velocity on the horizontal axis. The interpretation of this graphical representation follows Ramsay and Silverman [6], Ramsay and Ramsey [4].

In Figure 2 (left column) we give the phase plane plots of the MIB30 index in the three years 2000, 2001 and 2002. Concerning the year 2000, we note three large cycles surrounding zero, plus two small cycles that are much closer to the origin. The largest one starts in the beginning of February (Feb), with positive velocity but near zero acceleration. In this period the velocity of the index reaches its absolute maximum value in the year; the other maximum value of the velocity is roughly at the beginning of December. As remarked before, the size of the radius of the cycles is an important aspect of the plot. When the markets consider the economic situation good, they react with high level of confidence: the level of trading increases and the rate of change of the index is high. This large cycle ends to the middle of April (Apr). The cusp corresponding to the smallest cycle starts at the middle of May and lasts for about two months, until the middle of July (Jul). During this period both velocity and acceleration are very low, near zero. Note that the location of the center of this smallest cycle indicates a positive and decreasing velocity. The second largest cycle starts at the middle of October (Oct) and it continues until the end of the year.

As for the year 2001 there is just one evident large cycle. It is represented by the intermediate cycle starting in January till about the end of April. Note that its horizontal center is located to the left of the plot. The subsequent smaller cycle moves from the middle of May till the end of July. The largest cycle starts at the beginning of August with negative velocity and acceleration. The cycle moves clockwise through August and passes the horizontal zero acceleration line around the end of the month. We see now in the plot a bulge to the left starting from the middle of September, roughly when the catastrophic event took place: acceleration reached its maximum value but it was not followed by a positive velocity. Velocity stays negative until the beginning of October. This cycle closes at the end of October with very low velocity of the index. Note that in this plot the cusp now corresponds to the cycle spanning the last two months of the year and it is very small in size, with very low velocity and acceleration.

To summarize, the year 2001 started with a not very good performance of the index in terms of its rate of change, this is shown by the first two cycles corresponding to the first six/seven months of the year in the plot. After that it started to perform better; note, in fact, how the largest cycle is characterized by net negative velocity for its evident horizontal location of the center to the left, but at the same time net velocity increases for its vertical location of the center is above zero. The shock event in September interrupted some way this better performance: the year closed with a very small cycle.

Finally for the year 2002, we remark the rate of change of the index lasts in the following year. Note indeed the impressive change in the cycles in the year 2002 from the year 2000. Almost all cycles shift to the left; that is, they have now their horizontal centers on the left, which denotes a net negative velocity almost during the whole year. Further, the size of their radius with respect to the horizontal axes of the plot, denotes lower velocity of the index compared to the year 2000. Note that the cusp occurs again in the same period mid May - mid July, as in year 2000, but it has dramatically shifted to the left. Strictly speaking in year 2002, as a consequence of September 11th in the previous year, the overall slope of the index becomes more negative and the amplitude of the short term cycles shrinks.

The phase plane plot of the first PC concerning the years 2000, 2001 and 2002 is given in Figure 2 (right column). The interpretation of the plot is analogous as in the MIB30 case. However, the first PC "index" summarizes in a clearer way than the MIB30 index the dynamics of the events in the three years here considered. In fact, concerning the year 2000 the plot shows a very large cycle with a positive velocity during all the year but with an acceleration which is positive until the end of February only. Following the economic interpretation previously outlined, this kind of circle denotes a positive trend which is characterized however by a kinetic energy (the confidence of the financial operators in the strength of the economic situation) which starts to decrease already in the first part of the year. Concerning the year 2001, it is strongly evident in the plot the shock event of September 11th, since velocity even though positive during all the year, it is strongly increasing before that date then it is rapidly decreasing. For the year 2002 we remark the consequence of the last event in the rate of change of the first PC index. In fact the phase plane plot shows negative and decreasing both velocity and acceleration.

The results here presented suggest that the shares scores on this harmonic could be a good ingredient for a new family of financial indices trying to capture as most as possible of the variability of the prices in the share basket. This provides ideas for further developments of functional principal component techniques in the financial field.

References

[1] Costanzo G.D. (2003). *A graphical analysis of the dynamics of the MIB30 index in the period 2000-2002 by a functional data approach.* In: Atti Riunione Scientifica SIS 2003, Rocco Curto Editore, Napoli, 133 − 144.

[2] Elton E.J. and Gruber M.J. (1995). *Modern Portfolio Theory and Investment Analysis.* John Wiley & Sons, New York.

[3] Ingrassia S. and Costanzo G.D. (2003). *Functional principal component analysis of financial time series.* Book of Short papers of CLADAG 2003, 207 − 210.

[4] Ramsay J.O. and Ramsey J.B. (2002). *Functional data analysis of the dynamics of the monthly index on nondurable good production.* Journal of Econometrics, **107**, 327 − 344

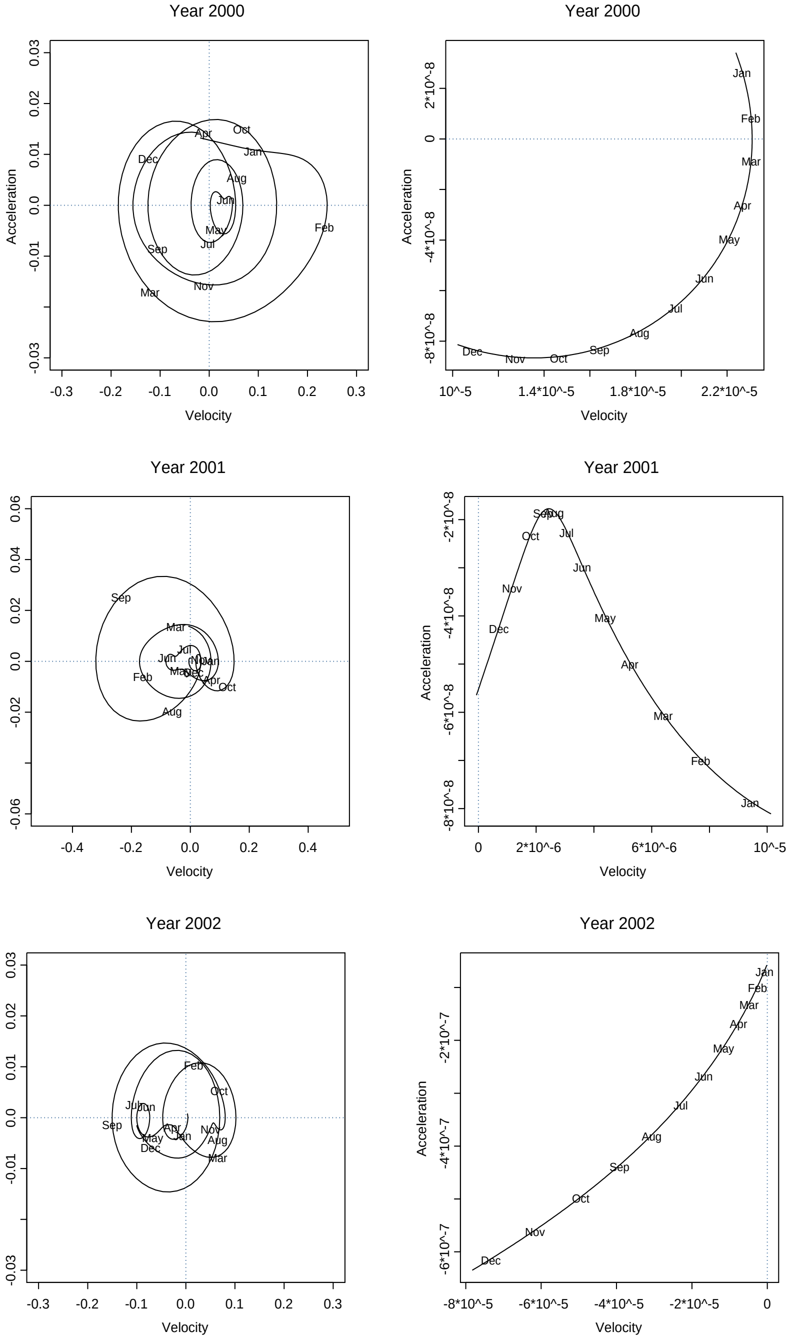
[5] Ramsay J.O. and Silverman B.W. (1997). *Functional Data Analysis.* Springer-Verlag, New York.

[6] Ramsay J.O. and Silverman B.W. (2002). *Applied Functional Data Analysis.* Springer-Verlag, New York.

Acknowledgement: Dataset used in this paper have been collected by the Italian Stock Exchange. The authors thank Research & Development DBMS (Borsa Italiana).

Address: G.D. Costanzo, S. Ingrassia, Dipartimento di Economia e Statistica, Università della Calabria

E-mail: `[dm.costanzo, s.ingrassia]@unical.it`



Year 2000

Year 2000

Year 2001

Year 2001

Year 2002

Year 2002

Figure 2: *Phase plane plot of the Mib30 index daily series (left column) and of the first functional PC of the MIB3 basket (right column) for the years 2000,2001 and 2002 (note the different scale of the acceleration in the plot concerning the year 2001).*

COMPSTAT'2004 Symposium

BAGGING A STACKED CLASSIFIER

**Christophe Croux, Kristel Joossens
and Aurelie Lemmens**

Key words: Classification, cross-validation, machine learning, resampling.

COMPSTAT 2004 section: Classification.

Abstract: Combining several classifiers to achieve a better predictive performance can be done by stacking. In this paper, we propose an algorithm to find an optimal weighted average of the different classifiers, yielding a systematically better cross-validated error rate on the training data. This combined classifier is then compared to other stacking methods. Furthermore, we study the potential of bagging as a way to improve the performance of stacked classifiers.

1 Introduction

Suppose that one needs to classify an incoming email as being spam or not. For doing so, several supervised two-class classification methods (e.g. logistic regression, k-nearest neighbors, neural networks, decision trees, discriminant analysis) can be used. However, it can be time consuming and quite difficult to select the most appropriate classifier. Each classifier has indeed its own advantages and disadvantages and could well be adapted to one specific task, but not recommended for another. This issue can be solved by combining, in a fully-automated way, different classifiers. The combined classifier should perform better than each single classifier.

In the classification literature, several classifier combination methods have already been proposed and can be categorized in two groups. Some algorithms combine classifiers of the same type. For example, the boosted tree is a linear combination of several decision trees (e.g. Hastie et al. [2]). A second category of combination methods handles classifiers of different types. Within this second class, we focus on a method introduced by Wolpert [5] in the neural network literature and called stacked generalization. In the statistical literature this method was considered by LeBlanc and Tibshirani [3], who called it stacked classification. A brief review of the stacking principle will be given in Section 2.

In this paper, we propose a new, easy-to-implement and flexible algorithm to combine different classifiers. It is constructed in such a way that the final error rate, as estimated by ten-fold cross-validation, will be minimized. Furthermore, in Section 3, we will apply the bagging technique to reduce the variability of the stacked classifier, hereby further improving its performance.

2 Stacked classification

Consider a training dataset $Z = \{(x_i, y_i) | 1 \leq i \leq n\}$, containing n observations. For the training data, we observe the values of the binary variable Y, indicating to which source population an observation belongs to. This variable Y needs to be predicted using X, a vector of p explicative variables. The realizations of X on the training set are $x_1, \ldots, x_n$. The aim is to predict the value of Y for a new instance x. A classifier $C(x; Z)$ is constructed on the basis of the training set. It takes values on the interval $[0, 1]$: the higher the value of $C(x; Z)$, the higher the likelihood that an observation with features x belongs to the source population with label "Y=1". The value of $C(x; Z)$ is called the "score" attached to x. Classifiers taking only the values 0 and 1 can be thought of as a special case were only the most extreme values are given as score.

Several supervised classification rules exist. In this paper, four well-known methods are considered: classification trees, discriminant analysis, logistic regression and k-nearest neighbors. Classification trees yield a non-parametric classification procedure, and are widely used in applied fields such as medicine and botany. Linear discriminant analysis is the most traditional supervised classification method, optimal if both source populations are normal with the same covariance matrix. It is a parametric method, where the scores are given by linear combinations of the explicative variables. Logistic regression is another parametric classification technique where the conditional probabilities $P(Y = 1|x)$ are modeled by means of the logit transformation. Finally, the k-nearest neighbors method consists of finding the k-closest neighbors to an observation x (according to a certain distance). The value $C(x; Z)$ is then given by the frequency of observations for which $y_i = 1$, among the k-nearest neighbors (we select $k = \sqrt{n}$). The different classifiers have been computed using the standard S-plus implementation. Of course, other methods such as Support Vector Machines or Random Forests, could also have been taken as level-zero classifiers.

It is well recognized that there does not exist a unique best classifier. Suppose that K different initial classifiers $C_1, \ldots, C_K$ are available. These initial classifiers can be combined in one single stacked classifier. Wolpert [5] called the initial classifiers the "level-zero" components. The idea is to use the outcomes of the "level-zero" classifiers as inputs of the "level-one" classifier. In other words, the training dataset for this "level-one" classifier is given by $\{(C^*(x_i; Z), y_i) | 1 \leq i \leq n\}$, where $C^*(x_i; Z) = (C_1(x_i; Z), \ldots, C_K(x_i; Z))^t$ is obtained by stacking the different level-zero classifiers. As level-one classifiers, one could again consider logistic regression, a decision tree etc.

Wolpert [5] and Leblanc and Tibshirani [3] advise to use predicted scores obtained by cross-validation instead of the $C_j(x_i, X)$, for $j = 1, \ldots, K$ and $i = 1, \ldots, n$. As such, less overfitting should occur, leading to a better performance on future data to classify. Moreover, the cross-validation reduces the risk of having high correlation between the level-zero classifiers. Hence, throughout this paper, the level-one classifiers will be constructed from the stacked vector $C^*(x_i; Z) = (C_1^-(x_i; Z), \ldots, C_K^-(x_i; Z))^t$, where $C_j^-(x_i; X)$ is obtained via cross-validation for $j = 1, \ldots, K$. To reduce the computation time, the $C_j^-(x_i; Z)$ were computed via 10-fold cross-validation. Formally, the training dataset Z is split into ten subsets, Z_1, …, Z_{10}. Predictions for the observations belonging to a subset Z_l are then based on a classifier constructed with $Z \setminus Z_l$ as training set, for $l = 1, \ldots, 10$.

The algorithm for obtaining an "optimal" level-one classifier will be outlined in the next subsection.

2.1 A new algorithm for combining classifiers

We restrict attention to stacked classifiers obtained as linear combinations of the component classifiers:

$$C_W(x; Z) = \beta_1 C_1(x; Z) + \cdots + \beta_K C_K(x; Z). \tag{1}$$

The coefficients $\beta_1, \ldots, \beta_K$ need to be chosen in such a way that the stacked classifier $C_W(x; Z)$ performs "better" than the initial classifiers. We require these coefficients to sum up to one and to be non-negative, such that they can be interpreted as the weights attached to each classifier. If a component classifier performs particularly bad, it receives a low or even zero weight, and has no impact on the final classification. The combined classifier C_W is therefore an "optimal" *weighted* average of the component classifiers. The weights $\beta_1, \ldots, \beta_K$ are now selected in such a way that they optimize a given criterion, such as the *error rate*, i.e. the percentage of misclassified observations. Other criteria do exist, such as the area under the receiver operating curve (ROC). The algorithm proposed in this paper works for any criterion of choice, and can be used for any set of component classifiers, making it very flexible. In this short paper, we restrict attention to the error rate as performance criterion. Note that Leblanc and Tibshirani [3] worked with the mean squared error as criterion, a natural choice in the regression setting but less comon in the classification literature. Since the population error rate is unknown, we work with the (10-fold) cross-validated error rate as an estimate of it. This latter is known to be an unbiased estimator of the error rate, in contrast to the apparent error rate. Finding the optimal values for $\beta_1, \ldots, \beta_K$ is not easy, certainly not for high values of K. Hence, we propose the following greedy algorithm:

- Compute, via 10-fold cross validation, the scores for every component classifier $C_j^-(x_1; Z), \ldots, C_j^-(x_n; Z)$ for $j = 1, \ldots, K$.
- Using the previously computed scores, compute the cross-validated error rate associated with the classifiers (without prior information, classify an observation in the first source population if the value of the score exceeds the threshold 0.5). These are used to sort the classifiers from

the smallest to the highest error rate. We denote the ordered classifiers as $C_{(1)}, \ldots, C_{(K)}$.

- Maximizing the criterion over the space of all possible values of $\beta_1, \ldots, \beta_K$ is difficult, since the objective function to minimize might be non differentiable. We propose an iterative procedure. First find α_1 such that the error rate of the combined classifier

$$C^1(x;Z) = \alpha_1 C_{(1)}(x;Z) + (1-\alpha_1)\, C_{(2)}(x;Z)$$

is minimized, where α_1 ranges over the interval $[0,1]$. A simple univariate optimization routine, like grid search, can be used. For $k = 2$ to $K-1$, find α_k such that the error rate of

$$C^k(x;Z) = \alpha_k C^{k-1}(x;Z) + (1-\alpha_k)\, C_{(k+1)}(x;Z)$$

is minimized. After a first cycle is completed, one could rearrange the terms and express C^{K-1} as a convex combination of all level-zero classifiers

$$\beta_1 C_1(x;Z) + \cdots + \beta_K C_K(x;Z).$$

A new cycle might then be computed, finding successively the best convex combination of $C_{(j)}$ and C^b, for $j = 1, \ldots, K$, where C^b is the currently "best" combination of level-zero classifiers. In the present version of the program, we perform three full cycles after which no significant further improvement was observed.

The coefficient $\beta_1, \ldots, \beta_K$ obtained in this way are used as weights for the final classifier C_W in (1). By construction, the classifier C_W always has a lower (10-fold cross-validated) error rate than each single level-zero classifier. Note that there is no guarantee, however, that the global optimum is reached.

To illustrate this procedure, we applied the greedy algorithm for finding C_W to 12 datasets available on the UCI repository [4]. The size n and the number of explicative variables p of these datasets are reported in Table 1. The datasets are previously cleaned from missing values. Each dataset is split into a training set Z (80% of the observations) and a test set T (20% of the observations). The training set Z is used to obtain both the level-zero classifiers and the stacked classifiers. The test sets will be used in the sequel.

Table 1 reports the cross-validated error rates for every level-zero classifier (classification tree, discriminant analysis, logistic regression, k-nearest neighbors (NN)) and for the optimal weighted average C_W of these four classifiers. No level-zero classifier clearly outperforms the others, justifying the choice for combining classifiers. As already mentioned 2, the error rates obtained by the C_W on the training sample are by construction always lower than those associated with the initial classifiers. The improvement is quite

				Level-Zero Classifiers				Stacked
	n	p		Tree	Disc	Logit	NN	W
1	690	14	Austral	17.03	13.59	12.50	28.80	11.78
2	156	5	Balloon	10.48	10.48	8.87	8.87	7.26
3	699	9	Breast	5.90	3.94	4.11	3.22	3.22
4	1.473	9	Cmc	30.31	32.60	32.34	28.27	27.16
5	653	15	Crx	18.20	12.84	13.60	31.23	12.26
6	351	33	Iono	12.50	14.29	15.36	16.07	8.57
7	8.124	21	Mushroom	0.15	6.03	3.46	4.39	0.06
8	768	8	Pimadiab	26.87	22.64	22.80	27.20	22.31
9	4.601	57	Spambase	8.78	11.68	7.28	25.38	6.03
10	958	9	Tictacto	17.10	32.64	32.90	14.88	12.79
11	569	30	Wdbc	6.81	4.40	5.93	6.81	3.30
12	198	31	Wpbc	26.58	20.25	22.78	22.78	17.09

Table 1: Cross-validated error rates (in %) of the level-zero classifiers and the optimal weighted average C_W. The first two columns contain the number of observations n and number of variables p of each dataset.

substantial, amounting until 31.44 % (Iono data) or even 60.00 % (Mushroom data) of relative improvement in error rate with respect to the best level-zero classifier. In one case (the Breast data), the k-nearest neighbors classifier performs equally as C_W. In this case, the weight associated with the k-nearest neighbors classifier turns out to be one.

At first sight, the results are very encouraging. However, one should not forget that the weights of the combined classifier are determined to minimize the cross-validated error rate computed on the training data. There is a risk that the reported error rates, albeit obtained by cross-validation, are still over-optimistic. Therefore, in the following subsection, we compute the error rates on the *test set*, and compare different level-one classifiers.

2.2 Empirical comparison of different stacking methods

Table 2 reports the error rates computed on the test set for the level-zero classifiers and the optimal weighted average C_W. One may observe from Table 2 that the weighted method does not systematically provide the lowest test error rates, even if it is still the case for 8 out of 12 datasets. When a single component performs better, the weighted combination has an error rate very close to it. Moreover, for each level-zero classifier, it is possible to find at least one dataset on which the level-zero classifier performs drastically worse than the stacked classifier C_W.

Table 2 also reports the results of other stacking methods. The level-one combination methods under consideration include a decision tree, linear discriminant analysis, logistic regression and k-nearest neighbors. One may

	Level-Zero Classifiers				Stacked Classfiers				
	Tree	Disc	Logit	NN	W	Tree	Disc	Logit	NN
1	14.4	15.9	14.4	31.1	15.9	20.2	14.4	13.7	12.3
2	9.38	9.38	9.38	9.38	9.38	12.5	12.5	9.38	9.38
3	6.43	9.29	8.57	6.43	6.43	4.29	6.43	5.71	7.14
4	29.8	32.8	32.2	28.1	28.8	32.8	29.4	28.4	28.4
5	17.5	16.0	14.5	33.5	16.0	16.7	16.0	15.2	15.2
6	19.7	14.0	12.6	16.9	12.6	9.86	11.2	9.86	11.2
7	0.25	5.05	2.71	4.62	0.12	0.25	0.25	0.12	0.25
8	25.9	20.7	20.7	21.4	20.7	27.2	21.4	22.0	20.7
9	6.19	10.1	7.17	24.65	5.54	4.89	5.32	5.32	5.97
10	13.5	26.5	26.0	13.0	11.9	8.33	7.29	7.81	5.73
11	2.63	2.63	6.14	6.14	3.51	1.75	2.63	0.88	3.51
12	27.5	20.0	17.5	27.5	15.0	22.5	12.5	12.5	15.0

Table 2: Test error rates in % for the 4 different level-zero classifiers, for the optimal weighted average C_W, and for stacking by 4 different level-1 classifiers.

observe that the different ways of combining the level-zero estimators yield different error rates, even if this heterogeneity is less pronounced than among the level-zero classifiers. The stacked logistic regression seems to be the best stacking method on the datasets under investigation, giving on the whole the smallest test error rates. Note that the weighted average does not outperform the other stacking methods anymore on the test set. In the next section, we will apply bagging on the stacked classifiers.

3 Bagging the stacked classifiers

Error rates esimated by cross-validation are known to be very volatile. The bagging technique, originating from machine learning, helps to reduce the variance and improves the performance of unstable classifiers [1]. Therefore, we propose to apply the bagging algorithm to the weighted classifier C_W, described in section 2.1. Practically, the bagging procedure consists of computing classifiers on B different bootstrap samples of size n drawn from the initial training set Z. These bootstrap samples are obtained by drawing with replacement observations from Z, yielding $Z^{*1},\ldots,Z^{*b}$. The bagged classifier is then obtained by averaging over all predicted scores:

$$C_{bag,B}(x;Z)=\frac{1}{B}\sum_{b=1}^{B}C_W(x;Z^{*b}).$$

Bagging is an easy-to-implement procedure and can be applied to any given classification method.

In Figure 1, the test error of $C_{bag,B}$ is plotted with respect to B, for

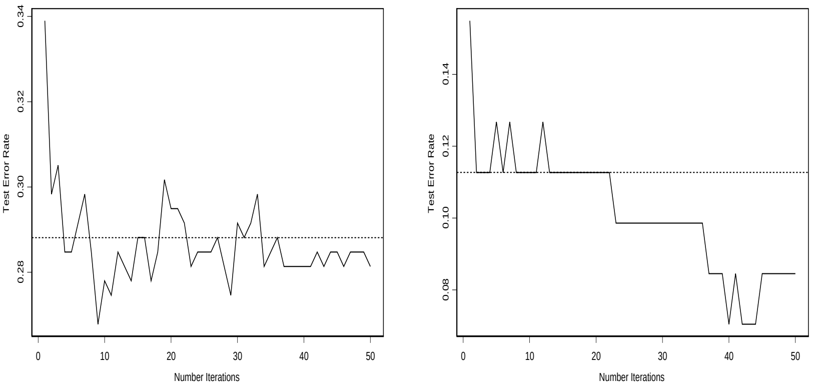


Figure 1: Test error rates w.r.t. the number of bootstrap steps B for stacked classification using as level one classifier (left panel) the weighted average C_W for the Cmc dataset (right panel) discriminant analysis for the Iono dataset. The dotted line corresponds to the error rate for the classification method without bagging.

	Bagged Stacked Classifiers				
	W	Tree	Disc	Logit	NN
1	11.5	12.3	12.3	12.32	12.3
2	9.38	9.38	12.5	9.38	12.5
3	5.71	4.29	5.00	5.00	6.43
4	28.1	28.8	28.4	28.4	28.8
5	14.5	14.5	14.5	14.5	14.5
6	7.04	8.45	8.45	7.04	7.04
7	0.00	0.12	0.12	0.00	0.12
8	21.4	22.0	20.7	20.1	23.3
9	5.10	4.89	5.32	5.21	5.10
10	7.81	3.13	3.13	4.69	5.21
11	1.75	0.88	0.88	0.88	0.88
12	12.5	15.0	15.0	15.0	15.0

Table 3: Test error rates of several bagged stacked classifiers (with $B = 50$): the weighted average C_W together with stacking by 4 different level-1 classifiers.

$B = 1, \ldots, 50$. In the left panel, bagging is applied on the weighted average classifier C_W for the Cmc dataset. One may observe that the test error of the bagged classifier improves the starting stacked classifier for large values of B. In the right panel of Figure 1, the performance gain obtained by bagging is illustrated even more clearly: here we bagged the stacked classifier using discriminant analysis in level-one. The initial stacked classifier has a test error rate of 12.6%, decreasing until 8.14% after $B = 50$ bagging steps.

Table 3 reports the test error rates of the combination methods after 50 bagging steps. The results indicate that, for most datasets and classifiers, bagging indeed reduces the test error rate of the stacked classifiers (45 times out of 60). The bagging also reduces the differences in performance across the different stacking methods. Moreover, it turns out that the bagged weighted combination method outperforms the others on 6 out of 12 datasets, and is comparable to the bagged stacked classifiers using decision tree or logistic regression as level-one classifier.

As a conclusion, this paper introduced a new combination method of classifiers which was compared to existing stacking methods. It was shown that this optimal weighted combination performed fairly well. Moreover, we also studied the possible advantage of applying bagging on stacked classifiers. It came out that the bagged weighted combination classifier outperformed (or, at least, performed comparatively as) the other combination methods (bagged or not) for most of the datasets under consideration.

References

[1] Breiman L. (1996). *Bagging predictors*, Machine Learning **26**, 123 – 140.

[2] Hastie T., Tibshirani R., Friedman J. (2001). *The elements of statistical learning: data mining, inference, and prediction.* Springer-Verlag, New York.

[3] LeBlanc M., Tibshirani R. (1996). *Combining estimates in regression and classification.* Journal of the American Statistical Association **91**, 1641 – 1647.

[4] Mertz C.J., Murphy P.H. (1996). *UCI repository of machine learning databases.* `http://www.ics.uci.edu/mlearn/MLRepository.html`, University of California.

[5] Wolpert D.H. (1992). *Stacked generalization.* Neural Networks **5**, 241 – 259.

Acknowledgement: We thank the reviewer for constructive comments. This research has been funded by the Research Fund K.U.Leuven and the "Fonds voor Wetenschappelijk Onderzoek"

Address: C. Croux, K. Joossens, A. Lemmens, K.U.Leuven, Department of Applied Economics, Naamsestraat 69, B-3000 Leuven, Belgium

E-mail: {christophe.croux,kristel.joossens}@econ.kuleuven.ac.be
aurelie.lemmens@econ.kuleuven.ac.be

DEVELOPING A MICROSIMULATION SERVICE SYSTEM

Joszef Csicsman and Cecilia Fenyes

Key words: Microsimulation, simulation.

COMPSTAT 2004 section: Simulations.

Abstract: The microsimulation procedure examines social and economic changes by assessing the effect of each provision with small units and the description of the overall effects is derived from these assessments. Relevance of the results relating to the society as a whole is ensured by the database which is a national representative sample of the units or households of such size that guarantees the required statistical reliability. Naturally, the range of social and economic changes that can be modelled in this way is defined by the information available on the microsimulation units in the database.

1 Introduction

In the past few years the Hungarian Central Statistical Office (KSH), the Ministry of Economic Affairs (GM), the Ministry of Finance (PM) and other departments of the government have had no opportunity to analyse large data sets to establish policy proposals.

The available analytical systems - usually based on EXCEL - are, in methodological respect, questionable, because it is hard to accept the reports based on small data sets both in mathematical and in economical sense.

A research group was founded for the project jointly by the Research Centre for Financial Economics of the Budapest University of Technology and Economics and Új Calculus Bt. The aim of the project is to develop and maintain a model system which allows analysing data collections and databases available in the administration for economists. Naturally, this methodology can also be used for processing other types of data too.

In accordance with the international applications the research group creates technical and methodological conditions for analysing large data sets and developing a SAS Software based microsimulation modelling system.

Till now a methodology to assess the impact of government programs has not been available.

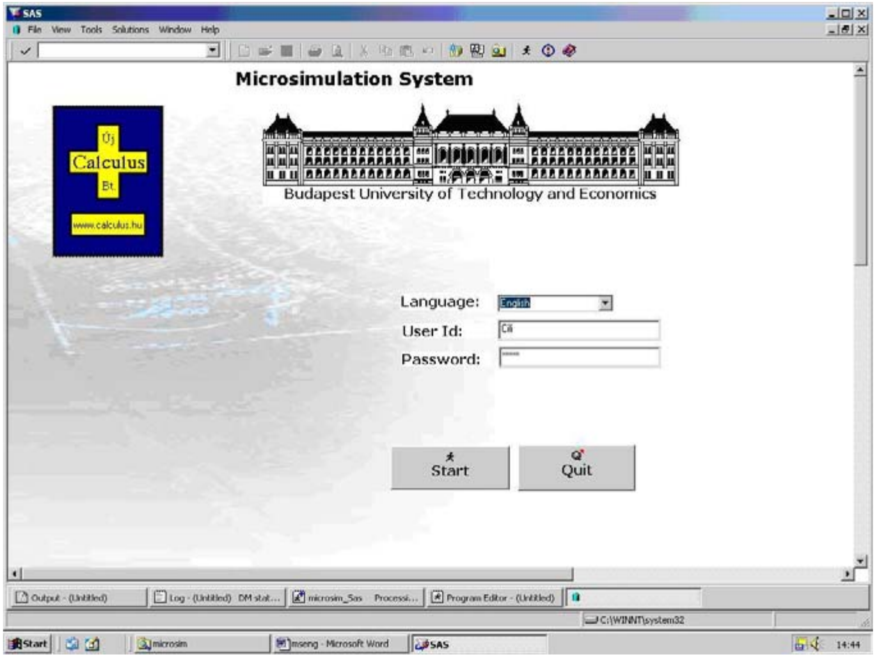
By utilizing research experience of the university more exact analyses could be prepared to qualify and quantify policy options.

As an outcome of this project a Microsimulation Modelling System will be developed which will be suitable for modelling the decisions of economic and social policy, and - complying with the national and international requirements - well-founded analyses can support the policy proposals of the government.

The following description will present the recent results of the project in a framework in which great number of university theses have been developed.

2 The Microsimulation Service System

As it can be seen in the Main menu, the language of the program is adjustable. At the moment English and Hungarian versions are available, but because the system works from a dictionary and the compilation does not need any special knowledge (only the command of the language), it is very easy to translate it into any other language.

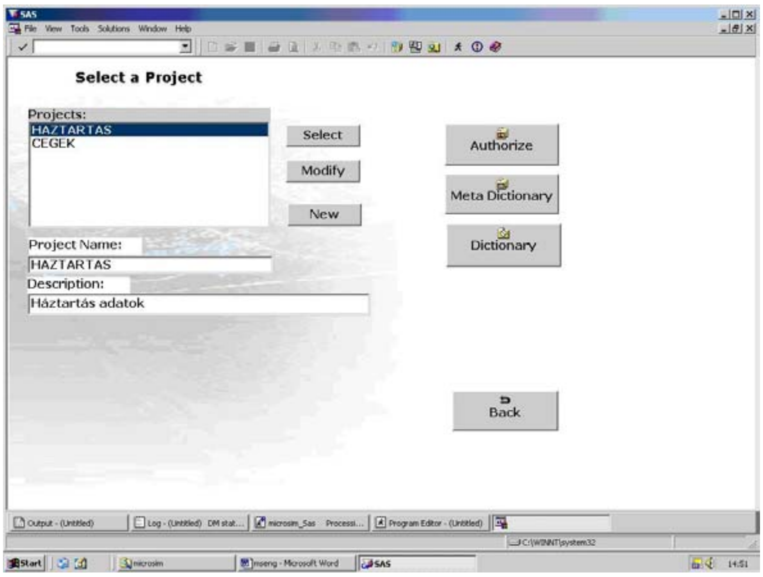


Language: English
User Id: Cili
Password: *****

2.1 Project selection

Prior to running a particular simulation a Project is to be selected (or created, if it does not exist), because the next steps are related to the data we want to use.

Without selecting a project 3 tasks can be chosen. By selecting ‘Authorize’ rights of the users can be set. By pushing the button ‘Meta dictionary’ the data’s of meta can be read or set, which are independent in any project (nomenclatures and numeric values). By pressing Dictionary the dataset of languages can be found. After selecting or creating the Project or selecting a task, the Project Handling frame will appear:

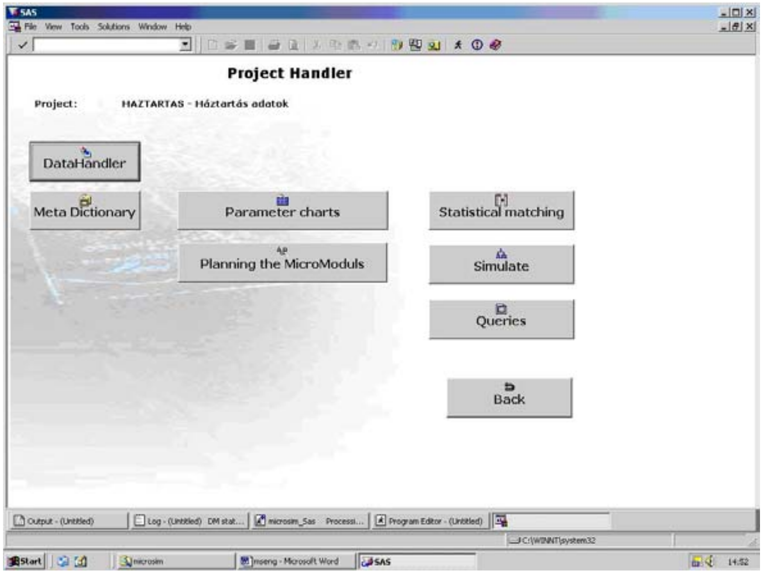


Select a Project

Projects:
HAZTARTAS
CEGEK

Project Name:
HAZTARTAS

Description:
Háztartás adatok



Project Handler

Project: HAZTARTAS - Háztartás adatok

2.2 Datahandling

a) Manipulating the input data

The system can handle text form and SAS files. Expanding the types of importable external files is one of the tasks of the next developing period.

b) Data protection

Due to the public environment use of a password entry and different levels of users is unavoidable.

2.3 Meta Dictionary

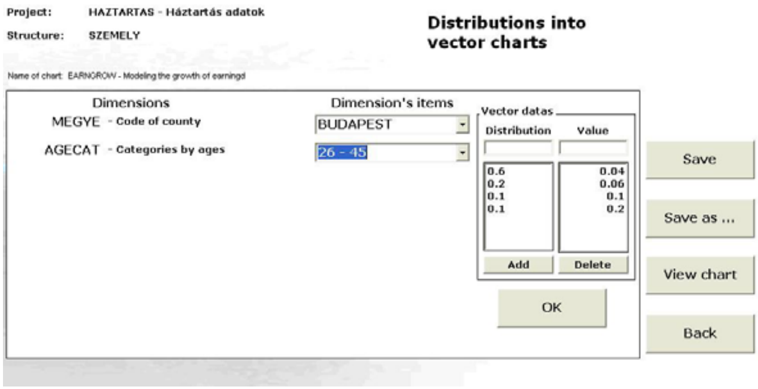
The Meta Dictionary contains all the information about the data and datasets: identifier, type, length and name of nomenclatures and pointers, structures of input datasets. Also the file catalog is the part of the meta. On the following screenshot the list of nomenclatures can be seen with a user friendly screen to view or modify it. Numeric values, structures of datasets, and the file catalog will be shown in similar frames.

2.4 Estimation algorithms

The parameter charts of estimation algorithms can be filled with the help of a graphical user interface, so economists can determine the internal algorithms without any SAS programming knowledge.

One of our most important goals is to complete these mathematical algorithms to provide opportunity for making appropriate analysis.

One of the frames, which helps to fill in the 'Value assignment upon distribution' parameter chart can be seen below:



Project: HAZTARTAS - Háztartás adatok

Structure: SZEMELY

Distributions into vector charts

Name of chart: EARNGROW - Modeling the growth of earningd

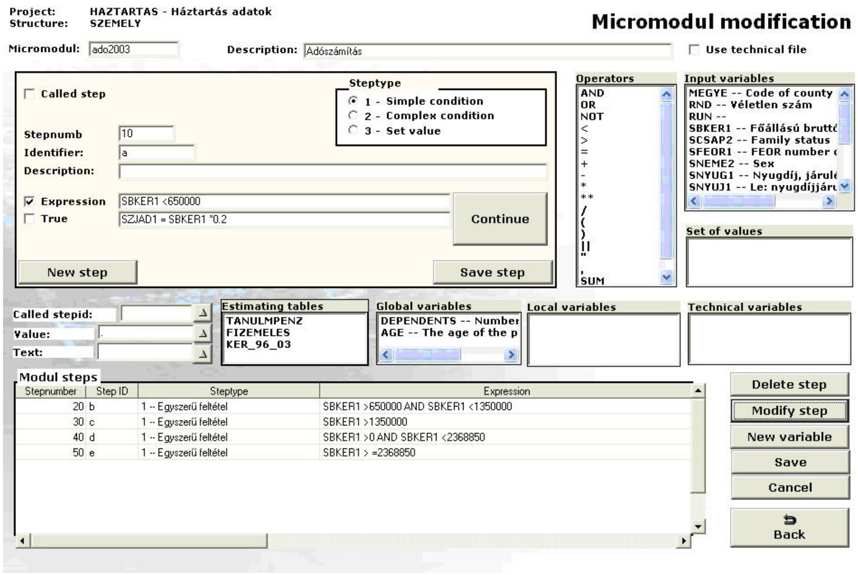
Dimensions	Dimension's items
MEGYE - Code of county	BUDAPEST
AGECAT - Categories by ages	26 - 45

Vector datas

Distribution	Value
0.6	0.04
0.2	0.06
0.1	0.1
0.1	0.2

2.5 Micromodules

Micromodules are Base Sas codes. During the simulation the selected micromodules will run on every record of the input dataset. This means that micromodules set the changes, which will occur. Here we prepared the platform for making micromodules without the knowledge of Base Sas. The code will be generated from the rows of modul steps (down, left) made by the user using almost only the mouse.



Project: HAZTARTAS - Háztartás adatok
Structure: SZEMELY

Micromodul: ado2003
Description: Adószámítás
□ Use technical file

□ Called step
Stepnumb: 10
Identifier: a
Description:
☑ Expression: SBKER1 <650000
□ True: SZJAD1 = SBKER1 *0.2

Steptype:
● 1 - Simple condition
○ 2 - Complex condition
○ 3 - Set value

Operators: AND, OR, NOT, <, >, =, +, -, *, **, /, (,), ||, ", ,, SUM

Input variables:
MEGYE -- Code of county
RND -- Véletlen szám
RUN --
SBKER1 -- Főállású bruttó
SCSAP2 -- Family status
SFEOR1 -- FEOR number c
SNEME2 -- Sex
SNYUG1 -- Nyugdíj, járulé
SNYUJ1 -- Le: nyugdíjjáru

Called stepid:
Value: .
Text:

Estimating tables:
TANULMPENZ
FIZEMELES
KER_96_03

Global variables:
DEPENDENTS -- Number
AGE -- The age of the p

Stepnumber	Step ID	Steptype	Expression
20	b	1 -- Egyszerű feltétel	SBKER1 >650000 AND SBKER1 <1350000
30	c	1 -- Egyszerű feltétel	SBKER1 >1350000
40	d	1 -- Egyszerű feltétel	SBKER1 >0 AND SBKER1 <2368850
50	e	1 -- Egyszerű feltétel	SBKER1 > =2368850

2.6 Running the simulation

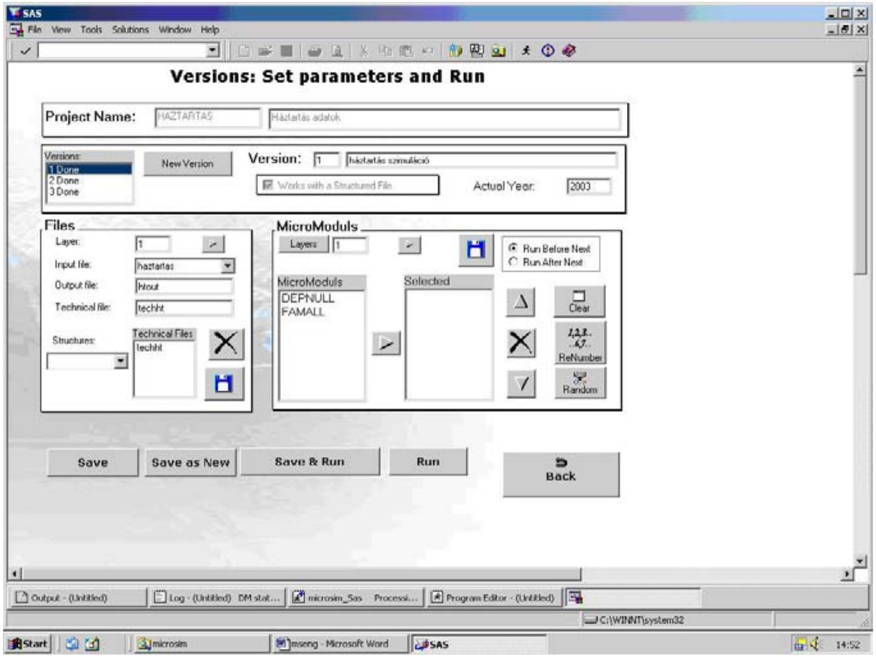
This frame is to set all the parameters of the simulation. Input file(s) can be chosen, name of the output file must be set, and here the micromodules can be selected to run from the list of the premade micromodules.

2.7 Analysing

After running the simulation it is very important to have opportunity to analyse the input and output data. The Analyse function of the system can help analysts, who are not experienced in SAS programming. For Analysis the procedures of SAS can be used very well.

2.8 Statistical matching

Statistical Matching is to pair the records of two data sets without having any key variable. The records of the secondary data sets are separated into groups



Versions: Set parameters and Run

Project Name: HAZTARTAS Háztartás adatok

Versions:
1 Done
2 Done
3 Done

Version: 1 háztartás szimuláció

☑ Works with a Structured File

Actual Year: 2003

Files

Layer: 1

Input file: haztartas

Output file: htout

Technical file: techht

Structures:

Technical Files: techht

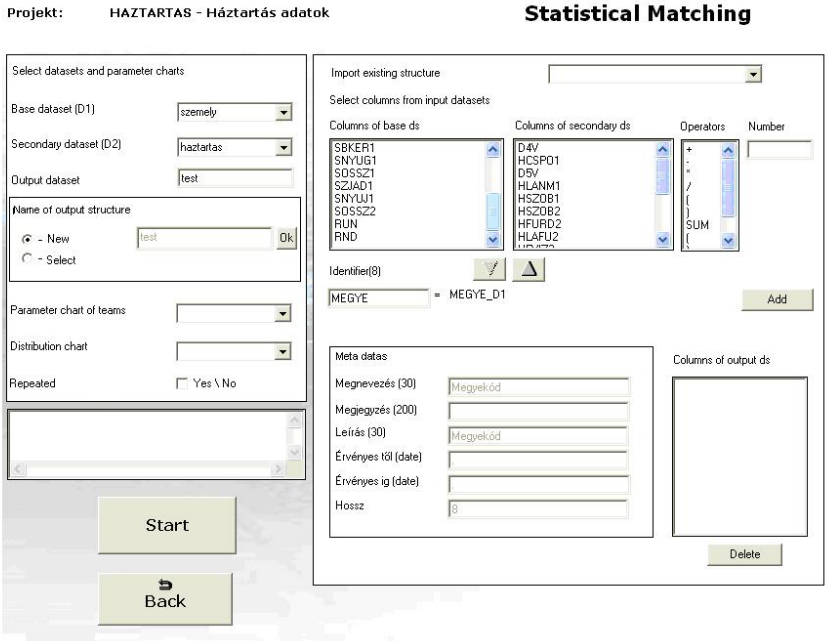
MicroModuls

Layers 1

● Run Before Next
○ Run After Next

MicroModuls: DEPNULL, FAMALL

Selected:



Projekt: HAZTARTAS - Háztartás adatok

Statistical Matching

Select datasets and parameter charts

Base dataset (D1) szemely

Secondary dataset (D2) haztartas

Output dataset test

Name of output structure

● - New test
○ - Select

Parameter chart of teams

Distribution chart

Repeated □ Yes \ No

Import existing structure

Select columns from input datasets

Columns of base ds: SBKER1, SNYUG1, SOSSZ1, SZJAD1, SNYUJ1, SOSSZ2, RUN, RND

Columns of secondary ds: D4V, HCSPO1, D5V, HLANM1, HSZOB1, HSZOB2, HFURD2, HLAFU2

Operators: +, -, *, /, (,), SUM, (

Number

Identifier(8)

MEGYE = MEGYE_D1

Meta datas

Megnevezés (30) Megyekód

Megjegyzés (200)

Leírás (30) Megyekód

Érvényes től (date)

Érvényes ig (date)

Hossz 8

Columns of output ds

by their selected attributes (they can be defined in the section of parameter charts, by selecting ‘statistical matching - parameter chart of teams’ from the list of type of table). The statistical matching goes through on every records of the primary data set. By the attributes of the record the member of the appropriate group (which is a record of the secondary data set) will be paired with this record.

If there is only one data set and the records of the same dataset are intended to be paired (like the simulation of marriage), it is also solved in the microsimulation system.

Address: J. Csicsman, C. Fenyes, Uj Calculus, 1457. Budapest Pf. 184., Hungary

E-mail: `csicsman@calculus.hu, fenyesc@itm.bme.hu`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

EFFECTIVENESS IN ENSEMBLE OF CLASSIFIERS AND THEIR DIVERSITY ON BIG MEDICAL DATA SET

Malgorzata Maria Cwiklinska-Jurkowska and P. Jurkowski

Key words: Fixed and trained combining classifiers, diversity, accuracy.

COMPSTAT 2004 section: Classification.

Abstract: The dependence of stacked combining accuracy on diversity was studied for real-life medical classification problem of thyroid disease recognition. The simple fixed (minimum, maximum, product, average, mean and majority vote) and also trained methods (with second step-the Bayesian discrimination) were examined. Base classifiers were both of type: crisp (classification trees, neural network) and fuzzy (parametric and nonparametric Bayesian discrimination).

We studied nine measures of diversity for oracle outputs of correct/ incorrect classification decisions. For trained ensemble methods we obtained smaller Spearman correlations between classification errors and diversity than for fixed combining.

1 Introduction

In order to avoid the possible loss of information, classifiers can be pooled. The recognition rate of a combination is usually better than that of each individual classifier [17]. Multiple classifier systems have been attempted in a variety of pattern recognition fields e.g. [4], [17]. Diversity among individual classifiers of the team is expected to be important for effectiveness in classifier combination [16]. The relationship between different combining accuracy and diversity of classifiers' ensemble was studied on generated data by Kucheva et al. [7], [8], [9] and Shipp et al. [14].

The purpose of the paper was the experimental examining dependence on diversity of accuracy of stacked fixed and also of stacked trained combining methods [4] on the basis of the real medical data set for thyroid disease recognition. The big data set with three groups was used [11]. The learning set consists of 3772 examples and testing set of 3428 examples. We employed all of 21 mixed variables, 15 binary and 6 continuous.

2 Constituent and combined classifiers

Let us denote as c- the number of classes $\pi_1, \ldots, \pi_c$, L-the number of classifiers $D_1, ..., D_L$. The L classifiers' outputs for classification to c groups of an observation $\boldsymbol{x}$ can be organized in a decision profile ***DP(x)*** matrix:

j-th (*j=1,. . . , c*) matrix column of ***DP(x)*** contains posterior probabilities (alternatively 0 or 1 for two applied crisp classifiers) that an observation $\boldsymbol{x}$ belongs to *j-th* population. For simple combining from the matrix ***DP(x)*** of the decision profile we obtain only one raw vector- of possibilities of belonging to c classes. For example, it contains averages (or minimum, maximum, median, product, majority vote) of column elements of matrix ***DP(x)***. For trained combination rule- all elements of the profile matrix ***DP(x)*** coming from first-step discrimination are features for second-step discrimination.

We used such constituent Bayesian discriminant classifiers as linear, quadratic, kernel and nearest neighbour. These base nonparametric classifiers were chosen with the best parameters, so are optimal in their classes: the kernel radius and the number of neighbours are optimized using a leave -one-out error estimation. Binary decision trees were also applied. To choose the most efficient one we tried CART [1] and QUEST tress [10] with different parameters. The artificial dipolar neural network, used in the research, consists of one layer of formal neurons with binary activation function [1].

The stacked combining- of different classifiers on the same feature set (all 21 mixed variables) - is examined. The error of classification was estimated by resubstitution, cross-validation method or by using the testing sample.

3 Diversity of classifiers

We applied 9 diversity measures [8], among them 4 averaged pairwise ones and 5 working on the whole group of classifiers. Let us consider oracle outputs of classification decisions (binary vectors: with values equal to 1 if correct decision of a classifier D_j for an observation $\boldsymbol{x}$ or zeros for incorrect classifications).

	$D_j$correct	$D_j$fail
$D_i$correct	a	b
D_i fail	c	d

Table 1: Summary table for relationship between two classifiers D_i and D_j with probabilities ($a+b+c+d$=1).

Pairwise measures

Q	Yule's Q statistic $Q=(ad\text{-}bc)/(ad+bc)$
D	disagreement measure $D = b + c$
DF	double-fault measure $DF=d$

Non-pairwise measures (comparing accuracies of several classifiers)

$E(ENTROPY\ )$	The entopy measure
HS	Hansen and Salamon measure of difficulty θ [5]
KW	Kohavi-Wolpert variance [6]
κ $(KAPPA\ )$	Measurement of interrater agreement
GD	Generalized diversity [13]
CFD	Coincident failure diversity [after: 8]

DF, *GD* and *CDF* are non symmetric measures (if we swap all values 1 by 0 and vice versa). Measures of similarity (the smaller the value the more diverse) are: Q, D, κ and *HS.*

For any L (number of all considered classifiers) it is fulfilled:

KW=(L-1) $D_{AVG}/2L$ (average of D is taken from all possible pairs of classifiers)

and for L=3 it holds: $D = 2/3E$ [8].

4 Results and discussion

Wang et al. [16] showed that when a high diversity between two or more different data mining techniques exists, combining them produce better classifications. They concluded that a multiple classifier system can only improve the performance when the members in the system are diverse from each other. Thus we added to combining some methodologically different techniques than Bayesian discrimination- the most efficient classification tree (CART with 18 nodes; test error =0.006) and the neural network. Accuracies of constituent classifiers are presented in Tab. 2. Duin et al. [4] showed that although the individual classification performances on the difficult datasets are poor, they still contain valuable information for the combination rules. When we rejected, from the set of all Bayesian constituent classifiers {1-5}, the worst basic quadratic discrimination, we obtained decrease of the correct classification rate- for the average and median rule. Most of considered combining teams however do not contain the worst quadratic method or the normal kernel with the pooled covariance matrices (Tab. 2).

Number	Method	Test-sample error
1.	Linear discrimination Pp	0.0619
2.	Quadratic discrimination Pe	0.5984
3.	Normal kernel r=0.5 ; Pp pooled covariance matrix	0.0614
4.	Normal kernel r=0.5 ; Pe; not pool.covariance matrices	0.4739
5.	6 Nearest Neighbour	0.0596
6.	Classification tree	**0.006**
7.	Neural Network	0.039

Table 2: Test-sample classification errors of individual classifiers.Pp: probabilities a priori proportional to sizes of groups. Pe: probabilities a priori equal in groups.

Standard statistical software SAS 8.2 was applied for building Bayesian classifiers, Statistica 5.0 for classification trees and we used also the program for the dipolar neural network [1]. We assessed fixed and trained ensemble errors and computed nine diversity measures [8] for different teams - subsets of above individual classifiers.

Combining classifiers is particularly useful for difficult problems – a large amount of noise, limited number of training data or unusually high dimensional patterns [15]. In used data the sizes of groups were much bigger than the number of variables, so the improvement, reached by some of examined combining *fixed* methods, was not apparent. On the other hand, *trained* combining gave much smaller classification errors - the smallest one for the kernel method (error=0.0005).

4.1 Relationships between measures of diversity

Teams 1-6 are the following: for Team 1- individual classifiers from Tab.2 are {1-6}; 2: {1-5}; 3: {1,2,3,5,6}; 4: {1,3,5,6}, 5:{1,3,5}; 6: {1,3,4,5,6}. Teams 7-12 are as above six ones, but with additionally 7th individual classifier from Tab. 1 (neural network). For these 12 teams of classifiers from learning sample and 12 teams from the testing one (polled into set of 24 teams) we examined the different relationships.

Q is the only measure that varies between -1 and 1 and attains zero for independent classifiers, like Pearson correlation coefficient, and additionally does not depend on accuracy of individual classifiers - taken with the same accuracy p [8]. Classifiers that tend to recognize the same objects correctly have positive values of Q and those which give errors for different objects deliver negative Q: thus negative dependence is preferable [7]. We obtained 6 values of Q near to 0 and the rest - positive.

For *HS* diversity measure we prefer teams of classifiers of the following property: the points that are difficult for some classifiers are easy for other classifiers (that means - we want *HS* to be small). This property, analogously to sets with negative Q, is called: "negative dependence": [8], though *HS* is always nonnegative as the variance. We examined *HS* measure for 12 subsets of classifiers for learning sample and 12 for testing one (Fig.1). Best by the criterion of *HS* have proved team 7 and 8 (with smallest *HS* equal to 0.063 and 0.071 for testing sample- Fig. 1 and with 0.07 and 0.08 for the learning sample).

Examined features- errors and diversities- are not symmetrically distributed - the Spearman rank order correlations were calculated. The cluster analysis was also performed on the basis of 24 combining teams (Fig. 2). For diversity measures the bigger cluster consists of all symmetric measures: *Q, KAPPA, HS, KW, D* and *ENTROPY*. Similarity measures: *Q, KAPPA* and *HS* are placed right in this cluster. Q is highly correlated with all measures from this cluster- negatively (with *D, KW* and *ENTROPY*) or positively (with *KAPPA* and HS). The asymmetric *DF* (double fault) measure can be regarded as one-element cluster. It can be explained by the fact that for *DF* we treat as similar the classifiers, only rejecting the same observations, without any knowledge of other classifying combinations. Very similar in constructing *GD* and *CDF* are close in the dendrogram. Strongly related measures: *D, KW* and *ENTROPY* are very highly positively correlated and

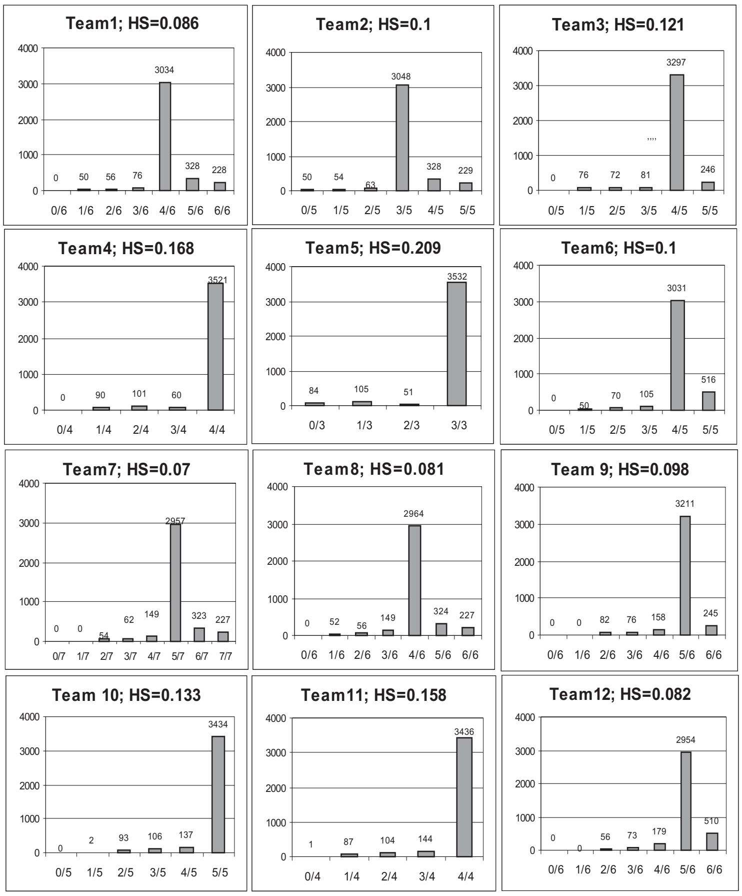


Figure 1: Hansen and Salamon measure of difficulty *HS* for 12 different teams of 7 individual classifiers. Histogram of the number of patients from testing sample that are correctly classified by i/L of L classifiers (horizontal axis). Classifiers performed on the testing sample.

lay near in the dendrogram. Some of the above effects for examined by us real medical data are consistent with results obtained by Shipp and Kuncheva [14] for big generated data set.

4.2 Simple fixed fusion

For most of classifiers teams - median and average fusion methods reached smallest errors and only for teams 4, 5 and 11- the most accurate was the maximum fusion method. Effectiveness of the product combining method is significantly positively correlated with minimum and maximum rule. Accu-

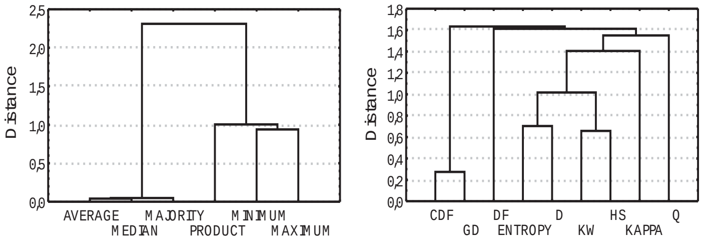


Figure 2: Cluster analysis. Left: classification errors of fixed combining methods. Right: measures of diversity

racy of average for considered medical data is highly positively correlated with median and majority vote. We obtain two clusters (Fig. 2): average with median and majority vote and the second one: product, minimum and maximum. Chen & Cheng [3] showed (in the two-class classification problem): median and majority vote produce the same result, but- for non symmetric distributions- average may yield a completely different decision (much more or much less correct), when a number of constituent classifiers becomes large.

The negative (positive) sign of highest correlations from Tab. 3 confirms the fact that a measure belongs to the similarity group (or respectively to the strict diversity measure group) with only the exception of *DF* measure- single cluster among diversity measures. The stronger relationship with accuracy we can observe for measures from the cluster $\{Q$, *KAPPA, HS, D, KW* and *ENTROPY*$\}$, i.e. for symmetric measures.

	Q	DF	κ	HS	D	KW	ENT	GD	CFG
Minimum	**-0.75**	**0.41**	**-0.55**	**-0.69**	**0.77**	**0.74**	**0.78**	0.36	0.36
Maximum	**-0.73**	0.37	**-0.55**	**-0.63**	**0.75**	**0.72**	**0.75**	**0.40**	**0.42**
Majority	0.36	**0.42**	0.49	0.10	-0.09	-0.06	-0.16	-0.22	-0.23
Product	**-0.78**	**0.42**	**-0.57**	**-0.64**	**0.78**	**0.76**	**0.78**	0.37	0.39
Median	0.41	0.40	0.51	0.21	-0.18	-0.14	-0.22	-0.21	-0.22
Average	**0.37**	**0.42**	0.48	0.17	-0.11	-0.07	-0.17	-0.21	-0.22

Table 3: Spearman correlations between measures of diversity and errors of fixed fusion methods (bolded- significant p=0.05).

4.3 Trained fusion

Similar to above analysis was also done for 14 trained fusion methods, with second-step discrimination. For probabilities a priori equal and proportional to sizes of groups we executed 7 discriminant Bayesian methods: parametric -linear and quadratic- and nonparametric ones-normal kernel with pooled and not pooled covariance matrices (radii $r = 1$ and $r = 0.5$) and nearest neighbour ($k = 6$ neighbours). For each individual fuzzy classifier- the vectors of posterior probabilities are linearly dependent. Similarly, the vectors coming from crisp classifiers (containing 1 or 0) are redundant: one vector can be removed. Thus, to remove the problem of pooled covariance matrix's singularity, one column (out of 3) of posterior probabilities for each classifier in trained fusion or one binary column of crisp classifier was not taken to calculations for second-step discrimination.

Spearman correlations between diversities and errors of trained fusion methods were significant only for *DF* diversity measure with linear, quadratic and nearest neighbour discrimination and also for *HS* measure with quadratic trained method (all methods with prior probabilities proportional to sizes of groups). Correlations between accuracies of trained combining separately were much higher than between simple fixed combined classifiers and we obtained two clusters of 14 second-step methods.

5 Concluding remarks

The correlation between diversity measures separately and also between accuracy of fixed combining was higher than between diversity measures and effectiveness of classification. For trained ensemble methods we obtained even smaller relationships between accuracy and diversity than for fixed combination while we achieved the highest correlations-of all considered kinds-between accuracy of trained methods separately. Considered diversity measures of combining classifiers are not very good indicators of effectiveness, especially for trained fusion methods.

It would be interesting to find measures that are higher correlated with accuracy of combining classification, because they might be used for constructing the efficient ensemble of classifiers.

References

[1] Bobrowski L., Kretowska M. (2000). *Dipolar pruning of neural layers.* Proceedings of the Fifth Conference on Neural Networks and Soft Computing. Zakopane 2000, 174 – 179.

[2] Breiman L. *Bagging predictions.* (1996). Machine Learning **24** (2), 123 – 140.

[3] Chen D., Cheng X. (2001). *An asymptotic analysis of some expert fusion methods.* Pattern Recognition Letters **22**, 901 – 904.

[4] Duin R.P.W., Tax D.M.J. (2000). *Experiments with classifier combining rules.* In: Multiple Classifier Systems, J. Kittler, F. Roli (Eds). Springer-Verlag, Berlin.16 – 29.

[5] Hansen L., Salamon P. (1990). *Neural network ensembles.* IEEE Transactions on Pattern Recognition and Machine Intelligence, **12** (10) 993 – 1001.

[6] Kohavi R., Wolpert D.H. (1996). *Bias plus variance decomposition for zero-one loss functions.* In L. Saitta (ed.), Machine Learning: Proc. 13^{th} International Conference, 275 – 283. Morgan-Kaufmann, 1996.

[7] Kucheva L.I., Whitaker C.J., Ship C.A (2000). *Is independence good for combining classifiers?* International Conference on Pattern Recognition (ICPR'00) Barcelona 2000 **2**, 168 – 171.

[8] Kuncheva L.I., Whitaker C.J. (2001). *Ten measures of diversity in classifier ensembles: limits for two classifiers.* Proc. IEE Workshop on Intelligent Sensor Processing, Birmingham, 10/1 – 10/6.

[9] Kucheva L.I., Whitaker C.J., Ship C.A. (2003). *Limits on the majority vote accuracy in classifier fusion.* Pattern Analysis and applications **6**, 22 – 31.

[10] Loh W.Y., Shih Y.S. (1997). *Split selection methods for classification trees.* Statistica Sinica **7**, 815 – 84.

[11] Machine Learning Repository. UCI, `www.ics.uci.edu/mlearn/MLRepository.html`

[12] Mc Lachlan G.J. (1992). *Discriminant analysis and statistical pattern recognition.* J. Wiley & Sons.

[13] Patridge D., Krzanowski W.J. (1997). *Software diversity: Practical statistics for its measurement and exploitation.* Information and Software Technology **39**, 707 – 717.

[14] Shipp C.A., Kuncheva L.I. (2002). *Relationships between combination methods and measures of diversity in combining classifiers.* Information Fusion **3** (2), 135 – 148.

[15] Tumer K., Ghosh J. (1996). *Analysis of decision boundaries in linearly combined neural classifiers.* Pattern Recognition **29** (2), 341 – 348.

[16] Wang W., Jones P., Partridge D. (2000). *Diversity between neural networks and decision trees for building multiple classifier systems.* In: Multiple Classifier Systems. Eds: J. Kittler, F. Roli. Springer-Verlag, Berlin, 240 – 250.

[17] Yu K. Jiang X., Bunke H.(1997). *Lipreading: A classifier combination approach.* Pattern Recognition Letters **18**, 1421 – 1426.

Acknowledgement: Thanks to Ludmila Kuncheva from University of Wales for Mathlab procedures measuring diversity and to Malgorzata Kretowska from Technical University in Bialystok for computing results of neural network program.

Address: M. Cwiklinska-Jurkowska, P. Jurkowski, Department of Theoretical Backgrounds and Department of Informatics and Research Methodology, The Ludwik Rydygier Medical University, ul.Jagiellonska 13; 86-067 Bydgoszcz, Poland

E-mail: `mjurkowska@amb.bydgoszcz.pl`

TEST OF CONTINUITY OF A REGRESSION FUNCTION

Václav Čapek

Key words: Nonparametric regression, change-point, M-smoothers, resampling.

COMPSTAT 2004 section: Nonparametrical statistics.

Abstract: A test of continuity of a regression function based on M-smoothers is proposed. Its limit behavior both under the null hypothesis (no jump) and the alternative (there is a jump) are presented. Critical values are obtained using bootstrap. The theoretical results are accompanied by a real data example.

1 Introduction

Consider the nonparametric regression model

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where m is an unknown function defined on $[0, 1]$ and $\varepsilon_1, \dots, \varepsilon_n$ are independent, identically distributed random variables having a symmetric distribution function F. Random variables $Y_1, \dots, Y_n$ are observed at design points $0 \le x_1 \le \cdots \le x_n \le 1$.

Suppose that the model is of the form

$$m(x) = m_0(x) + \Delta \mathbf{I}_{\{x > \tilde{x}\}}, \quad x \in (0, 1), \tag{2}$$

where m_0 is a smooth function and Δ and $\tilde{x}$ are unknown real parameters. The case $\Delta \neq 0$ and $\tilde{x} \in (0, 1)$ corresponds to a jump in the regression function m. We want to test whether m has a jump or not. Denote the null and the alternative hypotheses by

$$\begin{aligned} H_0 : \quad & \Delta = 0 && (3) \\ H_A : \quad & \Delta \neq 0 \text{ and } \tilde{x} \in (0, 1). && (4) \end{aligned}$$

There are not many papers concerning the problem of testing H_0 against H_A. Müller and Stadtmüller [7] developed a test for the case of a fixed design with regular grid on the basis of a quadratic statistic. Horváth and Kokoszka [5] considered a test procedure based on local polynomial smoothers for the same setup.

Our test is based on M-smoothers. We recall now the needed notion on M-smoothing techniques for a continuous regression (assume $\Delta = 0$). For the above model denote $F(y - m(x))$ by $F_x(y)$. For ψ nondecreasing and left-continuous and some function G put

$$\lambda_G(t) = -\int \psi(y-t)\, dG(y), \quad t \in \mathbf{R}$$

and

$$\lambda_G^{-1}(u) = \inf\{t :\, \lambda_G(t) \ge u\}.$$

It is easy to see that for ψ nondecreasing, left-continuous and antisymmetric with $\lambda'_G(0) > 0$, where $\lambda'_G(t)$ denotes the derivative, one gets

$$m(x) = \lambda_{F_x}^{-1}(0) \tag{5}$$

and that $\lambda_{F_x}^{-1}(0)$ is unique.

The M-smoother $m_n(x)$ is defined as an empirical analogue to (5)

$$m_n^h(x) = \lambda_{\hat{F}_x^h}^{-1}(0), \tag{6}$$

where $\hat{F}_x^h$ is a kernel estimator of F_x defined by

$$\begin{aligned} \hat{F}_x^h(y) &= \mathbf{I}_{\{Y_1 \le y\}} \int_{-\infty}^{x_1} h_n^{-1} K((x-z)/h_n)\, dz \qquad (7) \\ &\quad + \sum_{i=2}^{n-1} \alpha_{ni}^h(x) \mathbf{I}_{\{Y_i \le y\}} + \mathbf{I}_{\{Y_n \le y\}} \int_{x_{n-1}}^{\infty} h_n^{-1} K((x-z)/h_n)\, dz, \end{aligned}$$

with weights

$$\alpha_{ni}^h(x) = \int_{x_{i-1}}^{x_i} h_n^{-1} K((x-z)/h_n)\, dz, \quad i = 2, \dots, n-1.$$

Here $\mathbf{I}_{\{A\}}$ denotes an indicator of an event A, K is a kernel and $h = \{h_n\}_{n=1}^{\infty}$ is a sequence of bandwidths. The subscript n indicates a possible dependence of h_n on the number of observations.

In other words, the estimator $m_n^h(x)$ of the unknown regression function m at the point x is a solution of minimization problem (with respect to t)

$$\sum_{i=1}^{n} \rho\big(Y_i - t\big) \int_{x_{i-1}}^{x_i} K\Big(\frac{x-z}{h}\Big)\, dz,$$

where the function ρ is a loss function, assumed to be convex and differentiable.

When choosing $\rho(x) = x^2$ one gets here the classical kernel estimator of $m(x)$. In case $\rho(x) = |x|$ the problem leads to LAD procedures. The M-smoother $m_n^h(x)$ is a generalization of the M-estimator as well as of the kernel one.

Such setup was considered in Härdle and Gasser [4]. They derived results on the limit distribution of $m_n^h(x)$. Antoch and Janssen [1] established its limit probability inequalities. The random design case was considered for example in Härdle [3].

2 Procedure

The idea of detecting of a jump in the data is to use one-sided M-smoothers. We construct two different estimates, say m^h_{n+} and m^h_{n-} based on kernels

$$K_+(z) = 2K(z)\mathbf{I}_{\{z>0\}} \text{ and } K_-(z) = 2K(z)\mathbf{I}_{\{z<0\}}, \tag{8}$$

where K is a symmetric kernel. In this setup, $m^h_{n+}(x)$ is an estimate of the unknown function m at point x based only on observations (Y_i, x_i) having $x_i > x$, while $m^h_{n-}(x)$ is an estimate based only on observations (Y_i, x_i) having $x_i < x$.

We expect that $m^h_{n+}(x_i) - m^h_{n-}(x_i)$ will be sensitive to a possible jump in m. Therefore we propose to use the test statistic

$$T^h_n = \sqrt{nh_n} \max_{i=1,\dots,n} |m^h_{n+}(x_i) - m^h_{n-}(x_i)|. \tag{9}$$

Large values of T^h_n indicate possible discontinuity and so the critical region of the test will be of the form

$$T^h_n \geq \gamma_n(\alpha)\sqrt{\log n}, \tag{10}$$

where α is a level of the test.

An approximation for the critical value $\gamma_n(\alpha)$ is needed. It can be obtained using bootstrap or a limit distribution. In this paper we choose the bootstrap method. Let $m^h_n(x)$ be the estimator of the unknown regression function $m(x)$ defined in (6). Denote residuals by

$$\hat{\varepsilon}_i = Y_i - m^h_n(x_i), \quad i = 1, \dots, n. \tag{11}$$

Let $\varepsilon^*_1, \dots, \varepsilon^*_n$ be a bootstrap sample from $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ and denote

$$Y^*_i = m^h_n(x_i) + \varepsilon^*_i, \quad i = 1, \dots, n. \tag{12}$$

A bootstrap analogue to (5) is defined by

$$m^{*g}_n(x) = \lambda^{-1}_{\hat{F}^{*g}_x}(0), \tag{13}$$

where $\hat{F}^{*g}_x$ is computed similarly as $\hat{F}^h_x$ in (7), but for the observations $Y^*_1, \dots, Y^*_n$ and the sequence of bandwidths $g = \{g_n\}$.

The bootstrap version of the test statistic T^h_n is then

$$T^{g*}_n = \sqrt{ng_n} \max_{i=1,\dots,n} |m^{g^*}_{n+}(x_i) - m^{g^*}_{n-}(x_i)|, \tag{14}$$

where $m^{g^*}_{n+}$ and $m^{g^*}_{n-}$ are bootstrap estimators based on kernels K_+ and K_- defined in (8).

We expect that the behavior of T^{g*}_n is similar to the behavior of T^h_n and that the critical value $\gamma_n(\alpha)$ can be approximated by the appropriate quantile of the distribution of T^{g*}_n.

3 Theoretical results

Introduce now the set of conditions under which the results are given:

(C1) F is symmetric.
(C2) $\lambda_F \in \mathrm{Lip}(1)$ and $\lambda'_F(0) > 0$.
(C3) K is a symmetric density with support $[-L, L]$, $L > 0$ and $\|K\| = \sup_z K(z) = \mathbf{K} < \infty$.
(C4) ψ is nondecreasing, left continuous, antisymmetric and $\|\psi\| < \infty$.
(C5) $m_0 \in \mathrm{Lip}(1)$.
(C6) The design points $0 = x_0 \le x_1 \le \cdots \le x_n \to 1$, $n \to \infty$, satisfy $\delta_n = \max_{1 \le i \le n}(x_i - x_{i-1}) = O(n^{-1})$.
(C7) The sequence of bandwidths $h = \{h_n\}$ satisfies $\log n/(nh_n) \to 0$ and $nh_n^3 \to 0$ for $n \to \infty$.
(C8) The sequence of bandwidths $g = \{g_n\}$ satisfies $ng_n \to \infty$ and $ng_n^3 \to 0$ as $n \to \infty$.
(C9) The bandwidth sequences h and g satisfy $g_n = o(h_n/\log n)$ for $n \to \infty$.

Conditions (C8) and (C9) describe the necessary behavior of the bandwidth sequence g. Its proper choice is crucial for a good performance of the bootstrap.

The first Theorem describes the behavior of the test statistic T_n^h under the null hypothesis. It is a consequence of Theorem 2.1 in Antoch and Janssen [1].

Theorem 3.1. *Assume (C1)–(C7) and H_0. Then*

$$(\log n)^{-1/2}\, T_n^h \xrightarrow{a.s.} 0 \ \text{ as } n \to \infty.$$

The next theorem asserts that the statistic T_n^h can be taken as a diagnostic tool for the detection of a possible discontinuity of the regression function m. Moreover, since the maximum in the definition of T_n^h is, with probability tending to 1, reached near the point of the jump $\tilde{x}$, it can be used as a preliminary estimate of the location of the jump.

Theorem 3.2. *Assume (C1)–(C7) and H_A. Then*

$$\liminf_{n\to\infty} (\log n)^{-1/2}\, T_n^h = +\infty \quad a.s.$$

and moreover

$$(\log n)^{-1/2} \left[\sqrt{nh_n} \max_{i:|x_i - \tilde{x}| > Lh_n} |m_{n+}^h(x_i) - m_{n-}^h(x_i)| \right] \xrightarrow{a.s.} 0 \ \text{ as } n \to \infty.$$

Finally we present a result on the behavior of the bootstrap version of T_n^h. Note that the result does not depend on whether H_0 or H_A is true.

Theorem 3.3. *Assume (C1)–(C9). Then*

$$(\log n)^{-1/2} T_n^{g*} \xrightarrow{P^*} 0 \quad [\mathrm{P}] \ a.s.$$

From Theorems 3.2 and 3.3 it directly follows that whenever we use a quantile $\gamma_n^*(\alpha)$ of the distribution of T_n^{g*} as an approximation for the critical value $\gamma_n(\alpha)$ we get a consistent test.

Proofs of all the above theorems will be published elsewhere.

4 The Klementinum data

The data comprise 218 average annual temperatures measured in Prague since 1775 till 1992, see Figures 1 and 2. These data were analysed by many authors in order to detect climatic changes, see Jarušková [6], Horváth and Kokoszka [5] and Goderniaux [2].

We analyse the data using methods introduced above. For computations we chose combinations of the Epanechnikov kernel $K_e(z) = \frac{3}{4}(1 - z^2)\mathbf{I}_{\{|z| \le 1\}}$ and the uniform one $K_u(z) = \frac{1}{2}\mathbf{I}_{\{|z| \le 1\}}$ and the loss functions $\rho_1(x) = |x|$ and $\rho_2(x) = x^2$.

The most important step is the proper choice of the bandwidth. We have chosen the cross-validation method. We denote the cross-validation statistic as

$$CV(h) = \sum_{i=1}^{n} [Y_i - m_{n_{-i}}^h(x_i)]^2, \tag{15}$$

where $m_{n_{-i}}^h(x_i)$ is the leave-one-out M-smoother estimator (with the observation (Y_i, x_i) omitted). Values of $CV(h)$ over a grid of several h's are computed and the minimizer is used for evaluating the final estimate. Actually this procedure has been repeated a couple of times and in each step the grid has been softened around the actual optimum.

We started with a grid $(10, 11, 12, \ldots, 30)$. In the second step we used a grid $(15.0, 15.1, 15.2, \ldots, 20.0)$. The final grids were different for different choices of kernels and loss functions, see Table 1.

kernel	ρ_1	ρ_2
Epanechnikov	$(18.80, 18.81, \ldots, 19.00)$	$(19.00, 19.01, \ldots, 19.20)$
uniform	$(15.90, 15.91, \ldots, 16.10)$	$(17.30, 17.31, \ldots, 17.50)$

Table 1: Cross-validation grids for bandwidth selection.

The results are shown in Figures 1 and 2. The final bandwidth used is in the title of each figure. The dashed-and-dotted and dotted lines are the estimates m_{n+}^h (”right”) and m_{n-}^h (”left”). The values of $\sqrt{nh}\,|m_{n+}^h(x_i) - m_{n-}^h(x_i)|$ are suitably scaled, possitioned to the bottom of each figure and plotted with a solid line.

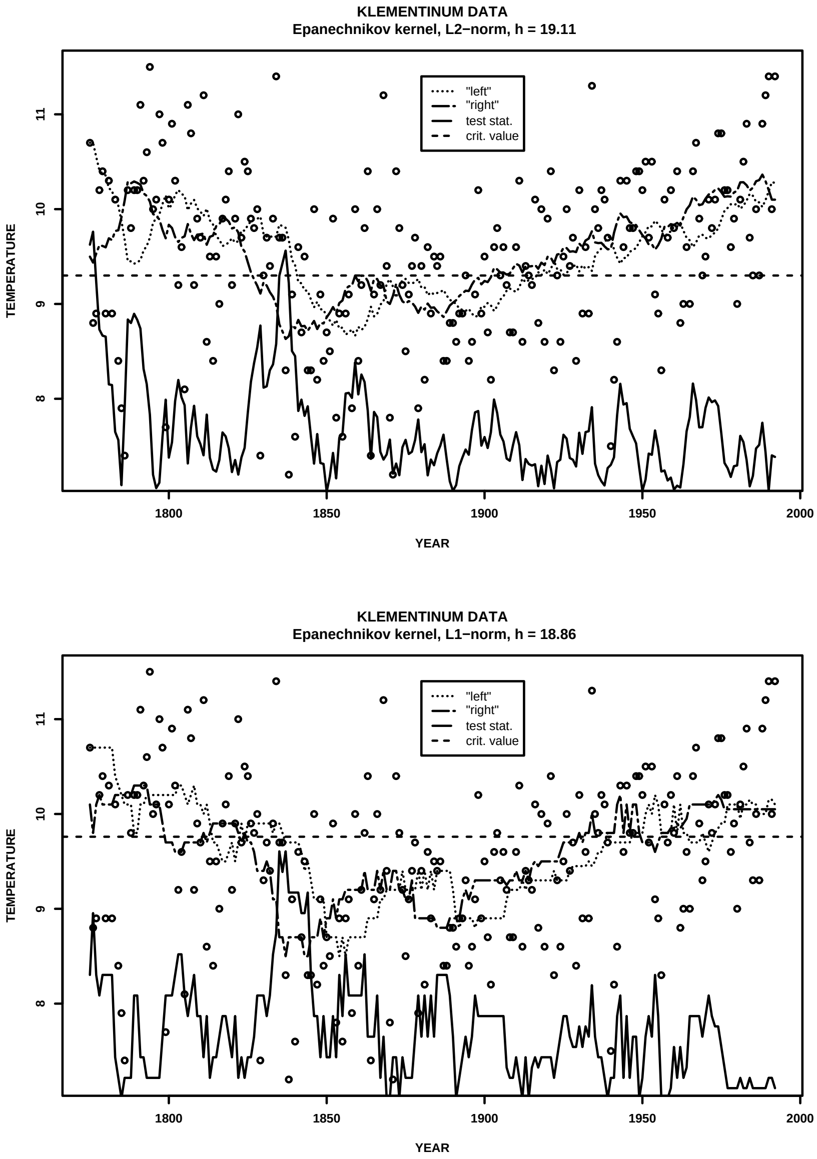
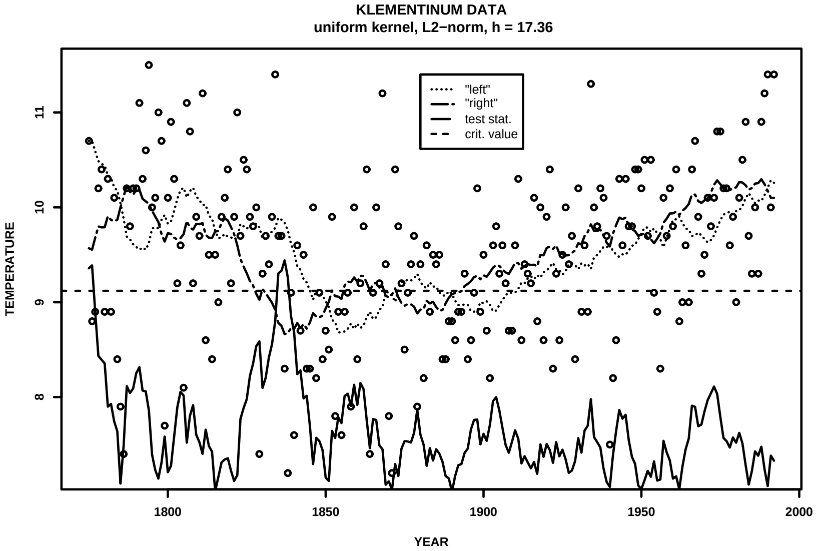


Figure 1: Klementinum data – Epanechnikov kernel.



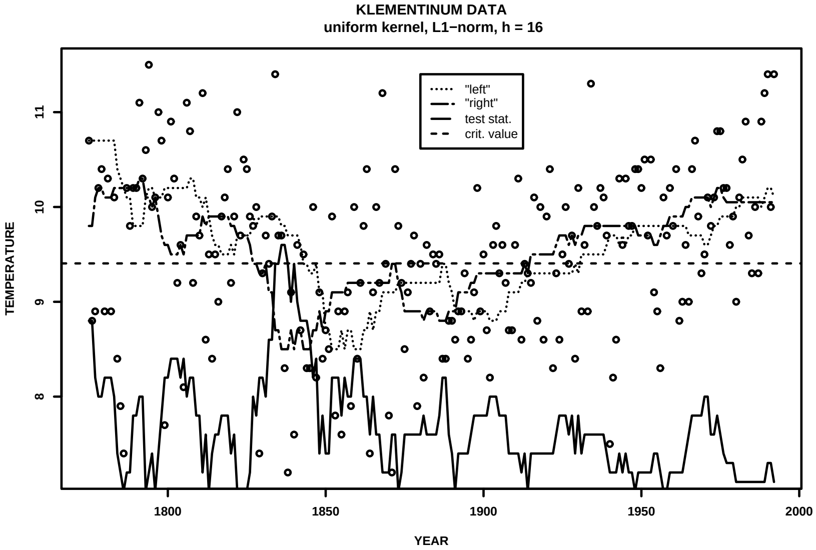


Figure 2: Klementinum data – uniform kernel.

The dashed line is the critical value for the test – the 95% quantile of the distribution of T_n^{g*}. It was estimated by resampling from $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. The resampling was repeated 200 times for each combination of kernels and loss functions. The value of the bandwidth g was calculated as $h/\log n$, see (C9).

Whenever the maximum of $\sqrt{nh}\,|m^h_{n+}(x_i) - m^h_{n-}(x_i)|$ exceeds the critical value we reject H_0. The null hypothesis was rejected in three of four cases, see Figures 1 and 2.

Peeks around the year 1836 clearly indicate a jump in the data. This result is in accord with results of other authors, see Jarušková [6] and Goderniaux [2].

Cases where ρ_1 is used do not indicate the jump as clearly as the cases using ρ_2. This may be due to the fact that LAD procedures are not so sensitive to outliers and there is probably an outlier around the year 1836. The critical value is in some figures exceeded also in the early measurements, before 1800. We are not sure about the jump in the regression at that point since it is to close to the beginning and some boundary effects of kernel estimators can play their role there.

Such result encourages the author to continue developing the method in detail and further studying the Klementinum data. The method could, for example, answer the question, whether the variance of the measurements in Klementinum is smooth along the years or not.

References

[1] Antoch J. and Janssen P. (1989), Nonparametric regression M-quantiles, *Statistics & Probability Letters*, **8**, 355 – 362

[2] Goderniaux A.-C. (2001), Automatic detection of change-points in nonparametric regression *PhD. Thesis*, Institute of Statistics, Université Catholique De Louvain-la-Neuve

[3] Härdle, W. (1984), Robust regression function estimation, *Journal of Multivariate Analysis*, **14**, 169 – 180

[4] Härdle, W. and Gasser, T. (1984), Robust nonparametric function fitting, *Journal of the Royal Statistical Society, Series B*, **46**, 42 – 51

[5] Horváth L. and Kokoszka P. (2002), Change-point detection with nonparametric regression. *Statistics*, **36**, 9 – 31

[6] Jarušková D. (1997), Some problems with application of change-point detection methods to environmental data, *Environmetrics*, **8**, 469 – 483

[7] Müller H. G. and Stadtmüller U. (1999), Discontinuous versus smooth regression, *Ann. Statist.*, **27**, 1, 299 – 337

Acknowledgement: The work was partially supported by GAČR 201/03/0945 and MSM 113200008.

Address: V. Čapek, Charles University in Prague, Department of Statistics,
Sokolovská 83, CZ – 186 75 Prague 8 – Karlín, Czech Republic

E-mail: capek@karlin.mff.cuni.cz

COMPSTAT'2004 Symposium

ROBUST ESTIMATION OF DIMENSION REDUCTION SPACE

Pavel Čížek

Key words: Dimension reduction, robust estimation, local M-estimation.

COMPSTAT 2004 section: Robustness, Dimensional reduction.

Abstract: Most dimension reduction methods based on nonparametric smoothing are highly sensitive to outliers and to data coming from heavy tailed distributions. We show that the recently proposed MAVE and OPG methods by Xia et al. [13] allow us to make them robust in such a way that preserves all advantages of the original approach. The proposed combination of MAVE and OPG with local one-step M-estimators is sufficiently robust to outliers and data from heavy tailed distributions, it is relatively easy to implement, and surprisingly, it performs as well as the original methods when applied to normally distributed data.

1 Introduction

In regression, we aim to estimate the regression function, which describes the relationship between a dependent variable $y \in \mathbb{R}$ and explanatory variables $X \in \mathbb{R}^p$. This relationship can be, without prior knowledge, modelled nonparametrically, but an increasing number of explanatory variables makes nonparametric estimation suffer from the curse of dimensionality. There are two main approaches to deal with high dimensional X variables: we can either assume a simpler form of the regression function or we can try to reduce the dimension of the space of explanatory variables. The latter, more general approach received a lot of attention recently, see Li [9] and Xia et al. [13], for instance, and it is also in the focus of our interest here.

A dimension-reduction (DR) regression model can be written as

$$y = g(B_0^\top X) + \varepsilon, \tag{1}$$

where g is an unknown smooth link function, B_0 represents a $p \times D$ orthogonal matrix, $D \leq p$, and $E(\varepsilon|X) = 0$ almost surely. Hence, to explain the dependent variable y, the space of p explanatory variables X can be reduced to a DR space given by $B_0^\top X$ with a smaller dimension D (the columns of B_0 are called directions in this context). The dimension reduction methods aim to find the dimension D and matrix B_0 defining the DR space.

Recently, Xia et al. [13] proposed a new method, the minimum average variance estimation (MAVE), that improves in several aspects over other existing estimators. First, MAVE does not need undersmoothing when estimating the link function g to achieve a faster rate of convergence. Second, MAVE can be applied to many models including time series data. Further, the MAVE approach renders generalizations of some other existing methods; for example, the outer product of gradients (OPG) estimator extends the average derivative estimator (ADE) of Härdle and Stoker [6].

Despite many features, MAVE does not seem to be robust to outliers in the dependent variable y as discussed in Čížek and Härdle [3] since it is based on local least-squares estimation. Similar sensitivity to outliers in the space of explanatory variables X (so-called leverage points), was observed and remedied in the case of the sliced inverse regression (SIR) by Gather et al. [5]. Because of many advantages that MAVE and OPG have, we address their low robustness to outlying observations and propose ways to improve them without affecting main strengths of MAVE and OPG.

The rest of the paper is organised as follows. First, we describe both the MAVE and OPG methods (Section 2). Then we propose their robust enhancements that overcome inherent computational difficulties (Section 3). Finally, we compare all estimators by means of simulations (Section 4).

2 Estimation of dimension reduction space

In this section, we present the MAVE and OPG methods (Sections 2.1 and 2.2) as well as a procedure for determining the effective DR by means of cross validation (Section 2.3).

2.1 The MAVE method

Let d represent the working dimension, $1 \leq d \leq p$. For an assumed number d of directions B in model (1), Xia et al. [13] proposed to minimize the unexplained variance $E\{y - g(B^\top X)\}^2$, where the unknown function g is locally replaced by a linear approximation; that is, $g(B^\top X) \approx a_{X_0} + b_{X_0}^\top B^\top (X - X_0)$ around some X_0. The novel feature of MAVE is that one minimizes simultaneously with respect to directions B and coefficients a_{X_0} and b_{X_0} of the local linear approximation. Hence, given a sample $(X_i, y_i)_{i=1}^n$, one minimizes

$$\min_{\substack{B:B^\top B = I_p \\ a_j, b_j, j=1,\ldots,n}} \sum_{i=1}^n \sum_{j=1}^n [y_i - \{a_j + b_j^\top B^\top (X_i - X_j)\}]^2 w_{ij}, \tag{2}$$

where w_{ij} are weights describing the local character of linear approximation. Initially, weights at any point X_0 are given by a multidimensional kernel function K_h, where h refers to a bandwidth: $w_{i0} = K_h(X_i - X_0)\{\sum_{i=1}^n K_h(X_i - X_0)\}^{-1}$; $i = 1, \ldots, n$. Additionally, once we have an estimate $\hat{B}$ of the DR space, it is possible to iterate using weights based on distances in the reduced space: $w_{i0} = K_h\{\hat{B}^\top (X_i - X_0)\}[\sum_{i=1}^n K_h\{\hat{B}^\top (X_i - X_0)\}]^{-1}$.

Xia et al. [13] also proposed a non-trivial iterative estimation procedure based on repeating two simpler optimizations of (2): one with respect to a_j, b_j given an estimate $\hat{B}$ and another with respect to B given estimates $\hat{a}_j, \hat{b}_j$. This computational approach greatly simplifies and speeds up estimation.

2.2 The OPG method

Based on the MAVE approach, Xia et al. [13] also generalised ADE of Härdle and Stoker [6] to more dimensions. Instead of using a moment condition for the gradient of the regression function g in model (1), $E\{\nabla g(X)\} = 0$, they start from the average outer product of gradients (OPG), $\Sigma = E\{\nabla g(X)\nabla^\top g(X)\}$. It can be shown that the DR matrix B consists of the d eigenvectors corresponding to the d largest eigenvalues of Σ. Recalling that local linear fitting solves for a given sample $(X_i, y_i)_{i=1}^n$

$$\min_{a_j, b_j} \sum_{i=1}^n [y_i - \{a_j + b_j^\top (X_i - X_j)\}]^2 w_{ij}, \tag{3}$$

we can estimate Σ by $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \hat{b}_j^\top \hat{b}_j$, where $\hat{b}_j$ are estimates of b_j from 3). Hence, the OPG method consists in estimating $\hat{\Sigma}$ and determining its d eigenvectors corresponding to the d largest eigenvalues. The choice of weights w_{ij} follows the same rules as for MAVE.

OPG does not exhibit the convergence rate of MAVE. On the other hand, it is easy to implement, fast to compute, and can be flexibly combined with robust estimation methods. Moreover, our simulations show that it can perform as well as MAVE if just one or two directions, $d \leq 2$, are of interest.

2.3 Dimension of effective reduction space

The described methods can estimate the DR space for a given dimension d. To determine d, Xia et al. [13] extend the cross-validation (CV) approach of Yao and Tong [14] and estimate d by $\hat{d} = \operatorname{argmin}_{0 \leq d \leq p} CV(d)$, where

$$CV(d) = \sum_{j=1}^n \left[y_j - \sum_{i=1, i\neq j}^n \frac{y_i K_h\{\hat{B}^\top (X_i - X_j)\}}{\sum_{i=1, i\neq j}^n K_h\{\hat{B}^\top (X_i - X_j)\}} \right]^2 \tag{4}$$

for $d > 0$ and $CV(0) = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$ to incorporate the possibility of y and X being independent. Note that this CV criterion can be also used to select bandwidth h, which results in two-dimensional CV over d and h. Although we use this time-demanding strategy in our simulations, in practice it is possible to consistently select bandwidth h^* in the DR space for the largest dimension $d = p$ and employ this bandwidth for all other dimensions.

3 Robust enhancements

As Čížek and Härdle [3] argued, both MAVE and OPG seem to be highly sensitive to outliers in data. Our aim is to propose robust enhancements of MAVE and OPG that should keep their present qualities, increase their robustness, and be computationally feasible. We discuss first general ways of making MAVE and OPG more robust (Section 3.1). Next, we address computational issues and try to propose robust MAVE and OPG that would be computationally feasible (Section 3.2). Finally, we adapt the CV procedure mentioned in Section 2.3 for robust estimation (Section 3.3).

3.1 Robust MAVE and OPG

There are many robust estimation methods alternative to least squares in linear regression. Using these methods in nonparametric regression adds a requirement on easy and fast implementation, and on the other hand, eliminates need for robustness against leverage points to some extent. During last decades, local L- and M-estimators have become particularly popular and well studied, see Boente and Fraiman [1], Čížek [2], and Härdle and Tsybakov [7], Fan and Jiang [4], respectively.

Theoretically, the application of local L- or M-estimation to MAVE and OPG reduces to replacing the squared residuals in (2) and (3) by a general convex function $\rho_{\hat{\sigma}}$ of residuals, where $\hat{\sigma}$ is an estimate of the variance of residuals. Many robust choices of $\rho_{\hat{\sigma}}$ depend on an estimate of residual variance, which is typically treated as a nuisance parameter or is set proportional to the median absolute deviation (MAD). To reflect local estimation window given by w_{ij} in (2) and minimize computational cost, we propose to estimate the variance of residuals in (2) and (3) by weighted MAD with weights w_{ij}.

In the case of OPG, this means that one applies a local polynomial M-estimator with a given function $\rho_{\hat{\sigma}}$ and a variance estimate $\hat{\sigma}$ and then searches the directions $\hat{B}$ as described in Section 2.2.

In the case of MAVE, having a general objective function $\rho_{\hat{\sigma}}$ leads to

$$\min_{\substack{B:B^\top B=I_p \\ a_j,b_j,j=1,\ldots,n}} \sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{\hat{\sigma}}\left[y_i - \{a_j + b_j^\top B^\top (X_i - X_j)\}\right] w_{ij}. \tag{5}$$

Although (5) cannot be minimized using the algorithm proposed by Xia et al. [13], the optimization can be still carried out by repeated estimation of (5) with respect to a_j, b_j given an estimate $\hat{B}$ and with respect to B having estimates $\hat{a}_j, \hat{b}_j$. The first step is just the already mentioned local M-estimation. The second step, estimating of B, requires a reformulation of (5). For $B = (\beta_1, \ldots, \beta_d)$ given $(a_j, b_j) = (a_j, b_{j1}, \ldots, b_{jd})$, (5) is equivalent to

$$\min_{B:B^\top B=I_p} \sum_{i=1}^{n}\sum_{j=1}^{n} \rho\left[y_i - \left\{a_j + \sum_{k=1}^{d} \beta_k^\top b_{jk}(X_i - X_j)\right\}\right] w_{ij}. \tag{6}$$

This represents a standard regression problem with n^2 observations and pd variables. Thus, it can be estimated by a regression M-estimator, and our simulations show that estimating MAVE this way leads to slightly better estimates than the original algorithm. Unfortunately, the size of regression (6) will be enormous as the sample size increases, which will hinder computation. For example, there are very fast algorithms available for computing least squares and L$_1$ regression in large data sets, see Koenker and Portnoy [8], and even in these two cases, computation becomes 10 to 20 times slower than the original algorithm for samples of 100 observations! Thus, this algorithm seems to be disqualified from practical use.

3.2 One-step estimation

To be able to employ the fast original MAVE algorithm, robustness has to be achieved only by modifying the weights w_{ij} in (2). To achieve this, we propose to use one-step M-estimators as discussed in Fan and Jiang [4] and Welsh and Ronchetti [12]. More precisely, we employ the (iteratively) reweighted least squares approach: using an initial highly robust estimate $\hat{\beta}^0 = \{\hat{B}^0, \hat{a}_j^0, \hat{b}_j^0\}$, we construct weights w_{ij} such that the objective function (2) is equivalent to (5) at $\hat{\beta}^0$ and then perform the original algorithm using the constructed weights. As the initial robust estimator, we use L$_1$ regression, which guarantees robustness against outliers and fast computation.[1]

3.3 Robust cross-validation

The robust estimation of the DR matrix B_0 is not sufficient if dimension d is not known. Using the CV criterion described in Section 2.3 can lead to a bias in dimension estimation (and bandwidth selection) even if a robust estimator of B_0 is used, see Ronchetti et al. [10]. Now, due to the local nature of nonparametric regression, we have to “protect” primarily against outliers in the y-direction. In this context, the L$_1$ estimator is highly robust and the same should apply to CV based on L$_1$ rather than L$_2$ norm, which was already studied by Wang and Scott [11]. Thus, we propose to use instead of (4)

$$CV(d) = \sum_{j=1}^{n} \left| y_j - \sum_{i=1, i \neq j}^{n} \frac{y_i K_h\{\hat{B}^\top (X_i - X_j)\}}{\sum_{i=1, i \neq j}^{n} K_h\{\hat{B}^\top (X_i - X_j)\}} \right|. \tag{7}$$

4 Simulations

In this section, we study finite sample properties of discussed methods by means of simulations. First, we introduce models used for simulations (Section 4.1). Next, we compare original OPG and MAVE and their modifications proposed in Section 3 (Section 4.2). Finally, we examine the performance of the L$_1$ and L$_2$ CV dimension selection (Section 4.3).

[1] Assuming a robust choice of bandwidth, one does not have to protect against leverage points: estimation is done in a local window given by the bandwidth and a kernel function.

4.1 Simulation models

Throughout this section, we consider the following nonlinear model

$$y_i = (X_i^\top \beta_1)^2 - (0.5 + X_i^\top \beta_2)^2 + 15\cos(X_i^\top \beta_3) + 0.5\varepsilon_i, \tag{8}$$

where all explanatory variables have the standard normal distribution in $\mathbb{R}^{10}$ and $\beta_1 = (1,2,3,0,0,0,0,0,0,0)/\sqrt{14}$, $\beta_2 = (-2,1,0,1,0,0,0,0,0,0)/\sqrt{6}$, and $\beta_3 = (0,0,0,0,0,0,0,1,1,1)/\sqrt{3}$. Thus, the true dimension of the reduced space is $D = 3$. To compare the robust properties of all estimators, we use three error distributions: $\varepsilon_i \sim N(0,1)$, $\varepsilon_i \sim t_1$, and $\varepsilon_i \sim 0.95N(0,1) + 0.05U(-600,600)$, where $N(\mu,\sigma), t_f$, and $U(a,b)$ refer to the normal, Student, and uniform distributions. For the sake of brevity, we refer further to these three cases as NORMAL, STUDENT, and OUTLIERS, respectively. Presented simulation results are done for sample size $n = 100$, 100 repetitions, the Gaussian kernel and bandwidth chosen by CV (7).

Let us note that we compare the performance of all methods by the distance of the estimated space $\hat{B}$ and the true space $B_0 = (\beta_1, \beta_2, \beta_3)$. It can be measured by $m_0(\hat{B}) = m(\hat{B}, B_0) = \|(I - B_0 B_0^T)\hat{B}\|$ for $d \leq D = 3$ and by $m_0(\hat{B}) = m(\hat{B}, B_0) = \|(I - \hat{B}\hat{B}^T)B_0\|$ for $d \geq D = 3$, see Xia et al. [13].

4.2 Estimation of dimension reduction space

Using simulations described in Section 4.1, we compare MAVE and its modifications (given the limited space, we omit results on OPG). The compared methods include MAVE defined by (5) using the following functions $\rho_{\hat{\sigma}}$: (i) $\rho_{\hat{\sigma}}(x) = x^2$ (LS, the original MAVE); (ii) $\rho_{\hat{\sigma}}(x) = |x|$ (L1), (iii) $\rho_{\hat{\sigma}}(x) = \int \operatorname{sgn}(x)\min(|x|, \hat{\sigma}))dx$ (HUBER), (iv) $\rho_{\hat{\sigma}}(x) = \int \operatorname{sgn}(x)\max\{0, \min(|x|, \hat{\sigma}) - \max(0, |x| - 2\hat{\sigma})\}dx$ (HAMPEL).[2] The function L1 was implemented by algorithm described in Section 3.1 and HUBER and HAMPEL as local one-step M-estimators discussed in Section 3.2.

The simulation results are summarized in Table 1 and clearly document the need for and advantages of robust modifications of MAVE. For data NORMAL, all estimates except for MAVE HAMPEL perform almost equally well. MAVE HAMPEL provides worse estimates since it fully rejects data points with residuals larger than $3\hat{\sigma}$. For data STUDENT, all robust versions of MAVE provide similar results, whereas the estimates by original MAVE LS exhibit rather large errors, especially in the first direction β_1. For data OUTLIERS, there is a large difference between non-robust MAVE LS and robust modifications of MAVE, which documents sensitivity of the original MAVE estimator to outliers. Although all robust estimators perform similarly, MAVE HAMPEL is best due to its full-rejection feature. Let us note that results can be further improved by adjusting function $\rho_{\hat{\sigma}}$.

[2] The functions named HUBER and HAMPEL correspond to the standard Huber and Hampel ψ-functions with parameters 1 and 1, 2, 3, respectively.

Data	**Method**	**Parameter**			
		$m_0(\hat{\beta}_1)$	$m_0(\hat{\beta}_2)$	$m_0(\hat{\beta}_3)$	$m_0(\hat{B})$
NORMAL	MAVE LS	0.004	0.089	0.094	0.147
	MAVE L1	0.005	0.061	0.097	0.149
	MAVE HUBER	0.005	0.100	0.095	0.164
	MAVE HAMPEL	0.006	0.116	0.155	0.252
STUDENT	MAVE LS	0.252	0.397	0.510	0.910
	MAVE L1	0.037	0.316	0.443	0.722
	MAVE HUBER	0.035	0.284	0.451	0.685
	MAVE HAMPEL	0.039	0.294	0.428	0.633
OUTLIERS	MAVE LS	0.752	0.682	0.680	0.976
	MAVE L1	0.029	0.156	0.218	0.465
	MAVE HUBER	0.014	0.228	0.202	0.416
	MAVE HAMPEL	0.010	0.151	0.161	0.305

Table 1: Median errors of MAVE estimates for dimension $d = 3$.

CV, CVA		**Estimated dimension** d				
Data	**Method**	1	2	3	4	≥5
NORMAL	LS	8, 8	48, 59	38, 31	5, 1	1, 1
NORMAL	HAMPEL	11, 8	48, 55	39, 35	2, 2	0, 0
OUTLIERS	LS	2, 14	1, 5	10, 8	12, 8	75, 65
OUTLIERS	HAMPEL	0, 9	0, 43	1, 24	13, 9	86, 15

Table 2: Estimates of the DR dimension by OPG LS and OPG HAMPEL using L_2 (CV, first number) and L_1 (CVA, second number) cross validation. Entries represent numbers of samples out of 100 with estimated dimension d.

4.3 Cross-validation simulations

Simulating again from model (8), we estimated the DR dimension $\hat{d}$ ($D = 3$) using OPG with $\rho_{\hat{\sigma}}$-functions LS and HAMPEL and the L_2- and L_1-based CV criteria (4) and (7). Results for data NORMAL and OUTLIERS are summarized in Table 2. Clearly, one can be almost indifferent between the CV criteria for data NORMAL. For data OUTLIERS, only the combination of a robust estimator with L_1-CV leads to results resembling the true model. When judging results in Table 2, one should take into account that this nonlinear inference is based on only 100 observations in $\mathbb{R}^{10}$. Nevertheless, results can be improved, for example, by using more time-demanding MAVE.

5 Conclusion

We proposed robust enhacements of MAVE and OPG that are as good as the original methods under 'normal' data, are robust to outliers and heavy-tailed distributions, and are easy to implement. Should we pick up one method for a general use, MAVE HUBER seems to be the most suitable candidate as (i) MAVE LS is not robust, (ii) MAVE L1 is slow to compute, see Section 3.1, and (iii) MAVE HAMPEL does not perform so well for normal data.

References

[1] Boente G., Fraiman R. (1994). *Local L-estimators for nonparametric regression under dependence.* J. Nonparametr. Stat. **4**, 91 – 101.

[2] Čížek P. (2004). *Smoothed local L-estimation with an application.* In Hubert M., Pison G., Struyf A., and Van Aelst S. (eds.), Theory and Applications of Recent Robust Methods, Birkhäuser, Basel, in press.

[3] Čížek P., Härdle W. (2003). *Robust adaptive estimation of dimension reduction space.* SFB 373 Discussion Paper 1/2003.

[4] Fan J., Jiang J. (1999). *Variable bandwidth and one-step local M-estimator.* Sci. China Ser. A **29**, 1 – 15.

[5] Gather U., Hilker T., Becker C. (2001). *A robustified version of sliced inverse regression.* In Fernholz et al. (eds.) Statistics in Genetics and in the Environmental Sciences, Birkhäuser, Basel, 147 – 157.

[6] Härdle W., Stoker T. M. (1989). *Investigating smooth multiple regression by method of average derivatives.* J. Amer. Statist. Assoc. **84**, 986 – 995.

[7] Härdle W., Tsybakov A. B. (1988). *Robust nonparametric regression with simultaneous scale curve estimation.* Ann. Statist. **16**, 120 – 135.

[8] Koenker R., Portnoy S. (1997). *The Gaussian hare and the Laplacian tortoise: computability of squared-error vs. absolute-error estimators.* Statist. Sci. **12**, 279 – 300.

[9] Li K. C. (1991). *Sliced inverse regression for dimension reduction.* J. Amer. Statist. Assoc. **86**, 316 – 342.

[10] Ronchetti E., Field C., Blanchard W. (1997). *Robust linear model selection by cross-validation.* J. Amer. Statist. Assoc. **92**, 1017 – 1023.

[11] Wang F. T., Scott D. W. (1994). *The L_1 method for robust nonparametric regression.* J. Amer. Statist. Assoc. **89**, 65 – 76.

[12] Welsh A.H., Ronchetti E. (2002). *A journey in single steps: robust one-step M-estimation in linear regression.* J. Statist. Plann. Inference **103**, 287 – 310.

[13] Xia Y., Tong H., Li W. K., Zhu L.-X. (2002). *An adaptive estimation of dimension reduction space.* J. Roy. Statist. Soc. Ser. B **64**, 363 – 410.

[14] Yao Q., Tong H. (1994). *On subset selection in nonparametric stochastic regression.* Statist. Sinica **4**, 51 – 70.

Acknowledgement: This work was supported by grant GA ČR 402/03/0084.

Address: P. Čížek, Charles University, Faculty of Social Sciences, Institute of Economic Studies; Opletalova 26, 110 00 Praha 1, Czech Republic

E-mail: `P.Cizek@uvt.nl`

COMPSTAT'2004 Symposium

NONPARAMETRIC UNSUPERVISED CLASSIFICATION OF SATELLITE WAVE ALTIMETER FORMS

Sophie Dabo-Niang, Frédéric Ferraty and Philippe Vieu

Key words: Curves data, curves classification, functional statistics, modal curve, altimetric curves, satellite data.

COMPSTAT 2004 section: Functional data analysis.

Abstract: We present a new unsupervised classification method for curves. The main feature of our approach is to take fully into account the continuous nature of these data. The presentation will be centered around a real data problem coming from geophysical sciences, and which consists in classifying a set of altimetric curves registered by the satellite Topex/Poseidon upon the Amazonian basin.

1 Introduction

Let us look at the following data, corresponding to altimetric measurements registered by the satellite Topex/Poseidon around an area of 25 kilometers upon the Amazon River. Each altimetric data, is represented by its wave on the range $(0, 70)$, and the satellite is registering 10 curves each second. Indeed we just kept a sample of 500 waves, randomly selected among all the numerous curves that we had at hand. It is known that each wave is linked with the kind of ground treated by the satellite, and the idea for the Amazonian basin is to use these altimetric data for hydrological purpose.

To fix the idea, we present in Figure 1 a set of 20 randomly selected among these curves. A deeper presentation of these data can be found in Frappart [2]. We can see clearly from Figure 1 that there are several different forms of waves, that could be interpreted by different kinds of grounds. With other words, in the Amazonian basin, these different forms are corresponding to different levels of water (lake, small river, large river, ...). Some of these curves are very flat (for instance curves numbers 13 and 42). They are known to be noise spanned by the satellite and they should be ignored in the remaining of the study. Other wave curves are of several different forms: there are curves with one peak and small tails (for instance curves numbers 23 and 24); there are other curves with one peak and high tails (for instance curves numbers 3 and 138); there are curves that look to have more than one peak (for instance curves numbers 1 and 7); there are also curves that looks without peak (for instance curves numbers 5 and 26); ... The reader can find in Frappart [2] some comments about the possible links between these different forms and different kind of waters.

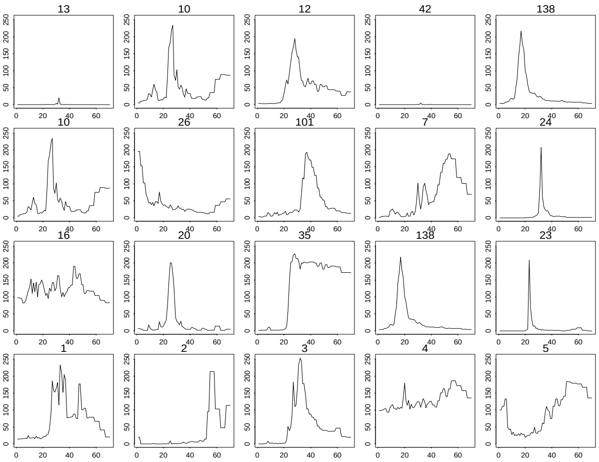


Figure 1: Some wave altimetric curves.

So, the questions are simple: Are these curves separable in different groups? And, if it is the case, how could we assign some given curve to one among these groups? With statistical words this problem can be summarized by the following question: How can we classify these altimetric curves data? This question is known as an unsupervised curves classification problem. It is an unsupervised problem because we do not have at hand any categorical response variable, and it is a curve classification problem because the data sets are clearly of functional nature. Advances on *Functional Statistics* (see Ramsay and Silverman [6] and Ferraty and Vieu [3] and [4], for references) allow to have at hand several different statistical procedures for dealing with such kind of data, including supervised curves classification problem. The unsupervised classification problem is much more difficult because the groups are unknown before the study.

2 The nonparametric functional methodology

The statistical modelling for treating curves data consist in looking at them as a sample of independent realizations $X_1, \ldots, X_n$ of some functional variable X taking values in some abstract infinite dimensional space (E, d), d being some measure of proximity between curves (for instance a metric or a semi-metric). For instance, in the altimetric data set discussed before, each X_i is a whole wave curve, $X_i = \{X_i(t), t \in (0, 70)\}$.

Let $S \subset E$ be some subset of curves. Keeping in mind what happens in the unfunctional case (that is for finite dimensional data), the most natural centrality index for classification purpose would be the notion of modal curve rather than the notion of mean curve or median curve. A first work in this direction has been done by Gasser et al. [5] and the recent paper by Dabo-

Niang et al. [1] gives asymptotic support to these ideas when d is a metric. According to these authors, the modal curve, denoted $X_{modal,S}$, is defined as the curve of the sample S which maximizes the quantity

$$Q_{mod}(x) = \sum_{\{i, X_i \in S\}} K\left(\frac{d(X_i, x)}{h}\right).$$

Let us just tell that in the altimetric application presented below we will use the classical quadratic weight function (called too *kernel*):

$$K(u) = \frac{3}{2}(1 - u^2)1_{u \in (0,1)}.$$

On the other hand, it is possible to extend to the functional case the notion of median curve, denoted $X_{median,S}$, by considering the one of the sample S which minimizes the quantity:

$$Q_{med}(x) = \sum_{\{i, X_i \in S\}} d(X_i, x).$$

Clearly, as in the unfunctional case, both the mean and the median curves are interesting only in the situation of homogeneous data, while the modal curve would be more useful for detecting any structural differences between data. This is exactly the idea that we are going to use for our classification purpose, and we will look for some difference (in the sense of the measure d) between the modal curve $X_{modal,S}$ and one among $X_{mean,S}$ or $X_{median,S}$. Because of the horizontal shift in the curves it is much more reasonable for these altimetric data to use $X_{median,S}$ rather than $X_{mean,S}$ (but the same methodology could be used with $X_{mean,S}$ and could be efficient for other functional data sets). We would like to consider the quantities:

$$\frac{d(X_{modal,S}, X_{median,S})}{d(X_{median,S}, 0) + d(X_{modal,S}, 0)},$$

but, in order to get a more stable criterion, we use L (randomly generated) subsamples $S^{(l)} \subset S$ ($l = 1, \ldots, L$, each subsample being of the same size M_S), and we introduce the following heterogeneity charateristic for the set of curves S:

$$HIR(S) = \frac{1}{L}\sum_{l=1}^{L} \frac{d(X_{modal,S^{(l)}}, X_{median,S^{(l)}})}{d(X_{median,S^{(l)}}, 0) + d(X_{modal,S^{(l)}}, 0)}.$$

Concerning our altimetric example, these parameters were taken to be $L = 50$ and $M_S = Card(S)/2$.

This quantity will be used to reply to the important question: Does a set of curves S need to be splitted into the K subgroups $S_1, \ldots S_K$ or does not? This can be done by heuristic considerations according to the experience of the user, but in other cases there is necessity for some automatic stop criterion. This automatic procedure is based on a comparison between $HIR(S)$ and its splitted version:

$$HIR(S; S_1, \dots S_K) = \frac{1}{Card(S)} \sum_{k=1}^{K} Card(S_k) HIR(S_k),$$

and we will use as a statistics to decide for splitting or not

$$GAIN = GAIN(S; S_1, \dots S_K) = \frac{HIR(S) - HIR(S; S_1, \dots S_K)}{HIR(S)}.$$

We will do (*respectively* undo) the splitting of S into $S_1, \dots S_K$ if $GAIN$ is (*respectively* is not) less that some fixed threshold γ. To fix the ideas, note that for our altimetric example below this threshold was taken to be $\gamma = 0$, but for other data sets this threshold can be taken larger.

As usual in nonparametric setting, as well for functional as for finite dimensional purposes, the choice of the smoothing factor is a crucial point to insure good behaviour of the procedure. So the last natural question to be answered before applying our methodology is the following one: How could we choose the bandwidth parameter h? To do that, the procedure has to take into account the objective of the study. Concretely here, that means that this parameter has to be chosen in function of some criterion which is adapted to the classification problem. This is done by estimating the small ball probabilities $P(X \in B(x,h))$, which play a key role in the theoretical properties of our mode estimate. Indeed, it has been shown by Dabo-Niang et al. [1, Theorem 4] that the modal curve $X_{modal,S}$ converges almost surely to the mode of the distribution of the X_i's with the rate $O(h) + O\left(\left(\frac{\log n}{n\psi(h)}\right)^{1/4}\right)$, where $\psi(h)$ is strongly linked to the quantity $P(X \in B(x,h))$. The bandwidth is selected such that these small ball probabilities exhibit the strongest heterogeneity (see Dabo-Niang et al. [1] for details).

3 Classification of altimetric curves

The only thing to be chosen now is the function d that has to be used for measuring proximity between curves. This function plays a great role to insure good properties of the modal curve (and therefore of our classification procedure), as illustrated in the asymptotic theory developped in Dabo-Niang et al. [1]. The choice of d can be driven from the practical problem which is investigated, that is from our altimetric context.

Choice of d. First of all look again at Figure 1 above, and note that there is clearly some horizontal shift between curves. Even the curves numbered 138, 24 and 23 that are apparently of the same kind, are affected by this horizontal shift. So, there is a real need for constructing a function d which is invariant by translation. In a second attempt, it is usual in functional data analysis to consider successive derivatives of the curves instead of themselves. But here, looking again at Figure 1, the altimetric curves are very irregular and so it makes sense to construct a criterion of comparison which is based directly on the curves themselves. For all these reasons, we decided to use as a proximity measure for these altimetric curves, the following function d:

$$d(x,y) \; = \; inf_{\alpha\in(-\tau,+\tau)}\frac{1}{b-a-2|\alpha|}\int_{a+|\alpha|}^{b-|\alpha|} \left(x(t+\alpha)\; - \; y(t-\alpha)\right)^2 \, dt,$$

where $(a,b) = (0,70)$ is the range of each wave altimetric curve, and where the rescaling parameter τ has been chosen such that $\tau = \frac{b-a}{5} = 14$.
Previous processing. This classification method has been used on our set of 500 altimetric curves. Note that, before doing that, we supressed (by hand) all the "flat" curves which are known to be noise of the sattelite (that is, all the curves looking like numbers 13 and 42 in Figure 1). We did that by hand, but obviously these curves could be easily automatically putten in some special group. Because they are very different from the other ones (flatness against strong irregularity), the same classification method but with some other measure of proximity d (for instance, with a new d based on the first derivative of the curve) could be used. We did not do this step here for two reasons. The first one is because it was very easy to detect by hand. Secondly, the non interest of these "flat" curves is very well-known by the geophysical scientists. So finally, we used our method on the $n = 472$ remaining curves. Let us denote by $GROUP0$ this first set of curves.
Classification in action. In Figure 2, the results of the different iterations of our procedure are presented. As we can see in Figure 2, at the first iteration our set of altimetric curves has been splitted into two groups, namely into $GROUP\,1$ and $GROUP\,2$. Because the $GAIN$ in terms of homogeneity index HIR was important (.44), this splitting has been accepted.

So, the second iteration of the method has been done twice: firstly for the $GROUP\,1$ and then for the $GROUP\,2$. The $GROUP\,1$ was also splitted into $GROUP\,11$ and $GROUP\,12$, and this splitting was again accepted since the homogeneity $GAIN$ was .70. The $GROUP\,2$ was also candidate to be splitted into two subgroups (that means that the corresponding density $\hat{d}_{GROUP\,2}$ had again two modes), but the corresponding subgroups led to a loss of homogeneity ($GAIN = -.22$) and so the splitting of $GROUP\,2$ was rejected.

The third iteration has only to be done twice: firstly for the $GROUP\,11$ and then for the $GROUP\,12$. As indicated in Figure 2 each of them was candidate to a new splitting . However, the corresponding splittings would lead to a loss of homogeneity ($GAIN = -.50$ or $-.11$) and so both of them have been rejected.

Finally, we got a classification of our set of curves into 3 subgroups, having respective sizes $Card(GROUP\ 2) \;=\; 214$, $Card(GROUP\ 11) \;=\; 92$, and $Card(GROUP\ 12) \;=\; 166$. Figure 3, 4, and 5 display random subsamples of each of the selected groups. It appears that $GROUPS$ 1.1 and 1.2 are quite homogeneous. The last one, $GROUP$ 2, contains all the curves presenting very different shapes. Obviously, it could be possible to split again this group (for instance by using the same method but with other d).

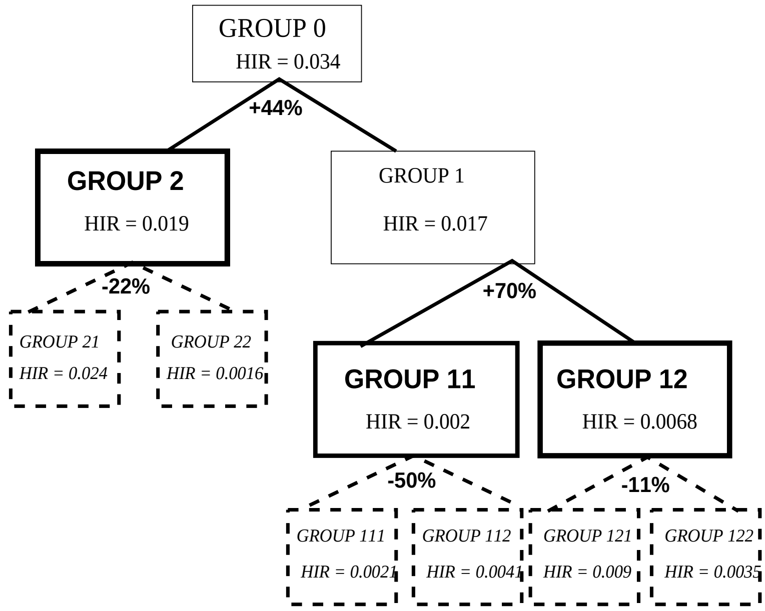


Figure 2: Results of the classification method for altimetric curves.

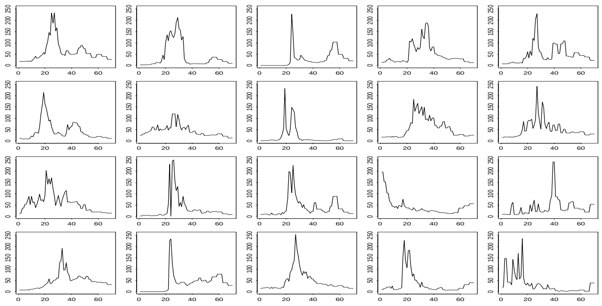


Figure 3: Subsample of GROUP 11.

4 Concludings

The results of this study show that the recent nonparametric approach for estimating the mode of a sample of curves provides a nice new tool

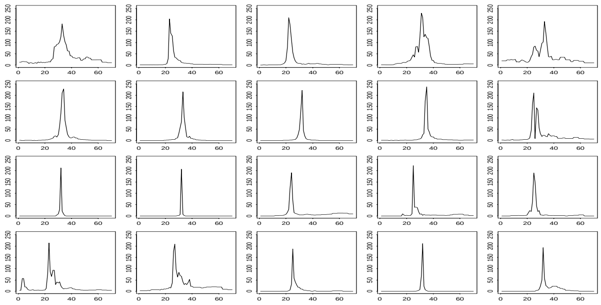


Figure 4: Subsample of GROUP 12.

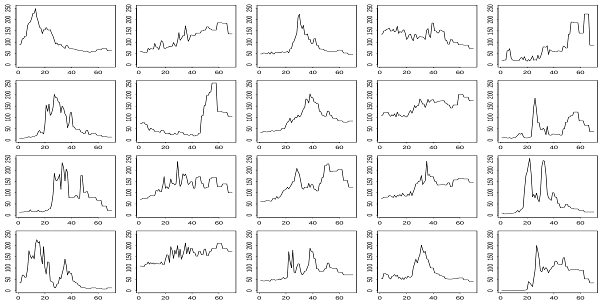


Figure 5: Subsample of GROUP 2.

for classifying curves data. Concerning the altimetric curves, this new functional approach for classification gives good results in the sense that the algorithm produces automatically homogeneous subgroups that can be easily interpreted by geophysicians in terms of difference of grounds (and therefore, in the Amazonian basin in terms of river, lake, ...).

References

[1] Dabo-Niang S., Ferraty F., Vieu P. (2004). *Mode estimation for functional random variable and its application for curves classification.* Preprint.

[2] Frappart F. (2003). *Catalogue des formes d'onde de l'altimètre Topex/Poséidon sur le bassin amazonien.* Technical Report, CNES, Toulouse, France.

[3] Ferraty F. and Vieu P. (2003a). *Functional nonparametric statistics: a double infinite dimensional framework.* In "Recent Advances and Trends in Nonparametric Statistics", M. Akritas and D. Politis (eds), Elsevier 61 − 79.

[4] Ferraty F. and Vieu P. (2003b). *Curves discrimination: a nonparametric functional approach.* Computational Statistics and Data Analysis **44**, 161 − 173.

[5] Gasser T., Hall P. and Presnell B. (1998). *Nonparametric estimation of the mode of a distribution of random curves*, J. R. Statist. Soc, B **60**, 4, 681 − 691.

[6] Ramsay J.O., Silverman B.W. (1997). *Functional dat a analysis.* Springer Series in Statistics.

Acknowledgement: The authors would like to thank all the participants to the working group STAPH, on Functional and Operatorial Statistics in Toulouse, for their continuous comments and encouragements, as well as Frédéric Frappard from the LEGOS-CNES who kindly provided us the altimetric data set.

Address: S. Dabo-Niang, Univ. Pierre et Marie Curie, Paris VI, France
F. Ferraty, Univ. Paul Sabatier, Toulouse 3 and Univ. Toulouse le Mirail, Toulouse 2, France
P. Vieu, Univ. Paul Sabatier, Toulouse 3, France

E-mail: `ferraty@cict.fr`

COMPSTAT'2004 Symposium

ROBUST REGRESSION QUANTILES WITH CENSORED DATA

Michiel Debruyne and Mia Hubert

Key words: Robust statistics, regression quantiles, regression depth, censored data.

COMPSTAT 2004 section: Robustness.

Abstract: In this paper we propose a method to robustly estimate linear regression quantiles with censored data. We adjust the estimator recently developed by Portnoy by replacing the Koenker-Bassett regression quantiles with the regression depth quantiles. The resulting optimization problem is solved iteratively over a set of grid points. We show on some examples that, contrary to the Koenker-Bassett approach, this estimator can resist bad leverage points.

1 Introduction

We consider the linear regression setting, in which we have to model the response variable Y at a certain p-dimensional vector $\mathbf{x}$. For each $0 < \tau < 1$, denote $Q_\tau(Y|\mathbf{x})$ the τth quantile of the conditional distribution of Y given $\mathbf{x}$. The regression quantile model states that this conditional distribution is linear in $\mathbf{x}$ or

$$Q_\tau(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_\tau. \tag{1}$$

We will try to estimate the parameters $\boldsymbol{\beta}_\tau$, given a data set which may include right censored data. This means that for every data point $1 \leq i \leq n$, the covariates $\mathbf{x}_i \in \mathbb{R}^p$ are measured, as well as censoring times c_i. Let y_i be the true response, possibly unobserved, then the observed responses $(\tilde{y}_i, \Delta_i)$ satisfy

$$\tilde{y}_i = \min\{y_i, c_i\} \qquad \text{and} \qquad \Delta_i = I(y_i \leq c_i)$$

with I the indicator function. So we only observe the smallest of y_i and c_i and we know whether y_i is censored or not. If there is no censoring, the observed $\tilde{y}_i$ are all equal to the y_i.

In Section 2 we will briefly discuss the estimator recently developed by Portnoy [4]. This estimator is based on the classical Koenker-Bassett estimator and therefore is quite sensitive to leverage points. In Section 3 we will discuss a more robust estimator, the deepest regression estimator [5], which has been defined for non censored data. In Section 4 we will extend this estimator to the censored case along the lines of Portnoy's estimator. In Section 5 we provide an example.

2 The Koenker-Bassett estimator

In case of uncensored data, we can use the Koenker-Bassett estimator [3] to estimate regression quantiles. This estimator is defined as

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum \rho_\tau(r_i(\boldsymbol{\beta}))$$

with $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ the ith residual and $\rho_\tau(u) = u(\tau - I(u < 0))$. For $\tau = 0.5$, $\hat{\boldsymbol{\beta}}_\tau$ coincides with the L_1-estimator which minimizes the sum of the absolute residuals.

In case of censored data a weighted version of this Koenker-Bassett estimator has been introduced by Portnoy [4]. The quantiles are iteratively estimated for $0 < \tau < 1$. At the start all observations have weight 1. When the ith censored observation is crossed for $\tau = \hat{\tau}_i$ (this means its residual with respect to the τth regression quantile is negative if $\tau > \hat{\tau}_i$), weights are defined in two pseudo-observations. One pseudo-observation coincides with the most recently crossed observation and receives a weight $w_i(\tau) = (\tau - \hat{\tau}_i)/(1 - \hat{\tau}_i)$. Note that this weight is an estimate of the probability of the censored observation lying in between the $\hat{\tau}_i$th and τth quantile. The second pseudo-observation is put arbitrarily far away and receives a weight $1 - w_i(\tau)$, which is an estimate of the probability of the censored observation lying above the τth quantile. In each step a weighted version of the objective function is minimized. More precisely, denote K the set of censored observations that have previously been crossed, then $\hat{\boldsymbol{\beta}}_\tau$ is chosen to minimize

$$\sum_{i\notin K} \rho_\tau(r_i(\boldsymbol{\beta})) + \sum_{i\in K} \{w_i(\tau)\rho_\tau(c_i - \mathbf{x}_i'\boldsymbol{\beta}) + (1 - w_i(\tau))\rho_\tau(y^* - \mathbf{x}_i'\boldsymbol{\beta})\}$$

over all hyperplanes $\boldsymbol{\beta}$ through p data points. The number y^* is any value sufficiently large to exceed all $\{\mathbf{x}_i'\boldsymbol{\beta} : i \in K\}$. Iterative computation can be done using a simplex pivoting algorithm, as described in [4].

3 Regression depth quantiles

Regression depth has been introduced by Rousseeuw and Hubert [5]. For each $\boldsymbol{\theta} \in \mathbb{R}^p$, the regression depth of $\boldsymbol{\theta}$ with respect to the data set $Z_n = \{(\mathbf{x}_i, y_i)\}$ is defined as

$$rdepth(\boldsymbol{\theta}, Z_n) = \min_{\boldsymbol{\lambda}\in\mathbb{R}^p} (\#\{\mathbf{x}_i : sgn(r_i(\boldsymbol{\theta})) \neq sgn(\mathbf{x}_i'\boldsymbol{\lambda})\})$$

with $sgn(u) = -1$ if $u < 0$, $sgn(u) = 0$ if $u = 0$ and $sgn(u) = 1$ if $u > 0$.

In case of simple regression this definition can be interpreted as follows. A line $\boldsymbol{\theta}$ can be rotated around any point v lying on $\boldsymbol{\theta}$, until it becomes a vertical line. This can be done clockwise or counterclockwise. Both ways, the minimum number of data points that is passed is counted. Finally this can be repeated for every point v and the overall minimum of crossed data

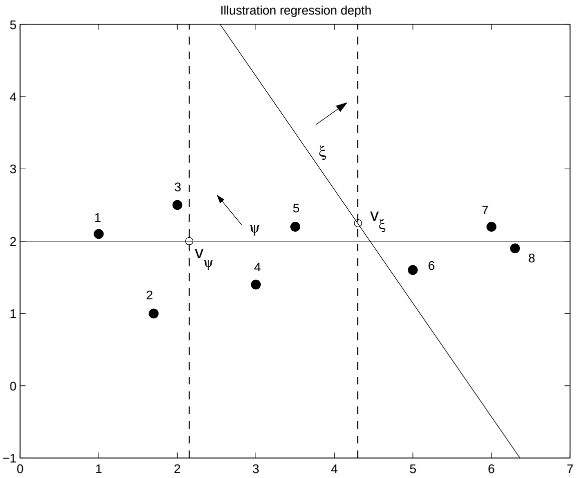


Figure 1: Example with 8 data points. The regression depth of the line ξ is 0, whereas the depth of the line ψ is 3.

points is retained. This is exactly the regression depth of $\boldsymbol{\theta}$. An illustration is given in Figure 1. The line ξ has regression depth 0, since it can be rotated clockwise around the point v_ξ without encountering any data points. The regression depth of the line ψ is 3, since we encounter 3 data points if we rotate it counterclockwise around v_ψ and we can find no lower number than this.

For general p, a hyperplane with high regression depth is well surrounded by data points as we will always find a large number of observations when it is rotated. So it can be expected that hyperplanes with high regression depth are quite good fits. Rousseeuw and Hubert ([5]) defined the deepest regression as

$$DR(Z_n) = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^p} rdepth(\boldsymbol{\theta}, Z_n).$$

It was shown in [2] that DR is a consistent estimator for the conditional median $\boldsymbol{\beta_\tau}$ with $\tau = 0.5$.

The deepest regression has a breakdown value of 33%, which means that it can resist up to 33% of outliers in the data set. This is not the case for the Koenker-Bassett estimator, which can be heavily influenced by even one single outlier. More specifically, the estimator can resist vertical outliers but not leverage points which are outlying in the $\mathbf{x}$-space.

For general τ the deepest regression can be generalized similar to the Koenker-Bassett estimator. This yields the regression depth quantiles:

$$\hat{\boldsymbol{\beta}}_\tau = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \inf_{\boldsymbol{\lambda}\in\mathbb{R}^p} \Big(\tau\ \#\{y_i : (r_i(\boldsymbol{\beta}) > 0, \mathbf{x}_i'\boldsymbol{\lambda} < 0)\} \qquad (2)$$
$$+(1-\tau)\ \#\{y_i : (r_i(\boldsymbol{\beta}) < 0, \mathbf{x}_i'\boldsymbol{\lambda} > 0)\}\Big)$$

4 Depth quantiles with censored data

The ideas to obtain Koenker-Bassett regression quantiles for censored data can now be extended to the regression depth quantiles. A schematic overview of the algorithm is as follows.

STEP 1 Choose a set of grid points $\{0 < \tau_1 < \ldots < \tau_m < 1\}$. The number of grid points m needed to find good estimates is quite low: Portnoy already suggested $\sqrt{n}$ in [4] and this was also sufficient in our examples.
Estimate the τ_1th regression quantile using the regression depth quantile for uncensored data (2). Crossed censored observations can be ignored since they almost do not contain any information.

STEP 2 Suppose we have estimated the τ_lth regression quantile $\hat{\boldsymbol{\beta}}_{\tau_l}$. Then we also know the set of crossed censored observations $K_{\tau_l} = \{(\mathbf{x}_i, c_i) : c_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\tau_l} \leq 0\}$. For each of these crossed censored observations a number $\hat{\tau}_i$ has been given following equation (3) that will be explained in step 3 of the algorithm. The according weight is

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i}.$$

STEP 3 Estimate the τ_{l+1}th regression quantile using a weighted version of (2).

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\tau_{l+1}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \inf_{\boldsymbol{\lambda}\in\mathbb{R}^p} \Big(&\tau\ \#\{\tilde{y}_i \notin K_{\tau_l} : (r_i(\boldsymbol{\beta}) > 0, \mathbf{x}_i'\boldsymbol{\lambda} < 0)\}\\ &+ (1-\tau)\ \#\{\tilde{y}_i \notin K_{\tau_l} : (r_i(\boldsymbol{\beta}) < 0, \mathbf{x}_i'\boldsymbol{\lambda} > 0)\}\\ &+ \tau\ w_i(\tau)\ \ \#\{\tilde{y}_i \in K_{\tau_l} : (r_i(\boldsymbol{\beta}) > 0, \mathbf{x}_i'\boldsymbol{\lambda} < 0)\}\\ &+ (1-\tau)w_i(\tau)\ \#\{\tilde{y}_i \in K_{\tau_l} : (r_i(\boldsymbol{\beta}) < 0, \mathbf{x}_i'\boldsymbol{\lambda} > 0)\}\\ &+ \tau\ (1 - w_i(\tau))\ \ \#\{\tilde{y}_i \in K_{\tau_l} : (\mathbf{x}_i'\boldsymbol{\lambda} < 0)\}\Big).\end{aligned}$$

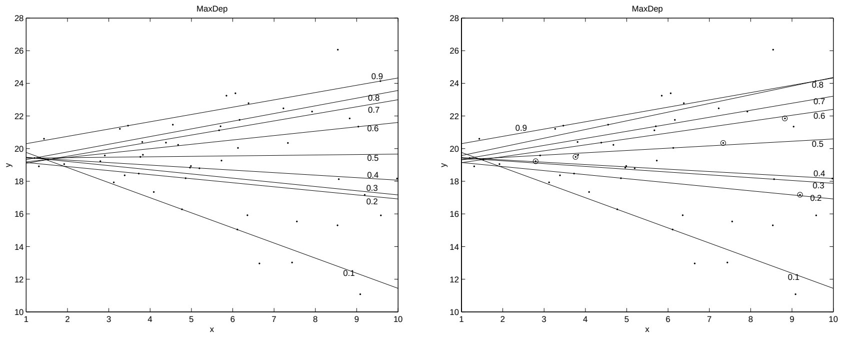
The maximization is performed on a random grid of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ vectors, similar to the algorithms described in [1]. Consider the set $K_{\tau_{l+1}} = \{(\mathbf{x}_i, c_i) : c_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\tau_{l+1}} \leq 0\}$.

IF $K_{\tau_{l+1}} = K_{\tau_l}$,

the current estimate $\hat{\boldsymbol{\beta}}_{\tau_l}$ was found using the correct weights and is therefore a correct solution.

IF $K_{\tau_{l+1}} \neq K_{\tau_l}$,

then the weights should be changed. Observations in $K_{\tau_l} \backslash K_{\tau_{l+1}}$ are censored observations that were crossed but are not anymore. These receive weight 1 again. Observations from $K_{\tau_{l+1}} \backslash K_{\tau_l}$ are censored observations that are



(a) (b)

Figure 2: Regression depth quantiles for a simulated example of 50 data points (a) without censoring, (b) with 5 censored (= circled) data points.

crossed just now, during the transition from τ_l to τ_{l+1}. We define the number

$$\hat{\tau}_i = \tau_l \tag{3}$$

for each of these observations. Their weight is then

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i} = \frac{\tau - \tau_l}{1 - \tau_l}.$$

The remaining weight $1 - w_i(\tau)$ is assigned to a pseudo-observation arbitrarily far away. Thus we find a new set of crossed censored observations. The regression quantile $\hat{\boldsymbol{\beta}}_{\tau_{l+1}}$ is then recomputed with this new set of weights. We repeat this step until we find an estimate for which the weights remain the same.

STEP 4 The algorithm stops when we have dealt with the last grid point τ_m.

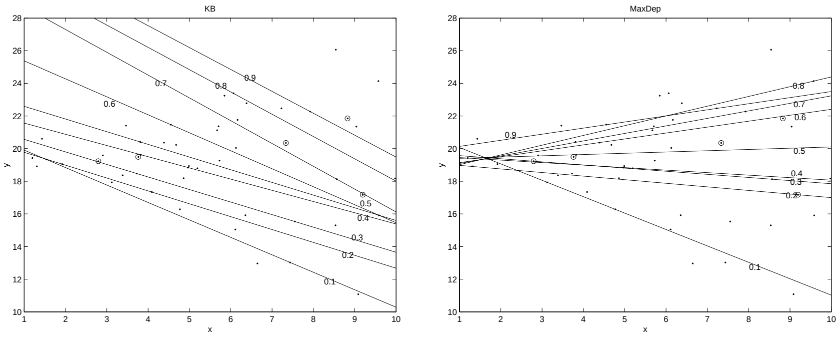
□

Remark that it is possible that no convergence is obtained in step 3. In the examples and simulations we studied so far, this rarely occurred. Hence it does not seem too much of a problem. If it happens, we just skip the grid point τ_{l+1} and continue with the next one τ_{l+2}.

Although the algorithm performs very well in our examples and simulations, its computational complexity is very high since we have to solve a nested optimization problem in each grid point. Therefore in our presentation we will discuss some possible improvements using updating methods, which seem to speed up the computation time.

5 Examples

Let us look at an example in case of simple regression. Data were simulated from a heteroscedastic linear regression model. More precisely data were



(a) (b)

Figure 3: (a) Koenker-Bassett quantiles, (b) Regression depth quantiles for the simulated example with one leverage point.

generated according to the model

$$Y = 20 + 0.5xN(0,1)$$

with the x uniformly distributed in $[1, 10]$. Figure 2(a) shows the resulting deciles for 50 uncensored data points. Then we have randomly chosen 5 data points to be censored with $c_i = y_i$ and obtained Figure 2(b). We see that the regression quantiles are slightly higher than in Figure 2(a) (this is especially clear for the 0.5-quantile), just as one expects.

Figure 3 shows the results when adding one leverage point at $(-50, 100)$. The Koenker-Bassett quantiles in Figure 3(a) are heavily affected, but the depth quantiles in Figure 3(b) only change slightly. This demonstrates the robustness of the latter.

A more realistic example is shown in Figure 4. 50 datapoints were generated according to the model

$$Y = 20 + (0.5N(0,1) + 2)x$$

with the x uniformly distributed in $[1, 10]$. Three outliers were added with coordinates around $(1, 35)$. Note that the x- nor the Y-value of these outliers is outlying. Yet, since these points do not follow the linear model, they have a bad impact on the Koenker-Bassett regression quantiles. The 0.7-quantile is already slightly biased and the 0.8-quantile is very badly estimated. The depth quantiles perform better. The 0.7- and 0.8-quantile are well estimated and even the 0.9-quantile is not completely tilted towards the outliers.

In our presentation we will further show the efficiency and the robustness of the algorithm by several simulation studies in varying dimensions p.

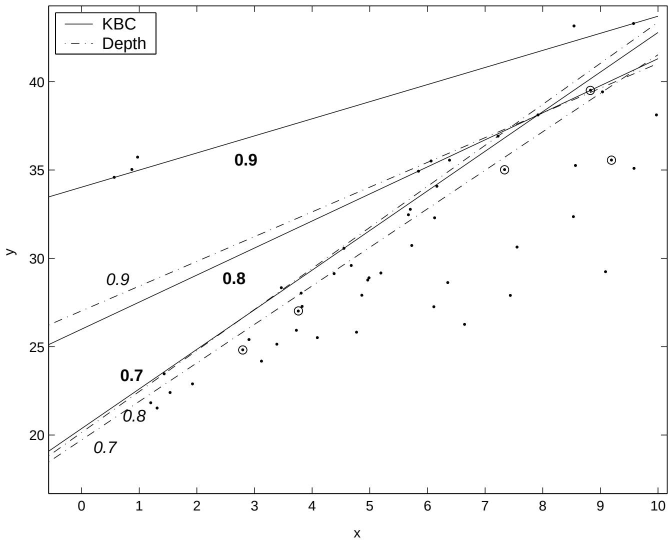


Figure 4: Comparison between Koenker-Bassett quantiles (solid lines) and depth quantiles (dashed-dotted lines) for a simulated example with three outliers.

References

[1] Adrover J., Maronna R.A., Yohai V.J. (2004). *Robust regression quantiles.* Journal of Statistical Planning and Inference **122**, 187 – 202.

[2] Bai Z.D., He X. (2000). *Asymptotic distributions of the maximal depth estimators for regression and multivariate location.* The Annals of Statistics **27**, 1616 – 1637.

[3] Koenker R., Bassett G.J. (1978). *Regression quantiles.* Econometrica **46**, 33 – 50.

[4] Portnoy S. (2003). *Censored regression quantiles.* Journal of the American Statistical Association **98**, 1001 – 1012.

[5] Rousseeuw P.J., Hubert M. (1999). *Regression depth.* Journal of the American Statistical Association **94**, 388 – 402.

Address: M. Debruyne, M. Hubert, K.U.Leuven, Department of Mathematics, W. De Croylaan 54, B-3001 Leuven, Belgium

E-mail: {`michiel.debruyne,mia.hubert`}`@wis.kuleuven.ac.be`

A MULTIVARIATE MODELLING METHOD FOR STATISTICAL MATCHING

Christian Derquenne

Key words: Statistical grafting, partial maximum likelihood, partial least squares, multivariate responses.

COMPSTAT 2004 section: Data imputation, Partial least squares.

Abstract: A better knowledge of our customers is one of the main major aim of Electricité de France, to propose services and products adapted to their needs. A feasible solution is to have a lot of information on each customer with the aid of internal database, because some of them are very interesting to increase the efficiency of operational marketing campaigns. Unfortunately, there is a high amount of missing data in these internal variables. On the other hand several external information are not available in this database, but in survey customers or census of French population. Then, the motivation is to predict these variables. Usually, the internal response variables are directly predicted with internal explanatory variables, whereas the external response variables are predicted by common explanatory variables between internal and external databases. The methods of statistical merging are used. We have chosen logit models adapted to the type of categorical response variable.

1 Context – problem – proposed method

There are three approaches to predict a set response variables. The first one named *Univariate* consists in building a model for each variable, but this approach does not take into account the relation (correlation) between the response variables. The second one, called *Sequential*, has been developed in frame of statistical merging [4]. The first step of this approach builds one model by response variable, as the previous method, then a first response variable is selected with the highest well-classified rate, then it is used as an explanatory variable, and so on. The last approach named *Multivariate*, allows to predict a set of response variables in one time. In this case, the Partial Least Squares Regression [8], [10] with several response variables (PLS2) can be used because it takes into account the correlation between variables. But this method has been developed to numeric response variables. However, this method has been extended to Generalized Linear Models for one response variable (PLS1), more especially for logistic regression [1], [9], [11]. In the other hand, another extension has been proposed of PLS [7] and is based on the maximum likelihood estimator.

But these two sets of methods only allow to predict one response variable. Then we have proposed, in frame of statistical merging [5] to transform categorical variables in function of their types. This method is named

Logit PLS2. For instance, let's be $Y = (Y_1, \ldots, Y_s, \ldots, Y_q)$, q ordinal response variables having respectively $R_1, \ldots, R_s, \ldots, R_q$ responses, and let's be $X = (X_1, \ldots, X_j, \ldots, X_p)$, p categorical explanatory variables, with m_j categories respectively. Then, the crossing of these variables give G categories named $v_1, \ldots, v_g, \ldots, v_G$. In this case, the transformation is the cumulative logit for each Y.

$$\tilde{y}_{s_g}^{(r_s)} = \log\left(\sum_{k=1}^{r_s} n_{s_g}^{(k)} / \left(n_g - \sum_{k=1}^{r_s} n_{s_g}^{(k)}\right)\right) \tag{1}$$

where n_g and $n_{s_g}^{(k)}$ are respectively the number of individuals in category v_g and the number of individuals who have given the answers (k) in category v_g. There are $R_s - 1$, $\tilde{y}_{s_g}^{(r_s)}$ numeric values for Y_s which correspond to the new response variables in PLS2 regression, where $(X_1, \ldots, X_j, \ldots, X_p)$ are used to categorical explanatory variables, as an ANOVA model.

Usually, this method provides enough good results in term of preservation of correlation between response variables, but poor in term of reconstitution of marginal distribution and percentage of well-classified rate. In consequence, this approach is only a first mean to resolve the multivariate case, but it is limited by the complexity of number of crossing of explanatory variables too. Then a new multivariate approach based on Partial Maximum Likelihood (PML) is proposed in this paper, to parallel this one with PLS. In opposite, PML is not really comparable with the approach introduced by Cox [2], [3]. Indeed, in our case "partial", as in PLS, comes from estimation process of the response variables which is based on p regressions (one by explanatory variable) and on the reference variable to estimate latent response variable.

2 A multivariate modelling approach based on the partial maximum likelihood

In the classical PLS2 regression, the first step S_0 is to begin with $X_0 = X$ and $Y_0 = Y$, and the following steps S_1,...,S_h,...,S_H, allow to build PLS components $(h = 1, 2, \ldots, H)$. In our approach, the step S_1 consists in the search of the first component $(h = 1)$ based on the Partial Maximum Likelihood Estimator, whereas the other components are built with the PLS2 regression on the residuals on X and Y. The different sub-steps are described below.

($\mathbf{S}_{1.1}$) A response variable is chosen as reference variable (for instance, Y_1), and corresponds to the first component PML.

($\mathbf{S}_{1.2}$) p ordinal logistic regressions (one by explanatory variable) are made, such as, $\forall j = 1, \ldots, p$:

$$\Pr(Y_1 \le r_1 / x_j = l) = \exp(\alpha_{r_1} + \beta_{jl} 1_{x_j = l}) / (1 + \exp(\alpha_{r_1} + \beta_{jl} 1_{x_j = l})) \tag{2}$$

(S$_{1.2.1}$**)** The coefficients β_{jl}'s are standardized to give a weights vector $\hat{w}_{(1)} = (\hat{w}_{jl(1)}; j = 1 \text{ to } p, l = 1 \text{ to } m_j)$ and $(\hat{w}_{jm_j(1)} = 0$, for $j = 1$ to $p)$:

$$\hat{w}_{jl(1)} = \hat{\beta}_{jl} / \sqrt{\sum_{j=1}^{p} \sum_{l=1}^{m_j} \hat{\beta}_{jl}^2} \tag{3}$$

(S$_{1.2.2}$**)** We calculate the first component PML $t_{(1)}$ with the X_j's such as:

$$t_{(1)} = \sum_{j=1}^{p} \sum_{l=1}^{m_j} \hat{w}_{jl(1)} 1_{x_j = l} \tag{4}$$

(S$_{1.2.3}$**)** We apply q logistic nominal regressions of $t_{(1)}$ on $(Y_1, \ldots, Y_q)$, such as, $\forall s = 1, \ldots, q$:

$$\Pr(Y_s = r_s / t_{(1)}) = \exp(\theta_{r_s} + \gamma_{r_s} t_{(1)}) / \left(1 + \sum_{r_s=1}^{R_s - 1} \exp(\theta_{r_s} + \gamma_{r_s} t_{(1)}) \right) \tag{5}$$

(S$_{1.2.4}$**)** The coefficients γ_{r_s}'s are standardized to give weights vector $\hat{c}_{(1)}$:

$$\hat{c}_{r_s(1)} = \hat{\gamma}_{r_s} / \sqrt{\sum_{s=1}^{q} \sum_{k=1}^{R_s} \hat{\gamma}_k^2} \tag{6}$$

where $\hat{c}_{(1)} = (\hat{c}_{r_s(1)}\ ;\ s = 1$ to q, $r_s = 1$ to $R_s\ /\ \hat{c}_{R_s(1)} = 0$, $s = 1$ to $q)$.

(S$_{1.2.5}$**)** The first component PML $u_{(1)}$ on the Y_q's is calculated such as:

$$\tilde{u}_{(1)} = \sum_{s=1}^{q} \sum_{k=1}^{R_s} \hat{c}_{r_s(1)} 1_{Y_s = r_s} \tag{7}$$

These calculus concern only the first iteration of $w_{(1)}$, then the following step S$_{1.3}$ splitted in six sub-steps provides the convergence of $w_{(1)}$:

(S$_{1.3.1}$**)** p standard linear regressions of $\tilde{u}_{(1)}$ on $(X_1, \ldots, X_p)$ are applied.

$$\tilde{u}_{(1)} = \beta_{jl} 1_{x_j = l} + \epsilon \tag{8}$$

where classically $\hat{\beta}_{jl} = \sum_{i=1}^{n} \tilde{u}_{i(1)} 1_{x_{ji} = l} / \sum_{i=1}^{n} 1_{x_{ji} = l}$ is the LS estimator.

(S$_{1.3.2}$**)** A new standardized weights vector $\hat{w}_{(1)}$ is calculated as in (3).

(S$_{1.3.3}$**)** A new first PML component is built $t_{(1)}$ on $(X_1, \ldots, X_p)$ as in (4).

(S$_{1.3.4}$**)** q logistic regressions of $t_{(1)}$ on $(Y_1, \ldots, Y_q)$ are applied as in (5).

($\mathbf{S}_{1.3.5}$) A new standardized coefficients vector $\hat{c}_{(1)}$ is calculated as in (6).
($\mathbf{S}_{1.3.6}$) The first PML component $\tilde{u}_{(1)}$ with $(Y_1, \ldots, Y_q)$ is built as in (7).
The step $\mathrm{S}_{1.3}$ is based on the Iterated Power Method [6].
The following step $\mathrm{S}_{1.4}$ consists in calculating the residuals on X and Y.
($\mathbf{S}_{1.4.1}$) p logistic regressions of $t_{(1)}$ on $(X_1, \ldots, X_p)$ are applied, $\forall j = 1, p$

$$\Pr(X_j = l/t_{(1)}) = \exp(\tau_{jl} + \xi_{jl} t_{(1)}) / \left(1 + \sum_{l=1}^{m_j - 1} \exp(\tau_{jl} + \xi_{jl} t_{(1)})\right) \qquad (9)$$

($\mathbf{S}_{1.4.2}$) We calculate the residuals on $(X_1, \ldots, X_p)$, such as: $f_{jl(1)} = f_{jl} - \hat{p}_{jl}$, where f_{jl} is the frequency of category l of the variable X_j in the sample and $\hat{p}_{jl}$ is the estimated probability, then $f_{jl(1)}$ is the associated residual. This residual corresponds to the first step $h = 1$.

($\mathbf{S}_{1.4.3}$) The residuals on $(Y_1, \ldots, Y_q)$, in category ν_g, are calculated in $\mathrm{S}_{1.3.5}$: $f^{(g)}_{r_s(1)} = f^{(g)}_{r_s} - \hat{q}^{(g)}_{r_s}$ where $f^{(g)}_{r_s}$ is the frequency of answer r_s of the variable Y_s in the sample and $\hat{q}^{(g)}_{r_s}$ is the estimated probability, then $f^{(g)}_{r_s(1)}$ is the associated residual, for the first step $h = 1$.

($\mathbf{S}_2,\ldots,\mathbf{S}_H$) For the estimation of the following PML components ($h = 2, \ldots, H$), we use a classical PLS2 regression of residuals scores of Y on discretized residuals of X (only the best explanatory variables, see details at the end of this section). Indeed, it had been more naturally to use directly the residuals of Y as a new response variables, but the range of these residuals is [-1;1], as the range of a probability is [0;1]. And as the classical PLS2 provides a numeric scale for estimated responses in |R, this constraint ([-1;1]), so has been not filled. Whereas if we use the residual score as response variable, we get round this problem. Then the new form (residuals scores) of a response variable Y_q in a category ν_g and for $h = 1$, is the following:

$$z^{(g)}_{r_s} = \log(f^{(g)}_{r_s} / f^{(g)}_{R_s}) - \log(\hat{q}^{(g)}_{r_s} / \hat{q}^{(g)}_{R_s}) \qquad (10)$$

We can remark that these scores correspond to the odds ratio in case of Multinomial Logit Model, where $f^{(g)}_{R_s}$ and $\hat{q}^{(g)}_{R_s}$ are respectively the observed reference frequency and the estimated reference probability. On the other hand, we have chosen to discretize the residuals of X (explanatory variables). Firstly, for a category l of a variable X_j, the value of residual $f_{jl(1)}$ given in ($\mathrm{S}_{1.4.2}$) varies in [-1 ; 1], as the residual of Y. Secondly, the value of $z^{(jl)}_{r_s}$ varies on |R (numeric scale), but if the category l has a discriminant power, the values of $z^{(jl)}_{r_s}$ are divided in three groups ($z^{(jl)}_{r_s} < -5$; $-5 \leq z^{(jl)}_{r_s} \leq 5$; $z^{(jl)}_{r_s} > 5$). Then, our discretization consists in six categories constituting the discretized residuals of X. This discretization allows to get around non linear problem. For instance, for the category l of X_j and the response r_s of Y_s, the new variable has the six following categories:

$$c^{(jl)}_{r_s} = (c^{jl(<0)}_{rs(<-5)}, c^{jl(<0)}_{rs(-5;5)}, c^{jl(<0)}_{rs(>5)}, c^{jl(>0)}_{rs(<-5)}, c^{jl(>0)}_{rs(-5;5)}, c^{jl(>0)}_{rs(>5)}) \qquad (11)$$

Consequently, by variable X_j, there are $m_j \sum_{s=1}^{q}(R_s - 1)$ new discretized variables having each six categories. However, we must keep in the mind that the number of new variables is $\sum_{j=1}^{p} m_j \sum_{s=1}^{q}(R_s - 1)$ and then the number of categories is $6(\sum_{j=1}^{p} m_j \sum_{s=1}^{q}(R_s - 1))$. As this number can become very high, then we choose to keep only the best explanatory variable in $(X_1, \ldots, X_p)$.

The "best explanatory variable" is selected by the following way : a PLS2 Regression is made on each explanatory variable $X_1, \ldots, X_p$ and we choose the high percentage of well-classified rate of predicted $\hat{Y}$ with respect of Y. The choice of an only explanatory variable is motivated by the fact that an additional explanatory variable is useless, because this does not improved the results of the modelling. Indeed, the space of data is sufficiently splitted to assure a reasonable quality of prediction.

Lastly, the PLS2 regression model has $\sum_{s=1}^{q}(R_s - 1)$ response variables (residuals scores of $Y_1, \ldots, Y_q$) and $m_j \sum_{s=1}^{q}(R_s - 1)$ explanatory variables with each six categories (as in (11)), if X_j is the best explanatory variable (for further information on PLS2 regression, see [8], [10]).

3 Validation of the model

The validation of the model is an essential step of modelling. This step contains not only the classical statistical tests on overall model and on each explanatory variable, but also the reconstitution of data with the model. In our case, the model based on Partial Maximum Likelihood ($h = 1$) and on PLS2 regression ($h = 2, \ldots, H$) is validated with cross-validation. The set of explanatory variables for the Y's response variables used in PML comes from initial modelling on each response variable. Indeed, this set of explanatory variables corresponds to the union of significant variables on all the Y's response variables. Then, as specified in part 2, we selected the "best explanatory variable" to apply the PLS2 regression. On the other hand, the model is also judged on three steps of validation in term of preservation of correlation between response variables and in term of reconstitution of marginal distribution, in addition of percentage of well-classified rate. However, we have constructed some statistical tests associated to each of the three steps of validation.

For the marginal distribution, we use the classical test of chi-square: This test consists in comparison between the estimated marginal distribution with the model $\hat{F}_{r_s}$ of $\hat{Y_s}$ and the observed marginal distribution on the data F_{r_s} of Y_s. We use the chi-square statistic to test the likeness of both distributions.

For the well-classified rate, we use a test of comparison of two proportions: This test consists in comparison between the proportion of well-classified rate observed and the proportion of well-classified rate at random. Indeed, r_{ran} corresponds to the maximum rule, that is to say the maximum percentage obtained on the R_s categories of Y_s response variable. In other words, it is the percentage obtained if no model was applied. Normally, if we use a "good" statistical model, the percentage of well-classified rate r_{obs} must be greater than r_{ran} statistically.

For the preservation of correlation between response variables: This test consists in comparison between the estimated crossed distributions of two responses variables $\hat{Y_s}$ and $\hat{Y_k}$ with the model $\hat{F}_{r_s/r_k}$ and the observed crossed distributions of the same responses variables Y_s and Y_k on the data F_{r_s/r_k}. We use the chi-square statistic to test the likeness of both distributions.

These tests are simply indexes because the associated statistics under the null hypothesis are not optimized by the PML estimator.

4 Application on survey satisfaction on electric heating

In 1998, the Marketing Department of EDF has conducted a survey on electric heating, to better know its customers and top of that to offer products and services suited to their expectations and needs. In frame of this paper, we use some variables coming from questionnaire. But, as specified in the introduction, this survey has served to enrich a big database of EDF'customers with statistical data merging approach.

There are three response variables: satisfaction (ordinal categorical variable: no satisfy ; near satisfy ; very satisfy), choice of future energy (boolean variable : electric heating ; other) and advice to friends or family circle (ordinal categorical variable : no electric heating ; perhaps electric heating ; yes electric heating).

There are seven explanatory variables (all categorical): year of first subscription with EDF, total electric consumption, payment code, payment period, customer's payment quality, tariff and housetype.

There are 7114 customers (individuals) in this survey and each of them has a sample weight. This dataset has been split in two samples: a training sample (80%) and a validation sample (20%). Then, we have applied four different approaches on these data: univariate approach (consists in building a ordinal logit model for each response variable), sequential approach, logit PLS2 approach based on cumulative logit transformation and multivariate approach based on PML + PLS2 on scores residuals of response variables. Lastly, we compare the results of these four approaches with the aid of three steps of validation described in part 3.

All explanatory variables are selected at least one time. Then, we are going to take into account all variables for multivariate approach with PML.

The table 1 gives the results of comparison between estimated marginal distribution with the model and observed marginal distribution on the data. In this case, three p-values, in parenthesis, are greater than 0.05. In addition, even for the significant tests (p-value $<$ 0.05), that only the both last approaches (logit PLS2 and Multivariate) have the values of chi-square statistic clearly under than other two first methods, excepted for the satisfaction.

Table 2 provides the well-classified rates on the validation sample. The

Approaches	**Satisfaction**	**Choice**	**Advice**
Univariate	530.59 (< 0.001)	35.87 (< 0.001)	60.75 (< 0.001)
Sequential	56.24 (0.3720)	45.84 (< 0.001)	428.18 (< 0.001)
Logit PLS2	304.06 (< 0.001)	0.10 (0.7508)	28.43 (< 0.001)
Multivariate	264.02 (< 0.001)	2.14 (0.1434)	31.69 (< 0.001)

Table 1: Results of marginal distribution on validation sample.

Approaches	**Satisfaction**	**Choice**	**Advice**
Max rule	51.54	57.65	39.15
Univariate	52.31 (0.2645)	63.92 (< 0.001)	46.26 (< 0.001)
Sequential	53.53 (0.0529)	64.28 (< 0.001)	40.72 (0.0945)
Logit PLS2	50.36 (0.8308)	66.74 (< 0.001)	45.74 (< 0.001)
Multivariate	50.40 (0.8230)	64.39 (< 0.001)	44.32 (< 0.001)

Table 2: Results of well-classified rates on validation sample.

first row corresponds to the maximum rule and the other rows are related to the different models. The satisfaction has bad well-classified rates for all approaches, because they are close to the percentage of maximum rule (non significant test). The choice of energy has good and similar well-classified rates. The advice has an only bad well-classified rate (sequential approach).

Lastly, the table 3 provides the results concerning the correlation, or more exactly the comparison between the estimated crossed distributions of two responses variables with the model and the observed crossed distributions of same responses variables on the data. Even if all differences are significant (p-value < 0.05), we can see that the values of chi-square statistic associated to our multivariate approach are very clearly under than the other three methods, except for a particular case : satisfaction crossed with choice for the sequential spproach.

Approaches	**Satis.×choice**	**Satis.×advice**	**Choice×advice**
Univariate	576.36 (< 0.001)	632.76 (< 0.000)	98.60 (< 0.001)
Sequential	183.44 (< 0.001)	467.29 (< 0.001)	693.26 (< 0.001)
Logit PLS2	325.95 (< 0.001)	352.52 (< 0.001)	51.10 (< 0.001)
Mulitvariate	282.81 (< 0.001)	312.43 (< 0.001)	47.82 (< 0.001)

Table 3: Results of correlation between response variables on valid sample.

5 Concluding remarks and research works

The results of this new method “Multivariate Approach” based on PML are very encouraging because it provides substantial better results in terms of preservation of correlation, with respect to the three other approaches. This property avoids a problem of bad coherence between the estimated responses with respect to the observed responses. This method has relatively good results in term of reconstitution of marginal distributions, and these ones are quite similar of logit PLS2 Approach. On the other hand the well-classified rates of the new approach remain comparable to the first three methods which offered good results. Then, there are two advantages of this new approach. First, the complexity of crossing variables has no influence, whereas it penalizes the logit PLS2 Approach. Secondly, the first PML component is based on maximum likelihood estimator, it allows to better take into account the type of response variables. Now, we work to improve this algorithm, notably in frame of missing data, on the properties of estimators and above all on the validity of estimated models.

References

[1] Bastien Ph., Espozito Vinzi V., Tenenhaus M. (2002), *Régression linéaire généralisée PLS*, Les Cahiers de la Recherche HEC, **766**, Groupe HEC.

[2] Cox D.R. (1972), *Regression Models and Life-Tables (with discussion)*, Journal of the Royal Statistical Society, Series B, **34**, 187 – 220.

[3] Cox D.R. (1975), *Partial Likelihood*, Biometrika, **62**, 2, 269 – 276.

[4] Derquenne Ch. (1999), *A Method of Generating a Sample of Artificial Data from Several Existing Data Tables*: *Application Based on the Residential Electric Power Market*, Proceeding of Statistics Canada Symposium 99, Combining Data from Different Sources.

[5] Fischer N., Derquenne Ch., Saporta G. (2001), *A method to match data set applied to electric market*, ETK-NTTS 2001, Creta.

[6] Hotteling H., (1936), *Simplified Calculation of Principal Components*, Psychometrika, vol 1., 27 – 35.

[7] Marx B.D. (1996), *Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression*, Technometrics, vol. **38**, 4, 374 – 381.

[8] Tenenhaus M. (1998), *La Régression PLS*: *Théorie et pratique*, Ed. Technip, Paris.

[9] Tenenhaus M. (2000), *La Régression Logistique PLS*, Journées d'Etudes en Statistique, Modèles Statistiques pour Données Qualitatives, 261 – 273.

[10] Wold S., Martens H., Wold H. (1983), *The Multivariate Calibration Problem in Chemistry Solved by PLS Method*, In Proc.Conf. Matrix Pencils, Ruhe A., Kågstrøm B. (Eds), March 1982, Lecture notes in mathematics, Springer Verlag, Heidelberg, 286 – 293.

[11] Wold S., Ruhe A., Wold H., Dunn III W.J (1984), *The Collinearity Problem in Linear Regression: The Partial Least Squares (PLS) Approach to Generalized Inverses*, SIAM J. Sci. Stat. Comput., vol. **5**, 3,. 735 – 743.

Address: C. Derquenne, EDF R&D - 1, av. du Général de Gaulle,
92141 Clamart, France

E-mail: `christian.derquenne@edf.fr`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

PERFORMANCE OF CONTROL CHARTS FOR SPECIFIC ALTERNATIVE HYPOTHESES

Alessandro Di Bucchianico, Marie Hušková,
Petr Klášterecký and William R. van Zwet

Key words: Control charts, alternative hypothesis, statistical process control, sequential tests, likelihood ratio tests, isotonic regression.

COMPSTAT 2004 section: Statistical process control, Simulations.

Abstract: We present statistical models in terms of hypothesis testing for practical out-of-control situations in SPC that extend the traditional mean shift or linear trend situations. Based on these explicit alternative hypotheses, we derive likelihood ratio tests. Simulations are used to obtain critical values and to study the performance (in terms of both mean and standard deviation of detection delays) of our procedures. We compare our control charts with a control chart proposed by Chang and Fricker. It turns out that smaller mean delays are not always preferable.

1 Introduction

Change detection lies at the heart of SPC (statistical process control). Nowadays, various procedures exist for change detection. As noted already by Shewhart, the use of control charts is connected to a specific way of hypothesis testing (either "normal" or sequential). Of course, a proper use of control charts entails much more than just hypothesis testing (see e.g. the interesting discussions in [2], [6], [7], [8]). It is standard practice in hypothesis testing to describe both null and alternative hypothesis. However, in the SPC literature alternative hypotheses are hardly mentioned nor used when developing control chart procedures. This is surprising, since a proper use of statistical process control techniques requires active involvement of process engineers. Hence, one often has detailed knowledge about departures from an in-control situation. Of course, the original goal of Shewhart was to provide simple monitoring tools to be used on the work floor. However, with the current state of computer technology, there is no excuse not to use advanced statistical techniques, if the outcomes can be presented in a way that is easy to interpret.

The goal of this paper is to derive and study likelihood ratio tests for specified alternative hypotheses. In particular, we develop on-line control charts based on likelihood ratios for several specific alternative hypotheses that are of practical value. Section 2 discusses some specific alternative hypotheses, which are additions to the common change of mean out-of-control situation.

Although our control charts are for on-line use (i.e., in a sequential setup), we present a hypothesis testing framework in Section 3. We adapt the notions of critical value and power to accommodate the use in a sequential way. Critical values of our procedures are reported in Section 4. Section 5 contains results and discussions on simulations of the performance (in terms of both mean and standard deviation of detection delays) of our procedures. It turns out that smaller mean delays are not always preferable.

2 Alternative hypotheses

The general model that we have in mind is the following. We are sequentially observing $X_1, \ldots, X_n$ $(n \leq N)$. The observations X_i are independent and

$$X_i = \mu_i + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where $e_1, \ldots, e_N$ are i.i.d. error terms with density function f symmetric around 0, and $\mu_1, \ldots, \mu_n$ are unknown parameters. As null hypothesis we consider the case when all μ_i's are equal. The most widely used out-of-control situation studied in the SPC literature is the situation of a sudden persistent change of the mean:

Persistent change of mean

$$\begin{aligned} &\mathrm{H}_0 : \mu_1 = \ldots \mu_n \\ &\mathrm{H}_1 : \begin{cases} \mu_i = \mu_0, & i = 1, \ldots, m, \\ \mu_i = \mu_0 + \delta, & i = m+1, \ldots, n, \end{cases} \end{aligned} \tag{2}$$

where μ_0 and δ are known.

In case of one-sided tolerances on the process characteristic (e.g., viscosity), one may not be interested in a target value for the process mean. Instead, one wishes that the process mean stays below or above a certain value. In such cases the following alternative hypothesis may be more appropriate than the persistent change of mean described above.

Persistent non-monotone threshold crossing

$$\begin{aligned} &\mathrm{H}_0 : \mu_i \leq \delta, \quad i = 1, \ldots, n, \\ &\mathrm{H}_1 : \begin{cases} \mu_i \leq \delta, & i = 1, \ldots, m, \\ \mu_i > \delta, & i = m+1, \ldots, n, \end{cases} \end{aligned} \tag{3}$$

where $\delta > 0$ is known.

A monotone version of this situation has been studied in [1]. This situation is a good model for tool wear.

Persistent monotone threshold crossing

$$\begin{aligned} &H_0 : \mu_1 \leq \ldots \leq \mu_n \leq \delta \\ &H_1 : \begin{cases} \mu_1 \leq \ldots \leq \mu_m < \delta & i = 1, \ldots, m \\ \delta < \mu_{m+1} \leq \mu_{m+2} \leq \ldots \leq \mu_n & i = m+1, \ldots, n \end{cases} \end{aligned} \tag{4}$$

where δ is known.

Of course, there are many other sensible alternatives like a temporary change of the mean (also called an epidemic alternative), which is useful in situations where a feedback controller is active. This controller tries to compensate changes in the process, giving rise to a temporary change of the process parameters (see the literature on integration of SPC and APC). Furthermore, hypotheses concerning the variance are definitely meaningful, although they hardly appear in the SPC literature. A nice list of practical out-of-control situations in terms of patterns on means and variances can be found in [4, Chapter 8] and [5, Chapter 6].

3 Procedures

Most of the standard procedures mentioned in Section 1 are related to the log likelihood ratio. Here we work out the alternative hypotheses "Persistent non-monotone threshold crossing" and "Persistent monotone threshold crossing" for normally distributed data with known variance.

The likelihood ratio statistic for the persistent non-monotone threshold crossing for normally distributed data with known variance is given by

$$Q_n = \max_{m<n} \frac{1}{2\sigma^2} \left\{ \sum_{i=m+1}^{n} (X_i - \delta)^2 \operatorname{sign}(X_i - \delta) \right\}, \quad n = 2, \ldots, N. \tag{5}$$

In order to obtain distributional results on this statistic, it is necessary to standardize. An obvious choice would be to standardize with the square root of N, the maximal number of observations. However, this would lead to poor performance in detection of early change points. Therefore next to Q_{nN}, we also study Q_{nn} in order to investigate the impact of standardizing with the square root of the actual number of observations.

$$\begin{aligned} Q_{nN} &= \frac{1}{\sqrt{N}} Q_n = \frac{1}{\sqrt{N}} \max_{0\le m<n} \frac{1}{2\sigma^2} \sum_{i=m+1}^{n} (X_i - \delta)^2 \operatorname{sign}(X_i - \delta) \qquad (6)\\ Q_{nn} &= \frac{1}{\sqrt{n}} Q_n = \frac{1}{\sqrt{n}} \max_{0\le m<n} \frac{1}{2\sigma^2} \sum_{i=m+1}^{n} (X_i - \delta)^2 \operatorname{sign}(X_i - \delta). \qquad (7) \end{aligned}$$

These choices yield expressions that need to be maximized once or twice with respect to the parameter m. Since this may be time consuming when performed on-line and may also make it difficult to obtain the asymptotic distribution of the test statistic, we will also consider so-called windowed likelihood ratio statistics where we do not maximize over all $m = 1, \ldots, n-1$, but only over $m = \max(1, n-G), \ldots, n-1$. The idea is that if G is reasonably large (moderate), then we reveal the change with delay smaller than G.

$$Q_{n,G} = \frac{1}{\sqrt{G}} \max_{n-G \leq m < n} \frac{1}{2\sigma^2} \sum_{i=m+1}^{n} (X_i - \delta)^2 \operatorname{sign}(X_i - \delta), \tag{8}$$

$$Q_{n,G}^{\text{simp}} = \frac{1}{\sqrt{G}} \frac{1}{2\sigma^2} \sum_{i=n-G+1}^{n} (X_i - \delta)^2 \operatorname{sign}(X_i - \delta). \tag{9}$$

In [1] a likelihood ratio procedure is proposed for the "Persistent monotone threshold crossing" alternative. It is based on isotonic regression

$$\begin{aligned} M_n &= \sum_{i=1}^{n} (X_i - Z_i)^2 - \sum_{i=1}^{n} (X_i - Y_i)^2 \\ &= \begin{cases} 0 & \text{for } Y_n \leq \delta \\ \sum\limits_{i=J}^{n} (X_i - \delta)^2 - \sum\limits_{i=J}^{n} (X_i - Y_i)^2 & \text{for } Y_n > \delta, \end{cases} \end{aligned} \tag{10}$$

where $Z_1, \ldots, Z_n$ denotes an isotonic regression restricted to increase up to at most δ, $Y_1, \ldots, Y_n$ is the corresponding unrestricted isotonic regression, and $J = \min\{i : Y_i > \delta\}$.

4 Critical values

In this section we give tables of critical values for our procedures based on simulations. We reject the hypothesis of no change if for some $n \leq N$ a critical value $c_{N,\alpha}$ is violated. Since we have a sequential procedure in a finite time frame, we must decide on a sensible way to control false alarms. The most common way to control false alarm in SPC is to use run lengths, in particular the in-control ARL (average run length). This is the expected value of the stopping time of the procedure under the null hypothesis. However, run length distributions are usually skewed. Hence, it is dangerous to judge the performance by considering ARL's (see [8] for some interesting views on a proper use of ARL's). Sometimes also the SRL (standard deviation of the run length) is taken into account, or even better one uses quantiles (see *e*.g., [3]).

The decision rule here: we reject the null hypothesis as soon as for some $n \leq N$

$$T_n > c_{N,\alpha} \tag{11}$$

where T_n denotes any of the statistics presented in Section 3 and $c_{N,\alpha}$ is chosen in such a way the significance level is α, i.e.

$$P_{H_0}\left(\max_{1 \leq n \leq N} T_n > c_{N,\alpha} \right) = \alpha.$$

Since the distribution of our statistics do not admit closed-form expression, these critical values have to be obtained by simulations. Work on asymptotic results on these distributions is in progress and will be published elsewhere.

We used the software R (see `www.r-project.org`), version 1.8.1 on a Unix platform. Its new built-in procedure `isoreg()` is much faster in computing the isotonic regression (needed for simulations of the M_n statistics) than the PAVA algorithm used in [1]. The source codes of all procedures we used are available at `http://www.karlin.mff.cuni.cz/~klaster/compstat04`.

α	N		
	10	100	1000
0.01	7.387	23.241	75.612
0.05	5.254	17.995	59.724
0.10	4.225	15.666	51.722

α	N		
	10	100	1000
0.01	12.235	20.260	28.045
0.05	8.209	16.147	23.581
0.10	6.862	14.387	21.809

Table 1: Simulated critical values $c^{\text{sim}}_{N,\alpha}$ for $\{Q_n\}_{n=1}^{N}$ (left-hand side) and $\{M_n\}_{n=1}^{N}$ (right-hand side).

α	N		
	10	100	1000
0.01	2.336	2.324	2.391
0.05	1.661	1.800	1.889
0.10	1.340	1.567	1.636

α	N		
	10	100	1000
0.01	3.260	3.416	3.544
0.05	2.221	2.560	2.715
0.10	1.776	2.202	2.411

Table 2: Simulated critical values $c^{\text{sim}}_{N,N,\alpha}$ of standardized version $\{Q_{nN}\}_{n=1}^{N}$ (left-hand side) and $c^{\text{sim}}_{N,n\,\alpha}$ of standardized version $\{Q_{nn}\}_{n=1}^{N}$ (right-hand side).

α	$G = 0{,}2N$, N:			α	$G = 0{,}15N$, N:		
	10	100	1000		10	100	1000
0.01	3.853	3.274	3.090	0.01	-	3.422	3.270
0.05	2.731	2.642	2.663	0.05	-	2.769	2.786
0.10	2.181	2.351	2.433	0.10	-	2.473	2.552
α	$G = 0{,}1N$, N:			α	$G = 0{,}05N$, N:		
	10	100	1000		10	100	1000
0.01	4.839	3.672	3.360	0.01	-	4.320	3.623
0.05	3.281	2.988	2.900	0.05	-	3.456	3.115
0.10	2.688	2.680	2.667	0.10	-	3.055	2.882

Table 3: Simulated critical values $c^{\text{sim}}_{N,G,\alpha}$ of $\{Q_{n,G}\}_{n=1}^{N}$.

α	$G = 0{,}2N$, N:			α	$G = 0{,}15N$, N:		
	10	100	1000		10	100	1000
0.01	3.853	3.226	3.038	0.01	-	3.384	3.210
0.05	2.712	2.593	2.585	0.05	-	2.709	2.727
0.10	2.160	2.290	2.346	0.10	-	2.417	2.470

α	$G = 0{,}1N$, N:			α	$G = 0{,}05N$, N:		
	10	100	1000		10	100	1000
0.01	4.839	3.642	3.299	0.01	-	4.295	3.574
0.05	3.281	2.952	2.838	0.05	-	3.426	3.071
0.10	2.688	2.633	2.602	0.10	-	3.032	2.831

Table 4: Simulated critical values $c^{\mathrm{sim}}_{N,G,\alpha}$ of $\{Q^{\mathrm{simp}}_{n,G}\}_{n=1}^{N}$.

5 Simulations

In this section we present simulations on the performance of our procedures. All simulations are for the case $N = 100$, $\alpha = 0.05$, $G = 0.1N = 10$, $\sigma = 1$. We investigate two types of changes, each of which has a "small" and a "large" version, corresponding to a change of mean of σ and 3σ, respectively:

gradual change mean increases on 5 equally spaced intervals before change point from μ to δ and on 5 equally spaced intervals after the change-point from δ to $2\delta - \mu$, where $\delta = 1$ and $\mu = 0.5$ (small change) or $\delta = 3$ and $\mu = 1.5$ (large change).

general change mean is uniformly distributed on $[\delta - r, \delta]$ before change-point, and uniformly distributed on $[\delta, \delta + r]$ after the change-point, where $r = 1$ (small change) or $r = 3$ (large change), respectively (so the expected mean increases by 1 and 3, respectively).

Note that because of the finite time interval, the detection lengths must be interpreted with care. If a change occurred at time point x and the number of observations equals N, then in our simulations the number $N - x$ indicates either a detection at time point N or no detection at all. We report mean and standard deviations below, but kept records of other summary statistics like quantiles as well during our simulations. In our comparisons, we first consider the mean. However, there are often considerable differences in the standard deviations. Therefore, we will see that in certain cases the preferred procedure may not be the one with the smallest mean.

For detection of a small change in both the general and the gradual case, we see that Chang and Fricker's M_n has the smallest mean. However, Q_{nN} has a considerably smaller standard deviation. Closer inspection of summary statistics reveals that M_n has better low quantiles and that Q_{nN} has better median and higher quantiles. Hence, in most cases Q_{nN} performs better, but M_n sometimes detects extremely fast.

	General change				
	Q_{nN}	Q_{nn}	$Q_{n,G}$	$Q_{n,G}^{\text{simp}}$	M_n
m	mean; sd	mean; sd	mean; sd	mean; sd	mean; sd
5	42.34; 15.07	37.08; 23.10	52.12; 31.69	51.54; 31.47	37.99; 20.81
15	42.95; 15.68	44.87; 23.05	48.79; 29.11	48.69; 28.99	38.67; 21.47
25	41.88; 14.38	47.51; 19.86	46.10; 25.10	45.68; 24.86	37.34; 19.10
40	41.35; 12.47	49.39; 13.89	41.80; 19.40	41.46; 19.32	36.45; 16.25
60	35.32; 6.56	38.16; 9.73	31.18; 11.96	31.06; 11.86	30.47; 10.88

	Gradual change				
	Q_{nN}	Q_{nn}	$Q_{n,G}$	$Q_{n,G}^{\text{simp}}$	M_n
m	mean; sd	mean; sd	mean; sd	mean; sd	mean; sd
5	74.96; 16.29	80.98; 20.88	82.96; 20.38	82.96; 20.09	75.19; 20.06
15	69.49; 14.55	76.26; 15.85	75.88; 16.58	75.67; 16.78	69.72; 17.41
25	63.37; 12.53	69.48; 13.36	66.67; 16.24	66.60; 16.13	62.79; 15.68
40	54.29; 9.03	57.96; 9.32	55.21; 11.38	55.24; 11.20	53.45; 11.51
60	38.43; 4.20	39.01; 8.77	37.30; 8.24	37.36; 7.89	37.06; 7.87

Table 5: Delay of detection, small change, $G = 10 = 0.1N$.

	General change				
	Q_{nN}	Q_{nn}	$Q_{n,G}$	$Q_{n,G}^{\text{simp}}$	M_n
m	mean; sd	mean; sd	mean; sd	mean; sd	mean; sd
5	10.45; 3.99	5.35; 3.33	6.22; 3.19	8.03; 2.52	6.32; 3.57
15	10.52; 3.90	7.42; 3.82	6.15; 3.32	8.15; 2.52	6.24; 3.53
25	10.60; 4.35	7.80; 3.22	5.70; 2.31	7.60; 1.71	5.30; 2.71
40	10.71; 3.93	10.64; 5.38	6.34; 3.20	8.18; 2.51	6.47; 3.67
60	10.57; 3.83	12.39; 5.58	6.16; 3.05	8.07; 2.38	6.29; 3.49

	Gradual change				
	Q_{nN}	Q_{nn}	$Q_{n,G}$	$Q_{n,G}^{\text{simp}}$	M_n
m	mean; sd	mean; sd	mean; sd	mean; sd	mean; sd
5	42.34; 8.60	39.47; 13.24	44.50; 15.26	44.07; 15.24	37.78; 10.78
15	39.69; 8.47	40.49; 10.99	39.98; 14.86	39.57; 14.75	34.83; 10.61
25	37.45; 7.38	40.24; 9.25	37.45; 12.88	37.20; 12.63	33.18; 9.51
40	33.29; 7.04	37.63; 7.99	32.06; 11.09	31.90; 10.90	28.77; 8.59
60	27.37; 5.17	32.27; 5.90	24.89; 7.79	24.84; 7.62	22.89; 6.61

Table 6: Delay of detection, large change, $G = 10 = 0.1N$.

For detection of a large change in the general case, we see that Chang and Fricker's M_n and $Q_{n,G}$ are performing best. They have comparable means and standard deviations. Closer inspection of summary statistics reveals that M_n has better quantiles. However, it must be said that in this case all procedures react fast.

For detection of a large change in the gradual case, Chang and Fricker's M_n is performing best as expected. Again $Q_{n,N}$ compensates a somewhat larger mean by a smaller standard deviation. However, closer inspection of the quantiles reveals that M_n is having smaller quantiles (both high and low). Hence, M_n is clearly the best procedure in this case.

References

[1] Chang J.T., Fricker R.D. (1999). *Detecting when a monotically increasing mean has crossed a threshold.* Journal of Quality Technology **31**, 217 – 234.

[2] Crowder S.V., Hawkins D.M., Reynolds M.R., Jr., Yashchin E.(1997). *Process control and statistical inference.* Journal of Quality Technology **29**, 134 – 139.

[3] Does R.J.M.M., Schriever B.F. (1992). *Variables control chart limits and tests for special causes.* Statistica Neerlandica **46**, 229 – 245.

[4] Gitlow H., Oppenheim A., Oppenheim R. (1989). *Quality management: tools and methods for improvement.* 2nd edition, Irwin.

[5] Griffith G.K. (1996). *Statistical process control methods.* ASQC Quality Press, Milwaukee.

[6] Stoumbos Z., Reynolds M.R., Jr., Ryan T.P., Woodall W.H. (2000). *The state of statistical process control as we proceed into the 21st century.* Journal of the American Statistical Association **95**, 992 – 998.

[7] Woodall W.H., Montgomery D.C. (1999). *Research issues and ideas in statistical process control.* Journal of Quality Technology **31**, 376 – 386.

[8] Woodall W.H. (2000). *Controversies and contradictions in statistical process control.* Journal of Quality Technology **32**, 341 – 378.

Acknowledgement: The work of the second and third author is partially supported by grants GAČR 201/03/0945 and MSM 113200008.

Address: A. Di Bucchianico, Eindhoven University of Technology, Eindhoven, Netherlands
M. Hušková, Charles University, Prague, Czech Republic
P. Klášterecký, Charles University, Prague, Czech Republic
V.R. van Zwet, Leiden University, Leiden, Netherlands

E-mail: a.d.bucchianico@tue.nl, huskova@karlin.mff.cuni.cz, klaster@karlin.mff.cuni.cz, vanzwet@math.leidenuniv.nl

DIMENSIONALITY PROBLEM IN TESTING FOR NONCAUSALITY BETWEEN TIME SERIES A PARTIAL SOLUTION

Francesca Di Iorio and Umberto Triacca

COMPSTAT 2004 section: Time series analysis.

Abstract: For Vector Autoregressive models, the problem of dimensionality, associated with an increasing dimension of the model, can affect the power of noncausality tests. In this paper, by a Monte Carlo study, we analyze the impact of high dimensionality on the power of noncausality test and we proposed a testing strategy that, under certain conditions, limit the negative effects of high dimensionality in the causality analysis.

1 Introduction

Vector autoregressive (VAR) models have become a dominant research strategy in econometrics since Sims [6] critique of traditional simultaneous equations econometric models. They are however subjected to so-called curse-of-dimensionality problem.

Consider a n-dimensional time series vector $\mathbf{y_t}$ generated by a vector autoregressive process of order p, denoted VAR(p) model,

$$\mathbf{y}_t = \mathbf{A}_1\mathbf{y}_{t-1} + ... + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t t = 1, ..., T \quad (1)$$

where the $\mathbf{A}_i$ are fixed $(n \times n)$ coefficient matrices and $\mathbf{u}_t$ is a n-dimensional white noise with non singular covariance matrix $\mathbf{\Sigma}$. In this VAR model, without deterministic terms, there are n^2p coefficients. The number of parameters to estimate grows rapidly as the number of variables in the model increases, each additional lag adds n^2 coefficients.

Abadir, Hadri and Tzavalis [1] and Gonzalo and Pitirakis [3] show that an increase in the dimension of a cointegrated VAR model can lead to very undesirable properties for both the usual test statistics and estimators. In particular, we are interesting to the dimensionality problem in testing for noncausality between time series. With sample sizes commonly used in applied modelling, the available degrees of freedom are often small. This could affect the power of noncausality tests. In this paper we propose a way to limit the negative effects of high dimensionality in the causality analysis.

The outline of the paper is as follows. In section 2 we evaluate the relevance of the empirical problem studied in this paper. In section 3 we discuss a possible solution of the dimensionality problem. The final section contains some conclusions.

2 Is there a dimensionality problem?

Let $\{\mathbf{y}_t, \;\; t \in I\}$ be a $n \times 1$ multivariate stationary stochastic process on the integers I, defined on some probability space (Ω, F, P) and write $\mathbf{y}_t = \left(\mathbf{y}^{'}_{1,t}, \mathbf{y}^{'}_{2,t}, \mathbf{y}^{'}_{3,t}\right)^{\prime}$ where $\mathbf{y}_{1,t} = (y^{(1)}_{1,t}, ..., y^{(1)}_{n_1,t})^{\prime}$, $\mathbf{y}_{2,t} = (y^{(2)}_{1,t}, ..., y^{(2)}_{n_2,t})^{\prime}$ and $\mathbf{y}^{'}_{3,t} = (y^{(3)}_{1,t}, ..., y^{(3)}_{n_3,t})^{\prime}$, with $n_1 + n_2 + n_3 = n$. Let L^2 (Ω, F, P) be the Hilbert space of real-valued square-integrable random variables on (Ω, F, P), with inner product $\langle x, y \rangle = E(xy)$. We denote by $H_{123}(t)$, $H_{12}(t)$, $H_{13}(t)$, and $H_1(t)$ the closures with respect to mean square convergence of the linear manifolds generated, respectively, by subsets

$$\left\{y^{(1)}_{1,\tau}, ..., y^{(1)}_{n_1,\tau}, y^{(2)}_{1,\tau}, ..., y^{(2)}_{n_2,\tau}, y^{(3)}_{1,\tau}, ..., y^{(3)}_{n_3,\tau}; \quad \tau \leq t\right\},$$
$$\left\{y^{(1)}_{1,\tau}, ..., y^{(1)}_{n_1,\tau}, y^{(2)}_{1,\tau}, ..., y^{(2)}_{n_2,\tau},; \quad \tau \leq t\right\},$$
$$\left\{y^{(1)}_{1,\tau}, ..., y^{(1)}_{n_1,\tau}, y^{(3)}_{1,\tau}, ..., y^{(3)}_{n_3,\tau}; \quad \tau \leq t\right\},$$
$$\left\{y^{(1)}_{1,\tau}, ..., y^{(1)}_{n_1,\tau}; \quad \tau \leq t\right\},$$

of L^2 (Ω, F, P). Let N and M be two closed subspaces of L^2 (Ω, F, P), we denote by $N \vee M$; the closure, with respect mean square convergence, of the linear manifold generated by $N \cup M$. For any closed subspace S of L^2 (Ω, F, P) and for $1 \leq l \leq n_i$, $(i = 1, 2, 3)$ we denote $P(y^{(i)}_{l,t+1}|S)$ the orthogonal projection of $y^{(i)}_{i,t+1}$ on S and $P(\mathbf{y}_{i,t+1}|S) = \left[P(y^{(i)}_{i,t+1}|S), ..., P(y^{(i)}_{n_i,t+1}|S)\right]^{\prime}$. In the sequel we will use the following definitions of non-causality.

Definition 1. The vector $\mathbf{y}_3$ does not Granger cause $\mathbf{y}_1$, with respect to $H_{123}(t)$ iff

$$P(\mathbf{y}_{1,t+1}|H_{123}(t)) = P(\mathbf{y}_{1,t+1}|H_{12}(t))\forall t.$$

Symbolically : $\mathbf{y}_3 \nrightarrow \mathbf{y}_1|H_{123}(t)$. Causality from $\mathbf{y}_3$ to $\mathbf{y}_1$, is symbolized by $\mathbf{y}_3 \rightarrow \mathbf{y}_1|H_{123}(t)$.

Definition 2. The vector $\mathbf{y}_3$ does not Granger cause $\mathbf{y}_1$, with respect to $H_{13}(t)$ iff

$$P(\mathbf{y}_{1,t+1}|H_{123}(t)) = P(\mathbf{y}_{1,t+1}|H_1(t))\forall t.$$

Symbolically: $\mathbf{y}_3 \nrightarrow \mathbf{y}_1|H_{13}(t)$. Causality from $\mathbf{y}_3$ to $\mathbf{y}_1$, is symbolized by $\mathbf{y}_3 \rightarrow \mathbf{y}_1|H_{13}(t)$.

Definition 3. The vector $\mathbf{y}_2$ does not Granger cause $\mathbf{y}_1$, with respect to $H_{12}(t)$ iff

$$P(\mathbf{y}_{1,t+1}|H_{12}(t)) = P(\mathbf{y}_{1,t+1}|H_1(t))\forall t.$$

Symbolically: $\mathbf{y}_2 \nrightarrow \mathbf{y}_1|H_{12}(t)$. Causality from $\mathbf{y}_2$ to $\mathbf{y}_1$, is symbolized by $\mathbf{y}_2 \rightarrow \mathbf{y}_1|H_{12}(t)$.

In order to test the null hypothesis that $\mathbf{y}_3 \not\to \mathbf{y}_1 | H_{123}(t)$ we can suppose that $\mathbf{y}_t = \left(\mathbf{y}_{1,t}^{'}, \mathbf{y}_{2,t}^{'}, \mathbf{y}_{3,t}^{'}\right)^{'}$ follows a VAR(p) model,

$$\begin{bmatrix} \mathbf{y}_{1,t} \\ \mathbf{y}_{2,t} \\ \mathbf{y}_{3,t} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}(L) & \mathbf{A}_{12}(L) & \mathbf{A}_{13}(L) \\ \mathbf{A}_{21}(L) & \mathbf{A}_{22}(L) & \mathbf{A}_{23}(L) \\ \mathbf{A}_{31}(L) & \mathbf{A}_{32}(L) & \mathbf{A}_{33}(L) \end{bmatrix} \begin{bmatrix} \mathbf{y}_{1,t-1} \\ \mathbf{y}_{2,t-1} \\ \mathbf{y}_{3,t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_{1,t} \\ \mathbf{u}_{2,t} \\ \mathbf{u}_{3,t} \end{bmatrix} \tag{2}$$

where $\mathbf{A}_{ij}(L) = \sum_{h=1}^{p} \mathbf{A}_{ij,h} L^{h-1}$, $\mathbf{A}_{ij,h}$ are fixed coefficient matrices, and $\mathbf{u}_t = \left(\mathbf{u}_{1,t}^{'}, \mathbf{u}_{2,t}^{'}, \mathbf{u}_{3,t}^{'}\right)^{'}$ is a n-dimensional white noise with non singular covariance matrix $\mathbf{\Sigma_u}$.

In this framework it is well known that $\mathbf{y}_3$ does not Granger cause $\mathbf{y}_1$ with respect to $H_{123}(t)$, if and only if $\mathbf{A}_{13}(L) = \mathbf{0}$. Thus the null hypothesis becomes

$$H_0 : \mathbf{A}_{13,h} = \mathbf{0}, \quad h = 1, ..., p \tag{3}$$

The usual test F can be computed to test this hypothesis. Stationarity ensures that this statistic is distributed asymptotically as a $F(j, T - np)$ random variable where j is the number of restrictions under the null hypothesis (3). The order of the dimensionality problem can be evaluated by the following simple Monte Carlo experiment on the power of the test F. Following Davidson and Mackinnon (1993, p. 419) if the alternative hypothesis is very general, the power may either rise or fall as we increase the VAR order (that is increasing the number of the parameters involved by the null hypothesis). Then we keep fixed the VAR order, and the dimension of the vector $\mathbf{y}_3$, and in particular we choose a VAR(1) and $n_3 = 1$.

Consider the following VAR models:

$$\begin{bmatrix} y_{1t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} .3 & 0 \\ .5 & -.4 \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{3t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{3t} \end{bmatrix} \quad \Sigma_u = \begin{bmatrix} 1.3 & -0.3 \\ -0.3 & 1.2 \end{bmatrix} \tag{4}$$

where, in this case, $\mathbf{y}_{1,t} = (y_{1t})$, and $\mathbf{y}_{3,t} = (y_{3t})$ with $n_1 = 1$, $n_2 = 0$, $n_3 = 1$, and

$$\begin{bmatrix} y_{1t} \\ y_{2t}^{'} \\ y_{2t}^{''} \\ y_{2t}^{'''} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} .3 & 0 & 0 & 0 & 0 \\ 0 & -.4 & .2 & .7 & 0 \\ .5 & .1 & .6 & 0 & .3 \\ .3 & 0 & .1 & .1 & -.3 \\ .6 & .2 & 0 & 0 & -.5 \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1}^{'} \\ y_{2t-1}^{''} \\ y_{2t-1}^{'''} \\ y_{3t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \\ u_{4t} \\ u_{5t} \end{bmatrix} \tag{5}$$

with

$$\Sigma_u = \begin{bmatrix} 1.3 & 0.2 & -0.3 & 0.7 & 0 \\ 0.2 & 1.7 & 0.5 & 0 & -0.6 \\ -0.3 & 0.5 & 1.2 & 0.3 & 0.2 \\ 0.7 & 0 & 0.3 & 1.5 & 0 \\ 0 & -0.6 & 0.2 & 0 & 1 \end{bmatrix}$$

and where, $\mathbf{y}_{1,t} = (y_{1t})$, $\mathbf{y}_{2,t} = (y'_{2t}, y''_{2t}, y'''_{2t})'$ and $\mathbf{y}_{3,t} = (y_{3t})$ with $n_1 = 1$, $n_2 = 3$, $n_3 = 1$.

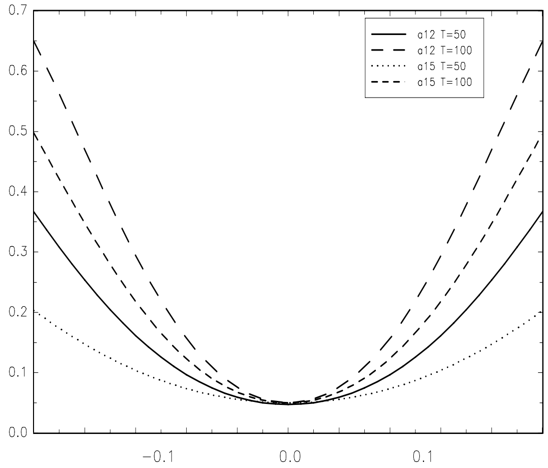
For these models the null hypothesis $\mathbf{y}_3$ does not Granger cause $\mathbf{y}_1$is respectively: $H_0 : a_{12} = 0$ and $H_0 : a_{15} = 0$.

The first model has four parameters and $(1, T-2)$ degree of freedom for the F-test while the second has 25 parameters and $(1, T-5)$ degree of freedom. It's clear that the dimensional problem is more relevant in small sample.

Generally, the applications in macro-economics consider sample size no longer that 150 observations (see, for the case of the analysis of money demand, Golinelli and Pastorello [2]).

We consider $T = 50$ and $T = 100$ as sample size and 1000 Monte Carlo replications for each experiments. The VARs are estimated by OLS,as usual [5]. Under the local alternative hypothesis the F-test is distributed as a non-central F distribution where the non central parameter depends on the local alternative (see for example Savin (1984)). Let $H_1 : a_{12} = \delta$ and $H_1 : a_{15} = \delta$ where $\delta \in [-0.2, 0.2]$. As we can see in figure 1 the loss in the power is clear. As expected the loss in power is reduced when the sample size increases.

Figure 1: Powers F−test



3 A possible solution

The usual solution for the problem showed above is to simplify the model. In our case, for example, we may specify a VAR model for the sub process $\left(\mathbf{y}_{1,t}^{'}, \mathbf{y}_{3,t}^{'}\right)^{'}$,

$$\begin{bmatrix} \mathbf{y}_{1,t} \\ \mathbf{y}_{3,t} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{11}(L) & \mathbf{B}_{12}(L) \\ \mathbf{B}_{21}(L) & \mathbf{B}_{22}(L) \end{bmatrix} \begin{bmatrix} \mathbf{y}_{1,t-1} \\ \mathbf{y}_{3,t-1} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_{1,t} \\ \mathbf{e}_{3,t} \end{bmatrix} \tag{6}$$

where $\mathbf{B}_{ij}(L) = \sum_{h=1}^{k} \mathbf{B}_{ij,h} L^{h-1}$, $\mathbf{B}_{ij,h}$ are fixed coefficient matrices, and $\mathbf{e}_t = \left(\mathbf{e}_{1,t}^{'}, \mathbf{e}_{3,t}^{'}\right)^{'}$ is a $(n - n_2)$-dimensional white noise with non singular covariance matrix $\mathbf{\Sigma_e}$. In this case, the null hypothesis $\mathbf{y}_3 \nrightarrow \mathbf{y}_1 | H_{13}(t)$ assumes the following form

$$H_0 : \mathbf{B}_{13,h} = \mathbf{0}, \quad h = 1, ..., k \tag{7}$$

Again, we can use a F statistic to test this hypothesis. The drawback of this procedure is that the omission of variables in $\mathbf{y}_2$ may produce an invalid causal inference. A classical example of this situation is the so-called problem of noncausality due to omitted variables [4]. However, under certain conditions, the causal inference is robust towards omission of the information in $\mathbf{y}_{2,\tau}$ $\tau \leq t$. More precisely, we have that if $\mathbf{y}_2 \nrightarrow \mathbf{y}_1 | H_{12}(t)$, then $\mathbf{y}_3 \rightarrow \mathbf{y}_1 | H_{123}(t)$ if and only if $\mathbf{y}_3 \rightarrow \mathbf{y}_1 | H_{13}(t)$. This result is shown in [7, p. 597]. According to this result we have that if the vector $\mathbf{y}_2$ does not cause the vector to be predicted, $\mathbf{y}_1$, with respect $H_{12}(t)$, then the inference on causality from $\mathbf{y}_3$ to $\mathbf{y}_1$ is invariant to the selection of $H_{123}(t)$ or $H_{123}(t)$ information set. Thus, in order to test the null hypothesis $\mathbf{y}_3 \nrightarrow \mathbf{y}_1 | H_{123}(t)$, first, we can test the hypothesis $\mathbf{y}_2 \nrightarrow \mathbf{y}_1 | H_{12}(t)$, if this hypothesis is accepted we can test the null hypothesis $\mathbf{y}_3 \nrightarrow \mathbf{y}_1 | H_{13}(t)$ in order to limit the negative effects of high dimensionality.

3.1 The Monte Carlo experiment

A natural way to evaluate the gain of power using the proposed strategy to test the Granger causality is to perform a Monte Carlo experiment. We are interested to test if $\mathbf{y}_3$ does not Granger cause $\mathbf{y}_1$.Taking into account the above observation about the power behavior with a general alternative hypothesis, we consider three stationary VAR(1) models with three, four and five variables, $\mathbf{y}_{1,t} = (y_{1t})$, $\mathbf{y}_{3,t} = (y_{3t})$ and $n_2 = 1, 2, 3$, respectively, to evaluate the performance of the proposed solution growing the number of variable in $\mathbf{y}_2$.

As seen before the Granger causality can be verified conducting a F-test just on a parameter. Sample size $T = 50$ and $T = 100$ and 1000 Monte Carlo replication for each experiments.

Consider the following VAR(1) model

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{bmatrix} = \begin{bmatrix} 0.3 & 0 & 0 \\ 0.5 & -0.4 & 0.7 \\ 0 & 0 & 0.3 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \end{bmatrix} \tag{8}$$

with

$$\Sigma_u = \begin{bmatrix} 1.3 & 0.2 & -0.3 \\ 0.2 & 1.7 & 0.5 \\ -0.3 & 0.5 & 1.2 \end{bmatrix}$$

that satisfy required condition: $\mathbf{y}_2 \not\rightarrow \mathbf{y}_1 | H_{12}(t)$, where $\mathbf{y}_2 = (y_2)$. The Granger causality test of y_3 on y_1 involves the null hypothesis $H_0 : a_{13} = 0$. Following the proposed solution the same Granger causality can be verified on the following simplified VAR(1):

$$\begin{bmatrix} y_{1,t} \\ y_{3,t} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{3,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{3,t} \end{bmatrix} \tag{9}$$

performing a F-Test where the null Hypothesis is $H_0 : b_{12} = 0$.
In figure 2 are reported the power functions of the F-test on $H_0 : a_{13} = 0$ and $H_0 : b_{12} = 0$ for $T = 50$ and $T = 100$ performed on (8) and (9), considering a local alternatives $H_1 : a_{13} = \delta$ and $H_1 : b_{12} = \delta$ where $\delta \in [-0.2, 0.2]$. As we can see, the gain in the power is clear especially when the alternative hypothesis is greater than ±0.1. The dotted vertical lines indicate ± the Monte carlo mean of the estimated standard error of a_{13} when $T = 100$.

Figure 2: Power F test Granger-Causality VAR(1)

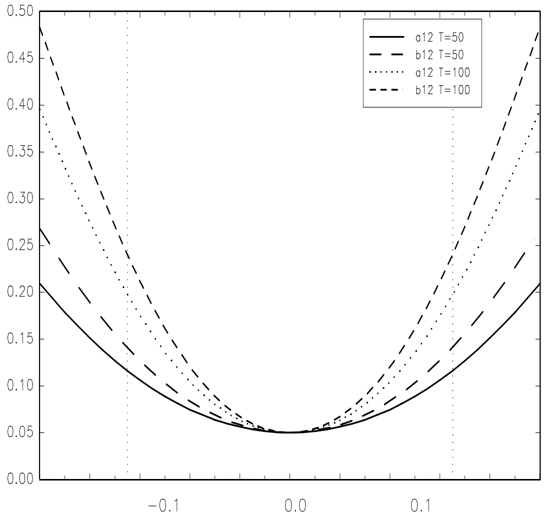


Figure 3: Power F test Granger–Causality VAR(1)

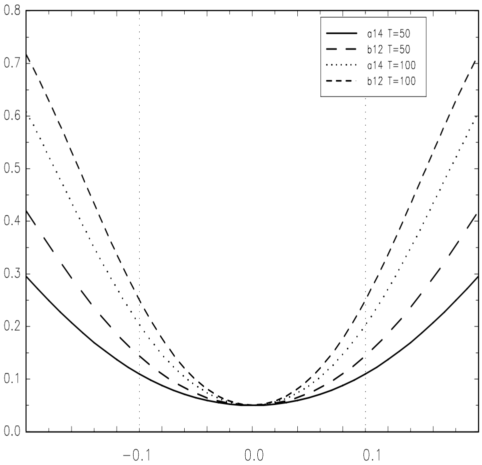
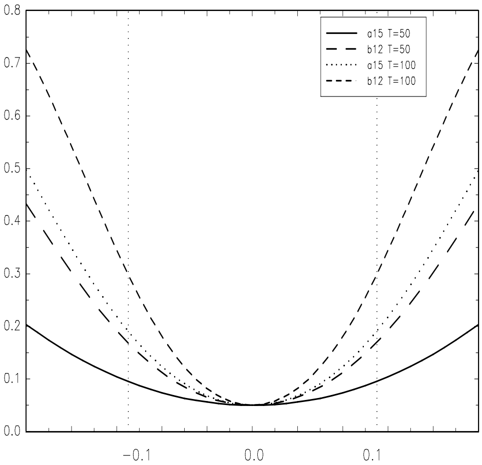


Figure 4: Power F test Granger–Causality VAR(1)



A similar experiment has been conducted on the VAR(1) model with four variables of interest:

$$\begin{bmatrix} y_{1,t} \\ y'_{2,t} \\ y''_{2,t} \\ y_{3,t} \end{bmatrix} = \begin{bmatrix} 0.3 & 0 & 0 & 0 \\ 0 & -0.4 & 0.2 & 0.7 \\ 0.3 & 0 & 0.1 & 0 \\ 0.6 & 0 & 0 & -0.5 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \\ y_{4,t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \\ u_{4,t} \end{bmatrix} \qquad \Sigma_u = I \tag{10}$$

that, again, satisfy required condition: $\mathbf{y}_2 \not\rightarrow \mathbf{y}_1 | H_{12}(t)$, where $\mathbf{y}_2 = (y'_2, y''_2)'$. The Granger causality test of y_3 on y_1 involves the null hypothesis $H_0 : a_{14} = 0$, and, as before, the same hypothesis can be conducted on the restricted

VAR(1) obtained from (10) deleting $\mathbf{y}_2 = (y_2', y_2'')'$. The power functions on the F-tests are reported in figure 3. As before the gain of the power become relevant near ± the Monte Carlo mean of the estimated standard error of a_{14} when $T = 100$ represented by dotted vertical lines.

The similar experiment can be conducted on the VAR(1) model with five variables (5). In this case the Granger causality test of y_3 on y_1 can be conducted on the restricted VAR(1) obtained from (5) deleting $\mathbf{y}_2 = (y_2', y_2'', y_2''')'$. The power functions on the F-tests are reported in figure 4. As expected, Figure 2-4 show that, as the number of variables in $\mathbf{y}_{2,t}$ raises, the power gain becomes more relevant.

4 Conclusion

The problem of dimensionality is very important for multivariate time series since for vector autoregressive models the number of parameters to estimate grows rapidly with the number of variables. In this paper, by a Monte Carlo study, we have analyzed the impact of high dimensionality on the power of noncausality test. We have, also proposed a testing strategy that, under certain conditions, enables a reduction of dimensionality. The gain of power of this solution has been evaluated, by a Monte Carlo simulation.

References

[1] Abadir K., Hadri K., Tzavalis E. (1999). *The influence of VAR dimensions on estimator biases.* Econometrica **67** (1), 163 − 182.

[2] Golinelli R., Pastorello S. (2002). *Modelling the demand for M3 in the euro area.* The European Journal of Finance **8** (4), 371 − 401.

[3] Gonzalo J., Pitirakis J.Y. (2000). *Dimensionality effect in cointegration analysis.* In R.F. Engle and H. White (eds), Cointegration, Causality and Forecasting, Festschrift in Honour of Clive W.J. Granger. Oxford University Press.

[4] Lütkepohl H. (1982). *Non-causality due to omitted variables.* Journal of Econometrics **19**, 367 − 378.

[5] Lütkepohl H. (1991). *Introduction to multiple time series analysis.* Springer. Berlin.

[6] Sims C. A. (1980). *Macroeconomics and reality.* Econometrica **48**, 1 − 48.

[7] Triacca U. (2002). *Selection of the relevant information set for predictive relationship analysis between time series.* Journal of Forecasting **21**, 595 − 599.

Address: F. Di Iorio, Dipartimento di Scienze Statistiche, Università degli Studi di Napoli, via L. Rodinò 22, Napoli, Italy
U. Triacca, Facoltà di Economia, Università di L'Aquila, Roio Poggio, L'Aquila, Italy

E-mail: `fdiiorio@unina.it`

A MIXTURE OF MIXTURE MODELS TO DETECT UNITY MEASURE ERRORS

Marco Di Zio, Ugo Guarnera and Roberto Rocci

Key words: Editing, systematic error, mixture models, EM algorithm.

COMPSTAT 2004 section: Clustering, Official statistics.

Abstract: The unity measure error is one of the most frequent systematic errors in surveys measuring quantitative variables. In this paper we reinterpret the identification of items in error as a clustering problem where each class is associated with a specific error pattern. In particular we use a two-level mixture approach where each class is modelled as a multivariate Gaussian mixture to allow effective classification in non-normal settings. Finally, an application of the method to the *1997 Italian Labour Cost Survey* is reported.

1 Introduction

In the production of Official Statistics, a very crucial point is the data editing phase. This phase consists of localising non-sampling errors in data (*editing*) and treating them, generally substituting the value with a more plausible one (*imputation*). Data editing is important both in terms of data quality, and survey costs. Thus the techniques introduced to clean data are essentially required to take into account both these aspects simultaneously, in order to balance the trade off between them [6]. Recently, with the advances in computers capabilities, the automatic editing approach, based on the Fellegi-Holt paradigm [5], [3], has increased its popularity. However, this approach is appropriate only for dealing with random errors and requires data free of systematic errors.

A particular systematic error, that frequently appears in surveys collecting numerical data, is the unity measure error times a constant factor (e.g. 100 or 1,000). This error is due to the erroneous choice, by some respondents, of the unity measure in reporting the amount of questionnaire items, e.g. a respondent is request to report the amount of money in thousands but he expresses (erroneously) in millions. This error highly affects both data accuracy (bias) and editing and imputation costs. In fact all the automatic data editing process cannot be performed in the right way if this error is not removed preliminarily. In the National Statistical Institutes, this error is generally treated through ad hoc solutions, using mainly graphical analyses and ratio edits, i.e. bounds on ratios between two variables. The limit of this approach is both in terms of quality and costs. Quality is limited by the fact that in the traditional approaches no more than a pairwise relationship between variables can be analyzed. Costs are essentially influenced by the fact that for each survey a new ad hoc procedure must be set up.

In this paper we present an approach to the localization of observations affected by systematic error based on mixture modelling [11]. The plan of the paper is the following. In section 2 we present a model which can be considered as a generalization of the work by Di Zio et al. [4]. In section 3 an EM algorithm is introduced to compute the maximum likelihood estimates of model parameters. Finally, in section 4 we describe an application of the method to the *1997 Italian Labour Cost Survey*.

2 The model

Recently, the systematic error problem have been treated by Di Zio et al. [4] in the probabilistic framework of mixture modelling. According to this approach, non erroneous data are considered as independent realizations from a random variable $\tilde{\mathbf{X}}$ with density $\tilde{g}_0(\mathbf{x})$. If $\tilde{\mathbf{X}}$ is a p-vector of variables $\tilde{X}_1, \ldots, \tilde{X}_p$, the unity measure error acts on each variable $\tilde{X}_j \;\; (j = 1, \ldots, p)$ through the transformation $\tilde{X}_j \rightarrow \tilde{c}\tilde{X}_j$ where $\tilde{c}$ is a constant factor (the generalization to the case of different constant factors for different variables is straightforward). In presence of non negative-valued variables (a typical situation in economic surveys), it is useful to work in the logarithmic scale: if we let $\mathbf{X} = \log(\tilde{\mathbf{X}})$, and denote by $g_0(\mathbf{x})$ the p.d.f. associated with $\mathbf{X}$, the unity measure error can be represented through the transformation $X_j \rightarrow X_j + c$ where $c = \log(\tilde{c})$. In this way, for each subset of indices $l = \{j_1, \ldots, j_k\} \subset \{1, \ldots, p\}$, the observations affected by a unity measure error in the variables $\mathbf{X}_l = [X_{j_1}, \ldots, X_{j_k}]'$ with $(j_k \leq p)$ define a cluster, that can be labelled by l, similar in shape to the non-erroneous units cluster, but spread around a different location. More precisely, the units of cluster l can be thought of as generated by the density $g_l(\mathbf{x}) = g_0(\mathbf{x} - \mathbf{c}_l)$ where $\mathbf{c}_l$ is a vector whose components $c_{lj}(j = 1, \ldots, p)$ equal c if $j \in l$, zero otherwise. Thus the data can be modelled trough the mixture density

$$f(\mathbf{x}) = \sum_{l=1}^{L} p_l g_l(\mathbf{x}), \tag{1}$$

where the summation is over L distinct error patterns. This approach allows us to reinterpret the localization of the systematic error as a clustering problem. In fact, through this model we can estimate for each observation $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ the "posterior" probability to belong to a particular cluster, i.e.

$$Pr(l|\mathbf{x}_i) = \frac{p_l g_l(\mathbf{x}_i)}{\sum_{l=1}^{L} p_l g_l(\mathbf{x}_i)}. \tag{2}$$

They can be used to classify the observations by the L different error patterns, by assigning each observation to the cluster, i.e. error pattern, corresponding to the maximum posterior probability. In this way, atypicality indices can be built to be successfully used through selective editing procedures in order to balance the trade off between accuracy and costs. In other words, between automatic and interactive review of the observations. In fact, the observations that will be classified as erroneous with low probability, are highly recommended to be clerically reviewed in order to be sure about the nature of the observation. Furthermore, it should be noted that this modelling exploits multivariate relationships among variables, thus increasing the quality of the procedure. It also allows to develop an algorithm that can be easily generalized to all the cases where the unity measure error arises, thus decreasing costs of the editing phase.

In Di Zio et al. [4] the "non erroneous data" density g_0 is taken as p-variate normal with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so that each component g_l is obtained from g_0 simply by substituting $\boldsymbol{\mu}$ with $\boldsymbol{\mu} + \mathbf{c}_l$. In the present approach the normality assumption for the data distribution is avoided by allowing more general forms for the functions g_l in (1). In particular, the density g_l is estimated from the data. This is accomplished by assuming that each density g_l is, in turn, expressed in the form of a mixture of M Gaussians

$$g_l(\mathbf{x}) = \sum_{m=1}^{M} q_m h(\mathbf{x} - \mathbf{c}_l; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{3}$$

where $h(\cdot; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is a p-variate normal density with mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$. From (1) and (3) it follows that the density of $\mathbf{X}$ can be written as

$$\begin{aligned} f(\mathbf{x}) &= \sum_{l=1}^{L}\sum_{m=1}^{M} p_l q_m h(\mathbf{x} - \mathbf{c}_l; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \\ &= \sum_{l=1}^{L}\sum_{m=1}^{M} p_l q_m h(\mathbf{x}; \boldsymbol{\mu}_m + \mathbf{c}_l, \boldsymbol{\Sigma}_m), \end{aligned} \tag{4}$$

where

$$\begin{aligned} h(\mathbf{x}; \boldsymbol{\mu}_m + \mathbf{c}_l, \boldsymbol{\Sigma}_m) &= (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_m|^{-\frac{1}{2}} \times \\ &\times \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_m - \mathbf{c}_l)'\boldsymbol{\Sigma}_m^{-1}(\mathbf{x} - \boldsymbol{\mu}_m - \mathbf{c}_l)\right]. \end{aligned}$$

Formula (4) is the density of a mixture of $L \times M$ Gaussians with mixing proportions and mean vectors suitably constrained. It is clear that mixture modelling is here used both for classification and density estimation. Our approach has the advantage to be mathematically tractable and, at the same time, it gives the possibility to model distribution that are far from normal [10].

3 Maximum likelihood estimates

By assuming the independence of observations, we can write the log-likelihood of the whole sample as

$$\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i) = \sum_{i=1}^{n} \log \left\{ \sum_{l=1}^{L} \sum_{m=1}^{M} p_l q_m h_{ilm} \right\}, \tag{5}$$

where $\boldsymbol{\vartheta}$ is the whole set of parameters to be estimated and $h_{ilm} = h(\mathbf{x}_i; \boldsymbol{\mu}_m + \mathbf{c}_l, \boldsymbol{\Sigma}_m)$. To compute the maximum likelihood estimates of model parameters, we note that (see [8] the maximization of (5) is equivalent to the maximization of the "fuzzy" function

$$\ell_{\mathrm{f}}(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \sum_{l=1}^{L} \sum_{m=1}^{M} u_{ilm} \log(p_l q_m h_{ilm}) - \sum_{i=1}^{n} \sum_{l=1}^{L} \sum_{m=1}^{M} u_{ilm} \log(u_{ilm}), \tag{6}$$

where the u_{ilm}'s are non-negative and such that $\sum_{lm} u_{ilm} = 1$ for $i = 1, 2, ..., n$. To maximize ℓ_{f} we adopt a coordinate ascent method, where in each step the objective function is maximized with respect to a subset of parameters given the current values of the others. In this way each parameter, or subset of parameters, is in turn updated and the algorithm increases the value of the objective function at each step. The algorithm stops when the function increment in a particular step is lower than a given threshold. The fundamental steps of our algorithm are the following:

(a) **Update of** u_{ilm}
It can be easily shown that (6) has a maximum with respect to the u's when

$$u_{ilm} = \frac{p_l q_m h_{ilm}}{\sum_{lm} p_l q_m h_{ilm}}; \tag{7}$$

(b) **Update of** p_l
By rewriting (6) as

$$\ell_{\mathrm{f}}(\boldsymbol{\vartheta}) = \sum_{ilm} u_{ilm} \log(p_l) + const, \tag{8}$$

where *const* indicates a term that does not depend on the p's, we deduce that (6) is maximized with respect to the p's when

$$p_l = \frac{1}{n} \sum_{im} u_{ilm}; \tag{9}$$

(c) **Update of** q_m
As in the previous step, it can be shown that (6) obtains a maximum with respect to the q's when

$$q_m = \frac{1}{n} \sum_{il} u_{ilm}; \tag{10}$$

(d) **Update of $\boldsymbol{\mu}_m$**
First we rewrite (6) as

$$\ell_{\mathrm{f}}(\boldsymbol{\vartheta}) = \sum_{ilm} u_{ilm} \left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_m - \mathbf{c}_l)' \boldsymbol{\Sigma}_m^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_m - \mathbf{c}_l) \right] + const, \tag{11}$$

where *const* indicates a term independent of the $\boldsymbol{\mu}$'s. Then it simply follows that

$$\boldsymbol{\mu}_m = \frac{1}{\sum_{il} u_{ilm}} \sum_{il} u_{ilm}(\mathbf{x}_i - \mathbf{c}_l); \tag{12}$$

(e) **Update of $\boldsymbol{\Sigma}_m$**
By rewriting (6) as

$$\ell_{\mathrm{f}}(\boldsymbol{\vartheta}) = -\frac{1}{2} \sum_{ilm} u_{ilm} [\log(|\boldsymbol{\Sigma}_m| + \mathbf{d}'_{ilm} \boldsymbol{\Sigma}_m^{-1} \mathbf{d}_{ilm})] + const, \tag{13}$$

where *const* indicates a term independent of the $\boldsymbol{\Sigma}_m$'s and $\mathbf{d}_{ilm} = \mathbf{x}_i - \boldsymbol{\mu}_m - \mathbf{c}_l$, we deduce that the update of $\boldsymbol{\Sigma}_m$ is

$$\boldsymbol{\Sigma}_m = \frac{1}{\sum_{il} u_{ilm}} \sum_{il} u_{ilm} \mathbf{d}_{ilm} \mathbf{d}'_{ilm}, \tag{14}$$

while in the homoscedastic case, i.e. $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m$, we have

$$\boldsymbol{\Sigma} = \frac{1}{n} \sum_{ilm} u_{ilm} \mathbf{d}_{ilm} \mathbf{d}'_{ilm}. \tag{15}$$

By iterating the above described steps we obtain a monotone algorithm which can be easily shown to be of ECM type [12].

In practical applications, it turns out that a crucial role is played by the choice of the starting points, as usual in the EM algorithms (see [1]. To overcome this problem, we developed two different strategies of initialization. The first consists in several "short run", in terms of number of iterations, of the algorithm from random initializations followed by a "long" run of EM from the solution maximizing the observed log-likelihood. The second is based on a constrained version of the k-means clustering technique that should approximate the mixture model, i.e. the $L \times M$ centroids are constrained to be of the form $\boldsymbol{\mu}_m + \mathbf{c}_l$.

4 Application

To illustrate the effectiveness and test the performance of our proposal, we carry out an experiment on a subset of the *1997 Italian Labour Cost Survey* (LCS). The LCS is a periodic sample survey that collects information on employment, hours worked, wages, salaries and labour cost on about

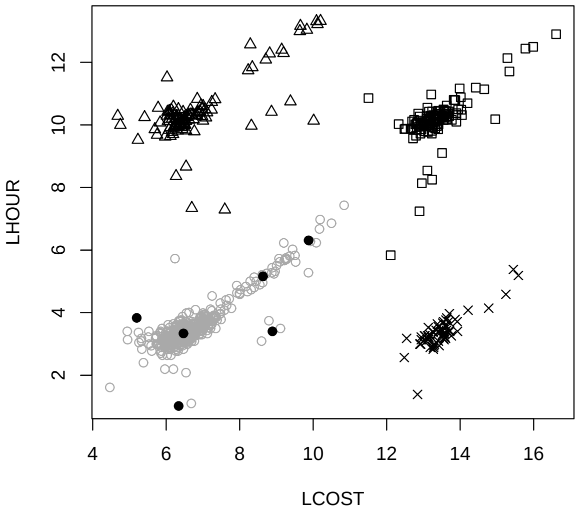


Figure 1: Classification through a mixture with $M = 6$ within cluster components.

12.000 firms with more than 10 employees. The survey is subject to a specific European Regulation requiring to all the European Community Member States to collect every four years detailed information about the labour cost and employment structure in some specific Industries. Our data-set consists of 744 units that belong to the metallurgic economic activity sector. In particular, we analyze two main variables measuring the *Total Labour Cost* and *Total Hours Worked*. These variables are affected by the *1000-factor error*, since some respondents have expressed the Total Labour Cost in thousand of Italian Lira instead of millions, and similarly the hours have not been reported in thousand as requested. Details on the error profile and the impact of systematic error on data accuracy can be found in Cirianni et al. [2].

We take the logarithmic transformation of the Total Labour Cost (LCOST), Total Hours Worked (LHOUR), and we define the clusters associated with the four different error patterns (cluster1 = no errors, cluster2 = only LHOUR in error, cluster3 = only LCOST in error, cluster4 = both variables in error).

In order to classify firms according to their unity measure error pattern, we follow the approach described in the previous sections, modelling data through a homoscedastic Gaussian mixture. The starting points for the EM algorithm are determined as described in the previous section. Different experiments are performed by varying the number M of the components of the within-cluster mixture. The optimal number of components is chosen according to the BIC criterion (see [9]. It results that the appropriate choice is $M = 6$. The resulting classification is reported in figure 1, where cluster1 is denoted by circles, cluter2 by triangles, cluster3 by crosses and cluster4 by squares. In addition the mean vectors of the within cluster mixture are depicted by solid circles. We remark that, for the sake of simplicity, only the means of the cluster1 are reported, while the others can be easily obtained by the appropriate translation.

It is worthwhile noting that, though the BIC criterion indicates that data are best fitted by the six-components mixture, in term of units classification the results do not depend appreciably on the mixture components number. However a stronger sensitivity to the components number is expected in presence of clusters more overlapping each other. The estimation of the mixture modelling parameters and the clustering has been done through an R-code available upon request.

5 Discussion

In this paper a method to identify observations affected by unity measure errors is proposed. The problem is reinterpreted in a probabilistic clustering framework. The p.d.f. of the observations is modelled as a finite mixture where each component corresponds to a particular error pattern. The density of each component is, in turn, estimated by using a finite mixture of Gaussians in order to allow a more general setting. A similar approach has been also used by Hastie et al. [7] in the discriminant analysis context.

The maximum likelihood estimates of model parameters are computed by using an EM algorithm and two different strategies of initialization have been developed to cope with the local optima problem. The two strategies have been tested in a simulation study not reported here for the sake of brevity. They perform quite well, but further research is needed especially to decrease the computational time. In the actual setting, the systematic error is supposed to be a priori known, however the model can be easily extended to the case where the exact mechanism of the error is unknown and it has to be estimated from data.

References

[1] Biernacki C., Celeux G., Govaert G. (2003). *Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models.* Computational Statistics & Data Analysis, **41**,561–575.

[2] Cirianni A., Di Zio M., Luzi O. and Seeber A.C. (2000). *The new integrated data procedure for the Italian Labour Cost survey: measuring the effects on data of combined techniques.* Proceedings of the Interantional Conference on Establishment Surveys II (ICES), June 17-21 2000, Buffalo, USA.

[3] De Waal T. (2003). *Solving The Error Localization Problem by Means of Vertex Generation.* Survey Methodology, **1**,71 – 79.

[4] Di Zio M., Guarnera U., and Luzi O. (2003). *Using Mixture Modelling to deal with Unity Measure Error.* UN/ECE Work Session on Statistical Data Editing, October 20-22 2003, Madrid, Spain (www.unece.org/stats/documents/2003.10.sde.htm).

[5] Fellegi I.P., and Holt D. (1976). *A systematic approach to edit and imputation.* Journal of the American Statistical Association, **71**, 17 – 35.

[6] Granquist, L. (1995). *Improving the Traditional Editing Process.* In B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge, and P.S. Kott (Eds) *Business Survey Methods*, Wiley, New York.

[7] Hastie T., and Tibshirani R. (1996). *Discriminant Analysis by Gaussian Mixtures.* Journal of the Royal Statistical Society (B), **58**, 155 – 176.

[8] Hathaway R.J. (1986). *Another interpretation of the EM algorithm for mixture distributions.* Statistics & Probability Letters, **4**, 53 – 56.

[9] Keribin C. (2000). *Consistent estimation of the order of mixture models.* Sankya: The Indian Journal of Statistics, **62**, 49 – 66.

[10] Marron J.S., and Wand M.P. (1992). *Exact Mean Integrated Squared Error.* The Annals of Statistics, **20**, 712 – 736.

[11] McLachlan G.J., and Basford K.E. (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

[12] Meng X.L., and Rubin D.B. (1993). *Maximum Likelihood Estimation via the ECM algorithm: a general framework.* Biometrika, **80**, 267 – 278.

Address: M. Di Zio, U. Guarnera, Istituto Nazionale di Statistica, via Cesare Balbo 16, 00184 Roma, Italy
R. Rocci, Università di Tor Vergata, via Columbia 2, 00133 Roma, Italy

E-mail: `dizio@istat.it`, `guarnera@istat.it`,
`roberto.rocci@uniroma2.it`

MULTIVARIATE TECHNIQUES FOR IMPUTATION BASED ON BAYESIAN NETWORKS

Marco Di Zio, Guiseppe Sacco, Mauro Scanu and Paola Vicard

Key words: Missing data, graphical models, multivariate analysis.

COMPSTAT 2004 section: Data imputation.

Abstract: In this paper, we compare two imputation procedures based on Bayesian networks. One method imputes missing items of a variable taking advantage only on information of its parents, while the other takes advantage of its Markov blanket. The structure of the paper is as follows. The first section contains an illustration of Bayesian networks. Then, we explain how to use the information contained in Bayesian networks in section 2. In section 3, we describe two evaluation indicators of imputation procedures. Finally, a Monte Carlo evaluation is carried on a real data set in section 4.

1 Bayesian networks

The methods for imputation proposed in this paper are based on *Bayesian networks* [1]. A Bayesian network (BN) is a graphical and numerical representation of the joint distribution of a set of variables, $(X_1, \dots, X_k)$ say. BNs are particularly important and useful to learn and represent the dependency structure of $(X_1, \dots, X_k)$. A BN is given by:

1. a directly acyclic graph (DAG) incorporating the set of dependencies among variables and
2. an inferential engine to make inference on the parameters of the model.

A DAG is a pair $G = (V, E)$ where V is the set of nodes — each node representing, in our case, a variable with a finite set of states — and E is the set of edges which are arrows linking pair of nodes. Cycles are forbidden, meaning that it is not possible to start form a node and then, following the direction of the arrows, end up in it. When two nodes, X_i and X_j, are connected by an arrow (i, j) pointing to X_j from X_i, we say that X_j is probabilistically dependent from X_i and that X_i is a parent of X_j. A conditional probability distribution is associated to every node. In particular, to a given variable X_j it is associated the conditional probability table of X_j given its parents $\mathrm{pa(X_j)}$. In this way the joint probability distribution can be factorised as follows:

$$\mathrm{P(X_1, \dots, X_k) = \prod_{i=1}^{k} P(X_i|pa(X_i))}. \tag{1}$$

In many contexts directed edges are interpreted in terms of causality and so there is wide debate about whether the data contain information able to suggest causal relations, i.e. choose the arrow direction. This problem is beyond the purposes of this paper since here the variable order denotes their reliability order (see section 2.1). Once the BN has been learnt, its modular structure can be exploited to apply fast and efficient algorithms.

2 Use of Bayesian networks for imputation

One of the most important features of BNs is the description of the information directly associated (that is the nodes directly connected) with each variable. This aspect represents one of the most important elements for an efficient imputation algorithm: the search of a complete and at the same time synthetic source of information for inferring a good value to impute to the missing items. Two methods have been proposed considering two different kinds of information: the parents [3] and the Markov blanket [2] of a variable.

In the first one, the variables are firstly ordered according to their reliability, where reliability is defined in terms of the overall quality of each variable (see section 2.1 for more details). The BN structure should respect this ordering (i.e. an arc always starts from a more reliable variable to a less reliable one). The distribution of a variable given its parents represents the uncertainty related to the true but unknown value of the missing item. Therefore the source of information used for imputing missing values of a variable is the distribution of that variable conditional on more reliable variables.

The second approach considers the fact that each variable is statistically related not only to its parents, but also to its children and the parents of its children, *i.e.* to its Markov blanket [1]. Then the information source for imputing missing items of a variable is the distribution of that variable conditional to its Markov blanket, *i.e.* the conditioning set now includes also less reliable variables. In fact the idea is that the conditioning variables should be a comprehensive set of all the variables able to give information on the variable to impute.

2.1 Two algorithms for imputation of missing values

In order to apply imputation by BNs it is preferable to order the variables at hand according to their *reliability*. A variable is more reliable when it is less affected by missing items and/or it is characterised by a higher accuracy and/or external sources are available. Actually the ordering should not be considered in a strict sense, *i.e.* more than one variable can be equally reliable.

Furthermore, most of the times the structure of the network is not intrinsic in the problem definition or given by previous analyses, so it has to be learnt from the data (with appropriate constraints). The same holds for the probability distributions. Under the previous reliability order, we have estimated a BN according to a two-steps procedure: at first we have estimated the structure of a BN by means of the PC algorithm [6] and secondly we have estimated the distributions of the BN (conditionally on the estimated structure) by means of the EM algorithm [4]. The network as well as the probability distributions are estimated here using the computer software Hugin (version 6.2, `www.hugin.com`; see also [5].

The procedures differ according to the imputation algorithm.

2.1.1 The BN imputation algorithm. In this case, the estimation of the BN structure is constrained to the reliability order among the variables. Let X_1 be the most reliable variable, X_2 be the second most reliable variable, $\dots$, X_k be the least reliable variable. Each missing item in X_1 is imputed by a random generation of a value from the marginal distribution of X_1. Once all the X_{t-1} $(t \leq k)$ missing items are imputed, each missing item in X_t is imputed by a random generation of a value from the distribution of X_t conditional on its parents. Note that X_t parents are always complete (i.e. all missing items have been imputed). This procedure is easily extended to the case two or more variables are equally reliable (see [3], for more details).

2.1.2 The MBBN imputation algorithm. In order to consider the distribution of the variable to impute conditional on its Markov blanket, the procedure needs more steps.

At first, we estimate on the data set with missing items as many BNs as the number of variables. Let G_t be the BN with the constraint that there are not any edges starting from X_t, $t = 1, \dots, k$. As a consequence, the Markov blanket for X_t in the graph G_t consists only of the X_t parents, $t = 1, \dots, k$.

Then, the previously estimated BNs are used sequentially in the imputation process in the following way. We impute the data set according to the BN imputation procedure by means of the BN whose structure is G_1. As a matter of fact, we impute a missing item for X_1 generating a value from the distribution of X_1 conditional on its parents (i.e. its Markov blanket). Note that X_1 parents may be either observed or imputed. However, once all X_1 missing items have been filled in, we suggest to retain just these values and discard the imputations for the other variables, $(X_2, \dots, X_k)$. We consequently obtain a new data set, D_1, where X_1 is complete. Imputations for X_2 missing items are performed on the data set D_1 with the BN whose structure is G_2. Once each X_2 missing item has been filled in, only the imputed values for X_2 are retained, obtaining a new data set D_2 where X_1 and X_2 are complete. This procedure is iterated for all the variables. Note that imputations for X_k are performed on a data set D_{k-1} where the other $k-1$ variables have been completed, and consequently the imputation step only involves the random generation from the distribution of X_k conditional on the (observed or previously completed) values of the variables in the Markov blanket of X_k in the BN whose structure is G_k.

3 Evaluation of the imputation procedures

In Di Zio et al. [3] we have considered two evaluation criteria for imputation techniques, useful in a simulated context.

A first criterion analyses the preservation of the joint multivariate distribution. The univariate version of the indicator is the following. Let us consider one variable X, with n^* missing items among the n units of the sample. Let us denote f_x as the relative frequency of category x of X in the true data set of the n^* missing items, and $\tilde{f}_x$ as the corresponding frequency after imputation. The indicator is:

$$\Delta = \frac{1}{2}\sum_x |f_x - \tilde{f}_x|, \tag{2}$$

where the sum is over the categories of X. Note that (2) takes values between zero and one. This indicator can be easily extended to the multivariate case, considering the comparison of the frequency distributions computed on the sub-data set where at least one of the variables is missing (for more details see [3]).

A second evaluation criterion considers the preservation of micro data in each single variable. In this case, the indicator evaluates the fraction of correct imputations out of all imputed values, *i.e.*

$$\xi = \frac{\sum I_{x_a}(\tilde{x}_a)}{n^*} \tag{3}$$

where n^* is the number of missing items in X, x_a is the true but unobserved value for unit a, $\tilde{x}_a$ is the corresponding imputed value, I is the indicator function and the sum is over the n^* units with X missing.

Since the BN imputation method gives explicitly the probability distribution used in the imputation process, it is also possible to evaluate the average value of (3). Without loosing in generality, let us consider X missing items refer to units $1, \ldots, n^*$. Then:

$$E(\xi) = \frac{\sum_{a=1}^{n^*} P(X = x_a | An(x_a))}{n^*}, \tag{4}$$

where $An(x_a)$ denotes the nearest ancestors of the variable X that, for unit a (where $X = x_a$), are observed. Consequently the probabilities in (4) may be obtained marginalising the distributions in the BN with respect to the unobserved parents and ancestors. Additionally, the set of conditioning variables may vary according to the pattern of missing items in the different records of the data set. A similar equation may be provided also for the expectation of ξ in the MBBN imputation procedure. It is enough to consider the probability distribution of the BN used in imputing variable X. In order to compute (4) easily, and to provide also expected values for Δ, reasonable approximations may be given by a Monte Carlo simulation.

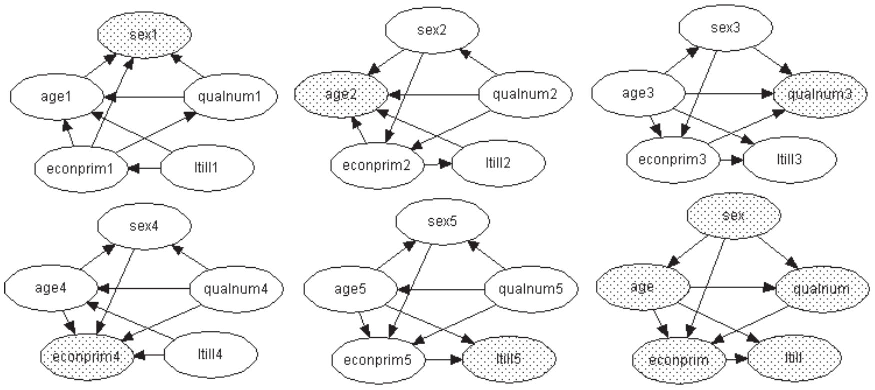


Figure 1: The first 5 BNs (indexed 1,2,3,4,5) are those used in the MBBN imputation algorithm. The last one is used in the BN imputation algorithm. The darker nodes are those whose imputed values are retained (in the BN imputation algorithm all nodes are imputed using the same BN).

4 Experimental results

The experiment has been carried out on categorical test data from a subset of 27,289 units of the 1991 U.K. Population Census Sample of Anonymised Records (SARS). Five variables have been analysed: age (6 categories), sex, primary economic position (econprim, 10 categories), long term illness (ltill, 2 categories), number of higher education qualifications (qualnum, 3 categories). We have introduced missing items according to the following scheme. All the variables have been contaminated according to an MCAR mechanism with different expected percentages of missingness: sex 7%, age 10%, qualnum 8%, econprim 20%, ltill 10%. Furthermore, the variables qualnum and ltill have been contaminated with a MAR mechanism following these rules:

$$P(qualnum = missing|age = i) = (2 \times i + 1)\%, \qquad i = 1, \ldots, 6,$$

$$P(ltill = missing|sex = 1) = 14\%, P(ltill = missing|sex = 2) = 20\%.$$

According to the previous missing mechanism we have obtained a perturbed data set with the following percentage of missing items: sex 7.19%, age 10.27%, qualnum 13.76%, econprim 19.78%, ltill 23.72%. As a result, the reliability order among the variables is: sex, age, qualnum, econprim, ltill. The missing values have been imputed according to both the BN and MBBN imputation algorithms. The necessary BNs have been learned on the perturbed data set with the software Hugin and the estimated structures are reported in figure 1. The expected values of Δ and ξ have been approximated through a Monte Carlo experiment consisting of $t = 1, \ldots, 1000$ replications of the BN and MBBN imputation algorithm on the perturbed data set. The approximations are:

$$\tilde{\Delta} = \frac{1}{1000}\sum_{t=1}^{1000}\Delta_t, \qquad \tilde{\xi} = \frac{1}{1000}\sum_{t=1}^{1000}\xi_t.$$

As far as the indicator Δ is concerned, its expectation is approximately 0.105 for the BN imputation procedure and 0.073 for the MBBN one. Thus the MBBN procedure is preferable in terms of the preservation of the multivariate distribution. This improvement is due to a remarkable increase in the expected number of correct imputations, in terms of the indicator ξ. The results are reported in table 1. It is important to underline that the largest improvement is reported for those variables that are near the root of the Bayesian network considered for the BN imputation procedure (i.e. sex and age). This advantage is less remarkable for econprim, and it is not evident in qualnum and ltill. The performance improvement is due to the use of the Markov blanket. In particular, imputation of the more reliable variables (sex and age in our example) takes advantage (with the MBBN method) of the fact that it is done on the basis of the conditional distribution given the parents respectively in structure 1 and 2, figure 1, *i.e.* exploiting all the information contained in their Markov blanket. When the variable to impute is the least reliable (ltill in our case), there is no improvement since its parents set in the BN method is the same as its Markov blanket in the MBBN method. As far as the variable econprim is concerned, the slight improvement in $\tilde{\xi}$ is due to the fact that its Markov blanket includes only one additional variable (ltill) if compared to its parents in structure 6. Note that the parents of the variable sex in structure 1, figure 1 (i.e. the variables age, econprim and

	$\tilde{\xi}$	$\min_t \xi_t$	$\max_t \xi_t$
sex (BN)	51.01	47.22	54.25
sex (MBBN)	61.26	58.02	64.39
age (BN)	19.09	16.90	21.38
age (MBBN)	26.40	24.08	28.85
qualnum (BN)	78.72	77.01	80.05
qualnum (MBBN)	78.71	77.01	80.13
econprim (BN)	43.08	41.28	44.78
econprim (MBBN)	44.78	43.23	46.64
ltill (BN)	83.09	82.08	84.72
ltill (MBBN)	83.42	82.22	84.88

Table 1: Average ($\tilde{\xi}$), minimum and maximum values of ξ for the 1000 experiments.

qualnum) are exactly the Markov blanket of sex in the BN learnt respecting the reliability order, as shown in structure 6, figure 1. The same holds for the other variables: the parent set of the variable to impute coincides with the Markov blanket of that variable in structure 6, figure 1.

Furthermore, it seems that there is less variability for ξ_t on sex, age and (at a lower extent) econprim. In other words, the use of a more complete information makes variability due to the MBBN imputation algorithm less important.

Another remark is that the MBBN imputation algorithm seems not affected by a particular order among the variables (as the reliability order), while the BN one makes explicit use of it. Further investigations on this point may be required.

The improvements performed by the MBBN imputation algorithm are particularly interesting if compared to other imputation practices. In the first experiments [3] the BN procedure has been compared with some hot-deck imputation practices: the random hot-deck and some random stratified hot-deck procedures. In particular the BN imputation procedure improves the Δ indicator with respect to all stratifications. BN and hot-deck procedures have the same performance only when the stratification is the one explicitly considering the MAR mechanism. As shown in section 4 the MBBN behaves better than the BN imputation procedure; therefore we are led to the conclusion that the MBBN procedure is better than the different random hot-deck procedures considered in [3]. Additional comparisons with other imputation techniques that are currently applied in National Statistical Institutes are under study.

As a final remark, although the implementation of the MBBN imputation algorithm is more computationally intensive than the BN imputation algorithm, it allows a simpler automatisation of the imputation procedure (in particular if the order among the variables is not important). The implementation of the MBBN imputation algorithm is just k times slower than the BN one (where k is the number of variables). The missing item imputations have been done with an ad hoc C++ code that makes use of the output of the software Hugin.

References

[1] Cowell R.G., Dawid A.B., Lauritzen S.L., Spiegelhalter D.J. (1999). *Probabilistic networks and expert systems.* Springer Verlag, New York.

[2] Di Zio M., Scanu M., Vicard P. (2003). *Open problems and new perspectives for imputation using Bayesian networks.* Proceedings of the SCO2003 Conference, Treviso, September 4-6 2003, Treviso (Italy), 170 – 175.

[3] Di Zio M., Scanu M., Coppola L., Luzi O., Ponti P. (2004). *Bayesian networks for imputation.* Journal of the Royal Statistical Society, A, **167**(2), 309 – 322.

[4] Lauritzen S. (1995). *The EM algorithm for graphical association models with missing data.* Computational Statistics and Data Analysis **19**, 191–201.

[5] Madsen A.L., Lang M., Kjaerulff U.B., Jensen F. (2003) *The Hugin tool for learning Bayesian networks.* In T.D. Nielsen and N.L. Zhang (Eds) Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Proceedings of the 7th European Conference, ECSQARU 2003, Aalborg, Denmark, July 2003.

[6] Spirtes P., Glymour C., Scheines R. (2000). *Causation, prediction and search.* 2nd edition. MIT Press, Boston.

Acknowledgement: This work was partially supported by MIUR-cofin 2002, MIUR-cofin 2003 and EUREDIT project financial support from the IST programme of the European Union.

Address: M. Di Zio, G. Sacco, M. Scanu, Istituto Nazionale di Statistica, via Cesare Balbo 16, 00184 Roma, Italy
P. Vicard, Università Roma Tre, via Ostiense 139, 00154 Roma, Italy

E-mail: `dizio, sacco, scanu@istat.it, vicard@uniroma3.it`

COMPSTAT'2004 Symposium

EXTENDING PLS1 TO PLAD REGRESSION AND THE USE OF THE L1 NORM IN SOFT MODELLING

Yadolah Dodge, Athanassios Kondylis and Joe Whittaker

Key words: Soft modelling, partial least squares (PLS) regression, LAD regression, model selection.

COMPSTAT 2004 section: Partial least squares.

Abstract: Soft modelling approaches have not yet taken into account the use of the L_1 *norm* on model building and model selection in PLS setting. Partial Least Absolute Deviations (PLAD) regression, as illustrated here, introduces the use of the L_1 *norm* in PLS regression modelling. PLAD is then used in model selection and proves to be a fast and reliable alternative.

1 Introduction

Regression modelling has been affected recently by new technologies in instrumentation and data collection. In fact, modern instruments permit the recording of a considerable high number of often interrelated variables. PLS techniques in regression, as well as in other statistical fields, permit researchers to deal with the new standards.

PLS regression seeks to provide statistical models which focus in the reduction of the space spanned by the often large number of correlated predictors in a lower dimensional space generated by derived PLS components. It is commonly expected that a small number of derived components will be finally used in the PLS regression model as regressors.

PLS techniques received mostly the interest of chemometricians who implemented them in various settings (eg multivariate calibration and inverse regression). PLS finally received the attention of the statistical community and they now have a central role in statistical research. PLS properties and the new perspectives arising from their applications, either in classification or in regression setting, were the main interest of relatively recent work. In chemometrics setting one can see [5], and [7] while for a statistical point of view one can see [5] and [2]. For a global review of PLS, their properties, and their relation to other multivariate techniques one can see [6].

The following notations are used to describe the PLS1 (which equals to PLS on one response) and PLAD regression methods:

-k denotes the current iteration of the algorithm,
-$\mathbf{X_k}$ denotes the predictor's matrix $(\mathbf{x_{1,k}}, \mathbf{x_{2,k}}, \ldots, \mathbf{x_{p,k}})$ at iteration k,
-$\mathbf{y_k}$ denotes the response vector at iteration k,
-$\mathbf{q}$ denotes the regression parameter vector $(q_1, \ldots, q_k)$,

-$\mathbf{T_k}$ denotes the $n \times k$ matrix with each column corresponding to successive derived components or score vectors $(\mathbf{t_1}, \ldots, \mathbf{t_k})$,
-$\mathbf{w}$ denotes the $p \times k$ weight matrix (or loading's matrix) where, for each k, the loadings are calculated according to: $\mathbf{w_k} = \{cov(\mathbf{x_{1,k}}, \mathbf{y_k}), \ldots, cov(\mathbf{x_{p,k}}, \mathbf{y_k})\}$,
-$E(\mathbf{x_j})$ is the expected value for variable $\mathbf{x_j}$,
-$s_{\mathbf{x_j}}$ is the standard deviation for variable $\mathbf{x_j}$,
-$E_{LS}(\mathbf{y}|t_k)$ is the least square fit resulting from the regression of $\mathbf{y}$ on the k^{th} derived component $\mathbf{t_k}$, and $E_{LAD}(\mathbf{y}|t_k)$ the corresponding Least Absolute Deviations (LAD) fit of $\mathbf{y}$ regressed on $\mathbf{t_k}$,
-$E_{LAD}(\mathbf{y_k}|\mathbf{t_1}, \ldots, \mathbf{t_k}) = E_{LAD}(\mathbf{y_k}|\mathbf{T_k})$ is the LAD fit of $\mathbf{y}$ on the k mutually orthogonal derived components $(\mathbf{t_1}, \ldots, \mathbf{t_k})$.

2 A review of PLS1 regression

PLS1 regression builds a linear model using the derived PLS1 components instead of the original predictors. The k^{th} derived component is denoted as $\mathbf{t_k}$. Each component is a linear combination of the original predictors. That is,

$$\mathbf{t_i} = w_1\mathbf{x_{1i}} + w_2\mathbf{x_{2i}} + \cdots + w_p\mathbf{x_{pi}}$$

for each k, where $w_{jk} = cov(\mathbf{x_{jk}}, \mathbf{y_k})$ (for the commonly used case of standardized data), while $j = 1, 2, \ldots, p$.

The PLS1 regression model can be written as:

$$E_{LS}(\mathbf{y_k}|\mathbf{t_1}, \ldots, \mathbf{t_k}) = \widehat{\mathbf{y}}_\mathbf{k} = \mathbf{T_k}\,\widehat{\mathbf{q}}_\mathbf{k},$$

or in an equivalent way:

$$\widehat{\mathbf{y}}_\mathbf{k} = \widehat{\mathbf{y}}_\mathbf{k-1} + \widehat{q}_k\,\mathbf{t_k}$$

where $\widehat{\mathbf{y}}_\mathbf{k}$ denotes the predicted response vector for a final model with k components, $\widehat{\mathbf{q}}_\mathbf{k} = (\widehat{q_1}, \widehat{q_2}, \ldots, \widehat{q_k})^T$ denotes the estimated regression coefficient vector for a model containing k components, and $\mathbf{T_k} = (\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_k})$ represents the k-PLS1 score matrix which is regressed oxn the target $\mathbf{y}$.

Since the value of k represents the number of PLS1 iterations in order to extract new components. The smaller the k the better the dimension reduction achieved by the use of $\mathbf{t's}$ instead of the original predictors $\mathbf{x_j}$. After k is fixed (using model selection criteria), the regression results are transformed back in terms of the original predictors. Then, PLS1 regression model can predict the response value for any given set of future observations. The final model is then selected. Cross-validation, the bootstrap, and penalized maximum likelihood are commonly used for model selection.

Amongst numerous good points of PLS1 regression the following are worth to notice:
-PLS1 regression uses orthogonal derived components $\mathbf{t_k}$ as regressors and reduces significantly the dimension of the regression problem, essential when p is large.

-PLS1 components satisfy: $arg\ max_{\alpha}\ \{corr^2(\mathbf{X}\alpha, \mathbf{y})var(\mathbf{X}\alpha)\}$ subject to $\alpha'\alpha = 1$, and forms a compromise between principal components regression and multiple linear regression.
-PLS1 regression estimates yield shrinkage properties.

3 PLAD regression

PLAD regression algorithm retains orthogonality and shrinkage properties of PLS1 while it regresses its derived components on the response $\mathbf{y}$ using LAD regression. PLAD algorithm extracts components' loadings using covariance estimate according to:

$$\mathbf{w} = cov_{MAD}(\mathbf{x_j}, \mathbf{y}) = \frac{1}{4}(MAD(\mathbf{x_j} + \mathbf{y}) - MAD(\mathbf{x_j} - \mathbf{y})),$$

where the abbreviation MAD is the *median absolute deviation* calculated according to: $MAD(y) = med(|\mathbf{y} - med(\mathbf{y})|)$. In this way PLAD scores are extracted using the median as the center location parameter (instead of the mean of PLS1), and the MAD estimate of dispersion. PLAD scores are thus expected to be resistant to outliers which may mislead the direction of PLS1 scores.

The fundamental feature in PLAD algorithm is the use of LAD estimates in regressing the derived components on the response. LAD regression has a long history, originated in the 18^{th} century when it was introduced by Roger Joseph Boscovich (1757). LAD minimizes the sum of the absolute values of the residuals ($\min_{\beta} \sum_{i=1}^{n} |r_i|$, for r_i the i^{th} residual). LAD is often referred as *regression to the median* or $L_1 regression$, which clearly distinguishes LAD regression from LS *regression to the mean.* For computational and inferential aspects on LAD one can see [1]. Below the algorithm of PLAD is schematically presented.

PLAD algorithm

- Standardize all variables so as to have zero center and unit length. Use, thus as $\mathbf{x_j}$ (columns of matrix $\mathbf{X}$) and response $\mathbf{y}$ their standardized values $\frac{\mathbf{x_{ij}} - E(\mathbf{x_j})}{s_{\mathbf{x_j}}}$ and $\frac{\mathbf{y_i} - E(\mathbf{y})}{s_{\mathbf{y}}}$ respectively, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.
  $k \leftarrow 1$
  $\mathbf{X_0} \leftarrow \mathbf{X}$, $\mathbf{y_0} \leftarrow \mathbf{y}$.

Step 1: Calculate vector $\mathbf{w_{k-1}} = cov_{MAD}(\mathbf{x_{j,k-1}}, \mathbf{y_{k-1}})$ and normalize in order to have $\mathbf{w}'\mathbf{w} = 1$.

Step 2: Extract $\mathbf{t_k} = \mathbf{X_{k-1}} \mathbf{w_{k-1}}$.

Step 3: α. Regress $\mathbf{y_{k-1}}$ to $\mathbf{t_k}$, calculate the residuals

$$\mathbf{y_k} = \mathbf{y_{k-1}} - E_{LS}(\mathbf{y_{k-1}} | \mathbf{t_k})$$

β. Regress each $\mathbf{x_{j,k-1}}$ to $\mathbf{t_k}$, calculate for each one the residuals

$$\mathbf{x_{jk}} = \mathbf{x_{j,k-1}} - E_{LS}(\mathbf{x_{j,k-1}}|\mathbf{t_k}) \quad \text{where} \quad j = 1, 2, .., p.$$

Step4: Regress the initial target $\mathbf{y}$ on the extracted components (PLAD regression step).
Using LAD regression, give the model $\widehat{\mathbf{y}} = E_{LAD}(\mathbf{y}|\mathbf{T_k})$.

- If a certain stopping criterion (to be discussed in what follows) is satisfied, give final PLAD regression model $\widehat{\mathbf{y}} = E_{LAD}(\mathbf{y}|\mathbf{T_k})$. Otherwise, return to Step 1 with $k \leftarrow k + 1$.

Model Selection

By model selection we mean the process of selecting the number of components which will be retained in the final model. The latter guarantees the dimension reduction by the use of a certain regression method. The comparison of the reduction performance between different methods and algorithms is very important in practice. As already seen cross-validation, the bootstrap, and penalized maximum likelihood criteria are commonly used in order to select the final model.

In standard PLS1 analysis cross-validation is commonly used in order to select the final model. Cross-validation is generally based on the idea to split the data set into training set (model development) and test set (model validation). For each split a loss function is computed while the model is cross-validated for various number of components (here $\mathbf{t_k}$).

The Root Mean Squared Error (RMSE) is a widely used loss function. The RMSE for a regression model containing k components (denoted as $RMSE_k$), is given by:

$$RMSE_k = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \widehat{y}_{ik})^2},$$

and is computed τ times, where τ is the number of times that data are split. For example, on leave-one-out cross validation τ equals n, the total number of observations (not to be confused with n_{test} which corresponds to the number of testing units in a given split, eg. on leave-one-out cross validation $n_{test} = 1$). The model for which the expected $RMSE_k$, as an average over the τ distinct splits, reaches its minimum, is finally selected. It indicates the number of components to be retained, that is k.

The use of the L_1 *norm* can be also extended to model selection. The median of the absolute deviations of each y_i from $\widehat{y}_{ik}$ replaces the squared loss of the $RMSE_k$. Thus, we select the final model which minimises the *Median Absolute Error*

$$MAE_k = med(|\mathbf{y} - \widehat{\mathbf{y}}_k|).$$

Both criteria will be used in the examples that follow with interest being focused on the *Median Absolute Error*.

4 Examples

The examples that follow are chosen in order to illustrate PLAD regression and compare it to PLS1 in settings when collinearity problems rise in the predictors space, and the sample units are not strongly larger than the number of the predictors. Octane data set, for example, includes a large number of recorded predictors for a limited number of observations. The latter is usually the case in NIR spectral analysis from which octane data are collected.

PLS1 and PLAD regression algorithms have been implemented in **S-Plus 6.1** and extended to include model selection techniques such as cross-validation.

- Diabetes Data

  The diabetes data consist of 442 diabetes patients measured on 10 baseline variables. These variables are: age, sex, body mass index, average blood pressure, and six blood serum measurements. Baselines were registered for each of the 442 diabetes patients. The response of interest was a quantitative measure of disease progression one year after. A prediction model for the response was built using PLS1 and PLAD. The first 150 sampling units constitue the test set for the model built on cases 151 to 442. More or less two thirds of the data set was used for training and one third for testing. The resulting loss for both squared and absolute loss functions is given in Table 1.

Loss	Algorithm	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$RMSE_k$	PLS1	0.8628	0.8122	0.8113	0.8114
	PLAD	0.8300	0.7978	0.7979	0.7976
MAE_k	PLS1	0.6238	0.5768	0.5804	0.5808
	PLAD	0.5855	0.5664	0.5662	0.5659

Table 1: Diabetes Data Set. $RMSE_k$ and MAE_k loss for both PLS1 and PLAD regression models.

  Both algorithms select two components in the final model. The addition of more than two components would just over-fit the final model. Moreover, for both models the expected loss from either the RMSE or MAE given $k = 2$.

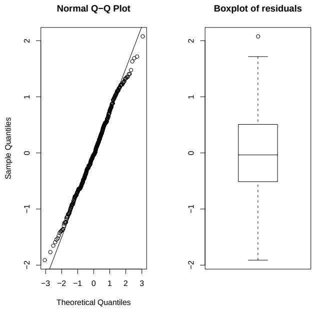


Figure 1: Diabetes data.

In Figure 1 the boxplot and the normal Q-Q plot of the residuals are given. They show symmetrically normally distributed residuals in a big data set. No strong outliers appear and both regression models are almost equal. The implied regression coefficients $\widehat{\beta}_{impl}$ (not shown here) almost coincide. In such cases PLS1 and PLAD regression models share the same statistical properties such as shrinkage and dimension reduction, and provide similar results.

- Octane Data

Octane data set arises from NIR experiments where PLS regression is widely applied. Octane data have been analyzed in [6] as well as in [3], so previous knowledge will be helpful to validate PLAD regression results. The data consists of 39 gasoline samples for which the octanes have been measured in 225 wavelengths (measured in nm). For wavelengths over 150nm outlying samples are detected. These samples contain alcohol. PLS1 and PLAD algorithms are implemented for octane data and cross-validated MAE_k are recorded. The MAE_k for the different splits used in cross-validation are averaged in order to obtain an appropriate aggregated measure of loss. The results of the routine are given in Table 2. Figure 2a and 2b plot the first PLS1 and PLAD components ($\mathbf{t_1}$) versus the response ($\mathbf{y}$). The dotted line in both panels sketch the resulting line after regressing $\mathbf{y}$ on $\mathbf{t_1}$. PLS1 plot is on the left panel, and PLAD on the right.

Figures 2a and 2b illustrate two main points. Firstly, they illustrate that PLS1 is sensitive to outliers. Regressing the response on derived components, instead of the original predictors, does not change the negative effects of the outliers on the regression. PLAD regression algorithm, in contrast to PLS1, is not affected by the outliers. This is mainly due to the use of the LAD. Secondly, from the above figure one can expect that PLS1 algorithm looses a run-iteration, since the regression results for $k = 1$ are badly influenced by the outliers.

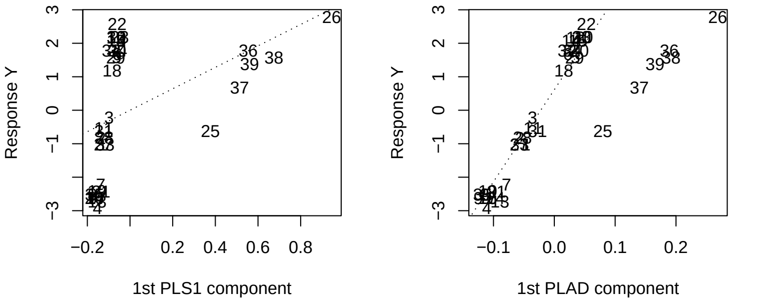


Figure 2: (a) *left panel:* Plot of 1st PLS1 component versus the response. PLS1 regression totally confused by the outliers. (b) *right panel:* Plot of 1st PLAD component versus the response. PLAD is not affected by the outliers.

Our beliefs are justified by the use of cross-validation in order to select the number of components to be finally retained. The results are given in Table 2. They show the effect of octane's outliers on model selection. PLAD regression model is built using two components since for $k = 2$ the loss function is minimized. PLS1 regression model is deceived by the six outlying values in octane data set, and needs at least an additional run (an additional iteration) in order to build the model with the minimum loss.

Risk	Algorithm	$k = 1$	$k = 2$	$k = 3$	$k = 4$
MAE_k	PLS1	1.8108	1.5032	1.3826	1.3553
	PLAD	1.5793	0.7311	0.8181	1.1244

Table 1: Octane Loss for PLS1 and PLAD regression algorithms.

5 Conclusions

PLAD retains PLS structure while it focuses on the use of the L_1 *norm* in model building and model selection. PLAD algorithm still benefits from PLS statistical properties as seen in section 2. As a conclusion, it is worth noticing a few fundamental arguments in favour of PLAD as they were seen throughout the examples. LAD estimates are well-suited to longer-tailed or asymmetric error distributions, something which is very common when small data sets contain outliers. PLAD estimates are not affected by the outliers and PLAD model selection is straight-forward.

PLAD regression algorithm can be applied to a big number of cases rising within soft modelling setting. We consider for example modelling data collected from microarray or NIR instruments, where hundreds of observed variables are being recorded for moderate sample sizes. In this context PLAD proves to be an effective solution as well as a promising perspective in soft modelling setting.

References

[1] Dodge Y., Jureckova J. (2002). *Adaptive regression.* Springer Verlag, Berlin.

[2] Frank I., Friedman J. (1993). *A statistical view of some chemometrics regression tools.* Technometrics **35**, 109 – 135.

[3] Hubert M., Vanden Branden K. (2003). *Robust methods for partial least squares regression.* Journal of Chemometrics, to appear.

[4] Martens H., Naes T. (1989). *Multivariate calibration.* Wiley, UK.

[5] Brooks R., Stone M. (1994). *Joint continuum regression for multiple predictands.* J. Amer. Statist. Assoc. **89**, 1374 – 1377.

[6] Tenenhaus M. (1998). *La régression PLS. Théorie et pratique.* Technip, Paris.

[7] Wold S., Sjostrom M., Eriksson L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems **58**, 109 – 130.

Address: Y. Dodge, A. Kondylis, Statistics Group, University of Neuchâtel, Espace de l'Europe 4, 2002, Neuchâtel, Switzerland
J. Whittaker, Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, England

E-mail: yadolah.dodge@unine.ch, atanassios.kondylis@unine.ch, joe.whittaker@lancaster.ac.uk

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

MINIMUM DISTANCE INFERENCE FOR SUNDT'S DISTRIBUTION

Louis G. Doray and A. Haziza

Key words: Statistical computing, numerical algorithm, simulations, iterative reweighted least-squares.

COMPSTAT 2004 section: Algorithms.

Abstract: The probability function of a discrete distribution belonging to Sundt's family satisfies a certain recursive relationship of order k. Maximum likelihood estimation of its parameters is difficult since there is no closed-form expression for the probability function. We propose an alternative method to estimate the parameters, based on the construction of a linear model and the minimization of a quadratic distance. The asymptotic properties of these estimators are investigated: asymptotic normality of their distribution, unbiasedness, efficiency.

The quadratic distance estimator (QDE) of the parameters can be calculated by using an iteratively reweighted least-squares algorithm. With simulated data from Sundt's family, we show how to implement this algorithm.

Another advantage of the minimum quadratic distance is that we can construct a test statistic easily computable with the QDE and derive its asymptotic distribution. This enables us to test a simple hypothesis for the parameter values as well as a composite hypothesis leading to a goodness-of-fit test.

1 Properties of Sundt's distribution

Sundt [9] has introduced the following family of discrete distributions. A discrete random variable N, taking non-negative values, belongs to Sundt's family if its probability function satisfies the following recursive relationship of order k

$$p_n = \sum_{i=1}^{k}(a_i + b_i/n)p_{n-i}, \quad k \geq 1, \quad p_{-1}, p_{-2}, \ldots = 0, \tag{1}$$

where $a = (a_1, \ldots, a_k)$ and $b = (b_1, \ldots, b_k)$ are parameter vectors of the distribution such that (1) defines a probability function, parameters that we want to estimate. We will denote the distribution of N by $\mathcal{R}_k[a, b]$. Let $\Psi(s)$ be the probability generating function (pgf) of N. Sundt [9] has shown that $N \in \mathcal{R}$ if and only if $\frac{d}{ds} \ln \Psi(s)$ can be written as the ratio of two polynomials, the one in the numerator being of degree $\leq k-1$ and the one in the denominator of degree $\leq k$ and with a constant term equal to 1.

Some well-known distributions such as the binomial, Poisson and negative binomial distributions belong to this family with $k = 1$. See Panjer [7] for the values of the parameters a and b of these distributions, composing Panjer's family. By using the pgf, Sundt [9] has shown the following results:

1. the sum of two independent random variables $\mathcal{R}_k[a, b]$ and $\mathcal{R}_l[c, d]$ also follows Sundt's distribution, but of order $k + l$.
2. the sum of two independent random variables $\mathcal{R}_k[a, b]$ and $\mathcal{R}_k[a, d]$ follows a $\mathcal{R}_k[a, e]$ distribution, where $e_i = ia_i + b_i + d_i$, for $i = 1, \ldots, k$.
3. the distribution of the m^{th} convolution of independent and identically distributed random variables $\mathcal{R}_k[a, b]$ follows a $\mathcal{R}_k[a, \beta]$ distribution where $\beta_i = (m - 1)ia_i + mb_i$, for $i = 1, \ldots, k$.
4. the convolution of two independent random variables $\mathcal{R}_1[a_1, b_1]$ and $\mathcal{R}_1[a_2, b_2]$ follows a $\mathcal{R}_2[(a_1 + a_2, a_1a_2), (b_1 + b_2, -(a_1b_2 + a_2b_1))]$ distribution.

2 Minimum quadratic distance estimation

The maximum likelihood estimates of the parameters for Sundt's family of order k are difficult to compute, because the roots of a polynomial of high degree need to be found and local maxima must be distinguished from the global maximum. For $k = 1$, which is Panjer's family, Luong and Garrido [6] have shown how to use the recursive relationship to estimate the parameters of the distribution. We will generalize their method to arbirary k. Let us first define the truncated $\mathcal{R}_k$ family.

Definition. A discrete random variable N with domain $0, 1, \ldots, w$ belongs to Sundt's truncated family of finite order k if its probability function satisfies the following recursive equation

$$p_n^* = \sum_{i=1}^{k}(a_i + b_i/n)p_{n-i}^*, \quad p_{-1}^* = \ldots = p_{-(k-1)}^* = 0, \quad n = 1, \ldots, w.$$

The theoretical probabilities p_n^* are estimated with the observed frequencies from the sample,

$$\hat{p}_n^* = \frac{f_n}{m}, \quad n = 0, \ldots, w,$$

where f_i is the number of observations equal to i in the sample of size m. That recursive equation, linear in the parameters $a_1, \ldots, a_k, b_1, \ldots b_k$, suggests the following linear regression model

$$\hat{p}_n^* = \sum_{i=1}^{k}(a_i + b_i/n)\hat{p}_{n-i}^* + \epsilon_n, \quad \hat{p}_{-1}^* = \ldots = \hat{p}_{-(k-1)}^* = 0, \quad n = 1, \ldots, w,$$

where ϵ_n is a random error.

Let us define the two vectors $\hat{Y} = (\hat{p}_1^*, \hat{p}_2^*, \ldots, \hat{p}_w^*)'$, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_w)'$ and matrix $\hat{X} =$

$$\begin{pmatrix} \hat{p}_0^* & \hat{p}_0^* & 0 & 0 & \ldots & 0 & 0 \\ \hat{p}_1^* & \hat{p}_1^*/2 & \hat{p}_0^* & \hat{p}_0^*/2 & \ldots & 0 & 0 \\ \hat{p}_2^* & \hat{p}_2^*/3 & \hat{p}_1^* & \hat{p}_1^*/3 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{p}_{w-1}^* & \hat{p}_{w-1}^*/w & \hat{p}_{w-2}^* & \hat{p}_{w-2}^*/w & \ldots & \hat{p}_{w-k}^* & \hat{p}_{w-k}^*/w \end{pmatrix}.$$

In matrix notation, this model can be rewritten as $\hat{Y} = \hat{X}\theta + \epsilon$, where θ is the parameter vector $(a_1, b_1, \ldots, a_k, b_k)'$. If $w > k$, matrix $\hat{X}$ is of full rank with probability 1, as $m \to \infty$, and it tends in probability to its theoretical part, denoted X. We also have $E(\hat{X}) = X$.

Since f_i follows a binomial (m, p_i^*) distribution, it can easily be shown that $E(\epsilon_i) = 0$ for $i = 1, \ldots, w$. Using the fact that $(f_i, f_j), i \neq j$, has a trinomial (m, p_i^*, p_j^*) distribution, $\text{Var}(\epsilon_i)$ and $\text{Cov}(\epsilon_i, \epsilon_j)$ can be obtained after some tedious calculation. Let us denote by Σ_θ the variance-covariance matrix of vector ϵ, and let us define $\Sigma_\theta^* = m\Sigma_\theta$ (see [5] for the terms of this matrix).

The minimum quadratic distance estimator (MQDE) of vector θ is the vector value which minimizes the expression $\epsilon'\Sigma_\theta^{*-1}\epsilon$, the solution of which is given by

$$\hat{\theta} = (\hat{X}'\Sigma_\theta^{*-1}\hat{X})^{-1}\hat{X}'\Sigma_\theta^{*-1}\hat{Y};$$

this is not an estimator in the usual sense, since Σ_θ^* is a function of the unknown vector θ. For that reason, an iteratively reweighted least-squares algorithm must be used:
Step 0: Set $i = 0$ and $\hat{\Sigma}_{\hat{\theta}_0}^* = I_w$, where I_w is the identity matrix of dimension w.
Step 1: Compute $\hat{\theta}_{i+1} = (\hat{X}'\hat{\Sigma}_{\hat{\theta}_i}^{*-1}\hat{X})^{-1}\hat{X}'\hat{\Sigma}_{\hat{\theta}_i}^{*-1}\hat{Y}$.
Step 2: Recalculate $\hat{\Sigma}_{\hat{\theta}_{i+1}}^*$
Step 3: Set $i \leftarrow i + 1$.
Go back to step 1 until convergence is attained.

Luong and Garrido [6] have shown, for $k = 1$, that $\hat{\theta}_i \xrightarrow{p} \theta$ and $\hat{\Sigma}_{\hat{\theta}_i}^{-1} \xrightarrow{p} \Sigma_\theta^{-1}, i = 1, 2, \ldots$, where $\xrightarrow{p}$ denotes convergence in probability. The proof remains valid for $k > 1$. For any value of k, Haziza [4] has shown the following results:

1. Vector ϵ has an asymptotic normal distribution $N(0, \Sigma_\theta)$, from which it follows that $\sqrt{m}\hat{\theta} \xrightarrow{\mathcal{L}} N(\theta, (X'\Sigma_\theta^{*-1}X)^{-1})$.
2. If $||(X'\Sigma_\theta^{*-1}X)^{-1})|| < \infty$, $\hat{\theta}$ is a consistent estimator of θ.
3. $\hat{\theta}$ is an aymptotically efficient estimator of θ.

Until now, we have assumed that the observations were coming from a truncated $\mathcal{R}_k$ family. If we assume instead that the sample comes from the $\mathcal{R}_k$ family, i.e. we let w tend to infinity, the consistency, asymptotic normality and asymptotic efficiency of $\hat{\theta}$ will remain valid. In practice, we can set $w = A$, for a large value of $A > 0$, and assume that any observation larger than A is an outlier, which is rejected. In this case, the aymptotic efficiency of $\hat{\theta}$ does not hold, but the two other properties still hold. The value of A should be chosen large enough to have the largest possible asymptotic effiency of $\hat{\theta}$.

In the truncated $\mathcal{R}_2$ family, Haziza [4] has also shown that the MQDE of θ has the robustness property, meaning that that the influence function is bounded (see [2] for an exhaustive presentation on the theory of robustness and [3] for an intuitive treatment of influence curves).

3 Tests of hypothesis

In this section, we will generalize tests considered by Doray and Huard [1] to distinguish between the Poisson and negative binomial distributions to tests applicable to Sundt's family.

Let us assume that the observed data $n_1, \ldots, n_m$ come from the $\mathcal{R}_k$ family truncated at w. To test the null hypothesis that the sample arose from that distribution with parameter vector $\theta_0 = (a_1^0, b_1^0, \ldots, a_k^0, b_k^0)'$, with all the parameter values specified, we calculate the following distance between the empirical and parametric cdf

$$d(F_m, F_{\theta_0}) = m(\hat{Y} - \hat{X}\theta_0)'\Sigma_{\theta_0}^{*-1}(\hat{Y} - \hat{X}\theta_0).$$

Haziza [4] has shown that, under H_0, as the sample size $m \to \infty$, the asymptotic distribution of $d(F_m, F_{\theta_0})$ is χ_w^2. We will therefore reject the null hypothesis H_0 at the approximate level α if $d(F_m, F_{\theta_0}) > \chi_{w;1-\alpha}^2$ where $\chi_{w;1-\alpha}^2$ is the $100(1-\alpha)$ percentile of a χ_w^2 distribution.

Suppose, on the other hand, we want to test the null hypothesis H_0 specifying that the data come from a truncated $\mathcal{R}_l$ family, where $l < k$ and l is a positive integer. Let H_0 be

$$H_0 : \theta = (a_1, b_1, \ldots, a_l, b_l, a_{l+1} = 0, b_{l+1} = 0, \ldots, a_k = 0, b_k = 0)',$$

where the first $2l$ parameters of H_0 are unknown.

Let $\tilde{\theta} = (\tilde{a}_1, \tilde{b}_1, \ldots, \tilde{a}_l, \tilde{b}_l, 0, \ldots, 0)'$ be the MQDE of θ obtained by initially setting $a_{l+1} = b_{l+1} = \cdots = a_k = b_k = 0$ and minimizing the distance

$$m(\hat{Y} - \hat{X}\theta)'\Sigma_{\theta}^{*-1}(\hat{Y} - \hat{X}\theta).$$

Under H_0, the distance

$$d(F_m, F_{\tilde{\theta}}) = m(\hat{Y} - \hat{X}\tilde{\theta})'\Sigma_{\tilde{\theta}}^{*-1}(\hat{Y} - \hat{X}\tilde{\theta})$$

follows an asymptotic χ^2 distribution with $w-l$ degrees of freedom (see [4] for the proof, based on the fact that $\Sigma^*_\theta \Sigma^{*-1}_{\tilde{\theta}}$ is an idempotent matrix with trace equal to $(w-l)$).

The test will therefore consist in rejecting H_0 at level α if $d(F_m, F_{\tilde{\theta}}) > \chi^2_{w-l;1-\alpha}$ where $\chi^2_{w-l;1-\alpha}$ is the $100(1-\alpha)$ percentile of a χ^2_{w-l} distribution.

4 Numerical example

In this section, we will illustrate with simulated data, the method developed to calculate the MQDE of the parameter vector for a special case of Sundt's family.

Let us consider the distribution obtained by truncating the convolution of a Poisson distribution with $\lambda = 2$ and a negative binomial distribution with $s = 2$ and $q = 2$. The domain of this truncated distribution is the set $\{0, 1, \dots, 8\}$ and its theoretical probabilities are given in Table 1. The sample size was set at $m = 15003$ and the observed frequencies also appear in Table 1.

n	p^*_n	f_n
0	0.0191	279
1	00636.	986
2	0.1145	1691
3	0.1499	2229
4	0.1616	2408
5	0.1540	2357
6	0.1352	1973
7	0.1123	1730
8	0.0898	1350

Table 1: Results of the simulation.

It is well known (see [8]) that the probability function of the above convolution will satisfy the recurrence equation

$$p^*_n = (a + b/n)p^*_{n-1} + (c/n)p^*_{n-2}, \quad p^*_{-1} = 0, \quad n = 1, \dots, 8.$$

Schröter's family is a special case of Sundt's family with $k = 2$ and parameter b_2 set equal to 0.

We obtain the regression model $\hat{Y} = \hat{X}\theta + \epsilon$, where

$$\begin{aligned} \hat{Y} &= (0.066, 0.113, 0.149, 0.161, 0.157.0.132, 0.115, 0.090)', \\ \theta &= (a, b, c)' \end{aligned}$$

and

$$\hat{X} = \begin{pmatrix} 0.019 & 0.019 & 0 \\ 0.066 & 0.033 & 0.009 \\ 0.113 & 0.038 & 0.022 \\ 0.149 & 0.037 & 0.028 \\ 0.161 & 0.032 & 0.030 \\ 0.157 & 0.026 & 0.027 \\ 0.132 & 0.019 & 0.022 \\ 0.115 & 0.014 & 0.016 \end{pmatrix}.$$

By setting $\hat{\Sigma}^*_{\hat{\theta}_0} = I_8$, we calculate, from step 1 of the algorithm in Section 2, a first estimate for θ, $\hat{\theta}_1 = (0.556, 2.647, -0.677)'$. Applying the iterative algorithm, convergence was attained after only 5 iterations; the MQDE is equal to $\hat{\theta} = (0.641, 2.624, -1.127)'$. The estimated variance-covariance matrix of the parameters is equal to

$$\mathrm{Var}(\hat{\theta}) = \begin{pmatrix} 0.0041 & 0.0021 & -0.0258 \\ 0.0021 & 0.0012 & -0.0135 \\ -0.0258 & -0.0135 & 0.1607 \end{pmatrix}.$$

We can test that the parameter c is significantly different from 0 in the model; the approximate 95% confidence interval for c is $[-1.786, -0.468]$.

References

[1] Doray L.G., Huard L. (2001). *On some new goodness-of-fit tests for the poisson distribution.* In: New Trends in Statistical Modelling, Proceedings of the 16^{th} International Workshop on Statistical Modelling, B. Klein and L. Korsholm (eds), Odense, Denmark, 429 – 435.

[2] Hampel F.R. (1974). *The influence curve and its role in robust estimation.* Journal of the American Statistical Association **69**, 383 – 393.

[3] Hampel F.R., Rochetti E.M., Rousseuw P.J., Stahel W.A. (1986). *Robust statistics: The approach based on influence functions.* Wiley: New York.

[4] Haziza A. (1997). *Inférence pour la famille de Sundt.* Mémoire de maîtrise, Département de mathématiques et de statistique, Université de Montréal, 77 p.

[5] Luong A., Doray L.G. (2002). *General quadratic distance methods for discrete distributions definable recursively.* Insurance: Mathematics and Economics **30**, 255 – 267.

[6] Luong A., Garrido J. (1993). *Minimum quadratic distance estimation for a parametric family of discrete distributions defined recursively.* Australian Journal of Statistics **35**, 59 – 67.

[7] Panjer H.H. (1981). *Recursive evaluation of a family of compound distributions.* Astin Bulletin **12**, 22 – 26.

[8] Schröter K.J. (1990). *On a family of counting distributions and recursions for related compound distributions*. Scandinavian Actuarial Journal, 161 – 175.

[9] Sundt B. (1992). *On some extensions of Panjer's class of counting distributions*. Astin Bulletin **22**, 61 – 80.

Acknowledgement: The authors gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada and of the Insurance Chair of Laval University.

Address: L.G. Doray, A. Haziza, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montréal, Qc, Canada, H3C 3J7

E-mail: `doray@dms.umontreal.ca`

COMPSTAT'2004 Symposium

OPTIMAL 2^2 FACTORIAL DESIGNS FOR BINARY RESPONSE DATA

Roberto Dorta-Guerra and Enrique González-Dávila

Key words: D-optimal design, logistic model, probit model.

COMPSTAT 2004 section: Design of experiments.

Abstract: In this contribution, we compute D-optimal 2^2 designs for binary responses, under first order generalized linear models, and we use them to illustrate what is different relative to the typical continuous response case. In a follow up extended paper, we consider the general 2^{k-p} case, and cover both the additive model as well as the model with interactions.

1 Introduction

Two-level factorial experiments are very useful in the early screening stages of the investigations and as building blocks for response surface exploration, the most frequent scenarios faced in industrial experimentation practice. Most often, the response of interest can be modelled through first order normal linear homochedastic regression models, and in that case, two-level factorial experiments are either optimal or close to optimal among all experiments with the same sample size, for a broad class of experimental regions and for most sensible design optimality criterion, including the determinant of the information matrix, (see e.g. [16] or [6]).

Under that normal linear model, the determinant of the information matrix satisfies a series of properties that make the choice of the best two-level factorial design under that criterion a non-issue, and that allows one to restrict consideration to the coded factor levels -1 and $+1$. In particular, that determinant does neither depend on where the factorial experiment is centered, nor on how it is oriented relative to the contour lines of the surface, and increasing the number of factors or including interaction terms does not lead to an alternative choice of a factorial experiment. Furthermore, the balanced allocation that assigns the same number of replicates to all factor combinations is always better than unbalanced allocations with the same total number of runs, (see e.g. [4]).

As a consequence, when planning for two-level factorial experiments in this linear normal setting, the only thing that matters is the range of variation of the factors involved; the larger that range, the larger the determinant of the information matrix, and the more informative the experiment.

Two-level factorial experiments are also very popular with discrete responses, (see e.g., [3], [7], [14], [21], and [15]), but the design issues involved are a lot more complicated than for continuous responses, because none or almost none of the properties listed above do hold anymore.

2 Generalized linear models for binary response

In a binary response experiment with two design variables, n_i subjects are administered dose levels $x_i = (x_{1i}, x_{2i})$, for $i = 1, \dots, q$, and the outcome is binary. Usually, the total number of subjects, $n = \sum_{i=1}^{q} n_i$, is specified, and one assumes that the number of successes on the n_i subjects at x_i, y_i, are conditionally independent binomial random variables, $y_i | x_i, \beta \sim \text{Binomial}(n_i, p(x_i; \beta))$. Here x_{1i} and x_{2i} are assumed continuous, and

$$p(x_i; \beta) = F(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) = F(z_i), \tag{1}$$

where $F(.)$ is a known cumulative distribution function, and $\beta = (\beta_0, \beta_1, \beta_2)$. That is, we assume that there exists a known link function, $F^{-1}(.)$, such that $z_i = F^{-1}(p(x_i; \beta)) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ is an unknown linear combination of $x_i = (x_{1i}, x_{2i})$, (see e.g. [11].

When (1) holds, the contour levels of $p(x; \beta)$ are straight lines, and one states that $x = (x_1, x_2)$ belongs to the $ED_{F(z)}$ line with $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, that is the set of design points with $p(x; \beta)$ equal to $F(z)$. Under this model, the Fisher information matrix for $\beta = (\beta_0, \beta_1, \beta_2)$ is:

$$I(\beta) = n \begin{pmatrix} \sum_i h(z_i, \lambda_i) & \sum_i x_{1i} h(z_i, \lambda_i) & \sum_i x_{2i} h(z_i, \lambda_i) \\ \sum_i x_{1i} h(z_i, \lambda_i) & \sum_i x_{1i}^2 h(z_i, \lambda_i) & \sum_i x_{1i} x_{2i} h(z_i, \lambda_i) \\ \sum_i x_{2i} h(z_i, \lambda_i) & \sum_i x_{1i} x_{2i} h(z_i, \lambda_i) & \sum_i x_{2i}^2 h(z_i, \lambda_i) \end{pmatrix}, \tag{2}$$

where $\lambda_i = n_i/n$ and $h(z_i, \lambda_i) = \lambda_i (F^{'}(z_i))^2 / (F(z_i)(1 - F(z_i)))$. Table 1 lists the most popular special cases of the framework considered here.

Model	$z_i = F^{-1}(p(x_i; \beta))$	$h(z_i, \lambda_i)$
Logistic	$\log(p(x_i; \beta)/(1 - p(x_i; \beta)))$	$\lambda_i e^{z_i}/(1 + e^{z_i})^2$
Probit	$\Phi^{-1}(p(x_i; \beta))$	$\lambda_i e^{-z_i^2}/(2\pi\Phi(z_i)\Phi(-z_i))$
Log-log	$-\log(-\log(p(x_i; \beta)))$	$\lambda_i e^{-2z_i}/(e^{e^{-z_i}} - 1)$
C. Log-log	$\log(-\log(1 - p(x_i; \beta)))$	$\lambda_i e^{2z_i}/(e^{e^{z_i}} - 1)$

Table 1: Link function and $h(z_i, \lambda_i)$ for the logistic, probit, log-log and complementary log-log models; $\Phi(.)$ is the cdf of the standard normal.

3 Design optimality criteria

The inverse of $I(\beta)$ is the asymptotic variance-covariance matrix of the maximum likelihood estimate of β, and it relates to the size and shape of the approximate confidence regions for β built based on these estimates. Other than for very special cases though, one can not compare experiments through the whole matrix $I(\beta)$, because matrices are not totally ordered. When the goal is to estimate any one of the individual components in β, or any real valued function of them, one typically chooses the q and the (x_i, λ_i)'s that minimize the asymptotic variance of the maximum likelihood estimate for that component, (see e.g. [20] or [10].

Instead, when the goal is to jointly estimate all the components of β, one typically chooses the q and (x_i, λ_i)'s that maximize the determinant of $I(\beta)$, and thus minimize the volume of the asymptotic confidence ellipsoids for β, obtaining what are called the D-optimal designs. The determinant of $I(\beta)$ is the default criteria, when one lacks detailed specifications about the goal of the experimenter. One appealing feature of this criteria, is the fact that it induces an ordering of experiments that is invariant under reparametrizations, (see e.g. [2] or [16].

In our binary response setting, the determinant of $I(\beta)$ is:

$$det(I(\beta)) = n^3 det(\sum_i h(z_i, \lambda_i)(1, x_{1i}, x_{2i})'(1, x_{1i}, x_{2i})), \tag{3}$$

and it depends on the unknown value of β through z_i. In the local D-optimal approach, one guesses β to be equal to a known $\beta^0 = (\beta_{00}, \beta_{10}, \beta_{20})$, and finds the q and the (x_i, λ_i)'s that maximize $det(I(\beta^0))$, possibly under the restriction that the x_i's belong to a pre-specified region of interest.

In the case of a single factor, the solution to the D-optimal design problem for an unbounded experimental region is finite, and it is supported on just two points, (see [1], [12], [8], [5], [19], [13], [17] and [10].

In the case of two factors, Sitter and Torsney [18] find that if the experimental region is unbounded, $det(I(\beta))$ can be made arbitrarily large, as in the case for normal linear models. They then go on to characterize the D-optimal solution for a carefully chosen finite experimental region. It turns though, that if one restricts consideration to two-level factorial designs, and searches for the D-optimal design in that class, the solution is finite even for unbounded regions, (except when either β_{10} or β_{20} are equal to 0).

4 D-optimal 2^2 factorial designs

Here we derive the D-optimal design for the binomial model with (1), within the class of designs supported on the four vertices of a rectangle in R^2. Let $x_1 = (x_{10} - R_1, x_{20} - R_2)$, $x_2 = (x_{10} + R_1, x_{20} - R_2)$, $x_3 = (x_{10} - R_1, x_{20} + R_2)$ and $x_4 = (x_{10} + R_1, x_{20} + R_2)$, where $x_0 = (x_{10}, x_{20})$ is the center point of the 2^2 experiment, and R_1 and R_2 are one half the range of the levels chosen for the two factors. Furthermore, let $z_0 = \beta_0 + \beta_1 x_{10} + \beta_2 x_{20}$, and therefore assume that the center point is on the $ED_{F(z_0)}$ straight line.

4.1 Determinant of $I(\beta)$ for the 2^2 factorial design

Proposition 4.1 The determinant of $I(\beta)$, for the 2^2 factorial design supported on the vertices of the rectangle, (x_1, x_2, x_3, x_4), defined above is:

$$det(I(\beta)) = 4^2 n^3 R_1^2 R_2^2 \sum_{i<j<r} h(z_i, \lambda_i) h(z_j, \lambda_j) h(z_r, \lambda_r), \tag{4}$$

where $z_1 = z_0 - \beta_1 R_1 - \beta_2 R_2$, $z_2 = z_0 + \beta_1 R_1 - \beta_2 R_2$, $z_3 = z_0 - \beta_1 R_1 + \beta_2 R_2$ and $z_4 = z_0 + \beta_1 R_1 + \beta_2 R_2$. Equivalently,

$$det(I(\beta)) = n^3 \frac{(z_2 - z_1)^2 (z_3 - z_1)^2}{(\beta_1 \beta_2)^2} \sum_{i<j<r} h(z_i, \lambda_i) h(z_j, \lambda_j) h(z_r, \lambda_r). \qquad (5)$$

Note that the summation is over 4 terms. It follows that for 2^2 experiments, $det(I(\beta))$ depends on the design points (x_1, x_2, x_3, x_4), only through (z_0, R_1, R_2) and therefore for a given set of $(n, \lambda_1, \lambda_2, \lambda_3, \lambda_4)$, it only depends on the $ED_{F(z_0)}$ line where the design is centered, and the range for the two factors. Therefore, all 2^2 designs centered on the same $ED_{F(z_0)}$ line and with the same $(R_1, R_2, n_1, n_2, n_3, n_4)$, have the same value for $det(I(\beta))$.

Likewise, (5) implies that $(\beta_1 \beta_2)^2 det(I(\beta))$ depends on (x_1, x_2, x_3, x_4), only through (z_1, z_2, z_3, z_4), and therefore it depends only on the $ED_{F(z_i)}$ lines where the four combinations of dose levels are placed. In fact, by substituting $z_3 = 2z_0 - z_2$, $z_4 = 2z_0 - z_1$ and $\lambda_4 = 1 - \lambda_1 - \lambda_2 - \lambda_3$ in (5), one finds that for any 2^2 design, $(\beta_1 \beta_2)^2 det(I(\beta))$ depends only of (z_0, z_1, z_2) and $(n, \lambda_1, \lambda_2, \lambda_3)$.

For the linear normal case, the balanced allocation with $\lambda_i = 1/4$ for $i = 1, 2, 3, 4$, is always better than any unbalanced allocation with the same n. For the binary models under consideration, the balanced allocation is not optimal anymore, and one is bound to consider balanced and unbalanced situations separately. Due to lack of space, next we report only on the D-optimal 2^2 designs under the balanced restriction. The results on the general case will be reported in the expanded version of the manuscript.

4.2 Optimal 2^2 design centered at x_0 and with $\lambda_i = 1/4$

The maximization of (5) over (z_1, z_2), for fixed z_0 and λ_i's, can be posed in terms of the solution of a set of two non-linear equations. Tables 2 to 4 list the values for (z_1, z_2, z_3, z_4) for the 2^2 design with $\lambda_i = 1/4$ and centered at a $x_0 = (x_{10}, x_{20})$ on the $ED_{F(z_0)}$ line, that maximizes (5). Observe that for the D-optimal 2^2 design centered on x_0, $z_2 = z_3 = z_0$, and therefore x_2 and x_3 is always to be placed on the $ED_{F(z_0)}$.

For any 2^2 design, $R_1 = |z_2 - z_1|/(2|\beta_1|)$ and $R_2 = |z_2 - z_1|/(2|\beta_2|)$, and therefore the fact that the D-optimal choice satisfies $z_2 = z_3$ implies that for that choice, $R_2 = R_1|\beta_1/\beta_2|$. For example, if $F(z_0) = .7$, one finds that if one assumes that $\beta_1 = 2$ and $\beta_2 = 3$, the D-optimal design for the logistic model has $R_1 = 1.014$ and $R_2 = .6759$, for the probit model it has $R_1 = .5747$ and $R_2 = .3831$, for the log-log model it has $R_1 = .5716$ and $R_2 = .3811$ and for the complementary log-log model it has $R_1 = 1.011$ and $R_2 = .6738$.

The 2^2 factorial design in Table 2 with the largest $detI(\beta)$, is the one centered at a point on the $ED_{.5}$ line. Indeed, it can be proven that it is actually the best possible 2^2 balanced design for logistic models. On the other hand for the probit model in Table 3, the designs centered on the $ED_{.5}$

line are not the best ones anymore; even though the probit link is symmetric, the best 2^2 designs are centered on the $ED_{.8130}$ line. The best 2^2 design for the log-log link are the ones centered on the $ED_{.1086}$ line, and therefore the best ones for the complementary log-log are centered on the $ED_{.8914}$ line.

$F(z_0)$	$\|z_2 - z_1\|$	$F(z_1)$	$F(z_2)$	$F(z_3)$	$F(z_4)$	$(\beta_1\beta_2)^2 det(I(\beta))$
.50	3.8837	.9798	.50	.50	.0202	.009471 n^3
.55	3.8951	.9836	.55	.55	.0243	.009441 n^3
.60	3.9287	.9871	.60	.60	.0287	.009343 n^3
.65	3.9828	.9901	.65	.65	.0334	.009149 n^3
.70	4.0556	.9926	.70	.70	.0388	.008810 n^3
.75	4.1474	.9948	.75	.75	.0453	.008262 n^3
.80	4.2621	.9965	.80	.80	.0534	.007421 n^3
.85	4.4120	.9979	.85	.85	.0643	.006191 n^3
.90	4.6301	.9989	.90	.90	.0807	.004482 n^3
.95	5.0337	.9997	.95	.95	.1101	.002258 n^3

Table 2: Values of $F(z_1), F(z_2), F(z_3)$ and $F(z_4)$ that characterize the 2^2 design with $\lambda_i = 1/4$ and centered at any point on the $ED_{F(z_0)}$ line, that maximizes the determinant of $I(\beta)$ for the logistic model.

5 General comments and extensions

Observe that even in this supposedly very simple two-factor setting, the design issues involved in the binary response case are a lot more complicated than for typical continuous responses. Tables 2 to 4 indicate that, different from what happens in the linear normal case, the performance of two-level factorial experiments for a binary response heavily depends on the location of its center point, and on its orientation relative to the contour levels of the model surface, and unless either $\beta_1 = 0$ or $\beta_2 = 0$, it does not hold anymore that the larger the range of variation for the two factors, the larger that determinant and therefore the more informative the experiment. That complicates the choice of a 2^2 factorial experiment for binary responses.

Furthermore, as soon as one is willing to consider unbalanced allocations, where $\lambda_i \neq 1/4$, Tables 2 to 4 do not apply anymore and one has to maximize (5) over both (z_1, z_2, z_3, z_4) and $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ at the same time. Note also that even though in this communication we do neither cover the case of more than two factors, nor models that include second order interactions, it should be clear from our discussion that those situations lead to alternative choices of two-level factorial experiments, something that does not happen in the continuous response case. We plan to report on the results for general 2^{k-p} factorial designs for binary responses, in a follow up manuscript.

$F(z_0)$	$\lvert z_2 - z_1 \rvert$	$F(z_1)$	$F(z_2)$	$F(z_3)$	$F(z_4)$	$(\beta_1\beta_2)^2 det(I(\beta))$
.50	2.0174	.9782	.50	.50	.0218	.032052 n^3
.525	2.0213	.9814	.525	.525	.0251	.032047 n^3
.55	2.0331	.9846	.55	.55	.0282	.032032 n^3
.575	2.0532	.9875	.575	.575	.0312	.032016 n^3
.60	2.0823	.9902	.60	.60	.0337	.032014 n^3
.625	2.1213	.9927	.625	.625	.0357	.032043 n^3
.65	2.1707	.9947	.65	.65	.0371	.032128 n^3
.675	2.2303	.9964	.675	.675	.0378	.032291 n^3
.70	2.2989	.9976	.70	.70	.0380	.032551 n^3
.725	2.3736	.9985	.725	.725	.0379	.032902 n^3
.75	2.4513	.9991	.75	.75	.0378	.033312 n^3
.775	2.5295	.9995	.775	.775	.0380	.033707 n^3
.80	2.6072	.9997	.80	.80	.0387	.033979 n^3
.825	2.6850	.9998	.825	.825	.0400	.033979 n^3
.85	2.7645	.9999	.85	.85	.0420	.033516 n^3
.875	2.8488	.9999	.875	.875	.0447	.032344 n^3
.90	2.9426	1.000	.90	.90	.0483	.030139 n^3
.925	3.0537	1.000	.925	.925	.0532	.026458 n^3
.95	3.1984	1.000	.95	.95	.0601	.020668 n^3

Table 3: Values of $F(z_1), F(z_2), F(z_3)$ and $F(z_4)$ that characterize the 2^2 design with $\lambda_i = 1/4$ and centered at any point on the $ED_{F(z_0)}$ line, that maximizes the determinant of $I(\beta)$ for the probit model.

6 Appendix: Proof of Proposition 4.1

It can be checked that for designs supported on the vertices of a rectangle centered on x_0, (x_1, x_2, x_3, x_4), the determinant of $I(\beta)$ can be written as

$$det(I(\beta)) = n^3 R_1^2 R_2^2 det(\sum_{i=1}^{4} h(z_i, \lambda_i) l_i l_i') \tag{6}$$

with $l_i' = (1, (-1)^i, (-1)^{[(i-1)/2]+1})$. Algebraic computations lead to

$$\begin{aligned} det(I(\beta)) \quad = \quad & n^3 R_1^2 R_2^2 \sum_{i,j,r} (1 + 2(-1)^{[(i-1)/2]+j+[(r-1)/2]+r} \\ & + (-1)^{j+r+1} + (-1)^{[(j-1)/2]+[(r-1)/2]+1} \\ & + (-1)^{[(j-1)/2]+j+[(r-1)/2]+r+1}) h(z_i, \lambda_i) h(z_j, \lambda_j) h(z_r, \lambda_r), \end{aligned}$$

where the summation is over all 4^3 possible combinations of the elements of $\{1, 2, 3, 4\}$, taken in groups of three. Equation (4) follows from this.

$F(z_0)$	$\lvert z_2 - z_1 \rvert$	$F(z_1)$	$F(z_2)$	$F(z_3)$	$F(z_4)$	$(\beta_1\beta_2)^2 det(I(\beta))$
.05	4.1024	.9517	.05	.05	.0000	.047701 n^3
.10	4.0799	.9618	.10	.10	.0000	.057366 n^3
.15	4.0665	.9680	.15	.15	.0000	.055130 n^3
.20	4.0568	.9725	.20	.20	.0000	.048741 n^3
.25	4.0491	.9761	.25	.25	.0000	.041165 n^3
.30	4.0429	.9791	.30	.30	.0000	.033676 n^3
.35	4.0375	.9816	.35	.35	.0000	.026833 n^3
.40	4.0328	.9839	.40	.40	.0000	.020858 n^3
.45	1.7261	.8675	.45	.45	.0113	.018032 n^3
.50	1.8055	.8923	.50	.50	.0147	.018894 n^3
.55	1.9003	.9145	.55	.55	.0183	.019485 n^3
.60	2.0104	.9339	.60	.60	.0221	.019751 n^3
.65	2.1379	.9505	.65	.65	.0259	.019631 n^3
.70	2.2864	.9644	.70	.70	.0299	.019055 n^3
.75	2.4625	.9758	.75	.75	.0342	.017935 n^3
.80	2.6775	.9848	.80	.80	.0389	.016156 n^3
.85	2.9532	.9916	.85	.85	.0444	.013575 n^3
.90	3.3387	.9963	.90	.90	.0513	.010009 n^3
.95	3.9932	.9990	.95	.95	.0619	.005273 n^3

Table 4: Values of $F(z_1), F(z_2), F(z_3)$ and $F(z_4)$ of the 2^2 design with $\lambda_i = 1/4$ and centered at any point on the $ED_{F(z_0)}$ line, that maximizes the determinant of $I(\beta)$ for the log-log model. For the complementary log-log model, the $F(z_i)$'s for the optimal design centered on the $ED_{F(z_0)}$ line, are 1 minus the $F(z_i)$ values in the line for $1 - F(z_0)$ of this table.

References

[1] Abdelbasit K.M., Plackett R.L. (1983). *Experimental design for binary data.* Journal of the American Statistical Association **78**, 90 − 98.

[2] Atkinson A.C., Donev A.N. (1992). *Optimum experimental designs.* Clarendon Press, Oxford.

[3] Bisgaard S., Fuller H.T. (1995). *Sample size estimates for 2^{k-p} designs with binary responses.* Journal of Quality Technology **27**, 344 − 354.

[4] Box G.E.P., Draper N. (1987). *Empirical model-building and response surfaces.* Wiley, New York.

[5] Ford I., Torsney B., Wu C.F.J. (1992). *The use of canonical form in the construction of locally optimal designs for non-linear problems.* Journal of the Royal Statistical Society (ser B) **54**, 569 − 583.

[6] Goel P.K., Ginebra J. (2003). *When is one experiment always better than another?.* Journal of the Royal Statistical Society (ser D) **52**, 515 − 537.

[7] Hamada M., Nelder J.A. (1997). *Generalized linear models for quality-improvement experiments.* Journal of Quality Technology **29**, 292 – 304.

[8] Khan M.K., Yazdi A.A. (1988). *On D-optimal designs for binary data.* Journal of Statistical Planning and Inference **18**, 83 – 91.

[9] Kiefer J. (1959). *Optimum experimental designs, (with discussion).* Journal of the Royal Statistical Society (ser B) **21**, 272 – 319.

[10] Mathew T., Sinha, B.K. (2001). *Optimal designs for binary data under logistic regression.* Journal of Statistical Planning and Inference **93**, 295 – 307.

[11] Mc Cullagh P., Nelder J.A. (1989). *Generalized linear models.* 2nd ed., Chapman Hall, London.

[12] Minkin S. (1987). *Optimal designs for binary data.* Journal of the American Statistical Association **82**, 1098 – 1103.

[13] Myers W.R., Myers R.H., Carter W.H. (1994). *Some alphabetic optimal designs for the logistic regression model.* Journal of Statistical Planning and Inference **42**, 57 – 77.

[14] Myers R.H., Montgomery D.C. (1997). *A tutorial on generalized linear models.* Journal of Quality Technology **29**, 274 – 291.

[15] Myers R.H., Montgomery D.C., Vinning G.G. (2002). *Generalized linear models.* Wiley, New York.

[16] Pukelsheim F. (1993). *Optimal design of experiments.* Wiley, New York.

[17] Sitter R.R., Fainaru I. (1997). *Optimal design for the logit and probit models for binary data.* Canadian Journal of Statistics **25**, 175 – 190.

[18] Sitter R.R., Torsney B. (1995). *Optimal designs for binary response experiments with two design variables.* Statistica Sinica **5**, 405 – 419.

[19] Sitter R.R., Wu C.F.J. (1993). *Optimal designs for binary response experiments: Fieller, D, and A criteria.* Scandinavian Journal of Statistics **20**, 329 – 342.

[20] Wu C.F.J. (1988). *Optimal design for percentile estimation of a quantal response curve.* In: Optimal Design and Analysis of Experiments (eds. Dodge, Y., Fedorov, V. and Wynn, H.P.), Elsevier, Amsterdam. 213 – 223.

[21] Wu, C.F.J., Hamada, M. (2000). *Experiments. Planning, analysis, and parameter design optimization.* Wiley, New York.

Acknowledgement: This project was supported by the Ministerio de Ciencia y Tecnología Research Grant with reference number BFM2001-2327. We thank Professor Josep Ginebra for their contributions regarding the material covered by this paper.

Address: R. Dorta-Guerra, E. González-Dávila, Dpto. de Estadística, Investigación Operativa y Computación, Universidad de La Laguna, C. Astrofísico Fco. Sánchez, 38271 La Laguna, Tenerife, Spain

E-mail: `rodorta@ull.es, egonzale@ull.es`

REDUCTION OF GIBBS PHENOMENON IN WAVELET SIGNAL ESTIMATION

Timothy R. Downie

Key words: Wavelets, thresholding, non-decimated wavelet transform, Haar wavelet, signal processing.

COMPSTAT 2004 section: Smoothing.

Abstract: Wavelet thresholding is an effective method for noise reduction of a wide class of naturally occurring signals. However, bias near to a discontinuity and Gibbs phenomenon are a drawback in wavelet thresholding. The Haar wavelet basis is good at approximating discontinuities, but is bad at approximating other signal artefacts.

A method of detecting jumps in a signal is developed that uses non-decimated Haar wavelet coefficients. This is designed to be used in conjunction with most existing thresholding methods. A detailed simulation study has been carried out. A summary of the results show that when discontinuities are present, a substantial reduction in bias can be obtained, leading to a corresponding reduction in mean square error.

1 Introduction

Signal estimation using wavelets, gives robust estimates with good results in terms of mean squared error. The usual justification given is that a *wide class of signals* observed in practice have an economical representation in the wavelet domain. Included in this wide class of signals are functions that include discontinuities.

A straightforward signal estimation routine is to threshold the wavelet coefficients. This handles signal discontinuities with only partial success. The location and the size of the jump are estimated well, compared to many non-wavelet methods. However there is usually an interval to either side of the jump that has noticeable oscillation (Gibbs phenomenon) and is a source of bias.

A graphical example of this can be seen in Figure 1. The dashed line is the true signal (the saw-tooth function) to which white noise has been added. A hard thresholding routine using the universal threshold was applied to the wavelet coefficients, to obtain the estimate shown as the solid line. In this example there is quite a considerable Gibbs phenomenon in the resulting estimate.

This bias occurs because a jump is encoded by a small number of wavelet coefficients, whose positions are determined by the location of the jump. The dominant coefficients are large, but the coefficients that give a smooth signal up to the point of the discontinuity will often be small with respect to the noise.

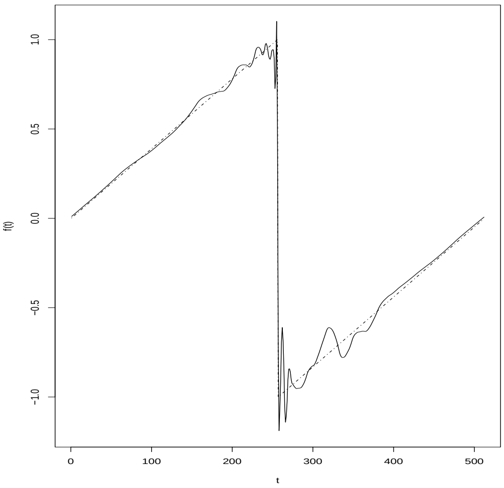


Figure 1: Simple thresholding of a linear function with discontinuity. Hard thresholding with the universal threshold was applied to a wavelet decomposition using the Daubechies Least Asymmetric wavelet with 5 vanishing moments. A root signal to noise ratio of 5 was used.

The method proposed in this paper utilises the ability of the Haar wavelet to efficiently represent discontinuities, and the ability of thresholding with smoother wavelets to efficiently represent the continuous part of the signal. The denoising of the continuous signal is independent from the jump detection routine, and the user is free to choose the method of his/her own choice.

2 Wavelet representation and wavelet thresholding

For an application based introduction to wavelet methods refer to Abramovich et al. [1]. For a more mathematical introduction see Daubechies [4] or Chui [3]. Percival and Walden [9] introduces this topic with respect to stochastic time series.

A mother wavelet ψ is a function that generates a wavelet basis by taking dilations and translations of this mother wavelet. Any one dimensional real function f in an appropriate functional space can be represented by this wavelet basis using $f(t) = \sum_{j,k} d_{j,k}\psi_{j,k}(t)$, where $d_{j,k}$ are real coefficients and $\psi_{j,k}(t) = 2^{\frac{j}{2}}\psi(2^j t - k)$ are the dilations and translations of the mother wavelet.

The value of each coefficient encodes the amount of energy present in the function $f(t)$ at a scale 2^j and at location $2^j t - k$.

A fast algorithm called the Discrete Wavelet Transform (DWT) can be used to compute the wavelet coefficients. For a given j, the the sequence of wavelet coefficents $d_{j,k}$ is called the jth resolution level. In the DWT, the number of coefficients in each resolution level decreases as j decreases.

The non-decimated DWT (ND-DWT) has the same number of wavelet coefficients at each resolution level, an over determined transformation. The ND-DWT is primarily used in analysing non-stationary time series [9] but can also be used for wavelet thresholding [2][7].

Suppose an observed function $g(t)$ is the true function $f(t)$ corrupted by Gaussian white noise, then a method of estimating f is wavelet thresholding [5]. The DWT of $g(t)$ is taken, to obtain the coefficients $d_{j,k}$. All the $d_{j,k}$ with a magnitude less than a threshold are set to zero, the larger coefficients are either unchanged (hard thresholding) or shrunk towards zero (soft thresholding). Applying the inverse DWT to the modified coefficients $(\hat{d}_{j,k})$ gives the estimate of f.

For a given threshold λ, the hard thresholding rule is $\hat{d}^{(H)}_{j,k} = d_{j,k} I(|d_{j,k}| > \lambda)$, and the soft thresholding rule is $\hat{d}^{(S)}_{j,k} = \text{sgn}(d_{j,k})(|d_{j,k}| - \lambda)_+$. A simple rule for choosing the threshold is the universal threshold [5]. The universal threshold is $\lambda_{(UV)} = \sigma\sqrt{2\log n}$, where n is the number of coefficents thresholded and σ^2 is the variance of the white noise.

3 Bias and Gibbs phenomenon in wavelet thresholding

3.1 Using hard thresholding

Each wavelet coefficient can be written as $d_{j,k} = \theta_{j,k} + \epsilon_{j,k}$, where $\theta_{j,k}$ is the wavelet coefficinet of the true function and $\epsilon_{j,k}$ is the noise in the wavelet domain. Provided the wavelet basis is orthonormal, $\epsilon_{j,k} \sim N(0, \sigma^2)$. The $\{j, k\}$ subscripts can be dropped when considering individual coefficients.

For a given θ, λ and σ the bias of $\hat{d}^{(H)}$, $\text{B}_{\hat{d}^{(H)}}(\theta, \lambda, \sigma)$, can be obtained [6].

The bias function is anti-symmetric in θ and tends to zero as $\theta \to \pm\infty$. For large values of θ the probability that the coefficient is thresholded is very low and so the bias is small, alternatively if θ is very close to zero, then the thresholded estimate, i.e. zero, will be close to θ. Large bias occurs when absolute value of the coefficient is between $\lambda/2$ and λ.

Considering the overall bias in the reconstructed signal, it is natural to take the mean squared bias (MSB) as a measure of overall bias which is $\frac{1}{n}\sum_{i=1}^{n} B(\hat{f}(t_i))^2 = \frac{1}{n}\sum_j \sum_k B_{\hat{d}^{(H)}_{j,k}}(\theta_{j,k}, \lambda, \sigma)^2$,

3.2 Example

As an example we use the sawtooth function as the signal. This function is linear everywhere except at the discontinuity. Noise is added to the signal, with a root signal-to-noise ratio (RSNR) of 5. We can compute the theoretical

MSE of the reconstructed signal and the theoretical MSB of the reconstructed signal due to thresholding. The implied thresholding method used was the universal hard threshold, with known variance,

The theoretical MSE of the noisy signal is 0.0134, The theoretical MSB of the reconstructed signal is 4.44×10^{-4}, and the theoretical MSE of the reconstructed signal is 1.57×10^{-3}, so the squared bias accounts for 28.2% of the error. A single simulated realisation of this example was used as the example in Figure 1.

A further consideration is that all the non-zero coefficients of the signal occur within in 5 coefficients' distance from the discontinuity. On reconstruction, the bias is concentrated in the interval around the discontinuity giving a poor visual representation.

3.3 Using other wavelet methods

So far we have only considered the most basic thresholding method, which allows us to compute the exact expected errors. There are however many adaptations to the basic method which may make the Gibbs phenomenon less stark, but it will still be present.

Other wavelet methods which have been considered are soft thresholding, level dependent thresholding, the choice of the mother wavelet used, the SURE-shrink method of choosing the threshold, and coefficient specific thresholds that depend on the magnitude of neighbouring coefficients. Details of the bias introduced by these methods can be found in Downie [6].

Using the ND-DWT rather than the DWT to obtain the wavelet coefficients reduces the amount of bias and Gibbs effect near a discontinuity, however it is still a problem and the Gibbs effect is persists for longer than when using the DWT.

4 Detection of discontinuities using Haar wavelets

4.1 The Haar wavelet

A Haar wavelet basis is defined by the mother wavelet $\psi(x) = I_{[0,0.5)}(x) - I_{[0.5,1)}(x)$ [4]. This wavelet has ideal time localisation, but the worst decay in the frequency domain and has zero vanishing moments. It also has a discontinuity at $x = 1/2$, which is usually considered a disadvantage for reconstruction, as the signal estimate exhibits many small discontinuities. This discontinuity, however, is an advantage when there is a large discontinuity in the signal.

Our aim is to take advantage of a Haar wavelet decomposition to identify discontinuities whilst keeping the advantages of smoother wavelets for all other aspects of denoising.

4.2 Description of method

When considering the Haar coefficients at one resolution level ($|d^{(h)}_{j',k}|$) and the corresponding Daubechies wavelet coeffcients ($|d^{(w)}_{j',k}|$), there is a sharp peak in the Haar wavelet coefficients at the location of a jump. However the wavelet coefficients have multiple peaks all smaller then the Haar coefficient peak.

The heuristic is that if the Haar coefficients in a resolution level exceed all the wavelet coefficents in that level, then the largest artefact will probably be a jump.

A *possible jump* can be identified at the location of the largest absolute Haar coefficient within a chosen resolution level (j') provided the largest coefficient is larger than the associated higher order wavelet coefficients. The magnitude of the jump can then be estimated either from the value of the largest coefficient, or from the observed signal in the time domain. Once a discontinuity has been located at t^* and estimated as δ then the underlying step function $s(t) = \delta I_{[t^*,n]}(t)$ can be removed from the signal. The residual signal (the observed signal with the step removed) can then be checked for further discontinuities. When no further discontinuities can be detected then the residual signal can be wavelet thresholded using the user's preferred method.

The above method, as it stands, will identify too many 'possible jumps', because non-zero Haar coefficients are required to fit linear and other low order locally polynomial signals. So another constraint is used to reduce the number of proposed jumps. Haar coefficients encode either a jump, a low order polynomial, or some other continuous artefact in the signal; large wavelet coefficients only encode either a jump, or some other continuous artefact. The wavelet coefficients for a continuous artefact will usually be larger than the Haar coefficients. If there are wavelet coefficients that exceed a threshold (i.e. $\max\{|d^{(w)}_{j',k}|\} > \lambda$) then we assume that there are one or more signal artefacts that are not locally low-order polynomials, and we only check for jumps if this condition is satisfied.

5 Simulation study

A comprehensive simulation study has been done comparing the jump detection method, with existing thresholding methods[6]. The Wavethresh package [8] was used to implement the wavelet transforms and the existing thresholding methods. Here we only present the comparison between different methods using one piecewise continuous signal.

For each simulation reported as *without jump detection*, a noisy piecewise continuous signal with a RSNR of 5 is simulated. The ND-DWT is applied, the thresholding method specified is applied to the wavelet coefficients and the resulting wavelet coefficients are transformed back to the time domain. Twenty simulations are run to obtain an estimate of the expected reconstructed signal, hence an estimate of the mean squared bias MSB. This process is repeated also twenty times, giving a sample of MSB estimates from which the average and standard error can be obtained. In each example the same is also carried out *with detection*, i.e. applying the jump detection method to the ND-DWT prior to the wavelet thresholding, as described in the previous section.

Table 1 compares the following different methods of thresholding. Without detection the lowest MSE was obtained using NeighCoeff, reducing the MSE of the noisy signal from 1.57 to 0.320, while the smallest MSB was obtained using SURE. Using jump detection with this signal gives a consistently better MSE and MSB for all methods. Universal hard and soft, SURE, SURE level dependent and NeighCoeff all give a similar MSE between 0.13 and 0.14. The MSB is always much lower with detection sometimes as low as a tenth of the MSB without jump detection. Even for the method that gives the lowest MSB without detection, there is a 79% reduction in MSB when using jump detection.

Method	MSE	s.e. (10^{-3})	MSB	s.e. (10^{-3})
	Without Detection			
Universal Hard	0.37	3.77	0.247	3.57
Universal Soft	0.833	4.25	0.754	4.42
Sure	0.354	2.76	0.157	2.47
Sure level dep.	0.342	3.19	0.171	2.26
NeighCoeff	0.32	3.04	0.194	2.7
	With Detection			
Universal Hard	0.133	2.36	0.0242	0.85
Universal Soft	0.147	3.08	0.0324	0.738
Sure	0.144	2.5	0.0327	0.581
Sure level dep.	0.137	2.32	0.0219	0.683
NeighCoeff	0.136	2.26	0.0194	0.929

Table 1: Comparing methods of thresholding.

We also investigate the contribution to the MSB within 32 points either side of each jump. There there is a huge decrease in the bias in the neighbourhood of a jump, for universal soft at the first jump the MSB with jump detection is over 300 times smaller than without jump detection. For SURE thresholding detection gives a 98% decrease in MSB in the vicinity of both jumps.

Figure 2 shows the expected signal obtained from averaging all 400 simulations giving an indication of the bias with and without jump detection. The expected estimate without jump detection is the grey line and the expected estimate with jump detection is the thick black line. The true signal

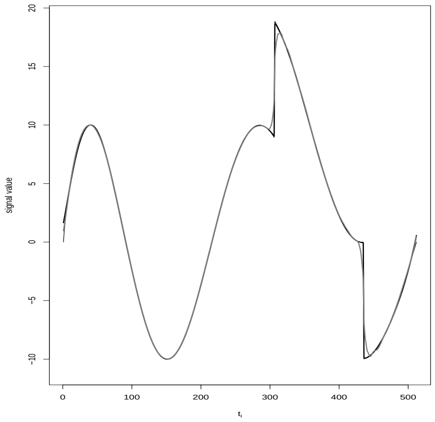


Figure 2: The expected signal estimates using SURE level dependent thresholding with no jump detection (grey line) and with jump detection (black line).

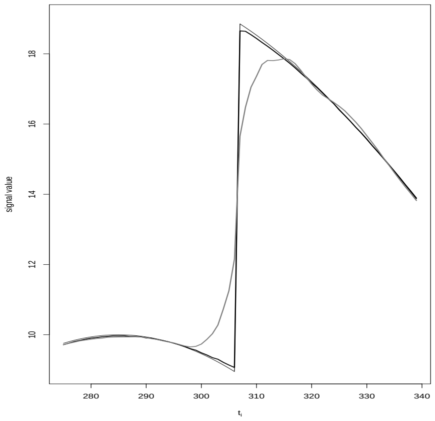


Figure 3: The expected signal estimates using SURE level dependent thresholding with no jump detection (grey line) and with jump detection (black line), with $275 \leq t_i \leq 339$.

is plotted using a thin line which on this scale is almost completely obscured by the thick black line. Figure 3 is a blow up of Figure 2, showing 32 time points either side of the first jump. The bias without using jump detection is clear, whereas using jump detection the only bias is caused by a small underestimation of the magnitude of the true jump, as shown by the thin line.

6 Conclusion

A method of detecting jumps in a signal is developed that uses non decimated Haar wavelet coefficients. The proposed method is aimed at reducing the bias due to wavelet thresholding in the neighbourhood to such jump. It can be considered as pre-thresholding step and the user can use their preferred thresholding method.

Simulations results show that when discontinuities are present a substantial reduction in bias can be obtained, leading to a corresponding reduction in mean square error. This reduction in bias occurs locally to the discontinuities.

References

[1] Abramovich F., Bailey T.C., Sapatinas T. (2000). *Wavelet analysis and its statistical applications.* Journal of the Royal Statistical Society, Series D The Statistician **48**, 1 − 30.

[2] Coifman R.R., Donoho D.L. (1995). *Translation-invariant de-noising.* Technical report, Yale University.

[3] Chui C.K. (1992). *An introduction to wavelets.* Academic Press, London.

[4] Daubechies I. (1992). *Ten lectures on wavelets.* Society for Industrial and Applied Math., Philadelphia.

[5] Donoho D.L., Johnstone I.M. (1994). *Ideal spatial adaption via wavelet shrinkage.* Biometrika **81**, 425 − 455.

[6] Downie T.R. (2003). *Acurate signal estimation near discontinuities.* Technical report 239, University College London, `http://www.ucl.ac.uk/Stats/Resrprts/abs03.html#239`.

[7] Nason G.P., Silverman B.W. (1995). *The stationary wavelet transform and some statistical applications.* In A. Antoniadis and G. Oppenheim, editors, Lecture Notes in Statistics 103, Wavelets and Statistics, Springer-Verlag, New York, 281 − 299.

[8] Nason G.P. (1998). *WaveThresh3 software.* Department of Mathematics, University of Bristol, Bristol, UK.

[9] Percival D. B., Walden A. T. (2000). *Wavelet methods for time series analysis.* Cambridge University Press, U.K.

Acknowledgement: This research was supported by a Nuffield Foundation Award NUF-NAL 02.

Address: T.R. Downie, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom.

E-mail: `tim@stats.ucl.ac.uk`

EXACT SIMULATION-BASED INFERENCE FOR AUTOREGRESSIVE PROCESSES BASED ON INDUCED TESTS

Jean-Marie Dufour and Malika Neifar

Key words: Autoregressive process, exact inference, induced test, test combination, Monte Carlo test.

COMPSTAT 2004 section: Time series analysis.

Abstract: In this paper, we propose exact inference methods for autoregressive models of given order p ($p \geq 1$), which may be nonstationary and include a drift term. We study the problem of testing any hypothesis that sets the complete vector of autoregressive coefficients. This is done by first transforming the model to eliminate serial dependence under the null hypothesis, and then testing whether autocorrelation remains present in the transformed data. Tests for dependence at different lags are then combined using the methods proposed by [16] and Fisher-Pearson [11], [14] for combining independent tests. In view of the dependence among the different tests, the size of the combined procedures is controlled by using Monte Carlo test techniques. The construction of valid confidence sets based on these tests is discussed. Numerical illustrations based on simulated data are also presented.

1 Introduction

Statistical inference (tests and confidence regions) on autoregressive (AR) models constitute a basic problem in time series analysis, statistics and econometrics. Applied methods are generally based on unreliable asymptotic approximations even under strong parametric assumptions (such as, Gaussian innovations). But the actual level of asymptotically based tests can differ markedly from the posted level even with reasonably large samples, especially for processes of order greater than one (see [2], [5] and [15]).

In this paper, we consider the problem of testing any hypothesis that sets the full vector of the autoregressive coefficients. We propose tests which are applicable on both stationary and nonstationary processes including a drift term, with possibly non-Gaussian errors. Further, we consider the problem of building exact confidence intervals and regions for the AR coefficients.

For that purpose, the data are first transformed under the null hypothesis so that the data become independent. Then the filtered data are tested for the presence of serial dependence at several lags. This raises the problem of combining tests against serial dependence at different lags. In [9], we considered a combination technique based on the Boole-Bonferroni inequality. This has the disadvantage of producing conservative tests, hence a power loss. Here, we consider induced tests based on two approaches originally suggested for independent test p-values, namely, the minimum criterion of [16] and the Fisher-Pearson product criterion [11] and [14]; for further discussion of induced tests, see [10] and [7]. Since the tests are not independent, the overall size of the test is controlled by the technique of Monte Carlo tests (see [6] and [4]). The technique of MC tests allows one to obtain provably exact randomized tests in finite samples using very small numbers of MC replications of the test statistics under null hypothesis.

This paper is organized as follows. In section 2, we describe the model and test problems studied. In section 3, the test statistics are derived. The implementation by the MC test technique is explained in section 4. Numerical illustrations based on simulated data are presented in section 5.

2 Framework

We consider a parametric autoregressive model of order p [AR(p)]:

$$y_t = \beta + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + u_t \ , \tag{1}$$

$$u_t = \sigma\varepsilon_t \text{ where } \varepsilon_t \overset{i.i.d.}{\sim} D\,,\ t = 1, \ldots, T\,, \tag{2}$$

where D is a completely specified distribution, the errors $\boldsymbol{u} = (u_1, \ldots, u_T)'$ are independent of the initial values $y_0, y_{-1}, \ldots, y_{-p+1}$, the parameters $\varphi_1, \varphi_2, \ldots, \varphi_p$, β and σ^2 are unknown, $T \geq p+1$ and $p \geq 1$. Assumption (2) means that the errors are independent and identically distributed (i.i.d.) according to a distribution which is specified up to an unknown scale parameter. An important special case of (2) is the one where the errors are i.i.d. Gaussian:

$$u_t \overset{i.i.d.}{\sim} [0,\ \sigma^2],\ t = 1, \ldots, T\,. \tag{3}$$

But other distributions could be considered, such as more heavy-tailed distributions (e.g., a Cauchy distribution).

It will useful here to reparameterize this model as in [3] and [1]. This yields the equivalent form:

$$\begin{aligned} y_t &= \beta + \big(\sum_{j=1}^{p}\varphi_j\big)y_{t-1} + \sum_{j=1}^{p-1}\big(-\sum_{i=j+1}^{p}\varphi_i\big)[y_{t-j} - y_{t-(j+1)}] + u_t \\ &= \beta + \theta_1 y_{t-1} + \sum_{j=2}^{p}\theta_j \triangle y_{t-(j+1)} + u_t\,, \qquad t = 1, \ldots, T\,, \end{aligned} \tag{4}$$

where $\theta_1 = \varphi_1 + \varphi_2 + \cdots + \varphi_p$, $\theta_j = -\sum_{i=j}^{p}\varphi_i$, and $\triangle y_{t-j} = y_{t-j} - y_{t-(j+1)}$, $j \geq 2$. Now, unknown parameters are $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$, β and σ.

We wish to study the problem of testing any hypothesis which sets the complete vector of autoregressive coefficients $\boldsymbol{\theta}$ in (4) at any specified value

$\boldsymbol{\theta}_0$ [in an admissible set S]:

$$H_0(\boldsymbol{\theta}_0) : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{against} \quad H_a(\boldsymbol{\theta}_0) : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \,. \tag{5}$$

3 Test statistics

In order to test $H_0(\boldsymbol{\theta}_0) : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, we consider the following transformation of the data:

$$z_t(\boldsymbol{\theta}_0) = y_t - \theta_{10} y_{t-1} - \theta_{20} \triangle y_{t-1} - \cdots - \theta_{p0} \triangle y_{t-p+1} = \beta + u_t \,,\ t = 1, \ldots, T \,. \tag{6}$$

Thus, under $H_0(\boldsymbol{\theta}_0)$, the filtered variables $z_t(\boldsymbol{\theta}_0)$, $t = 1, \ldots, T$ are distributed like u_t, $t = 1, \ldots, T$, except possibly for the addition of an unknown constant β. Under the alternative hypothesis $H_a(\boldsymbol{\theta}_0)$, the same variables are autoccorrelated, following an ARMA(p, p) process. Consequently, we can test $H_0(\boldsymbol{\theta}_0)$ by testing whether the variables $z_t(\boldsymbol{\theta}_0)$, $t = 1, \ldots, T$, are mutually independent. Since the autocorrelation structure of an ARMA(p, p) process is completely specified by its first p autocorrelations, it will be sufficient to test the independence of $z_t(\boldsymbol{\theta}_0)$, $t = 1, \ldots, T$, against autocorrelations at lags $1, 2, \ldots, p$ (see [13] for more details).

To do so, we can use p statistics of the form:

$$D_j(\boldsymbol{\theta}_0) = \frac{\boldsymbol{z}(\boldsymbol{\theta}_0)' \, \bar{\boldsymbol{A}}_j \, \boldsymbol{z}(\boldsymbol{\theta}_0)}{\boldsymbol{z}(\boldsymbol{\theta}_0)' \, \bar{\boldsymbol{B}}_j \, \boldsymbol{z}(\boldsymbol{\theta}_0)} \,,\ j = 1, \ldots, p \,, \tag{7}$$

where $\bar{\boldsymbol{A}}_j = \boldsymbol{M}\boldsymbol{A}_j\boldsymbol{M}$, $\bar{\boldsymbol{B}}_j = \boldsymbol{M}\boldsymbol{B}_j\boldsymbol{M}$, $\boldsymbol{M} = \boldsymbol{I}_T - \frac{1}{T}\boldsymbol{\iota}'\boldsymbol{\iota}$, $\boldsymbol{\iota} = (1, 1, \ldots, 1)'$ and $\boldsymbol{B}_j$ is a positive definite matrix. $D_j(\boldsymbol{\theta}_0)$ may be interpreted as a test statistic designed to especially sensitive to serial dependence at lag j. Many commonly used autocorrelation tests – for example, tests based on Durbin-Watson statistics and sample autocorrelations – involve statistics of the form (7). But the null hypothesis we focus on is independence (which entails the absence of serial dependence at all lags) between $z_t(\boldsymbol{\theta}_0)$, $t = 1, \ldots, T$.

Under the assumptions (1) - (2) and $H_0(\boldsymbol{\theta}_0)$, $D_j(\boldsymbol{\theta}_0)$ has a distribution which does not depend on nuisance parameters (β and σ) and can be easily simulated by MC techniques (or calculated for example by Imhof's algorithm [12] in the Gaussian case). Further, this distribution does not depend on $\boldsymbol{\theta}_0$. Note that an error normality assumption is not required for this invariance to hold.

In order to test the hypothesis of independence, we will combine tests based on $D_j(\boldsymbol{\theta}_0)$, $j = 1, \ldots, p$, in a way that will control the overall level of the procedure. Specifically, we consider two methods for combining the tests (or p-values) without using a conservative bound, such as the Bonferroni inequality: namely, the minimum criterion of [16] and the Fisher-Pearson product criterion [11] and [14].

For that purpose, we start by considering the survival function (or p-value function) of $D_j(\boldsymbol{\theta}_0)$ under $H_0(\boldsymbol{\theta}_0)$:

$$G_j(x) = \mathsf{P}[D_j(\boldsymbol{\theta}_0) \geq x] .$$

On evaluating this function at $x = \hat{D}_j(\boldsymbol{\theta}_0)$, the observed value of $D_j(\boldsymbol{\theta}_0)$, we get the marginal significance level of the test of $H_0(\boldsymbol{\theta}_0)$ based on $D_j(\boldsymbol{\theta}_0)$:

$$p_j = G_j[\hat{D}_j(\boldsymbol{\theta}_0)] .$$

For one-sided tests, the critical region with level α_j $(0 < \alpha_j < 1)$ for each test $D_j(\boldsymbol{\theta}_0)$ has the form

$$p_j \leq \alpha_j ,$$

$j = 1, \ldots , p$. Similarly, to get a two-sided test with level α_j based on $D_j(\boldsymbol{\theta}_0)$, we can take

$$\min\{p_j,\, 1 - p_j\} \leq \alpha_j/2 . \tag{8}$$

If the exact distribution of the statistic $D_j(\boldsymbol{\theta}_0)$ is unknown, we can also use an asymptotic approximation or simulate it through a Monte Carlo experiment independent of the data. For example, we can consider $t_j = |\sqrt{T}\, r_j(\boldsymbol{\theta}_0)|$ which is asymptotically N[0, 1] under H_0, where $r_j(\boldsymbol{\theta}_0)$ is the lag j sample autocorrelation based on the transformed data. To the extent that the joint distribution of the statistics $D_j(\boldsymbol{\theta}_0)$, $j = 1, \ldots , p$, is free of nuisance parameters and can be simulated (in the present case, this is easy to do), the fact that individual p-values are only approximate can be dealt with automatically by the MC test technique described in the following section.

A difficulty we meet here consists in combining multiple tests based on different statistics of the form $D_j(\boldsymbol{\theta}_0)$, $j = 1, \ldots , p$. We will consider here two ways of combining multiple test p-values, the minimum criterion of [16] and the Fisher-Pearson product criterion [11] and [14], which can be described as follows:

1. Tippett's minimum p-value criterion is based on

   $$F_{\min} = \inf_{j=1, \ldots , p} \{p_j\},$$

   or

   $$\bar{F}_{\min} = 1 - \inf_{j=1, \ldots , p} \{p_j\} ;$$

   the null hypothesis is rejected when $F_{\min}$ is small or, equivalently, when $\bar{F}_{\min}$ is large;

2. the Fisher-Pearson product criterion is based on

   $$F_{\times} = \prod_{j=1}^{p} p_j$$

or

$$\bar{F}_{\times} = 1 - \prod_{j=1}^{p} p_j \,;$$

the null hypothesis is rejected when $F_{\times}$ is small or, equivalently, when $\bar{F}_{\times}$ is large.

Even though the distribution of the above induced test statistics may be quite difficult to establish analytically, they do not involve nuisance parameters and can be easily simulated. Consequently, we can apply to them the technique of MC tests [4], which is described in section 4.

Confidence sets can be obtained by "inverting" the above proposed tests for hypotheses $H_0(\boldsymbol{\theta}_0)$, i.e., by finding the set I of values $\boldsymbol{\theta}_0$ which are not rejected at level α. This yields a joint confidence set I with level $1-\alpha$ for $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$:

$$P[\boldsymbol{\theta} \in I] \geq 1 - \alpha. \tag{9}$$

By projection, we can also build simultaneous confidence intervals for the individual coefficients θ_i, $i = 1, \ldots, p$; for more details, see [8], [6] and [4].

4 Finite-sample Monte Carlo tests

Consider a statistics, say S, for testing $H_0(\boldsymbol{\theta}_0)$ and assume its distribution is continuous. Let S_0 be the test statistic computed from the observed data. In view of the fact that S has a distribution that does not involve nuisance parameters under the null hypothesis, we can use Monte Carlo methods to simulate N i.i.d. replications $S_1, \ldots, S_N$ of S under $H_0(\boldsymbol{\theta}_0)$, independently of S_0. Since $S_0, S_1, \ldots, S_N$ are i.i.d., all rankings of these $N+1$ variables are equally probable. It follows that the rank $R_N(S_0)$ of S_0 has a uniform distribution over the integers $1, \ldots, N+1$.

Let us define the p-value function

$$p_N(x) = \frac{N G_N(x) + 1}{N+1} \tag{10}$$

where

$$G_N(x) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{[0,\,\infty)}(S_j - x), \quad \mathbf{1}_A(x) = \begin{cases} 1, & \text{if } x \in A\,, \\ 0, & \text{if } x \notin A\,. \end{cases}$$

If N is chosen so that $\alpha(N+1)$ is an integer (e.g., for $\alpha = 0.05$, we can take $N = 19, 99, 199$, etc.), the above uniformity result entails that

$$\mathsf{P}[p_N(S_0) \leq \alpha] = \alpha \tag{11}$$

under the null hypothesis. Thus the critical region $p_N(S_0) \leq \alpha$ has level α. In other words, the randomized critical region $p_N(S_0) \leq \alpha$ control the level of the test and has the same level as the critical region $G(S_0) \leq \alpha$. For further details and references, see [8], [6] and [4].

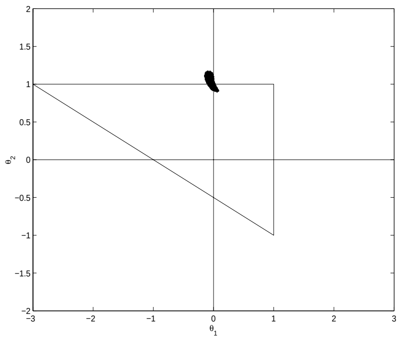


Table 1: Confidence set for model M1.

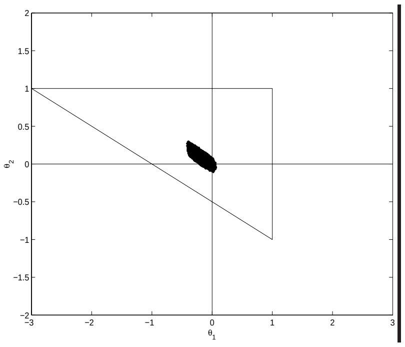


Table 2: Confidence set for model M2.

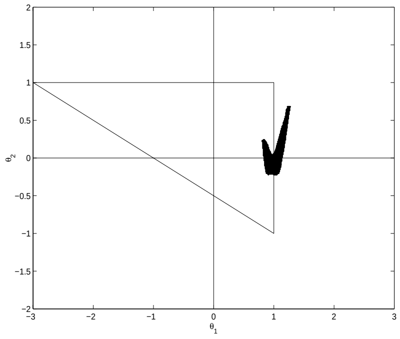


Table 3: Confidence set for model M3.

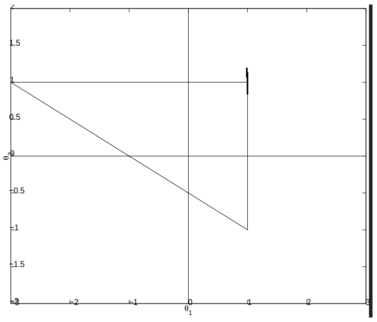


Table 4: Confidence set for model M4.

5 Numerical illustration

In order to illustrate the above methodology, we generated data using AR(2) models with the following coefficients:

Model	M_1	M_2	M_3	M_4
Parameter values	$\theta_1 = 0$, $\theta_2 = 1$	$\theta_1 = 0$, $\theta_2 = 0$	$\theta_1 = 1$, $\theta_2 = 0$	$\theta_1 = 1$, $\theta_2 = 1$

where $u_t \overset{i.i.d.}{\sim} \mathrm{N}[0,\ \sigma^2],\ t = 1, \ldots, T,\ T = 52,\ \beta = 0,\ (y_0,\ y_{-1}) = (0,\ 0)$. For level $\alpha = 0.05$ and using $N = 19$ replications, we applied the method described in the previous sections to test hypotheses on $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, using the Fisher-Pearson combination criterion ($\bar{F}_\times$), and build the corresponding confidence set for $\boldsymbol{\theta}$ (with level $1 - \alpha$). The p-values for individual tests were approximated form a preliminary simulation (independent of the rest) using 199 replications. The joint confidence sets for (θ_1, θ_2) appear (in black) in tables 1 to 4. The triangles represent the boundaries of the region inside which the process is stable (i.e., the roots of the AR polynomial are outside the unit circle). The corresponding projection-based confidence intervals are given in the following table:

Model	M_1	M_2	M_3	M_4
θ_1	$[-0.14, 0.08]$	$[-0.42, 0.07]$	$[0.81, 1.27]$	$[0.99, 1.01]$
θ_2	$[0.91, 1.17]$	$[-0.11, 0.30]$	$[-0.22, 0.68]$	$[0.84, 1.19]$

We see from these numerical results that the confidence sets are quite precise and (as expected) cover the true parameter values. Of course, more precise confidence sets can be obtained by using larger numbers of Monte Carlo replications.[1]

References

[1] Beveridge S., Nelson C. (1981). *A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycles'*. Journal of Monetary Economics **7**, 151 – 174.

[2] Blough S.R. (1992). *The relationship between power and level for generic unit root tests in finite samples.* Journal of Applied Econometrics **7**, 295 – 308.

[3] Dickey D.A. (1976). *Estimation and testing of non stationary time series.* PhD thesis, University of Iowa.

[4] Dufour J.-M. (2002). *Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics.* Journal of Econometrics, forthcoming.

[5] Dufour J.-M. (2003). *Identification, weak instruments and statistical inference in econometrics.* Canadian Journal of Economics **36** (4), 767 – 808.

[6] Dufour J.-M., Khalaf L. (2001). *Monte Carlo test methods in econometrics.* In B. Baltagi, (ed.), Companion to Theoretical Econometrics, Blackwell Companions to Contemporary Economics, Basil Blackwell, Oxford, U.K., chapter 23, 494 – 519.

[7] Dufour J.-M., Khalaf L. (2002). *Exact tests for contemporaneous correlation of disturbances in seemingly unrelated regressions.* Journal of Econometrics **106** (1), 143 – 170.

[8] Dufour J.-M., Kiviet J.F. (1998). *Exact inference methods for first-order autoregressive distributed lag models.* Econometrica **66**, 79 – 104.

[1] Due to space limitations and the computer intensive nature of the procedures proposed above, we cannot present here a full-fledged power simulation or a detailed empirical application.

[9] Dufour J.-M., Neifar M. (2001). *Méthodes d'inférence exactes pour des processus autorégressifs: une approche fondée sur des tests induits.* L'Actualité économique **78** (1), 19 – 40.

[10] Dufour J.-M., Torrès, O. (1998). *Union-intersection and sample-split methods in econometrics with applications to SURE and MA models.* In D.E.A. Giles, A. Ullah (eds), Handbook of Applied Economic Statistics, Marcel Dekker, New York, 465 – 505.

[11] Fisher R.A. (1932). *Statistical methods for research workers.* Oliver and Boyd, Edinburgh.

[12] Imhof P.J. (1961). *Computing the distribution of quadratic forms in normal variables.* Biometrika **48**, 419 – 426. Corrigendum 49 (1962), 284.

[13] Neifar M. (1996). *Méthodes d'inférence exactes pour des modèles de régression avec erreurs autoregressives et applications macroéconomiques.* PhD thesis, Université de Montréal.

[14] Pearson K. (1933). *On a method of determining whether a sample of size n supposed to have been drawn from a parent population.* Biometrika **25**, 379 – 410.

[15] Stock J.H. (1994). *Unit root, structural breaks and trends.* In R.F. Engle, D.L. McFadden (eds), Handbook of Econometrics Volume IV, North-Holland, Amsterdam, chapter 46, 2740 – 2841.

[16] Tippett L.H. (1931). *The methods of statistics.* Williams and Norgate, London.

Acknowledgement: This work was supported by the Canada Research Chair Program (Chair in Econometrics, Université de Montréal), the Alexander-von-Humboldt Foundation (Germany), the Institut de Finance mathématique de Montréal (IFM2), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Canada Council for the Arts (Killam Fellowship), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Fonds de recherche sur la société et la culture (Québec), and the Fonds de recherche sur la nature et les technologies (Québec).

Address: J.-M. Dufour, Canada Research Chair Holder (Econometrics), Centre interuniversitaire de recherche en analyse des organisations (CIRANO), Centre interuniversitaire de recherche en économie quantitative (CIREQ), and Département de sciences économiques, Université de Montréal, C.P. 6128 succursale Centre-ville, Montréal, Québec, Canada H3C 3J7,
`http://www.fas.umontreal.ca/SCECO/Dufour`
M. Neifar, Département des méthodes quantitatives et technologie de l'information, Institut Supérieur de Gestion de Sousse, Rue Abdel Aziz Elbahi C.P.763, Sousse 4000, Tunisie

E-mail: `jean.marie.dufour@umontreal.ca`,
`malika.neifar@isgs.rnu.tn`, `malika.neifar@topnet.tn`

A KIND OF PISA-SURVEY AT UNIVERSITY

Christine Duller

Key words: Mathematical comprehension, survey.

COMPSTAT 2004 section: Teaching statistics.

Abstract: Do students have a head for figures, do they have mathematical comprehension? The samples show that most of the participating students have serious difficulties in interpreting "statistical" diagrams and also problems in very basic mathematics. In my opinion especially interpreting diagrams is very important in professional and daily life. Therefore I think we have to focus not only on technical skills in statistics, but also in correct interpretations of results and diagrams.

1 Intention for the survey

PISA is the OECD Programme for International Student Assessment, this is a three-year survey of the knowledge and skills of 15-year-olds in different countries. The survey was conducted in 2000 and in 2003, about 265.000 students from 32 countries took part. (More information about PISA on `www.pisa.oecd.org`) The results of PISA 2003 were not very glorious for austria, especially in mathematics. Therefore the question occurs, if results at university would be as bad as those at school. Do students have a head for figures, do they have mathematical comprehension?

2 The sample

The data were collected from students of economics and social science at the University of Linz, Austria, in March 2001 (n = 607), March 2003 (n = 349) and January 2004 (n = 259). The samples were no random selection, the students were chosen, because they took part in basic courses in statistics. Therefore the results are not representative for any really interesting population, but still very meaningful. The questionnaire contained eleven questions dealing with mathematics or statistics, most of them multiple choice with six different possibilities of response. The topics were interpretation and calculation (without calculator) of percentages and fractions, interpretation of graphics and (rough) estimation of square roots and percentages.

3 Results

The results at university were not very glorious, too. Even majority is not always right, as shown in some examples. The first part of this chapter shows results for the different samples (2001, 2003, 2004), the second part differentiates them to several interesting points of view.

3.1 Main results

The first question about percentages was “How much is 30% of 70%?”, the options for response were 3%, 17%, 21%, 30%, 37% and 70%. Only 75.6% of the whole sample (all percentages refer to valid cases only) did manage this question, there were hardly any differences between the sample 2001 (75.7% correct), the sample 2003 (74.6% correct) and the sample 2004 (76.8% correct).

Even worse was the result of the question “The fraction 1/40 can also be written as …” with the options 0.40, 4/100, 0.25, 0.040, 1/25 and 0.025. Only about two third (66.3%) did know the correct answer. Again there were hardly any differences between the group of 2001 (64.2% correct), the group of 2003 (66.6% correct) and the group of 2004 (70.8% correct).

The result of the question “Figure 1 shows the price-quality-index of two products. Which product would you prefer?” (see Figure 1) shows us that even majority is not always right.

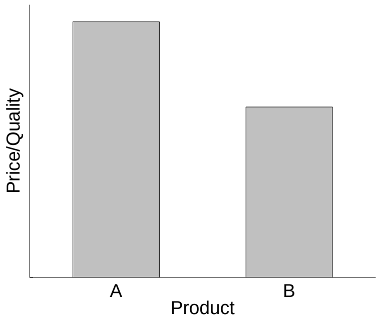


Figure 1: Which product would you prefer?

The figure shows the ratio price to quality , so product A would be more expensive for the same quality as product B, but only 40.3% of the students chose product B. There were diffenrences in the various years (2001: 38.1%, 2003: 48.4% and 2004: 34.4%), but majority was wrong in each year.

There were better results for the question “40% stands for …” with the possibilities “every fourth”, “one out of forty”, “four out of ten”, “one fourth”, “1/25” and “ten out of fourhundred”. 86.1% of the sample were able to find the right interpretation. This seems to be a good result, but on the other hand this means, that about 14% of the students do not understand anything about percentages, so how should they manage probability?

Another bad highlight was the result of the question “The square root of 0.5 is …” with the options “bigger than 0.5”, “equal 0.5” and “smaller than 0.5”. Only 50.7% of the whole sample chose the first option, 1.0% voted for the option “equal” and 48.3% voted for the last response. So it seems to be very difficult for the students to estimate the square root of a decimal number. Estimating the square root of 14641 with the options 11, 71, 121, 235, 550, 739 was also a big problem (56.9% correct answers).

For the last question students had to decide on which figure shows a bigger increase (with possibilities "the first", "the second" and "both are equal").

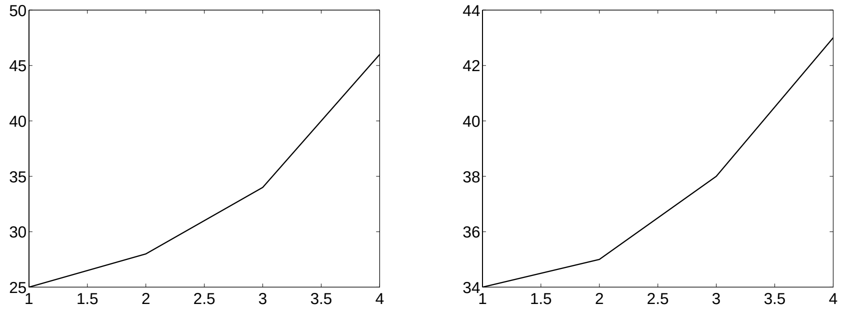


Figure 2: Which figure shows a bigger increase?

In the sample 2001 45.4% voted for the correct figure, the result of the sample 2003 was a little bit better with 51.2% correct answers, in the sample of 2004 only 45.9% chose the correct figure. This result is disappointing, because one of the topics in the basic courses in statistics is how to look at figures.

There were eleven questions dealing with mathematics and statistics, two of them concerning graphics. 19.9% of the sample did manage both of the graphical questions, 15.3% did manage all non-graphical questions, and only 4.5% gave correct answers for all eleven questions. Detailed information about the differences in the various samples is given in Table 1.

Part	**2001**	**2003**	**2004**	**All**
Figures (2 questions)	18.5%	24.3%	17.4%	19.9%
Non-figures (9 questions)	14.8%	13.5%	19.0%	15.3%
All questions	4.2%	5.1%	4.7%	4.5%

Table 1: Correct answers in different years.

3.2 Some interesting details

In this section the results are checked on differences between women and men and differences between various secondary schooltypes.

Table 2 shows results by sex for the questions mentioned above. Some results are very close for men and women, but others differ very much. For

Question/Part	**Male**	**Female**	**Total**	**Total size**
Bigger increase	54.2%	42.1%	47.1%	1205
Procuct	38.5%	41.7%	40.3%	1190
30% of 70%	78.8%	73.3%	75.6%	1188
fraction 1/40	76.4%	58.9%	66.3%	1203
40% stands for	92.3%	81.6%	86.1%	1195
$\sqrt{0.5}$	62.1%	42.4%	50.7%	1202
$\sqrt{14641}$	66.9%	49.7%	56.9%	1195
Figures	22.1%	18.4%	19.9%	1183
Non-Figures	22.9%	9.8%	15.4%	1120
All questions	7.0%	2.7%	4.5%	1103

Table 2: Correct answers by sex.

example estimating the square root of 0.5 seems to be a bigger problem for women (42.4% correct) than for men (62.1% correct). As a matter of fact only the question about the product-price-index was a smaller problem for women (41.7%) than for men (38.5%).

Students were asked about their admittance for university, with the following options (all schools are with school-leaving exams):

- AHS: Academic secondary school
- HAK: Secondary colleg for business administration
- HBLA: Advanced-level secondary vocational school, several vocations (f.e. tourism)
- HTL: Advanced-level secondary technical school
- SBL: “Studienberechtigungslehrgang” Special exam to get an admittance for a certain study at university for persons without school-leaving exam

The results for schooltype by sex is shown in Table 3.

Sex	**AHS**	**HAK**	**HBLA**	**HTL**	**SBL**	**Other**
Male	40.6%	46.7%	11.0%	88.9%	49.4%	31.4%
Female	59.4%	53.3%	89.0%	11.1%	50.6%	68.6%
Size	465	366	136	72	79	86

Table 3: Schooltypes, sex-proportion.

The results for the different questions or parts can be seen in Table 4 up to Table 6.

Question	**AHS**	**HAK**	**HBLA**	**HTL**	**SBL**	**Other**
Bigger increase	48.8%	53.0%	37.5%	47.2%	48.1%	29.4%
Procuct	39.8%	39.3%	37.0%	54.2%	39.5%	40.2%

Table 4: Correct answers in various schooltypes, graphical questions.

Question	**AHS**	**HAK**	**HBLA**	**HTL**	**SBL**	**Other**
30% of 70%	74.3%	81.2%	74.2%	80.6%	61.5%	66.7%
fraction 1/40	68.7%	69.2%	52.6%	76.4%	62.3%	54.7%
40% means	88.5%	87.3%	80.9%	88.9%	79.2%	80.0%
$\sqrt{0.5}$	53.4%	50.7%	45.9%	56.9%	43.6%	44.7%
$\sqrt{14641}$	57.0%	63.0%	43.9%	61.1%	57.1%	45.8%

Table 5: Correct answers in various schooltypes, non-graphical questions.

Part	**AHS**	**HAK**	**HBLA**	**HTL**	**SBL**	**Other**
Figures	21.4%	21.9%	13.3%	30.1%	16.0%	8.6%
Non-Figures	14.8%	17.6%	13.3%	19.7%	8.5%	11.0%
All questions	4.2%	5.8%	0.8%	11.3%	1.4%	1.4%

Table 6: Correct answers in various schooltypes.

In general HTL seems to be the school-type with the best results, followed by AHS and HAK; students with HBLA, SBL or other kind of admittance had worst results. Because of the varying sex-proportions in different schooltypes, results are divided by sex within different schooltypes.

But also within different schooltypes women had worse results than men (f.e. see Table 7), again only the question about the product-price-index was less problem for women than for men (see Table 8).

$\sqrt{14641}$	**AHS**	**HAK**	**HBLA**	**HTL**	**SBL**	**Other**
Male	66.0%	67.7%	66.7%	64.1%	75.7%	61.5%
Female	50.9%	59.0%	41.0%	37.5%	40.0%	38.6%
Total	57.0%	63.0%	43.9%	61.1%	57.1%	45.8%
Size	461	362	132	72	77	83

Table 7: Correct answers $\sqrt{14641}$ by sex and schooltype.

The next interesting point of view would be the differences in results for various schooltypes within female or male. Unfortunately it is not possible to give a detailed serious view on that point, because there are only very few

Product	**AHS**	**HAK**	**HBLA**	**HTL**	**SBL**	**Other**
Male	36.7%	36.9%	33.3%	51.6%	30.8%	38.5%
Female	41.9%	41.5%	37.5%	75.0%	48.6%	41.1%
Total	39.8%	39.3%	37.0%	54.2%	39.5%	40.2%
Size	455	361	135	72	76	82

Table 8: Correct answers product by sex and schooltype.

female in HTL (8 of 1215) and also very few male in HBLA (15). For other schooltypes results by sex had the same trend as the results not diveded by sex.

4 Summary

The samples show that most of the participating students have serious difficulties in interpreting "statistical" diagrams and also problems in very basic mathematics. In my opinion especially interpreting diagrams is very important in professional and daily life. Therefore I think we have to focus not only on technical skills in statistics, but also in correct interpretations of results and diagrams.

References

[1] Borovcnik M. (1984). *Was bedeuten statistische Aussagen.* Teubner, Wien. Schriftenreihe Didaktik der Mathematik. Band 8.

[2] Krämer W. (2002). *So lügt man mit Statistik.* 3. Auflage. Piper, München.

[3] OECD-Organisation for Economic Co-operation and Developement (2002). *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills.* OECD, Paris.

Address: C. Duller, University of Linz, Department for Applied Statistics, Altenbergerstrasse 69, 4040 Linz, Austria

E-mail: `christine.duller@jku.at`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

DISCUSSIONS IN A BASIC STATISTICS CLASS

William H. Eichhorn

Key words: Estimation, causality, observational, confidence, interval, P-value.

COMPSTAT 2004 section: Teaching statistics.

Abstract: Different topics in the study and understanding of statistics are discussed and demonstrated for students in a basic statistics class. They include:
Criticism of media reports:

- On the effect of prescription drugs on safe driving.
- On what roll Ohio plays in presidential elections.

Provoking thoughts:

- How many decimal places in a recorded measure?
- A different view of unbiased estimation.
- Judging estimates from the performance of the estimators.
- Observational vs. experimental studies.
- Another way of introducing confidence intervals with floating transparencies.
- Finding the significance level of a test.
- Finding the rejection region in a test of hypothesis.
- Computer produced P-values and the significance level of a test.

1 Prescription drugs vs. alcohol

In the news: Prescription drugs may affect driving ability more than alcohol. This conclusion was derived from the following study. A large number of drivers which were involved in accidents were chosen at random. The drivers were tested, their blood was checked for having taken any of certain prescription drugs, or for having an alcohol level over the legal limit.

It was found that among those drivers that were involved in accidents, there were more "under the influence" of prescription drugs than "under the influence" of alcohol.

It was concluded that the chance of being involved in an accident is higher when driving while using one of these prescription drugs than it is when driving under the influence of alcohol.

Students bring up the following argument questions:

1. Who is to blame for the accident?
2. Why not separate between different drugs?

3. Why don't we consider the ages of the drivers?
4. We should consider road conditions, weather, time of day.
5. How much driving experience did each driver have?

We show that all these considerations can be ignored in a random statistical study. Differences will be part of the error term. The mistake that was made by the media in this case was to consider the wrong ratios.

We want to compare the conditional probabilities of having an accident when:

a) Under the influence of prescription drugs, and when:
b) Under the influence of alcohol.

We have here the wrong ratios: We found that:

The number of people under prescription drugs over the number of people in accidents is larger than the number of people under influence of alcohol over the number of people in accidents, instead of the number of people in accident over the number of people under influence of . . . etc.

Example: Hypothetical study in Sweden.
Among a large number of people who suffered a heart attack it was found that 95 % had blue eyes. Is having blue eyes a risk factor?
Students suggested to conduct a study on a *simulator* using either *2 independent* samples, or *matched pairs* – the same sample *once* under the influence of *drugs* and *once* under the influence of *alcohol*.

As Ohio goes, so goes the Nation

2 Confusing association with causality

Here is another example of drawing the wrong conclusion from statistical information:
In 23 of the last 25 presidential elections, Ohio has voted for the winner. Knowing this fact, both candidates are trying very hard to bring Ohio to their side, hoping that winning Ohio will lead them to an overall victory in the Presidential race.
Ohio has 21 electoral votes but other states have more, yet Ohio's results are seen as a predictor of the results of the whole nation and that is why the candidates feel that Ohio is so important to them.
The question is, does it make sense to try harder in Ohio?
Will an intervention to change the vote in Ohio lead to a change of the vote for the whole nation? Probably not!
It is like noting that most basketball players are very tall, so if one wants his son to be tall he should teach him to play basketball.

3 Order of magnitude and data reading

When data is recorded, how many decimal places are appropriate?
At my doctor's office, they put me on the scale in my clothes with my shoes on, then they record my weight up to an ounce. Should they not round it off?
I tell the story about the two friends at the British museum looking at the dinosaurs. One says: this dinosaur is 50 million and 3 years old.
"How do you know?" asked his friend.
"I was here 3 years ago and they told me then that it was 50 million years old."
Also on the news, a far galaxy was discovered 10,000 light years away from the United States. How far is it from Rider University?

4 Unbiased estimation

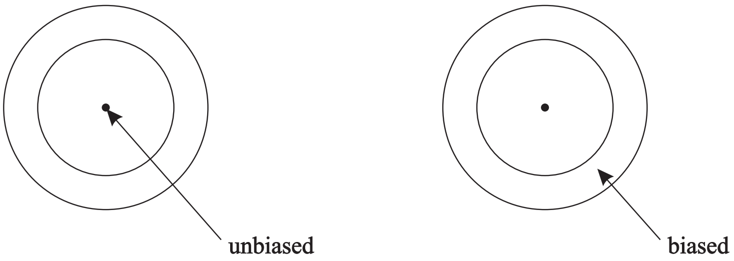
Estimator $\hat{\theta}$ of θ is unbiased if

$$E(\hat{\theta}) = \theta\,.$$

If $E(\hat{\theta}) - \theta = B \neq 0$ it is biased.

The larger B is, it seems the larger the error in estimation.

It looks like we aim at a target, unbiased estimators aim at the center of the target. Biased estimators aim at a distance B from the target.

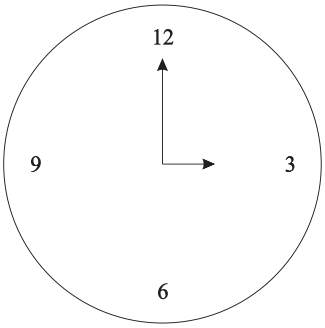


So is it always better to have an unbiased estimator?

If we use the mean square error MSE $= E(\hat{\theta}-\theta)^2$ as a criterion to judge the efficiency of an estimator, we know that biased estimator can have a smaller MSE than unbiased ones.

So there must be an additional reason to prefer unbiased estimators and aspire to have a MVUE. This is the connection between the distribution of the estimator $\hat{\theta}$ and the value of the parameter θ. Let us consider the following example:
In a class each student has a watch which is almost accurate. Let us assume that the average time of all students' watches is precise.
If X is the time on the watch of a randomly chosen student, then X is an unbiased estimator of the time T.

On the wall we have a clock which has stopped and shows 3:00. Let Y denote the time on this clock.
Y is a biased estimator of T.
Is X *always* a better estimator of T than Y?



Is X *always* closer to T than Y?
Not **always**.
Twice a day, at 3:00 a.m. and 3:00 p.m., Y is going to be closer to T than X. Yet as an estimator Y is useless. You need to know T to tell whether Y is the better estimate.
Y has nothing to do with T. It does not change when T changes.
One important (and usually overlooked) property of an unbiased estimator is that it "moves along" with the parameter.
When T changes, the center of the distribution of X changes with T.
We chose to obtain this property by requiring that the expectation of $\hat{\theta}$ be equal θ.
This property would be true also if we asked for the median of $\hat{\theta}$'s distribution to equal θ, but this will make the mathematics much harder.

This choice *rules out* those estimators, like Y, that have "nothing to do" with the parameter. We can this way justify using "s" to estimate "σ" even if it is not unbiased, it still has the property of "moving along" with the parameter.

5 Estimators and estimates

When we have a random sample from a normal population, the sample mean $\overline{X}$ is a MVUE of μ, the sample median $\tilde{X}$ is also an unbiased estimator of μ but has a larger variance.

From such a sample we have the following results:

$$\overline{x} = 185 \quad \tilde{x} = 174$$

Questions:

a) Which of these values is closer to μ?
b) Which one should we use to estimate μ?

The answer to a) is, we don't know.
The answer to b) is we choose $\overline{x} = 185$.
This shows that we choose an *ESTIMATE* by the performance of the *ESTIMATOR.*

6 Observational study and "cause and effect"

In a hypothetical study in San Diego, a large number of males were observed. The subjects were divided into two groups:

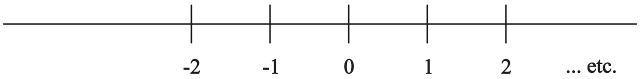
1. Those who swam regularly in the ocean.
2. Those who did not.

They were observed as to how much hair did they lose on their heads. It turned out that those who swam regularly in the ocean kept more of their hair than those who did not. Can we conclude that swimming in the ocean helps you keep your hair? Clearly not, this is an observational study and probably those who swam were younger and kept more of their hair.
How do we know?
Story: A man dreamt that he was being executed. At the moment that he saw the guillotine fall, he was so shocked that he got a heart attack and died. This story does not sound real, had this happened how could we know about his dream? Compare it to the following item in the news:
Seven million Americans have diabetes but do not know it. We can question this statement in a similar way: If they don't know it, how can we know? The answer is, of course, that this is an estimate based on statistical analysis.

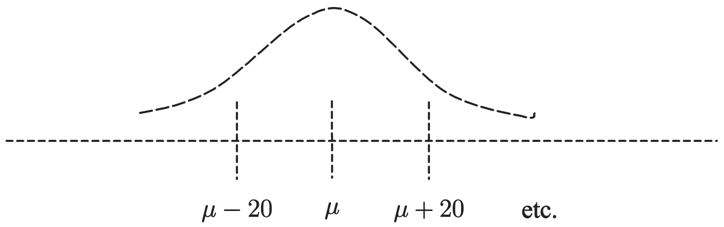
7 Confidence intervals

(Red lines will be represented by dashed lines in this printing.) Given a sample of $n = 100$ from a population with mean μ (unknown) and $\sigma = 200$. Find A $1 - \alpha = .95$ confidence intervals for μ.

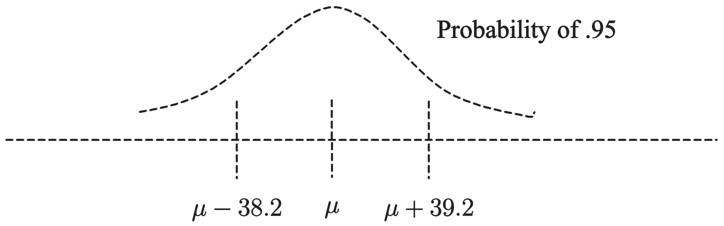
Let us consider the black line indicating the x axis with values marked on it.



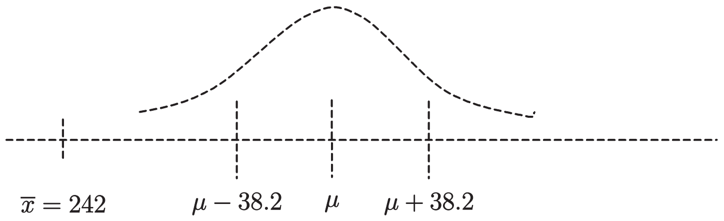
As the value of μ is unknown, we are going to mark it in red on a "floating" transparency, that can float over the black line. $\overline{X}$ the sample mean will have a normal distribution with mean μ (unknown) and standard deviation of $\frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{100}} = 20$ it's distribution can be marked around μ on the red transparency.



A $1-\alpha = .95$ prediction interval for $\overline{X}$ will be a symmetric interval around μ of size $2 \cdot (1.96)\frac{\sigma}{\sqrt{n}} = 78.4$ or $\mu \pm 39.2$. We take a transparency with the prediction interval marked on it in red and let it slide over the x axis.



Let $\overline{x}$ be observed to equal 242. We mark it on the black line, put the transparency above it. Is $\overline{x}$ inside the prediction interval?



We don't know, but have strong confidence that it is. If $\overline{x}$ is in the prediction interval μ has to be inside the interval $\overline{x} - 38.2$, $\overline{x} + 38.2$ which is the confidence interval.

8 Testing hypotheses

You own a company that uses batteries. You are offered a new brand of batteries, you are willing to buy them if the batteries' mean lifetime, μ, is longer than 8 hours.

After consulting a statistician (for \$ 1,000) you set up a test of the null hypothesis $H_0 : \mu \leq 8$ vs. $H_1 : \mu > 8$.

You instruct your employees to take a large sample of n batteries and check their lifetime. Then compute the sample mean $\overline{x}$ and variance s^2. You take this data back to the statistician for a decision.

The employees gave you their results, $\overline{x}$ was 8.5 hours s^2 and n were given. The statistician said, "Go ahead buy the new brand."
When you come back, the employees told you that there was a mistake, $\overline{x}$ should have been 8.7 hours but s^2 was correct. Should you go back to the statistician? (\$ 1,000)?

9 Type I error

You are given the following test.

$$H_0 : \mu \leq 100$$
$$H_1 : \mu > 100$$

n is large and σ given.

Your test will reject H_0 when $\overline{x} \geq 102.5$. If $\mu = 98$, you can compute the probability of making a type I error.
If $\mu = 99$, you can show that this probability will be larger.

What about the probability of making a type I error when $\mu = 101$? Will it be larger? Or? *Why should we determine the significance level α before finding the P-value of a test?* If we test a null hypothesis say

$$H_0 : \mu = \mu_0$$

We determine a significance level α. We find the observed significance probability, P-value and reject H_0 if the P-value is less than α. The probability of making a type I error is going to be α. However, when computers are used to perform the test, we are not required to determine the significance level α and feed it into the computer.

The computer will give us the p-value of our test and *leaves to us* the decision whether to reject H_0.

This may enable us to *manipulate* the results. If the p value falls in a moderate range, say between 0.01 and 0.05 we could then set our α to either reject or accept H_0 to suit our interest. We then may claim this to be the type I error probability. If our interest is in rejecting H_0, our stated type I error may actually be less than the true one, on average as small as *one half of it.*

$$\text{Let} \qquad H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

If we choose α to be 0.05 and if H_0 is true, we still have a 0.05 probability of rejecting H_0. If this happens (given we reject H_0), the P-value will have a uniform distribution over 0 to 0.05 and its expected value will be 0.025.

If we have the intention to reject H_0 and be willing to do it if the P-value is less than 0.05, we'll on average claim an α of 0.025.

Address: B.H. Eichhorn, Rider University, 2083 Lawrenceville Road,
Lawrenceville, NJ 08648

E-mail: Eichhorn@rider.edu

COMPSTAT'2004 Symposium

FAST CROSS-VALIDATION IN ROBUST PCA

Sanne Engelen and Mia Hubert

Key words: Cross-validation, robustness, fast algorithm.

COMPSTAT 2004 section: Partial least squares.

Abstract: One of the main issues in Principal Component Analysis (PCA) is the selection of the number of principal components. To determine this number, the Predicted Residual Error Sum of Squares (PRESS) value is frequently used [2], [8]. It can be computed on a validation set, but such a new data set is not always available and many data sets in chemometrics or bioinformatics have a too small size to split into a training set and a validation set. Therefore, a cross-validated version of the PRESS statistic can be used, which is obtained by removing one observation at a time from the total data set. This technique is however rather time consuming. The computational complexity increases even more when robust PCA methods are used, such as the MCD estimator [12] for low-dimensional data or the ROBPCA method [4] for high-dimensional data. In this paper we introduce faster algorithms to compute the cross-validated PRESS value for these two methods. We evaluate the developed procedures by means of simulated and real data in both low and high dimensions. We also extent the methodology to high-breakdown regression methods such as the LTS estimator [15], MCD-regression [13], robust principal component regression [7] and robust PLS regression [6].

1 Introduction

The comparison of different estimators is a very important issue in statistics. One possibility is to compare them based on their predictive ability. For this purpose the Predicted Residual Error Sum of Squares (PRESS) is very well suited (see e.g. [1] and [11]). In general it is defined as the sum of the squared residuals from a validation set. The model parameters on the other hand are estimated from an independent training set. However, a validation (or test) set is not always available and many data sets in chemometrics or bioinformatics have a too small size to split into a training set and a validation set. Therefore, a cross-validated version of the PRESS statistic can be used. To compute the residual of case i, this observation is first removed from the data set before the parameters are estimated. This one-fold cross-validation is very popular, but is very time consuming. Only in very specific cases, like linear regression, the cross-validated residuals can be calculated using closed formulas. Therefore one-fold cross-validation is usually applied to smaller data sets. Note that at larger data sets, m-fold cross validation is a valuable alternative.

In this paper we concentrate on two robust PCA methods: the MCD estimator [12] for low-dimensional data and the ROBPCA method [4] for high-dimensional data. Contrary to classical PCA, they are resistant to outliers in the data. Although their computation time is reasonable (e.g. it takes 4.16 seconds to run ROBPCA on a data set with $n = 180$ observations in $p = 750$ dimensions on a Pentium IV with 2.4 GHz), it is no longer feasible to execute the full algorithm n times. Fast cross-validation is thus certainly needed for these robust methods.

In the next section, we first define a robust cross-validated PRESS value. In Section 3 we introduce fast algorithms for its computation. In Section 4 and 5 we illustrate the performance of our method by means of simulations and examples.

Matrices will be denoted by capital letters. Our data matrix $X_{n,p}$ has n observations and p dimensions. A vector is always indicated with a bold symbol e.g. $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$ stands for the ith observation.

2 A robust PRESS value

PCA [8] is a well-known dimension reduction technique where a k-dimensional loading matrix $P_{p,k}$ and scores $\boldsymbol{t}_i$ are constructed such that

$$\boldsymbol{t}_i = P'_{k,p}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}) \tag{1}$$

with $\hat{\boldsymbol{\mu}}$ an estimate of the center of the data. In classical PCA, $\hat{\boldsymbol{\mu}}$ is given by the mean of the data and the k loading vectors are the eigenvectors of the empirical covariance matrix of the data that belong to the k largest eigenvalues. Details of the construction of $\hat{\boldsymbol{\mu}}$ and P for robust PCA methods are given in Section 3 for the MCD-algorithm and in reference [4] for the ROBPCA method. An estimate of $\boldsymbol{x}_i$ in the k-dimensional PCA-space is given by

$$\hat{\boldsymbol{x}}_{i,k} = P_{p,k}\boldsymbol{t}_i + \hat{\boldsymbol{\mu}}. \tag{2}$$

The cross-validated PRESS_k value is then defined as

$$\text{PRESS}_k = \sum_{i=1}^{n} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_{-i,k}\|^2 \tag{3}$$

with $\hat{\boldsymbol{x}}_{-i,k}$ the estimate of the ith observation based on a PCA model with k components constructed from the $n-1$ other samples.

Even if the fitted values $\hat{\boldsymbol{x}}_{-i,k}$ in (3) are based on a robust PCA method, the PRESS value is not robust as it also includes the prediction error of the possible outliers. A robust version of the PRESS is obtained by adding weights to each observation:

$$\text{R-PRESS}_k = \sum_{i=1}^{n} w_i \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_{-i,k}\|^2. \tag{4}$$

These weights w_i are determined before the cross-validation is started. Based on the PCA estimates on the full data set for $k = 1, \ldots, k_{max}$ components, outliers for every model under investigation are marked. For details about the definition of an outlier in PCA, we refer again to [4]. If an observation is an outlier in one or more models, its weight w_i equals 0. Samples that are never detected as an outlier, obtain weight $w_i = 1$. Doing so, the robust PRESS$_k$ value is based on the same set of observations for each k. This definition is similar to the robust RMSECV value that is defined for robust PCR [7] and robust PLS regression [6].

3 Fast cross-validation

The MCD (Minimum Covariance Determinant) estimator [12] is a highly robust method to estimate multivariate location and scatter parameters when the number of variables p is smaller than half the number of samples n. The principal components can be considered as the eigenvectors of the robust covariance estimate. ROBPCA [4], ROBust Principal Components Analysis, on the other hand is a technique where projection pursuit ideas are combined with the MCD estimator and is in particular appropriate for high-dimensional data.

In order to explain how the calculation of the PRESS values can be speeded up for MCD, we first describe the original FAST-MCD algorithm [14]. Then we present the adapted version and indicate where some time improvements of the original procedure are made. For a detailed description of the ROBPCA algorithm and its changes towards fast cross-validation (which are comparable with those of MCD), we refer to [4] and [3]. Here, we will only show the numerical results.

3.1 The MCD estimator

The objective of the raw MCD is to find $h > \frac{n}{2}$ observations out of n whose covariance matrix has the smallest determinant. Its breakdown value is $\frac{[n-h+1]}{n}$, hence the number h determines the robustness of the estimator. For its computation, the FAST-MCD algorithm [14] can be used. It roughly proceeds as follows :

1. Many random $(p+1)$-subsets are drawn, which are enlarged to initial h-subsets using a C-step as explained in the next step. If it is computationally feasible, all possible $(p+1)$-subsets are used. Else 500 $(p+1)$-subsets are drawn.

2. Within each h-subset, two C-steps are performed. Basically a C-step consists of computing first the classical center $\hat{\boldsymbol{\mu}}_0$ and the classical covariance matrix $\hat{\Sigma}_0$ of the h observations. Then the robust distance

(which depends on $\hat{\boldsymbol{\mu}}_0$ and $\hat{\Sigma}_0$) of each point is computed as:

$$\mathrm{RD}_{\hat{\boldsymbol{\mu}}_0,\hat{\Sigma}_0}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0)'\hat{\Sigma}_0^{-1}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0)}. \tag{5}$$

A new h-subset is formed by the h observations with smallest robust distance.

3. For the 10 h-subsets with the best value for the objective function, C-steps are performed until convergence. Other time saving techniques can be applied and are described in [14].

4. The raw estimates of location $\hat{\boldsymbol{\mu}}_{\text{raw}}$ and scatter $\hat{\Sigma}_{\text{raw}}$ are the classical mean and classical covariance matrix (multiplied by a consistency factor) of the h-subset H_0 with lowest objective function.

5. Next, a reweighting step is applied for efficiency purposes. Every observation is multiplied by a weight based on its robust distance $\mathrm{RD}_{\hat{\boldsymbol{\mu}}_{raw},\hat{\Sigma}_{raw}}(\boldsymbol{x}_i)$. When this squared distance is larger than the 0.975 quantile of the χ^2_k distribution, the weight is set equal to 0 and else to 1. The classical mean and covariance matrix of the weighted observations are the final robust center $\hat{\boldsymbol{\mu}}_{MCD}$ and scatter matrix $\hat{\Sigma}_{MCD}$.

6. Finally, the principal components are defined as the k eigenvectors of $\hat{\Sigma}_{MCD}$ which belong to the k largest eigenvalues of $\hat{\Sigma}_{MCD}$. These principal components are stored in the loading matrix $P_{p,k}$. Analogously to (2), an estimate of $\boldsymbol{x}_i$ in the k-dimensional space spanned by these components is given by

$$\hat{\boldsymbol{x}}_{i,k} = P_{p,k}P'_{k,p}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{MCD}) + \hat{\boldsymbol{\mu}}_{MCD}. \tag{6}$$

3.2 The approximate algorithm

In order to compute the robust PRESS value (4), we need $\hat{\Sigma}_{-i}$ and $\hat{\boldsymbol{\mu}}_{-i}$, which are the MCD estimates of the covariance matrix and center of the data set without observation i. In the naive approach the FAST-MCD algorithm is fully applied n times. This takes a lot of time, as the algorithm is based on random resampling (see step 1). Therefore we have developed an approximate estimate for $\hat{\Sigma}_{-i}$ and $\hat{\boldsymbol{\mu}}_{-i}$. First note that $\hat{\Sigma}_{-i}$ and $\hat{\boldsymbol{\mu}}_{-i}$ are the MCD estimates for a data set with $n-1$ observations. In order to retain a breakdown value as close as possible to the breakdown value of the full MCD, being $\frac{[n-h+1]}{n}$, we define the raw MCD estimator on $n-1$ points as the mean and covariance matrix of the $h-1$ observations with smallest covariance determinant.

The approximate algorithm proceeds as follows:

1. Perform the MCD algorithm on the whole data set. We store the h-subset H_0, the center $\hat{\boldsymbol{\mu}}_{raw}$ and covariance matrix $\hat{\Sigma}_{raw}$ before weighting and the center $\hat{\boldsymbol{\mu}}_{MCD}$ and covariance matrix $\hat{\Sigma}_{MCD}$ after weighting.

2. Repeat the next steps for each sample $i = 1, \cdots, n$.

   (a) We remove sample i from the set of n observations.

   (b) Now we have to find the $(h-1)$-subset $H_{-i,0}$ with lowest objective function. Instead of obtaining it by resampling, we just use an update of H_0. This yields an approximate, but a very fast solution. When H_0 contains the ith observation, we take the remaining $(h-1)$ points of H_0. On the other hand, when sample i does not belong to H_0, the $(h-1)$ points of H_0 with the smallest robust distance $\mathrm{RD}_{\hat{\boldsymbol{\mu}}_{raw},\hat{\Sigma}_{raw}}(\boldsymbol{x}_i)$ are used to form $H_{-i,0}$. Denote $\boldsymbol{x}_r$ as the observation which has been removed from H_0, or $H_{-i,0} = H_0 \backslash \{\boldsymbol{x}_r\}$. Remark that for all observations i outside H_0, $H_{-i,0}$ needs to be computed only once.

   (c) Next, we compute $\hat{\boldsymbol{\mu}}_{-i,0}$ and $\hat{\Sigma}_{-i,0}$ as the mean and covariance matrix of the $(h-1)$ points from $H_{-i,0}$. This can be performed quickly using updates of $\hat{\boldsymbol{\mu}}_{raw}$ and $\hat{\Sigma}_{raw}$:

   $$\begin{aligned}\hat{\boldsymbol{\mu}}_{-i,0} &= \frac{n}{n-1}(\hat{\boldsymbol{\mu}}_{raw} - \frac{1}{n}\boldsymbol{x}_r)\\ \hat{\Sigma}_{-i,0} &= \frac{n-1}{n-2}\hat{\Sigma}_{raw} - \frac{n-1}{n(n-2)}\Big((\hat{\boldsymbol{\mu}}_{-i,0} - \mathbf{x}_r)(\hat{\boldsymbol{\mu}}_{-i,0} - \mathbf{x}_r)^t\Big)\end{aligned}$$

   (d) To improve this solution, we apply two C-steps starting from $\hat{\boldsymbol{\mu}}_{-i,0}$ and $\hat{\Sigma}_{-i,0}$, yielding $\hat{\boldsymbol{\mu}}_{-i,raw}$ and $\hat{\Sigma}_{-i,raw}$.

   (e) Finally, we perform a reweighting step based on $\hat{\Sigma}_{-i,raw}$ and $\hat{\boldsymbol{\mu}}_{-i,raw}$ as described in step 5 of the MCD algorithm. This yields $\hat{\boldsymbol{\mu}}_{-i,MCD}$ and $\hat{\Sigma}_{-i,MCD}$ whose k dominant principal components are stored in P_{-i}.

   (f) As in equation (6) an estimate of $\boldsymbol{x}_{i,k}$ is then given by:

   $$\hat{\boldsymbol{x}}_{-i,k} = P_{-i}P'_{-i}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{-i,MCD}) + \hat{\boldsymbol{\mu}}_{-i,MCD}$$

3.3 The ROBPCA method

Very shortly written, the ROBPCA method proceeds as follows. First, a singular value decomposition is performed on the data in order to reduce their data space to the affine subspace spanned by the n observations. In a next step a measure of outlyingness is computed for every point. The h observations with smallest outlyingness are collected into H_0. Then, all the data are projected onto the k-dimensional subspace spanned by the dominant eigenvectors of the covariance matrix of the points in H_0. Finally a slightly adapted version of MCD is performed. This results in a loading matrix $P_{p,k}$ and an estimate of $\boldsymbol{\mu}$ from which the fitted value (2) and the robust R-PRESS can be computed.

To perform fast cross-validation with ROBPCA we apply similar time reduction techniques as for the MCD estimator in the different stages of the algorithm. Details of this approach are described in [3].

4 Simulations

We have performed many simulations to investigate the time reduction and the precision of the approximate MCD and ROBPCA algorithms. Because of lack of space, we only report one simulation setting. The results of the simulations are shown graphically by plotting the R-PRESS curves for the naive and the approximate algorithm in one figure. The curve marked with a $\bullet$ symbol represents the approximate method, while the one with the $\times$ markers stands for the naive approach. If the approximate method works well, the curves should be close to each other. The time to run the program is also stored.

For the MCD approach we have simulated a data set of $n = 100$ observations in $p = 10$ dimensions. The data were generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\Sigma = diag(10, 9, 7.5, 5, \ldots)$. The dots indicate eigenvalues that are negligibly small. So the optimal PCA-space has dimension $k = 4$. We have also generated 10% outliers in the data.

To test the ROBPCA algorithm, we have generated a 100×500 data matrix from a multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and Σ a diagonal matrix with eigenvalues (10, 7.5, 5, 3,...) where the dots indicate again very small numbers. Again 10% outliers were included in the data.

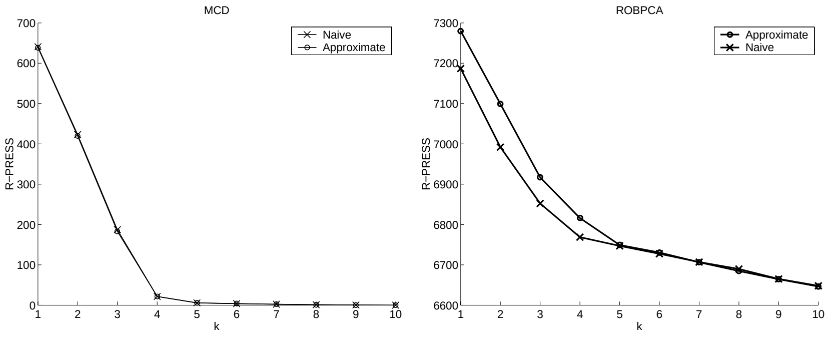
The resulting curves for MCD can be found in Figure 1(a). It took 4.25 seconds to run the approximate method versus 214.25 seconds for the naive approach. The curves for the naive and approximate ROBPCA algorithm can be found in Figure 1(b). Here the approximation only needed 23.9 seconds versus 4251.5 seconds in case of the naive approach.

We see that both curves are close to each other (for MCD they are even indistinguishable), whereas we made a huge reduction in the computation time.

5 Examples

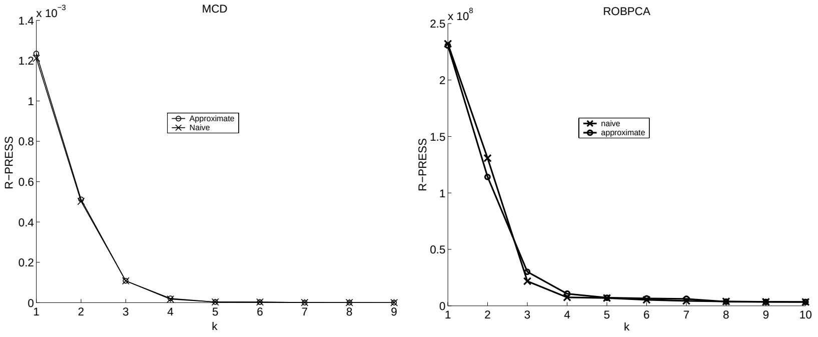
We illustrate the accuracy of the approximate methods by means of two data sets. The MCD algorithm is tested on the *Fish* data [10]. This data set consists of highly multicollinear spectra at $p = 9$ wavelengths. Measurements are taken for $n = 45$ animals. For ROBPCA we use the *Glass* data, which contain EPXMA spectra over $p = 750$ wavelengths collected on $n = 180$ different glass samples [9]. The naive and approximate R-PRESS curves are shown in Figure 2 and are again very similar.

For the Fish data, it takes 2.84 seconds to run the approximate method,



(a) (b)

Figure 1: The naive and approximate R-PRESS curves for (a) the MCD estimator and (b) the ROBPCA method.



(a) (b)

Figure 2: R-PRESS curves for (a) the fish data and for (b) the glass data.

while 70.52 seconds are needed for the naive one. Also for the Glass data set the computation time of the approximate PRESS values (85.25 seconds) is much more favorable than for the naive method which takes 9969.7 seconds.

6 Conclusions and outlook

The simulations and examples show us that the approximate techniques lead to almost the same cross-validated R-PRESS curves as the naive ones, but in a much faster way.

We have also constructed fast cross-validation methods for several regression estimators such as the LTS estimator [15], MCD-regression [13], the robust PCR [7] and the robust PLS regression [6] method. They will be described in a forthcoming paper.

References

[1] Baibing L. et al (2002). *Model selection for partial least squares regression.* Chemometrics and Intelligent Laboratory System **64**, 79 – 89.

[2] Eastment H.T., Krzanowski W.J. (1982). *Cross-validatory choice of the number of components from a principal component analysis.* Technometrics **24**, 73 – 77.

[3] Hubert M., Engelen S. (2004). *Fast cross validation for high breakdown resampling algorithms.* In preparation.

[4] Hubert M., Rousseeuw P.J., Vanden Branden K. (2002). *ROBPCA: a new approach to robust principal component analysis.* To appear in Technometrics. Available at *http://www.wis.kuleuven.ac.be/stat/robust.html.*

[5] Hubert M., Rousseeuw P.J., Verboven S. (2002). *A fast robust method for principal components with applications to chemometrics.* Chemometrics and Intelligent Laboratory Systems **60**, 101 – 111.

[6] Hubert M., Vanden Branden K. (2003). *Robust methods for Partial Least Squares regression.* Journal of Chemometrics **17**, 537 – 549.

[7] Hubert M., Verboven S. (2003). *A robust PCR method for high-dimensional regressors.* Journal of Chemometrics **17**, 438 – 452.

[8] Joliffe I.T. (2002) *Principal Component Analysis, 2nd edition.* Springer-Verlag, New York.

[9] Lemberge P., De Raedt I., Janssens K.H., Wei F., Van Espen P.J. (2000). *Quantitative Z-Analysis of the 16 - 17th Century Archaelogical Glass Vessels using PLS Regression of EPXMA and -XRF Data.* Journal of Chemometrics **14**, 751 – 763.

[10] Naes T. (1985). *Multivariate calibration when the error covariance matrix is structured.* Technometrics **27**, 301 – 311.

[11] Ronchetti E., Field C., Blanchard W. (1997). *Robust linear model selection by cross-validation.* Journal of the American Statistical Association **92**, 1017 – 1023.

[12] Rousseeuw P.J. (1984). *Least median of squares regression.* Journal of the American Statistical Association **79**, 871 – 880.

[13] Rousseeuw P.J., Van Aelst S., Van Driessen K., Agulló A. (2004). *Robust multivariate regression.* Technometrics, to appear. Available at *http://www.agoras.ua.ac.be.*

[14] Rousseeuw P.J., Van Driessen K. (1999). *A fast algorithm for the minimum covariance determinant estimator.* Technometrics **41**, 212 – 223.

[15] Rousseeuw P.J., Van Driessen K. (2000). *An algorithm for positive-breakdown methods based on concentration steps.* In Data Analysis: Scientific Modeling and Practical Application (W. Gaul, O. Opitz, and M. Schader, eds.), Springer-Verlag, New York, 335 – 346.

Address: S. Engelen, M. Hubert, Katholieke Universiteit Leuven, Department of Mathematics, W. De Croylaan 54, B-3001 Leuven, Belgium

E-mail: {sanne.engelen,mia.hubert}@wis.kuleuven.ac.be

AN APPLICATION TO LOGISTIC REGRESSION WITH MISSING LONGITUDINAL DATA

Manuel Escabias, A.M. Aguilera and M.J. Valderrama

Key words: Functional principal components, logistic regression, longitudinal data, B-spline basis.

COMPSTAT 2004 section: Functional data analysis.

Abstract: The purpose of this paper is to model a binary response variable as lupus sprout in terms of functional data given by daily stress level. We will board the importance of using a functional basis approximation for longitudinal data when there is missing data and different time points of observation for the subjects of a sample. On the other hand we will state too the importance of estimating accurately the parameter function of a functional logistic regression model by using functional principal component analysis. Finally we will interpret the relationship between the response and the predictor functional variables from the estimated parameter function.

1 Introduction

Recent years functional models have emerge as an alternative to historical methods for modelling longitudinal data. These data consist on a set of observations of a single variable repeatedly over the same subject at different time points, what makes them to be correlated. In order to use these data with the objective of predicting a response, different methods, that take into account the dependence framework between the observations and include this dependence into the corresponding model, have been used. See for example Frank and Friedman [8] for multiple response regression or Marx and Eilers [11] and Liang and Zeger [10] for the case of generalized linear models. On the other hand functional models consider longitudinal data as observations of smooth curves in some time points and include them into a model after reconstructing their functional form. We can find many situations like this in literature as for example in Zeger and Diggle [16] who consider that both response and predictor are functional variables. In this line of study Valderrama et al. [15] make a review of principal component prediction (PCP) models to forecast a functional variable in a future period from its recent past. On the other hand Ferraty and Vieu [7], Ratcliffe et al. [14], James [9], and Escabias et al. [6] introduce different types of functional regression models by considering only the predictor as functional variable. In this paper we deal with the last situation. A general overview on functional data analysis (FDA) and its applications can be seen in Ramsay and Silverman [12] and [13]. More details about functional forecasting are given in Bosq [4].

Most of the cited papers, and many others in literature, consider equally spaced time observations of the functional predictor or at least the same number of observations at the same time points on each subject. This situation is not usual because of the existence of missing data or simply the impossibility of observing a variable a specific day on certain subject. An application of PCP model with unequally spaced data has been developed in Aguilera et al. [2]. Opposite to many papers we consider that the predictor functional variable can be observed at different time points for each individual as in Besse et al. [3]. In order to introduce this kind of functional data in a regression model we use a least squares approximation of each individual functional observation in terms of basic functions, that avoids the need of observing the functional variable at the same time points for all subjects. In this sense we could consider the methodology proposed in this paper as a method of working with missing data in functional data different to other possible solutions as for example the EM algorithm. Moreover, we do not have to remove incomplete observations and we use all the available information.

We focus our attention on the functional logistic regression model with the scope of modelling lupus patient data. The functional data consist on observations of the stress level for lupus patients during eighteen days before a blood test has been made. Many patients have observations of the stress level for all days and the others have less than eighteen with the missing values at different time points. The objective of this study is to determine if the stress level is predictive of the occurrence of a lupus sprout. Then, the response variable will take the value one when an individual has had a sprout and zero in other case.

In literature we can find examples of functional logistic regression [9] and [14], most of them focused on the prediction of the response variable. In order to estimate this model it is usual to transform the functional model into a multiple one after considering the sampled functional observations expressed in terms of the elements of a basis of functions. As it is shown in Escabias et al. [6] the fitted model obtained in this case provides good predictions and goodness of fit measures (correct classification rate (CCR) or deviance) but unfortunately the estimated parameter function can be unaccurated because of multicollinearity. This fact affect to the interpretation of the parameter function in terms of odds ratios. To solve this problem we use a principal component based solution in order to estimate accurately the parameter function of the functional logistic regression model and interpret the true relation between stress level and lupus sprout.

2 Reconstructing the functional form of data

As we have stated before, longitudinal data are observations of a single variable for a specific subject of a sample at different time points $x_{i1}, x_{i2}, \ldots, x_{im_i}$ $i = 1, \ldots, n$. In order to use longitudinal data for predicting a response variable Y, we could use a standard regression method with longitudinal data as predictors (see for example Liang and Zeger [10]). In this situation it would be necessary to have at least the same number of observations for each subject of the sample ($m_i = m, \ \forall i$). The functional solution to this situation consists on considering longitudinal data as observations of a set of curves at different time points, that is, for a subject $x_{ik} = x_i(t_{ik}), \ k = 1, \ldots, m_i, \ i = 1, \ldots, n$ with $x_i(t)$ its corresponding curve, and to propose a functional model with functional predictor $x_i(t)$. Because of the impossibility of observing each $x_i(t)$ continuously we have to reconstruct the functional form of each curve from the observations $x_{i1}, x_{i2}, \ldots, x_{im_i}$. Let us consider that each curve can be expressed in terms of a basis of functions $\{\phi_1(t), \ldots, \phi_p(t)\}$ as $x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t)$ and that there is some error in the observations $x_{ik} = x_i(t_{ik}) + \varepsilon_k$, with ε_k normally distributed with zero mean and constant variance. Then we could reconstruct their functional form by least squares with the basis coefficients obtained as $a_i = (a_{i1}, \ldots, a_{ip})' = (\Phi'\Phi)^{-1}\Phi' x_i$ with $\Phi_{m_i \times p} = (\phi_j(t_{ik}))$ and $x_i = (x_{i1}, x_{i2}, \ldots, x_{im_i})'$. As we obtain the basis coefficient independently for each sampled observation, we have not to consider the same number of observations in all of them and not too that they are observed in the same nodes t_k.

In order to express the sample curves in terms of a basis of functions we can use different type of basis. For example Ratcliffe et al. [14] use Fourier basis and James [9] uses B-spline basis. In this sense, the nature of the sampled curves ought to guide us to select the best type of basis. In order to model the lupus data we have used cubic B-spline basis defined by a set of nodes, $\tau_0 < \ldots < \tau_q$, because cubic splines are smooth functions with good local behaviour. In this case the basis dimension will be $q + 3$. In the application of Lupus patient that we include at the end of this paper we have used eight definition nodes in which case the basis dimension has been ten.

3 Principal component functional logistic regression model

As we have stated in the previous section, longitudinal data are commonly used to model the relationship between a response variable and a functional predictor by using a functional model. If the response is a binary variable, the most used model is the functional logistic regression model.

In order to formulate the model let us consider $x_1(t), \ldots, x_n(t)$ a sample of observations of a functional covariate, that can be seen as the reconstruction of longitudinal data in terms of a basis as in the past section. Let $y_1, y_2, \ldots, y_n$ be a set of observations of a binary response variable associated with the sample paths. Then the functional logistic regression model is defined as $y_i = \pi_i + \epsilon_i, \ i = 1, \ldots, n$ where $\pi_i = \exp(l_i)/[1 + \exp(l_i)]$, with l_i being the logit transformation given by

$$l_i = \ln\left[\frac{\pi_i}{1 - \pi_i}\right] = \alpha + \int_T x_i(t)\,\beta(t)\,dt, \ i = 1, \ldots, n, \tag{1}$$

α a real parameter, $\beta(t)$ a parameter function and ϵ_i zero mean independent random errors.

We cannot estimate its parameter function by the maximum likelihood method as it is shown for example in Ramsay and Silverman [12]. The most proposed solution for the estimation of this model consists on considering that the parameter function can be expressed in terms of the same basis than the predictors. Then $\beta(t) = \sum_{k=1}^{p} \beta_k \phi_k(t)$ so that the functional model in terms of the logit transformation (1) turns to a multiple one $L = \mathbf{1}\alpha + A\Psi\beta$ with $L = (l_1, \ldots, l_n)'$, A the matrix that has as rows the coordinates of the sample paths, Ψ the one that has as entries the inner products between the basic functions, β the vector that has the coordinates of the parameter function and $\mathbf{1} = (1, \ldots, 1)'$. Then, by estimating this multiple model by maximum likelihood we can obtain an estimation of the parameter function.

In spite of this estimation of the model gives us good predictions of the response and goodness of fit measures, the parameters of this multiple model have not an accurated estimation because of the great multicollinearity that there is between the columns of the design matrix [1]. This bad estimation has an adverse effect on the interpretation of the parameter function in terms of odds ratios. One solution pointed by Escabias et al. [6] consists on using as covariates of the multiple logistic model a reduced number of principal components (p.c.'s) of the columns of its design matrix and then to reconstruct the original parameters in terms of the ones given by the model in terms of the p.c.'s.

Then, if we denote by $\Gamma = (\xi_{ij})_{n\times p}$ the matrix of p.c.'s of $A\Psi$, and V the one that has as columns the corresponding eigenvectors, the multiple model in terms of all p.c.'s is $L = \mathbf{1}\alpha + \Gamma\gamma$ with $\gamma = (\gamma_1, \ldots, \gamma_p)'$ the vector of its parameters. The reconstructed original parameters (parameter function basis coefficients) are then obtained from $V'\beta = \gamma$. It can be proved that multiple principal component analysis (PCA) of the $A\Psi$ matrix is equivalent to a functional PCA of the sample paths with a new inner product different to the usual one in $L^2(T)$, with eigenfunctions $f_r(t) = \sum_{j=1}^{p} f_{rj}\phi_j(t)$, $r = 1, \ldots, p$ where its matrix of basis coefficients is obtained as $F = (f_{rj})_{p\times p} = \Psi^{-1}V$. A different functional PCA for functional linear regression model purpose can be seen in Cardot et al. [5].

If we consider all the p.c.'s in the model the estimation of the β coefficients is the same as if we do not use them, but if we use a reduced set of p.c.'s the reconstructed β parameters are more accurated. Let $\Gamma_{(s)}$ the matrix with the first s p.c.'s of $A\Psi$ and $V_{(s)}$ its corresponding eigenvector one, then the functional principal component logistic regression model is defined as $l_{i(s)} = \ln\left[\pi_{i(s)}/\left(1-\pi_{i(s)}\right)\right] = \alpha_{(s)} + \sum_{j=1}^{s} \xi_{ij}\gamma_{j(s)}$ with $\alpha_{(s)}$ being a real parameter and $\gamma_{(s)} = \left(\gamma_{1(s)}, \ldots, \gamma_{s(s)}\right)'$ a vector of parameters. As the p.c.'s are uncorrelated, we can estimate this model accurately. Then we can obtain the following estimation of the parameter function $\widehat{\beta}_{(s)}(t) = \sum_{k=1}^{p} \widehat{\beta}_{k(s)}\phi_k(t)$ from its estimated coordinates $\widehat{\beta}_{(s)} = \left(\widehat{\beta}_{1(s)}, \ldots, \widehat{\beta}_{p(s)}\right)' = V_{(s)}\widehat{\gamma}_{(s)}$. The estimated variance of this $\widehat{\beta}_{(s)}$ vector is given by expression $V_{(s)}\left(\Gamma'_{(s)}\widehat{W}_{(s)}\Gamma_{(s)}\right)^{-1}V'_{(s)}$ with $\widehat{W}_{(s)}$ the diagonal matrix of $\widehat{\pi}_{i(s)}\left(1-\widehat{\pi}_{i(s)}\right)$.

In matrices $\Gamma_{(s)}$ and $V_{(s)}$ we have considered the first p.c.'s. There is a lot of examples in literature in which the last p.c.'s may be more predictive for the response than the first ones. In this sense Escabias et al. [6] have proposed an alternative method of including p.c.'s in the model different to the natural order given by explained variance. This method consists on including them accordingly with their statistical significance given by stepwise method based on conditional likelihood ratio tests. They have been shown that it provides a bigger dimension reduction and similar accuracy in the estimations.

4 Modelling lupus sprout from stress level

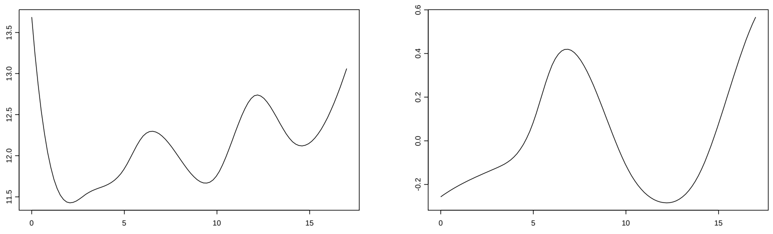
As the Lupus Foundation of America defines on its web, lupus is a chronic inflammatory disease that can affect various parts of the body and may cause serious and even life-threatening problems ('http://www.lupus.org/'). A lupus sprout is a state of the illness that can be cause serious consequences on the patient, including sometimes the death, and that is detected by a blood test. We try to put in relation the probability of a lupus patient to suffer a sprout with the daily stress level suffered by this patient during some time.

The data consist on 44 patients of lupus who where required for measuring their stress level 18 days before the blood test was done. Not all the people have the same number of observations because some of them did not answered the stress test some days. This data set has been provided by the Autoimmune Diseases unit of the Internal Medicine Service of the Ruiz de Alda hospital of Granada (Spain).

The occurrence of a sprout has been considered as response that takes value 1 if the corresponding patient has had it and 0 if he (or she) has not had a sprout. On the other hand we have considered the stress level as the functional covariate. In order to approximate the discrete observations into a basis, we have considered the cubic B-spline definition nodes $t_1 = 0$, $t_4 = 3$, $t_6 = 5$, $t_7 = 6$, $t_{11} = 10$, $t_{13} = 12$, $t_{15} = 14$, $t_{18} = 17$. Finally we have obtained independently each basis coefficients by least squares from its observations between t_1 and t_{18}. From this approximation the mean level of stress $\overline{x}(t) = n^{-1}\left(\sum_{i=1}^{n} x_i(t)\right)$ can be seen in Figure 1.a

After approximating the sample paths, we have fitted the functional logistic regression model from the multiple one that we obtain after considering the sample paths and the parameter function in the same B-spline basis. The estimated parameter function obtained has been drawn Figure 2.a

This parameter function has not an easy interpretation and make us to suspect that it has little accuracy because of the multicollinearity that there is in the multiple model. In order to solve this problem we have obtained



(a) (b)

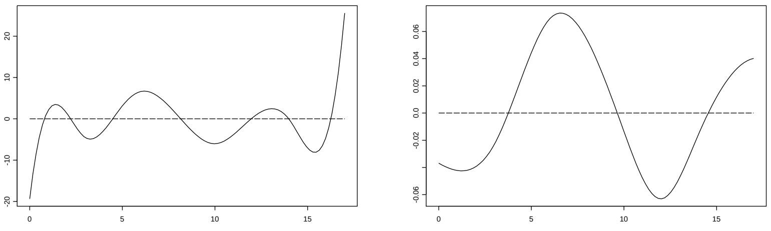
Figure 1: Mean function (a) and third eigenfunction (b)

the p.c.'s of $A\Psi$ matrix. The percents of variance explained by the first three p.c.'s have been 94.76, 98.06 and 99.12 percent.

Finally we have fitted the principal component logistic regression models by introducing p.c.'s in the order given by stepwise method based on conditional likelihood ratio test, and reconstructing the parameter function basis coefficients in each fit. In order to choose the best estimation of the parameter function it is usual to consider the model with a number of p.c.'s previous to a significant increase in the estimated variance of the estimated parameters [6]. In our case the best estimation has been the one with only the third p.c. in the model that is the only significative p.c. for the response. The eigenfunction associated to this p.c. in the corresponding functional PCA can be seen in Figure 1.b. The parameter function reconstructed by this method has been drawn in Figure 2.b.

The CCR have been 97.73% with 0.5 as cutpoint, and 84.09% with the ratio of ones in the sample as cutpoint. Both classifications are high. The deviance statistic has been 10.39 that provides a p-value nearly 1. Finally the variance of the estimated parameter $\widehat{\beta_{(s)}}$ has been 1.08 far from 2.23e+12 of the model with all p.c.'s. All these goodness of fit measures show that the model fits well.

In multiple logistic regression the accurated estimation of the parameters is important because the exponential of each one can be interpreted as the change in the odds of occurrence of success produced by a one unit change in the covariate. In the functional case high absolute values of the parameter function indicate times with a large influence on a lupus sprout whereas small values represent times with little influence. This means (see Figure 2.b) that people with high stress level around the fifth and eighteenth days would have higher probability of suffering a lupus sprout whereas absolute high level around the second and twelfth days would reduce this probability. As a result, we could interpret that consequences of high stress level on a lupus sprout have approximately a five days lag.



(a) (b)

Figure 2: Estimated parameter function of the functional logistic regression model without using p.c.'s (a) and the model with only the third p.c. (b)

References

[1] Aguilera A.M., Escabias M. (2000). *Principal component logistic regression.* Proceedings in computational statistics 2000, Physica-Verlag, 175 – 180.

[2] Aguilera A.M., Ocaña F.A., Valderrama M.J. (1999). *Forecasting with unequally spaced data by a functional principal component approach.* Test. **8** (1), 233,- 253.

[3] Besse P., Cardot H., Ferraty F. (1997). *Simultaneous nonparametric regressions of unbalanced longitudinal data.* Computational Statistics and Data Analysis **24**, 255,- 270.

[4] Bosq D. (2000). *Linear processes in function spaces.* Lecture notes in Statistics 149, Springer-Verlag.

[5] Cardot H., Ferraty F., Sarda P (2003). *Spline estimators for functional linear model.* Statistica Sinica **13**, 571,- 591.

[6] Escabias M., Aguilera A.M., Valderrama M.J. (2004). *Principal component estimation of functional logistic regression: discussion of two different approaches.* Journal of nonparametric statistics. In press.

[7] Ferraty F., Vieu P. (2002). *The functional nonparametric model and application to espectrometric data.* Computational Statistics **17**, 545,- 564.

[8] Frank I.E., Friedman J.H. (1993). *A statistical view of some chemometric regression tools.* Technometrics **35**, 109,- 148.

[9] James G.M. (2002). *Generalized linear models with functional predictors.* Journal of the Royal Statistical Society, Series B **64** (3), 411,- 432.

[10] Liang K.-Y., Zeger S.L. (1986). *Longitudinal data analysis using generalized linear models.* Biometrika **73** (1), 13,- 22.

[11] Marx B., Eilers P. (1999). *Generaized linear regression on sampled signals and curves: a P-spline approach.* Technometrics **41**, 1,- 13.

[12] Ramsay J.O., Silverman B.W. (1997). *Functional data analysis.* Springer-Verlag.

[13] Ramsay J.O., Silverman B.W. (2002). *Applied functional data analysis.* Springer-Verlag.

[14] Ratcliffe S.J., Leader L.R., Heller G.Z. (2002). *Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression.* Statistics in medicine **21**, 1115,- 1127.

[15] Valderrama M.J., Aguilera A.M., Ocaña F.A. (2000). *Predicción dinámica mediante análisis de datos funcionales.* La Muralla-Hespérides, Madrid.

[16] Zeger S.L., Diggle P.J. (1994). *Semiparametric models for longitudinal data with applications to CD4 cell numbers in HIV seroconverters.* Biometrics **50**, 689,- 699.

Acknowledgement: This research has been funded by BFM2000-1466 project from the Spanish Ministry of Science and Technology. The authors would like to thank professor Miguel Pérez (Department of Personality, Testing and Psychological Treatment of the University of Granada (Spain)) for providing us the data.

Address: M. Escabias, A.M. Aguilera, M.J. Valderrama, University of Granada, Department of Statistics; O.R. Faculty of Farmacy, Campus de Cartuja S/N, 18071 Granada

E-mail: `escabias@ugr.es`

COMPSTAT'2004 Symposium

CORE FUNCTION AND PARAMETRIC INFERENCE

Zdeněk Fabián

Key words: Score function, basic data characteristics.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: At bounded or unbounded intervals of the real line, we introduce classes of regular statistical families, called Johnson families, because they are constructed by making use of the Johnson transformation and show that they contain a lot of important statistical models. The system of Johnson parametric families with unified meaning of parameters is based on the formerly introduced concept of a core function of distribution, a modification of the score function. The Johnson families have a new parameter called Johnson location, which can be treated as a 'centre of the mass' of the distribution. There are many members of Johnson families with non-existing mean or variance. We show that random samples taken from them can be properly characterized by the estimates of the Johnson location together with estimates of the newly proposed value, the core variance.

1 Johnson transformation

For every open interval $\mathcal{X} = (a, b) \subseteq \mathbb{R}$, $\Pi_{\mathcal{X}}$ is a class of Lebesgue dominated probability measures (distributions) supported by $\mathcal{X}$ with well defined derivatives of the respective Lebesgue densities.

An increasing one-to-one mapping $\psi : \mathcal{X} \to \mathbb{R}$ given by formula

$$\psi(x) = \begin{cases} \log(x-a) & \text{if} \quad \mathcal{X} = (a, \infty) \\ \log\dfrac{(x-a)}{(b-x)} & \text{if} \quad \mathcal{X} = (a, b) \\ \log(b-x) & \text{if} \quad \mathcal{X} = (-\infty, b) \end{cases} \tag{1}$$

will be called a *Johnson function.* It is just the reversed Johnson transformation $\psi^{-1} : \mathcal{X} \to \mathbb{R}$ (see [4], [5]), generalized for arbitrary interval support different from the whole real line.

Given $G \in \Pi_R$, distribution $F = G\psi$ will be called the Johnson induced distribution on $\mathcal{X}$. Its density is

$$f(x) = g(\psi(x))\psi'(x), \tag{2}$$

where g is the density of G and $\psi'(x) = d\psi(x)/dx$. Denote by y^* the mode of g (the solution of $g'(y) = 0$ if it is unique). The mode of g we regard as a 'centre of the mass' of G.

Let $F \in \Pi_{\mathcal{X}}$ be a distribution with partial support $\mathcal{X} \neq \mathbb{R}$. Let us construct the *Johnson prototype distribution* $G = F\psi^{-1}$, find its mode y^* and put

$$x^* = \psi^{-1}(y^*).$$

The image of the centre of the mass of the prototype distribution will be considered as a centre of the mass of F. It always exists, in contrast with the mean or with the mode of distributions supported by $\mathcal{X} \neq \mathbb{R}$.

2 Johnson location

Let the parameter space Θ_R of distributions with support $\mathcal{X} = \mathbb{R}$ be in the form $\Theta_R = \mathbb{R} \times (0,\infty) \times \Lambda$ where $\Lambda = (0,\infty) \subset R^p$, so that $\theta_R \in \Theta_R$ is $\theta_R = (\mu, \sigma, \lambda)$ where $\mu = y^* \in \mathbb{R}$ is the location, $\sigma \in (0,\infty)$ the scale and $\lambda \in \Lambda$ a vector of shape parameters. By introducing a *pivotal function*

$$u_R = u_{\mu,\sigma}(y) = \frac{y-\mu}{\sigma},$$

the members of parametric family $\{G_{\theta_R} : \theta_R \in \Theta_R\}$ can be written as $G_{\theta_R}(y) = G_\lambda(u_R)$ and their densities as $g_{\theta_R}(y) = \sigma^{-1} g_\lambda(u_R)$.

Let us construct an image on $\mathcal{X}$ of a given system of parametric families $\mathcal{G}_{\theta_\mathcal{R}} \subset \Pi_R$. Put $x = \psi^{-1}(y)$ and, for any $G_\lambda(u_R) \in \mathcal{G}_{\theta_\mathcal{R}}$, put

$$\tau = \psi^{-1}(\mu). \tag{3}$$

Let us call τ the *Johnson location,* and fix the parameter space Θ of the intended system on $\mathcal{X}$ in the form $\Theta = \mathcal{X} \times (0,\infty) \times \Lambda$, so that $\theta \in \Theta$ can be written as $\theta = (\tau, \sigma, \lambda)$. By defining the *pivotal function on* $\mathcal{X}$,

$$u_{\mathcal{X}} = u_{\psi(\tau),\sigma}(x) = \frac{\psi(x) - \psi(\tau)}{\sigma}, \tag{4}$$

we obtain a system of distribution families

$$(\mathcal{F}_{\mathcal{X}})_\theta = \{F_\theta : F_\theta(x) = G_\lambda(u_{\mathcal{X}}), \theta \in \Theta\}$$

with densities

$$f_\theta(x) = dG_\lambda(u_{\mathcal{X}})/dx = \sigma^{-1} g_\lambda(u_{\mathcal{X}})\psi'(x). \tag{5}$$

For instance, if $\mathcal{X} = (0,\infty)$ the pivotal function (4) is (since $\psi(x) = \log x$ and $\psi'(x) = 1/x$)

$$u_{(0,\infty)} = \log\left(\frac{x}{\tau}\right)^{1/\sigma}$$

and the density (5) is $f_\theta(x) = \frac{1}{\sigma x} g_\lambda(\log(x/\tau)^{1/\sigma})$. The reciprocal scale $\beta = 1/\sigma$ is usually often taken for a shape parameter. It should be noted that some model distributions with support $\mathcal{X} = (0,\infty)$, discussed in statistic literature (e.g. exponential, Weibull, Wald and extreme value distributions), do have the Johnson location parameter. This parameter is often taken for a scale. Other distributions (e.g. gamma, chi-squared, Maxwell and beta-prime) distributions), which do not posses the Johnson location parameter, can be reparametrized. Instead of describing a general way of reparametrization, let us present a characteristic example.

Example 1. Consider a gamma distribution with $\mathcal{X} = (0, \infty)$ and with density $f_{\alpha,\gamma}(x) = \frac{\gamma^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\gamma x}$. Its Johnson prototype has density $g_{\alpha,\gamma}(y) = \frac{1}{x} \frac{\gamma^\alpha}{\Gamma(\alpha)} e^{\alpha y} e^{-\gamma e^y}$ with mode $y^* = \log \frac{\alpha}{\gamma}$. By setting $\mu = y^*$ and transforming $g_{\alpha,\gamma}(y - \mu)$ into $\mathcal{X}$ we obtain a reparametrized form of the density of the gamma distribution with Johnson location $\tau = \alpha/\gamma$,

$$f_{\tau,\alpha}(x) = \frac{\alpha^\alpha}{x\Gamma(\alpha)} \left(\frac{x}{\tau}\right)^\alpha e^{-\alpha \frac{x}{\tau}}. \tag{6}$$

3 Core function

Let Y be a random variable with distribution $F \in (\mathcal{F}_\mathcal{X})_\theta$ having prototype $G = F\psi^{-1}$ and pivotal variable $u_\mathcal{X}$ given by (4). The *core function* T_F of Y was defined by [2] as a score for the pivotal function on $\mathcal{X}$, i.e.,

$$T_F(x; \theta) = -\frac{g'_\lambda(u_\mathcal{X})}{g_\lambda(u_\mathcal{X})}. \tag{7}$$

Let us consider special cases:

i/ $\mathcal{X} \neq \mathbb{R}, F = G\psi$. Let us put $y = \psi(x)$. By (2) and the chain rule for differentiation,

$$T_F(x) = -\frac{1}{g(y)} \frac{d}{dy} g(y) = \frac{1}{f(x)} \frac{d}{dx} \left(\frac{1}{\psi'(x)} f(x)\right).$$

ii/ $\mathcal{X} \neq \mathbb{R}, F_\theta(x) = G_\lambda(u_\mathcal{X})\psi$. Making use of (5), in the general parametric case it holds that

$$\frac{\partial}{\partial \tau} \log f_\theta(x) = \frac{g'_\lambda(u_\mathcal{X})}{g_\lambda(u_\mathcal{X})} \frac{\partial u_\mathcal{X}}{\partial \tau} = \frac{\psi'(\tau)}{\sigma} T_F(x; \theta). \tag{8}$$

Core functions of parametric distributions are proportional to the scores for Johnson location.

iii/ $\mathcal{X} = \mathbb{R}$. Distribution $G \in \Pi_R$ can be considered as transformed by the identical mapping. Using (8), $\frac{\partial}{\partial \mu} \log g_\lambda(u_R) = \frac{1}{\sigma} T_G(y; \theta_R)$. The core function of normal distribution is thus $T_G(u_R) = u_R$.

It follows from (8) that the core function is proportional to the influence function of the maximum likelihood estimator of parameter τ. From the point of view of the estimation of the centre of mass of the distribution, a crucial property of core functions is their boundedness or unboundedness when x approaches to the lower and upper end of the support interval.

4 Model distributions

Based on the above considerations, distributions can be classified into three classes:

Class I: distributions with unbounded core functions.

Class II: distributions with bounded core functions.

Class III: distributions with 'semibounded' core functions.

Core functions and densities of parametric families, which we call a 'Johnson system of distributions', are given in Table 1.

Class	$T_F(z)$	$f(z)$	Name
Ia	$\frac{\alpha}{2}[z - \frac{\nu}{z} + \nu - 1]$	$\frac{\kappa\nu^{-\rho/2}z^\rho}{2K_\rho(\alpha\sqrt{\nu})}e^{-\frac{\alpha}{2}(z+\nu/z)}$	GIG
Ib	$\log z$	$\frac{\kappa}{\sqrt{2\pi}}e^{-\frac{1}{2}(\log z)^2}$	Normal
IIa	$\alpha\frac{z-1}{z+1/\nu}$	$\frac{\kappa}{\nu^\alpha B(\nu\alpha,\alpha)}\frac{z^{\nu\alpha}}{(z+1/\nu)^{[1+\nu]\alpha}}$	Transf. beta
IIb	$\frac{2\alpha\log z}{1+(\log z+\nu)^2}$	$\frac{\kappa C}{(1+(\log z+\nu)^2)^\alpha}e^{2\nu\alpha\tan^{-1}(\log z+\nu)}$	Pearson IV
IIIa	$\alpha(z-1)$	$\frac{\kappa\alpha^\alpha}{\Gamma(\alpha)}z^\alpha e^{-\alpha z}$	Gen. gamma
IIIb	$\alpha(1-1/z)$	$\frac{\kappa\alpha^\alpha}{\Gamma(\alpha)}z^{-\alpha}e^{-\alpha/z}$	Extr. val.

Table 1: Core functions and densities of Johnson system of distributions.

Shape parameters of the families in Table 1 have a unique meaning: α characterizes the excess and ν the non-symmetry of both core functions and densities. The Johnson location and scale parameters are introduced by means of pivotal variable $u_\mathcal{X}$ for the given support. We denoted

$$z = e^{u_\mathcal{X}}$$

in order that the densities in Table 1 optically resemble the densities of distributions with support $\mathcal{X} = (0,\infty)$. Further, we denoted $\kappa = \frac{\psi'(x)}{\sigma}$. The explicit values of z and κ for different supports are given in Table 2.

$\mathcal{X}$	z	κ
$\mathbb{R}$	$e^{\frac{y-\mu}{\sigma}}$	σ^{-1}
$(0,\infty)$	$(\frac{x}{\tau})^{1/\sigma}$	$(\sigma x)^{-1}$
(a,b)	$\left(\frac{(x-a)(b-\tau)}{(\tau-a)(b-x)}\right)^{1/\sigma}$	$\frac{(b-a)}{\sigma(x-a)(b-x)}$

Table 2: Values of z and κ for distributions supported by $\mathcal{X}$.

From (7) and (4) it follows that $T_F(\tau;\theta) = 0$ and hence $T_F(z) = 0$ for $z = 1$. It is apparent from Table 1 that the core functions are often simpler and more understandable than the densities.

Family Ia with unbounded core functions (where $K_\rho(\cdot)$ is a modified Bessel function of the third kind, $\rho = \frac{\alpha}{2}(1-\nu)$ and $\nu > 0$) is a reparametrized Generalized inverse Gaussian (GIG) family [5], which turns, if $\nu = 1$, into Wald distribution. The members of family Ib are normal, lognormal and Johnson U_B distributions. Family IIa with bounded increasing core functions (B denotes the beta function, $\nu > 0$) is the reparametrized Transformed Beta family [6] with many members (logistic, log-logistic, Fisher-Snedecor, beta, Burr III, Burr XII etc.). Class IIb is the Pearson IV family (with not specified constant C and $\nu \in R$), it has asymmetric redescending core functions and its members if $\nu = 0$ are Cauchy, log-Cauchy and Student distributions (with $C = 1/B(\frac{1}{2}\alpha - \frac{1}{2})$). Finally, non-symmetric distributions of Class III (where Γ is a gamma function) have prototypes with densities mutually symmetric according to $y-$axis (Gumbel and extreme value I). Distributions of Class IIIa with support $\mathcal{X} = (0, \infty)$ are frequently used (gamma, Maxwell, chi-squared, Weibull), in contrast with their 'counterparts' from Class IIIb with often non-existing usual moments (examples: extreme value II and Gompertz).

Actually, there are only few standard model distributions not contained in Table 1 (we mention the Laplace and Pareto distributions with discontinuous core functions). Any other system of distributions, built perhaps by means of a mapping ψ different from (1), should contain distributions with all the types of core functions discussed.

5 Parametric estimation

Core moments of random variable X with distribution F are the moments of core function T_F, for $k \in \mathcal{N}$ given by $\mathcal{M}_k(F) = E_f T_F^k$. Making use of (7) and (5), we obtain a relation

$$\mathcal{M}_1(F_\theta) = \int_a^b T_F(x;\theta) f_\theta(x)\, dx = \int_{-\infty}^{\infty} -g'_\lambda(u_\mathcal{X})\, du_\mathcal{X} = 0, \tag{9}$$

saying that the core moments are the central moments around τ.

Given a random sample $\mathbb{X}_n = (X_1, ..., X_n)$ from F_θ, the natural estimators of core moments are core moments of the sample distribution function. The core moment estimate $\hat{\theta}_{CM}$ of the true value of parameter θ is thus the solution of the system

$$\frac{1}{n}\sum_{i=1}^{n} T_F^k(x_i;\theta) = \mathcal{M}_k(F_\theta), \qquad k = 1, \ldots, m. \tag{10}$$

The core moment (CM) estimates were shown by [2] to be consistent and asymptotically normal. Moreover, in cases of bounded core functions (this can be done artificially by procedures used in robust statistics) the CM estimates are insensitive to outliers and their asymptotic relative efficiencies are near to one.

Due to (8), the CM and maximum likelihood (ML) estimates are identical if the estimated parameter is the Johnson location.

Example 2. Consider Lomax distribution L_α with density $f_\alpha(x) = \frac{\alpha}{(x+1)^{\alpha+1}}$. Its parameter α is not the Johnson location. The Johnson prototype of the Lomax distribution has density $g_\alpha(y) = \frac{\alpha e^y}{(e^y+1)^{\alpha+1}}$ with mode $y^* = \log\frac{1}{\alpha}$. The reparametrized Lomax density

$$f_{\tau,\alpha}(x) = \frac{\alpha^{\alpha+1}}{x}\frac{x/\tau}{(x/\tau+\alpha)^{\alpha+1}} \tag{11}$$

belongs to class IIa and can be expressed as $f_{1/\alpha,1,\alpha,1/\alpha}(x)$, where $f_{\tau,\sigma,\alpha,\nu}(x) = f_{\alpha,\nu}(z)$ is density IIa in Table 1. The score for α is $s_\alpha(x) = \frac{1}{\alpha} - \log(x+1)$ and $\hat{\alpha}_{ML} = n/\sum_{i=1}^n \log(1+x_i)$. The core function is $T_{F_\alpha}(x) = (\alpha x - 1)/(x+1)$ and the core moment estimate is also given explicitly by

$$\hat{\alpha}_{CM} = \frac{\sum_{i=1}^n 1/(x_i+1)}{\sum_{i=1}^n x_i/(x_i+1)}.$$

Both estimates were compared in the following simulation experiment: 1000 samples of length n from L_α were generated and average values of both $\hat{\alpha}_{ML}$ and $\hat{\alpha}_{CM}$ were calculated. The results are given in Table 3. Although the average values of the standard deviations of the CM estimates (not printed) are a little higher than those of the ML estimates (the asymptotic relative efficiency of α_{CM} is $\alpha(\alpha+2)/(\alpha+1)^2)$), their values are systematically nearer to the true value $\alpha = 2$. The CM estimates are better than the ML ones even when the data are not contaminated! The reason is the bounded core function of the Lomax distribution, suppressing the 'outliers' produced by the data generating process.

n	15	30	60	100
$\bar{\alpha}_{ML}$	2.133	2.082	2.035	2.018
$\bar{\alpha}_{CM}$	2.100	2.060	2.025	2.016

Table 3: Average values of estimates of α.

6 Characteristics of the sample

Taking the expectation of both sides of equation (8) squared, we obtain a relation between the Fisher information for τ and the second core moment in the form

$$I_\tau(\theta) = \left(\frac{\psi'(\tau)}{\sigma}\right)^2 \mathcal{M}_2(F_\theta). \tag{12}$$

The existence of $I_\tau(\theta)$ (and, therefore, of $\mathcal{M}_2(F_\theta)$) is secured by the usual type of regularity conditions imposed on distributions.

In [1], reasons are given for taking the function $i_F(x) = T_F^2(x)$ as the information function of the distribution, expressing the density of information about Johnson location contained in x. In parametric cases $i_F(x) = (\psi'(\tau)/\sigma)^2 T_F^2(x|\theta)$; its mean value is $I_\tau(\theta)$. (12) thus can be viewed as the mean information of distribution F and its reciprocal value

$$V = I_\tau(\theta)^{-1} \tag{13}$$

as a '*core variance*' characterizing the dispersion of the distribution. We suggest taking the sample core variance $\hat{V} = \hat{\tau}^2\hat{\sigma}^2/M_2(F_{\hat{\theta}})$ as a value characterizing the dispersion of data $\mathbb{X}_n$ around its centre of the mass $\hat{\tau}$.

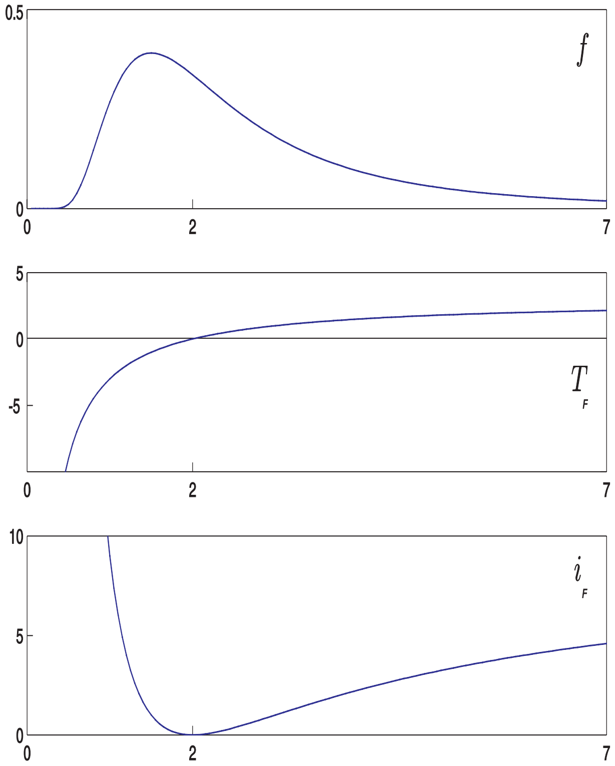


Figure 1: Density, core function and information function of reciprocal gamma distribution.

Example 3. Distribution F from Class IIIb with density

$$f_{\tau,1,\alpha}(x) = \frac{\alpha^\alpha}{x\Gamma(\alpha)}\left(\frac{\tau}{x}\right)^\alpha e^{-\alpha\frac{\tau}{x}} \tag{14}$$

can be taken as 'reciprocal gamma', a counterpart of the gamma distribution. It has Johnson location $\tau = \gamma/\alpha$ and core function $T_F(x) = \alpha(1-\tau/x)$. Density f, core function T_F and information function $i_F = T_F^2$ of the reciprocal gamma distribution for $\tau = 2$ and $\alpha = 3$ are plotted on Fig. 1.

The usual k-th moments of the distribution exist only if $k < \alpha$. For instance, if $\alpha \leq 1$, the distribution has neither a mean nor a variance. The sample mean and the sample variance of $\mathbb{X}_n$ from F do not characterize the data in any way, as well as the ML estimates of parameters α and γ. On the other hand, $\mathcal{M}_2(F) = \alpha$ and the core moment equations (10) are

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \alpha(1-\frac{\tau}{x_i}) &= 0 \\ \frac{1}{n}\sum_{i=1}^n \alpha^2(1-\frac{\tau}{x_i})^2 &= \alpha. \end{aligned}$$

The estimate $\hat{\tau}_{CM}$ from the first equation is the harmonic mean $\hat{\tau}_{CM} = \hat{\tau}_n = n/\sum_{i=1}^n 1/x_i$, and $\hat{\alpha}_{CM} = n/\sum(1-\hat{\tau}_{CM}/x_i)^2$. By (12) and (13) , $V = \tau^2/\alpha$ and the couple characterizing the sample $\mathbb{X}_n$ is

$$(\hat{\tau}, \hat{V}) = \left(\hat{\tau}_n, \hat{\tau}_n^2 \frac{1}{n}\sum_{i=1}^n (1-\frac{\hat{\tau}_n}{x_i})^2\right).$$

References

[1] Fabián Z. (2001). *Information and entropy of continuous random variables.* IEEE Trans.on Information Theory **43**, 1080 – 1083.

[2] Fabián Z. (2001). *Induced cores and their use in robust parametric estimation.* Commun. Statist.-Theory Meth. **30**, 537 – 556.

[3] Fabián Z., Vajda I. (2003). *Core functions and core divergences of regular distributions.* Kybernetika **39**, 29 – 42.

[4] Johnson N.L. (1949). *Systems of frequency curves generated by methods of translations.* Biometrika **36**, 149 – 176.

[5] Johnson N.L., Kotz S., Balakrishnan N. (1994,1995). *Continuous univariate distributions 1, 2.* Wiley, New York.

[6] Klugmann S.A., Panjer H.H., Willmott G.E. (1998). *Loss models. From data to decisions.* Wiley, New York.

Acknowledgement: The work was supported by the Grant Agency of ASCR under Grant Number IAA1075403.

Address: Z. Fabián, Institute of Computer Science CAS, Pod Vodárenskou věží 2, Prague 8, Czech Republic

E-mail: `zdenek@cs.cas.cz`

ANALYSIS OF THE ORGANIZATIONAL CULTURE AT A PUBLIC UNIVERSITY BY APPLICATION OF MULTIVARIATE TECHNIQUES

Karmele Fernández-Aguirre, P. Mariel and A. Martín-Arroyuelos

Key words: Multiple correspondence analysis, classification, probit regression, discriminant analysis.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: The objective of the paper is to analyze organizational aspects at The University of the Basque Country paying special attention to the organizational culture. We apply several techniques to analyze the type of culture: from the most classical obtaining summates scales to the multiple exploratory analysis (Multiple Correspondence Analysis (MCA) and Classification (CA)) in a complementary way ([4]) and inferential techniques (Probit Regression and Fisher's Discriminant analysis).

1 Introduction

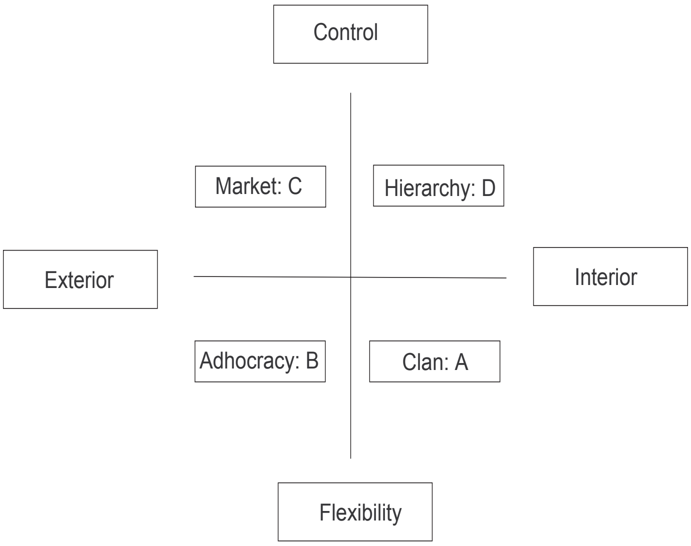
We use a data set obtained from a survey which collects answers of 600 lecturers from a total of 2900 which work at The University of the Basque Country. This survey was designed for the specific purpose to analyze the organizational aspects and its questions were posed using a preliminary half hour long interview about the organizational structure of the university with 20 individuals with high ranking posts. The model we adopt in our analysis is the *Model of Values in Competition* ([3]) which is based on two bipolar dimensions. The first one opposes the organizational position towards *interior* against *exterior* and the second one opposes *flexibility* against *control.* These two axes form four quadrants in which the following organizational position (type of culture) can be placed: *clan, hierarchy, market* and *adhocracy or innovation.* An institution can stand out in various or no position.

The conclusions we obtain indicate that the *flexible* culture prevails at The University of the Basque Country the and that the *clan*-type of culture with a high percentage of *innovation* is the most perceived one. These conclusions support the hypothesis about the opening of the rigid university structure through high level quality research groups.

2 The model

The model we adopt in our analysis is the *Competing Values Framework* from Cameron and Quinn ([3]) which is based on two bipolar dimensions.

The first one opposes the organizational position towards *interior* against *exterior* and the second one opposes *flexibility* against *control.* As we can see in the following diagram, these two axes form four quadrants in which the following types of culture can be placed.



The competing values framework can be used in constructing an organizational culture profile, and the central issue associated with organizational culture is its linkage with organizational performance.

The name *clan*, or A is used for its similarity to the family organization. It is an organization that concentrates on **internal** maintenance with flexibility, concern for people, and sensitivity for customers. This type of organization is characterized by shared values, shared objectives, cohesion, participation and a very strong group feeling of “we”. The clan culture represents a friendly place to work.

Adhocracy, or innovation, or B, is an organization that concentrates on **external** positioning with a high degree of flexibility and individuality. It means that the most important task of management is to stimulate knowledge, risk and creativity. This type of culture is based on improvement groups which better employ the basic procedures of an organization to achieve adaptability, flexibility and creativity.

The organizational form called *market*, or C, is based on “management by objectives” and “cost transaction”. It is an organization that concentrates on **external** maintenance with a need for **stability and control**.

Finally, the type of culture D, compatible with *hierarchy*, is defined as an organization that focuses on **internal** maintenance with a need for **stability and control**. It is also a space of formalized and structured work, where the formal rules and policies are pillars of the organization.

3 Data and results

We use a data set obtained from a survey which collects the answers of 600 lecturers from a total of 2900 which work at The University of the Basque Country. This survey was designed for the specific purpose of analyzing the organizational aspects and its questions were posed using a preliminary half-hour long interview about the organizational structure of the university with 20 individuals from high ranking posts.

In the 2001-2002 academic year we applied random stratified sampling according to 5 demographic variables: sex, age, standing, if they take part in a research group or not and their area of knowledge. The survey is part of a large project concerning the process of scientific knowledge research, development and transfer.

We apply classification over the first factors obtained, from a Multiple Correspondence Analysis ([4]) centered on the 18 questions about characteristics of the formation and dissolution of research groups. The answer to all of them will be from strongly disagree to strongly agree in a five points scale.

For instance: "A research group arises when there is an interest to research and to increase new knowledge" If the answer is 4 or 5, that is, agree or strongly agree, the culture is perceived as B. Remember that B was adhocracy or innovation type of culture.

In the second part of the work we have applied Probit Regression and Fisher's Discriminant Analysis using these 18 questions about the formation and dissolution of research groups as explanatory variables, and taking 4 dependent variables one for each type of culture A, B, C, D. These 4 variables have been previously formed as summated scales with internal consistency calculated using Cronbach's alpha coefficient. When we have a variable generated from a set of questions that returned a stable response, then our variable is said to be reliable. Cronbach's alpha is an index of reliability associated with the variation accounted by the true score of the hypothetical variable that is being measured ([2]).

We apply hierarchical classification, using Ward's criterion, over the first 7 factors of a MCA, which accounts for nearly 30 % of the total inertia. We obtain 9 classes of individuals, characterized by the categories of the proposed questions, according to a test-value from [5]. This test-value is associated with the deviation of the rate of individuals choosing a certain category within each class, from the same rate in the whole sample.

The classes we have obtained are:

- *Class 1*: (14.26%) Include individuals not very well defined. The most characteristic answers are neither agree or disagree with respect to A and C, also their perception is against D.
- *Class 2*: (25.96%) Agree with C and A type of culture.
- *Class 3*: (18.28%) Agree with B and against D culture.
- *Class 4*: (17%) Agree with A and against B, C or D culture.

- *Class 5*: (7.5%) Strongly agree with D and C and against B
- *Class 6*: (7.31%) Strongly agree with A and against B and C.
- *Class 7*: (2.93%) Strongly agree with B and A: flexibility.
- *Class 8*: (1.28%) Strongly disagree with B and C: interior.
- *Class 9*: (5.48%) Strongly agree with B.

As we can see, *Clan* and *Market* type of culture are mixed in a class that accounts for 25.96% of the individuals and *Market* and *Hierarchy* are mixed in a class that accounts for 7.5%, but the findings of this study are in agreement with the fact that approximately 40% of them perceive the *Clan* culture type, 33% perceive the *Adhocracy*, 24% *Market* and approximately 3% is perceive as *Hierarchy*.

The conclusions we have obtained indicate that the *flexible* culture prevails over the *rigid* one and that the *clan*-type of culture with a high percentage of *innovation* is the most perceived. These conclusions support the hypothesis about the possibility of opening of the rigid university structure through high level quality research groups.

The following step was to treat the current situation of the institution with the objective to improve the organizational culture. We have taken four dependent variables formed as summated scales, each one about a type of culture, A, B, C, D and the same 18 questions about FORmation and DISolution of research groups as explanatory variables. Over these 18 variables we have applied Factor Analysis obtaining 7 variables called facFOR and facDIS, respectively. Applying Probit regression we have obtained about B type of culture or Adhocracy the following result:

B typ of culture: Estimated Probit Model (Maximum Likelihood)

Parameter	Estimate	Standard Error
CONST	0.6010930	0.0982470
facfor1	0.4696450*	0.0589993
facfor2	-0.0989266	0.0743018
facfor3	-0.0766089	0.0964508
facdis1	0.0578748	0.0690571
facdis2	0.0729662	0.0804176
facdis3	0.0431982	0.0720618
facdis4	-0.0475905	0.0973738

* significant at 5%
Estrella=0,452

We can see only one explanatory variable significant at 5% and the Estrella statistic ([1]) whose value is reasonable high, supports the importance of this variable. We have also applied linear discriminant analysis using Fisher discriminant function and have obtained the same results which confirm us our previous Probit analysis.

B typ of culture: Fisher's Discriminant Analysis

Standarized Discr. Factor		Correlation between Var. and St. D.F	
facFor1	0.951	facFor1	0.970
facFor2	-0.121	facFor2	-0.225
facFor3	-0.105	facFor3	0.070
facDIS1	0.102	facDIS1	0.214
facDIS2	0.126	facDIS2	0.146
facDIS3	0.098	facDIS3	0.128
facDIS4	-0.018	facDIS4	-0.264

Squared canonical correlation = 0.4021

4 Conclusions

The applied methodology (complementary use of MCA and CA) is relatively unknown to social psychologists. Our aproach provided us with a better description of the culture profile than the usual methodology does. We refer to the use of the scores of summated scales from the items with internal consistency calculated using Cronbach's alpha coefficient. We can conclude that multivariate exploratory methods as complementary use of MCA and CA are suitable to analyze the culture profile of an organization.

Concerning the second part of the work, we can conclude that the only significant variable is that which represents the fact that research groups are formed because of the need for research. That indicates that groups are formed because of the existing conciousness that research must take place. University lecturers should be encouraged in becoming conscious of the need for high quality research.

References

[1] Estrella A. (1998). *A new measure of fit for equations with dichotomous dependent variables.* Journal of Business & Economic Statistics **19**, 198 – 205.

[2] Hatcher L. (1994). *A step-by-step approach to using the SAS(R) system for factor analysis and structural equation modeling.* Cary, NC:SAS.

[3] Cameron K.S., Quinn R.E (1999). *Diagnosing and changing organizational culture.* Reading, MA: Addison Wesley Longman.

[4] Lebart L. (1994). *Complementary use of correspondence analysis and cluster analysis.* Correspondence Analysis in the Social Sciences, M. Greenacre and J. Blasius (eds), Academic Press, London, 162 – 178.

[5] Morineau A. (1984). *Note sur la caractérisation statistique d'une classe et les valeurs-tests.* Bull. Tech. du CESIA **1**, 9 – 12.

Address: K. Fernández-Aguirre, P. Mariel, A. Martín-Arroyuelos, Departamento de Econometría y Estadística UPV, Facultad de Ciencias Económicas y Empresariales Avda. Lehendakari Aguirre 83 E48015 Bilbao, Spain

E-mail: `etpfeagk@bs.ehu.es`

RIDGE-PARTIAL LEAST SQUARES FOR GENERALIZED LINEAR MODELS WITH BINARY RESPONSE

Gersende Fort and Sophie Lambert-Lacroix

Key words: Partial least squares, ridge-penalized logistic regression, classification of microarrays.

COMPSTAT 2004 section: Partial least squares.

Abstract: An extension of PLS for regression and dimension reduction in logit models is derived, an extension that still works when the number of covariates is far larger than the number of observations. It is applied to classification of Microarray.

1 Introduction

Partial Least Squares (PLS), first introduced in chemometrics [12], [9] is both used as a dimension reduction tool and as a linear regression tool. The goal of the present contribution is to extend its application to regression in univariate Generalized Linear Models (GLM) with binary response, an extension that covers the case where the length p of the covariate vector is larger or equal to the number of observations n.
PLS constructs predictive models by exhibiting latent covariates (or scores) that account for most of the variation in the response. Unlike Principal Component Regression (PCR), the definition of the scores is based both on the covariates and on the response variable $\mathbf{Y}$, and in that sense, PLS looks more appropriate than PCR to overcome the problems involved by the large number of covariates and their high collinearity. Nguyen and Rocke [10] combines PLS and Iteratively Reweighted Least Squares (IRLS, [6]) *i.e.* PLS and a regression analysis based on the Maximum Likelihood (ML) method; they determine the first κ PLS components from $\mathbf{Y}$ and the initial design matrix; then a regression onto these scores is performed in the ML sense. Besides the question on the pertinence of applying the PLS machinery with a categorical response vector, this algorithm has convergence weaknesses since the ML estimate does not necessarily exist. A second try for extension of PLS to GLM can be found in Marx [8] in a two step procedure. The first step is to exhibit κ PLS scores at convergence of a PLS-within-IRLS algorithm; the second one runs a ML regression onto these scores. In many applications, the Marx algorithm is nothing else than the Nguyen and Rocke algorithm and thus inherits its drawbacks, as discussed in [4].

This is the reason why we introduce an extension, called *Ridge-PLS algorithm*, that can be summarized as a weighted PLS algorithm in which the categorical response variable $\mathbf{Y}$ is replaced with a continuous-valued *pseudo-variable* that captures the information contained in $\mathbf{Y}$. Roughly speaking, *Ridge* fights the multicollinearity while *PLS* is the dimension reduction part. In this contribution, the method is derived for logit models. We show how Ridge-PLS can be used for supervised classification of Microarray data, characterized by a number of covariates far larger than the number of observations.

2 Heuristic of the Ridge-PLS algorithm

We postpone the algorithmic description to Section 4, and start with a naive description. Ridge-PLS is based on the following observation : Least Squares inference and ML inference coincide for regression in a normal linear model which is both a linear model and a GLM. For canonical GLM, the ML estimate $\hat{\theta}^{\mathrm{ML}}$ is the weighted least squares estimate when regressing a pseudo-response variable ψ onto the columns of the design matrix; ψ is obtained at convergence of an IRLS procedure, and for normal models, is equal to $\mathbf{Y}$ [6]. As a consequence, our extension of PLS consists in applying PLS by replacing $\mathbf{Y}$ with the pseudo-variable at convergence of IRLS. Nevertheless, this rough idea has to be made robust in order (i) to be valid when $\hat{\theta}^{\mathrm{ML}}$ does not exist and (ii) to take into account the heteroscedasticity of the pseudo-variable. This is done by respectively (i) substituting $\hat{\theta}^{\mathrm{ML}}$ for a penalized ML estimator, namely the Ridge one, and (ii) introducing a Weighted PLS (WPLS) algorithm. Before deriving Ridge-PLS for logit models, we introduce notations and basic algorithmic ingredients.

3 Some basic ingredients

For a column-vector u, $\|u\|$ is the Euclidean norm, $u_{1:p}$ collects the first p components of u. For a matrix A, A' is the transpose, A_{ij} denotes the entry (i,j), and $A_{\cdot,1:r}$ the matrix that contains the first r columns of A. $\mathbb{1}_n$ is the vector $(1,\cdots,1)'$ of length n and $J^{(r)}$ is a diagonal $(r+1)\times(r+1)$-matrix with $J^{(r)}_{11}=0$ and $J^{(r)}_{kk}=1$ otherwise.

Logit model and Logistic discrimination rule The observations consist of n independent $\{0,1\}\times\mathbb{R}^p$-valued pairs $(\mathbf{y}_i, X_{i\cdot})$ where given $X_{i\cdot}$, the conditional mean of $\mathbf{y}_i$ is π_i, which is related to the linear predictor η_i by $\pi_i = (1+\exp(-\eta_i))^{-1}$, or equivalently $\eta_i = \ln(\pi_i/(1-\pi_i))$. η_i depends on the design vector $Z_{i\cdot} := [1\ X'_{i\cdot}]'$ through the relation $\eta_i = Z'_{i\cdot}\theta$, where $\theta\in\mathbb{R}^{p+1}$ is the unknown parameter. The n response variables (resp. conditional means) are collected in the vector $\mathbf{Y}$ (resp. Π). The $n\times(p+1)$ design matrix is denoted by $Z=[\mathbb{1}_n\ X]$.

For a given estimate $\hat{\theta}$, and a new design vector z, the binary variable $\hat{\mathbf{y}}$ is predicted by applying the logistic discrimination rule, *i.e.* $\hat{\mathbf{y}}=1$ if $\hat{\eta}:=z'\hat{\theta}\geq 0$, and $\hat{\mathbf{y}}=0$ otherwise.

The Ridge-ML estimator When $n > \text{rank}(Z)$, $\hat{\theta}^{\text{ML}}$ is unique when it exists. Unfortunately, the likelihood may be maximal on the boundary of $\mathbb{R}^{p+1}$ so that $\|\hat{\theta}^{\text{ML}}\| = +\infty$ [11]. When $n = \text{rank}(Z)$ - which occurs if and only if $n \leq (p+1)$ and Z has full rank - the solution to the normal equation yields $\|\hat{\theta}^{\text{ML}}\| = +\infty$.
Hence, inference of the parameter necessitates the introduction of a regularization method; we opt for a *Ridge*-penalized ML approach, which shrinks the coefficients towards zero (except the intercept one θ_1). The Ridge estimator $\hat{\theta}^{\text{R}}$ is defined as the maximum of the penalized log-likelihood l^*

$$l^*(\theta) = \sum_{k=1}^{n} \{\mathbf{y}_k Z'_{k\cdot}\theta - \ln(1 + \exp(Z'_{k\cdot}\theta))\} - \frac{\lambda}{2}\theta'\Sigma^2\theta, \tag{1}$$

where $\lambda > 0$ is a *shrinkage* parameter, and Σ is a diagonal matrix taking into account the non-standardization of the covariate matrix : $\Sigma^2_{11} = 0$ and $\Sigma^2_{kk} = \sum_{j=1}^{n}(Z_{j,k} - \mathbb{I}'_n Z_{\cdot,k}/n)^2$ for $k \in [2, p+1]$. $\hat{\theta}^{\text{R}}$ exists, is unique and is computed by the (iterative) Newton-Raphson algorithm, each iteration of which is a weighted Ridge-regression of a pseudo-variable onto the columns of Z.

WPLS algorithm For a given $\mathbb{R}^n$-valued observation ψ, a covariate matrix X, and a positive-definite symmetric weight matrix W, the PLS scope is to convey the relation between ψ and X through the definition of κ scores $(t_j)_{1\leq j\leq\kappa}$. These are linear combinations of the columns of the design matrix Z such that for all j, $\mathbb{I}'_n W t_j = 0$ and for all $j \neq k$, $t'_j W t_k = 0$. This yields the decomposition $\psi = q_0 \mathbb{I}_n + q_1 t_1 + \cdots + q_\kappa t_\kappa + f_{\kappa+1}$ where $f_{\kappa+1}$ is W-orthogonal to the vectors $(\mathbb{I}_n, t_1, \cdots, t_\kappa)$. The pairs (q_j, t_j) are recursively computed as follows

1. $t_0 = \mathbb{I}_n$; $E_0 = X$; $f_0 = \psi$.
2. For $j = 0, \cdots, \kappa$, set $q_j = t'_j W f_j/(t'_j W t_j)$, $f_{j+1} = f_j - q_j t_j$, $E_{j+1} = E_j - t_j t'_j W E_j/(t'_j W t_j)$, $t_{j+1} = E_{j+1}E'_{j+1}W f_{j+1}$.

We refer to the literature for an interpretation of the above algorithm and a discussion on the maximal number of W-orthogonal scores $\kappa_{\max}$ [7]. WPLS, read as a regression method, yields a PLS estimate $\hat{\theta}^{\text{PLS},\kappa}$ through the relation $\hat{\psi}_\kappa = \psi - f_{\kappa+1} = Z\hat{\theta}^{\text{PLS},\kappa}$.

4 The Ridge-PLS algorithm, $n \leq p+1$

Given $(\mathbf{Y}, X)$, for the parameters (λ, κ),

A. Determine ψ : compute $\hat{\theta}^{\text{R}}$, the limiting value of $(\theta^{(t)})_t$ where

$$\theta^{(t+1)} := \left(Z'W^{(t)}Z + \lambda\Sigma^2\right)^{-1} Z'W^{(t)}\psi(\theta^{(t)}), \tag{2}$$

$$\psi(\theta^{(t)}) := Z\theta^{(t)} + \left[W^{(t)}\right]^{-1}\left(\mathbf{Y} - \Pi^{(t)}\right), \tag{3}$$

$Z := [\mathbb{1}_n\ X]$, $\Pi^{(t)}$ is the mean vector Π computed at the current value of the parameter and $W^{(t)}$ is a diagonal matrix with $W^{(t)}_{kk} := \Pi^{(t)}_k(1-\Pi^{(t)}_k)$. Set $\psi := \psi(\hat{\theta}^{\mathrm{R}})$ and $W := W^{(\infty)}$.

B. Run the WPLS with κ components for the variables (ψ, X, W) and compute $\hat{\theta}^{\mathrm{PLS},\kappa}$ as described in Section 3.

Step A builds a continuous response variable ψ whose expected value has linear relationship with the covariates, for the input of PLS; conditionally to $\hat{\theta}^R$, the dispersion matrix of ψ is W^{-1}, which explains the call, in Step B, to a weighted PLS procedure with weight W.

Implementation The procedure, presently derived in $\mathbb{R}^{p+1}$ can be equivalently derived in $\mathbb{R}^{r+1}$ where $r+1 := \mathrm{rank}(Z) \leq n$. To that goal, compute UDV', the singular values decomposition (svd) of $(X - \mathbb{1}_n\mathbb{1}_n'X/n)\Sigma^{-1}$, the standardized covariate matrix, and set $\Xi := (UD)_{\cdot,1:r}$ so that $Z\theta = [\mathbb{1}_n\ \Xi]\gamma$ for some $\gamma \in \mathbb{R}^{r+1}$; it is readily seen that the above procedure, run by replacing (X, Σ^2) by $(\Xi, J^{(r)})$, yields an estimate $\hat{\gamma}^{\mathrm{PLS},\kappa}$ uniquely related to $\hat{\theta}^{\mathrm{PLS},\kappa}$ by the formulas

$$\hat{\theta}_1 = \hat{\gamma}_1 - \mathbb{1}_n'X\hat{\theta}_{2:p+1}/n \qquad \hat{\theta}_{2:p+1} = \left(\Sigma_{2:p+1,2:p+1}\right)^{-1} V_{\cdot,1:r}\hat{\gamma}_{2:r+1}.$$

Hence, up to a single svd, the procedure is independent of p which is of computational importance.

In the application, λ is chosen as the value λ_{opt} in a given range $\mathcal{R}$ minimizing the BIC criterion $-2\hat{l}+\log(n)\mathrm{Dim}$ where $\hat{l}$ is the log-likelihood for the value $\hat{\theta}^R$ of the parameter, and Dim is the trace of $Z\left(Z'WZ + \lambda\Sigma^2\right)^{-1} Z'W$.

5 Application to binary classification

We apply the above procedure to supervised classification of Microarray data; the data set *Leukemia*[1], contains 72 samples divided into 47 cases of acute lymphoblastic leukemia, labeled 0, and 25 cases of acute myeloid leukemia, labeled 1. Each sample consists in a $\{0,1\}$-valued label and 7129 gene expression levels (see Golub *et al.* [5] for a description of the data set). We perform an out of sample (OS) analysis on 100 random partitions of the data set into a learning set and a test set. The learning set contains 27 samples type 0 and 11 samples type 1. We report in Table 1, row "*RPLS* κ" the mean number (and the standard deviation) of misclassified samples in the test set, when the classification rule is determined on the learning set $[\kappa = 1, \cdots, 6]$. Regression is not performed with the 7129 initial covariates; some of them are irrelevant and are deleted following the pre-processing method described in Dudoit *et al.* [2]. We stress that this filtering and the number of remaining genes depend on the learning set. We test the procedures by considering different values of p $(> n = 38)$ and select the p most pertinent covariates as

[1] available at http://www.broad.mit.edu/cgi-bin/cancer/publications

advocated in Dudoit *et al.* [2]. We run the OS analysis for the classification rule induced by the Ridge estimator $\hat{\theta}^{\mathrm{R}}$ (row "*Ridge*", Table 1); the results outline the interest of a dimension reduction step after the regularization one. Eilers *et al.* [3] propose a method quite similar to the Ridge analysis. They compute $\hat{\theta}$ as maximizing the criterion (1) in which Σ is replaced by $J^{(p)}$ (although Z is not standardized); then, their classification rule is based on a Bayes risk : $\hat{\mathbf{y}} = 1$ iff $\hat{\pi}$ is greater than the empirical mean of the observations in the learning set. We run their algorithm and report the results in row "*Eilers*", Table 1. Ridge-PLS yields better results; nevertheless, this assertion has to be nuanced since for less "regular" data sets, Ridge-PLS and the Eilers *et al.* 's method may have an equivalent behavior.

For each partition, λ_{opt} is determined as described above, over 51 $\log_{10}$-linearly spaced points in $\mathcal{R} = [10^{-2}, 10^{3}]$. The mean value of λ_{opt} over the 100 partitions is given in Table 2 for the Eilers *et al.* 's algorithm (λ_{E}) and the Ridge and Ridge-PLS algorithms (λ_{R}). Whatever p, $\lambda_{\mathrm{E}} > \lambda_{\mathrm{R}}$, which is due to the standardization of the design Z.

method	p=50	p=100	p=300	p=500	p=1000
Ridge	1.52 (1.11)	1.35 (1.09)	1.62 (1.05)	1.89 (1.21)	2.83 (1.37)
RPLS 1	1.24 (0.93)	1.18 (0.98)	1.12 (0.86)	1.20 (0.97)	1.45 (1.07)
RPLS 2	1.36 (0.98)	1.24 (0.91)	1.15 (0.93)	1.08 (0.79)	1.27 (0.96)
RPLS 3	1.43 (1.01)	1.32 (0.91)	1.10 (0.77)	1.06 (0.79)	1.14 (0.82)
RPLS 4	1.40 (0.94)	1.34 (0.93)	1.09 (0.79)	1.12 (0.85)	1.39 (0.94)
RPLS 5	1.40 (0.95)	1.33 (0.96)	1.08 (0.80)	1.12 (0.77)	1.21 (0.74)
RPLS 6	1.43 (0.97)	1.27 (0.89)	1.12 (0.79)	1.13 (0.79)	1.25 (0.77)
Eilers	1.44 (1.00)	1.52 (1.00)	1.48 (0.94)	1.42 (0.90)	1.45 (0.95)

Table 1: Mean number of misclassified samples (standard deviation between parentheses).

	p=50	p=100	p=300	p=500	p=1000
λ_{E}	12.16	26.06	78.60	131.20	269.60
λ_{R}	0.38	0.94	3.76	7.30	18.42

Table 2: Mean value of λ_{opt}.

6 Conclusion

We derived an extension of PLS to GLM for logit models. The numerical results show the pertinence of the combination of a regularization step and a dimension reduction step. The technique can be easily adapted to other GLM models such as the multivariate ones, and this will be done in a forthcoming paper. Future research will concern the choice of the regularization method (based for example on the Firth's penalty, as proposed in [1], private communication), and the variable selection and the model selection themes in order to determine optimal values for (λ, κ).

References

[1] Ding B., Gentleman R. (2003). *Classification using generalized partial least squares.* Work in progress.

[2] Dudoit S., Fridlyand J., Speed T. (2002). *Comparison of discrimination methods for the classification of tumors using gene expression data.* J. Amer. Stat. Assoc. **97**, 77 – 87.

[3] Eilers P., Boer J., Van Ommen G., H. Van Houwelingen H. (2001). *Classification of microarray data with penalized logistic regression.* In Proceedings of SPIE. Progress in biomedical optics and images, **4266**, 187 – 198.

[4] Fort G., Lambert-Lacroix S.(2003). *Classification using partial least squares with penalized logistic regression.* Technical report, IAP Network, TR 0331.

[5] Golub T., Slonim D., Tamayo P., Huard C., Gaasenbeek M., Mesirov J., Coller H., Loh M., Downing J., Caligiuri M., Bloomfield C., Lander E. (1999). *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science **286** (5439), 531 – 537.

[6] Green P. (1984). *Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives.* J.R. Statist.Soc. B, **46** (2), 149 – 192.

[7] Helland I. (1990). *Partial least squares regression and statistical models.* Scand. J. Stat. **17** (2), 97 – 114.

[8] Marx B. D. (1996). *Iteratively reweighted partial least squares estimation for generalized linear regression.* Technometrics **38** (4), 374 – 381.

[9] Naes T., Martens H. (1985). *Comparison of prediction methods for multicollinear data.* Commun. Stat., Simulation Comput. **14**, 545 – 576.

[10] Nguyen D., Rocke D. (2002). *Tumor classification by partial least squares using microarray gene expression data.* Bioinformatics **18** (1), 39 – 50.

[11] Santner T., Duffy D. (1986). *A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models.* Biometrika, **73** (3), 755 – 758.

[12] Wold H. (1975). *Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach.* In Perspect. Probab. Stat., Pap. Honour M. S. Bartlett Occas. 65th Birthday, 117 – 142.

Acknowledgement: We are really grateful to A. Antoniadis for constructive and fruitful discussions. This work is supported by the project ASBGEN and the IAP research network P5/24.

Address: G. Fort, S. Lambert-Lacroix, CNRS/LMC, 51, rue des Mathématiques, BP 53, 38041 Grenoble Cedex 9, France

E-mail: `Gersende.Fort,Sophie.Lambert@imag.fr`

COMPSTAT'2004 Symposium

NONPARAMETRIC ESTIMATION OF THE VOLATILITY FUNCTION WITH CORRELATED ERRORS

Mario Francisco-Fernández and J.M. Vilar-Fernández

Key words: Local polynomials, autoregressive process, conditional variance, heteroscedasticity.

COMPSTAT 2004 section: Smoothing.

Abstract: In this paper, we consider a fixed regression model where the errors are a strictly stationary process and in which both functions, the conditional mean and the conditional variance (volatility), are unknown. Two nonparametric estimators of the volatility function based on local polynomial fitting are studied. The asymptotic normality is shown for both estimators. A simulation study and an analysis with real economic data illustrate the behavior of the proposed nonparametric estimators.

1 Introduction

Nonparametric methods are appropriate tools used to perform exploratory analyses, because they do not require selecting a specific parametric shape before fitting the data. In the context of the regression function estimation, a popular nonparametric method is the local polynomial regression (LPR) estimator. We refer to Wand and Jones [5] for an overview of this method.

Sometimes, it cannot be assumed that the observations in the sample data are independent, for example, if they are gathered sequentially in time. In this case, the statistical properties of the LPR estimator obtained under independence can change. Some related works in this setting of dependence are Masry and Fan [4] and Härdle and Tsybakov [3]. In these papers, a regression model considering a random data sample, $\{(X_t, Y_t)\}_{t=1}^{n}$, satisfying some mixing conditions, was used. However, in Francisco-Fernández and Vilar-Fernández [2], a regular fixed design regression model with short-range correlated errors was considered. In this case, while the asymptotic bias of the regression estimator is exactly the same as that obtained under independence, the asymptotic variance of the estimator changes.

In this paper, we consider the same framework as that used in Francisco-Fernández and Vilar-Fernández [2], but now the regression model is heteroscedastic. In this case, the aim is not only to estimate nonparametrically the regression function but also the volatility function. This kind of regression models frequently arise in economic studies, in the analysis of growth curves and usually in the study of time series with deterministic trend and non-constant conditional variance (risk, in financial terminology). Considering this model, we study two estimators of the volatility function previously studied in Härdle and Tsybakov [3] and Fan and Yao [1], respectively, in different contexts. They consider a dynamic regression model of a mixing process and a two-dimensional strictly stationary and absolutely regular process, respectively, in their approaches. As it will be seen, the leading term of the asymptotic variance of these estimators in our model is different from the ones obtained in those papers.

It is assumed that univariate data $Y_{1,n}, \cdots, Y_{n,n}$ are observed, and that

$$Y_{t,n} = m(x_{t,n}) + s(x_{t,n})\, \varepsilon_{t,n}, \quad 1 \leq t \leq n \tag{1}$$

where $m(x)$ and $s(x)$ are "smooth" functions defined on $[0,1]$, with $s(x) > 0$. The errors are a sequence of unobserved random variables with $\mathrm{E}(\varepsilon_{t,n}) = 0$ and $\mathrm{E}\left(\varepsilon^2_{t,n}\right) = 1$, where, for each n, $\{\varepsilon_{i,n}\}_{i=1}^n$ have the same joint distribution as $\{\epsilon_i\}_{i=1}^n$, with $\{\epsilon_t\}_{t \in Z}$ being a strictly stationary stochastic process. Also, it is assumed that the design is a regular fixed design generated by a design density f. We study the problem of estimation of the volatility function $v(x) = s^2(x)$, given a sample $\{(x_t, Y_t)\}_{t=1}^n$.

The organization of the work is as follows: In Section 2, two estimators for the volatility function, $v(x) = s^2(x)$, are introduced. In Section 3 asymptotic properties of both estimators are provided. In Section 4 the estimators studied are compared via a simulation study and, finally, in Section 5 they are used to analyze a real economic data set.

2 The estimators

Due to the simple decomposition $v(x)$ $=\mathrm{E}\left(Y^2 \middle| x\right) - m^2(x)$ and following the idea of Härdle and Tsybakov [3], the first estimator of the volatility function is defined by $\hat{v}_n^S(x) = \hat{g}_n(x) - \{\hat{m}_n(x)\}^2$, where $\hat{g}_n(x)$ is an estimator of $g(x) = E\left(Y^2 \middle| x\right) = m^2(x) + s^2(x)$, and $\hat{m}_n(x)$ is an estimator of $m(x)$. We will use estimators $\hat{m}_n(x)$ and $\hat{g}_n(x)$ based on the LPR estimator. So, assuming that the $(p+1)$th derivatives of $m(x)$ and $g(x)$ exist and are continuous, local polynomial fitting permits estimating the parameter vectors $\vec{\beta}(x) = (\beta_0(x), \ldots, \beta_p(x))^t$, where $\beta_j(x) = m^{(j)}(x)/(j!)$, and $\vec{\gamma}(x) = (\gamma_0(x), \ldots, \gamma_p(x))^t$, where $\gamma_j(x) = g^{(j)}(x)/(j!)$, with $j = 0, 1, \ldots, p$ by using local and weighted least squares methods.

The solution of these minimizations problems are, respectively

$$\hat{\beta}_{(n)}(x) = \left(X^t_{p,(n)} W_{(n)} X_{p,(n)}\right)^{-1} X^t_{p,(n)} W_{(n)} \vec{Y}_{(n)} \tag{2}$$

$$\hat{\gamma}_{(n)}(x) = \left(X^t_{p,(n)} W_{(n)} X_{p,(n)}\right)^{-1} X^t_{p,(n)} W_{(n)} \vec{Y}^2_{(n)}, \tag{3}$$

where $\vec{Y}_{(n)} = (Y_1, \ldots, Y_n)^t$, $\vec{Y}^2_{(n)} = \left(Y_1^2, \ldots, Y_n^2\right)^t$, $X_{p,(n)}$ is a matrix with the ith row equal to $(1, \ldots, (x_i - x)^p)$ and $W_{(n)}$ is the $(n \times n)$ diagonal array whose ith diagonal element is $n^{-1} h_n^{-1} K(h_n^{-1}(x_i - x))$, with K being a kernel function and h_n the bandwidth or smoothing parameter.

The estimator $\hat{v}_n^S(x)$ of $v(x)$ is defined as

$$\hat{v}_n^S(x) = \hat{\gamma}_{(n)}(x)^t e_1 - \left\{ \hat{\beta}_{(n)}(x)^t e_1 \right\}^2, \tag{4}$$

where e_1 is the $(p+1) \times 1$ vector given by $(1, 0, \ldots, 0)$.

On the other hand, Fan and Yao [1] suggest an approach asymptotically fully adaptive to the unknown conditional mean. This consists in, first, obtaining the residuals from a nonparametric fit (using, for instance, the LPR estimator with a kernel L_1, bandwidth h_{1n} and polynomial degree p_1), squaring them, $\hat{r}_t = \{Y_t - \hat{m}_{h_{1n}}(x_t)\}^2$, $\quad t = 1, 2, \ldots, n$, and finally defining the estimator of the volatility function as the LPR estimator of the regression function with kernel L_2, bandwidth h_{2n} and polynomial degree p_2, using $\{\hat{r}_t\}_{t=1}^n$ as the response variables. Then, the second estimator considered in this paper is defined by

$$\hat{v}_n^D(x) = e_1^t \left(X_{p_2,(n)}^t W_{2(n)} X_{p_2,(n)} \right)^{-1} X_{p_2,(n)}^t W_{2(n)} \hat{R}_{(n)}, \tag{5}$$

where $\hat{R}_{(n)} = (\hat{r}_1, \ldots, \hat{r}_n)^t$ and $W_{2(n)}$ is as $W_{(n)}$, but now with L_2 and h_{2n}.

3 Theoretical results

In this section, the asymptotic normality of estimators (4) and (5) are obtained. The following assumptions will be needed in our analysis:

A1 The kernel functions, $K(\cdot)$, $L_1(\cdot)$ and $L_2(\cdot)$ are symmetric, with bounded support, and Lipschitz continuous.

A2 The sequence of bandwidths, $\{h_n^*\}$, satisfies $h_n^* > 0$, $h_n^* \downarrow 0$, $nh_n^* \uparrow \infty$, where the sequence $\{h_n^*\}$ can be $\{h_n\}$, $\{h_{1n}\}$ or $\{h_{2n}\}$. Moreover $h_n^* = O\left(n^{-1/(2p^*+3)}\right)$, where h_n^* and p^* can be h_n and p, or h_{2n} and p_2.

A3 The errors satisfy $\mathrm{E}\left(\varepsilon_i^2\right) = 1$, $\mathrm{E}(\varepsilon_i) = \mathrm{E}\left(\varepsilon_i^3\right) = 0$ and $\mathrm{E}|\varepsilon_t|^{2(2+\delta)} < \infty$ for some $\delta > 0$. Denoting $c(k) = \mathrm{Cov}(\varepsilon_i, \varepsilon_{i+k})$, $k = 0, \pm 1, ...$, then $\sum_{k=1}^{\infty} k\,|c(k)| < \infty$, and $d(\varepsilon) = \sum_{k=-\infty}^{\infty} \mathrm{Cov}\left(\varepsilon_i^2, \varepsilon_{i+k}^2\right) < \infty$. We also assume that $\{\varepsilon_t\}$ is $\alpha-$mixing with mixing coefficients such that $\sum\limits_{t=1}^{\infty} \alpha(t)^{\delta/(2+\delta)} < \infty$ and a sequence of positive integers $\{s_n\}$, $s_n \to \infty$, with $s_n = o\left(\left(nh_n^{*3}\right)^{1/2}\right)$, such that $\left(nh_n^{*-1}\right)^{1/2} \sum\limits_{t=s_n}^{\infty} \alpha(t)^{1-\gamma} < \infty$, with $\gamma = 2/(2+\delta)$, exists, where h_n^* can be h_n or h_{2n}.

A4 $h_n^* = O\left(n^{-1/(2p^*+3)}\right)$, where h_n^* and p^* can be h_n and p, or h_{2n} and p_2.

A5 $\left\{ h_{1n}^{2(p_1+1)} + (nh_{1n})^{-1} \right\} = o\left(h_{2n}^{p_2+1}\right)$.

Let $K_p(u) = (j! \, |M_p(u)| / |S|) \, K(u)$, where S is the $(p+1) \times (p+1)$ array with $\mu_{i+j}(K)$, $0 \leq i, j \leq p$ as the $(i+1, j+1)$th element, where $\mu_r(K) = \int u^r K(u) du$, and $M_p(u)$ is the same as S with the first column replace by $(1, u, \ldots, u^p)$. We also denote $R(K) = \int R(u)^2 du$. The same kind of functions as $K_p(u)$, denoted by $L_{1,p_1}(u)$ and $L_{1,p_2}(u)$, and notations, but using kernels $L_1(\cdot)$ and $L_2(\cdot)$ will be also used.

Theorem 3.1. *Under A1-A4, for $x \in (h_n, 1-h_n)$, then as $n \to \infty$,*

$$\sqrt{nh_n} \left(\hat{v}_n^S(x) - v(x) - b_S(x) \right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_S^2(x)\right),$$

where

$$b_S(x) = \frac{h_n^{p+1}}{(p+1)!} \left(v^{(p+1)}(x) + \left(m^2(x)\right)^{(p+1)} - 2m(x) m^{(p+1)}(x) \right) \mu_{p+1}(K_p) \tag{6}$$

and

$$\sigma_S^2(x) = \frac{v^2(x)}{f(x)} d(\varepsilon) R(K_p). \tag{7}$$

With respect to estimator $\hat{v}_n^D(x)$, in the following Theorem its asymptotic normality is established. For this, the following additional assumption is used:

A6 The errors $\{\varepsilon_t\}$ follow an MA(∞) process, $\varepsilon_t = \sum_{-\infty}^{\infty} \Psi_j e_{t-j}$, with kurtosis of the white noise $\{e_j\}$ being equal to 0.

Theorem 3.2. *If assumptions A1-A6 are fulfilled and* $\max\{h_{1n}, h_{2n}\} < x < \min\{1-h_{1n}, 1-h_{2n}\}$, *we have*

$$\sqrt{nh_{2n}} \left(\hat{v}_n^D(x) - v(x) - b_D(x) \right) \xrightarrow{\mathcal{L}} N\left(0, \sigma_D^2(x)\right),$$

as $n \to \infty$, where

$$b_D(x) = \frac{h_{2n}^{p_2+1}}{(p_2+1)!} v^{(p_2+1)}(x) \mu_{p_2+1}(L_{2,p_2}) \tag{8}$$

and

$$\sigma_D^2(x) = \frac{v^2(x)}{f(x)} d(\varepsilon) R(L_{2,p_2}). \tag{9}$$

Remark 3.1. *Asymptotic expressions obtained in Theorems 3.1 and 3.2 show that, if the same kernel and bandwidth were used in both estimators ($\hat{v}_n^S(x)$ and $\hat{v}_n^D(x)$), while their asymptotic variances are exactly the same, their asymptotic biases are different, an additional term appearing in the bias of $\hat{v}_n^S(x)$ which can have an adverse effect on the bias of this estimator. This effect is also observed through simulations in the following Section. Moreover, comparing the results obtained in this paper with those obtained by Härdle and Tsybakov [3] for $\hat{v}_n^S(x)$ with random and α-mixing observations* *and Yao and Fan [1] for $\hat{v}_n^D(x)$ with random and absolutely regular observations, respectively, it can be observed that in both cases, while the bias is exactly the same in the random and the fixed design, the variance of both estimators changes, now the term $d(\varepsilon) = \sum_{k=-\infty}^{\infty} Cov(\varepsilon_i^2, \varepsilon_{i+k}^2)$ appearing instead of simply $Var(\varepsilon^2)$ as occurs in the random case.*

4 Simulation study

In this Section, the performance of estimators $\hat{v}_n^S(x)$ and $\hat{v}_n^D(x)$ is illustrated through a simulation study. To observe the real influence of using the residuals instead the errors in $\hat{v}_n^D(x)$ with finite samples, we have also considered in the study the same estimator as $\hat{v}_n^D(x)$, but using the squared errors instead of the squared residuals. This estimator is denoted by $\hat{v}_n^b(x)$.

We simulated 300 samples of size 100 from a fixed and equally spaced model like (1) in the interval $[0,1]$ with errors following an AR(1) process, with $N(0,1)$ distribution. We considered different values of the autocorrelation coefficient, $\rho = -0.9$, -0.6, -0.3, 0, 0.3, 0.6, 0.9 to study the influence of the dependence of the observations. Optimal bandwidths by minimizing the Mean Integrated Squared Error (MISE) were computed. Using Monte Carlo approximations, the integrated squared bias, the integrated variance and the MISE of $\hat{v}_n^S(x)$, $\hat{v}_n^D(x)$ and $\hat{v}_n^b(x)$ were then approximated. For a more comprehensive study, we considered three different problems: to estimate the volatility function on interval $[0,1]$ (global region), to estimate the function on the central region $[0.2, 0.8]$ and to estimate the function on the boundary region $[0, 0.2] \cup [0.8, 1]$. In each case, the optimal bandwidths obtained are different. The kernel function used was the quartic kernel. For brevity, we present here only some representative results obtained.

Table 1 shows the results for $m(x) = \sin(\pi x)$ and $s(x) = \frac{1}{2}x$, when $\rho = 0.3$ (small positive correlation) and $\rho = 0.9$ (strong positive correlation). The optimal bandwidths used to obtain the residuals needed to compute $\hat{v}_n^D(x)$ were 0.2636 and 0.3454 for $\rho = 0.3$ and $\rho = 0.9$, respectively. So, the optimal bandwidths appearing in Table 1 for $\hat{v}_n^D(x)$ and $\hat{v}_n^b(x)$ refer to those used in the second step of these estimators to fit the squared residuals and the squared errors, respectively.

In Table 2, the results for $m(x) = \sin(\pi x)$ and $s(x) = \sin(\pi x)$, when $\rho = -0.6$ (moderate negative correlation) and $\rho = 0.6$ (moderate positive correlation) are presented. Here, the optimal bandwidths needed to obtain the residuals were 0.2525 and 0.4545 for $\rho = -0.6$ and $\rho = 0.6$, respectively.

Finally, Figure 1 shows the evolution of the MISE of $\hat{v}_n^S(x)$, $\hat{v}_n^D(x)$ and $\hat{v}_n^b(x)$ as a function of ρ, in the central part of the interval ($[0.2, 0.8]$) and in the whole interval ($[0,1]$), when $m(x) = 5x$ and $s(x) = \sin(\pi x)$.

Overall, both estimators ($\hat{v}_n^S(x)$ and $\hat{v}_n^D(x)$) had similar, good performance, although, in general, $\hat{v}_n^D(x)$ gave better results, especially in cases where the additional term appearing in the leading term of the bias of $\hat{v}_n^S(x)$ could have an important effect on the estimation. For instance, this happens

	$\rho = 0.3$			$\rho = 0.9$		
Central	$\hat{v}_n^S(x)$	$\hat{v}_n^D(x)$	$\hat{v}_n^b(x)$	$\hat{v}_n^S(x)$	$\hat{v}_n^D(x)$	$\hat{v}_n^b(x)$
h_{MISE}	0.1591	0.4061	0.3737	0.2000	0.3990	0.5050
$\int Bias^2$	0.00008	0.00002	0.00002	0.00091	0.00163	0.00004
$\int Var$	0.00083	0.00036	0.00043	0.00150	0.00034	0.00249
$MISE$	0.00091	0.00038	0.00045	0.00241	0.00197	0.00253
Global						
h_{MISE}	0.1636	1.0000	1.0000	0.1727	1.0000	1.0000
$\int Bias^2$	0.00045	0.00030	0.00016	0.00418	0.00442	0.00019
$\int Var$	0.00228	0.00085	0.00105	0.00198	0.00048	0.00475
$MISE$	0.00273	0.00115	0.00121	0.00616	0.00490	0.00494
Bound.						
h_{MISE}	0.1636	1.0000	1.0000	0.1545	1.0000	1.0000
$\int Bias^2$	0.00098	0.00064	0.00013	0.00884	0.00845	0.00028
$\int Var$	0.00447	0.00149	0.00185	0.00278	0.00077	0.00801
$MISE$	0.00545	0.00213	0.00198	0.01162	0.00922	0.00829

Table 1: $m(x) = \sin(\pi x)$, $s(x) = 0.5\,x$, $\rho = 0.3$ and $\rho = 0.9$.

	$\rho = -0.6$			$\rho = 0.6$		
Central	$\hat{v}_n^S(x)$	$\hat{v}_n^D(x)$	$\hat{v}_n^b(x)$	$\hat{v}_n^S(x)$	$\hat{v}_n^D(x)$	$\hat{v}_n^b(x)$
h_{MISE}	0.4747	0.3990	0.4040	0.4192	0.3232	0.3959
$\int Bias^2$	0.01674	0.01404	0.01327	0.02470	0.02411	0.01613
$\int Var$	0.03095	0.03836	0.03785	0.03481	0.03788	0.04187
$MISE$	0.04769	0.05240	0.05112	0.05951	0.06199	0.05800
Global						
h_{MISE}	0.4343	0.3939	0.3959	0.3434	0.3232	0.3879
$\int Bias^2$	0.01232	0.00876	0.00815	0.01444	0.01465	0.00978
$\int Var$	0.02364	0.02582	0.02574	0.02849	0.02430	0.02789
$MISE$	0.03596	0.03458	0.03389	0.04293	0.03895	0.03767
Bound.						
h_{MISE}	0.1364	0.2576	0.2505	0.1162	0.2626	0.2343
$\int Bias^2$	0.00109	0.00100	0.00106	0.00050	0.00026	0.00075
$\int Var$	0.00745	0.00564	0.00571	0.00535	0.00397	0.00527
$MISE$	0.00854	0.00664	0.00677	0.00585	0.00423	0.00602

Table 2: $m(x) = \sin(\pi x)$, $s(x) = \sin(\pi x)$, $\rho = -0.6$ and $\rho = 0.6$.

in the third model considered, as seen in Figure 1 (see (6) to understand the effect of the regression function on the bias of $\hat{v}_n^S(x)$). On the other hand, significant differences between $\hat{v}_n^D(x)$ and $\hat{v}_n^b(x)$ do not seem to exist and the different results obtained for these estimators are possibly a random sample effect.

Another interesting point observed is that when $|\rho|$ increases, the MISE

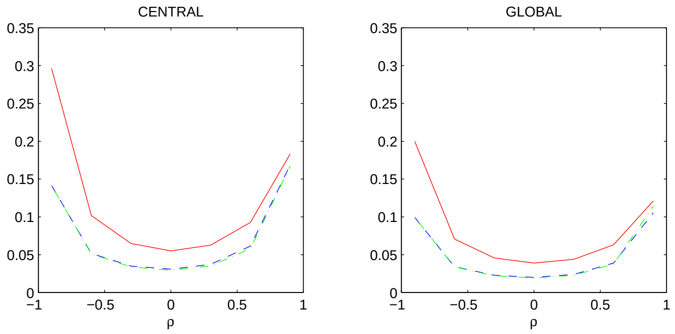


Figure 1: MISE of $\hat{v}_n^S(x)$ (solid line), $\hat{v}_n^D(x)$ (dash-dot line) and $\hat{v}_n^b(x)$ (dashed line) as a function of ρ. Interior points at left and whole interval at right.

of the estimators also increases. This is due to the behavior of the variance and it is compatible with the asymptotic expressions obtained in Section 3. Finally, we have observed that $\hat{v}_n^D(x)$ worked better than $\hat{v}_n^S(x)$ at boundary values. This could be due to the way $\hat{v}_n^D(x)$ is constructed: the known automatic boundary correction proven for the local linear estimator of the regression function is now carried over to the estimation of the conditional variance function when $\hat{v}_n^D(x)$ is used.

5 Example

In this Section, $\hat{v}_n^S(x)$ and $\hat{v}_n^D(x)$ are used to study a real economic data set. The sample data are 222 quarterly observations of the real change in private inventories in the USA, from 1947 to 2002. Each observation indicates the seasonally adjusted annual rate, measured in billions of chained 1996 dollars (in 2002, only two observations are available). The source of these data is the U.S. Department of Commerce, Bureau of Economic Analysis, obtained from the web page: http://www.research.stlouisfed.org/fred/data/gdp.html.

A fixed regression model can be fitted to these data, considering an equally spaced design in $[0, 1]$, that is,

$$Y_t = m\,(t/222) + s(t/222)\varepsilon_t, \quad t = 1, 2, \ldots, 222.$$

The aim is to estimate the volatility function $v(x) = s^2(x)$ using the nonparametric estimators $\hat{v}_n^S(x)$ and $\hat{v}_n^D(x)$. Only the linear case, $p = p_1 = p_2 = 1$, will be considered for both estimators. To compute $\hat{v}_n^D(x)$, first, it is necessary to obtain the residuals from a nonparametric fit. At this point, the bandwidth needed to estimate the regression function was computed by the time series cross-validation (TSCV) method proposed by Hart (1994), producing the result $\hat{h}_{TSCV1} = 0.2229$. The TSCV method was again used, producing the bandwidth $\hat{h}_{TSCV2} = 0.1618$, in the second step of this estimator. Figure 2 shows, on the left part, the sample data and the local linear estimator of $m(x)$ using $\hat{h}_{TSCV1}$, and on the right part, both estimators of the volatility function, $\hat{v}_n^S(x)$ being the dashed line and $\hat{v}_n^D(x)$ the solid line. For simplicity, the bandwidth needed to compute $\hat{v}_n^S(x)$ was selected by ordinary cross-validation, giving as result $\hat{h}_{CV} = 0.1172$. Both estimators have a similar shape, although there is a boundary effect when $\hat{v}_n^S(x)$ is used, especially at values near one, where the estimated volatility decreases. This effect seems not to be present for $\hat{v}_n^D(x)$, as explained in the previous Section.

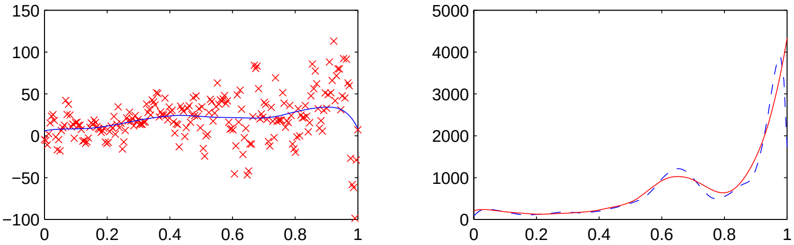


Figure 2: Left picture: observations and local linear estimator of $m(x)$ with $\hat{h}_{TSCV} = 0.2229$. Right picture: $\hat{v}_n^S(x)$ (dashed line) and $\hat{v}_n^D(x)$ (solid line).

References

[1] Fan J., Yao Q. (1998). *Efficient estimation of conditional variance functions in stochastic regression.* Biometrika **85**, 645 – 660.

[2] Francisco-Fernández M., Vilar-Fernández J.M. (2001). *Local polynomial regression estimation with correlated errors.* Communications in Statistics. Theory and Methods **30**, (7), 1271 – 1293.

[3] Härdle W., Tsybakov A. (1997). *Local polynomial estimators of the volatility function in nonparametric autoregression.* Journal of Econometrics **81**, 223 – 242.

[4] Masry E., Fan J. (1997). *Local polynomial estimation of regression functions for mixing processes.* Scandinavian Journal of Statistics **24**, 165 – 179.

[5] Wand M.P., Jones M.C. (1995). *Kernel smoothing.* Chapman and Hall. London.

Acknowledgement: This work was partially supported by Grants BFM2002-00265 (European FEDER support included), PGIDIT03PXIC10505PN and PGIDT01PXI10505PR.

Address: M. Francisco-Fernández, J.M. Vilar-Fernández, Departamento de Matemáticas, Universidad de A Coruña, A Coruña 15071, Spain

E-mail: `mariofr@udc.es`

COMPSTAT'2004 Symposium

BINARY FACTORIZATION OF TEXTUAL DATA BY HOPFIELD-LIKE NEURAL NETWORK

Alexander A. Frolov, Dušan Húsek, Pavel A. Polyakov, Hana Řezanková and Václav Snášel

Key words: Neural networks, binary factorization, application.

COMPSTAT 2004 section: Neural networks and machine learning.

Abstract: We suggest a procedure of binary factorization of signals of large dimension and complexity. The procedure is based on the search of attractors in Hopfield-like associative memory. Starting from random initial state, network activity stabilizes in some attractor which corresponds to one of factors (a true attractor) or one of spurious attractors. Separation of true and spurious attractors is based on calculation of their Lyapunov function. Being applied to textual data the procedure showed sensitivity to the context in which the words were used.

1 Introduction

Factor analysis is one of the most efficient method to overcome informational redundancy of high-dimensional signals. Factors extraction is a procedure which maps original signals into the space of factors. The principal component analysis (PCA) is a classical example of such mapping in the linear case. Linear factorization implies that each original signal can be presented as

$$\mathbf{X} = \mathbf{FS} \tag{1}$$

where $\mathbf{F}$ is a matrix $N \times L$ of factor loadings and $\mathbf{S}$ is a vector of factor scores. Columns of $\mathbf{F}$ represent factors in the original signal space and each component of $\mathbf{S}$ gives contribution of corresponding factor in the original signal. Mapping of original space to the factor space means that signals are represented by vectors $\mathbf{S}$ instead of original vectors $\mathbf{X}$.

In PCA vectors of factor loadings are eigenvectors of covariation matrix $\mathbf{J} = \mathcal{M}\{\mathbf{XX^T}\}$, and dispersions of factor scores are eigenvalues of $\mathbf{J}$. Eigenvector $\mathbf{f}^1$ with the highest eigenvalue Λ_1 (factor with the highest contribution to the total variance of signals $\mathbf{X}$) can be easily obtained [6] by the iterative procedure

$$\mathbf{X}(t+1) = N(\mathbf{h}(t)) \qquad \mathbf{h}(t) = \mathbf{JX}(t) \tag{2}$$

where $N(\mathbf{h}) = \mathbf{h}/|\mathbf{h}|$ denotes vector normalization. Starting from random initial vector $\mathbf{X}_{\text{in}}$, $\mathbf{X}(t)$ tends to $\mathbf{f}^1$. During the iterative procedure Lyapunov function $\Lambda = \mathbf{X}^{\mathrm{T}}\mathbf{JX}$ monotonically increases and reaches Λ_1. When $\mathbf{f}^1$ is obtained, the iterative procedure can be applied to matrix $\mathbf{J} - \Lambda_1 \mathbf{f}^1 \mathbf{f}^{1^{\mathrm{T}}}$ to obtain the next eigenvector of matrix $\mathbf{J}$, and so on.

This procedure can be obviously described in terms of neural network approach. Covariation matrix $\mathbf{J}$ corresponds to a matrix of synaptic connections obtained by Hebbian learning. The iterative procedure corresponds to evolution of activity in neural network with parallel dynamics where $\mathbf{h}$ is a vector of synaptic excitations. And substraction of the found factor from covariation matrix corresponds to Hebbian unlearning. Linear and even some nonlinear PCA procedures have been actually realized by neural network approach [4], [5], but only for special cases of nonlinearity.

One particular form of nonlinear factorization is a binary one, where a complex vector signal (pattern) has a form of the Boolean sum of weighted binary factors:

$$\mathbf{X} = \bigvee S_l \mathbf{f}^l. \tag{3}$$

In this case, original signals, factor scores and factor loadings are binary, i.e. possess the values 0 or 1. It was a challenge [3] to utilize for binary factorization the Hopfield-like neural network with parallel dynamics because it has a lot of similarities with the iterative procedure described above for linear factorization. First, the connection matrix of this network is a covariation matrix of input signals obtained by Hebbian learning:

$$J_{ij} = \sum_{m=1}^{M} (X_i^m - q^m)(X_j^m - q^m),\ i{\neq}j,\ J_{ii} = 0, \tag{4}$$

where M is the number of patterns in the learning set and $q^m = \sum\limits_{i=1}^{N} X_i^m / N$ is the total activity of the m-th pattern. Second, its activity is determined by the same iterative procedure (2) except that normalization of vector of synaptic excitations $\mathbf{h}$ is replaced by its binarization:

$$X_i(t+1) = \Theta(h_i(t) - T(t)),\ i = 1, \cdots, N \tag{5}$$

where Θ - step function, and $T(t)$ - activation threshold. And third, its activity has almost the same Lyapunov function

$$\Lambda(t+1) = \mathbf{X}^{\mathrm{T}}(t+1)\mathbf{J}\mathbf{X}(t). \tag{6}$$

This formula slightly differs from that of linear case because activity of Hopfield-like network with parallel dynamics converges not only to point attractors but also to cyclic attractors of the length two [2].

Theoretical analysis and computer simulation performed by Frolov et al. [3] completely confirmed the validity of Hopfield-like network for binary factorization. Since neurons that represent one common factor tend to fire together, they become more tightly connected than neurons belonging to different factors. Hence, factors create attractors of the network dynamics similarly to eigenvectors of the correlational matrix in iterative procedure for linear case. However, Hopfield-like network has one principal peculiarity. The network dynamics converges to one of the factors (true attractor) only when initial state falls inside its attraction basin. Otherwise it converges to one of the spurious attractors. Note that for linear case it converges to one of the factors starting from any random initial state. Thus binary factorization requires special recall procedure to separate true and spurious attractors. In this study we describe one of these procedures and computer experiments with some kinds of artificial and natural binary signals.

2 Recall procedure

To separate true and spurious attractors we suggest two-run recall procedure. Its initialization starts by presentation of random initial pattern $\mathbf{X}_{\text{in}}$ with $k_{\text{in}} = r_{\text{in}}N$ active neurons. On presentation of $\mathbf{X}_{\text{in}}$, network activity $\mathbf{X}$ evolves to some attractor. The evolution is determined by equation (5). On each time step k_{in} "winners" (neurons with the greatest synaptic excitation) are chosen and only they are active on the next time step. When activity stabilizes at the initial level of activity k_{in}, $k_{\text{in}}+1$ neurons with maximal synaptic excitation are chosen for the next iteration step, and network activity evolves to some attractor at the new level of activity $k_{\text{in}}+1$. Then level of activity increases to $k_{\text{in}}+2$, and so on, until number of active neurons reaches the final level $r_{\text{f}}N$. Thus, the whole procedure (one trial) contains $(r_{\text{f}} - r_{\text{in}})N$ iteration steps and several time steps inside each iteration step to reach some attractor for fixed level of activity.

At the end of each iteration step a relative Lyapunov function was calculated by formula: $\lambda = \Lambda/(rN)$ where Λ is given by (6). The relative Lyapunov function gives a mean synaptic excitation of active neurons. The time course of the relative Lyapunov function along the recall trajectory provides criterion for separation of true and spurious attractors (see later). Attractors with the highest Lyapunov function would be obviously winners in the most trials of the recall process. Thus, more and more trials are required to obtain new attractor with relatively small value of Lyapunov function. To overcome this problem attractors with high Lyapunov function should be deleted from the network memory. The deletion was performed according to Hebbian unlearning rule by substraction $\Delta J_{ij}, j \neq i$ from synaptic connections J_{ij} where

$$\Delta J_{ij} = \frac{\eta}{2} J(\mathbf{X})[(X_i(t-1) - r)(X_j(t) - r) + (X_j(t-1) - r)(X_i(t) - r), \quad (7)$$

$J(\mathbf{X})$ is the average synaptic connection between active neurons of the attractor, $\mathbf{X}(t-1)$ and $\mathbf{X}(t)$ are patterns of network activity at last time steps of iteration process, r is the level of activity, and η is an unlearning rate. For point attractor $\mathbf{X}(t) = \mathbf{X}(t-1)$ and for cyclic attractor $\mathbf{X}(t-1)$ and $\mathbf{X}(t)$ are two states of attractor.

3 Computer simulation

To reveal peculiarities of true and spurious attractors we performed computer experiments with artificial signals. Each pattern of the learning set is supposed to be a Boolean superposition of exactly C factors, where each factor is supposed to contain exactly $n = pN$ entries of value 1 and $(1-p)N$ entries of value 0. Thus, each factor $\mathbf{f}^l \in B_n^N$ and for each pattern of the learning set, vector of factor scores $\mathbf{S} \in B_C^L$ where $B_n^N = \left\{ \mathbf{X} | X_i \in \{0,1\}, \sum_{i=1}^{N} X_i = n \right\}$. We supposed factor loadings and factor scores to be statistically independent. As an example, Fig. 1 demonstrates changes of relative Lyapunov function for $N = 3000$, $L = 5300$, $p = 0.02$ and $C = 10$. Recall process started at $r_{\text{in}} = 0.005$.

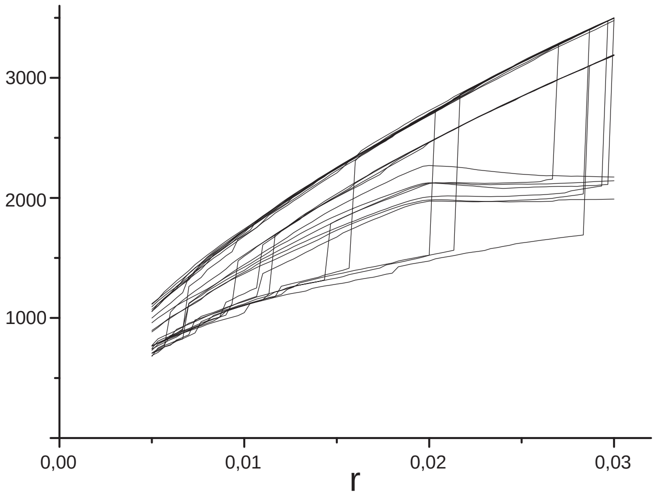


Figure 1: Relative Lyapunov function λ in dependence on the relative network activity r for artificial input signals with $p = 0.02$, $C = 10$, $N = 3000$ and $L = 5300$.

As shown in Fig. 1, neurodynamic trajectories form three clearly separated modes by distribution of Lyapunov function. The mode with highest values of Lyapunov function contains only two trajectories. It appeared that attractors creating these trajectories consist of neurons which were most often and most rare activated in the learning set. For large r all trajectories converge to these attractors. The middle mode contains true attractors and the mode with smallest Lyapunov function contains local spurious attractors. Along the recall trajectories there exist many transitions from attractors with low to those with high Lyapunov function. For true trials the change of relative Lyapunov function in dependence on r has specific breaking at the point

$r = p$. For $r < p$ the increase of r provides the increase of relative Lyapunov function (a mean synaptic excitation of active neurons) due to increase of the number of active neurons which are all tightly connected inside some factor. For $r > p$ the increase of r occurs due to joining of neurons not belonging to factor and therefore slightly connected with neurons of factor. That is why a mean synaptic excitation decreases when r increases. In the recall procedure namely this breaking at the curve $\lambda(r)$ was used as an index of true trials to distinguish them from spurious ones.

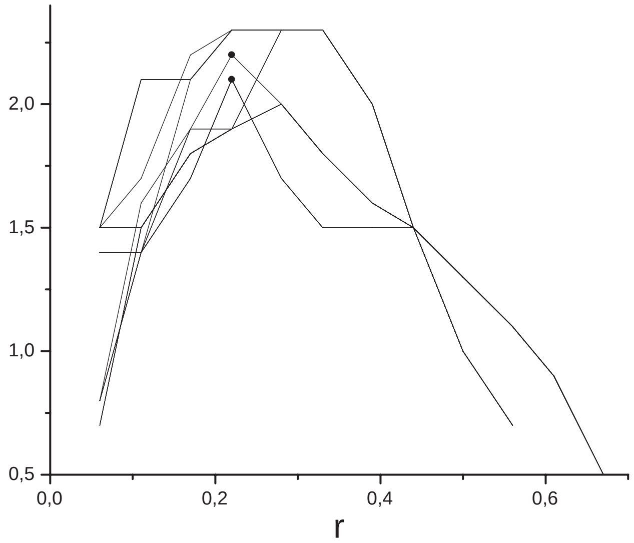


Figure 2: Relative Lyapunov function λ in dependence on the relative network activity r for 15 titles of medical articles presented in [1]. Circles are points of breaking which were identified as indexes of factors.

Fig. 2 demonstrates example of binary factorization over the list of titles of 15 medical articles presented in [1]. The titles were transformed to binary vectors with 18 component. The obtained binary codes of the titles were stored in the network of 18 neurons according to (4). Each trial was initiated by activation of one of 18 neurons. Thus the total recall procedure includes only 18 trials. Only two factors were revealed according to the use criterion. The first factor contains words: blood, close, disease and pressure. The second: fast, rats, rise and pressure. It is interesting that the words “culture”, “discharge” and “patients” do not create a factor in spite of the fact that they are included into two first titles and, hence, one can expect that they should be tightly connected. However in these titles the word “culture” has different meaning and its banding with words “discharge” and “patients” is not reasonable. Thus our method happened to be sensitive to the context in which the words are used.

We also applied our method to the set of 21000 messages of Reuters agency [7], [8]. The used vocabulary contained 3000 the most often words in the set (consequently network contained 3000 neurons). Each message was transformed to binary code dependently on presence or absence of words in the message. Each found factor was deleted from the network memory according to (7) with $\eta = 1$. Fig. 3 demonstrates the first 14 trials which were identified as true. Circles mark the points of curve breaking. All factors happened to be reasonable and mirror the content of the corresponding messages. For example the factor with the highest Lyapunov function contains 16 words: April, co, company, corp, cts, inc, loss, net, note, qtr, revs, share, shares, shr, stock, York. The words corresponds to the context of stock exchange. It is interesting that several words which are synonyms (share, shares, shr) and which were rarely included into the same message, are included in this factor. This is another example that our method combines words in factors not only according to the frequency of their appearance together at the messages but mainly according to their appearance at the same context.

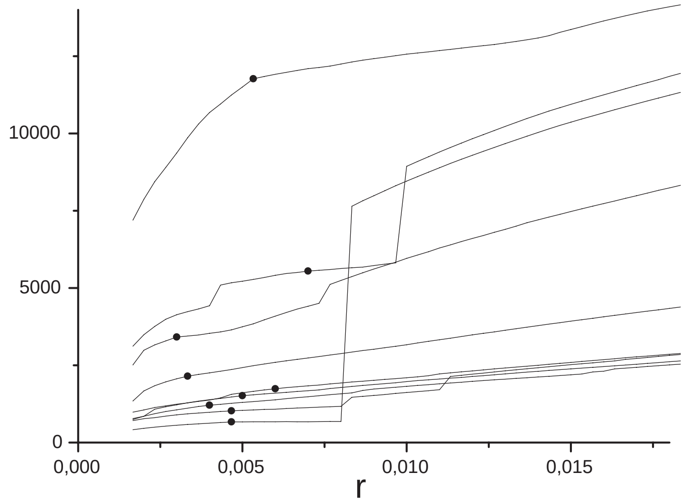


Figure 3: Relative Lyapunov function λ in dependence on the relative network activity r for 21000 messages of Reuters agency. Circles are points of breaking which were identified as indexes of factors.

4 Conclusion

Hopfield-like neural network is capable of performing binary factorization of the signals of high dimension and complexity. Being applied to textual data our method showed sensitivity to the context in which the words were used.

References

[1] Berry M.W., Dumais S.T., Letsche T.A. *Computational methods for intelligent information access.* `http://www.cs.utk.edu/ berry/sc95/sc95.html`.

[2] Goles-Chacc E., Fogelman-Soulie F., Pellegrin, D. (1985). *Decreasing energy functions as a tool for studying threshold networks.* Discrete Mathematics **12**, 261 – 277.

[3] Frolov A.A., Sirota A.M., Husek D., Muraviev I.P., Polyakov P.J. (2004). *Binary factorization in Hopfield-like neural networks: single-step approximation and computer simulations.* Neural Networks World (in press).

[4] Karhunen J., Joutsensalo J. (1994). *Representation and separation of signals using nonlinear PCA type learning.* Neural Networks **7**, 113 – 127.

[5] Oja E., Ogawa H., Wangviwattana J. (1991). *Learning in nonlinear constrained Hebbian network.* Proc. ICANN-91, Espoo: Finland, 385 – 390.

[6] Watkins D.S. (2002). *Fundamentals of matrix computations* (Second edition). John Wiley & Sons, Inc., N.Y. T.G. Rose, M.

[7] Reuters `http://about.reuters.com/researchandstandards/corpus/`

[8] Stevenson, Whitehead, M. (2002). *The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources* [245k PDF]. In Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, 29-31 May 2002.

Acknowledgement: This work was supported by Ministry of Education of the Czech Republic under No. LN00B096.

Address: A.A. Frolov, Institute of Higher Nervous Activity and Neurophysiology of the Russian Academy of Sciences, Butlerova 5a, Moscow, Russia
D. Húsek, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic
P.A. Polyakov, Institute of Optical Neural Technologies of the Russian Academy of Sciences, ul. Vavilova 44, korpus 2, 119333 Moscow, Russia
H. Řezanková, University of Economics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
V. Snášel, Dept. of Computer Science, Faculty of Electrical Engineering and Computer Science, VŠB – Technical University of Ostrava, 17.listopadu 15, 708 33 Ostrava-Poruba, Czech Republic

E-mail: `aafrolov@mail.ru, dusan@cs.cas.cz, pavel@8ka.mipt.ru, rezanka@vse.cz, vaclav.snasel@vsb.cz`

POSSIBILITIES AND PROBLEMS OF THE XML-BASED GRAPHICS IN STATISTICS

Tomokazu Fujino, Yoshikazu Yamamoto and Tomouki Tarumi

Key words: Statistical graphics, Web-based application, XML, e-learning.

COMPSTAT 2004 section: Internet based methods.

Abstract: In this paper, we will discuss the possibilities and problems with the XML-based graphics format in statistical visualization through application examples.

1 Introduction

XML (eXtensible Markup Language) has been widely prevalent as a standard data format on the Web since W3C (World Wide Web Consortium) issued a recommendation for XML1.0 in 1998. We pay particular attention to the XML-based vector graphics format in many XML-based technologies. In this paper, we will discuss the possibilities and problems of these in statistical visualization with a focus on SVG (Scalable Vector Graphics) and X3D (eXtensible 3D), which are XML-based two-dimensional and three-dimensional vector graphics formats, respectively.

The typical solution for statistical visualization on the Web would be to display the image file generated in advance. But the dynamic creation of a statistical graph by the programming language like Java Applet, Java Script and PHP has recently been a mainstream of the solution. These solutions have some problems like a reliance on a specific graphics library, cumbersome programming and so on. XML-based graphics has gotten a lot of attention recently as which the format can clear up such problems. There are three main reasons for applying the XML-based vector graphics format to the statistical visualization instead of existing raster graphics format such as JPEG, GIF, PNG and so on. The first reason is that it is the vector graphics format. In this kind of format there is no deterioration of the image quality by zooming operation. The second reason is the fact that it is XML-based format. It turns out that many people can take an active interest in developing its related tools and receive the benefit from them because it is open and standard technology. In addition, XML-based means that it is easy to cooperate with other XML-based format by XSLT(XSL Transformations) and DOM(Document Object Model) and to implement interactivity by using DOM. The third reason is that it is the plane text format, so that it can hold the statistical data for the graph, which is accessible to users and developers and can easily generate programming languages and statistical languages such as S and R without relying on any one library for graphics.

The point that we would like to emphasize here is its portability. Many features such as cooperation with other XML-based format and interactivity can be included into one graphics file so it can be used not only as a part of an application but also as an independent application itself. Even if it is generated on server side, there is no traffic between the server and the client when the interactive function works. The most applications by using XML-based vector graphics can be also created by using Macromedia FLASH. The big difference between FLASH and XML-based vector graphics would be that one is closed binary file, and the other is XML format as open and standard specification. FLASH requires commercial applications like FLASH MX for its development but XML-based vector graphics can be generated and manipulated by any programming language and edited even by usual text editor because it is the plane text file. We suppose that the division of roles between FLASH and XML-based vector graphics would proceed in the future, that is, FLASH would specialize in Web contents which has rich visual effects and interactivity and XML-based vector graphics would be used in GIS application and statistical graph.

We will explore the possibility of XML-based graphics through some practical example in following sections. Some of them can be viewed at `http://www.fwu.ac.jp/fujino/Xg4stat/`.

2 SVG

2.1 Outline of SVG

SVG is a XML format to describe two-dimensional vector graphics. In the days before SVG, microsoft-led VML(Vector Markup Language) and adobe-led PGML(Precision Graphics Markup Language) was proposed to W3C. SVG1.0 was released in September 2001 by W3C as integrated format from them. The current version is SVG1.1 and its recommendation was released in January 2003. Some software and plug-ins for displaying SVG are now available. The most popular plug-in would be Adobe SVG Viewer, which works on Internet Explorer on Windows98-XP and MacOS8.6-9.1. However, the MacOS version only supports static SVG. In this paper, we will assume Adobe SVG Viewer 3.0 for Windows as the environment for displaying SVG. On the other hand, some of authoring tools for statistical graph released by many software companies have begun to support SVG recently.

Figure 1 is the SVG scatter plot on Internet Explorer with SVG Viewer 3.0.The first line of the source code is the XML declaration, and in the second line the document type definition of SVG is described. The subsequent lines make up the XML instance of SVG.

The interactivity of SVG can be realized through SMIL (Synchronized Multimedia Integration Language) Animation or DOM. DOM is a specification that defines the API for processing XML documents and the object model processed by the interface. SVG version 1.0 supports the main

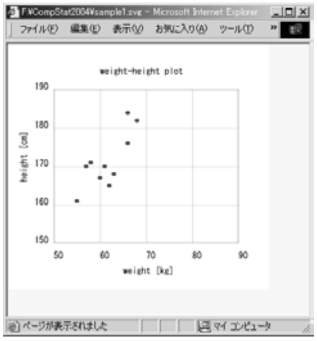

F:¥CompStat2004¥sample1.svg - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T)
weight-height plot
190
180
170
160
150
height [cm]
50
60
70
80
90
weight [kg]
ページが表示されました
マイ コンピュータ

Figure 1: Scatter plot by SVG (whplot.svg)

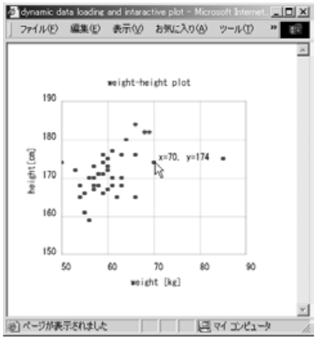

dynamic data loading and interactive plot - Microsoft Internet...
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T)
weight-height plot
190
180
170
160
150
height [cm]
x=70, y=174
50
60
70
80
90
weight [kg]
ページが表示されました
マイ コンピュータ

Figure 2: Implementation of interactivity in SVG

parts of the DOM level 1 and 2, and DOM in SVG is now implemented by Java(ECMA)Script.

2.2 Specific example of DOM in SVG

2.2.1 Dynamic loading XML format data When building a Web application including a visualization of statistical data, it is not desirable that the drawing program and graphics themselves contain data. In other words, it is important for such a system to create the mechanism that a program loads independent external data files according to their needs and draw graphics based on the data.

In section 2.1, the example of the static SVG scatter plot was illustrated. We will give a specific example of the dynamic loading of the external XML document, including the data for the scatter plot. Actually, SVG currently doesn't have a function for dynamic loading of external XML documents. But it is possible to achieve dynamic loading by using the DOM in the HTML. The specific procedure is that the script in the HTML document (`plotmain.html`) loads the external XML documents (`plotdata.xml`) and reflects the loaded data in the embedded SVG document (`whplot.svg`) as the external file by manipulating the DOM tree of the SVG.

2.2.2 Interactive graph The typical interactivity of a scatter plot is to obtain the coordinates of the plot point in response to the position of the mouse pointer. We will show the way to achieve such a feature by using DOM. Figure 2 illustrates the embedded scripts in the above-mentioned "`whplot.svg`". In addition, the text tag `<text id="info" x="0" y="0"> </text>` for displaying the coordinates of the plot point must be added to the "`whplot.svg`" and the description `onmouseover="mouse_over()"`, which calls the script the moment the mouse pointer passes over an object, must be added to the circle tag in "`whplot.svg`" which has `id="pt"`. When that script is called by the mouse movement, the values of the attribute x and y of the use tag are set to the text element of the text tag which has `id="info"`, and the x-coordinate and y-coordinate of the current mouse pointer are set to the attribute x and y of the text tag.

2.3 Inline SVG

In many cases, SVG files are displayed when they are invoked from "embed" tag in a HTML file. Inline SVG is the technique to display SVG by writing the SVG code directly into HTML. Some builds of Mozilla and Amaya, Web browsers released by W3C, have native rendering engine to display SVG, so inline SVG can be displayed in these Web browsers. However, Internet Explorer doesn't support native rendering of SVG, so to display inline SVG it needs to call the plug-ins, such as Adobe SVG Viewer, as an ActiveX object. There are some examples of inline SVG that can be displayed by using Internet Explorer 5 or later in a combination with Adobe SVG Viewer 3.0 in Schepers.cc. In general, a web page including image files consists of some files but a web page can be made by a single file by introducing inline SVG. This would be an effective technique for dynamically generating Web pages including graphics. However, there are some drawbacks to inline SVG, that is, users can't use the zoom-in/out function and the user's environment for displaying SVG is restricted.

The statistical software we are now developing displays all results of computations as HTML by using a component of Internet Explorer to achieve visual appeal and understandability. When the results include the statistical graph, inline SVG is used. This makes it possible to display the results without generating temporary files.

2.4 SVG in R

In this section we will discuss the possibilities of SVG in R, which is a statistical computing environment. R has basic functions for statistical graphics but most of them display statistical graphics that have only the bare visual effects in the default. Users then have to make full use of low-level drawing functions in order to display the statistical graphics with rich visual effect. When it comes to interactive statistical graphics, it is possible to know the coordinates of plot points on a mouse pointer by using the identify and locator function but it can be used only in the environment of R.

RsvgDevice, the library for using SVG as an output device for the graphics of R, is released by T.J.Ruciani. By using this library the statistical graphics of R can be generated in SVG format by the same operation as other output devices such as PostScript, JPEG and graphics window. This output is not only of high quality but supports zoom-in/out so it can be used for presentations and immediate Web delivery. However, the output of RSvgDevice doesn't fully exploit the advantages of SVG, that is, RSvgDevice cannot make SVG file including scripts for interactivity, and our requirement that users can obtain graphics with good visual effects and interactivity in simple operation cannot be achieved using only RSvgDevice. To solve this problem, we have begun to develop the original library "SvgOutput" for generating SVG statistical graphics for R. SvgOutput now supports basic statistical graphics. For example, the command `SvgPlot(rnorm(100),rnorm(100))` generates scatter plot, including a function that can display the coordinates of the plot point on the mouse pointer like in Figure 3. Of course text labels, the size of the graphics, the color of the plot points and the background and enabling interactive function can be specified through options. In future releases, low level drawing functions and functions to display graphics in real time using SVG Viewer from R will be implemented.

3 X3D

X3D is a virtual reality modeling language and an open standard for 3D on the Web. X3D replaces VRML(Virtual Reality Modeling Language) but also provides compatibility with existing VRML content and browsers. The X3D Graphics Working Group is designing and implementing X3D specifications. X3D Encoding has both an XML encoding and a classic VRML encoding. Shared virtual worlds written in X3D format is achieved by the X3D plug-in or the X3D browser. Internet Explore 6.0 is equipped with a X3D plug-in, the MediaMachines Flux X3D/VRML97 plug-in. FreeWRL is a plug-in for Mac OS and Linux OS. As well as SVG, it is possible to implement interactive facilities. An X3D language binding standard will specify bindings for the API to the ECMAScript and Java languages.

The advantage of statistical graphics using X3D is that it doesn't require the programming for the two-dimensional representation of three-dimensional data. In other words, a programmer can refer to three-dimensional coordinate directly. This leads to a reduction of his burden. Moreover, the biggest advantage of the graphical expression by X3D is that virtual space can be turned and moved by the X3D plug-in without programming. The 3D scatter plot displayed by MediaMachines Flux on Internet Explorer can be rotated by dragging the mouse (Figure 4).

4 Cooperation with GIS

There have been many studies about the effective use of XML based vector graphics format in GIS (Geographic Information System) in recent years. The traditional web-based GIS application using raster graphics has never satisfied users until now because a lot of traffic between a server and clients, that arises when zooming in/out or moving and displaying the related information of the map, would prevent the smooth manipulation. In addition, the developers of such web-based application have to do many tasks like making the boundary information for a clickable map and developing the application for clients. XML-based graphics format would easily solve such kinds of problems. Moreover, many studies about the cooperation between XML-based

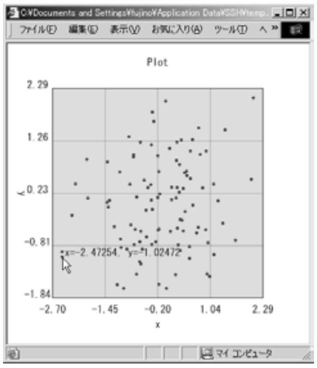

C:¥Documents and Settings¥fujino¥Application Data¥SSH¥temp...
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T)
Plot
2.29
1.26
0.23
-0.81
-1.84
y
-2.70
-1.45
-0.20
1.04
2.29
x
x=-2.47254, y=-1.02472
マイ コンピュータ

Figure 3: Scatter plot by SvgOutput

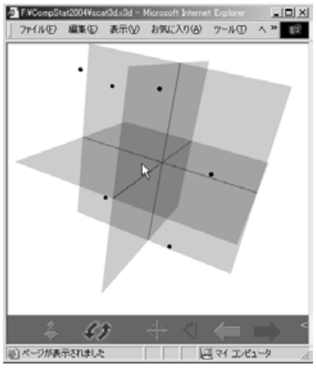

F:¥CompStat2004¥scat3dx3d - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T)
ページが表示されました
マイ コンピュータ

Figure 4: Three-dimensional scatter plot by X3D

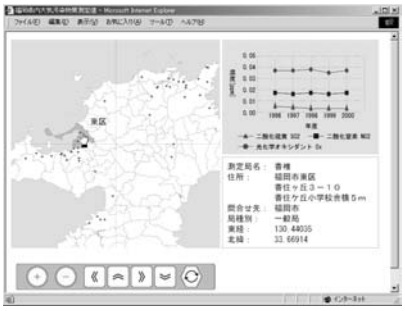
vector graphics format and GML (Geography Markup Language), which is XML encoding for describing geographic information, are actively carried out. Future developments in these studies are anticipated.

As an example of the utilization of SVG in the GIS sector we will show our web application. SVG graphics in Figure 5 illustrates the observation station for the concentration of air pollutants in Fukuoka Prefecture in western Japan and observational data at each station. The points on the map are observation stations. The line plots above right are the secular variations of the concentration of the three kinds of air pollutants. The address, the station type and information on the details of the station are displayed at lower right. When placing the mouse pointer on the map, the name of the municipality indicated by the mouse pointer appears near the pointer, and when placing the mouse pointer on the observation point, the concentration of the air pollutants and their detailed information will appear on the right side. This application consists of five files, which are one HTML file, two SVG files (map area and another area) and two XML data files (location data of observation stations and their concentration data).

5 Utilization of statistical education

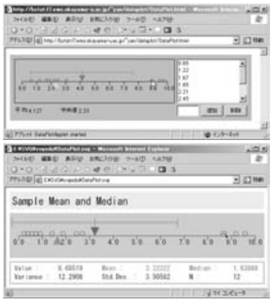
There are many contents for statistical education on the web at present. Some projects provide statistical courses, including interactive objects, in order to make these contents more effective and appealing. Most programs are written in Java and perform as Java Applet. However, complex actions may not be required for these kinds of interactive objects so it will be possible for XML-based vector graphics format to play that role. This will bring about an improvement in development efficiency and the quality of images and animations.

The CASE Project in Japan has been developing the contents for statisti-



測定局名： 香椎
住所： 福岡市東区
香住ヶ丘３－１０
香住ケ丘小学校舎積５ｍ
問合せ先： 福岡市
局種別： 一般局
東経： 130.44035
北緯： 33.66914

Figure 5: The concentration of air pollutants



Sample Mean and Median

Value :	8.48578	Mean :	3.22222	Median :	1.83889
Variance :	12.2908	Std.Dev :	3.50582	N :	12

Figure 6: Simulation of Median and Mean by Java and SVG

cal education using Java and Flash. Figure 6 is a screen shot of SVG version of a Java Applet to help understand the properties of mean and median, which were developed by the project. Users can put points on the one-dimensional coordinate axis freely and experience the change of the mean and median depending on the distribution of the points. In fact, the SVG version was even better than Java Applet version in terms of time and energy to develop them and its quality.

Fortunately, MathML, which is the XML encoding for describing equations, and DocBook, which is the XML encoding for scientific papers and books, have become popular in the general Internet environment. We have a plan to make the text for statistics on the web using such XML technologies. This will play a role in the interactive contents on the web and a printed textbook for statistical education at the same time.

6 Conclusion

So far, we have only emphasized the merit in the case of applying the XML-based vector graphics format to statistical visualization. However, there are also some problems. One of them is security. The mechanism of displaying XML-based vector graphics is that a viewer or plug-in on a client side interprets the received text files and displays them as graphic data, so it is possible for users to refer to the text data directly. To hide the text data from view might be easy technically, however, it would not attack the root of the problem. This is a serious problem in a web application that provides statistical information without users knowing about detailed data by displaying graphics only. As mentioned above, a XML-based graphics format could possibly play an important role in cooperation with GIS. However, to generate graphics including a map dynamically, the use of such a format would lead to the problem of allowing reuse of the map data that has commercial value itself. In future versions of SVG, the encryption framework would be contained to specifications.

Source code

whplot.svg

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 20010904//EN"
 "http://www.w3.org/TR/2001/REC-SVG-20010904/DTDsvg10.dtd">
<svg width="340" height="320">
<text id="title" x="120" y="40">weight-height plot</text>
<g transform="translate(60,60)">
<svg width="240" height="200">
<rect x="0" y="0" width="240" height="200" stroke="black"
 fill="none"/>
<line x1="0" y1="50" x2="240" y2="50"/>
 :
<line x1="180" y1="0" x2="180" y2="200" />
</svg>
<svg width="240" height="200" viewBox="50 -190 40 40"
 preserveAspectRatio="none">
<g id="dpt" transform="scale(1,-1)">
<defs>
<circle id="pt" cx="0" cy="0" r="0.5" fill="red" />
</defs>
<use xlink:href="#pt" x="60" y="167" />
 :
<use xlink:href="#pt" x="61" y="170" />
</g></svg></g>
<text id="xlab" x="150" y="300">weight [kg]</text>
<text x="60" y="280">50</text>
 :
<text x="300" y="280">90</text>
<text id="ylab" x="25" y="180"
 transform="rotate(-90,25,180)">
height [cm]</text>
<text x="35" y="260">150</text>
 :
<text x="35" y="60">190</text>
</svg>
```

plotdata.xml

```
<?xml version="1.0"?>
<plotdata>
<title>weight-height plot</title>
<xlab>weight [kg]</xlab>
<ylab>height[cm]</ylab>
```

```
<x>
 60 66 66 57 68 55 58 62 63 61
 60 57 57 59 60 53 60 56 57 58
 66 85 55 54 59 59 54 68 50 70
 55 59 61 61 58 63 56 69 63 64
</x>
<y>
 167 184 176 170 182 161 171 165 168 170
 175 173 168 171 173 172 172 159 167 168
 165 175 166 168 174 176 165 182 174 174
 161 166 168 177 171 170 170 182 176 180
</y>
</plotdata>
```

plotmain.html

```
<?xml version="1.0"?>
<!DOCTYPE html
   PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
 xml:lang="ja" lang="ja">
<head>
<title>dynamic data loading and intaractive plot</title>
<script type="text/javascript">
<!--
 function onload(){
   plotdata = new ActiveXObject("Microsoft.XMLDOM");
   plotdata.async=false;
   plotdata.load("plotdata.xml");
   pdata = plotdata.documentElement;
   splot = document.svgplot.getSVGDocument().
                    documentElement;
   title = pdata.getElementsByTagName("title");
   splot.getElementById("title").firstChild.
         setNodeValue(title[0].firstChild.nodeValue);
   xlab = pdata.getElementsByTagName("xlab");
   ylab = pdata.getElementsByTagName("ylab");
   splot.getElementById("xlab").firstChild.
         setNodeValue(xlab[0].firstChild.nodeValue);
   splot.getElementById("ylab").firstChild.
         setNodeValue(ylab[0].firstChild.nodeValue);
   x = pdata.getElementsByTagName("x")[0].
             firstChild.nodeValue.split(" ");
   y = pdata.getElementsByTagName("y")[0].
```

```
             firstChild.nodeValue.split(" ");
   tmp = splot.getElementById("dpt").
               getChildNodes.item(3);

   for(var i=0;i<x.length;i++){
     insnode = tmp.cloneNode(true);
     insnode.setAttribute("x",x[i]);
     insnode.setAttribute("y",y[i]);
     splot.getElementById("dpt").
           insertBefore(insnode,tmp);
   }
 }
//-->
</script></head>
<body onload="onload()">
<embed name="svgplot" width="340"
       height="320" src="whplot.svg" />
</body></html>
```

Script for interactivity

```
<script type="text/javascript"><![CDATA[
function mouse_over(evt){
   valueX = evt.target.getAttribute("x");
   valueY = evt.target.getAttribute("y");
   info = document.documentElement.getElementById("info");
   info.setAttribute("x",evt.clientX+5);
   info.setAttribute("y",evt.clientY);
   info.firstChild.setNodeValue("x="+valueX+", y="+valueY);
}
]]></script>
```

References

[1] *World Wide Web Consortium*, http://www.w3c.org/

[2] *XML 1.0 (Second Edition)*, http://www.w3c.org/TR/REC-xml/

[3] *SVG 1.1 Specification*, http://www.w3.org/TR/SVG11/

[4] *Web3D Consortium*, http://www.web3d.org/

Address: T. Fujino, Fukuoka Women's Univ., 1-1-1 Kasumigaoka, Fukuoka 813-8529, Japan
Y. Yamamoto, Tokai University, 1117 Kita-Kaname, Hiratsuka 259-1292, Japan
T. Tarumi, Okayama Univ., 1-1-1 Tsushima-Naka, Okayama 700-8530, Japan
E-mail: fujino@fwu.ac.jp, yoshiro@yamamoto.name
tarumi@ems.okayama-u.ac.jp

COMPSTAT'2004 Symposium

COMPARISON OF SOME RATIO AND REGRESSION ESTIMATORS UNDER DOUBLE SAMPLING FOR NONRESPONSE BY SIMULATION

Wojciech Gamrot

Key words: Regression estimator, double sampling, nonresponse.

COMPSTAT 2004 section: Simulations.

Abstract: The two-phase (or double) sampling scheme is one of the techniques used to reduce the bias due to nonresponse. Several ratio-type estimators for double sampling under nonresponse were proposed by Rao [2] and properties of these estimators were discussed for the deterministic nonresponse. In this paper some unbiased estimators of population variance and covariance under double sampling for nonresponse are considered. A regression estimator for two-phase sampling under nonresponse is then constructed. Its approximate bias and Mean Square Error for the deterministic nonresponse mechanism are derived using Taylor linearization technique. Further, a simulation study using Monte Carlo method is carried out to determine the behaviour of both estimators and compare their properties under stochastic-type logistic nonresponse mechanism. The data from 1996'Polish Agricultural Census are used in the simulation process.

1 Deterministic nonresponse model

Let us assume that the mean value $\overline{Y}$ of some characteristic Y in the finite and fixed population U of the size N is to be estimated and that nonresponse mechanism is deterministic. Therefore, the population may be divided into two non-overlapping strata U_1 and U_2, of unknown sizes N_1 and N_2 respectively, such that population units belonging to U_1 always provide data if contacted whereas units from U_2 always refuse to co-operate in the survey. Let us also denote $W_1 = N_1/N$ and $W_2 = N_2/N$.

2 Two-phase sampling scheme

The survey is executed in two phases. In the first phase a simple random sample s of the size n is drawn without replacement from the population, according to the sampling design:

$$P_1(S) = \binom{N}{n}^{-1}. \tag{1}$$

In the sample s some units belong to the stratum U_1 and hence they respond. However some units belong to U_2 and consequently do not respond. So the sample is randomly divided into two subsets $s_1 \subset U_1$ and $s_2 \subset U_2$, of the sizes n_1 and n_2 such that $s_1 \cup s_2 = s$, $s_1 \cap s_2 = \emptyset$ and $n_1 + n_2 = n$. It is worth noting that subset sizes n_1 and n_2 are random variables following hypergeometric distribution function and cannot be controlled directly. In the second phase of the survey a subsample s' of the size $n' = cn_2$ (where $0 < c < 1$) is drawn without replacement from s_2, with conditional probability:

$$P_2(s'|n_2) = \binom{n_2}{n'}^{-1}. \tag{2}$$

Another contact attempt is then undertaken for each subsampled unit and it is assumed that all these units respond in the second phase.

3 Linear estimator and ratio estimator

We will now present the linear and ratio estimators. Let us define

$$C_U(X,Y) = \frac{1}{N-1}\sum_{i\in U} x_i y_i - \frac{1}{N(N-1)}\sum_{i,j\in U} x_i y_j \tag{3}$$

$$S_U^2(X) = \frac{1}{N-1}\sum_{i\in U} x_i^2 - \frac{1}{N(N-1)}\sum_{i,j\in U} x_i x_j \tag{4}$$

$$C_{U_2}(X,Y) = \frac{1}{N_2-1}\sum_{i\in U_2} x_i y_i - \frac{1}{N_2(N_2-1)}\sum_{i,j\in U_2} x_i y_j \tag{5}$$

$$S_{U_2}^2(X) = \frac{1}{N_2-1}\sum_{i\in U_2} x_i^2 - \frac{1}{N_2(N_2-1)}\sum_{i,j\in U_2} x_i x_j \tag{6}$$

$$\overline{y}_1 = \frac{1}{n_1}\sum_{i\in s_1} y_i \qquad \overline{y}_{s'} = \frac{1}{n'}\sum_{i\in s'} y_i \tag{7}$$

$$w_1 = n_1/n. \qquad w_2 = n_2/n. \tag{8}$$

It is well known, that the following statistic is an unbiased estimator of the population mean under nonresponse:

$$\widehat{\overline{y}} = w_1\overline{y}_1 + w_2\overline{y}_u \tag{9}$$

An important fact is that the unbiasedness of this estimator does not depend on the nonresponse mechanism, which has been proven by Särndal et al. [3]. Its variance is given by:

$$V(\overline{y}_w) = \frac{N-n}{Nn}S_U^2(Y) + \frac{W_2}{n}\left(\frac{1-c}{c}\right)S_{U_2}^2(Y) \tag{10}$$

Let us denote $R = \overline{Y}/\overline{X}$. When the population mean of an auxiliary characteristic X is known, Rao [2] proposed the following ratio estimator of the population mean:

$$\widehat{\overline{y}}_{ratio} = \frac{\widehat{\overline{y}}}{\widehat{\overline{x}}}\overline{X} \tag{11}$$

For this estimator and deterministic nonresponse he derived approximate expression for the bias:

$$B(\widehat{\overline{y}}_{ratio}) = \frac{1}{\overline{X}}\frac{N-n}{Nn}\left(RS_U^2(X) - C_U(X,Y)\right) + \\ + \frac{1}{\overline{X}}\frac{1-c}{c}\frac{W_2}{n}\left(RS_{U_2}^2(X) - C_{U_2}(X,Y)\right) \tag{12}$$

and approximate expression for the MSE:

$$MSE(\widehat{\overline{y}}_{ratio}) = \frac{N-n}{Nn}\left(R^2S_U^2(X) - 2RC_U(X,Y) + S_U^2(Y)\right) + \\ + \frac{1-c}{c}\frac{W_2}{n}\left(R^2S_{U_2}^2(X) - 2RC_{U_2}(X,Y) + S_{U_2}^2(Y)\right) \tag{13}$$

4 Regression estimator

We will now introduce the regression estimator. Let us define:

$$\overline{X}_{s_1} = \frac{1}{n}\sum_{i\in s_1} x_i \qquad \overline{Y}_{s_1} = \frac{1}{n}\sum_{i\in s_1} y_i \tag{14}$$

$$\overline{X}_{s'} = \frac{1}{n}\sum_{i\in s'} x_i \qquad \overline{Y}_{s'} = \frac{1}{n}\sum_{i\in s'} y_i \tag{15}$$

$$C_{s_1}(X,Y) = \frac{1}{n_1-1}\sum_{i\in s_1}(x_i - \overline{X}_{s_1})(y_i - \overline{Y}_{s_1}) \tag{16}$$

$$C_{s'}(X,Y) = \frac{1}{n'-1}\sum_{i\in s'}(x_i - \overline{X}_{s'})(y_i - \overline{Y}_{s'}) \tag{17}$$

And consider the following statistic:

$$\widehat{C}(X,Y) = \frac{n_1-1}{n-1}C_{s_1}(X,Y) + \frac{cn_2(n-1)-n_1}{n(n-1)c}C_{s'}(X,Y) + \\ + \frac{nn_1}{2n_2(n-1)}\left(\overline{X}_{s_1} - \widehat{\overline{x}}_{s_1}\right)\left(\overline{Y}_{s_1} - \widehat{\overline{y}}_{s_1}\right) + \\ + \frac{nn_2}{2n_1(n-1)}\left(\overline{X}_{s'} - \widehat{\overline{x}}_{s'}\right)\left(\overline{Y}_{s'} - \widehat{\overline{y}}_{s'}\right) \tag{18}$$

It may be proven that under two-phase sampling scheme described above, the statistic (18) is an unbiased estimator of $C_U(X,Y)$. Consequently, under this two-phase sampling scheme, the statistic:

$$\widehat{S}^2(X) = \frac{n_1-1}{n-1}S^2_{s_1}(X) + \frac{cn_2(n-1)-n_1}{n(n-1)c}S^2_{s'}(X) +$$
$$+\frac{nn_1}{2n_2(n-1)}\left(\overline{X}_{s_1} - \widehat{\overline{x}}_{s_1}\right)^2 + \frac{nn_2}{2n_1(n-1)}\left(\overline{X}_{s'} - \widehat{\overline{x}}_{s'}\right)^2$$

is an unbiased estimator of $S^2_U(X)$. This allows us to construct the following regression estimator:

$$\widehat{\overline{y}}_{reg} = \widehat{\overline{y}} + \frac{\widehat{C}(X,Y)}{\widehat{S}^2(X)}(\overline{X} - \widehat{\overline{x}}) \tag{19}$$

Let us denote $B = \widehat{C}(X,Y)/\widehat{S}^2(X)$ and $Z = BX - Y$. Using second order Taylor linearization technique its approximate bias under deterministic nonresponse and two-phase sampling scheme may be expressed in the form:

$$B(\widehat{\overline{y}}_{reg}) = \frac{N-n}{n(N-2)}\frac{1}{S^2_U(X)}\left(C_U(X,XZ) - \overline{Z}S^2_U(X) - \overline{X}C_U(X,Y)\right) +$$
$$+\frac{1}{n-1}\frac{1-c}{c}\frac{1}{S^2_U(X)}\left(\left(\frac{W_2}{n} - W_2^2\right)\left(\overline{X}_{U_1}C_{U_2}(X,Z) + \overline{Z}_{U_1}S^2_{U_2}(X)\right) +\right.$$
$$\left.+W_2C_{U_2}(X,XZ) + \frac{W_2}{n}\left(\overline{Z}_{U_2}S^2_{U_2}(X) + 3\overline{X}_{U_2}C_{U_2}(X,Z) - C_{U_2}(X^2,Z)\right)\right) \tag{20}$$

Its approximate MSE under two-phase sampling is:

$$MSE(\widehat{\overline{y}}_{reg}) = \frac{N-n}{Nn}S^2_U(BX-Y) + \frac{W_2}{n}\frac{1-c}{c}S^2_{U_2}(BX-Y). \tag{21}$$

The expressions (13) and (21) may be used to compare the accuracy of both estimators.

5 Estimators under stochastic nonresponse - simulation results

Estimators (11) and (19) were constructed under assumption that nonresponse is deterministic. This does not have to be true. In more general case, we treat the response or lack of response as a random event, and assign some probability ρ_i of responding to each i-th population unit. In such case, the expressions given above, describing the biases and MSE's of these estimators are not applicable.

To determine the properties of these estimators under some stochastic nonresponse model, a simulation study was carried out. The aim of the study was to determine how the bias and MSE of both estimators depend on initial sample size n, and to compare the bias and MSE of both estimators under this model. In other words, the simulations were executed to determine the properties of estimators in some case of nonresponse model misspecification.

During the simulations, the data obtained during 1996' agricultural census for certain municipalities in the Dabrowa Tarnowska district represented the population under study. The total of 2422 units were used in simulation. The variable under study, denoted by Y was total sales of the farm in the year 1995. As an auxiliary variable X the farm area (in acres) was used. For every population unit the response probabilities were generated according to arbitrarily chosen logistic model, given by expression:

$$\rho_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \tag{22}$$

with arbitrarily chosen parameters $\beta_0 = -4$ and $\beta_1 = -0.003$, which resulted in average reponse probability in population equal to $0,89$. For simplicity, only the first-phase response was treated as a random event and it was assumed that in the second phase each unit responds. So the response probabilities corresponded only to the first-phase unit behaviour. The assumed nonresponse model implies, that for units with large values of auxiliary variable the response probability is lower. This might be justified by the reluctance to disclose the higher incomes or sales, e.g. due to high crime rate or tax evasion.

The experiments were carried out by repeatedly drawing without replacement simple random samples from the population. To represent the stochastic nonresponse mechanism, for each unit included in the initial sample an independent random trial was executed with the probability of success equal to this unit's response probability. A unit was assumed to respond if the outcome of the trial was a success and treated as nonrespondent otherwise. For the resulting set of nonrespondents a simple subsample of the size equal to the 30% of the first-phase nonrespondent number was drawn without replacement, and treated as responding. On the basis of each sample-subsample pair generated this values of linear estimator, ratio estimator and regression estimator were computed. For every value of n a total of 200000 samples were drawn from the population. On the basis of of empirical distribution of estimates, the bias and mean square error of each estimator were evaluated. Simulations were executed for initial sample size $n = 20, 40, \ldots, 400$.

The dependence between initial sample size n and relative accuracy of ratio and regression estimator (the proportion of MSE of ratio/regression estimator to the MSE of linear estimator) is shown on Figure 1. It is easy to notice, that both ratio and regression estimators are more accurate than linear estimator, hence their relative accuracy is below one for any value of

n. Except for the value of $n = 40$ the regression estimator was more accurate than ratio estimator but the difference was modest and did not exceed 4%.

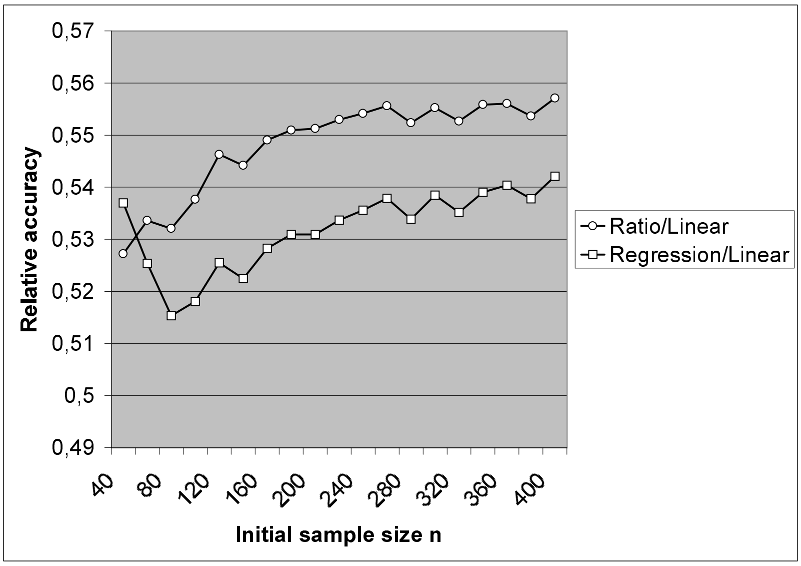


Figure 1: Relative accuracy of estimators as a function of initial sample size.

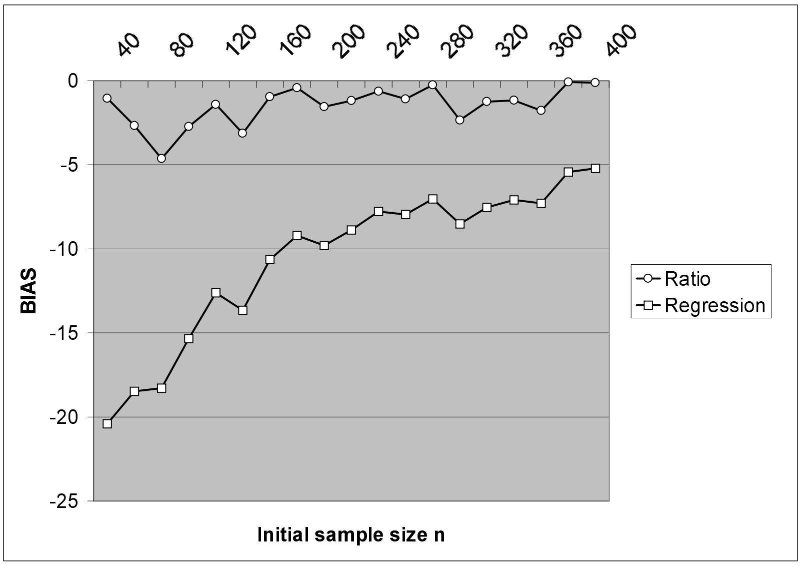


Figure 2: Biases of estimators as a function of initial sample size.

The dependence between initial sample size n and biases of both estimators is shown on Figure 2. Both estimators are biased negatively. The bias of regression estimator is much greater than the bias of ratio estimator. With the increase of initial sample size, biases of both estimators diminish. These results also indicate, that for $n > 40$ the variance of regression estimator is lower than the variance of ratio estimator. To investigate the impact of biases on the MSE of both estimators, the proportion of bias to the MSE as a function of initial sample size is shown on Figure 3. As it can be seen on the graph this proportion does not exceed 0,01% for ratio estimator and 0,1% for regression estimator. These results indicate that for both estimators the main source of error is their non-systematic variability represented by variance, and that in practice the bias of both estimators may be treated as negligible.

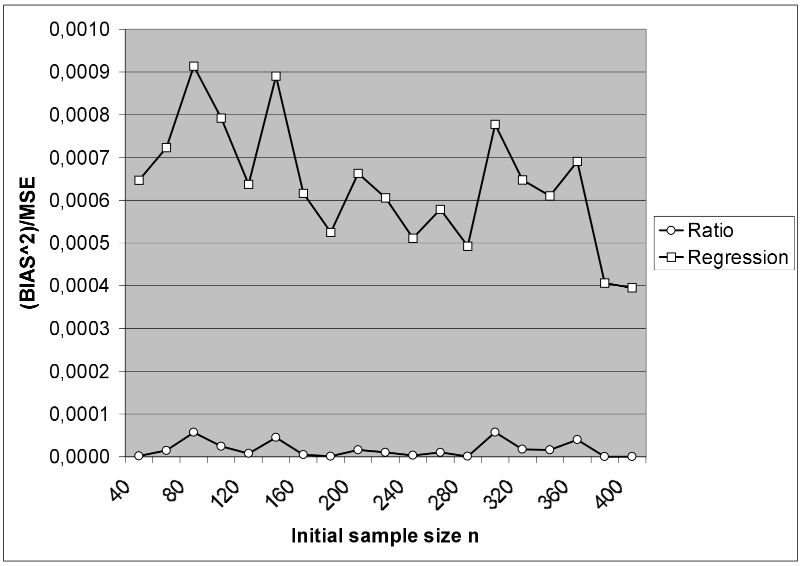


Figure 3: Biases of estimators as a function of initial sample size.

6 Conclusions

In this paper a regression estimator under two-phase sampling for nonresponse was introduced. Its bias and MSE under deterministic nonresponse mechanism were derived using Taylor linearization technique. For deterministic nonresponse the MSE of regression estimator is not greater than the MSE of the ratio estimator proposed by Rao [2]. In order to compare the properties of both estimators under some stochastic-type logistic nonresponse model a simulation study was performed.

The simulation results suggest that ratio and regression estimators of the population mean under two phase sampling and stochastic-type logistic nonresponse may be more accurate (in terms of MSE) than the linear estimator. It was also observed, that the contribution of the bias in the MSE of estimators was very modest, which suggests, that in practice both of them may be treated as nearly unbiased.

It should however be stressed, that the simulation results and conclusions are based on the assumption that the arbitrary logistic nonresponse model holds, and that the distribution of auxiliary variable is close to the distribution used in simulations. Although the simulation results sched some light on the properties of estimators, care should be taken when generalizing these conclusions to different situations.

References

[1] Gamrot W. (2004). *On some ratio estimatiors in two-phase sampling for nonresponse.* To be published in: Prace Naukowe AE Katowice.

[2] Rao P.S.R.S. (1986). *Ratio estimation with subsampling the nonrespondents.* Survey Methodology **12**, 217 − 230.

[3] Särndal C.E., Swensson B., Wretman J.H. (1992). *Model assisted survey sampling.* Springer-Verlag New York.

Address: W. Gamrot, Department of Statistics, University of Economics, Bogucicka 14, 40-226 Katowice, Poland

E-mail: gamrot@ae.katowice.pl

COMPARISON OF THREE STATISTICAL CLASSIFIERS ON A PROSTATE CANCER DATA

Eva Gelnarová and Libor Šafařík

Key words: Early detection, prediction, prostate cancer, cross-validation, logistic regression, neural networks, k-nearest neighbour, Gower coefficient of similarity.

COMPSTAT 2004 section: Applications.

Abstract: *Introduction:* The dataset of 826 patients who were suspected of the prostate cancer was examined. The best single marker and the combination of markers which could predict the prostate cancer in very early stage of the disease were looked for. *Methods:* For combination of markers the logistic regression, the multilayer perceptron neural network and the k-nearest neighbour method were used. 10 models for each method were developed on the training data set and the predictive accuracy verified on the test data set. *Results and conclusions:* The ROCs for the models were constructed and AUCs were estimated. All three examined methods have given comparable results. The medians of estimates of AUCs were 0.775, which were larger than AUC of the best single marker.

1 Introduction

There are approximately 1 100 deaths per year caused by prostate cancer in The Czech Republic [5]. The prostate cancer is a common disease affecting 14% of men over 50 and 28% over 70 years of age, who apparently do not suffer from malignant disease, yet. The most treacherous sign of the early prostate cancer is that it is asymptomatic for a long period of time to strike heavily only in advanced stages, when it is impossible to get cure, so far. Due to the unsatisfying outcomes of treatment of the late stages of prostate cancer, the physicians currently look more intensively on the early detection/screening models, since the results of treatment of the early stages of prostate cancer are much more successful.

The natural reguierement on models of early detection would be the easiness of obtaining satisfactory successful predictors. The prostatic specific antigen (PSA) and its derivatives - easy obtained from the blood serum, are considered as good markers. Unfortunately all these markers are organ specific and not cancer specific. They indicate not only the prostate cancer - malignant disease but also the benign prostatic hyperplasia (BPH). When the level of such a marker is increased - it is a task to predict/discriminate which disease this patient suffers from. It is obvious that the predictive power would increase when all the available information is used. For combination of all single markers and other quantities one can utilize the classical statistical methods, newly described artificial neural networks [8], [3] and others methods of data mining [2].

So, the first aim of this work is to describe a particular patient population. The second aim is to find the best method of discrimination between malignant disease and BPH using easy obtained markers and demonstrate its abilities on the particular patient population.

2 Description of the data set

Men treated in The Urological Clinic in Prague from April 1999 to Decembre 2003 with PSA between 0-20 ng/ml were included into the study. The number of patients with complete records (no missing values) was 826. The discrimination properties of several methods will be demonstrated on this data set.

The following markers and quantities were observed for each patient: *Prostate volume*(Volume) is a size of the gland measured in cm^3. *Prostatic specific antigen* (total PSA, PSA) is a glycoprotein that is produced primarily by the epithelia cells of the prostate gland, PSA consists of two forms - free (freePSA) and complexed. PSA in the blood serum was measured in ng/ml. *Density of PSA* (PSAD) is a ratio of total PSA and Volume,

$$PSAD = \frac{PSA}{Volume}.$$

Fraction is a percentage of freePSA contained in PSA,

$$fraction = 100\frac{freePSA}{PSA}.$$

Patients were examined by *Digital Rectal Examination*(DRE) which can be positive or negative. *Age* was approximated in years and also categorized: younger that 60, 60 - 70, older than 70 (treated as 3 dummy variables: $I_{age\leq 60}$, $I_{60<age<70}$, $I_{70\leq age}$ (0/1)). The malignant disease was *detected* if the evaluation of any transrectal ultrasound (TRUS) biopsy speciment was positive.

Each patient is representented as a 9-dimensional random vector. For purpose of the analysis let us treat the detection as the dependent variable and denote Y. Volume, PSA, PSAD, fraction, DRE and categorized age will be treated as the independent variables and overall denoted by $\mathbf{X}$.

3 Methods

To distinguish the malignant disease from the BPH using the easy obtained markers and quantities is a classification problem. Several methods have been developed, we have investigated the following three ones:

LR - Logistic regression - we have modeled the conditioned probability $\hat{p}(j)$ of having a malignant disease for a j^{th} patient,

$$\hat{p}(j) = P(Y(j) = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp\left(\sum_{i=1}^{m} \alpha_i x_i(j)\right)}{1 + \exp\left(\sum_{i=1}^{m} \alpha_i x_i(j)\right)}.$$

The classification itself was then realized by comparing $\hat{p}(j)$ with an appropriate cut point.

NN - Neural networks - we have concentrated on the special type of the NN - the multilayer perceptron neural network, *MLP*. The number of perceptrons in an input layer was 8 (perceptrons with the values of Volume, PSA, PSAD, fraction, DRE, $I_{age \leq 60}$, $I_{60 < age < 70}$, $I_{70 \leq age}$). There was one hidden layer with 10 perceptrons and one perceptron in an output layer (see the Figure 1).

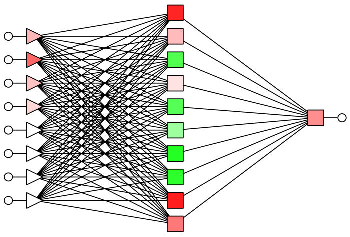


Figure 1: Structure of the used multilayer perceptron neural network.

k-Nearest Neighbour (k-NN) - the prediction for a certain patient is derived from the value of the k-nearest neighbours (in sence of a certain metric) from the set with known disease (train set). The *Gower coefficient of similarity*[1] between patients i and j, $S(i, j)$, was used as a metric in this case. $S(i, j)$ is able to combine several types of variables (dichotomous, qualitative and quantitative). Let us define $S(i, j)$ by

$$S(i, j) = \frac{1}{m} \sum_{l=1}^{m} s_l(i, j),$$

where m is the dimension of the vector of independent variables, which characterizes a patient.

Let us suppose that l-th independent variable is dichotomous or qualitative. Than $s_l(i, j) = 1$ in case the values $x_l(i)$ and $x_l(j)$ are the same or $s_l(i, j) = 0$ in case the values differ. In case that l-th independent variable is quantitave, we set

$$s_l(i,j) = 1 - \frac{|x_l(i) - x_l(j)|}{R_l},$$

where R_l is $\max_{i,j \in D}(|x_l(i) - x_l(j)|)$ on the investigated data set D.

Let us go back to the description of the k-NN method. The value of dependent quantity ($\hat{y}(j)$) is derived from the kernel function, which combines the values $y(i),\ i = 1, \ldots, k$, of the k-nearest neighbours ($y(i) = 1$ if the i^{th} patient suffers from malignant disease, $y(i) = 0$ if the i^{th} patient suffers from BPH). The used kernel function was following

$$\hat{y}(j) = \sum_{i=1}^{k} S(i,j) y(i).$$

The classification itself was then realized by comparing $\hat{y}(j)$ with an appropriate cut point.

The ideal marker should have high *sensitivity* and high *specificity* as well. Sensitivity is the probability that the prediction is positive in case that patient really suffers from malignant disease. Specificity is the probability that the prediction is negative in case that patient suffers only from the BPH.

For markers measured on a continuous scale it is convenient to construct the empirical *Receiver Operating Characteristics*, ROC, and to use the *Area Under a ROC Curve*, AUC. AUC is a measure of the diagnostic accuracy for that marker. The larger AUC, the better the marker separates.

The estimate of AUC, $\widehat{AUC}$, was computed with the usage of the Wilcoxon statitistics [6] (in case of logistic regression and k- nearest neighbour).

The predictive power of the model should be examined on a new data set. The idea of *cross-validation* is to divide randomly the dataset into n equal sized disjunct groups. Each group is once used for testing when as the model is constructed from the remaining groups. Cross-validation was used in case of LR, k-NN, and even in case of study of single markers. Cross-validation was used also in case of NN when each group was once used for testing and once used for validation.The remaining groups served as a training group. The dataset was divided into 10 groups with 82 or 83 patients.

The data were analysed using the software R, version 1.8.1 for Windows (logistic regression, k-nearest neighbour) and Statistica 6.0 (neural networks).

4 Results and discussion

The general description of the whole patient dataset (N=826) is displayed in the Table 1. (The notation SD means standart deviation. The values of independent variables for patients with the malignant disease and BPH were compared with two-sample nonparametric Mann-Whitney test.) As can be seen there are significant differences between the values of the patients with malignant disease and the BPH - the p-values of Mann-Whitney test are very low for all measured single markers. The discrimination ability of DRE was following: sensitivity was 0.817 and specificity 0.410. These independent variables were used as the entries to the three procedures mentioned above. The AUCs for the single markers(independent variables) are presented in the top part of Table 2. Median(AUC) is the median of 10 values of AUCs obtained by cross-validation. MAD is the median absolut deviation. Notice that PSA gives the poorest results (0.587), oppositely the best performance show PSAD (0.703) and Volume (0.710).

Table 1	Benign Disease(BPH) 502 patients mean (SD)	Malignant Disease 324 patients mean (SD)	p- value of M-W test
Age($years$)	65.7 (8.7)	69.9 (8.1)	< 0.001
Volume(cm^3)	49.7 (23.5)	36.5(20.8)	< 0.001
Serum PSA(ng/ml)	7.91 (3.2)	9.25(4.18)	< 0.001
PSAD($ng/(ml \cdot cm^3)$)	0.192(0.120)	0.310(0.193)	< 0.001
fraction(%)	18.1(9.14)	15.12(9.94)	< 0.001
DRE(No. of positives)	59	133	-

Table 2	median(AUC)	MAD(AUC)
Age	0.651	0.046
Volume	0.710	0.036
PSA	0.587	0.04
PSAD	0.703	0.0176
fraction	0.635	0.055
LR - aditive model	0.775	0.02
MLP (8-20-10-1)	0.775	0.054
29-NN	0.775	0.063

When the single markers were combined with the logistic regression (the bottom part of Table 2), the aditive model without any interactions was developed, the median of AUCs has substantially improved, comparing with the median of AUCs for PSA. The median (0.775) was even larger than the median of AUCs of the best single markers. The variable selection has not been applied. All following independent variables - even they were correlated (Spearman correlation coefficient for PSA and PSAD was 0.5) - were used: PSA, PSAD, log(Volume), fraction, DRE, $I_{60<age<70}$, $I_{70\leq age}$. When any independent variable would be dropped out of the model, the median of AUCs would decreas. The ROCs for LR are plotted in the Figure 2.

Almost identical results as in case of LR were obtained by running the MLP. The median of AUCs was the same (0.775), the MAD was a bit larger (0.054).

The dependance of the estimate of AUC on the number of the neighbours k is presented in the Figure 3. The optimal number of neighbours seems to be about 29. The results are presented in the bottom line of Table 2. The median for AUCs is also 0.775, but the MAD is 3 times larger than MAD of LR. The ROCs for 29-NN are plotted in the Figure 4.

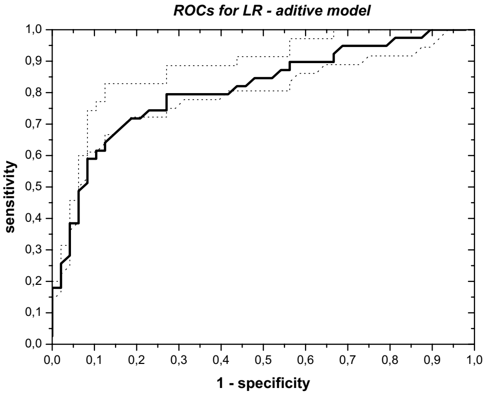


Figure 2: ROCs for the LR-aditive model. Solid line is the ROC, with the median AUC. Dotted lines represent ROCs with the minimum and maximum AUCs.

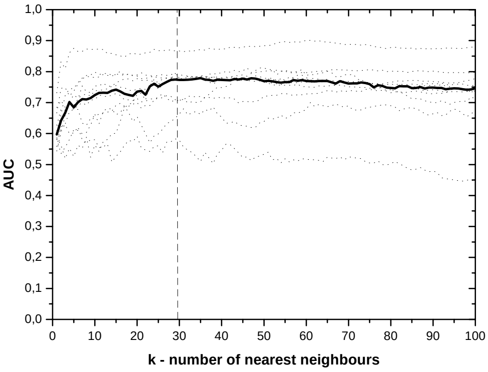


Figure 3: The dependence of the estimate of AUC on the number of neighbours. Dotted lines represent the estimates on each of the 10 disjunct groups (cross-validation). Solid line represent the medians of values on 10 sets for each k.

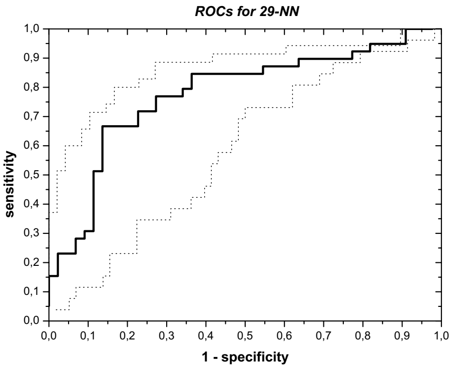


Figure 4: ROCs for the 29-NN. Solid line is the ROC, with the median AUC. Dotted lines represent ROCs with the minimum and maximum AUCs.

All these three methods gave the comparable results, which is surprising because each of then is sensitive on different type of disturbance. NN requires the high quality training data (correctly classiffied cases); estimates of coefficients of LR can be influenced by the presence of outliers; on the contrary the k-NN is robust and classification not influenced by outliers.

Unfortunately the prostate data are contamined with the gross-errors as missplaced decimal points or typing errors. There have occured many outliers (due to the natural biological variability, e.g. very large prostates with volumes larger then 200 cm^3, or due to typing errors).

But main source of distortion was the presence of wrongly classified patients. As it has been already described - a patient is said to suffer from malignant disease when any speciment obtained by TRUS biopsy is positive. But let us realize that TRUS biopsy is random and there is always chance that the biopsy needle misses the tumor despite the fact that patient suffers from malignant disease, all speciments could be negative. This phenomenon was deeply investigated by Stricker [7], his simulations showed among others that the probability of cancer detection by TRUS biopsy is 0.44 if the tumor volume is 5% of prostate volume. Rousseeuw [4] treats this problem theoreticaly as a problem of a model under which the observed response variable is strongly related but not equal to the unobservable true response.

In our data set there were 109 patients with negative biopsy who underwent the rebiopsy and 29 of the rebiopsies were positive. Rebiopsy was after a time period, which is, unfortunately, unknown and various for each patient. In case, when a rebiopsy is positive and biopsy was negative, it can not be decided, if the first biopsy failed or malignant disease striked the patient after the first biopsy. So the data can not be modified.

References

[1] Gower J.C.(1971). *A general coefficient of similarity and some its properties.* Biometrics **27**, 857 – 871.

[2] Mařík V., Štěpánková O., Lažanský J. et al. (2003). *Umělá inteligence(4).* Academia.

[3] Remzi M., Anagnostou T., Ravery V., Zlotta A., Stephan C., Marberger M., Djavan B. (2003). *An artificial neural network to predict the outcome of repeat prostate biopsies.* Urology **62** (3), 456 – 460.

[4] Rousseeuw P., Christmann A. (2002). *Robustness against separation and outliers in binary regression.* ICORS.

[5] Sabra R. et al. (1996). *Karcinom prostaty do roku 2000.* Maxdorf.

[6] Skalská H. (2003). *ROC curve and estimates of AUC.* 2nd International Conference APLIMAT, 645 – 649.

[7] Stricker H.J., Ruddock J.Wan, Belville W.D. (1993). *Detection of nonpalpable prostate cancer. A mathematical and laboratory model.* British Journal of Urology **71**, 43 – 46.

[8] Wei J.T., Zhang Z., Barnhill S.D., Madyastha K.R., Zhang H., Oesterling J.E. (1998). *Understanding artificial neural networks and exploring their potential applications for the practicing urologist.* Urology **52** (2), 161 – 172.

Acknowledgement: This work was supported by project GA CR 402/03/0084.

Address: E. Gelnarová, Charles University, Prague, Faculty of Mathematics and Physics, Department of probability and mathematical statistics, Sokolovská 83, Prague, Czech Republic

L. Šafařík, Charles University, Prague, 1st School of Medicine, Czech Republic

E-mail: `eva.gelnarova@matfyz.cz`

KNOWLEDGE DISCOVERY WITH CLUSTERING: IMPACT OF METRICS AND REPORTING PHASE BY USING KLASS

Karina Gibert, R. Nonell, J.M. Velarde and M.M. Colillas

Key words: Ill-structured domains, heterogeneous data matrices, clustering, mixed metrics, class-interpretation, software.

COMPSTAT 2004 section: Clustering, Statistical data mining, statistical software, data visualization.

Abstract: One of the features involved in clustering is the evaluation of distances between individuals. This paper is related with the use of different mixed metrics for clustering messy data. Indeed, in real complex domains it becomes natural to deal with both numerical and symbolic attributes. This can be treated on different approaches. Here, the use of mixed metrics is followed. In the paper, impact of metrics on final classes is studied. The application relates to clustering municipalities of the metropolitan area of Barcelona on the bases of their constructive behavior, the number of buildings of different types being constructed, or the politics orientation of the local government. Importance of the *reporting* phase is also faced in this work. Both clustering with several distances and the interpretation oriented tools are provided by a software specially designed to support Knowledge Discovery on real complex domains, called *KLASS*.

1 Introduction

It is clear that nowadays analysis of complex systems is an important handicap in Statistics, Artificial Intelligence, Information Systems, Data visualization, and other fields. Describing the structure or obtaining knowledge of complex systems is known as a difficult task. *Knowledge Discovery and Data Mining, (KDD)* is a research area where all those fields interact in order to extract useful knowledge from data [4]. Besides, clustering is one of the most used Data Mining techniques since it permits data separation into groups.

In fact, we agree with the idea that a number of real applications in *KDD* either require a clustering process or can be reduced to it [19]. Also, formation of classes is one of the basic methods used by human beings in aprehending the world. That's why, many expert systems [22] are indeed classifiers.

However, when facing *ill-structured domains, (ISD)* as mental disorders, sea sponges, disabilities or municipalities behaviours... clustering has to be done on data matrices where both numerical and categorical variables exist. For short, in *ISD* [5] [11] consensus among experts is weak —and sometimes non-existent; quantitative and qualitative information coexist in *heterogeneous* data bases; and, even more, the more the expert knows about the domain, the greater is the number of modalities he uses for categorical variables.

This work refers clustering with heterogeneous data matrices and this requires special attention when facing *ISD*. Standard clustering methods were originally conceived for numerical variables. Upon [1], three main strategies, more extensively discussed in [11], may be followed: *i) Partitioning* the variables upon their type, then analysing the dominant type [18]; *ii) Converting* all variables to a unique type, conserving as much original information as possible[1]; many authors[1] [5], discuss on this line; *iii) Compatibility measures* covering any combination of variable types; the idea is to allow clustering on heterogeneous data matrices without transforming the variables themselves. Main advantages of this approach are that it is respecting the original nature of data, there is not loss of information, no need to take previous arbitrary decisions which can bias results, enables study of all types of variables together, enables analysis of interactions between variables of different types.

Upon discussions on [11], [5] and [14] the last proposal is also our approach. Since in the core of clustering distances between individuals are needed, a function to do it with heterogeneous data is required. In the literature, several proposals are found: Gower 71 [16], Gowda & Diday 91 [15], Gibert 91 [11], [10], [6], Ichino & Yaguchi 94 [17], Ralambondrainy 95 [20]. In this work, five mixed metrics are used for clustering real heterogeneous data, all successfully implemented in a system called *KLASS* described below.

Actually, this paper covers two complementary goals. First, to study the behavior of different mixed metrics on a set of real heterogeneous data, in order to see if relevant differences appear in the resulting clusters. Formal approach to this problem requires complex theoretical development. That's why an applied approach is followed in this research. Previously [12] some proves restricted to Gibert's [11][6] and Ralambondrayni's [20] metrics were made with experimental data; a similar experiment is in [3], where the performance in clustering of metrics defined in [16] and [15] are compared. Next step is to compile comparative results of several real applications for inducing theoretical hypothesis to be later tested. This works goes in this direction.

Second goal, also based on *KLASS*, is to emphasize the *reporting* phase, so needed in a *KDD* system. Actually, given a partition of a large set of objects tools for assisting the user in the interpretation tasks are needed, in order to establish the *meaning* of the resulting classes. Often, and especially in a *KDD* context, it is not enough for the user to automatically obtain the classes, but to understand *why* those classes where detected. Indeed, this is a key point of a *KDD* system [4], since is the bridge to make the extracted knowledge explicit, which, we think it is as important as the analysis itself.

[1] In Statistics, traditionally, symbolic variables have been converted to a set of binary variables; then, clustering with χ^2 metrics is suitable [18]. In Artificial Intelligence (AI), grouping of quantitative values into a qualitative one is much more popular.

This paper is organized as follows: after the general introduction, the metrics used in this work are in 2. In 3 the application and main results are presented. Paper ends with some conclusions and future work.

2 The metrics involved

The standard input of a clustering algorithm is a data matrix with the values of K variables $X_1 \dots X_K$ (numerical or not) observed over a set $\mathcal{I} = \{1, \dots n\}$ of individuals. Variables are in columns, while individuals in rows. Cells contain the value (x_{ik}), taken by individual $i \in \mathcal{I}$ for variable X_k, $(k = 1 : K)$. The metrics involved in this research are taken from the literature:

- Gower 71 [16]. Gower works were the first on the direction of defining similarities, and afterwards distances, for spaces where numerical and qualitative variables coexists. Briefly, Gower's metrics combines a normalized distance on the absolute values (first order normalized Minkowsky metrics) for the numerical variables with equality or not for qualitative, taking into account missing data.

- Gowda & Diday 91 [15]. It is built with three components called *position, span and content*:

  $D_k(i, i') = D_p(i, i') + D_s(i, i') + D_c(i, i')$. In fact, D_p is the same component for numerical variables as Gower, and for qualitative ones it is null, while D_s, D_c are considering the number of objects in each modality as well as their intersections. Also suitable for other types of data, not considered in this research.

- Gibert 91. Introduced in [11, 6]: $d^2_{(\alpha,\beta)}(i, i') = \alpha d^2_{\zeta}(i, i') + \beta d^2_{\mathcal{Q}}(i, i')$, being $(d^2_{\zeta}(i, i'))$ the normalized euclidean metrics for numerical variables, and $(d^2_{\mathcal{Q}}(i, i'))$ the χ^2 metrics[2] for qualitative. Actually, it is a family of metrics indexed on $(\alpha, \beta) \in [0, 1]^2$; there is a proposal for weighting on the basis of dimensionality of both spaces and on a normalizing factor [14]. Successfully applied to several real *ISD* [7], [13], [9], [14], [10].

- Ichino & Yaguchi 94 [17]. A generalization of Minkowsky metrics based on a new formal model supporting a single expression for all variables: $\Phi(x_{ik}, x_{i'k}) = |x_{ik} \oplus x_{i'k}| - |x_{ik} \otimes x_{i'k}| + \gamma(2|x_{ik} \otimes x_{i'k}| - |x_{ik}| - |x_{i'k}|)$ on the basis of the operators $\oplus, \otimes$ defined in the model, $\gamma \in [0, 0.5]$. For numerical variables, definition coincides with Minkowsky metrics; for qualitative ones, it only distinguishes equality or not.

- Ralambondrainy 95 [20]. As a matter of fact, it is defined in the same way as Gibert's metrics, but the proposed coefficients for weighing the two components of the metrics are calculated on the basis of previous

[2] This means that difference between qualitative values refers the quantity of information involved [2].

works of the same author[21], using a much more formal paradigm, taking into account the inertia of the groups or the norm of operators.

Technical details or complete definition of those metrics is no possible here, owing to space limitations. Several references are provided in order not to enter in deeply presentation. In this paper they will all of them be called mixed metrics for short, although some are only dissimilarity coefficients.

3 The application

The behavior of different mixed metrics in the clustering will be teste on a real heterogeneous data set, which is presented in this section. Data relates to the 163 municipalities of the metropolitan area of Barcelona, the capital of Catalonia Autonomous Community (in Spain), refers to their growing situation taking into account the building of new houses of different types. It also includes information about the composition of local government. Data is provided by the most recent updates (1999) of the Data Bases from National Statistical Institute (INE), National Statistical Institute of Catalonia (IDESCAT), National Institute of Public Health (INSS)...

The municipalities are described by 16 variables of different kinds relative to their constructive status, together with other general characteristics: **i)** *Comarca*[3]: qualitative; it contains the 7 *regions* surrounding Barcelona. **ii)** *Sector*: Main activity sector of the municipality; qualitative, with three modalities: Industry, Services, Construction. **iii)** *Politic trend*: The party which got most votes in the last local elections (1999); qualitative: CIU, PSC, PP, IC, ERC, Others. **iv)** *Population*: Number of inhabitants. Next variables refers the number of houses started or finished *(ACA,* **v***)* along 1999 (given per 100 inhabitants to be comparable), considering two criteria: the type of house (*Council houses (PO,* **vi, vii***)*, **viii-x:***Isolated, Semidetached, Multifamilial*), and its size (from *Very small,* **xi** to *Very Big,* **xvi**).

This data set contains heterogeneous variables and the use of mixed metrics for clustering is required in order to extract knowledge about the behavior of Barcelona metropolitan area in terms of constructive growing.

Methods and software For this work a single clustering algorithm is used: a hierarchical reciprocal neighbors with Ward criteria [18]. Although it is known that for huge data matrices (usual in *KDD*) cheap clustering algorithms (like partition methods, of linear cost), perform better, for the present application it is still suitable a hierarchical method, of quadratic cost, which do not require to input the number of desired classes, since the goal of the paper is to study the impact of the metrics, rather than the performance of the algorithm itself. Later, other linear algorithms may be studied.

The clustering was done using the software *KLASS* [5], [11], specially designed for *KDD* in *ISD*, although *SODAS* is another software that allow

[3] It is the minimum administrative unit in Catalonia (little than a province).

working with other clustering methods and some of the metrics used here. Originally implemented in LISP[4], *KLASS* provides, among others, tools for descriptive data analysis, clustering and *reporting*, offering a friendly graphical interface. It allows clustering with heterogeneous data matrices using any of the metrics introduced before, among others. It may graphically represent the resulting dendrogram and it can recommend the final number of classes using an heuristic based on maximum distinction and homogeneity of classes.

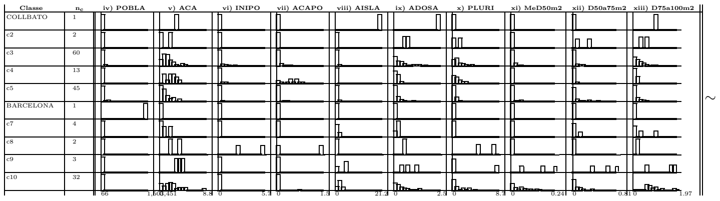
After that, *interpretation* of the results is required. When no previous knowledge on the existing profiles is available, like in this case, it seems reasonable to determine good behavior on the basis of the *meaning* of the resulting classes. It is known that this process is quite difficult and the software should provide tools helping the expert to *understand* which clusters were formed and *why* as easily as possible, although finally, the interpretation itself must be done by the expert in a non-systematic way.

Class-pannel graph *KLASS* is providing different interpretation-oriented tools; among them, tools for descriptive analysis of the resulting classes. In this paper *class-pannel graph* is presented (fig.1): given a partition $\mathcal{P}$ of the set of objects $\mathcal{I}$, it automatically displays at once a panoramic representation of conditional distributions of any type of variables per classes. The user must provide a variables selection among those used for clustering, or other ones, just used as illustrative variables [18] to enrich interpretation of classes. It is remarkable that, like most of the outputs of the system, especially the graphical ones, the *class-pannel graph* is directly generated as font LaTeX code, which can be easily incorporated to reports or other documents. Immediately after generating the LaTeX file, *KLASS* automatically compiles and visualizes it. Combined with other provided results, presented in previous works [8], as well as with automatic generation of final LaTeX reports, also available, *class-pannel graph* supports the later interpretation process made by the expert. Using this representation, it is quite easy to identify class particularities. On our experience, it use to be long and tedious to obtain such a quick overview with standard statistical packages. *KLASS* generates it at once and, if preferred, box-plots may also be chosen.

3.1 Results

Using *KLASS*, the municipalities were clustered using the 5 metrics. Upon the resulting dendrogram, and *KLASS* recommendations, the number of classes was *a posteriori* decided, as usual in hierarchical clustering. As a result, 5 classifications were obtained all between 8 and 11 classes (see table 1): P_{Gw} was obtained with Gower's metrics, P_{GD} with Gowda& Diday's, P_{Gb}

[4] Although LISP provides an excellent support for qualitative variables management, limitations on portability motivated development of a new version in JAVA. At present half of the system is ready in this new version.



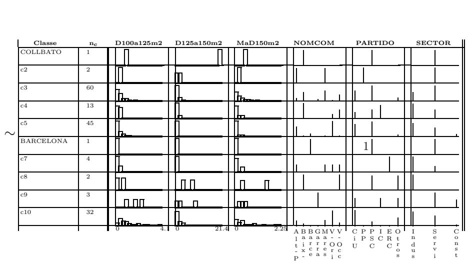


Figure 1: Class-pannel graph for P_{Gb}.

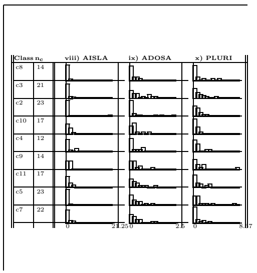


Figure 2: Part of the class-pannel graph of P_I.

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12
P_{Gw}	10	17	19	21	22		15	17	2			
P_{GD}	11	30	25	10	23		33	11	2	10		
P_{Gb}	1	2	60	13	45	1	4	2	3	32		
P_I		23	21	12	23		22	14	14	17	17	
P_R	1		64		71	1		2	10	10		4

Table 1: Classes sizes obtained with the different metrics.

with Gibert's, P_I with Ichino& Yaguchi, P_R with Ralambondrayni. For every one, descriptive analysis of classes was done, mainly using basic summary statistics local to classes as well as the above presented *class-pannel graph*. Figure 1, shows the *class-pannel graph* provided by *KLASS* upon P_{Gb}.

Clustering is expected to provide the profiles of cities surrounding Barcelona in terms of constructive growing §3. Upon the 5 class-pannel graphs of $P_{Gw}, P_{GD}, P_{Gb}, P_I, P_R$ it can be said that two behaviors are identified. On the one hand, P_{Gb} and P_R; on the other one, the remaining clusterings.

Both P_{Gb} and P_R identify Barcelona (C6) as a single cluster, as usual in any catalan study including it, since it concentrates the 25% of global catalan population by itself. Also, some clear groups are identified, as cities with extremely high construction rates (Collbató, c1) or classes with towns sharing the same political trend (PP in c2, ERC in c7), or classes with some variable with different distribution from the other classes (as c4 in ACAPO). Results for P_R are not identical, but the global idea is very close.

For P_{Gw}, P_{GD} and P_I, in general, means of numerical variables in different classes show quite similar values, which make difficult to distinguish one class from the others. Besides, distributions of variables seems to be really similar in all the classes (a sample is shown in fig. 2). However, for P_{Gw} and P_{GD} it is still possible to distinguish two groups of classes upon the qualitative variable *Sector*: one group of classes containing cities of Services and the other with Industry. Unfortunately, in spite of that, it is difficult to understand why those two main groups were subdivided in several classes in both partitions.

4 Conclusions and future work

As in previous works, this research shows that changing metrics produces real and relevant effects on clustering results. It is then important to know when different metrics have better behavior for recognizing real classes.

For this particular application, partitions P_{Gb} and P_R contain more interpretable classes than the others. For the remaining metrics, distribution of all the variables are very similar for all the classes (fig. 2), which makes really difficult to guess which was the underlying clustering criteria.

It has to be taken into account that Gibert's and Ralambondrayni's metrics has two common characteristics that may explain this behavior: **i)** they use a second order euclidean metrics on numerical variables while the others are based on the differences absolute value; we presume that the use of absolute value reduces outliers identification and, in consequence, it might increase variability within classes, so producing decrease on distinction between classes. **ii)** quantity of information is related with the use of χ^2 metrics in clustering [2], which is the second component of Gibert and Ralambondrayni proposals; this may also explain that P_{Gb} and P_R capture a lot of information, producing more interpretable classes; in fact this property of χ^2 metrics justifies its use in multivariate analysis of qualitative variables; this is a preliminary result that, of course, requires to go in depth.

Next step is to complete the experiment in order to consolidate the knowledge about metrics behavior, using other real and simulated data sets previously studied with Gibert's metrics, as well as some huge real datasets involved with *KDD* applications, or even other very quick clustering algorithms with good performance in huge datasets.

In the future, extension of the experiment to the use of *clustering based on rules* [9] (*KLASS* was actually developed to implement it), which takes into account *a priori* expert knowledge for biassing clustering and guarantees semantic meaning of final classes, will be also faced, in order to improve some clustering limitations in the context of ISD.

The process followed by experts for interpreting the classes upon the *class-pannel graph* and other elements generated by *KLASS* [8] is being considered for designing an automatic method for finding characteristic values of the classes. At present, some preliminary results are already presented in this line[13]. Later, the consolidation of the discovered knowledge — using an automatic rules generation approach [13]— will be faced. This is useful for later predictive goals; for this particular, prototype generation is also important. Our goal is providing human interpretable class descriptions.

References

[1] Anderberg M.R. (1973). *Cluster analysis for aplications.* Academic Press.

[2] Benzecri J.P. (1980). *L'analyse des données.* Dunod, Paris.

[3] Diday E. et al. (1984). *Rapport 289 Centre de Rocquencourt.* INRIA.

[4] Fayyad U. et al. (1996). *Advances in KD and DM.* R. AAAI/MIT.

[5] Gibert K. (1994). Ph. D. thesis. Dep. Statistics and Op. Res., UPC.

[6] Gibert K. (1997). **Weighing quantitative and. . . .** MATHWARE **10** (4), 251 – 266.

[7] Gibert K., Annicchiarico R. et al. (2003). *ITI*, Croatia, 2003, 181 – 186.

[8] Gibert K., Aluja T., Cortés U. (1998). LNAI, Springer-Verlag **1510**, 83 – 92.

[9] Gibert K., Cortés U. (1998). *Computación y Sistemas.* **1** (4), 213 – 227.

[10] Gibert K., Cortés U. (1994). LNStats, Springer-Verlag, NY **89**, 351 – 360.

[11] Gibert K., Cortés U. (1992). *KLASS:. . . .* In: IBERAMIA-92, proc., Noriega, Méx, 483 – 497.

[12] Gibert, K., Nonell R. (2003). LNCS, Springer, Berlin **2905**, 464 – 471.

[13] Gibert K., Rodríguez I. (2000). ECAI Works.BESAI, Berlin**9**, 1 – 10. Berlin, 2000

[14] Gibert K., Sonicki Z. (1999). JASMDA, Wiley **15** (4), 319 – 324.

[15] Gowda K.C., Diday E. (1992). IEEE Trans. SMC. **22** (2), 368 – 378.

[16] Gower J.C. *A general coefficient for similarity.* Biometrics **27**, 857 – 872.

[17] Ichino M., Yaguchi H. (1994). IEEE Trans. SMC **22** (2), 146 – 153.

[18] Lebart L. et al. (1982). *Traitement des données statistiques.* DUNOD.

[19] Nakhaeizadeh G. (1996). *Classification as. . . .* In IFCS'96 Proc., Kobe, 17 – 20.

[20] Ralambondrainy H. (1995). *A conceptual...* Pat. Rec. Let. **16**, 1147 – 1157.

[21] Ralambondrainy H. (1998). *A clustering method for nominal...* Elsevier.

[22] Shortlife E.H. (1976). *MYCIN: A rule-based. . . .* Ph. D. Standford.

Acknowledgement: Project *CICYT'2000* partially financed this research.

Address: Dep. of Statistics and Operation Research, Universitat Politècnica de Catalunya, C. Pau Gargallo 5, Barcelona 08028, SPAIN.

E-mail: `karina@eio.upc.es`

NEURAL NETWORK SIEVE BOOTSTRAP FOR NONLINEAR TIME SERIES

F. Giordano, M. La Rocca and C. Perna

Key words: Neural networks, bootstrap, Monte Carlo.

COMPSTAT 2004 section: Resampling techniques.

Abstract: In this paper a sieve bootstrap scheme, the Neural Network Sieve bootstrap, for nonlinear time series is proposed. The approach, which is non parametric in its spirit, does not have the problems of other nonparametric bootstrap techniques such as the blockwise schemes. The procedure performs similarly to the AR-Sieve bootstrap for linear processes while it outperforms the AR-Sieve and the moving block bootstrap for nonlinear processes, both in terms of bias and variability.

1 Introduction

Bootstrap techniques are powerful nonparametric methods for estimating the distribution of a given statistic. The method in its classical form was designed for application to samples of independent data. Extensions to dependent data are not straightforward and modifications of the original procedure are needed in order to preserve the dependence structure of the original data in the bootstrap samples [13]. When dealing with stationary time series, two different classes of bootstrap methods have been proposed. One is a model based approach where the dependence structure of the series is modeled explicitly and the bootstrap sample is drawn from the fitted model. Obviously, these procedures are sensitive to model misspecification, in which case they lead to bootstrap estimators which are not consistent.

Alternatively, nonparametric purely model-free bootstrap schemes for stationary observations can be used. They are based on resampled overlapping blocks of consecutive observations, with the block length growing with the sample size at a proper rate. They enjoy the property of being robust against misspecified models and are valid under weak conditions. However, the resampled series exhibits spurious features which are caused by randomly joining selected blocks. As a consequence, the asymptotic variance-covariance matrices of the estimators based on the original series and those based on the bootstrap series are different and a modification of the original scheme is needed. A possible solution is the matched moving-block bootstrap [5], based on a quite complex procedure which resamples the blocks according to a Markov chain whose transitions depend on the data. A further difficulty is that the bootstrap sample is not (conditionally) stationary. This can be overcome by taking blocks of random length, as proposed by Politis and Romano [11] but, a recent study of Lahiri [8] shows that this approach is much less efficient than the original one.

In this paper we propose a novel bootstrap scheme, the Neural Network Sieve bootstrap (NN-Sieve), for nonlinear time series which is is non parametric in its spirit but it does not have the problems of blockwise schemes. The technique is based on the idea of sieve approximation where an infinite dimensional nonparametric model is approximated by a sequence of finite-dimensional parametric models. This approach has been first proposed by Kreiss [7] and extensively studied by Bühlmann [1], [2]. It uses a sequence of approximating autoregressive models, $AR(p)$, for the data generating process with order p that increases as a function of the sample size. This bootstrap scheme is based on a parametric model but it is nonparametric in its spirit, being model free within the class of linear processes. For nonlinear processes the AR-Sieve is not consistent but the key idea can be preserved by considering a different sequence of approximating models. Here we propose to use a class of feedforward neural networks as approximating models since they can be considered "universal approximators" for a general class of nonlinear functions.

The paper is organized as follows. Section 2 introduces the general idea of the sieve bootstrap with a neural network model and shows its application as an inference tool. The results of a Monte Carlo study are reported in section 3. Some remarks close the paper.

2 The neural network sieve resampling scheme

Let $\{Y_t, t \in \mathbb{Z}\}$ a real valued stationary stochastic process, modeled as

$$Y_t = g\left(Y_{t-1}, \ldots, Y_{t-d}\right) + \epsilon_t$$

where ϵ_t is a sequence of *iid* random variables with zero mean and finite variance σ^2. We also suppose that ϵ_t is independent of $\{Y_{t-d}, d \geq 1\}$, the distribution function of ϵ_t is absolutely continuous with respect to a Lebesgue measure and its density function is continuous and positive on its support. The function $g(\cdot)$ is a nonlinear, unknown function satisfying some regularity conditions as in [14]. Under these assumptions, the Markov chain associated to the process Y_t is geometrically ergodic and Y_t is ϕ-mixing with geometrically decreasing mixing coefficients [12].

The process $\{Y_t, t \in \mathbb{Z}\}$ can be approximated by a family of parametric models $\{M_r,\ r \in \mathbb{N}\}$, equipped with a model selection rule, such that $\bigcup_{r=1}^{\infty} M_r$ contains in some sense the original process. Hence, a key issue is the selection of a proper model family.

For general stationary categorical processes, Bühlmann [3] proposes the Variable Length Markov Chain, a flexible class of Markov models that allows for parsimonious structure.

For linear processes a straightforward choice is the class of $AR(p)$ models with finite unknown p, assuming that some consistent estimator is available [1]. The latter approach performs better than other bootstrap techniques if the data generating process is linear, representable as an $AR(\infty)$ process.

The method is easy to implement, due to the simplicity of fitting an AR model.

If the model is nonlinear, the AR-sieve bootstrap is not asymptotically consistent and its success is related to the closeness of the underlying process to the $AR(\infty)$ representation. Bühlmann [1], [2] shows by simulation that for EXPAR(2) models and for some classes of SETAR models the AR-sieve bootstrap estimation exhibits a bias which does not decrease with increasing sample size. Therefore, an alternative approach for nonlinear data generating processes is needed. We propose to use the class of one layer feedforward neural network models, $NN(d,r)$ defined as

$$f\left(y_{t-1}\dots,y_{t-d};\eta\right)=\sum_{k=1}^{r}c_k\phi\left(\sum_{j=1}^{d}a_{kj}y_{t-j}+a_{k0}\right)+c_0 \tag{1}$$

where d is the number of input neurons (the order of the autoregression), r is the hidden layer size, a_{kj} is the weight of the connection between the j-th input neuron and the k-th neuron in the hidden level; c_k, $k=1,\dots,r$ is the weight of the link between the k-th neuron in the hidden layer and the output; a_k and c_0 are respectively the bias term of the hidden neurons and of the output; $\phi(\cdot)$ is the activation function of the hidden layer. We define $\eta=(c_0,c_1,\dots,c_r,\mathbf{a}_1',\mathbf{a}_2',\dots,\mathbf{a}_r')$ where $\mathbf{a}_i'=(a_{i0},a_{i1},\dots,a_{id})$ and we suppose that $\eta\in\mathbb{R}^{r(d+2)+1}$.

As usual in neural network applications, we will assume a sigmoidal activation function such as the logistic or the hyperbolic tangent function. In this case, the hypotheses on the function $g(\cdot)$ and on the process Y_t guarantee that single hidden layer neural networks are “universal approximators” [10], in that they can arbitrarily closely approximate, in an appropriate metric, to the unknown function $g(\cdot)$. Moreover, the stochastic process Y_t defined as

$$Y_t=f\left(Y_{t-1},\dots,Y_{t-d};\eta\right)+\epsilon_t$$

shares the same probabilistic structure as the original process, being stationary and ϕ-mixing with geometrically decreasing mixing coefficients.

Therefore, in the NN-sieve bootstrap scheme, we consider feedforward neural networks with fixed number of neurons in the input layer so that the class of models M_r is defined as $NN(d^*,r)$; we assume that the finite unknown hidden layer size r can be estimated consistently.

Given the time series $\{Y_1,\dots,Y_T\}$, let θ a finite dimensional parameter of interest and $\hat{\theta}_T=h\left(Y_1,\dots,Y_T\right)$ a scalar-, vector- or curve-valued estimator, which is a measurable function of the data. Inference on θ can be gained by using the NN-Sieve bootstrap approach which runs as follows.

- Select the input size d^* and estimate the hidden layer size $\hat{r}$ by using some kind of order selection criteria or some inferential procedures [9].

- Estimate the neural network model $NN(d^*, \hat{r})$ by minimizing a loss function such as the mean square error.
- Define the empirical distribution function of the centered residuals $\hat{\varepsilon}_t$

  $$\hat{F}(x) = (T - d^*)^{-1} \sum_{t=d^*+1}^{T} I(\hat{\varepsilon}_t \leq x)$$

  where $I(\cdot)$ denotes the indicator function.
- Draw a resample ε_t^* of *iid* observations from $\hat{F}$ and define

  $$Y_t^* = f\left(Y_{t-1}^*, \ldots, Y_{t-d}^*; \hat{\eta}\right) + \varepsilon_t^*$$

  with the first d observations fixed to the mean value of Y_t and $t = 1, \ldots, T + n$. The first n observations are discarded in order to make negligible the effect of starting values.
- Compute $\hat{\theta}_T^* = h(Y_1^*, \ldots, Y_T^*)$ on the resampled series $Y_1^*, \ldots, Y_T^*$.
- Replicate B times the procedure and obtain B bootstrap replicates of the statistic of interest $\{\hat{\theta}_{T,b}^*, b = 1, \ldots, B\}$. The empirical distribution function

  $$\hat{F}^*(x) = B^{-1} \sum_{b=1}^{B} I\left(\hat{\theta}_{T,b}^* \leq x\right)$$

  can be used to approximate the unknown sampling distribution of the estimator $\hat{\theta}_T$ (or at least of its standard error).

3 Simulation results

To investigate the performance of the procedure in finite samples, a simulation experiment has been implemented to study and compare the proposed scheme with the AR-Sieve bootstrap and the moving block bootstrap (MBB). The most important difficulty with the latter technique is that it generates series that are less dependent than the original data and this can lead to very bad resampling approximation. The “whitening” effect can be removed by resampling block of blocks [6] but it requires a quite complex modification of the original resampling mechanism. The NN-Sieve, instead, still shares the same logic and simplicity as the standard model-based bootstrap schemes.

We first considered an AR(1) model with gaussian innovations specified as follows:

(M1) $Y_t = -0.8Y_{t-1} + \epsilon_t$

where $\epsilon_t \sim N(0, 1)$.

Since the AR-Sieve bootstrap relies on linear approximation, we use it as a benchmark for our proposal. In this case, the AR-Sieve approach is expected to work satisfactorily.

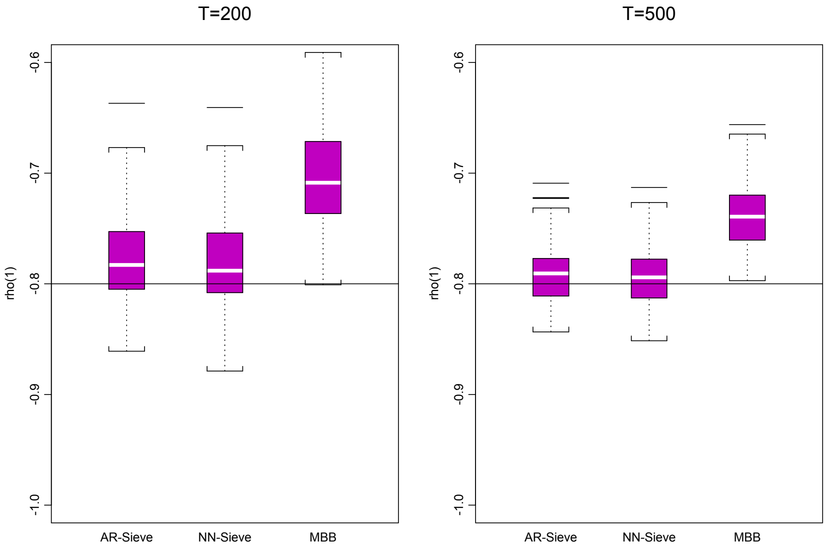


Figure 1: Distribution of the sample autocorrelation at lag 1 on the bootstrap replicates. The horizontal line is the “true” autocorrelation at lag 1.

To study the “whitening effect”, we generated 100 Monte Carlo series from model M1 and for each series we generated 100 bootstrap replicates under different resampling plans. In figure 1 we reported the empirical distribution of the bootstrap estimates of the autocorrelation at lag 1, when using the three alternative resampling schemes. For the MBB scheme, the optimal block length has been determined by using the procedure proposed by Bühlmann and Künsch [4]. Clearly, the MBB shows the “whitening effects” whereas the AR-Sieve and the NN-Sieve generate resampled series with correlation structure similar to the original data.

To study the performance of the proposed resampling scheme for nonlinear time series, we considered three alternative models specified as follows:

(M2) $Y_t = 0.3\epsilon_{t-1}Y_{t-2} + \epsilon_t$

(M3) $Y_t = (1.5 - 0.9Y_{t-1})\, I_t + (-0.4 - 0.6Y_{t-1})\,(1 - I_t) + \epsilon_t$

(M4) $Y_t = \left(0.5 + 0.9\exp\left(-Y_{t-1}^2\right)\right) Y_{t-1} - \left(0.8 - 1.8\exp\left(-Y_{t-1}^2\right)\right) Y_{t-2} + \epsilon_t$

where I_t is an indicator function defined as $I_t = 1$ if $Y_{t-1} \leq 0$, $I_t = 0$ otherwise.

Model M2 has a bilinear structure; it is nonlinear but it can be confused with a white noise if we focus only on the first two moments. M3 is

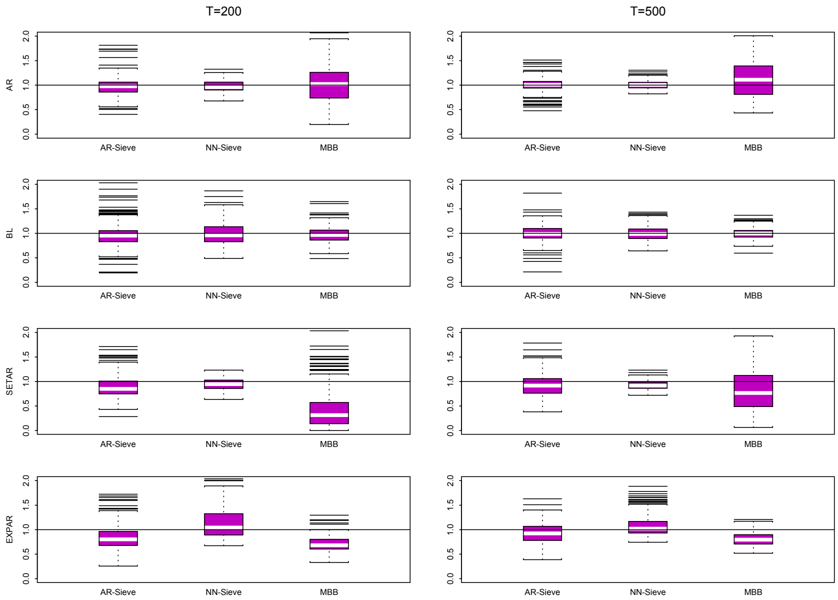


Figure 2: Distribution of the estimated bootstrap variance of the estimator of the mean. The horizontal line is the “true” variance value.

a SETAR(2;1,1) model representing a nonlinear process with non gaussian, strongly bimodal, marginal distribution; M4 is an EXPAR(2) model and, like M3, cannot be represented as a linear model. As pointed out by Bühlmann [2], in these cases the AR-Sieve bootstrap has a bias which does not decrease with increasing sample size and, as a consequence, the procedure is expected to behave poorly.

As parameter of interest θ, we consider the mean and the median, a nonlinear functional, of the underlying generating process $\{Y_t, t \in \mathbb{Z}\}$. Here we focus on the estimation of the bootstrap variance of $\hat{R}_T = \left[\hat{\theta}_T - \mathbb{E}\left(\hat{\theta}_T\right)\right]/\sigma_T$ where σ_T is the true standard error of $\hat{\theta}_T$. In the simulation study the unknown quantity σ_T has been estimated by 10,000 Monte Carlo runs.

In the Monte Carlo experiment, for each model we simulated $N = 200$ series of length $T = 200$ and $T = 500$. We generated $B = 999$ bootstrap replicates $Y_1^*, \ldots, Y_T^*$ with the AR-Sieve, the NN-Sieve and the MBB method. For the AR-Sieve the order p of the AR model has been chosen by using the Akaike information criterion. For the NN-Sieve the hidden layer size has been fixed again by using the Akaike information criterion. For each simulated series we estimated the statistic $\hat{R}_T^* = \left[\hat{\theta}_T^* - \mathbb{E}\left(\hat{\theta}_T^*\right)\right]/\sigma_T$ for each bootstrap replicate and for each method.

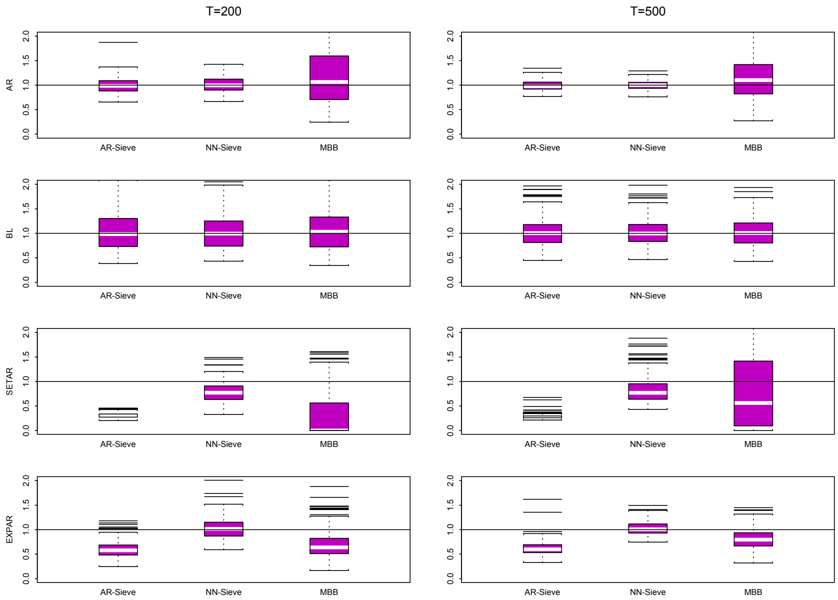


Figure 3: Distribution of the estimated bootstrap variance of the estimator of the median. The horizontal line is the "true" variance value.

In figure 2 we reported the distribution of the estimated bootstrap variance for the mean of the data generating process. The AR-Sieve and the NN-Sieve perform almost equivalently while the MBB shows a much greater variability. However, when the data generating process is nonlinear, the NN-Sieve bootstrap outperforms the AR-Sieve in all the cases considered. Moreover, the novel procedure seems to give better results then the MBB technique. It is also clear that when the sample size increases, both the bias and the variability reduce, as expected for consistent estimators.

This behavior is even more clear, and the performance of the NN-Sieve is even better, when considering a nonlinear functional, such as the median of the data generating process (see figure 3).

4 Concluding remarks

In this paper we propose a nonparametric bootstrap scheme, the Neural Network Sieve bootstrap, which does not have the problems of the blockwise schemes. The proposed NN-Sieve procedure yields satisfactory results in a simulation study for finite sample sizes. It performs similarly to the AR-Sieve bootstrap for linear processes while it outperforms the AR-Sieve and the MBB for nonlinear processes, both in terms of bias and variability.

References

[1] Bühlmann P. (1997). *Sieve bootstrap for time series.* Bernoulli **3**, 123–148.

[2] Bühlmann P. (2002). *Bootstraps for time series.* Statistical Science **17**, 52–72.

[3] Bühlmann P. (2002). *Sieve bootstrap with variable-length Markov chains for stationary categorical time series.* Journal of the American Statistical Association **97**, 443–471.

[4] Bühlmann P., Künsch H.R. (1999). *Block length selection in the bootstrap fro time series.* Computational Statistics and Data Analysis **31**, 295–310.

[5] Carlstein E., Do K.A., Hall P., Hesterberg T., Kunsch H.R. (1998). *Matched block bootstrap for dependent data.* Bernoulli **4**, 305–328.

[6] Davison A. C., Hinkley D. V. (1997). *Bootstrap methods and their application.* Cambridge University Press.

[7] Kreiss J.P. (1992). *Bootstrap procedures for AR*(∞) *processes.* In Jackel K.H., Rothe G., Sendler W. (eds.), Bootstrapping and Related Techniques, Springer, Heidelberg, 107–113.

[8] Lahiri S.N. (1999). *Theoretical comparisons of block bootstrap methods.* The Annals of Statistics **27**, 386–404.

[9] La Rocca M., Perna C. (2004). *Variable selection in neural network regression models with dependent data: a subsampling approach.* To be published on Computational Statistics and Data Analysis.

[10] Leshno M., Lin V., Pinkus A., Schocken S. (1993). *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function.* Neural Networks **6**, 861–867.

[11] Politis D.N., Romano J.P. (1994). *The stationary bootstrap.* JASA **89**, 1303–1313.

[12] Rosenblatt M. (1971). *Markov processes. Structure and asymptotic behavior.* Springer.

[13] Shao J., Tu D. (1995). *The Jackknife and the bootstrap.* Springer.

[14] Tong H. (1990). *Non-linear time series a dynamical system approach.* Clarendon Press, Oxford.

Address: F. Giordano, M. La Rocca, C. Perna, Dept. of Economics and Statistics, University of Salerno, via Ponte don Melillo, 84084 Fisciano (SA) – Italy

E-mail: `[giordano,larocca,perna]@unisa.it`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

INDIRECT METHODS OF IMPUTATION IN SAMPLE SURVEYS

S. González, Maria M. Rueda, A. Arcos, Yolanda Román and M.D. Martínez

Key words: Horvitz-Thompson estimator, auxiliary information, missing data.

COMPSTAT 2004 section: Data imputation.

Abstract: One of the most difficult problems confronting investigators who analyze data from surveys is how to treat missing data. Many statistical procedures cannot be used immediately if any values are missing. Imputation of missing data before starting statistical analysis is then necessary. This paper proposes imputation methods of the mean based on indirect estimators of available cases. A complete simulation study was performed to test the proposed techniques.

1 Introduction

Surveys are a common method of data collection in economics and social sciences, but they often suffer from the problem of nonresponse. The reasons for this vary; the respondent may not be present at the time of survey, or a certain class of respondent may not form part of the survey at all. Some of these factors affect the quality of the data. One obvious consequence of nonresponse is that the actual sample size is less than the planned one, which can produce biases in estimations and growth in sampling variance.

The problem of missing data can be addressed by various methods during the stages of data collection and processing. The aim of all these methods is to obtain a precise and complete data set. Nevertheless, it is still possible for errors and losses of entries to occur even once the data has been collected and filtered.

An initial option is to carry out a complete case analysis. Methods based on completely recorded units create a rectangular data set by discarding parts of the data. This is the simplest and most common approach to nonresponse. Among the advantages of this kind of analysis are its simplicity and the fact that different univariate statistics can be compared. However, it also has numerous disadvantages. Little and Rubin [5] pointed out the problems of methods that ignore incomplete observations. While these methods may be satisfactory when the percentage of incomplete cases is low, in general terms they lead to biased estimations, since they assume that the loss of data takes place in a completely random way. King et al. [3] illustrate how methods of complete cases are prone to serious errors. Thus, this practice may introduce bias into the estimate and increase sampling variance due to a reduction in sample size, see, e.g, Brik and Kalton [2], Schafer [11].

As a second option, we could try to improve the precision of the estimators by including all cases available for their calculation, see, e.g, Toutenburg and Srivastava [12] and Rueda and González [10] for an interesting account.

Alternatively, an imputation method to find substitutes for missing observations could be employed. By treating these imputed values as true observations, statistical analysis may be carried out using the standard procedures developed for data without any missing observations. In this study we propose a method for imputation of the mean based on indirect estimators of available cases. The procedure consists of making use of the information available from incomplete observations, and thus improving the precision of the estimator. When a complete data set is available, a simulation study is performed to test the functioning of the proposed techniques.

2 Indirect estimators based on available cases

Consider a population of N units from which a random sample, s, of fixed size, n is drawn according to a sample design $d = (S_d, P_d)$, with first order inclusion probabilities π_i. For this sample we observe the values of two variables, (y_i, x_i), $i = 1, \ldots, n$, for the estimation of the population mean, $\bar{Y}$.

It is assumed that a set of $(n - p - q)$ complete observations of selected units in the sample are available. In addition to these, observations of the x characteristic of p units in the sample are available but the corresponding observations for the y characteristic are missing. Similarly, we have a set of q observations of the y characteristic in the sample but the associated values of the x characteristic are missing. Furthermore, p and q are assumed to be integer numbers verifying $1 \leq p, q < \frac{n}{2} - 1$. For the sake of simplicity, the unit of the sample s is separated into three disjoint sets $s_1 = \{i \in s/x_i, y_i$ are available$\}$, $s_2 = \{i \ \in \ s/x_i$ are available, but y_i is not$\}$ and $s_3 = \{i \in s/y_i$ are available, but x_i is not$\}$.

The Horvitz-Thompson estimators based on these samples are:

$$\hat{\bar{y}}^1_{HT} = \frac{1}{N} \sum_{i \in s_1} \frac{y_i}{\pi_i}, \ \ \hat{\bar{y}}^3_{HT} = \frac{1}{N} \sum_{i \in s_3} \frac{y_i}{\pi_i}, \ \ \hat{\bar{x}}^1_{HT} = \frac{1}{N} \sum_{i \in s_1} \frac{x_i}{\pi_i} \ \ \hat{\bar{x}}^2_{HT} = \frac{1}{N} \sum_{i \in s_2} \frac{x_i}{\pi_i} \ \ .$$

The following indirect estimators for the population total based on complete cases can be formulated:

$$\hat{\bar{y}}_{r1} = \frac{\widehat{\bar{y}}^1_{HT}}{\widehat{\bar{x}}^1_{HT}} \bar{X}, \ \ \widehat{\bar{y}}_{d1} = \widehat{\bar{y}}^1_{HT} + (\bar{X} - \widehat{\bar{x}}^1_{HT}), \ \ \hat{\bar{y}}_{Reg1} = \widehat{\bar{y}}^1_{HT} + b(\bar{X} - \widehat{\bar{x}}^1_{HT}) \quad (1)$$

where b may be fixed and either known or unknown. In the latter case, if it is minimized the error obtained, $b = \dfrac{\text{Cov}(x, y)}{\text{Var}(x)}$, must be estimated.

All these estimators discard the information available on incomplete cases. This practice can introduce bias and errors into the estimation, and so the following classes of estimators, incorporating all the available observations, are proposed:

$$\hat{\bar{y}}_{r2} = \frac{\alpha_r \widehat{\bar{y}}^3_{HT} + (1-\alpha_r)\widehat{\bar{y}}^1_{HT}}{\beta_r \widehat{\bar{x}}^2_{HT} + (1-\beta_r)\widehat{\bar{x}}^1_{HT}} \bar{X} \tag{2}$$

$$\widehat{\bar{y}}_{d2} = \alpha_d \widehat{\bar{y}}^1_{HT} + (1-\alpha_d)\widehat{\bar{y}}^3_{HT} + (\bar{X} - (\beta_d \widehat{\bar{x}}^1_{HT} + (1-\beta_d)\widehat{\bar{x}}^2_{HT})) \tag{3}$$

$$\hat{\bar{y}}_{Reg2} = \alpha_{reg}\hat{\bar{y}}^1_{HT} + (1-\alpha_{reg})\hat{\bar{y}}^3_{HT} + b\left[\bar{X} - (\beta_{reg}\hat{\bar{x}}^1_{HT} + (1-\beta_{reg})\hat{\bar{x}}^2_{HT})\right] \ . \tag{4}$$

In the case of the regression estimator, if b is unknown, we can proceed as in the case of no nonresponse. Thus, two possible estimators for b are presented:

$$\hat{b}_1 = \frac{\mathrm{Cov}_{i\in s_1}(x,y)}{\mathrm{Var}_{i\in s_1}(x)} \text{ and } \hat{b}_2 = \frac{\mathrm{Cov}_{i\in s_1}(x,y)}{\mathrm{Var}_{i\in s_1\cup s_2}(x)} \tag{5}$$

which will generate two regression estimators $\hat{\bar{y}}_{Reg21}$ and $\hat{\bar{y}}_{Reg22}$.

The estimators with subindex 1 are the traditional ratio, difference and regression estimators, which are based on complete observations and ignore the incomplete pairs of observations. We propose the estimators with subindex 2 which incorporate all the available observations.

The following step is to look for the estimators with the best behaviour among the proposed classes of estimators. These choices are made in order to minimize the estimation error. The expressions of the mean squared errors of the estimators are easily obtained; by minimizing these errors, the estimator expressions with minimum error are derived. The optimal coefficients $\alpha_{r_{opt}}$, $\beta_{r_{opt}}$, $\alpha_{d_{opt}}$, $\beta_{d_{opt}}$, $\alpha_{reg_{opt}}$ and $\beta_{reg_{opt}}$ are given by:

$$\begin{aligned}
\alpha_{r_{opt}} &= \frac{-C_r + (E_r B_r - \frac{C_r}{A_r}B_r^2)/(D_r - B_r^2/A_r)}{A_r} \\
\beta_{r_{opt}} &= \frac{-E_r + \frac{C_r}{A_r}B_r}{D_r - B_r^2/A_r} \\
\alpha_{d_{opt}} &= \frac{A_d - \frac{C_d D_d - A_d B_d}{E_d C_d - B_d^2}B_d}{C_d} \\
\beta_{d_{opt}} &= \frac{C_d D_d - A_d B_d}{E_d C_d - B_d^2} \\
\alpha_{reg_{opt}} &= \frac{-C_{reg}}{A_{reg}} - \frac{B_{reg}}{A_{reg}} \frac{B_{reg}C_{reg} - A_{reg}E_{reg}}{A_{reg}D_{reg} - B_{reg}^2} \\
\beta_{reg_{opt}} &= \frac{B_{reg}C_{reg} - A_{reg}E_{reg}}{A_{reg}D_{reg} - B_{reg}^2}
\end{aligned}$$

where $R = \overline{Y}/\overline{X}$ and

$$\begin{array}{lcl}
A_r & = & 2R^2 \operatorname{Var}(\widehat{x}^2_{HT}) + 2R^2 \operatorname{Var}(\widehat{x}^1_{HT}) - 4R^2 \operatorname{Cov}(\widehat{x}^2_{HT}, \widehat{x}^1_{HT}) \\
B_r & = & -2R \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) + 2R \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^1_{HT}) + \\
 & & 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) - 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) \\
C_r & = & -2R^2 \operatorname{Var}(\widehat{x}^1_{HT}) + 2R^2 \operatorname{Cov}(\widehat{x}^2_{HT}, \widehat{x}^1_{HT}) - \\
 & & 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) + 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) \\
D_r & = & 2 \operatorname{Var}(\widehat{y}^3_{HT}) + 2 \operatorname{Var}(\widehat{y}^1_{HT}) - 4 \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{y}^1_{HT}) \\
E_r & = & -2 \operatorname{Var}(\widehat{y}^1_{HT}) + 2 \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{y}^1_{HT}) - \\
 & & 2R \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^1_{HT}) + 2R \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) \\
A_d & = & \operatorname{Var}(\widehat{y}^3_{HT}) - \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{y}^3_{HT}) + \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) - \\
 & & \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) \\
B_d & = & -\operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^1_{HT}) + \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{x}^2_{HT}) + \\
 & & \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^1_{HT}) - \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) \\
C_d & = & \operatorname{Var}(\widehat{y}^1_{HT}) + \operatorname{Var}(\widehat{y}^3_{HT}) - 2 \operatorname{Cov}(\widehat{y}^1_{HT}, \widehat{y}^3_{HT}) \\
D_d & = & \operatorname{Var}(\widehat{x}^2_{HT}) - \operatorname{Cov}(\widehat{x}^1_{HT}, \widehat{x}^2_{HT}) + \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^1_{HT}) - \\
 & & \operatorname{Cov}(\widehat{y}^3_{HT}, \widehat{x}^2_{HT}) \\
E_d & = & \operatorname{Var}(\widehat{x}^2_{HT}) + \operatorname{Var}(\widehat{x}^1_{HT}) - 2 \operatorname{Cov}(\widehat{x}^1_{HT}, \widehat{x}^2_{HT}) \\
A_{reg} & = & 2 \operatorname{Var}(\hat{y}^1_{HT}) + 2 \operatorname{Var}(\hat{y}^3_{HT}) - 4 \operatorname{Cov}(\hat{y}^1_{HT}, \hat{y}^3_{HT}) \\
B_{reg} & = & 2b \left[-\operatorname{Cov}(\hat{y}^1_{HT}, \hat{x}^1_{HT}) + \operatorname{Cov}(\hat{y}^1_{HT}, \hat{x}^2_{HT}) + \right. \\
 & & \left. \operatorname{Cov}(\hat{y}^3_{HT}, \hat{x}^1_{HT}) - \operatorname{Cov}(\hat{y}^3_{HT}, \hat{x}^2_{HT}) \right] \\
C_{reg} & = & -2 \operatorname{Var}(\hat{y}^3_{HT}) + 2 \operatorname{Cov}(\hat{y}^1_{HT}, \hat{y}^3_{HT}) + \\
 & & 2b \left[-\operatorname{Cov}(\hat{y}^1_{HT}, \hat{x}^2_{HT}) + \operatorname{Cov}(\hat{y}^3_{HT}, \hat{x}^2_{HT}) \right] \\
D_{reg} & = & b^2 \left[2 \operatorname{Var}(\hat{x}^1_{HT}) + 2 \operatorname{Var}(\hat{x}^2_{HT}) - 4 \operatorname{Cov}(\hat{x}^1_{HT}, \hat{x}^2_{HT}) \right] \\
E_{reg} & = & -2b^2 \operatorname{Var}(\hat{x}^2_{HT}) + 2b^2 \operatorname{Cov}(\hat{x}^1_{HT}, \hat{x}^2_{HT}) - \\
 & & 2b \operatorname{Cov}(\hat{y}^3_{HT}, \hat{x}^1_{HT}) + 2b \operatorname{Cov}(\hat{y}^3_{HT}, \hat{x}^2_{HT})
\end{array}$$

Unfortunately, these optimum values depend on theoretical variances and covariances among the Horvitz-Thompson estimators, which are generally unknown. However, they can be estimated when the sample is drawn. Furthermore, these values would be estimated by replication methods. The expressions of these variances and covariances for the case of simple random sampling without replacement and stratified sampling can be seen in Rueda and González [10]. These estimations allow us to obtain approximate values, $\widetilde{\alpha}_r$, $\widetilde{\beta}_r$, $\widetilde{\alpha}_d$, $\widetilde{\beta}_d$, $\widetilde{\alpha}_{reg}$ and $\widetilde{\beta}_{reg}$, and to build the correspondence estimators $\widetilde{\overline{y}}_{r2}$, $\widetilde{\overline{y}}_{d2}$ and $\widetilde{\overline{y}}_{Reg2}$. These estimators do not coincide with the theoretical estimators $\hat{\overline{y}}_{r2}$, $\hat{\overline{y}}_{d2}$ and $\hat{\overline{y}}_{reg2}$ in expressions (2) ,(3) and (4), but, using the results obtained by Randles [7], who derived the asymptotic distribution of estimators with estimated parameters, it could be proved that asymptotically they have the same distribution. For this reason, it is reasonable to assume that their sampling errors will be close to the theoretical ones.

Finally, the usual estimators are included in the proposed classes of estimators, and so the estimators obtained by minimizing the errors in these classes will be better, in terms of precision, than the traditional ones.

3 Imputation methods proposed

When an imputation method is applied, the set of complete data is specified by:

$$z_i = \begin{cases} y_i & \text{si} \quad i \in s_1 \cup s_3 \\ \\ \tilde{y}_i & \text{si} \quad i \in s_2 \end{cases}$$

where $\tilde{y}_i$ is the imputed value, and from these data the necessary estimations can be calculated. Thus, the following estimators, among others, are obtained:

Parameter	**Estimator**
Mean	$\widehat{\bar{y}}_{imp} = \dfrac{1}{N} \sum_{i \in s} \dfrac{z_i}{\pi_i}$
Total	$\hat{y}_{imp} = \sum_{i \in s} \dfrac{z_i}{\pi_i}$
Distribution function	$\hat{F}_{imp}(t) = \dfrac{1}{N} \sum_{i \in s} \dfrac{\Delta(t - z_i)}{\pi_i}$

By using indirect estimation methods, the traditional ratio, difference and regression estimators of the mean can be used as the imputed values. However, if a large proportion of the data is missing, the usual estimators will be based on a relatively small sample and their precision will be reduced correspondingly. We propose, therefore, methods for imputation of the mean in which indirect estimation is applied in the cases of non-response described in the above section. Thus, the following imputation procedures are proposed:

- Procedure based on a ratio estimator: in this situation, specify the complete data set using the estimator $\hat{\bar{y}}_{r2}$ for the estimated value

- Procedure based on a difference estimator: in this situation, specify the complete data set using the estimator $\hat{\bar{y}}_{d2}$ for the estimated value

- Procedure based on a regression estimator with b unknown: in this case, we propose two regression estimators using the two possible estimators of the regression coefficient. Thus we derive two imputation procedures, with $\hat{\bar{y}}_{reg21}$ and $\hat{\bar{y}}_{reg22}$ being the imputed values.

After having used one of these imputation methods and specified the corresponding complete data set, the relevant inferences can be made. The functioning of some of the estimations that could be produced is discussed in Sect. 4 by means of a simulation study.

4 Simulation study

This section describes estimator properties by applying a simulation study.

The FAM1500 population, taken from Fernández and Mayor [4], consists of 1500 families in Andalusia (Spain) The variable of interest, y, denotes family income and the auxiliary x denotes expenditure on food and drink.

The second population was used by Meeden [6]. For the purposes of the simulation, a superpopulation model is considered in which it is assumed that for each i, $y_i = bx_i + u_i e_i$, where $u_i = \sqrt{x_i}$ are known constants, and e_i are independent, identically distributed random variables with zero expectations. In this population, SIM1, the x_i's were a random sample from a gamma distribution with shape parameter twenty and scale parameter one. Then, given x_i the conditional distribution of y_i was normal with mean $1.2x_i$ and variance x_i. This simulated population contained 500 units.

The following algorithm is used for populations with several sample sizes:

- STEP 1: Take a sample of size n according to the procedure of simple random sampling without replacement.
- STEP 2: Choose two random numbers, p and q, verifying $\frac{n}{2} - 1 > p$, $q \geq 1$
- STEP 3: Randomly eliminate from the sample p elements of the auxiliary characteristic and q elements of the study characteristic, and define the subsamples s_1, s_2 and s_3.
- STEP 4: Calculate:

$$\hat{\bar{y}}_{r1},\ \hat{\bar{y}}_{r2},\ \hat{\bar{y}}_{Reg11},\ \hat{\bar{y}}_{Reg21},\ \hat{\bar{y}}_{Reg12},\ \hat{\bar{y}}_{Reg22},\ \hat{\bar{y}}_{d1},\ \hat{\bar{y}}_{d2}$$

- STEP 5: Build the complete data set obtained by inputting missing data for the estimators obtained in step 4.
- STEP 6: Obtain the mean and median estimators based on the complete data set obtained in step 5.
- STEP 7: Use the values obtained in 1000 items for the calculation of the relative efficiency of the estimators obtained in step 6, using the classic mean imputation technique as the basic method.

Results of the application of this algorithm can be seen in Tables 1 and 2.

This study of the Fam1500 population shows that the method of estimating complete cases produces estimations of the mean and of the median that on many occasions do not improve on the accuracy obtained by the classical method for imputation of the mean. On the other hand, the proposed imputation methods always lead to an improvement on the methods based on complete cases and, except for one case for each parameter (for a sample size of 25), always improve on the basic imputation method. This gain in precision is more evident when the parameter to be estimated is the median, and rises to over 13 per cent for a sample size of 100, for all the methods proposed.

For the Sim1 population, the estimations obtained by the proposed imputation methods are always better than the corresponding ones for complete cases, and also than those of the basic method (except for the mean, in one case). We also found that when the median was estimated, the increase in the precision of the estimators was greater, often more than 20 per cent better.

n	25	50	75	100	25	50	75	100
$\tilde{y}_i$	Mean Estimation				Median Estimation			
$\hat{y}_{r1}$	1.017	1.030	1.013	1.013	1.042	1.062	1.049	1.051
$\hat{y}_{r2}$	0.975	0.946	0.963	0.903	0.943	0.878	0.956	0.872
$\hat{y}_{d1}$	1.001	0.999	1.001	1.001	1.001	1.000	1.002	0.999
$\hat{y}_{d2}$	1.001	0.969	0.985	0.911	0.975	0.915	0.935	0.876
$\hat{y}_{reg11}$	1.037	1.016	1.011	0.987	1.102	1.031	1.019	0.985
$\hat{y}_{reg21}$	0.987	0.943	0.965	0.884	0.977	0.855	0.958	0.822
$\hat{y}_{reg12}$	1.071	1.032	1.033	0.985	1.142	1.058	1.046	0.982
$\hat{y}_{reg22}$	0.999	0.948	0.972	0.882	1.007	0.866	0.967	0.822

Table 1: Ratio of the mean squared error of imputation based on $\hat{\tilde{y}}_i$ to the mean squared error of the imputation based on $\bar{y}$, Fam1500 population.

n	25	50	75	100	25	50	75	100
$\tilde{y}_i$	Mean Estimation				Median Estimation			
$\hat{y}_{r1}$	0.976	0.986	0.968	0.987	0.961	0.972	0.939	0.986
$\hat{y}_{r2}$	0.907	0.945	0.936	0.907	0.772	0.791	0.759	0.721
$\hat{y}_{d1}$	1.016	1.016	1.027	1.017	1.043	1.049	1.069	1.047
$\hat{y}_{d2}$	0.921	0.972	0.987	0.960	0.809	0.879	0.878	0.875
$\hat{y}_{reg11}$	0.997	0.999	1.002	1.003	1.022	1.037	1.021	1.010
$\hat{y}_{reg21}$	0.916	0.966	0.943	0.925	0.846	0.848	0.787	0.759
$\hat{y}_{reg12}$	1.019	1.758	0.999	1.019	1.024	1.095	1.021	1.043
$\hat{y}_{reg22}$	0.918	1.168	0.943	0.922	0.846	0.925	0.792	0.756

Table 2: Ratio of the mean squared error of imputation based on $\hat{\tilde{y}}_i$ to the mean squared error of the imputation based on $\bar{y}$, Sim1 population.

5 Conclusions

For many years, studies concerning the sampling of finite populations were aimed at determining optimal strategies; on the one hand, to select a sampling design adapted to the population, and on the other, to obtain a suitable estimator to be used with such a sampling design. All these studies were based on the assumption that, for all the units selected in the sample, information was available concerning all the variables considered and that this information was free of errors.

However, time has revealed the impracticability of these studies, as they do not consider the possibility that information might not be obtained for some of the individuals selected in the sample.

This is why a change has occurred in the last few years concerning the focus of studies carried out in the field of sampling among finite populations. Although several ways to deal with non- response have been tried, the most fully developed has been the automatic imputation of data. Various authors have concentrated on defining efficient imputation techniques in order to obtain a matrix of the complete data and thus apply all the results of strategy optimality. One of these imputation methods is the imputation of the mean, which is frequently used due to its simplicity. In the present study, we propose to use the estimations obtained by various indirect methods as the imputed values, and these are subsequently modified in order to use all the information provided by the sample.

The positive qualities of these estimators lead us to believe that the proposed imputation techniques will lead to increased efficiency with respect to traditional methods.

After carrying out a complete simulation study, we conclude that although the methods for imputation of the mean that apply ratio, difference and regression estimators of complete cases present no obvious advantages over the classical method for imputation of the mean, the pattern changes considerably if we take into account the cases when part of the data is missing from the indirect estimations of the mean. The proposed imputation methods are a little more complex to apply, but they produce an increase in efficiency that is considerable in the estimation of such important parameters as the mean and the median.

References

[1] González S. (2002). *El problema de la falta de respuesta: Alternativas para su tratamiento en la construcción de estimadores indirectos.* Tesis Doctoral, University of Granada.

[2] Brick J. M., Kalton G. (1996). *Handling missing data in survey research.* Statistical methods in medical research **5**, 215 – 238.

[3] King G., Honaker J., Joseph A., Scheve K. (1996). *Listwise deletion is evil: what to do about missing data in Political Science.* **78**, unpublished document.

[4] Fernández F.R., Mayor J.A. (1994). *Muestreo en poblaciones finitas: curso básico.* Ed. PPU.

[5] Little R.J.A., Rubin D.B. (1987). *Statistical analysis with missing data.* John Wiley, New York.

[6] Meeden G. (1995) *Median estimation using auxiliary information.* Survey Methodology **21**, 71 − 77.

[7] Randles R.H. (1982). *On the asymptotic normality of statistics with estimated parameters.* The Annals of Statistics **10**, 462 − 474.

[8] Rubin D.B. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley, New York.

[9] Rueda M., González S. (2000). *Estimadores de razon con datos faltantes.* Metodología de Encuestas **2**, 261 − 272.

[10] Rueda M., González S. (2004). *Missing data and auxiliary information in surveys.* Computational Statistic, in press.

[11] Schafer J.L. (1997). *Analysis of imcomplete multivariate data.* Chapman and Hall, London.

[12] Toutenburg H., Srivastava V.K. (1998). *Estimation of ratio of population means in survey sampling when some observations are missing.* Metrika **48**, 177 − 187.

Address: M.M. Rueda, Dpto. de Estadistica e I.O. Universidad de Granada, Spain

E-mail: mrueda@ugr.es

ORDINAL VARIABLES IN ECONOMIC ANALYSIS

Laura Grassini

Key words: Concentration index, Lorenz curve, ordinal variable.

COMPSTAT 2004 section: Applications.

Abstract: This paper is concerned with the use of the concentration index in measuring the dependence of a continuous variate on an ordinal variable. This approach seems to be particularly suitable in economic analysis, where Gini index is widely used as a measure of variability. After a simulation exercise, an empirical application on family budget data is performed.

1 Introduction

In many areas of economics, relationships among different economic variables must be analyzed. For example, economic theory suggests a relation between total household expenditure and income. A number of contributions, starting with Mahalanobis [4], Blitz and Brittain [1] up to Kakwani [2], [3] and Taguchi [6], proposed to extend and generalize the concept of the Lorenz curve in order to measure the relationships among economic variables. This approach mainly deals with problems of consumer behavior patterns with respect to different commodities.

Specifically, Kakwani derived the expenditure elasticities of a specified item from two concentration curves: the Lorenz curve of total expenditure and a generalized Lorenz curve, which relates the proportion of expenditure on that item with the proportion of consumption units, up to a given level of the total expenditure. Taguchi introduced the concept of concentrative linear regression, where the regression coefficient of the linear model $y = a + bx + error$ is defined in terms of the concentration index of the generalized Lorenz curve for y and the Gini index of the Lorenz curve for x.

From these contributions, an index measuring the level of monotonicity of the relationships between two variables can be derived. This index seems to be suitable for estimating the dependence of a continuous variable on an ordinal variable. This index might be useful, for example, in the analysis of family budgets, where household income is often measured on a ranking scale.

This paper is structured as follows. In the second section the basic concept of Lorenz and concentration indexes are introduced. In that, we follow [6], where the concept of *mean codifference* between two variables was introduced. Hence, we define an index, which expresses the strength of the monotonous dependence of a continuous variable Y on a variate X which can be ordinal. The third section shows the results of a simulation exercise with X ordinal and Y continuous. The last paragraph contains the results of an empirical application on Italian family budget data, to analyze the dependence of consumption expenditures on income, which is an ordinal variate.

2 Basic definitions

This section provides a brief introduction of concentration indexes and related concepts. A more detailed discussion can be found in [3] and [6]. Specifically, in this presentation, we follow the results in [6].

Let us suppose that *(X, Y)* is a non negative bivariate random variable with finite and non-zero mean *(E(X), E(Y))* and joint probability density function *f(x,y)*.

The mean codifference of Y on X is defined as:

$$D_{y|x} = E_1 E_2[(Y_1 - Y_2) \;\; sgn(X_1 - X_2)] \tag{1}$$

where *sgn(w)* is 1 if $w > 0$, -1 if $w < 0$, 0 otherwise, (X_1, Y_1) and (X_2, Y_2) are mutually independent random vectors having the same distribution as *(X, Y)*. A special case of (1) is the mean difference of Y:

$$D_y = E_1 E_2[(Y_1 - Y_2) \;\; sgn(Y_1 - Y_2)] \tag{2}$$

Similarly we can define the mean codifference $D_{x|y}$ of X on Y and the mean difference D_x of X. These variability indexes are strictly connected with the concentration curves (for details see [3]). Namely:

$$G_x = \frac{D_x}{2E(X)} \qquad G_y = \frac{D_y}{2E(Y)} \tag{3}$$

are, respectively, the Gini indexes for X and Y. Moreover, the concentration indexes of X on Y and of Y on X (which are a generalization of the related Gini indexes) are:

$$C_{x|y} = \frac{D_{x|y}}{2E(X)} \qquad C_{y|x} = \frac{D_{y|x}}{2E(Y)} \tag{4}$$

From [3] and [6], we derive the following relation (similarly it holds for X):

$$-D_y \leq D_{y|x} \leq D_y \qquad - G_y \leq C_{y|x} \leq G_y \tag{5}$$

Specifically, if Y is constant, the concentration index is zero; if Y=kX where k is any positive constant, $D_{y|x}$ is equal to the Gini index of X. In general, if *Y=g(.)* and *g(X)* is an increasing function, X and *g(X)* (or Y) will have exactly the same ranking; in this case, $D_{y|x}$ will be equal to D_y (i.e.: $C_{y|x}$ will be equal to the Gini index of Y). Similarly it works for decreasing transformations.

In the followings, we will limit our discussion to situations of positive relationship among two variables.

From ([6],p. 76), if

$$E(Y|X=x) = a + bx \qquad E(X|Y=y) = a' + b'y \tag{6}$$

then

$$b = \frac{D_{y|x}}{D_x} \qquad b' = \frac{D_{x|y}}{D_y} \quad \rho^2_{xy} = \frac{D_{x|y}D_{y|x}}{D_x D_y} \tag{7}$$

Other interesting results arise in the case of bivariate normal distribution. If *(X, Y)* is bivariate normal with correlation coefficient ρ_{xy} and standard deviations σ_x, σ_y, the following relations hold ([6],p. 71):

$$D_y = \frac{2\sigma_y}{\sqrt{\pi}} \qquad D_x = \frac{2\sigma_x}{\sqrt{\pi}} \qquad D_{y|x} = \frac{2\rho_{xy}\sigma_y}{\sqrt{\pi}} \qquad D_{x|y} = \frac{2\rho_{xy}\sigma_x}{\sqrt{\pi}} \tag{8}$$

Hence:

$$\frac{D_{y|x}}{D_y} = \frac{D_{x|y}}{D_x} = \rho_{xy} = R_{x|y} = R_{y|x} \tag{9}$$

From these results, the following index can be proposed for measuring the strength of the monotonous dependence of Y on X

$$R_{y|x} = \frac{D_{y|x}}{D_y} = \frac{C_{y|x}}{G_y} \qquad -1 \le R_{y|x} \le 1. \tag{10}$$

3 Simulation exercise

In order to analyze some properties of $R_{y|x}$ in measuring the dependence between variables, we have conducted a simulation exercise by considering two positively related variables. The simulation here reported does not mean to be exhaustive, but the only purpose is to show some features of the procedure.

The aim of the simulation exercise is to investigate the variation of $R_{y|x}$ with varying strength and type of the relationship between Y and X. We simulated 1000 values for the correlation coefficient ρ_{xy}, assuming a uniform distribution over (0,1). For each value of ρ_{xy}, a sample of 1000 units was generated, under the hypothesis of (X, Y) bivariate normal. Simulated data are shifted to get positive values.

Values used for the response variable are: the original values (identity transformation), the logarithmic and the exponential transformations of Y.

Observations were aggregated into 10 equal sized classes with respect to increasing values of X. This classification represents the situation where X is an ordinal variable.

Ordinal labels from 1 up to 10 are attributed to X, according to this criterion (i.e.: 1 for the first class which contains the 10% smallest values of X, and so on). In fact, ordinal data are generally characterized by a limited number of levels. For example, in Italian family budget data, household income is measured on a ranking scale of 14 levels. This data structure implies that values on Y (which is assumed to be a continuous variable) related with the same value of X, are *smoothed*: the original values of Y relating to the same ordinal value of X, are substituted by their mean.

The computation of $R_{y|x}$ is made through formula (1) on the *smoothed* values of Y. This index is labelled as $R^o_{y|x}$, where the subscript *'o'* means '*ordinal*'. Specifically, we can write:

$$R^o_{y|x} = \frac{D^o_{y|x}}{D_y} = \frac{C^o_{y|x}}{G_y} \tag{11}$$

where it is intended that $D^o_{y|x}$ (and, obviously, $C^o_{y|x}$) is computed on the smoothed values of Y.

Now, the question could be: is the relation $-1 \leq R^o_{y|x} \leq 1$ verified as it is for $R_{y|x}$? To answer this question one must consider the value of G_y decomposed in terms of the classification with respect to increasing values of X. In this case, G_y can be expressed as the sum of three nonnegative components [5]:

$$G_y = Within\, groups(G_y) + Between\, groups\, means(G_y) + Interaction \tag{12}$$

where the *Interaction* term is determined by the presence of overlapping values among groups.

The corresponding $C^o_{y|x}$ is computed exclusively on the groups means of Y (the *smoothed* values) since the same value occurs within each groups and, moreover, the interaction term is absent. Hence, in general:

$$-Between\, groups\, means(G_y) \leq C^o_{y|x} \leq Between\, groups\, means(G_y) \tag{13}$$

Table 1 shows the results of the simulation experiment. In this table, we consider also $|R^o_{y|x}|$ to make a comparison with *ETA*, where:

$$ETA = \sqrt{\frac{Between\ class\ variance\ of\ Y}{Total\, variance\ of\ Y}} \tag{14}$$

In fact, even if simulated values of ρ_{xy} are positive, it is possible that, for any values of ρ_{xy} close to zero, the sample $R^o_{y|x}$ is negative, whereas *ETA* is always nonnegative.

From Table 1, we can see that $|R^o_{y|x}|$ assumes, on the average, values close to *ETA* for the identity and log transformations; values larger than *ETA* in the case of exponential function. Note that, for exponential data, in 884 occurrences $|R^o_{y|x}| > ETA$ (378 in the case of original data, 396 in the case of *log* data). Anyway, $|R^o_{y|x}|$ is always strictly correlated with *ETA*.

4 Application on family budgets

The index $R^o_{y|x}$ is computed on ISTAT family budget data in order to estimate the dependence of a specific consumption expenditures (Y) on income (X),

Function	*Index*	*Mean*	*Variance*	*CV*	*Correlation*
					with ETA
Identity	*ETA*	.4988	.0717	.5366	1
	$R^o_{y\|x}$	.4884	.0802	.5797	.9981
	$\|R^o_{y\|x}\|$	.4890	.0795	.5768	.9987
Exp	*ETA*	.4128	.0511	.5477	1
	$R^o_{y\|x}$	.5058	.0810	.5629	.9832
	$\|R^o_{y\|x}\|$	.5070	.0798	.5570	.9849
Log	*ETA*	.4986	.0716	.5366	1
	$R^o_{y\|x}$	.4885	.0801	.5795	.9981
	$\|R^o_{y\|x}\|$	.4891	.0795	.5766	.9987

Table 1: Simulation results(CV: coefficient of variation).

which is measured on a ranking scales with 14 levels. We considered both total and per capita household expenditures.

Table 2 gives a picture of the distribution of the independent variable: monthly household income. The grouping structure is different from the one used in the simulation experiment, where the main purpose was to show some features of the procedure.

Income class	Income class		
(1000 Lire)	(Euro)	N	%
0 ⊣ 600	0⊣ 309.87	327	1.56
600 ⊣ 1000	309.87⊣ 516.46	1533	7.32
1000 ⊣ 1500	516.46⊣ 774.69	2514	12.01
1500 ⊣ 2000	774.69⊣ 1032.91	3756	17.95
2000 ⊣ 2500	1032.91⊣ 1291.14	3144	15.02
2500 ⊣ 3000	1291.14⊣ 1549.37	2688	12.84
3000 ⊣ 4000	1549.37⊣ 2065.83	3428	16.38
4000 ⊣ 5000	2065.83⊣ 2582.28	1908	9.12
5000 ⊣ 6000	2582.28⊣ 3098.74	786	3.76
6000 ⊣ 7000	3098.74⊣ 3615.20	345	1.65
7000 ⊣ 8000	3615.20⊣ 4131.66	190	0.91
8000 ⊣ 10000	4131.66⊣ 5164.57	136	0.65
10000 ⊣ 12000	5164.57⊣ 6197.48	49	0.23
over 12000	over 6197.48	125	0.60

Table 2: Distribution of monthly household income (year 1999).

As it can be derived from Table 3, the indices exhibit larger values on total expenditure data because of the underlying effect of household size on expenditure.

	Total		Per capita	
Expenditure	$R^o_{y\|x}$	*ETA*	$R^o_{y\|x}$	*ETA*
Food	.3345	.3146	-.1113	.1033
Beverages	.2527	.2195	-.0071	.1033
Tobacco	.1660	.1397	.0341	.0334
Clothing and shoes	.4009	.3059	.2283	.1654
Housing	.4543	.3171	.0307	.0927
Furniture, home appliances	.3334	.1341	.1795	.0693
Health	.2147	.1092	.0102	.0284
Transportation, communications	.4908	.2574	.3283	.1545
Culture and recreation	.4110	.3197	.2027	.1418
Education	.4186	.1761	.3754	.1506
Other goods and services	.4442	.3102	.1727	.1310
Total expenditure	.5688	.4689	.1910	.1605

Table 3: Analysis of Italian family budget data (year 1999).

Looking at the results on per capita data, greater dependence is observed for Education, Transportation and communications, Clothing and shoes.

In most situations, $|R^o_{y|x}|$ is greater than *ETA*, as it occurred on simulated data, in the case of exponential transformation. However this comparison is not completely correct, because of the different grouping of data.

References

[1] Blitz R.C., Brittain J.A.(1964). *An extension of the Lorenz diagram to correlation of two variables.* Metron **33**, 137 – 143.

[2] Kakwani N.C. (1977). *Applications of Lorenz curves in economic analysis.* Econometrica **45**, 719 – 27.

[3] Kakwani N.C. (1980). *Income inequality and poverty.* Oxford University Press.

[4] Mahalanobis P.C. (1960). *A Method of fractile graphical analysis.* Econometrica **28**, 325 – 51.

[5] Pyatt G. (1976). *On the interpretation and disaggregation of Gini coefficient.* Economic Journal **86**, 243 – 55.

[6] Taguchi T. (1981). *On a multiple Gini's coefficient and some concentrative regressions.* Metron **1**, 69 – 97.

Address: L. Grassini, Dipartimento di Statistica, University of Florence, Italy

E-mail: grassini@ds.unifi.it

HIGH-DIMENSIONAL PROBABILISTIC CLASSIFICATION FOR DRUG DISCOVERY

A. Gray, P. Komarek, T. Liu and A. Moore

Key words: Virtual screening, high-throughput screening, classification, nonparametric, metric learning, nonisotropic, kernel density estimation.

COMPSTAT 2004 section: Biostatistics.

Abstract: Automated high-throughput drug screening constitutes a critical emerging approach in modern pharmaceutical research. The statistical task of interest is that of discriminating active versus inactive molecules given a target molecule, in order to rank potential drug candidates for further testing. Because the core problem is one of ranking, our approach concentrates on accurate estimation of unknown class probabilities, in contrast to popular non-probabilistic methods which simply estimate decision boundaries. While this motivates nonparametric density estimation, we are faced with the fact that the molecular descriptors used in practice typically contain thousands of binary features. In this paper we attempt to improve the extent to which kernel density estimation can work well in high-dimensional classification settings. We present a synthesis of techniques (SLAMDUNK: Sphere, Learn A Metric, Discriminate Using Nonisotropic Kernels) which yields favorable performance in comparison to previous published approaches to drug screening, as tested on a large proprietary pharmaceutical dataset.

1 Introduction: classification for drug screening

Virtual screening refers to the use of statistical and computational methods for prioritizing candidate molecules for biological testing for their possible use as drugs. Because these assays are time-consuming and expensive, accurate "virtual" assays, or prioritization of molecules by computer, has direct impact in cost savings and more rapid drug development. Virtual screening, part of the more general enterprise of *high-throughput screening*, has thus become an increasingly pressing new component of modern drug development research.

The classification problem. In this paper we are concerned with the scenario of a large pharmaceutical research and development laboratory, which is as follows: We assume there is a single *target* molecule. There are multiple molecules which are known to interact in the desired fashion with the target molecule, *i.e.* are *active* with respect to the target, and a generally larger number of molecules known to be *inactive* with respect to the target. The task is to predict whether a previously unseen molecule will be active with respect to the target.

The features. The structure of a molecule determines its interaction with a target molecule – whether and how it will interlock, or "dock" with the target – but the interaction is itself a complex dynamic process whose complete characterization remains an oustanding problem of science. Thus, molecular descriptions used in virtual screening typically contain hundreds or thousands of binary (0/1) features, collecting all manner of both generic and target-specific properties which might be relevant to the classification task. Typical binary features record the absence or presence of a certain kind of atom or substructure, proximity relationship, and so on.

The goal. Our goal to design a classifier with the best possible prediction performance based on a proprietary commercial training set of 26,733 molecules, 6,348 binary features, and one output variable ("active" or not).

Recent work in virtual screening. Most of the well-known classification methods have been proposed for the virtual screening problem, including decision trees, neural networks, naive Bayes classifiers, and support vector machines (SVM) ([16]), which are currently considered to be one of the most empirically successful in general. Our work is strongly motivated by two of the most recently published comparisons of classification methods for virtual screening ([17],[10]), which reveal two slightly lesser-known winners. One is the 'binary kernel discriminator' (BKD) of [9], a simple kernel estimator for classification using a kernel based on the Hamming distance. (We note that the BKD is not formulated directly in terms of decision theory.) In [17], a fairly extensive comparison (by a different group of researchers than the ones who first proposed BKD's for this problem) between SVM's and BKD's was performed, demonstrating surprisingly clear superiority in the performance of BKD's over SVM's. In that work, molecule descriptions containing up to about 1,000 features were used. In [10], which performed experiments using the *same dataset* used in this paper, a conjugate gradient-based logistic regression (LR) method was demonstrated to have consistently favorable performance compared with several popular methods including SVM's with both linear and nonlinear (radial basis function) kernels, decision trees, naive Bayes classifiers, and k-nearest-neighbor classifiers. Our work ultimately contains aspects of both BKD and LR, achieving a method with performance superior to either one.

Ranking versus binary decision-making. To score the ranking performance of a classifier, we use the standard device of *receiver operating characteristic* (ROC) curves ([3]), which captures more information than simply the percentage of correctly-classified data.[1] The starting point for the approach of this paper is that the ranking problem is more difficult than the standard classification problem because the quantity of interest is the posterior class probability rather than simply the error rate of making binary decisions. A classifier may estimate class probabilities with very large bias, but

[1] An ROC curve is constructed by sorting the data according to the predicted probability for the "active" class, *i.e.* $P(C_1|x)$. Starting at the origin and stepping through the data in order of decreasing "active" probability, a point on the curve is plotted by moving up one unit if the true label was actually "active" and moving right one unit if the prediction was incorrect. A summary of an ROC curve is the *area under the curve* (AUC), which is 0.5 for a classifier which guesses randomly and 1.0 for one which ranks perfectly.

still perform well when scored in terms of accuracy in binary decision-making as long as the order relation between the class probabilities is maintained.

In this work we pursue the extent to which direct estimation of posterior class probabilities, as opposed to pure classification designed to minimize the binary error rate, might yield superior ranking performance. There are additional practical advantages to obtaining accurate class-conditional densities. Among them: imputation of missing data is naturally treated, outliers are more naturally identified, and difficult-to-classify data are easily isolated.

2 General approach

Decision theory. The motivation above leads us naturally to the general framework of statistical decision theory. The posterior class probability $P(C_1|x)$, is expressed in terms of the class-conditional density $p(x|C_1)$:

$$P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)} \tag{1}$$

where C_1 and C_2 are the two classes. If the class-conditional distributions on the right-hand side are known, the Bayes error rate is achived, *i.e.* the best possible performance is obtained.

Nonparametric density estimation. We consider the classifier obtained by estimating $p(x|C_1)$ and $p(x|C_2)$ with minimal assumptions, using the nonparametric *kernel density estimator* (KDE):

$$\hat{p}(x) = \frac{1}{N}\sum_{i}^{N} K_h(x, x_i) \tag{2}$$

where N is the number of data, $K()$ is called the kernel function and satisfies $\int_{-\infty}^{\infty} K_h(z)dz = 1$, and h is a scaling factor called the bandwidth. Kernel density estimation is the most widely-used and well-studied method for nonparametric density estimation, owing to both its simplicity and flexibility, and the many theorems establishing its consistency for near-arbitrary unknown densities and rates of convergence for its many variants ([14],[13]). We refer to the resulting classifier as a *nonparametric Bayes classifier* (NBC), for lack of a standard name. The standard form of kernel which is most often used is the *product kernel*, in which

$$K_h(x, x_i) = \prod_{d}^{D} K_d\left(\frac{\|x - x_i\|}{h}\right), \tag{3}$$

where D is the number of dimensions, *i.e.* the kernel function is a product of D univariate kernel functions, and all share the same bandwidth h. Though we could consider a setup in which separate bandwidths can be adjusted for each dimension, this creates a combinatorial problem which is intractable in our high-dimensional setting. If we ensure that the scales of the respective features are roughly the same, we need only adjust a single parameter h.

Thus, our classifier has two parameters, the bandwidth for each class. These are found by first estimating the optimal bandwidth for each density independently using least-squares cross-validation ([14]), then scoring bandwidth pairs nearby these values using the leave-one-out error score.

3 SLAMDUNK: sphere, learn a metric, discriminate using nonisotropic kernels

The nonparametric Bayes classifier arises naturally when considering the best available method for accurate estimation of class probabilities with minimal assumptions. However, its power comes at potentially severe costs. The SLAMDUNK methodology consists of a set of procedures designed to mitigate the traditional limitations of nonparametric density estimation in the setting of high-dimensional classification, so that its distinct advantages may be exploited. We now treat in turn three significant roadblocks.

3.1 Fast algorithms for kernel density estimation

Estimation of the density at each of the N points, when performed in the straightforward manner, has $O(N^2)$ computational cost. Computational intractability impacts statistical inference quality directly – for example in [17] only 200 data were subsampled for each class to form the training set, due to the computational cost of BKD. In our experiments we use the entire set of 26,733 data. Any high-dimensional context demands the use of as much data as possible, forcing the computational issue. Fortunately, this problem has been largely mitigated in very recent work presenting a fast algorithm yielding simultaneously fast and accurate computation of kernel density estimates ([8]). The algorithm reduces the $O(N^2)$ cost to $O(N)$. Further, it is shown empirically in [8] that the algorithm's time complexity is not exponential in the dimension D, but instead appears to depend on the *intrinsic dimensionality*, the local dimensionality of the manifold upon which the data lies ([5]) (see below). However, the computational geometry methods employed by the algorithm require that the underlying distance be a true metric, which will constrain our methodology below.

3.2 Nonstationary and nonisotropic estimators

A major perceived obstacle is the statistical inefficiency of KDE in high dimensions. Theoretical bounds establish that in the worst case, the number of samples required for accurate kernel density estimation rises exponentially with the dimension ([14]). We now consider more realistic alternative choices for the kernel functions in KDE.

Nonstationary estimators. It has long been noted that the assumption of spatial *stationarity*, or a single scale h holding across the entire space is deficient. Visually it is clear that smoothing with a fixed bandwidth is unappealing when the dataset contains regions of differing density, which is inevitable in practice. Adaptive (or variable-kernel) kernel density estimators have been studied and shown to be more effective than fixed-width kernel density estimators in experimental studies, *e.g.* [15]. In these estimators, the variable bandwidth h_i for each point x_i is obtained by scaling the single global bandwidth h by a factor

$$\lambda_i \propto \{\tilde{p}(x_i)\}^{-1/2} \tag{4}$$

where $\tilde{p}()$ is a pilot estimate of the density, to which the overall estimator is largely insensitive ([1]). Many simple choices can be used for this pilot estimate, including adaptive Gaussian mixture models or variable kernel estimators based on nearest-neighbor distances ([2]).

Nonisotropic estimators. It has been noted by many authors (particularly in the field of machine learning, in which high-dimensional data classification and clustering is routinely performed) that in practice it is virtually never the case that a dataset's intrinsic dimensionality is equal to its explicit dimensionality D, *e.g.* [4]. With the assumption that the data lie on a linear manifold, the dimension of the subspace can be estimated using the eigenspectrum from a principal components analysis ([5]). However in general the data may lie on a nonlinear manifold ([12]). A common way estimator of the intrinsic dimension with minimal assumptions has been called, among other things, the correlation dimension ([7]), but amounts to the 2-point correlation function used in spatial statistics. Very often in practice the intrinsic dimension $D' << D$, regardless of which variant of its definition is used.

With this in mind, the standard product kernel, which is *isotropic*, *i.e.* has equal extent in all directions, is a poor match to realistic high-dimensional data. Further, as noted earlier, the behavior of volumes in high dimensionalities, rising exponentially in D, is disastrous when D is large. Instead we use an estimator in which the univariate bandwidths h_i are replaced by matrices H_i, resulting in a multivariate kernel such as the multivariate Gaussian

$$K_{H_i}(x, x_i) = \frac{1}{(2\pi)^{D/2}|H_i|^{1/2}} \exp\left\{-\frac{1}{2}(x - x_i)^T {H_i}^{-1}(x - x_i)\right\} \tag{5}$$

where $H_i = h\lambda_i\Sigma_k$, with Σ_k the covariance matrix estimated from x_i and its k nearest neighbors of x_i. Such estimators have received relatively little study, though one example showing their consistency is [6]. By allowing increased sensitivity to the local manifold of the data, or correlations in the feature space, we deflate the worst-case curse of dimensionality in KDE, relative to the naive product kernel estimator.

3.3 Coordinate transformation and metric learning

One of Vapnik's central arguments for the non-probabilistic approach [16] underlying the SVM is that if the error rate is the desired quantity to be minimized, estimation of entire densities rather than simpler decision boundaries is unnecessary and wasteful of modeling capacity ([16]). We now introduce a methodology for essentially focusing less modeling effort on directions that are less relevant with respect to the decision boundary.

Metric learning. An implicit part of the kernel estimator is the underlying metric used to obtain the distances. The standard Euclidean distance is used by default. It can be seen as a special case of a more general weighted Euclidean distance

$$d(x, y) = \|x - y\| = \sqrt{(x - y)^T W (x - y)} \tag{6}$$

in which the matrix W is diagonal containing all 1's. We instead consider the metric weight matrix W as a free parameter to adjust to maximize the performance of our estimator. We refer to this as "learning the metric".[2] This functional form retains the metric properties, for the purpose of using the fast algorithm described above.

The linear discriminant metric. We propose a form of W which is diagonal, and relates the metric to the decision boundary. We obtain the vector w (the diagonal of W) which is the result of a linear classifier such as logistic regression or a linear support vector machine (we use logistic regression based on the favorable experimental results described earlier). The weight vector w describes a classifier where the class prediction for x is obtained by computing wx and comparing it to a threshold w_0. Thus if two points x and y lie on the decision boundary of the classifier, we have that $w^T(x - y) = 0$, *i.e.* the vector w is orthogonal to the decision boundary. By taking the metric formed by the norm

$$d(x, y) = \|w^T (x - y)\| \tag{7}$$

we obtain a metric which measures distance along w, or between the class means (with the appropriate Gaussian assumptions). This can be interpreted as measuring the extent to which the linear discriminant prefers class 1 or class 2. It can be regarded as an implicit form of dimensionality reduction, by realizing that values of w tending to zero will cause the metric to assign negligible weight in those directions, which is tantamount to removing the corresponding features.

Sphering. Our diagonal restriction on W motivates the removal of correlation between the features in advance. Normalizing each feature so that they all have roughly the same scale is also important for KDE as noted earlier. For this reason we perform these operations (*sphering* the data) as the first step of our methodology using principal component analysis (PCA). We also take the opportunity at this stage to examine the resulting eigenspectrum and remove low-eigenvalue features. Our overall dimension reduction scheme thus includes two kinds of steps: this PCA-based explicit feature removal, which aims to 'denoise' the data, and the use of discriminant information to replace isotropy in the metric.

[2] Although asymptotic results imply that the choice of metric does not affect performance, finite-sample experiments show that marked improvements can be made by adjusting the metric to the task at hand [4], [11].

4 Experimental results

Our dataset contains 26,733 rows and 6,348 attributes, and is sparse, containing 3,732,607 non-zero input values. It has 804 positive output values ("active" class). A pre-analysis of the data, however, reveals that 2290 columns are empty. Furthermore, 388 out of 8,235,711 pairs of columns are identical. These are also removed. Among the remaining columns, a column reduction scheme also reveals linear dependencies. Removal of 406 columns from the remaining 3,871 columns is performed. This leaves about half of the original dimensions. We then perform PCA, keeping only 100 of these dimensions. This value was chosen to correspond roughly to the inflection point of the eigenspectrum, as per common practice, and captured 97% of the variance in this case. All experiments were performed using 10-fold cross-validation and testing (evaluation) is performed on one of them while training is performed on the other 9 put together.

Method	AUC
k-nearest neighbors	0.862 ± 0.017
Bayes classifier	0.891 ± 0.012
decision tree	0.893 ± 0.011
linear support vector machine	0.918 ± 0.010
RBF support vector machine	0.927 ± 0.013
logistic regression	0.931 ± 0.012
SLAMDUNK fixed isotropic kernel	0.933 ± 0.017
SLAMDUNK fixed isotropic + metric learning	0.937 ± 0.012
SLAMDUNK variable nonisotropic + metric learning	0.940 ± 0.012

The first part of the table lists the results of the experimental evaluation of [10] performed on the same data. Each method was tested both with and without the use of PCA projecting to 100 dimensions. The table shows only the better of the two procedures, for each method. The second part of the table shows the SLAMDUNK results on this data. We see a progression in the AUC score when metric learning is performed, and when variable and nonisotropic kernels are used, showing that these techniques each contribute to increased prediction quality in a non-conflicting and additive manner. [3]

5 Conclusion

We have presented a methodology called SLAMDUNK which we have designed to have favorable properties for the problem of virtual screening. We have demonstrated its favorable performance on a real pharmaceutical dataset in current use for drug discovery, providing evidence that this line of thinking may hold promise for this important contemporary problem. More generally, this work explores the extent to which fully probabilistic methods can be successful in high-dimensional problems.

[3] Test results are not shown for each of the possible combinations of the sub-techniques of SLAMDUNK for lack of space.

References

[1] Abramson. I. (1982). *On bandwidth variation in kernel estimation – a square root law.* Annals of Statistics **10**, 1217 – 1223.

[2] Breiman L., Meisel W., Purcell E. (1977). *Variable kernel estimates of multivariate densities.* Technometrics **19**, 135 – 144.

[3] Duda R.O., Hart P.E. (1973). *Pattern classification and scene analysis.* John Wiley & Sons.

[4] Friedman J.H. (1994). *Flexible metric nearest neighbor classification.* Technical report, Stanford University.

[5] Fukunaga K. (1990). *Introduction to statistical pattern recognition.* 2nd ed. Academic Press.

[6] Givens G. H. (1994). *Consistency of the local kernel density estimator.* Technical report, Colorado State University.

[7] Grassberger P., Procaccia I. (1983). *Measuring the strangeness of strange attractors.* Physica D, 189 – 208.

[8] Gray A.G., Moore A.W. (2003). *Very fast multivariate kernel density estimation via computationalgeometry.* In Joint Statistical Meeting 2003. To be submitted to JASA.

[9] Harper G., Bradshaw J., Gittins J.C., Green D.V.S. (2001). *Prediction of biological activity for high-throughput screening using binary kernel discrimination.* J. Chem. Inf. Comput. Sci. **41**, 1295 – 1300.

[10] Komarek P., Moore A.W. (2003). *Fast robust logistic regression for large sparse datasets with binary outputs.* In Workshop on AI and Statistics.

[11] Minka T.P. (2000). *Distance measures as prior probabilities.* Technical report, Massachusetts Institute of Technology.

[12] Roweis S., Saul L. (2000). *Nonlinear dimensionality reduction by locally linear embedding.* Science **290** (5500).

[13] Scott D.W. (1992). *Multivariate density estimation.* J. Wiley.

[14] Silverman B.W. (1986). *Density estimation.* Chapman and Hall, New York.

[15] Terrell G., Scott D.W. (1992). *Variable kernel density estimation.* Annals of Statistics **20** (3), 1236 – 1265.

[16] Vapnik V.N. (1995). *The nature of statistical learning theory.* Springer-Verlag.

[17] Wilton D., Willett P. (2003). *Comparison of ranking methods for virtual screening in lead-discovery programs.* J. Chem. Inf. Comput. Sci. **43**, 469 – 474.

Address: A. Gray, P. Komarek, T. Liu, A. Moore, Carnegie Mellon University, Robotics Institute, 5000 Forbes Avenue, Pittsburgh PA 15221

E-mail: agray@cs.cmu.edu

COMPSTAT'2004 Symposium

DETERMINATION OF CONSTRAINED MODES OF A MULTINOMIAL DISTRIBUTION

Marian Grendar

Key words: Kullback-Leibler distance, information projection, μ-projection, Gibbs' conditioning, maximum probability method.

COMPSTAT 2004 section: Algorithms.

Abstract: Modes of the multinomial distribution can be found either by means of Moran's bounds or by Finucan's or Le Gall's iterative procedures. A tool for determination of constrained modes of a multinomial distribution is presented here. It is based on Gibbs' conditioning and Moran's bounds.

1 Motivation

Let $\{X\}_{l=1}^{n}$ be a finite sequence of identically distributed random variables with a common law (called source or generator) q on a measurable space (Ω, S). Let the measure q be concentrated on m atoms from a set $\mathcal{X} \triangleq [x_1, x_2, \ldots, x_m]$, called alphabet or support. A type (n-type, or 'empirical measure') ν^n is defined as vector of the following elements: $\nu_i^n \triangleq 1/n \sum \delta_{X_l}$, $i = 1, 2, \ldots, m$; where δ_y denotes a Dirac measure at y. Multiplicity $\Gamma(\nu^n)$ of a type ν^n is defined as number of sequences $\{X\}_{l=1}^{n}$ which induce the type; i.e., $\Gamma(\nu^n) \triangleq \frac{n!}{n_1!\, n_2! \cdots, n_m!}$.

Let $\mathcal{H}_n$ be a set of n-types. Given $\{\mathcal{X}, n, q, \mathcal{H}_n\}$, Boltzmann-Jaynes inverse problem (BJIP) amounts to selection of a type (one or more) from the set $\mathcal{H}_n$. It is assumed that there is no restriction on multiplicities of types from $\mathcal{H}_n$.

One possible approach to the BJIP is based on Maximum Probability method (MaxProb), see [7]. Given the information-quadruple $\{\mathcal{X}, n, q, \mathcal{H}_n\}$ MaxProb selects from the feasible set of types $\mathcal{H}_n$ just the type(s) $\hat{\nu}^n \triangleq \arg\max_{\nu^n \in \mathcal{H}_n} \pi(\nu^n|q)$ which attains the highest value of $\pi(\nu^n|q) \triangleq \Gamma(\nu^n) \cdot \Pi_{i=1}^{m} q_i^{n\nu_i}$. In other words, MaxProb method selects the mode(s) of a multinomial distribution $\pi(\nu^n|q)$ under the constraint $\nu^n \in \mathcal{H}_n$. For an application of MaxProb to BJIP see [11]; for a bayesian interpretation of MaxProb see [10]. In what follows, $\hat{\nu}^n$ will be called μ-projection of the source q on the set $\mathcal{H}_n$. When n approaches infinity, μ-projection(s) converges to I-projection(s) $\hat{p}$ of q on $\mathcal{H}_\infty$, see [7]. Recall that I-projection $\hat{p}$ of a probability mass function (pmf) r on a set $\mathcal{G}$ of pmf's is defined as $\hat{p} \triangleq \arg\inf_{r \in \mathcal{G}} I(r||q)$, where $I(r||q) \triangleq \sum_{\mathcal{X}} r \log(r/q)$ is the information divergence, or Kullback-Leibler distance, or minus relative entropy. This motivates a need for a general way of finding the MaxProb type(s).

2 Determination of unconstrained modes of a multinomial distribution

When the feasible set of types is defined as $\mathcal{U}_n \triangleq \{\nu^n : <\nu^n, 1> = 1\}$, where $<,>$ denotes scalar product, 1 is an m-element vector of ones, then the problem of finding MaxProb type(s) reduces to the problem of finding standard, "unconstrained" mode(s) of the multinomial distribution $\pi(\nu^n|q)$. This problem can be approached in two ways: 1) via Moran's bounds (cf. [5]), or 2) via Finucan's algorithm (cf. [6]), which was recently extended by Le Gall [13].

Moran's bounds are based on the following property (cf. [13]), which is straightforward to establish:

Property. Let $\mathcal{U}_n$ be the feasible set of n-types. The probability $\pi(\nu^n|q)$ attains its maximum on $\mathcal{U}_n$ at any $\hat{\nu}^n \in \mathcal{U}_n$ for which the following holds:

$$q_i \hat{n}_j \leq q_j(\hat{n}_i + 1) \qquad \forall\ i, j \in 1, 2, \dots, m \quad \text{with} \quad \hat{n}_j > 0 \tag{1}$$

The Moran's lower bound results from summing both sides of (1) over j. The upper bound can be obtained by summing a rearrangement of (1):

$$\hat{n}_i \geq \frac{q_i}{q_j}\hat{n}_j - 1$$

over all $i \neq j$, keeping j fixed (cf. [5]). This leads to the Moran's bounds:

$$nq_i - 1 \leq \hat{n}_i \leq (n + m - 1)q_i$$

Note that the lower bound can be sharpened to $(n+1)q_i - 1 \leq \hat{n}_i$.

Thus, provided that a unique type $\hat{\nu}^n \in \mathcal{U}_n$ lays within the Moran's bounds, it is the sought mode of the multinomial distribution. If m is not small, several types usually fall within the Moran's bounds and mode(s) should be isolated among them by an enumerative procedure (see [12]). Even for moderately big m there can be thousands of types within the bounds. In this case, a faster way of finding the mode is provided by Finucan's iterative algorithm, or its recent extension/alternative due to Le Gall.

It is worth noting that if 1) the source is rational (i.e., $q_i \in \mathrm{Q}, i = 1, 2, \dots, m$) and 2) $n = kw$, where w is the common denominator of q's and k is a natural number, then the unconstrained mode of the multionomial distribution $\pi(\nu^n|q)$ is $\hat{\nu}^n = nq$ and it is unique. If $n \neq kw$, or q is not rational, then the mode is in general not unique. Finucan's/Le Gall's algorithm iteratively searches for such value(s) of $\rho \in \mathrm{R}$ for which

$$\sum_i \lfloor \rho q_i \rfloor = n \tag{2}$$

and this way determines the mode(s). If there is an interval of values of ρ for which (2) holds, then the mode is unique.

3 Determination of μ-projections

Here we are interested in determination of mode of multinomial distribution, when types are restricted to lay in a set $\mathcal{H}_n \subseteq \mathcal{U}_n$.

With no loss for the points to be made, let the feasible set $\mathcal{H}_n$ be hereafter defined as $\mathcal{H}_n \triangleq \{\nu^n : \nu^n n \in \mathrm{Z}^m, <\nu^n, 1>= 1, <\nu^n, x>= a\}$, where x is a known m-element vector and a is a fixed, known number.

Example 1. Let $q = [0.13\ \ 0.09\ \ 0.42\ \ 0.36]$. Let $x = [1\ \ 2\ \ 3\ \ 4]$. Let $n = 10$ and $a = 3.2$. This information defines a feasible set $\mathcal{H}_{10}$ which comprises 10 types (cf. [7]). It is desired to find μ-projection(s) of q on $\mathcal{H}_{10}$. $\diamond$

The μ-projection can be in principle found directly by a brute force method: for each type in $\mathcal{H}_n$ calculate the probability π (which involves calculation of factorials) and select the type(s) which has the highest value of the probability. If the feasible set is not given in the form of list of all its elements rather it is given by a defining formula (as in the Ex. 1), then it is necessary to construct the list of types which satisfy the definition. This might be not an easy task. However, the main practical problem is that a feasible set of types could comprise thousands of types, and the direct approach would be quite computer resources consuming/demanding.

3.1 Gibbs' conditioning

Gibbs' conditioning principle (see [2], [3], [4]) states that I-projection of q on $\mathcal{H}_\infty$ can be viewed for sufficiently large n as an almost iid source of types. In other words, the Gibbs' conditioning implies that for large n any sequence (and hence also type) from the feasible set can be viewed almost as if it was

drawn iid from the I-projection (the 'almost' qualifier comes from the 'end effect', see [1]).

It is then tempting (but erroneous) to expect that unconstrained n-mode of multinomial distribution with the source q replaced by the I-projection $\hat{p}$ of q on $\mathcal{H}_\infty$ will be identical with the sought μ-projection of q on $\mathcal{H}_n$. This is not the case, as the next Example illustrates.

Example 1 (cont'd):

I-projection of q on $\mathcal{H}_\infty$ is $\hat{p} = [0.0826\ \ 0.0709\ \ 0.4103\ \ 0.4361]$, cf. [7]. The unconstrained mode of the multinomial distribution $\pi(\nu^{10}|\hat{p})$ is $[1\ \ 0\ \ 4\ \ 5]$. The unique mode was found by Finucan's algorithm (with $\rho \in (12.10, 12.18)$). The unconstrained mode is different than the sought constrained mode (i.e., μ-projection): $[1\ \ 0\ \ 5\ \ 4]$. $\diamond$

3.2 Algorithm: Gibbs + Moran

Yet, it is reasonable to expect (and the above calculation supports it) that the μ-projection should lay close to the unconstrained mode of $\pi(\nu^n|\hat{p})$. Moran's bounds can be used to guide search for the μ-projection. From this observation arises a tool for determination of the μ-projection. The algorithm can be divided into five steps:

1. Find the I-projection $\hat{p}$ of q on $\mathcal{H}_\infty$.

2. Calculate Moran's bounds for $\pi(\nu^n|\hat{p})$.

3. List all n-types which fall within the bounds and select those of them which also belong to the feasible set $\mathcal{H}_n$.

4. If none of the listed types belong to the feasible set, loosen the bounds, and repeat Steps 2,3.

5. If there are several such types, then calculate their probabilities $\pi(\nu^n|q)$ and select the most probable type(s).

Example 2. a) Let $n = 100$, and all the other things let be the same as in the Ex. 1. $\mathcal{H}_{100}$ comprises 574 types (cf. [7]). We would like to find the μ-projection of q on $\mathcal{H}_{100}$.

Steps: 1) The I-projection is: $\hat{p} = [0.0826\ \ 0.0709\ \ 0.4103\ \ 0.4301]$. 2) Moran's bounds on $\pi(\nu^{100}|\hat{p})$ are: $x_1 \in [7\ \ 8]$, $x_2 \in [6\ \ 7]$, $x_3 \in [40\ \ 42]$ and $x_4 \in [42\ \ 44]$. 3) Among the types which fall into the bounds, the following types satisfy the adding-up constraint: $\nu_1 = [7\ \ 7\ \ 42\ \ 44]$, $\nu_2 = [8\ \ 6\ \ 42\ \ 44]$, $\nu_3 = [8\ \ 7\ \ 41\ \ 44]$ and $\nu_4 = [8\ \ 7\ \ 42\ \ 43]$. The last type satisfies also the mean constraint (i.e., belongs to $\mathcal{H}_{100}$). Since loosening of the bounds leads to types with lower value of $\pi(\nu^{100}|q)$, the type ν_4 is the sought μ-projection of q on $\mathcal{H}_{100}$.

b) For $n = 1000$ there are 53734 types in $\mathcal{H}_{1000}$. μ-projection should be selected among them. Steps: 2) Moran's bounds are: $x_1 \in [81\ \ 82]$, $x_2 \in [69\ \ 71]$, $x_2 \in [409\ \ 411]$ and $x_4 \in [435\ \ 437]$. 3) Among them there are the following four types which satisfy the adding up constraint (i.e., belong to $\mathcal{U}_{1000}$): $\nu_1 = [81\ \ 71\ \ 411\ \ 437]$, $\nu_2 = [82\ \ 70\ \ 411\ \ 437]$, $\nu_2 = [82\ \ 71\ \ 410\ \ 437]$, and $\nu_4 = [82\ \ 71\ \ 411\ \ 436]$. None of them however satisfy the mean constraint (i.e., belong to $\mathcal{H}_{1000}$). 4) Loosen the bounds. Loosening of the lower bound brings no effect. Loosening of the upper bound $ub(i) = \lfloor (n + m - 1)\hat{p}(i) \rfloor$ into $ub(i) = \lfloor (n + m)\hat{p}(i) \rfloor$ brings no effect. But $ub(i) = \lfloor (n + m + 1)\hat{p}(i) \rfloor$ leads to the following two types which satisfy also the mean constraint: $\nu_1 = [83\ \ 70\ \ 411\ \ 436]$ and $\nu_2 = [83\ \ 71\ \ 409\ \ 437]$. 5) Now it remains to find which of them has higher value of the probability $\pi(\cdot|q)$. Form a ratio of probabilities of $\pi(\nu_1|\cdot)$ to $\pi(\nu_2|\cdot)$: $71 * 437/(410 * 411) q_2^{-1} q_3^2 q_4^{-1}$ which is 1.0025. So, ν_1 is a candidate for μ-projection. Further loosening of the bounds leads to types with lower probabilities, which permits us to conclude that ν_1 is the μ-projection and it is unique. ⋄

While there is abundance of types within Moran's bounds in the unconstrained case when m is even moderately large, this doesn't happen in the constrained case, as the next Example illustrates.

Example 3. Let there be q, $n = 17$ as in Le Gall [13], Table 1; case 5. Let in addition, the support be $\mathcal{X} = [1\ \ 2, \ldots, 6]$ and the mean value $a = 2$. The I-projection of q on $\mathcal{H}_{17}$ can be found to be: $\hat{p} = [0.4382\ \ 0.3171\ \ 0.1147\ \ 0.0830$

0.0300 0.0169]. By the algorithm/tool with 'tight' Moran's bound the following two types were trapped: [7 5 3 2 0 0] and [6 7 2 2 0 0]. Further loosening of the bounds trapped other 6 types, among which the following was found to have the same probability as the two already found: [7 6 2 1 1 0]. Direct 'brute force' approach revealed that there are 119 types in the feasible set and that the three MaxProb types which were found by the 'algorithm' are indeed the sought μ-projections. $\diamond$

4 Notes

The presented tool for determination of constrained mode(s) of the multinomial distribution relays on Gibbs' conditioning (and Moran's bounds) and thus critically depends on the assumption that I-projection on the set $\mathcal{H}_\infty$ can be numerically 'easily' obtained. If the feasible set is defined by the usual moment constraints (of which the case of first moment was considered here, in detail) then the I-projection can be indeed easily found. However, for a non-linear moment constraints (cf. [9]) this is not the case.

Applicability of the tool is not restricted to the case of unique I-projection. Even if $\mathcal{H}_\infty$ admits several I-projections, the types concentrate on them (cf. [8]). An extension of the Gibbs' conditioning principle then supports the tool also in this case.

R function `expand.grid` can be used to accomplish Step 3 efficiently.

Application of the tool to numerous setups permits to make the following conjecture: if sole type falls between the Moran's bounds for $\pi(\nu^n|\hat{p})$ and the type at the same time belongs to the feasible set $\mathcal{H}_n$, then it is μ-projection (of q on $\mathcal{H}_n$) and it is unique.

Since I-projection of q on $\mathcal{U}_\infty$ is just q, the presented tool reduces to the standard way of finding the unconstrained mode(s) of the multinomial distribution $\pi(\nu^n|q)$ via Moran's bounds, when $\mathcal{H}_n$ reduces to $\mathcal{U}_n$.

5 Summary

A tool for determination of constrained mode(s) of the multinomial distribution (i.e., μ-projection(s) of a source q on a feasible set of types $\mathcal{H}_n$) was presented. The tool is based on Gibbs' conditioning and Moran's bounds. It permits to find μ-projections without listing types which belong to the feasible set $\mathcal{H}_n$. The combination of Gibbs' conditioning principle with Moran's bound substantially narrows the original feasible set of types and at the same time directs the search for μ-projections to the relevant subset of the feasible set. Several examples were developed to illustrate application of the tool. The tool could help promote exploitation of Maximum Probability method (cf. [7]) in the broad area of BJIP.

To mar, in memoriam.

References

[1] Cover T., Thomas J. (1991). *Elements of information theory.* Wiley (NY).

[2] Csiszar I. (1984). *Sanov property, generalized I-projection and a conditional limit theorem.* Annals of Probability **12**, 768 – 793.

[3] Csiszar I. (1998). *The method of types.* IEEE IT **44**, 2505 – 2523.

[4] Dembo A., Zeitouni O. (1998). *Large deviations techniques and applications.* Springer (NY).

[5] Feller W. (1968). *An introduction to probability theory and its applications.* Wiley (NY).

[6] Finucan H.M. (1964). *The mode of a multinomial distribution.* Biometrika **51**, 512 – 517.

[7] Grendar M. Jr., Grendar M. (2001). *What is the question that MaxEnt answers? A probabilistic interpretation.* CP 568, 83 – 94, AIP (Melville).

[8] Grendar M. Jr., Grendar M. (2003). *Asymptotic equiprobability of I-projections.* Acta U. Belii Ser. Math. **10**, 3 – 8.

[9] Grendar M. Jr., Grendar M. (2004). *Maximum entropy method with non-linear moment constraints: challenges.* To appear at proceedings of MaxEnt 23, AIP (Melville).

[10] Grendar M. Jr., Grendar M. (2004). *Maximum probability and maximum entropy methods: bayesian interpretation.* To appear at proceedings of MaxEnt 23, AIP (Melville). Also at arxiv:physics/0308005

[11] Grendar M. Jr., Grendar M. (2004). *Maximum Probability/Entropy translating of contiguous categorical observations into frequencies.* To appear at Appl. Math. Comp. Also available at IDEAS.

[12] Johnson N.L., Kotz S., Balakrishnan N. (1997). *Discrete multivariate distributions.* Wiley (NY).

[13] Le Gall F. (2003). *Determination of the modes of a multinomial distribution.* Statistics & Probability Letters **62**, 325 – 333.

Acknowledgement: Supported by the grant VEGA 1/0264/03 from the Scientific Grant Agency of the Slovak Republic.

Address: M. Grendar, Institute of Mathematics & Computer Science, Severna 5, 974 00 Banska Bystrica, Slovakia & Institute of Measurement Science, Dubravska cesta 9, 841 04, Bratislava, Slovakia

E-mail: `marian.grendar@savba.sk`

COMPSTAT'2004 Symposium

BOOTSTRAPPING FINITE MIXTURE MODELS

Bettina Grün and Friderich Leisch

Key words: Finite mixture models, bootstrapping, identifiability, regression.

COMPSTAT 2004 section: Clustering, Resampling methods.

Abstract: Finite mixture regression models are used for modelling unobserved heterogeneity in the population. However, depending on the specifications these models need not be identifiable, which is especially of concern if the parameters are interpreted. As bootstrap methods are already used as a diagnostic tool for linear regression models, we investigate their use for finite mixture models. We show that bootstrapping helps in revealing identifiability problems and that parametric bootstrapping can be used for analyzing the reliability of coefficient estimates.

1 Introduction

During the last decades it has become popular to include covariates into mixture models leading to mixture regression models. Applications can be found, e.g., in marketing [15] or for clinical trials [13], where unobserved heterogeneity of the population is present in the data and should therefore be taken into account in the modelling process.

The mixture regression models we consider can be formulated by

$$H(y|\mathbf{x}, \Theta) = \sum_{k=1}^{K} \pi_k F(y|\mathbf{x}, \theta_k, \phi_k), \quad 0 < \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

where H is the mixture distribution, $\mathbf{x}$ are the regressors, y the responses, K the number of components, F the component distribution functions, θ_k the regression coefficients, ϕ_k the (possible) dispersion parameters, π_k the prior class probabilities and Θ the vector of all parameters. If the component distribution functions F are from the exponential family the generalized linear modelling framework [7] can be used leading to the so-called GLIMMIX models [14].

A very popular way of estimating mixture models is the EM method, which is a class of iterative algorithms for maximum likelihood estimation in problems with incomplete data [2]. It has been shown that during the classical EM algorithm the values of the likelihood are monotonically increased. The likelihood is in general multimodal with a unique internal global maximum (if the model is identifiable) and several local maxima. Unboundedness of the likelihood might occur at the edge of the parameter space [5]. Then, the solution found by the EM algorithm depends on its initialization.

One possibility to initialize the EM algorithm is to use a partition of the data into the number of requested segments [8]. This partition can be generated randomly or by applying some clustering algorithm, such as, e.g., k-means. In order to ensure that the global maximum is found the EM algorithm is in general run several times with different initializations.

The solution of the EM algorithm does also depend on the given data set. For eliminating random effects of a given data set the results for different samples from the same data generating process (DGP) can be compared. However, in applications there is in general only one data set available. A remedy can be then to draw samples from the empirical distribution of the given data set, i.e., bootstrapping [1]. Furthermore, the parametric bootstrap can be used to assess the stability of the estimated parameters.

2 Identifiability

Finite mixture models are trivially not identifiable with respect to the ordering of the segments and to overfitting, as this leads to empty segments or to several segments having the same parameters. By imposing constraints, e.g., on the ordering of the components, these identifiability problems can be eliminated.

It has been shown that except for these identifiability problems finite mixture distributions of several popular continuous distributions are generically identifiable, as e.g., the (multivariate) normal, gamma and exponential distribution [11], [12], [16]. A discrete identifiable distribution is the Poisson distribution [10]. In contrast the discrete and the continuous uniform distributions are not generically identifiable. The binomial and the multinomial distributions are identifiable if the number of segments is limited with respect to the repetition parameter [11], [3].

In a first analysis of mixture regression models it has been shown that the identifiability of standard linear regression models is not guaranteed even if the regressor matrix has full rank [4], which is also shown by the example given in Section 3.1.1. Furthermore, multinomial mixture regression models were analyzed in [3].

The identifiability of a mixture regression model depends on the distribution of the dependent variable, the maximum number of segments allowed, the available information per object and the regressor matrix. With respect to the regressor matrix identifiability problems might arise if there are only a limited number of different covariate points and if in addition there is only very limited information per person available. Such problems might occur in applications because the covariates are often categorical variables, as, e.g., gender, promotion in marketing, likes/dislikes, test and control group in clinical trials . . . . These variables are in general coded as dummy variables.

3 Simulation

Two examples are presented where bootstrap methods are applied to finite mixture models. A standard linear regression example is used for demonstrating that bootstrap samples can reveal identifiability problems, whereas a Poisson regression example shows that by parametric bootstrapping the stability of the estimated coefficients can be challenged.

Our simulation was performed using the R environment for statistical computing [9]. For EM estimation the contributed package `flexmix` [6] was taken, which implements a general framework for finite mixtures of regression models.

As the EM algorithm might be trapped in local maxima, we always made five initializations with random partitions of the data and considered only the best result with respect to the log-likelihood. Hence, the result of one "run" of the EM algorithm refers to the best result out of these five repetitions.

3.1 Bootstrapping global maxima

We investigate the convergence of the EM algorithm to different global maxima with respect to a simple standard linear mixture regression example with two global maxima.

3.1.1 Normal mixture example Assume we have a standard linear mixture regression with one measurement per object and two different covariate points $\mathbf{x}_1 = (1, 0)'$ and $\mathbf{x}_2 = (1, 1)'$. Furthermore, let the mixture consist of two components with equal prior class probabilities.

The mixture regression can be formulated as

$$H(y|\mathbf{x}, \Theta) = \frac{1}{2}N(\mu_1, 0.1) + \frac{1}{2}N(\mu_2, 0.1)$$

where $\mu_i(\mathbf{x}) = \mathbf{x}'\theta_i$ and $N(\mu, \sigma^2)$ is the normal distribution.

As Gaussian mixture distributions are generically identifiable the means, variances and prior class probabilities are uniquely determined in each covariate point given the mixture distribution. If we assume that $\mu_1(\mathbf{x}_1) = 1$, $\mu_2(\mathbf{x}_1) = 2$, $\mu_1(\mathbf{x}_2) = -1$ and $\mu_2(\mathbf{x}_2) = 4$, the two possible solutions for θ are:

$$\begin{aligned} \theta_1^1 &= (2, \ \ 2)', & \theta_2^1 &= (1, -2)' \quad \text{and} \\ \theta_1^2 &= (2, -3)', & \theta_2^2 &= (1, \ \ 3)' \end{aligned}$$

A balanced sample of length 100 has been generated and can be seen together with the regression lines corresponding to the two different solutions in Figure 1.

3.1.2 Simulation results As the EM algorithm converges nearly always to the same global maximum for a given data set, we eliminate the influence

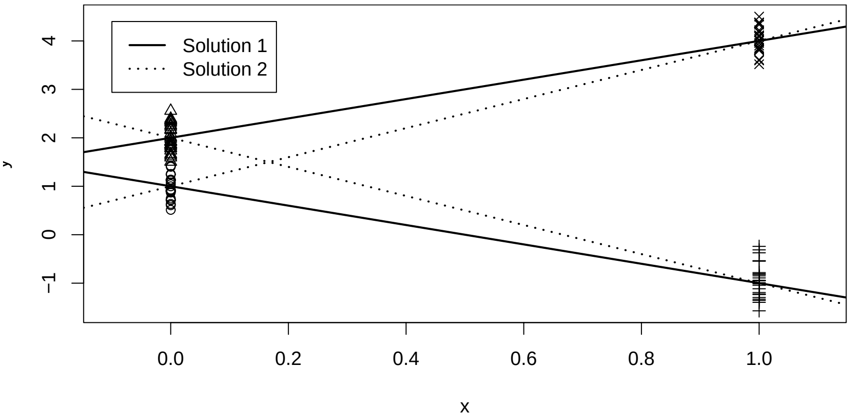


Figure 1: Data and theoretical solutions.

of a given data set by using different samples from the DGP and several bootstrap samples from a given data set. As the unidentifiability depends on the equality of the prior class probabilities we disturb them slightly in order to assess the sensitivity with respect to this parameter.

Thus, we generated 1000 balanced samples of size 100 from the mixture regression model specified in Section 3.1.1 for the DGP analysis. For the bootstrapping analysis we generated 20 samples in the same way and to each data set we added 49 bootstrap samples of the same length. For the sensitivity analysis of the prior class probabilities we applied the same setup except that we used prior class probabilities $(0.6, 0.4)$ and $(0.7, 0.3)$ respectively and the coefficients from solution 1 in Section 3.1.1 for sampling from the DGP. We made 10 runs of the EM algorithm for each sample (from the DGP or bootstrap) in order to show that if the sample is fixed the EM algorithm converges in most of the cases to the same global maximum.

We decided that a result of the EM algorithm is equal to one of the solutions if the maximum distance between the coefficients is less than 0.25 after ordering the components with respect to the intercept which is necessary due to label switching. Then we determined how often which solution was detected and how often during all ten runs only one, none and both of the solutions were found (cp. Table 1).

It can be seen that the results are similar for bootstrapping and sampling from the DGP except that the solutions are always found less often for the BS samples than the DGP samples. Even though this behavior is intuitive, it nevertheless causes that the percentage where none of the solutions are found is considerably higher for the BS samples. If the prior class probabilities are not equal but are 0.6 and 0.4, solution 2 is still found in about 20% of the runs and it is the only solution found during 10 repeated runs for nearly the same percentage. Because of the unequal priors solution 2 is only a local maximum, but obviously the attraction area is large enough that it is relatively often the best solution found in one run of the EM algorithm. The convergence

Priors	**Equal** **DGP**	 **BS**	**0.6/0.4** **DGP**	 **BS**	**0.7/0.3** **DGP**	 **BS**
Overall fraction of						
Solution 1	0.49	0.38	0.74	0.58	0.87	0.68
Solution 2	0.46	0.41	0.20	0.20	0.01	0.04
Fraction over 10 runs of						
Only solution 1	0.45	0.36	0.72	0.56	0.87	0.66
Only solution 2	0.43	0.39	0.18	0.18	0.01	0.04
None of them	0.05	0.20	0.06	0.22	0.12	0.28
Both solutions	0.06	0.04	0.03	0.03	0.00	0.01

Table 1: Simulation results.

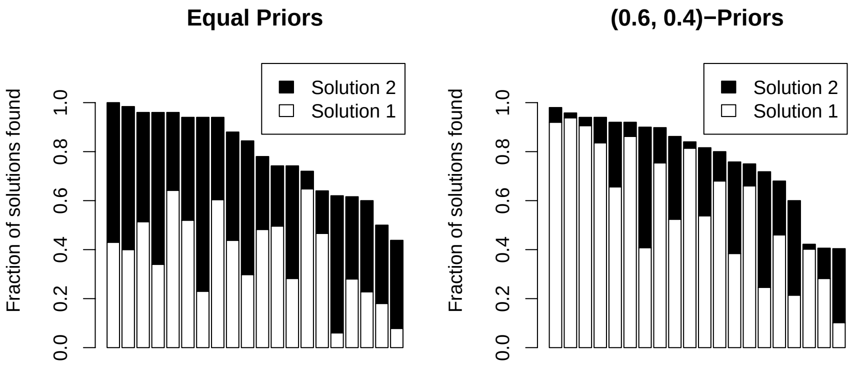


Figure 2: Fraction of only solution 1 or 2 found separately for each of the 20 data sets generated from the DGP when BS.

to solution 2 gets less the more the prior class probabilities deviate from 0.5 such that for $(0.7, 0.3)$ solution 2 is hardly ever found.

In Figure 2 the bootstrap results are analyzed separately for the 20 different data sets generated from the DGP. It can be seen how often which solution was the only one found during 10 runs. Even though the results are varying a lot for the different data sets, it can nevertheless be seen that both valid solutions are found every time by bootstrapping the data for equal priors and that for unequal priors there are data sets, where both solutions are nearly equally often found.

3.2 Parametric bootstrapping

Given a solution bootstrapping from the estimated distribution can be used for assessing the stability of the estimates. Note that we do not intend to replace the standard tools for estimating standard deviations but we propose that by additionally applying the parametric bootstrap it can be assessed if the standard asymptotic theory is appropriate. We use a Poisson mixture regression model to analyze the application of this method to finite mixture models.

3.2.1 Poisson mixture In [13] a Poisson mixture regression is fitted to data from a clinical trial where the effect of intravenous gammaglobulin on suppression of epileptic seizures is investigated. The data used were 140 observations from one treated patient, where treatment has started on the 28th day. In the regression there were three independent variables included: treatment, trend and interaction treatment-trend. Treatment is a dummy variable indicating if the treatment period has already started. Furthermore, the number of parental observation hours per day were available and it is assumed that the number of epileptic seizures per observation hour follows a Poisson mixture distribution. The fitted mixture distribution consisted of two components which can be interpreted as representing 'good' and 'bad' days of the patients.

The mixture model can be formulated by

$$H(y|\mathbf{x}, \Theta) = \pi_1 P(\lambda_1) + \pi_2 P(\lambda_2)$$

where $\lambda_i = e^{\mathbf{x}'\theta_i}$ for $i = 1, 2$ and $P(\lambda)$ is the Poisson distribution.

By reestimating this model we became nearly equal results than in [13]. Our solution for θ_i $i = 1, 2$ with the corresponding standard deviations is:

$$\begin{aligned}
\theta_1 &= (2.84, 1.30, -0.41, -0.43)' \quad \text{with } \mathrm{SD}(\theta_1) = (0.23, 0.47, 0.09, 0.13)' \\
\theta_2 &= (2.07, 7.43, -0.27, -2.28)' \quad \text{with } \mathrm{SD}(\theta_2) = (0.09, 0.52, 0.04, 0.14)'
\end{aligned}$$

The size of the first component representing 'bad' days is 0.28.

3.2.2 Simulation results We generated 100 samples from the estimated mixture distribution with the same structure as the sample used in [13] and applied the EM algorithm to them.

In Figure 3 the theoretical means are given together with the 95% confidence intervals derived with standard asymptotic theory in both plots. The means estimated for the bootstrap samples classified with respect to their prior class probability have been separated. While the confidence intervals for baseline and treatment period are of similar width, it can clearly be seen from the bootstrap application that estimation in the baseline period is much less stable than in the treatment period.

Furthermore, there can be component label switching observed. As the dummy variable treatment is included with its interaction terms and the component sizes are not separated enough, there are solutions which join days with low numbers of seizure episodes during baseline period with those with high numbers during treatment period.

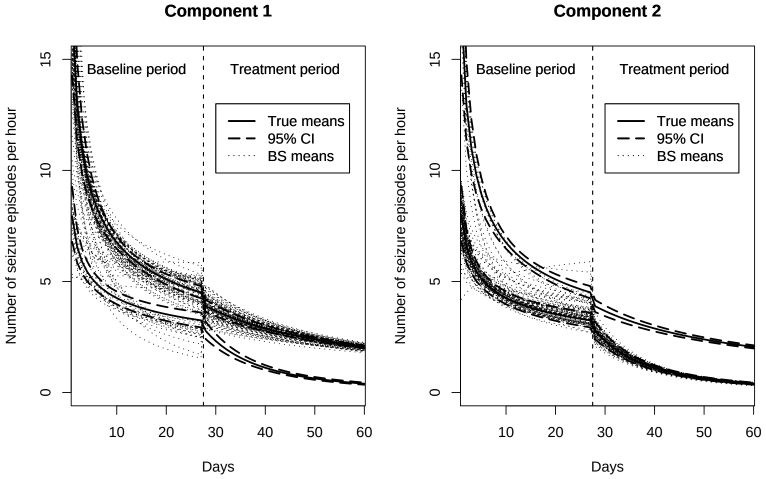


Figure 3: Theoretical means and estimated means of bootstrap samples.

4 Conclusion and future research

We showed that bootstrapping can be a valuable diagnostic tool when estimating finite mixture models as it can reveal identifiability problems and give further insight into the stability of parameter estimates.

Obviously, there exist data sets where the attraction area to one global maximum is that large that only this solution is found by using different initializations to the EM algorithm. Bootstrapping can be used in such a situation for revealing other global maxima. Furthermore, the stability of a solution can be analyzed by parametric bootstrap. In an example it was shown that by including categorical variables with their interaction terms label switching can also occur within components, because there are solutions found where components are joined differently for the different values of the factor due to the flexibility of the estimated model.

As computers are nowadays fast enough to repeat the EM algorithm with different initializations and different input samples within a reasonable amount of time, we recommend to use bootstrapping in addition to the already commonly accepted strategy to use different initializations, in order to ensure a higher stability of the method with respect to the dependency on a certain initialization and data set. Furthermore, by applying the parametric bootstrap additional insights can be gained on the stability of the estimates complementing the results derived with standard asymptotic theory.

References

[1] Davison A.C., Hinkley D.V. (1997). *Bootstrap methods and their application.* Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, Cambridge, UK.

[2] Dempster A., Laird N., and Rubin D. (1977). *Maximum likelihood from incomplete data via the EM-alogrithm.* Journal of the Royal Statistical Society B **39**, 1 – 38.

[3] Grün B. (2002). *Identifizierbarkeit von multinomialen Mischmodellen.* Master's thesis, Technische Universität Wien, Kurt Hornik and Friedrich Leisch, advisors.

[4] Hennig C. (2000). *Identifiability of models for clusterwise linear regression.* Journal of Classification **17**, 273 – 296.

[5] Kiefer N.M. (1978). *Discrete parameter variation: Efficient estimation of a switching regression model.* Econometrica **46** (2), 427 – 434.

[6] Leisch F. (2003). *FlexMix: A general framework for finite mixture models and latent class regression in R.* Report 86, SFB "Adaptive Information Systems and Modeling in Economics and Management Science".

[7] McCullagh P., Nelder J. (1989). *Generalized linear models.* Chapman and Hall.

[8] McLachlan G., Peel D. (2000). *Finite mixture models.* John Wiley and Sons Inc.

[9] R Development Core Team. (2003). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

[10] Teicher H. (1960). *On the mixture of distributions.* Annals of Mathematical Statistics **31**, 55 – 73.

[11] Teicher H. (1963). *Identifiability of finite mixtures.* Annals of Mathematical Statistics **34**, 1265 – 1269.

[12] Titterington D., Smith A., Makov U. (1985). *Statistical analysis of finite mixture distributions.* Chichester: Wiley.

[13] Wang P., Puterman M., Cockburn I., Le N. (1996). *Mixed poisson regression models with covariate dependent rates.* Biometrics **52**, 381 – 400.

[14] Wedel M., DeSarbo W. (1995). *A mixture likelihood approach for generalized linear models.* Journal of Classification **12**, 21 – 55.

[15] Wedel M., Kamakura W. (2001). *Market segmentation — conceptual and methodological foundations.* Kluwer Academic Publishers.

[16] Yakowitz S., Spragins J. (1968). *On the identifiability of finite mixtures.* The Annals of Mathematical Statistics **39** (1), 209 – 214.

Acknowledgement: This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modeling in Economics and Management Science').

Address: B. Grün, F. Leisch, Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Wien, Austria

E-mail: `Bettina.Gruen@ci.tuwien.ac.at,`
`Friedrich.Leisch@ci.tuwien.ac.at`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

AN ALGORITHM FOR OBTAINING STRATA WITH EQUAL COEFFICIENTS OF VARIATION

P. Gunning and J.M. Horgan

Key words: Stratification, skewed population, geometric progression.

COMPSTAT 2004 section: Algorithms.

Abstract: We present an algorithm for the construction of stratum boundaries, which is easier to implement than the commonly used cumulative root frequency method.

1 Introduction

A stratified sampling design partitions a population of size N into L mutually exclusive strata containing N_h elements, $(h = 1, 2, \dots L)$, and samples of size n_h are drawn from each stratum independently.
Stratification differs with respect to:

1. the number L of strata used;
2. how the sample is allocated among the strata;
3. the construction of the stratum boundaries.

The objective is to decide on a stratification strategy to minimise the variance of the estimated population parameter.

In the present paper we concentrate on (3), and develop a new method for the construction of stratum boundaries in skewed populations, which has practical benefits. In what follows we derive the algorithm, and test it on appropriate finite populations.

2 The algorithm

Dalenius [3] derived equations for determining boundaries so that the variance of the mean under optimum allocation is minimised. He pointed out that these equations are troublesome to solve because of dependencies among the components. Cochran [1] observed that with near-optimum boundaries, the coefficients of variation are often found to be approximately the same in all strata. It is on this observation that our algorithm is based.

For any given minimum and maximum data points, k_0 and k_L, we seek stratum breaks $(k_1, \dots, k_{L-1})$ which divide the population into L strata so that the $CV_h = S_h/\overline{X_h}$ are the same for $h = 1, 2, \ldots, L$, where X is the auxiliary variable:

$$\frac{S_1}{\overline{X}_1} = \frac{S_2}{\overline{X}_2} = \ldots = \frac{S_L}{\overline{X}_L} \ . \tag{1}$$

Here S_h is the standard deviation, and $\overline{X}_h$ the mean in stratum h:

$$S_h = \sqrt{\frac{\sum_{i=1}^{N_h}(X_{hi} - \overline{X}_h)^2}{N_h - 1}} \ , \ \overline{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} X_{hi} \ . \tag{2}$$

If we make the assumption that the distribution within each stratum is approximately uniform we may write

$$\overline{X}_h \approx \frac{k_h + k_{h-1}}{2}, \tag{3}$$

$$S_h \approx \frac{1}{\sqrt{12}}(k_h - k_{h-1}). \tag{4}$$

As an approximation to coefficients of variation, this gives

$$CV_h \approx \frac{(k_h - k_{h-1})/\sqrt{12}}{(k_h + k_{h-1})/2} \ . \tag{5}$$

With equal CV_h we have

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}} \ . \tag{6}$$

This recurrence relation reduces however to something familiar:

$$k_h^2 = k_{h+1}k_{h-1} \ . \tag{7}$$

The stratum boundaries are the terms of a geometric progression:

$$k_h = ar^h \ (h = 0, 1, \ldots, L) \tag{8}$$

In particular $a = k_0$, the minimum value of the variable, and $ar^L = k_L$, the maximum value of the variable. It follows that the constant ratio can be calculated as $r = (k_L/k_0)^{1/L}$.

For a numerical example take

$$L = 4 \ ; \ k_0 = 5 \ ; \ k_4 = 50{,}000 \ : \tag{9}$$

thus $k_h = 5.10^h$ $(h = 0, 1, 2, 3, 4)$ and the strata form the ranges

$$5 - 50; \ 50 - 500; \ 500 - 5{,}000; \ 5{,}000 - 50{,}000. \tag{10}$$

Clearly this is an extremely simple method of obtaining stratum breaks.

2.1 Justification

The relationship in (7) depends critically on the assumption that the distributions within stratum are uniformly (3),(4) distributed. This assumption may be justified for skewed populations by the following heuristic argument. When the parent distribution is positively skewed, then the low values of the variable have a high incidence, and the incidence decreases as the variable values increase. It is therefore appropriate to take small intervals at the beginning and large intervals at the end. This is what happens with a geometric series of constant ratio greater than one. In the lower range of the variable, the strata are narrow so that an assumption of a rectangular distribution in them is not unreasonable. As the value of the variable increases, the stratum width increases geometrically. This coincides with the decreased rate of change of the incidence of the positively skewed variable, so here also the assumption of uniformity is reasonable.

3 Performance of the algorithm

To test our algorithm, we implement it on four positively skewed populations. Our first, Population 1, is an accounting population of debtors in an Irish firm, detailed in Horgan [5]. An audit often consists of a take-all stratum of very large items that are audited on an 100% basis, and a take-none stratum of very small items that are not audited at all. In our case the take-all stratum consists of all items over 28,000 euro, and the take-none stratum consists of all items under 40 euro. We use the remainder of the population, which has a skewness coefficient of 6.44 to test our algorithm.

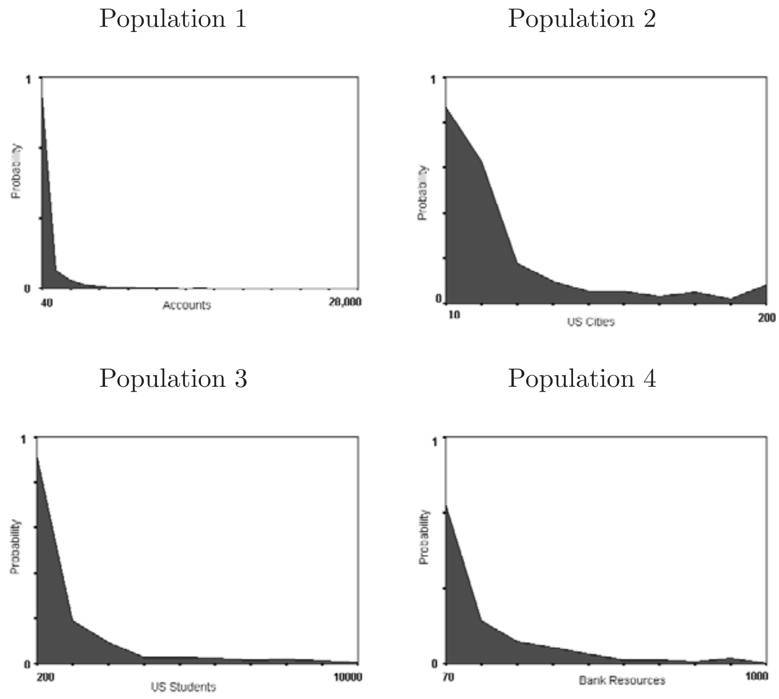
In addition, we take three of the populations from Cochran [1]:

The population of US cities, in thousands (Population 2);
The number of students in four-year US colleges (Population 3);
The resources of a large commercial bank in the US, in millions of dollars (Population 4).

All four populations are illustrated in Figure 1 and summarised in Table 1 in order of decreasing skewness.

3.1 Stratum boundaries: A comparison with the cumulative root frequency method

The most frequently used method for obtaining the stratum boundaries is referred to as the cumulative root frequency ($cum\sqrt{f}$), and was developed by Dalenius and Hodges [4] as an approximation to the optimum solution. This method involves dividing the populations into numerous classes, obtaining the root of the frequency in each class, adding the root of the frequencies, and obtaining strata so that there are equal intervals on the cumulative root frequencies. It is clear that the geometric method, outlined in (8), is much easier to implement than this.


Population 1

Population 2

Population 3

Population 4

Figure 1: Populations.

Population	N	Range	Skewness	Mean	Variance
1	763	40-28,000	6.44	838.64	3,511,827
2	1,038	10-200	2.88	32.57	924
3	677	200-10,000	2.46	1563.00	3,236,602
4	357	70-1,000	2.08	225.62	36,274

Table 1: Summary statistics for four real populations.

A comparison of the two methods is carried out in Tables 2 and 3, where the four populations summarised in Table 1 are divided into 4 and 5 strata using both methods of obtaining the breaks.

Inspection of the coefficients of variation in Tables 2 and 3 suggests that the geometric method is more successful than the cumulative root frequency method in obtaining near-equal strata CV_h in most cases. For example, in

Popul.	Stratif.		Stratum			
	Method		1	2	3	4
1	Cum $\sqrt{f}$	Range	40-558	559-1117	1118-2795	2796-28000
		% of Pop.	69%	14%	10%	7%
		Coeff. of Var.	.70	.19	.27	.69
	Geometric	Range	40-205	206-1057	1058-5443	5444-28000
		% of Pop.	42%	41%	14%	3%
		Coeff. of Var.	.45	.44	.48	.50
2	Cum $\sqrt{f}$	Range	10-19	20-38	39-85	86-200
		% of Pop.	35%	41%	15%	6%
		Coeff. of Var.	.20	.17	.25	.26
	Geometric	Range	10-20	21-43	44-93	94-200
		% of Pop.	44%	38%	13%	5%
		Coeff. of Var.	.22	.20	.22	.22
3	Cum $\sqrt{f}$	Range	200-689	690-2,159	2,160-5,099	5,100-10,000
		% of Pop.	35%	47%	11%	7%
		Coeff. of Var.	.31	.33	.29	.19
	Geometric	Range	200-526	527-1,386	1,387-3,653	3,654-10,000
		% of Pop.	20%	51%	19%	10%
		Coeff. of Var.	.27	.26	.26	.27
4	Cum $\sqrt{f}$	Range	70-162	163-255	256-488	489-1,000
		% of Pop.	58%	16%	10%	10%
		Coeff. of Var.	.23	.11	.18	.24
	Geometric	Range	70-134	135-261	262-504	505-1,000
		% of Pop.	44%	30%	18%	8%
		Coeff. of Var.	.18	.19	.19	.20

Table 2: Boundaries with 4 Strata.

Population 1 the CV_h differ substantially from each other when the cumulative root frequency method has been used to make the breaks, while the geometric method appears to achieve near equal CV_h in all cases. The homogeneity of CV_h between strata is better when $L = 5$ than when $L = 4$; this is to be expected since the validity of the assumption of uniformity of the distribution of elements within stratum is strengthened with increased number of strata.

3.2 Relative efficiency

One of the objectives of stratification is to reduce the variance of the sample estimator. In this section we examine how the geometric method of stratum construction compares with the cumulative root frequency method in terms of the variance of the stratum mean with a sample of size $n = 100$ allocated optimally. The comparison is in terms of the relative efficiency or variance ratio:

$$eff(\overline{x}_{st}) = \frac{V_{geom}(\overline{x}_{st})}{V_{cum}(\overline{x}_{st})} \ , \qquad (11)$$

Popul.	Stratif.				Stratum		
	Method		1	2	3	4	5
1	Cum $\sqrt{f}$	Range	40-279	280-838	839-1677	1678-4193	4194-28000
		% of Pop.	49%	30%	10%	7%	4%
		Coeff. of Var.	0.52	0.30	0.20	0.25	0.57
	Geometric	Range	40-147	148-549	550-2037	2038-7552	7553-28000
		% of Pop.	31%	37%	22%	8%	2%
		Coeff. of Var.	0.37	0.38	0.40	0.37	0.41
2	Cum $\sqrt{f}$	Range	10-28	29-38	39-57	58-104	105-200
		% of Pop.	70%	9%	9%	8%	4%
		Coeff. of Var.	0.28	0.08	0.11	0.16	0.16
	Geometric	Range	10-17	18-32	33-59	60-108	109-200
		% of Pop.	35%	40%	13%	8%	4%
		Coeff. of Var.	0.18	0.14	0.15	0.16	0.15
3	Cum $\sqrt{f}$	Range	200-1179	1180-1669	1670-3139	3140-6079	6080-10,000
		% of Pop.	67%	7%	14%	7%	5%
		Coeff. of Var.	0.40	0.09	0.20	0.19	0.13
	Geometric	Range	200-433	434-941	942-2043	2044-4434	4435-10,000
		% of Pop.	14%	38%	29%	11%	8%
		Coeff. of Var.	0.22	0.21	0.24	0.21	0.21
4	Cum $\sqrt{f}$	Range	70-162	163-255	256-395	396-627	628-1,000
		% of Pop.	58%	16%	10%	10%	6%
		Coeff. of Var.	0.23	0.11	0.10	0.13	0.11
	Geometric	Range	70-118	119-200	201-339	340-576	577-1,000
		% of Pop.	32%	32%	18%	11%	7%
		Coeff. of Var.	0.14	0.14	0.17	0.12	0.16

Table 3: Boundaries with 5 strata.

where $V_{geom}(\overline{x}_{st})$ and $V_{cum}(\overline{x}_{st})$ are the variances of the mean respectively with the geometric and the cumulative root frequency method.

The variance calculations are based on the auxiliary variable x, and since this is assumed to be highly correlated with the unknown survey variable y, we anticipate that $eff(\overline{y}_{st}) \approx eff(\overline{x}_{st})$. Table 4 gives the variance ratio when the number of strata L = 4 and 5.

	Population			
Strata	1	2	3	4
4	1.02	.81	1.00	.90
5	.85	1.06	.60	.75

Table 4: Relative efficiencies.

From Table 4 we see that in all except three cases, $eff(\overline{x}_{st}) < 1$, implying that, in these cases the new method is more efficient that the cumulative root frequency method of stratum construction. Large gains in efficiency are observed with the new method when $L = 5$ in Populations 1, 3 and 4: here the relative efficiencies are .85, .60 and .75 respectively. The relative efficiencies that are greater than 1 are only marginally so, 1.02 and 1.06 in Population 1 with 4 strata, and Population 2 with 5 strata respectively. In Population 3, $eff(\overline{x}_{st}) = 1$ when $L = 4$, indicating that the variance are the same for both methods of stratification.

4 Summary

This paper derives a simple algorithm for the construction of stratum boundaries in positively skewed populations so that the coefficients of variation are approximately equal in all strata. We show that near equality of the CV_h may be achieved, with positively skewed populations, by using the geometric progression to make the breaks. The algorithm was implemented on four real populations divided into four and five strata, and compared with the cumulative root frequency method of stratum construction. Substantial gains in the precision of the estimator of the mean with the proposed method of obtaining strata boundaries compared with the cumulative root frequency method were observed in nearly all cases; the greatest gains occurred when the number of strata was five.

References

[1] Cochran W.G. (1961). *Comparison of methods for determining stratum boundaries*. Bulletin of the International Statistical Institute **32**, 345–358.

[2] Cochran W.G. (1977). *Sampling techniques* 3rd Edition, New York:Wiley.

[3] Dalenius T. (1950). *The problem of optimum stratification*. Skandinavisk Aktuarietidskrift, 203–213.

[4] Dalenius T., Hodges J.L. (1957). *The choice of stratification points*. Skandinavisk Aktuarietidskrift, 198–203.

[5] Horgan J.M. (2003). *A list sequential sampling scheme with applications in financial auditing*. IMA Journal of Management Mathematics **14**, 1–18.

Acknowledgement: This work was supported by a grant from the Irish Research Council for Science, Engineering and Technology.

Address: P. Gunning, J.M. Horgan, Dublin City University, Glasnevin, Dublin 9, Ireland

E-mail: `pgunning@computing.dcu.ie, jhorgan@computing.dcu.ie`

COMPSTAT'2004 Symposium

SCHWARZ INFORMATION CRITERION IN THE PRESENCE OF INCOMPLETE-DATA

B. Hafidi and A. Mkhadri

Key words: SIC, MDL, Missing-data, Model selection, EM algorithm, Regression.

COMPSTAT 2004 section: Model selection.

Abstract: This note is concerned with an evaluation of a Schwarz information (SIC) criterion, one of the most widely known and used tools in statistical model selection. The SIC criterion serves as an asymptotic approximation to a transformation of the Bayesian posterior probability of candidate model. However, the latter is based on the observed-data empirical log-likelihood which may be problematic to compute in the presence of incomplete-data. We derive and investigate a variant of SIC criterion, for model selection in the settings where the observed-data is incomplete. We examine the performance of our criterion relative to other well known criteria in a large simulation study based on multiple and multivariate regression modeling.

1 Introduction

The selection of an appropriate model from a potentially large class of candidate models is an issue that is central in statistical modeling. This selection can be often be facilitated through the use of an information theoretic criteria, which assigns a score to every fitted model in a candidate class based on some underlying statistical principle.

The Schwarz information criterion (SIC) introduced by Schwarz [17] as a competitor to the Akaike [1], [2] information (AIC) criterion, is one of the most popular and effective of the criteria used for model selection. Its original derivation is based on establishing that the criterion serves as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. Akaike [3] and Schwarz [17] introduced equivalent consistent model selection criteria from a Bayesian perspective. Schwarz derived SIC for the case of independent identically distributed observations, under the assumption that the likelihood is from the regular exponential family. While, Akaike derivation is outlined for model selection in linear regression.

Moreover, Haughton [9] extends SIC to a context where the likelihood is from the curved exponential family. Recently, Neath and Cavanaugh [15] suggested several SIC variants based on the inclusion of other terms in the criterion, which are asymptotically negligible, and suggest that the applicability of SIC extends to a very wide range of modeling settings. Cavanaugh and Neath [7] present a derivation which does not require that the likelihood has any particular form, but only assumes that it satisfies a set of non restrictive regularity conditions. Additional generalization of Schwarz's derivation are considered by Stone [19], Leonard [13], Kashyap [10] and Kass [11]. But, a rigorous generalization of Schwarz's development seems to be lacking from the literature (cf. [12, p. 779]).

Furthermore, the SIC criterion is equivalent to Rissanen's minimum discreption length (MDL) criterion (1989). But, the derivation and justification of MDL are difficult to follow without a strong background in coding theory. The SIC criterion and its variants are based on the observed-data empirical log-likelihood, which may be problematic to compute in a large variety of practical problems in the presence of missing-data. In contrast, this computation is often straightforward using the complete-data tools. In settings where the observed-data are incomplete, Shimodaira [18] proposed a naturel extension of AIC (PDIO) and Cavanaugh and Shumway [6] derived a variant of AIC (AICcd). These criteria can be evaluated using only complete-data tools, readily available through the EM (expectation, maximization) algorithm [8] and the Supplemented EM algorithm [14]. Moreover, Bueso, Qian and Angulo [5] proposed a variant of MDL for model selection in the presence of missing-data. Biernacki, Celeux and Govaert [4] presented an integrated completed likelihood (ICL) criterion to assess a mixture model in a cluster analysis setting. The SIC criterion provides, under regularity conditions, a reliable approximation to the integrated likelihood. Although, these conditions for SIC do not hold for assessing the number of components in mixture model, there is an increasing practical support for its use in this context (cf. [4]).

In this note, we derive and investigate a variants of SIC, for model selection in the presence of incomplete-data. In Section 2, we briefly describe model selection based on SIC. We derive our criterion SIC_{cd} and its variants in Section 3. In Section 4, we compare the performance of our criteria to SIC and its variants, MDL and MDL_c in a large simulation study involving, multiple and multivariate regression modeling. We ends this note with a small discussion.

2 Schwarz information criterion (SIC)

Let $\mathbf{Y}_o$ the vector of observed-data or incomplete-data of dimension n_o. Assuming that $\mathbf{Y}_o$ is to be described using a model M_k selected from a set of candidate models M_1, M_2,..., M_L. Assumed that each model M_k $(1 < k < L)$ is characterized by the probability density $f_k(\mathbf{Y}_o|\theta_k)$, where θ_k is an element of the parameter space $\Theta(k)$. Assume that the derivatives of $f_k(\mathbf{Y}_o|\theta_k)$ up the order three exist with respect to θ_k, and continues and suitably bounded for all $\theta_k \in \Theta(k)$.

Let $\pi(\theta_k|k)$ $(1 < k < L)$ denote a prior distribution for parameter vector θ_k under model M_k. Let $Pr(M_k)$ the prior probability for model M_k. Also, let $\hat{\theta}_k$ denote the estimator of θ_k obtained by maximizing the likelihood $f_k(\mathbf{Y}_o|\theta_k)$ over $\Theta(k)$, and d_k denote the dimension of M_k.

The Bayes approach for model selection is to choose the model M_k with the posterior probability among a set of candidate models. The posterior probability of the model M_k is given by

$$Pr(M_k|\mathbf{Y}_o) = h(\mathbf{Y}_o)^{-1}Pr(M_k)\int f_k(\mathbf{Y}_o|\theta_k)\pi(\theta_k|k)d\theta_k,$$

where $h(\mathbf{Y}_o)$ denotes the marginal distribution of $\mathbf{Y}_o$.

This quantity is not possible to evaluate directly. However, Schwarz [17] showed that the criterion

$$SIC = -2\ln f_k(\mathbf{Y}_o|\hat{\theta}_k) + d_k \ln n_o$$

provides a large sample approximation to

$$-2\ln Pr(M_k|\mathbf{Y}_o) - 2\ln h(\mathbf{Y}_o), \tag{1}$$

where $f_k(\mathbf{Y}_o|\hat{\theta}_k)$ represent the incomplete-data empirical likelihood.

Moreover, Neath and Cavanaugh [15] proposed other variants of SIC based on the inclusion of some terms which are discarded in the original derivation of the SIC criterion and its subsequent extensions, as being asymptotically negligible. These variants are defined by
$SIC_f = SIC + \ln|\mathbf{I}_{ob}(\hat{\theta}_k|\mathbf{Y}_o)|$; $SIC_p = SIC - 2\ln Pr(M_k)$ and
$SIC_{fp} = SIC + \ln|\mathbf{I}_{ob}(\hat{\theta}_k|\mathbf{Y}_o)| - 2\ln Pr(M_k)$, where
$\mathbf{I}_{ob}(\theta_k|\mathbf{Y}_o) = \frac{-1}{n_o}\frac{\partial^2 \ln f_k(\mathbf{Y}_o|\theta_k)}{\partial\theta_k\partial\theta_k^t}$ is the observed Fisher information matrix.

Both, criteria are based on the observed-data empirical log-likelihood which may be problematic to compute in a large variety of practical problems in the presence of missing-data. In contrast, the evaluation of this quantity is often straightforward using the complete-data tool by the EM algorithm [8].

In the next section, we propose a version of SIC for model selection in settings where the observed-data is incomplete. We derive this version by considering the expected of (1) when replace $\mathbf{Y}_o$ by $\mathbf{Y}$ the vector of complete-data, with respect to density of the missing-data given the observed-data.

3 SIC for complete-data

In this section, we assume that the vector of complete-data has the following form $\mathbf{Y} = (\mathbf{Y}_o, \mathbf{Y}_m)$ of dimension n, where $\mathbf{Y}_o$ is the vector of the observed-data, and $\mathbf{Y}_m$ is the vector of the missing-data. In this setting, the complete-data density $f_k(\mathbf{Y}|\theta_k)$, for model M_k, is composed of the product of the incomplete-data density $f_k(\mathbf{Y}_o|\theta_k)$ and the conditional density of the missing-data $\mathbf{Y}_m$ given the incomplete-data $\mathbf{Y}_o$;i.e.

$$f_k(\mathbf{Y}|\theta_k) = f_k(\mathbf{Y}_o|\theta_k)f_k(\mathbf{Y}_m|\mathbf{Y}_o, \theta_k). \tag{2}$$

We will assume that the fitted parameter $\hat{\theta}_k$ is obtained by using the EM algorithm [8] over $\Theta(k)$.

Let

$$\begin{aligned} Q_k(\theta_1|\theta_2) &= \int_{\mathbf{Y}_m} \{\ln f_k(\mathbf{Y}|\theta_1)\} f_k(\mathbf{Y}_m|\mathbf{Y}_o,\theta_2)\mathbf{dY}_m, && (3)\\ \mathbf{I}_{oc}(\theta_k|\mathbf{Y}_o) &= \int_{\mathbf{Y}_m} \left\{\frac{-1}{n}\frac{\partial^2 \ln f_k(\mathbf{Y}|\theta_k)}{\partial\theta_k\partial\theta_k^t}\right\} f_k(\mathbf{Y}_m|\mathbf{Y}_o,\theta_k)\mathbf{dY}_m. && (4)\end{aligned}$$

The posterior probability of the model M_k for complete-data is given by

$$Pr(M_k|\mathbf{Y}) = h(\mathbf{Y})^{-1})Pr(M_k)\int f_k(\mathbf{Y}|\theta_k)\pi(\theta_k|k)d\theta_k,$$

Now, we consider minimizing

$$\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)}\{-2\ln Pr(M_k|\mathbf{Y})\}, \qquad (5)$$

where $\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)}$ denotes the expectation under $f_k(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)$.
We have

$$-2\ln Pr(M_k|\mathbf{Y}) = 2\ln h(\mathbf{Y}) - 2\ln Pr(M_k) - 2\ln\int f_k(\mathbf{Y}|\theta_k)\pi(\theta_k|k)d\theta_k \quad (6)$$

The first term on the right-hand side of (6) can be discarded, since it is constant with respect to k. Therefore, we have the proportionality

$$-2\ln Pr(M_k|\mathbf{Y}) \quad\propto\quad -2\ln Pr(M_k) - 2\ln\int f_k(\mathbf{Y}|\theta_k)\pi(\theta_k|k)d\theta_k$$

As Neath and Cavanaugh [15], using Laplace's methods for integrals in the Bayesian framework developed by Tierney and Kadane (1986) and Kass and Rafferty [12], we have under the noninformative prior $\pi(\theta_k|k) = 1$, an asymptotic approximation

$$-2\ln Pr(M_k|\mathbf{Y}) \approx -2\ln f_k(\mathbf{Y}|\hat{\theta}_k) + d_k\ln(\frac{n}{2\pi}) + \ln|\mathbf{I}_c(\hat{\theta}_k|\mathbf{Y})| - 2\ln Pr(M_k), \qquad (7)$$

where $\mathbf{I}_c(\theta_k|\mathbf{Y}) = \frac{-1}{n}\frac{\partial^2 \ln f_k(\mathbf{Y}|\theta_k)}{\partial\theta_k\partial\theta_k^t}$. We can therefore rewrite (5) as follows

$$\begin{aligned}&\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)}\big\{-2\ln Pr(M_k|\mathbf{Y})\big\} = && (8)\\ &\quad \mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)}\big\{-2\ln f_k(\mathbf{Y}|\hat{\theta}_k) + d_k\ln(\frac{n}{2\pi}) + \ln|\mathbf{I}_c(\hat{\theta}_k|\mathbf{Y})| - 2\ln Pr(M_k)\big\}.\end{aligned}$$

We have

$$\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)}\Big\{-2\ln f_k(\mathbf{Y}|\hat{\theta}_k)\Big\} \quad=\quad -2Q_k(\hat{\theta}_k|\hat{\theta}_k). \qquad (9)$$

The evaluation of $\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)} \ln|\mathbf{I}_c(\hat{\theta}_k|\mathbf{Y})|$ is not easy in general. In the next section, We will evaluated this expression for regression models. Now, substituting (9) on the second term of (8), we obtain

$$\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)}\Big\{-2\ln Pr(M_k|\mathbf{Y})\Big\} = \tag{10}$$
$$-2Q_k(\hat{\theta}_k|\hat{\theta}_k) + d_k\ln(\frac{n}{2\pi}) + \mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)} \ln|\mathbf{I}_c(\hat{\theta}_k|\mathbf{Y})| - 2\ln Pr(M_k).$$

As Neath and Cavanaugh [15], ignoring terms in the preceding expression that are bounded as the sample size grows to infinity, we obtain a criterion

$$SIC_{cd} \approx -2Q_k(\hat{\theta}_k|\hat{\theta}_k) + d_k\ln n. \tag{11}$$

By analogy with the variants of SIC for observed-data, we obtain other variants of SIC_{cd} which are defined by

$$\begin{aligned} SIC_{cd}^{f} &= SIC_{cd} + \mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)} \ln|\mathbf{I}_c(\hat{\theta}_k|\mathbf{Y})| \\ SIC_{cd}^{p} &= SIC_{cd} - 2\ln Pr(M_k) \\ SIC_{cd}^{fp} &= SIC_{cd} + \mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)} \ln|\mathbf{I}_c(\hat{\theta}_k|\mathbf{Y})| - 2\ln Pr(M_k). \end{aligned}$$

The first term of both criteria is based on the complete-data function $Q_k(\hat{\theta}_k|\hat{\theta}_k)$, which is the principal tool used by the EM algorithm. The evaluation of this term is easy after the last iteration of the EM algorithm. In contrast, we can't said the same thing for the first term of SIC and its variants, which is based on the incomplete-data log-likelihood $\ln f_k(\mathbf{Y}_o|\hat{\theta}_k)$. We should also note that the computation of SIC_{cd}, SIC_{cd}^{f}, SIC_{cd}^{p} and SIC_{cd}^{fp} involves only complete-data quantities which arise in the execution of the EM algorithm.

4 Numerical experiments

We carried out a fairly large simulation study of the performance of the SIC_{cd},SIC_{cd}^{f}, SIC_{cd}^{p} and SIC_{cd}^{fp} criteria against SIC,SIC_f, SIC_p, SIC_{fp} MDL, and MDL$_c$ criteria. This simulation study focuses on two important modeling frameworks: the multiple and multivariate regression models.

4.1 Multiple regression

One of the most important problems of model selection is the multiple regression problem defined by

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{Y}$ is an nx1 observation vector, $\mathbf{X}$ is a known nxp design matrix, β is a px1 vector and ϵ is a Gaussian random vector with zero mean and variance-covariance matrix equal to $\sigma^2\mathbf{I}$. We assume that the candidate models (for

$p = 1$ to $p = 8$) are nested. This corresponds to practical setting where the predictor variables can be listed in some order of importance.

We present the first example considered by Neath and Cavanaugh [15] to investigate the performance of the Schwarz information criteria.

For each example, 1000 samples of size $n = 20$ are generated from a model $\mathbf{Y}_0 = \mathbf{X}_0\beta_0 + \epsilon$, with dimension $p_0 = 4$. In the first example of our simulations, the values of the first column of $\mathbf{X}$ are fixed equal to 1. The values of the other columns are generated from uniform distribution on the interval $(0, 6)$. The p_0x1 parameter vector β_0 has all elements set equal to 1. Moreover, in some sets, certain data within each sample is made incomplete by eliminating, according to specified discard probabilities, some observations in $\mathbf{Y}_0$. The probability of missing-data is set at 0, 0.15, 0.30 and 0.40 respectively.

As Neath and Cavanaugh [15] we use for the prior $Pr(M_k)$ a poisson distribution where the mean is set equal to the true number of predictor variables in the true model. The distribution is truncated at 8.

For each of the 1000 samples in a set, all eight models in the candidate class are fit to the data using EM algorithm. Over the 1000 data sets, the selections are summarized in Tables 1 and 2. In these tables, the discard probability for the sample are listed in the first column, and the dimension of generating model is listed in the second column.

In this setting the evaluation of $\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)} \ln |\mathbf{I}_c(\hat{\theta}_k|\mathbf{Y})|$ is easy and it's equal to

$$-(p+1)\ln\hat{\sigma}^2 - p\ln n + \ln|\mathbf{X}^{'}\mathbf{X}| - \ln 2$$

Using this expression for criteria with missing data, the dimension selections are grouped in three categories:"$< p_0$"(underfitting), "p_0"(correct dimension), and "$> p_0$"(overfitting).
When there are none missing-data, SIC, SIC_f, SIC_p, SIC_{fp} are equivalent to SIC_{cd}, SIC^f_{cd}, SIC^p_{cd} and SIC^{fp}_{cd} respectively. Therefore, each two criteria yield the same selection results. The same for MDL and MDL_c. In this setting, both criteria SIC_f, SIC_{fp} outperforms their counterparts that assume a uniform prior (SIC, SIC_p). The same remark is pointed out by Neath and Cavanaugh [15]. But the SIC_f, SIC_{fp} and MDL greatly outperforms SIC, SIC_p.

When the discard probabilities are increased, the SIC_{fp} criterion outperforms all other criteria, followed by SIC_f, MDL and then SIC^{fp}_{cd}, SIC^f_{cd}, and MDL_c.

As the discard probabilities are increased, SIC_{fp}, SIC_f, and MDL never exhibit any tendency, while the other criteria, particularly SIC, SIC_p, SIC_{cd} and SIC^p_{cd}, becomes more prone towards selecting greater dimensional in set 4. Thus, all criteria for complete-data tends to overfit to a strong degree than their correspondent criteria for the observed-data. As SIC_f and SIC_{fp}, in this setting SIC^f_{cd} and SIC^{fp}_{cd} outperform their counterparts that assume a uniform prior (SIC_{cd} ,SIC^p_{cd}).

		Criteria									
Pr(y$_{mis}$)	dim	MDL	MDL$_c$	SIC	SIC$_f$	SIC$_p$	SIC$_{fp}$	SIC$_{cd}$	BIC$^f_{cd}$	BIC$^p_{cd}$	BIC$^{fp}_{cd}$
	< 4	0	0	0	0	0	0	0	0	0	0
0.00	$= 4$	977	977	771	986	858	990	771	986	858	990
	> 4	23	23	229	14	142	10	229	14	142	10
	< 4	0	0	0	0	0	0	0	0	0	0
0.15	$= 4$	981	963	701	989	797	992	668	978	752	984
	> 4	19	37	299	11	203	8	332	22	248	16
	< 4	0	0	0	0	0	0	0	0	0	0
0.30	$= 4$	984	913	619	991	730	997	540	953	629	964
	> 4	16	87	281	9	270	3	460	47	271	36
	< 4	0	0	0	0	0	0	0	0	0	0
0.40	$= 4$	979	831	540	989	656	993	389	894	507	911
	> 4	21	169	460	11	344	7	611	106	493	89

Table 1: Selected dimension for multiple regression, $n = 20$ and $\sigma = 0.25$.

4.2 Multivariate regression

Another important setting of application of models selection is the multivariate regression model defined by $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$, where the rows of $\mathbf{Y}_{n\times p}$ correspond to p response variables on each of n individuals. $\mathbf{X}_{n\times m}$ is a known matrix of covariate values, and $\beta_{m\times p}$ is a matrix of unknown regression parameters. The rows of the error matrix $\mathbf{U}_{n\times p}$ are assumed to be independent, with identical $\mathcal{N}_p(0, \Sigma)$ distribution with $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$. The number of unknown parameters in this setting is $d = pm + 0.5p(p + 1)$.

We consider a setting where $p = 2$, so that the rows of $\mathbf{Y}$ represents bivariate data pairs. There were eight candidate models stored in an nx8 matrix $\mathbf{X}$, with a column of ones, followed by seven columns of independent measurements on a random variable having an uniform distribution on the interval (0,5). The candidate models include the columns of $\mathbf{X}$ in a sequentially nested fashion; that is, columns 1 to m define the design matrix for the candidate model with m covariates. The design matrix for the true model was the first m_0 columns of $\mathbf{X}$. One thousand sets of data are generated from a model having the form $\mathbf{Y}_0 = \mathbf{X}_0\beta_0 + \mathbf{U}$, where $\sigma_{11} = 4$,$\sigma_{22} = 16$ and $\sigma_{12} = \sigma_{21} = 7$. For our simulation, we chose $m_0 = 3$. All elements of the m_0x2 parameter matrices β_0 are fixed equal to 1. We consider a collection of five simulation sets are run with the pair of discard probabilities (Pr(y$_{1mis}$), Pr(y$_{2mis}$)) set at $(0.00, 0.00)$, $(0.00, 0.60)$, $(0.20, 0.40)$, $(0.40, 0.20)$, $(0.60, 0.00)$.

As in multiple regression, the evaluation of $\mathbf{E}_{(\mathbf{Y}_m|\mathbf{Y}_o,\hat{\theta}_k)} \ln |\mathbf{I}_c(\hat{\theta}_k|\mathbf{Y})|$ is easy and it's equal to

$$-(m + p + 1)\ln|\hat{\Sigma}| - mp\ln n + p\ln|\mathbf{X}^{'}\mathbf{X}| - \frac{p(p + 1)}{2}\ln 2.$$

The selected dimensions are grouped in three categories:"$< d_0$"(underfitting), "d_0"(correct dimension), and "$> d_0$"(overfitting). Over the 1000 data sets, the selections are summarized in Table 2.

Pr(y$_1$ mis),	True	Criteria									
Pr(y$_2$ mis)	Dim	MDL	MDL$_c$	SIC	SIC$_f$	SIC$_p$	SIC$_{fp}$	BIC$_{cd}$	BIC$^f_{cd}$	BIC$^p_{cd}$	BIC$^{fp}_{cd}$
0.00,0.00	< 9	0	0	0	0	0	0	0	0	0	0
	$= 9$	816	861	970	956	980	978	970	956	980	991
	> 9	139	139	30	44	20	22	30	44	20	9
0.00,0.60	< 9	1	2	2	5	5	9	3	3	4	6
	$= 9$	897	505	852	941	915	968	641	677	738	783
	> 9	102	493	143	54	80	23	356	320	358	216
0.20,0.40	< 9	2	0	6	11	5	9	5	5	6	4
	$= 9$	894	522	875	949	926	965	671	721	772	820
	> 9	104	478	119	40	69	26	324	274	222	176
0.40,0.20	< 9	5	0	11	13	9	10	5	2	5	4
	$= 9$	889	537	844	938	907	960	669	714	756	797
	> 9	106	463	145	49	84	30	326	284	239	199
0.60,0.00	< 9	22	2	36	68	37	57	11	10	13	12
	$= 9$	869	486	817	883	884	912	621	663	725	753
	> 9	109	512	147	49	79	31	368	327	262	235

Table 2: Selected dimension for multivariate regression, $n = 50$.

We have obtained the same results as in the precedent example. But, the correct selection rates decline slightly from the first example.

We have considered other example of simulation (not reported here), where the value of sample size is set to $n = 50, n = 100, n = 200$. We have obtained that the Sic criterion and its variants are approximatively equivalent to the SIC_{cd} and its variants.

5 Conclusion

In this note, we have derived the Schwarz information criterion for model selection, SIC_{cd} in application where the observed-data is incomplete. Our criterion is an asymptotic approximation to the conditional expected of a transformation of the Bayesian posterior probability of a candidate model. As SIC, we have proposed corrected variants of SIC_{cd} which are based on the inclusion of the two asymptotically negligible terms; One which depends the Fisher information matrix of complete-data for the model parameters, the other involves a prior over the collection of candidate models. The inclusion of these terms improves the performance of the criterion.

Unlike SIC, the SIC_{cd} criterion is based entirely on complete-data tools, and does not require the evaluation of the observed-data empirical log-likelihood, which may be difficult to compute. Moreover, KICcd may be evaluated in this framework by the EM algorithm.

In our simulation the SIC criterion and its variants performed very well than SIC_{cd} and its variants. These later tends to overfit to a strong degree than their correspondent criteria for the observed-data. Although the expansion of SIC$_{cd}$ depend on regularity condition that do not hold for finite mixture models, we will try to investigate framework for future work.

References

[1] Akaike H. (1973). *Information theory and an extension of the maximum likelihood principle.* In: B. N. Petrov and F. Csaki, Eds., Second International Symposium Information Theory, Akademia Kiado, Bubapest, 267 – 281.

[2] Akaike H. (1974). *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, **AC-19**, 716 – 723.

[3] Akaike H. (1978). *Time serie analysis and control through parametric methods.* Applied Times Series Analysis (D. Findley, Ed.). New York: Academic Press.

[4] Biernacki C., Celeux G. and Govaert G. (2000). *Assessing a mixture model for clustering with the integrated classification likelihood.* IEEE Transactions on Pattern Analysis and Machine Intelligence, **22(7)**, 719 – 725.

[5] Bueso M.C., Qian G. and Angulo J.M. (1999). *Stochastic complexity and model selection from incomplete data.*Journal of Statistical Planning and Inference, **76**, 273 – 284.

[6] Cavanaugh J. E. and Shumway R. H. (1998). *An Akaike information criterion for model selection in the presence of incomplete data.*Journal of Statistical Planning and Inference, **67**, 45 – 65.

[7] Cavanaugh J.E. and Andrew A. N.(1999). *Generalizing the derivation of the Schwarz information criterion.* Communication in Statistics-Theory and Methods, **28**, 49 – 66.

[8] Dempster A. P., Laird N. M. and Rubin D. B., (1977). *Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion).* Journal of the Royal Statistical Society, Series B **39**, 1 – 38.

[9] Haughton D.M.A. (1988). *On the choice of a model to fit data from an exponential family.* The Annals of Statistics, **6** 342 – 307.

[10] Kashyap R. L. (1982). *Optimale choice of AR and MA parts in autoregressive moving-average models.* IEEE transactions on Pattern Analysis and Machine Intelligence, **4**, 99 – 104.

[11] Kass R. E. (1983). *Bayes factors in practice.* The Statistician, **42**, 551 – 560.

[12] Kass, R. E. and Rafferty, A. E. (1995). *Bayes factors.* Journal of the American Statistical Association, **90**, 773 – 795.

[13] Leonard T. (1982). *Comments on 'A simple predictive density function,' by M. LeJeune and G. D. Faulkenberry.* Journal of the American Statistical Association, **77**, 657 – 658.

[14] Meng X.L. and Rubin D. B. (1991). *Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm.* Journal of the American Statistical Association, **86**, 899 – 909.

[15] Neath A. A. and Cavanaugh J. E. (1997). *Regression and time series model selection using variants of the Schwartz information criterion.* Communication in Statistics-Theory and Methods, **26**, 559 – 580.

[16] Rissanen J. (1989). *Stochastic complexity* Singapore:World Scientific Publishing.

[17] Schwarz G. (1978). *Estimating the dimension of a model.* The Annals of Statistics, **6**, 461 − 464.

[18] Shimodaira H. (1994). *A new criterion for selecting models from partialy observed data.* In: P. Cheeseman and R. W. Oldford, Eds. Selecting Models from Data: Artificial Intelligence and Statistica IV, Lecture Notes in Statistics **89**, Springer-Verlag, New York, 21 − 29.

[19] Stone M. (1979). *Comments and model selection criteria of Akaike and Schwarz.* Journal of the Royal Statistical Society, B **41**, 276 − 278.

Acknowledgement: The second author is supported by TWAS Research grant 01-159 RG/MATHS/AF/AC.

Address: B. Hafidi, A. Mkhadri, University Cadi-Ayyad, Faculty of sciences Semlalia, Department of Mathematics, PB.2390 Marrakech, Morocco

E-mail: `b.hafidi@ucam.ac.ma and mkhadri@ucam.ac.ma`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

REGRESSION OF A MULTI-SET BASED ON AN EXTENSION OF THE SVD

M. Hanafi and R. Lafosse

Key words: Singular values decomposition (SVD), multivariate regression, redundancy analysis.

COMPSTAT 2004 section: Statistical software.

Abstract: A multivariate regression of a multi-set on a multi-set is proposed directly from an extension to several matrices of the SVD.

1 Introduction

This paper is in keeping with some regression methods directly based on an extension of the singular values decompositions (SVD). The usual SVD of a matrix splits the link between two subspaces, spanned from the rows and the columns. The links between two data sets of centered variables measured on a same set of individuals are contained in a cross covariances matrix. The SVD of this matrix leads to define the pairs of components of the interbattery analysis of Tucker [6] which sum up these links. By considering that each pair is an explanatory-explained component pair, one deduces directly a multivariate regression of one set of variables on another [3], without mixture between the different pairs since a explanatory component of any pair is zero correlated with all the explained components of the other pairs. A such regression may be considered for example when the explanatory matrix contains many variables with high correlations, so that the redundancy analysis [5], [7] becomes not realistic in practice.

The SVD of a matrix has first been extended for splitting the link of one subspace with a set of subspaces, subspaces spanned from several matrices with a common number of rows [4]. The multivariate regression of one multiset of variables by one set directly deduced from this extension was defined by Hanafi and Lafosse [1]. Each explanatory component is associated to several explained components and may be used for pointing out what matrix is the best explained and then for selecting variables.

The previous SVD has been again extended for splitting the link of a set of subspaces with another set, subspaces spanned from any matrices [2]. The directly deduced multivariate regression is the one of a multi-set of variables on another multi-set, with some properties about the non-mixture between the solution components. Such a regression may be used for pointing out a best pair "explanatory set - explained set" among all the pairs and for selecting variables. By standardizing each explanatory matrix, this multivariate regression appears as an extension of the redundancy analysis.

2 Two successive extensions of the SVD notion

2.1 Introduction

One introduces the usual SVD of a matrix in a form which is forwards straight extended. The SVD of a matrix A, $p \times q$, consists in defining successive pairs (u, v) summarizing the link in A between the p rows and the q columns. The first best solution which associates a vector $v \epsilon R^q$ to a vector $u \epsilon R^p$ is obtained by maximizing the following criterion:

$$f(u, v) = (u'Av)^2, \tag{1}$$

subject to the constraint that vectors u and v have length one. The pair of vectors (u_1, v_1) associated to the optimum is the singular pair associated to the square of the greatest singular value. The other singular pairs (u_s, v_s), $s > 1$, may be successively obtained with the same criterion and by adding an orthogonality constraint between the solution vectors v_s of R^q. The orthogonality of the system of vectors u_s then is implicit. The number of solutions, where the criterion is different from zero, is equal to $rank(A)$.

2.2 SVD of matrices linking one space with several spaces

Now let $A_h,\ p \times q_h,\ \ h = 1, \dots, N$, be N matrices. The extended SVD introduced in this section is relative to the set of matrices A_h which have the same rows. One solution of this SVD is given by the $N + 1$ vectors $(u,\, v_1, v_2, \dots, v_N)$ that summarize the simultaneous link expressed by the matrices A_h $h = 1, \dots, N$ between the same p rows and the N sets of columns $q_h,\ \ h = 1, \dots, N$. The first best solution which associates N vectors of columns $v_h \epsilon R^{q_h}$ to one vector $u \epsilon R^p$ is obtained by maximizing the following criterion :

$$f(u,\, v_1, v_2, \dots, v_N) = \sum_{h=1}^{N} (u' A_h v_h)^2, \tag{2}$$

subject to the constraints that vectors $(u,\, v_1, v_2, \dots, v_N)$ have length one. One denotes $A = [A_1\ A_2\ \cdots\ A_N]$ the matrix $p \times q$ obtained from the concatenation of the matrices A_h. Let (u_1, v_1) be the pair of normed singular vectors associated to the greatest singular value s_1 of A. Let b_{h1} be the N block-vectors of v_1 of dimension q_h respectively. From the concor analysis of Lafosse and Hanafi [4], the solution of (2) is obtained for $u = u_1$ and $v_{h1} = \frac{b_{h1}}{|b_{h1}|}$, $\forall h$, and each term of the sum (2) then verifies

$$(u_1' A_h v_{h1})^2 = s_1^2\ |b_{h1}|^2\ ,\ \forall h.$$

The successive solutions $(u_s,\, v_{1s}, v_{2s}, \dots, v_{Ns})$ are obtained by using again the criterion (2) after deflations of the respective matrices A_h, for having orthonormed each system of vectors $\{v_{hs}\}$, $s = 1, 2, \dots r$, with the number r equal to $min(rank(A_h))$. Then the orthogonality of the system $\{u_s\}$ and the one of $\{v_s\}$ are implicit.

2.3 SVD of matrices linking two sets of subspaces

The SVD considered in this section is an extension of the previous one and has been defined by Kissita and al. [2], as an analysis named concorGM. The association of N metric spaces R^{q_h} with M metric spaces R^{p_k} is defined here by $N \times M$ block matrices A_{hk}, $h = 1, \dots, N$, $k = 1, \dots, M$. By construction of the convergent algorithm, this extension consists in the obtention of M vectors u_k of R^{p_k}, each of them being associated with N vectors $v_1, v_2, \dots, v_N$, from the N matrices A_{hk}, $h = 1, \dots, N$. At the same time, each of the N vectors v_h is associated with the vectors $u_1, u_2, \dots, u_M$ from the M matrices A_{hk}, $k = 1, \dots, M$. To obtain these vectors, the criterion maximized is the following one

$$f(u_1, u_2, \dots, u_M,\ v_1, v_2, \dots, v_N) = \sum_{k=1}^{M} \sum_{h=1}^{N} (v'_h A_{hk} u_k)^2, \tag{3}$$

subject to the constraint that all vectors have length one. Each solution vector u_{k1} is the first left singular vector of the concatenated matrix $p_k \times N$

$$[A'_{1k} v_{11}\ A'_{2k} v_{21}\ \dots\ A'_{Nk} v_{N1}], \tag{4}$$

and each solution vector v_{h1} is the first left singular vector of the concatenated matrix $q_h \times M$

$$[A_{h1} u_{11}\ A_{h2} u_{21}\ \dots\ A_{hM} u_{M1}]. \tag{5}$$

The successive solutions are obtained from the same criterion, with $M+N$ supplementary orthogonality constraints in each R^{p_k} and in each R^{q_h}. So, $M + N$ orthonormal systems of r vectors $\{u_{ks}\}$ and $\{v_{hs}\}$ may be defined, with $s = 1, \dots, r$, where $r = min(rank(A_{hk}))$.

3 Applications

3.1 Regression of a set on another by the interbattery

Let X, $n \times p$, and Y, $n \times q$, be two sets of centered variables. The multivariate regression of Y on X is built with $A = X'Y/n$ from section 2.1. So the criterion (1) maximized with norm constraints is

$$f(u, v) = cov^2(Xu, Yv).$$

The solution pairs (u_s, v_s) define the pairs of components (Xu_s, Yv_s) of the interbattery analysis of Tucker [6]. The vectors v_s verifying

$$v_s = \frac{Y'Xu_s}{|Y'Xu_s|}, \tag{6}$$

by noting y_l, $l = 1, \ldots, q$, the columns of Y, from (6) and $v'_s v_s = 1$ we have

$$\rho^2(Xu_s, Yv_s)\, var(Yv_s) = \sum_{l=1}^{q} \rho^2(Xu_s, y_l)\, var(y_l), \tag{7}$$

the relation (7) shows how the explanations of all the variables of Y are packed with Yv_s.

In the same way the variables of X are packed with the explanatory components Xu_s. That may be wanted when the variables of X are numerous with high correlations. An explanatory component only built for maximizing the packing (7) is not always easily definable in practice. In any case, one may also have an interest for some explanatory components which represent X because they catch a lot of correlations between the variables of X. The vector v_s are wanted orthogonal for defining different parts $var(Yv_s)$ of the variance of Y. So the explained variances $\rho^2(Xu_s, Yv_s)\, var(Yv_s)$ are also different. No mixture exists between an explanatory component with an explained component relative to another pair, since

$$cov(Xu_s, Yv_{s'}) = 0,\ \forall s' \neq s.$$

* *Remarks* Each vector u_s verifies

$$u_s = \frac{X'DYv_s}{|X'DYv_s|}. \tag{8}$$

The equality (8) is not dependent on the norm of Yv_s, and nothing changes by supposing $var(Yv_s) = 1$. Denoting x_j the (standardized) column j of X, and $\rho_j = \rho(x_j, Yv_s)$ the linear correlation which estimates the predict power of x_j, the relation (8) leads to the fact that the terms u_j of u_s are equal to

$$u_j = \frac{\rho_j}{\sqrt{\sum \rho_j^2}},\ \forall j. \tag{9}$$

When the standardized variables of X are uncorrelated, the explanatory components Xu_s are also standardized and uncorrelated. So, when X is the matrix of the standardized principal components of an initial set of variables, this regression corresponds to the redundancy analysis of Y on the initial set [5], [7].

3.2 Regression of a multi-set on one set by concor

Let Y_h, $n \times q_h$, and X, $n \times p$, be $N + 1$ data sets of centered variables measured on a common set of n individuals. The simultaneous regression of the matrices Y_h on X has been defined by Hanafi and Lafosse [1] and may be deduced directly from the section 2.2, by considering $A_h = Y_h' X/n$. The criterion (2) becomes

$$f(u,\, v_1, v_2, \ldots, v_N) = \sum_{h=1}^{N} cov^2(Xu, Y_h v_h). \tag{10}$$

and each of the s successive solutions defines N components $Y_h v_{hs}$ explained by one component Xu_s. By denoting y_{hl} the columns of Y_h. As (6) and (7) we have

$$\rho^2(Xu_s, Y_h v_{hs})\, var(Y_{hs} v_{hs}) = \sum_{l=1}^{q_h} \rho^2(Xu_s, y_{hl})\, var(y_{hl}). \tag{11}$$

One denotes $Y = [Y_1\ Y_2\ \ldots.\ Y_N]$ the concatenated matrix $n \times q$,and $Z_s = [Y_1 v_{1s}\, Y_2 v_{2s}\, \ldots\, Y_N v_{Ns}]$ the concatenated matrix $n \times N$. The N components $Y_h v_{hs}$ define a mean component

$$Yv_s = \sum_{h=1}^{N} cov(Xu_s, Y_{hs} v_{hs})\, Y_{hs} v_{hs}, \tag{12}$$

where v_s is the right singular vector of $X' Z_s/n$. That means that a component Yv_s also packs the components $Y_h v_{hs}$

$$\rho^2(Xu_s, Yv_s)\, var(Yv_s) = \sum_{h=1}^{N} \rho^2(Xu_s, Y_{hs} v_{hs})\, var(Y_{hs} v_{hs}). \tag{13}$$

In the same time, because the criterion (10), the packing of the correlations of the variables of X also is considered.

The non-mixture between the solutions comes from the equalities

$$cov(Xu_s, Y_h v_{hs'}) = 0,\ \forall s' < s, \forall h.$$

$$cov(Xu_s, Yv_{s'}) = 0,\ \forall s' \neq s.$$

When X is the matrix of the standardized principal components of an initial set of variables, this regression corresponds to a generalized redundancy analysis of $Y_1,\ Y_2,\ \ldots.\ Y_N$ on the initial set.

3.3 Regression of a multi-set on another by concorGM

One considers M data matrices X_k, $n \times p_k$, $k = 1, \ldots, M$, and N data matrices Y_h, $n \times q_h$, $h = 1, \ldots, N$, which are $M+N$ sets of centered variables measured on a common set of n individuals. The explanatory matrices X_k may contain high correlations. From the section 2.3, by denoting$A_{hk} = Y_h' X_k/n$, the criterion (3) becomes

$$f(u_1, u_2, \ldots, u_M,\, v_1, v_2, \ldots, v_N) = \sum_{k=1}^{M} \sum_{h=1}^{N} cov^2(X_k u_k, Y_h v_h). \tag{14}$$

Then, from (4), a solution s leads to M matrices, for $k = 1, \ldots, M$

$$X_k' D [Y_1 v_{1s} \, Y_2 v_{2s} \, \ldots . \, Y_N v_{Ns}] = X_k' D Z_{Ys}.$$

One denotes by u_{ks} the M first respective left singular vectors and b_{ks} the M respective first right. All the components $X_k u_{ks}$are associated with the same set of components $\{Y_h v_{hs}\}$, and each $X_k u_{ks}$ is linked to a particular mean component

$$Z_{Ys} b_{ks} = \sum_{h=1}^{N} cov(X_k u_{ks}, Y_h v_{hs}) \, Y_h v_{hs},$$

which packs the components $Y_h v_{hs}$, these last being packing the variables of the respective matrices Y_h. In the same way from (5), a solution s leads to N matrices, for$h = 1, \ldots, N$

$$Y_h' D [X_1 u_{1s} \, X_2 u_{2s} \, \ldots . \, X_M u_{Ms}] = Y_k' D Z_{Xs}.$$

One denotes by v_{hs} the N respective first left singular vectors and a_{hs} the N respective first right. All the component $Y_h v_{hs}$ are associated to the same set of components $\{X_k u_{ks}\}$, and each $Y_h v_{hs}$is linked to a particular mean component

$$Z_{Xs} a_{hs} = \sum_{k=1}^{M} cov(Y_h v_{hs}, X_k u_{ks}) \, X_k u_{ks},$$

which packs the components $X_k u_{ks}$, these last being packing the variables of the respective matrices X_k.

In the previous section 3.2 only the matrices Y_h were deflationed, but the deflation of X was then implicit. But here, it is not the case for each X_k which must be deflationed. Only when the easy interpretation of the scores given by (9) is strongly wanted, these deflations might not be done.

The following properties of the multivariate regression based on concorGM increase the non-mixture between the successive explanations:

$$cov(Z_{Ys} b_{ks}, X_k u_{ks'}) = 0, \; \forall s' > s, \; \forall k.$$

$$cov(Z_{Xs} a_{hs}, Y_h v_{hs'}) = 0, \; \forall s' > s, \; \forall h.$$

A proposal for a generalized redundancy analysis of N matrices on M initial matrices consists in this type of regression of the N matrices on M standardized matrices, each of them being the matrix of the standardized principal components of the respective initial matrix.

4 Illustration

We consider matrices of 8 chemical variables for 30 different types of wine (5 types of stem combined with 6 types of graft), collected in the 5 years

1983, 1985, 1988, 1989, and 1990, contained in column-centered-standardized matrices $X_1, \ldots, X_5$ of order 30x8. These measurements were obtained just after harvesting when these wines were still *young*. The measurements of the 30 wines were repeated after five years, obtaining an new set of matrices $Y_1, \ldots, Y_5$ referred to as *old* wines. The measurements of the old wine can then be explained from the measurements of the young wine, and the X_i's matrices are the explanatory matrices. In the tables below, we report the explained variances and the squared correlations of the two first solutions obtained from the concorGM regression described in the section 3.3.

	Component 1					Component 1				
	1983	1985	1988	1989	1990	1983	1985	1988	1989	1990
1983	**3.10**	.63	.25	.29	.46	**.84**	.21	.89	.41	.82
1985	1.44	**1.76**	1.14	1.01	.90	.56	**.13**	1.11	.36	.60
1988	.08	.74	* **1.75**	1.54	.83	.51	.34	**1.22**	.33	.60
1989	.47	.62	.99	**2.49**	.45	.65	.19	1.24	**.55**	.97
1990	.98	1.24	1.39	1.16	**1.07**	.58	.17	.88	.38	**1.34**

Table 1: Explained variances.

	Component 1					Component 1				
	1983	1985	1988	1989	1990	1983	1985	1988	1989	1990
1983	**.82**	.23	.11	.09	.22	**.64**	.19	.52	.45	.42
1985	.38	**.66**	.51	.30	.43	.43	**.12**	.64	.40	.31
1988	.02	.28	**.78**	.46	.39	.39	.32	**.70**	.37	.31
1989	.12	.23	.44	**.74**	.21	.49	.18	.72	**.61**	.50
1990	.26	.46	.62	.34	**.50**	.44	.16	.51	.42	**.68**

Table 2: Squared correlations.

Principally for the first component solution, the tables 1 and 2 reveal higher values on the diagonal. When all the young wines are set in competition for explaining all the old wines, the explanations between the wines of the same year are the best. For each harvesting, some chemical measurements might be used for predicting the state old from the state young.

References

[1] Hanafi M., Lafosse R. (2001). *Généralisations de la régression linéaire simple pour analyser la dépendance de K ensembles de variables avec un K+1 eme.* Rev. Stat. Appliquée **49** (1), 5 − 30.

[2] Kissita G., Cazes P., Hanafi M., Lafosse R. (2004). *Deux méthodes d'analyse factorielle du lien entre deux tableaux de variables partitionnés.* To appear in Rev. Stat. Appliquée.

[3] Lafosse R. (1997). *Analyse de concordance de deux tableaux: monogamies, simultanéités et découpages.* Rev. Stat. Appliquée **45** (3), 45 – 72.

[4] Lafosse R. et Hanafi M. (1997). *Concordance d'un tableau avec K tableaux: définition de K+1 uplés synthétiques.* Rev. Stat. Appliquée **45** (4), 111 – 126.

[5] Rao C.R. (1964). *The use and the interpretation of principal component analysis in applied research.* Sankya, ser. A **26**, 329 – 358.

[6] Tucker L. R. (1958). *An interbattery method of factor analysis.* Psychometrika **23**, 111 – 136.

[7] Wollenberg A.L. (1977). *Redundancy analysis. An alternative for canonical analysis.* Psychometrika **42**, 207 – 219.

Address: M. Hanafi, Unité Mixte de Recherche en Sensométrie et Chimiométrie, (ENITIAA-INRA), Nantes, France
R. Lafosse, Laboratoire de Statistique et Probalilités, Université Paul Sabatier, Toulouse, France

E-mail: `hanafi@enitiaa-nantes.fr, lafosse@cict.fr`

AN AID TO ADDRESSING TOUGH DECISIONS: THE AUTOMATION OF GENERAL EXPRESSION TRANSFER FROM EXCEL TO AN ARENA SIMULATION

William V. Harper

Key words: Simulation, automation, data transfer.

COMPSTAT 2004 section: Simulations.

Abstract: A large segment of quantitative teaching involves decision methodologies. These techniques are challenging for many students but yet are sometimes felt to be too academic and not real world oriented by some. This paper demonstrates the development of Visual Basic for Applications (VBA) automation based methods to allow both students and practitioners to more easily apply simulation techniques to decision problems. This flexibility can be powerful in addressing numerous real world problems in a timely manner.

1 Introduction

This paper has a particular focus on automating the interface between the simulation package Arena and database oriented packages such as Microsoft Access and Excel. Such automation is not commonly used and many hours/ days may be spent entering data into simulation packages by hand. Using Microsoft's Visual Basic for Applications (VBA), the paper demonstrates how the data transfer can be automated to aid in making simulation-based decision supporting methods more accessible.

Automating transfer to or from programs such as Excel to Arena can be a tremendous aid. Transferring numeric data during run time may be accomplished without VBA in many circumstances. For example, Arena's ReadWrite module may be sufficient for one's needs. VBA may be necessary in some applications or to provide a desired user interface.

What if the need is not the transfer of numeric results but text strings? This paper looks at a particular application involved the transfer of general expressions that can vary from simple numeric values to more complicated expressions such as "(PartType = = 1) * (EXPO(6)) + (PartType = = 2) * TRIA(4, 7, 9)". This requires not only VBA but also proper execution timing to ensure that compilation checks are complete but transfer of such strings is performed just prior to the actual beginning of the run mode of Arena.

Arena [1] is a well known simulation package commonly used for discrete-event simulation. While the focus of this paper is primarily on Arena and

Excel, the methods shown can be adapted to other software packages. Readers are encouraged to contact the author with suggestions or questions. From this point on, the steps described are from the perspective of running from the Arena simulation package.

Many external application programs such as Excel, Access, Oracle, and Arena supply a VB/VBA object library. It is necessary to add the object library to the Visual Basic application so that the object library becomes available to your program. Select the Tools/References option from Visual Basic application and a list of selected and available objects is displayed. To add the Excel object library, scroll through the list of available libraries and check the "Microsoft Excel 9.0 Object Library." or whatever the relevant version is.

2 Populate an Arena array with general expressions

The process for this example models a two-machine finishing plant. Unfinished furniture enters the system and proceeds to a sanding machine. The sanding machine prepares the furniture for painting; 10 different types of furniture enter the system. Each product type requires a different delay for sanding. Sanded furniture proceeds to a painting machine that finishes the furniture; the 10 different types of furniture also requiring different paint times.

In this example the data transfer from Excel will be much more general than simple numeric values to be used as population parameters for predefined distributions in Arena. Instead the user may put any valid Arena expression into Excel and then let it control the processing times in Arena. Below are the major steps involved with this process.

1. Call a VBA user form before the run mode of Arena is initiated and request the user to select an appropriate Excel file name.

2. Open an instance of Excel.

3. Call a form that will load expressions from the Excel workbook and copy it into Arena.

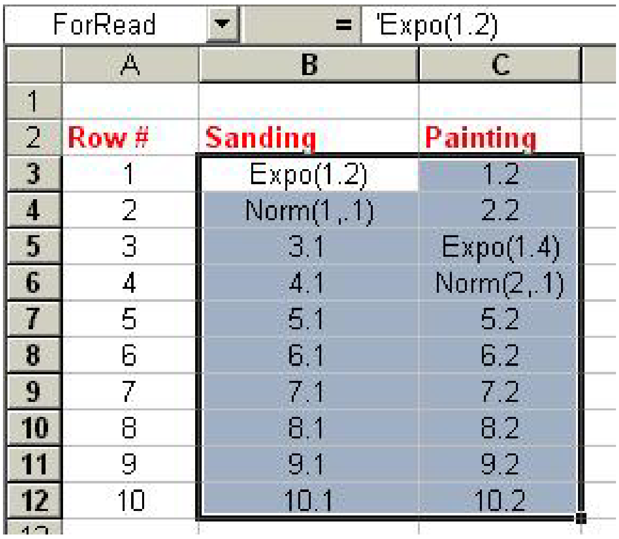
4. Use a VBA form that queries the user for which workbook to load. Use a common dialog control to model the file selection form.

5. Use a VBA form that informs the user that information is being imported and the spreadsheet that is currently active. Use the form activate event to retrieve the data from Excel and populate the Arena variable.

6. Close the Excel file.

7. Show the VBA form filled with VBA expressions.

Open the Excel file (Figure 1) containing the source data with sanding and painting times, Example 4 Reading Expressions.xls. Note that a Named Range called *ForRead* that will be used to link this Excel file to our VBA code. Close the Excel file.



ForRead = 'Expo(1.2)

	A	B	C
1			
2	Row #	Sanding	Painting
3	1	Expo(1.2)	1.2
4	2	Norm(1,.1)	2.2
5	3	3.1	Expo(1.4)
6	4	4.1	Norm(2,.1)
7	5	5.1	5.2
8	6	6.1	6.2
9	7	7.1	7.2
10	8	8.1	8.2
11	9	9.1	9.2
12	10	10.1	10.2

Figure 1: Excel file highlighted named range.

Run the model to see the sequence of events:

- The *Select Input File* form is displayed. Click the **Select Excel File** button.
- Browse to the **Example 4 Reading Expressions.***xls* file and click **Open**.
- Click **OK** in the Select Input File form.
- The *Loading Data from Excel* form is displayed briefly.
- A form showing the transferred Expressions is shown.
- Then the simulation run commences.

3 Solution overview

When the Arena Run tool button is clicked, Arena initiates the RunBegin event and any associated VBA code. This is where it is essential to perform all the VBA transfer of text strings. Otherwise once the RunBegin event closes and the RunBeginSimulation event, the SIMAN object takes control for the running of the simulation. When SIMAN is active, it is not possible to have the editing capabilities similar to using a mouse while editing Arena. Prior to SIMAN becoming active, VBA can perform many of the activities that one would manually do in Arena. At the end of the simulation having SIMAN active is a big plus on the other hand as all the simulation result data is available for whatever purposes one may have.

At the beginning of the simulation run in the RunBegin event, the LoadData form is displayed. This form includes the code to ask the user for a file name, read the expression data from Excel, and initialize the Arena Expressions. The resulting transferred expression form is shown to the user before the run mode of Arena is started.

A key part is the timing. In Arena the compilation step converting the Arena objects to SIMAN code and the initialization of variables and expressions is performed in the RunBegin event. After successful completion of RunBegin, SIMAN is activated and the model run in the RunBeginSimulation. Our VBA task occurs in the RunBegin event so that we can still interact directly with the Arena objects and transfer the desired expressions. Once the subsequent RunBeginSimulation event initiates, the SIMAN object is active and considerably less control of the simulation is possible during this run time. While numeric data values can be transferred during the run time, it is not possible to edit Arena objects or to transfer expressions with text strings. Below is the simple VBA code for the RunBegin event.

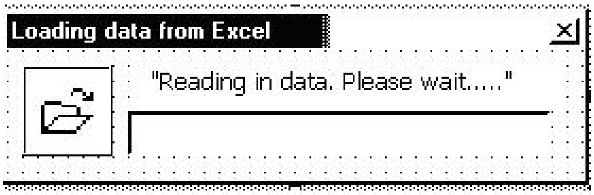
```
Option Explicit
Private Sub ModelLogic_RunBegin ()
‘Show the LoadData form to read the information from Excel
    LoadData.Show
      ’Show that the 20 expressions have successfully migrated.
          ShowExp.Show
End Sub
```

The approach to reading the data from Excel requires the user to select the Excel file, then read the data from the file (with a status form, the LoadData form, displayed). This uses the three form files GetFile.frm, LoadData.frm, and ShowExp.frm developed for this application.

4 LoadData: Read Excel input and assign Arena expressions

The *LoadData* form will be displayed while data is being read from Excel. The VBA code associated with this form (Figure 2) may be obtained from the author as the conference page limits will not allow the full listing. However comments on the highlights of the code are given below.

- **sExcelFile variable:** Declared as Public so that the GetFile form can set its value based on the selection made by the user.



"Reading in data. Please wait....."

Figure 2: LoadData User Form.

- **Excel variables:** The oExcelAppl, oWorkbook, and oWorksheet variables are Global variables we defined in this project. They are used here to start Excel, open a file, and read data from a worksheet. They will be cleared in RunEnd as part of our cleanup code.
- **CreateObject:** This is a built-in Visual Basic function that starts an application (or class), returning an object variable that points to the program. A similar function, GetObject, returns a pointer to a copy of the program that is already running. With Microsoft Excel, the application starts but is not displayed (i.e., not visible). If you want to see Excel, set its visible property to true, e.g., **oExcelApp.Visible = True.**
- **Me.Repaint:** Refreshes the form. Windows sometimes does not immediately display forms; in cases, where there is a lot of processing, the form may not be painted at all without forcing a Repaint.
- **Cells property:** The Cells property returns or sets data values in Excel cells. It takes two arguments, the row and column numbers.
- **Range object:** Excel's Range object can be used to point to a range of cells. We use it here so that we can name a range in Excel (using the *Insert, Names, Define* menu), rather than hard-coding a specific starting and ending cell.
- **oRange.Cells:** When using the Cells property of a Range object (instead of the Worksheet), the row and columns are offset relative to the top, left cell of the range, starting with 1.
- **oTextBoxes(10, 2) As TextBox:** This allows all 20 text boxes on the ShowExp user form to be easily populated using this doubly dimensioned array via simple looping in the VBA code.

5 GetFile: Selecting the Excel file

The **GetFile** form will be used to allow the user to browse to and select the Excel file containing the source data with the desired expressions. Below is the graphic of the form (Figure 3) followed by the associated VBA code.

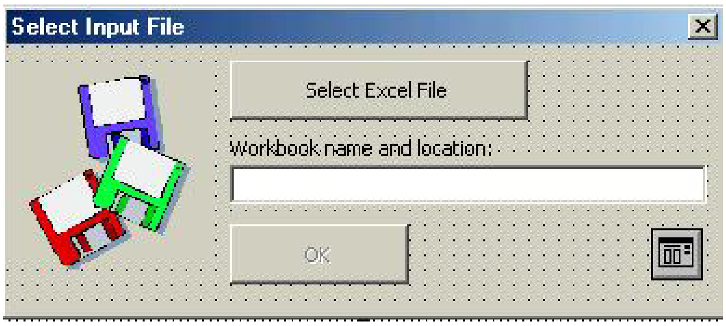


Figure 3: GetFile User Form.

```
Option Explicit
Private Sub cmdSelectExcelFile_Click()
    With cdlgListExcelFiles
        ' Set the filter for the common dialog control
        .Filter = "All Files (*.*)|*.*|Excel Files (*.xls)|*.xls|"
        ' Set the common dialog filter index to the second item
               in the filter list
        .FilterIndex = 2
        ' Set the common dialog style to Open file.
        ' This command also activates the common dialog control.
        .Action = 1
    End With
    ' When control is returned back to this form, query the common
    '  dialog control and retrieve the file name.
    LoadData.sExcelFile = cdlgListExcelFiles.FileName
    'If a filename was retrieved, enable the OK button
    If Len(LoadData.sExcelFile) > 0 Then
        cmdOK.Enabled = True
        ' Display the filename and path in the text box so the
               user can validate it.
        txtExcelFilename.Text = cdlgListExcelFiles.FileName
    End If
End Sub
Private Sub cmdOK_Click()
    Unload Me
End Sub
```

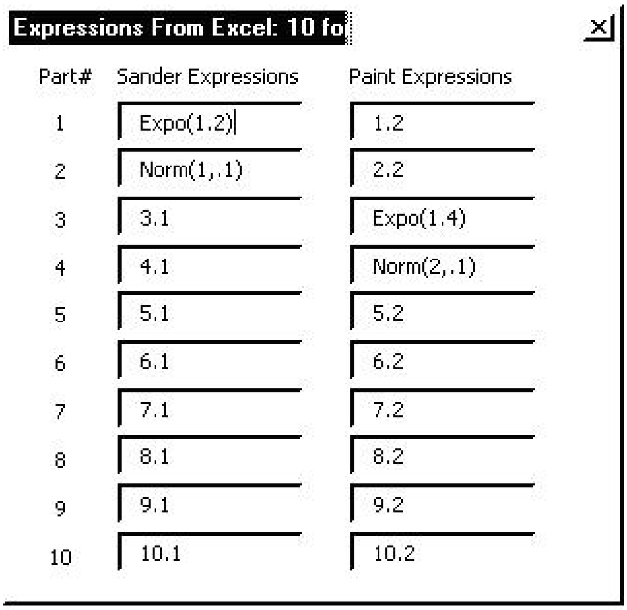
5.1 Comments on GetFile form

Common Dialog control This control is a Microsoft control providing standard features for file browsing. To attach it to your VBA project, right-click in the Toolbox and select the Additional Controls menu item. This displays a list of controls that are installed on your computer. Click the check box next to **Microsoft Common Dialog Control, version 6.0** or whatever is the current version to add the Common Dialog to your Toolbox. This is an ActiveX control that you can use in your projects.

6 Form to show the transferred expressions

Information transfer to the *ShowExp* form is handled by the *LoadData* form VBA code. This user form has no direct VBA code associated with it; however, care was taken to label the text boxes and control their properties accordingly.

Below (Figure 4) is the fully populated form with the expressions successfully transferred from Excel. Once the user closes this form, the simulation will change to the run mode in RunBeginSimulation and use these expressions to control the 20 processing times.



Part#	Sander Expressions	Paint Expressions
1	Expo(1.2)	1.2
2	Norm(1,.1)	2.2
3	3.1	Expo(1.4)
4	4.1	Norm(2,.1)
5	5.1	5.2
6	6.1	6.2
7	7.1	7.2
8	8.1	8.2
9	9.1	9.2
10	10.1	10.2

Figure 4: ShowExp User Form.

7 Summary

This paper illustrates the use of Visual Basic for Applications to automate the transfer of expressions from Excel to the simulation package Arena. This approach allows complex simulations to be easily adapted to many scenarios by changing the expressions in Excel which is generally easy for most users. Thus a given simulation can be adapted to other needs without the need of the user being forced to learn the simulation package just for data and expression entries.

References

[1] Kelton, David W., Sadowski R.P., Sturrock D.T. (2004). *Simulation with Arena*. Third Edition, McGraw-Hill, New York.

Acknowledgement: Thanks to my family for their tolerance with me while working hard to learn automation methods such as described here.

Address: William V. Harper, Mathematical Sciences, Otterbein College, Towers Hall 139, One Otterbein College, Westerville, OH 43081-2006 USA
614-823-1417 Fax: 614-823-3201
Faculty page: `http://www.otterbein.edu/home/fac/WLLVHRPR`

E-mail: `WHarper@otterbein.edu`

TWO CLASSIFICATION METHODS FOR EDUCATIONAL DATA AND IT'S APPLICATION

Atsuhiro Hayashi

Key words: Rule space method, educational statistics, cognitive science.

COMPSTAT 2004 section: Classification.

Abstract: Both methods, Rule Space Method (RSM) and Neural Network Model (NNM) are techniques of statistical pattern recognition and classification approaches developed from different fields; one is for behavioral and the other is for neural sciences. RSM is developed in the domain of the educational statistics. It starts from the use of an incidence matrix Q that characterizes the underlying cognitive processes and knowledge (Attribute) involved in each Item. Examinee's mastered/non-mastered states (Knowledge State) for each attribute is determined from item response patterns. RSM uses the multivariate decision theory to classify individuals, and NNM that is considered as a nonlinear regression method uses the middle layer of the network structure as classification results. We have found some similarities and differences between the results from the two approaches, and moreover the both methods have supplemental characteristics to each other when applied to the practice. In this paper, we compare the both approaches by focusing on the structures of NNM and on knowledge States in RSM. And finally, we show an application result of RSM for a reasoning test in Japan.

1 Introduction

A Neural Network model was proposed for the purpose of modeling the information processing in person's brain in the 1940s. Neurons (nerve cell elements) are considered as the minimum composition unit of cerebral functions that entangled in complicated and organic manners. The model shows that all the logical reasoning can be described in a finite size of the number of neurons and connections [2]. The model enables us to express acquisition of new knowledge from learned examples in the past, therefore it can be used to help to solve one of the weaknesses in constructing an AI (Artificial Intelligence) system. It is known that expressing knowledge acquisition in an AI system is extremely difficult.

On the one hand, Rule Space Method is a technique of clustering examinees into one of the predetermined latent Knowledge States (KS) that are derived logically from expert's hypotheses about how students learn. The method can be considered as a statistical testing technique of expert's hypotheses. These hypotheses are expressed by an item-attribute matrix (incidence matrix Q) where attributes are representing underlying knowledge and cognitive processing skills required in answering problems [1]. A Knowledge State consists of mastered/non-mastered of attributes, and a list of all the possible Knowledge States can be generated algorithmically by applying Boolean Algebra to the incidence matrix Q. This method is fairly new but has lately started getting some attention because it is possible to provide diagnostic scoring reports for a large-scale assessment [3]. We have found there are similarities between the results from the two approaches, and moreover they have complementary characteristics when applied to the practice. In this paper, we discuss the comparisons of both approaches by focusing on the structure of the Neural Network (NN) and of Knowledge States in the RSM. And we show an application result for a reasoning test.

2 Feed-forward neural network model

In spite of that the mathematical formulization of the Feed-Forward NN is simple, any nonlinear functions can be used by selecting deferent numbers of middle layers and connections between neurons. When we apply this technique to existing data obtained from learning processes, we can use this model to search for the strategy of any joint intensity between units.

From statistical point of view, NNs are nonlinear regression equation models. In this paper Feed-Forward NN is considered as a model-fitting procedure to estimate the optimum values of parameters in regression equations [4].

This procedure is called parameter estimation in statistics, but is called a learning algorithm in NN. One of the learning algorithms commonly used is Back Propagation (BP) that is a learning method by passing on errors to previous layers. BP is an adaptation of the steepest descent method to the NN field. This method has a reducible faculty of the convergence to the local minimum point.

3 Rule space method

RSM is a technique developed in the domain of the educational statistics [7]. It starts from the use of an incidence matrix Q that characterizes the underlying cognitive processes and knowledge (Attribute) involved in each Item. It is a grasping method of each examinee's mastered/non-mastered learning level (Knowledge State, KS) from item response patterns. Up to now, the results of examinees' performance on a test are reported by total scores or scaled scores. However, if this technique is used in educational practices, it is possible to report which attributes each student mastered or non-mastered, in addition to his/her total scores. It is often true that the same total score may have several different Knowledge States. By reporting detailed information of his/her Knowledge State, learning can be facilitated more effectively than just providing total scores only.

4 Science reasoning test

The Science Reasoning Test (SR-Test) is an entrance examination test that measures the student's interpretation, analysis, evaluation, reasoning, and problem-solving skills required in the natural sciences [5].

Since we got the ACT's (American College Testing, Inc.) cooperation, we used one open-form of their ACT Assessment tests for our experimentation. The test is based on units containing scientific information and a set of multiple choice questions about the scientific information. Calculators are not permitted to be used for the test. The scientific information for the test is provided in one of three types of formats.

The first format, data representation, presents graphic and tabular material similar to that found in science journals and texts. The questions associated with these format measure skills such as graph reading, interpretation of scatter plots, and interpretation of information presented in tables. The second format, research summaries, provides students with descriptions of one or more related experiments. The questions focus upon the design of experiments and interpretation of experimental results. The third format, conflicting viewpoints, presents students with expressions of several hypotheses or views that, being based on differing premises or on incomplete data, are inconsistent with one another. The questions focus upon the understanding, analysis, and comparison of alternative viewpoints or hypotheses.

The Science Reasoning Test questions require students to use the scientific method to answer the questions. The students are required to recognize and understand the basic features of, and concepts related to, the provided information; to critically examine the relationships between the information provided and the conclusions drawn or hypotheses developed; and to generalize from given information to gain new information, draw conclusions, or make predictions.

5 Numerical examples

We applied the RSM to a data of fraction addition problems, and got a tree structure of Knowledge State. We related RSM that derives the Knowledge State from an incidence matrix Q, to the Feed-Forward NNM. For that, we designed the network of the three-layer structure in which items were assigned to the input layer and Attributes were to the output layer. The Knowledge States in the RSM were considered to correspond to the middle layers of NNM. We applied several numerical examples to the both methods, and found close similarities in their results although they were not identical.

And we applied the RSM to a data of Science Reasoning Test of 286 Japanese students. The number of attributes and items are 12 and 18, respectively. Figure 1 is the tree representation of the Knowledge States that shows the examinee's mastered/non-mastered learning level. In this figure, each circle is the Knowledge State, and the numbers in the circle are the IDs of non-mastered attribute. Or the number in the parenthesis is the number of examinee classified in this Knowledge State. We can find the fact that the main solving attribute IDs are 6, 8 and 9, and secondary attribute are 2 and 5. The total examinee classified in these Knowledge States is 225, which is about 80% of all. The main streams to reach the full mastered state are three Knowledge States of left-hand side in the third layer from the top.

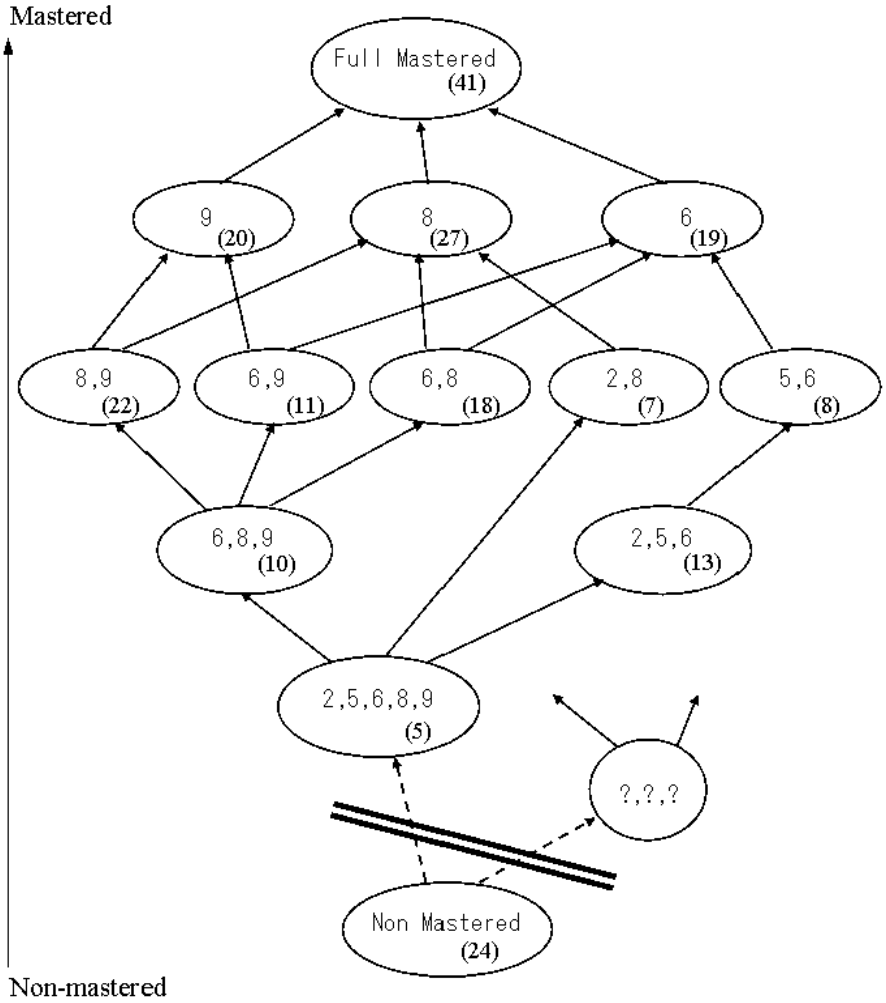


Figure 1: A tree representation of Knowledge States for the SR-Test data.

6 Discussion and conclusions

We investigated the relationship between the characteristics of the middle layer of NN and the Knowledge States in the RSM, and discussed their similarities and usefulness at the weaknesses existing in the RSM.

It is well known that the composition of an incidence matrix Q in the RSM is a very laborious task, requires experts' intense cooperation. The experts identify attributes involved in each item and express them in an incidence matrix Q. It needs to investigate multiple numbers of solution strategies for each item. This is extremely hard work. If an examinee's mastering level (cluster) is known to some extent from past experiences, it is also possible to construct a network in which these clusters are assigned to the output layer of NNM. The middle layer drawn from this model is expected to correspond to Attributes. It may be possible to use this result for replacing a task analysis required in making an incidence matrix Q in RSM.

We plan to clarify the difference and similarities of the two models with numerical examples, or will get useful results to apply these methods for the SR-Test data and our real examination data.

References

[1] Klein M.F., Birenbaum M. et al. (1981). *Logical error analysis and construction of tests to diagnose student "Bugs" in addition and subtraction of fractions.* University of Illinois, Computer-based Education Research Laboratory, Research Report 81 – 6.

[2] Kurita T., Motomura Y. (1993). *Feed-forward neural networks and their related topics.* Japanese Journal of Applied Statistics **22** (3), 99 – 114, (in Japanese).

[3] Tatsuoka K.K. (1995). *Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach.* Paul D. Nichols et al. (eds.), Cognitively Diagonostic Assesment, 327 – 359, Lawrence Erlbaum Associates.

[4] Hayashi A., Baba Y. (1998). *An analysis of university entrance examination data by using neural network models.* Compstat 1998, Physica-Verlag, Bristol, Short Communications, 45 – 46.

[5] Maxey J. (2000). *Introduction to the ACT assessment, international comparison study for university entrance examinations.* M. Huzii et al. (eds.), 42 – 55.

[6] Hayashi A., Tatsuoka K.K. (2000). *A comparison of rule space method and neural network model for classifying individuals.* International Conference on Measurement and Multivariate Analysis and Dual Scaling Workshop(ICMMA), Volume two, 226 – 228.

[7] Tatsuoka K.K., Hayashi A. (2001). *Statistical method for individual cognitive diagnosis based on latent knowledge state*. Journal of The Society of Instrument and Control Engineers **40** (8), 561 – 567, (in Japanese).

Acknowledgement: We are grateful to Dr. James Maxey and ACT that gave us their permission of one open-form in their ACT Assessment tests for our experimentation.

Address: A. Hayashi, The National Center for University Entrance Examinations, 2-19-23 Komaba, Meguro-Ku, Tokyo, 153-8501, Japan

E-mail: `hayashi@rd.dnc.ac.jp`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

PROTECTION OF CONFIDENTIAL DATA WHEN PUBLISHING CORRELATION MATRICES

Jobst Heitzig

Key words: Statistical computing, multivariate correlation, confidentiality.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: When publishing Pearson correlations of confidential data, how to ensure that the individual data remain unknown even to people with additional knowledge? The paper gives a first answer by proving an "inference interval" for each data cell and outlining some viable algorithms for confidentiality protection, one of which might also be applicable to other statistical measures.

1 Introduction

Like other national statistical offices, the Federal Statistical Office Germany (and also the Statistical Offices of the Länder) is trying to improve scientists' access to official microdata. Besides the distribution of anonymised microdata files (so-called "Scientific Use Files"), our *Research Data Centre* also allows researchers to submit analysis programs (using SPSS or SAS syntax, for instance) which are executed and whose results are then returned to the scientist.

Doing so, one must of course make sure that the individual respondent's confidential data remain protected after publication of the results. That is, no-one must be able to infer anything useful about an individual by combining the returned information with any other knowledge he might possess. This *additional knowledge* could for example include a part of some individual's data. It is therefore of essential importance to study the inferences that can be drawn from various kinds of statistical measures in combination with partial knowledge of the underlying data.

This paper deals with the idealized case where the requested information only depends upon first and second moments of the common empirical distribution of (subsets of) the data. In other words, we ask what one can infer from *means, variances, and Pearson correlations.* After proving an encouraging first result in section 2, section 3 presents an algorithm of acceptable computational efficiency which ensures confidentiality protection for the correlation matrix of some given multivariate data. Section 4 discusses the more complicated case where further additional knowledge must be assumed — when the data is known to respect some boundary conditions, for instance. Finally, I try to stimulate further research along these lines by pointing out a number of open questions.

2 Possible values of the target cell

Assume that $Z_1, \dots, Z_k$ are the standardized versions of our multivariate data, consisting of n observations. That is, each Z_i is an n-dimensional column data vector whose sum is zero and whose square sum $n-1$. Let $\varrho = (\varrho_{ij})_{ij}$ be their correlation matrix, that is, $\varrho_{ij} = Z_i^\top Z_j/(n-1)$.

In our first, worst-case-style scenario, the *additional knowledge* consists of all but one values for the data of some "target" individual. Without loss of generality, we can assume that the target individual is represented by observation 1 and the unknown target data is represented by variable k. Hence the question is:

What can we infer about Z_{k1} knowing ϱ and $(Z_{11}, \dots, Z_{(k-1)1})$?

Note that we do not ask for an estimation of Z_{k1} and the corresponding estimation error. Instead, we need to know all possible values of Z_{k1}:

Theorem 2.1. *Given ϱ, Z_{known}, and $n \geq k+2$, the possible values of Z_{k1} are given by the interval with boundary*

$$\zeta \quad \pm \quad \sqrt{\frac{\det \varrho}{\det \varrho_{\text{known}}}} \sqrt{\frac{(n-1)^2}{n} - Z_{\text{known}}^\top (\varrho_{\text{known}})^{-1} Z_{\text{known}}}, \tag{1}$$

where

$$\begin{aligned} \zeta &= Z_{\text{known}}^\top (\varrho_{\text{known}})^{-1} \varrho_{\text{dep.}}, \\ Z_{\text{known}} &= (Z_{11}, \dots, Z_{(k-1)1}), \\ \varrho_{\text{known}} &= (\varrho_{ij})_{i,j=1\dots k-1}, \\ \textit{and} \quad \varrho_{\text{dep.}} &= (\varrho_{ik})_{i=1\dots k-1}. \end{aligned}$$

Proof. Given ϱ and the complete first observation $(Z_{11}, \dots, Z_{k1})$, one can easily compute the correlation matrix ϱ_{rest} corresponding to the remaining $n-1$ observations as

$$(\varrho_{\text{rest}})_{ij} = \frac{(n-1)^2 \varrho_{ij} - n Z_{i1} Z_{j1}}{\sqrt{(n-1)^2 - n Z_{i1}^2} \sqrt{(n-1)^2 - n Z_{j1}^2}}.$$

As this is a valid correlation matrix, it must be positive semidefinite. On the other hand, given ϱ, $Z_{\text{known}} = (x_1, \dots, x_{k-1})$, and some candidate value y for Z_{k1}, assume that $\varrho_{\text{rest}} =: \varrho_{\text{rest}}(y)$ turns out positive semidefinite. As is well-known, $\varrho_{\text{rest}}(y)$ is then a feasible $k \times k$ correlation matrix, meaning it can be realized with m observations whenever $m > k$ (this follows, for instance, from Menger's 1928 Theorem on distance matrices [2]). Since $n-1 > k$, there thus exist some values $(Z'_{ia})_{i=1\dots k, a=2\dots n}$ which have $\varrho_{\text{rest}}(x)$ as their correlation matrix and which fulfil $\sum_{a=2}^n Z'_{ia} = -Z'_{i1}$ and $\sum_{a=2}^n {Z'_{ia}}^2 = (n -$

$1) - {Z'_{i1}}^2$. Together with $(Z'_1, \ldots, Z'_k) := (x_1, \ldots, x_{k-1}, y)$ they consequently build standardized data vectors $Z'_1, \ldots, Z'_k$ whose correlation matrix is ϱ. In other words: y is a possible value for Z_{k1} consistent with the given knowledge if and only if $\varrho_{\text{rest}}(y)$ is positive semidefinite. This will lead us to the proposed interval quite soon.

Note that $\varrho_{\text{rest}}(y)$ without its last row and column is just the correlation matrix of variables 1 to $k-1$ in observations 2 to n, hence it is positive semidefinite. Thus, $\varrho_{\text{rest}}(y)$ itself is positive semidefinite if and only if $\det \varrho_{\text{rest}}(y) \geq 0$. Putting $r_{ij} := \varrho_{ij}(n-1)^2/n$ for $i, j = 1 \ldots k$, the latter condition is equivalent to

$$\begin{aligned}
0 \;&\leq\; \det(r_{ij} - Z_{i1}Z_{j1})_{i,j=1\ldots k} = \begin{vmatrix} & \vdots & & \vdots \\ \cdots & r_{ij} - x_i x_j & \cdots & r_{ik} - x_i y \\ & \vdots & & \vdots \\ \cdots & r_{kj} - y x_j & \cdots & r_{kk} - y^2 \end{vmatrix} \\
&= y^2 \begin{vmatrix} & \vdots & & \vdots \\ \cdots & r_{ij} - x_i x_j & \cdots & x_i \\ & \vdots & & \vdots \\ \cdots & x_j & \cdots & -1 \end{vmatrix} - 2y \begin{vmatrix} & \vdots & & \vdots \\ \cdots & r_{ij} - x_i x_j & \cdots & x_i \\ & \vdots & & \vdots \\ \cdots & r_{kj} & \cdots & 0 \end{vmatrix} \\
&\quad + \begin{vmatrix} & \vdots & & \vdots \\ \cdots & r_{ij} - x_i x_j & \cdots & r_{ik} \\ & \vdots & & \vdots \\ \cdots & r_{kj} & \cdots & r_{kk} \end{vmatrix} =: \alpha y^2 - 2\beta y + \gamma.
\end{aligned}$$

A little linear algebra shows that

$$\begin{aligned}
\alpha \;&=\; \begin{vmatrix} & \vdots & & \vdots \\ \cdots & r_{ij} & \cdots & x_i \\ & \vdots & & \vdots \\ \cdots & 0 & \cdots & -1 \end{vmatrix} = -\left(\frac{(n-1)^2}{n}\right)^{k-1} \det \varrho_{\text{known}} \leq 0, \\
\beta \;&=\; \begin{vmatrix} & \vdots & & \vdots \\ \cdots & r_{ij} & \cdots & x_i \\ & \vdots & & \vdots \\ \cdots & r_{kj} & \cdots & 0 \end{vmatrix} = \alpha Z_{\text{known}}^\top (\varrho_{\text{known}})^{-1} \varrho_{\text{dep.}} = \alpha\zeta, \quad \text{and}
\end{aligned}$$

$$\begin{aligned}
\gamma &= \begin{vmatrix} & \vdots & & \vdots & \\ \cdots & r_{ij}-x_ix_j & \cdots & r_{ik} & x_i \\ & \vdots & & \vdots & \\ \cdots & r_{kj} & \cdots & r_{kk} & 0 \\ \cdots & 0 & \cdots & 0 & 1 \end{vmatrix} = \begin{vmatrix} & \vdots & & \vdots & \\ \cdots & r_{ij} & \cdots & r_{ik} & x_i \\ & \vdots & & \vdots & \\ \cdots & r_{kj} & \cdots & r_{kk} & 0 \\ \cdots & x_j & \cdots & 0 & 1 \end{vmatrix} \\
&= \left(\frac{(n-1)^2}{n}\right)^{k-1} \det \varrho \left(\frac{(n-1)^2}{n} - Z_{\text{known}}^{\top}(\varrho^{-1})_{\text{trunc.}} Z_{\text{known}}\right),
\end{aligned}$$

where

$$(\varrho^{-1})_{\text{trunc.}} \quad = \quad \big((\varrho^{-1})_{ij}\big)_{i,j=1\dots k-1} = \varrho^{-1} \text{ without its last row and column.}$$

Solving the inequality for y, we get the interval with bounds

$$\begin{aligned}
&\zeta \pm \sqrt{\zeta^2 + \frac{\det \varrho}{\det \varrho_{\text{known}}}\left(\frac{(n-1)^2}{n} - Z_{\text{known}}^{\top}(\varrho^{-1})_{\text{trunc.}} Z_{\text{known}}\right)} \\
&\quad = \quad \zeta \pm \sqrt{\frac{\det \varrho}{\det \varrho_{\text{known}}}} \sqrt{\frac{(n-1)^2}{n} - Z_{\text{known}}^{\top} M Z_{\text{known}}}, \quad \text{where} \\
M \quad &= \quad (\varrho^{-1})_{\text{trunc.}} - \frac{\det \varrho_{\text{known}}}{\det \varrho} (\varrho_{\text{known}})^{-1} \varrho_{\text{dep.}} \varrho_{\text{dep.}}^{\top} (\varrho_{\text{known}})^{-1}.
\end{aligned}$$

It turns out that M equals $(\varrho_{\text{known}})^{-1}$, since

$$\begin{aligned}
&\big((\varrho^{-1})_{\text{trunc.}} \varrho_{\text{known}}\big)_{ij} = \sum_{l=1}^{k-1} (\varrho^{-1})_{il} \varrho_{lj} = \delta_{ij} - (\varrho^{-1})_{ik} \varrho_{kj} \\
&= \quad \delta_{ij} - \frac{(-1)^{i+k}}{\det \rho} \det(\varrho_{ab})_{a \neq i, b \neq k} \cdot \varrho_{kj} \\
&= \quad \delta_{ij} - \frac{(-1)^{i+k}}{\det \varrho} \sum_{l=1}^{k-1} (-1)^{k-1+l} \varrho_{kl} \det(\varrho_{ab})_{a \notin \{i,k\}, b \notin \{k,l\}} \cdot \varrho_{kj} \\
&= \quad \delta_{ij} - \frac{(-1)^{i+k}}{\det \varrho} \sum_{l=1}^{k-1} (-1)^{k-1+l} \varrho_{kl} (-1)^{i+l} \det \varrho_{\text{known}} \big((\varrho_{\text{known}})^{-1}\big)_{il} \cdot \varrho_{kj} \\
&= \quad \left(1 + \frac{\det \varrho_{\text{known}}}{\det \varrho} (\varrho_{\text{known}})^{-1} \varrho_{\text{dep.}} \varrho_{\text{dep.}}^{\top}\right)_{ij}.
\end{aligned}$$

This finishes the proof of the theorem.

Let us designate the *"inference" interval* with bounds (1) by the symbol $I(\rho, Z_{\text{known}}, k, n)$. It is not surprising that its centre ζ is just the multilinear regression estimator of Z_{k1}, while the first factor of the interval width,

$\sqrt{\det \varrho / \det \varrho_{\mathrm{known}}}$, is the standard deviation of the corresponding residuals. The summand $Z_{\mathrm{known}}^{\top}(\varrho_{\mathrm{known}})^{-1}Z_{\mathrm{known}} =: \varepsilon_{k1}$ in the second factor can be interpreted as a measure of "extremity" of the first observation, corrected for correlations among the variables and ignoring its Z_k-value. It is non-negative and, when computing this measure for all other observations, too, its average is constant:

$$\begin{aligned}\sum_{a=1}^{n} \frac{\varepsilon_{ka}}{n} &= \sum_{i,j=1}^{k-1} \left((\varrho_{\mathrm{known}})^{-1}\right)_{ij} \sum_{a=1}^{n} Z_{ia}Z_{ja}/n \\ &= \sum_{i,j=1}^{k-1} \frac{1}{\det \varrho_{\mathrm{known}}}(-1)^{i+j} \det(\varrho_{ab})_{a\notin\{i,k\},b\notin\{j,k\}}(\varrho_{\mathrm{known}})_{ij}\tfrac{n-1}{n} \\ &= \sum_{i=1}^{k-1} \frac{1}{\det \varrho_{\mathrm{known}}} \det \varrho_{\mathrm{known}}\tfrac{n-1}{n} = (k-1)(n-1)/n.\end{aligned}$$

This implies that for a randomly chosen observation, the probability of a small inference interval is small. For example, less than ten percent of all observation's Z_k-values have an interval width below

$$2\sqrt{\tfrac{\det \varrho}{\det \varrho_{\mathrm{known}}}}\sqrt{n-10k+8}.$$

3 A viable protection algorithm

Now our task is this: publish the means, variances and correlations of Z_1, ..., Z_k in such a way that no-one can determine any of the values Z_{ka} more accurately than up to some δ, given the values of Z_{ia} for $i = 1\ldots k-1$. More precisely, we require that, given this knowledge, there always be at least two feasible values for Z_{ka} which differ by at least δ, the latter being a pre-specified "minimal uncertainty". A simple and fast procedure for this follows.

Given the standardized n-dimensional data vectors $Z_1, \ldots, Z_k$, and $\delta > 0$, we first compute ϱ, which runs in $O(nk^2)$ time. Then put

$$\begin{aligned} \overline{\varepsilon}_k &:= \frac{(n-1)^2}{n} - \frac{\delta^2}{4}\frac{\det \varrho_{\mathrm{known}}}{\det \varrho} \\ \text{and} \qquad \lambda_k &:= \text{smallest eigenvalue of } \varrho_{\mathrm{known}}. \end{aligned}$$

Now all values Z_{ka} for which ε_{ka} is at most $\overline{\varepsilon}_k$ have an inference interval width of at least δ. Since computation of ε_{ka} needs $O(k^2)$ time, the test for $\varepsilon_{ka} \leq \overline{\varepsilon}_k$ can be sped up considerably by first computing the value $\varepsilon'_{ka} := Z_{\mathrm{known}}^{\top}Z_{\mathrm{known}}/\lambda_k \geq \varepsilon_{ka}$, which only needs $O(k)$ time, and testing for $\varepsilon'_{ka} \leq \overline{\varepsilon}_k$.

Those Z_{ka} failing both tests are "at risk" and require special treatment. We can enlarge their inference interval by publishing the mean μ_k, the variance σ_k^2 and the correlations ϱ_{ik} only with some imprecision. For each such

Z_{ka}, replace Z_{ka} by $Z'_{ka} := Z_{ka} - \delta/2$ and re-calculate all these statistics, giving a triple $(\mu,\sigma,\varrho)'_{ka}$. Likewise, calculate the triple $(\mu,\sigma,\varrho)''_{ka}$ using $Z''_{ka} := Z_{ka} + \delta/2$ instead. Since the original statistics are known, these calculations can be done in $O(k)$ time.

Simultaneously, calculate the element-wise minimum and maximum of (μ,σ,ϱ), all $(\mu,\sigma,\varrho)'_{ka}$, and all $(\mu,\sigma,\varrho)''_{ka}$, resulting in bounds $(\underline{\mu},\underline{\sigma},\underline{\varrho})$ and $(\overline{\mu},\overline{\sigma},\overline{\varrho})$ which can finally be published.

Given this published information, no-one can determine some Z_{ka} more accurately than up to δ because then each $(\mu,\sigma,\varrho)'_{ka}$ and $(\mu,\sigma,\varrho)''_{ka}$ remains a possible version of the statistics, hence each Z'_{ka} and Z''_{ka} remains a possible value of Z_{ka}.

The whole algorithm runs in two sequential passes through the data, both needing $O(k^2)$ space and at most $O(nk^2)$ time. When we protect the other variables $Z_1,\dots,Z_{k-1}$ also in the same way in the second pass, its time complexity increases to $O\big(nk(1+k(r_1+\dots+r_k))\big)$, where $r_i \le 1$ is the proportion of observations failing the test $\varepsilon'_{ia} \le \overline{\varepsilon}_i$. Because ε'_{ia} averages to $(k-1)(n-1)/n\lambda_i$, we have $r_i < k/\lambda_i\overline{\varepsilon}_i$.

4 More additional knowledge

The above considerations only hold when the assumed additional knowledge is restricted to the values of one target individual. In practice, though, several other things are also known about the data.

Additionally known observations. First of all, some additional individuals' data might be known. This will be the case when the person interested in the data himself belongs to the sample, or when certain "prominent" members of the sample are publicly well-known. One can adapt the theorem to this case by replacing ϱ by the correlation matrix of the sample without the known observations, which can be computed from ϱ and these observations. But since it is not clear beforehand which observations might be known, a corresponding protection algorithm would have to consider all possible sets of, say, s additionally known observations — resulting in a time complexity increased by a factor of $O(n^s)$.

Boundary conditions. Secondly, often statistical data are known to be non-negative, bounded, or integer multiples of a certain unit. Hence, given the means and variances, one often knows Z_{ia} to belong to some set S_i of a priori feasible values. In this case, the theorem cannot be adapted as easily since it is not clear which matrices are feasible correlation matrices of variables with such restrictions.

When S_i is the set of integers, one could try to modify the theorem and algorithm by considering how much ϱ_{rest} may change when its realizing data vectors Z'_i were rounded to the nearest integer. This could then be transformed into some suitable additional imprecision for the publication.

It seems that the resulting additional imprecision needed for ϱ_{ij} is approximately $(2\varrho_{ij}^2 + |\varrho_{ij}|)(\sigma_i^{-1} + \sigma_j^{-1}) \leq 6/\sigma_{\text{min.}}$.

There is also a trivial way to protect confidentiality in cases with boundary conditions. Apply the algorithm, but assume all cells "at risk" and choose Z'_{ia} and Z''_{ia} from S_i always. The resulting imprecision is easily seen to be of order $O(1/n)$, hence this is a viable method when n is large enough.

A somewhat more sophisticated version of the latter involves some compensation of the effect of replacing Z_{ka} by $Z_{ka} \pm \delta/2$. When we find an observation b which is distinct from but similar to a, and simultaneously replace Z_{kb} by $Z_{kb} \mp \delta/2$, we can hope for the effects of both replacements to eliminate each other. In case of small n, one can use always the best such b, but this increases the running time by a factor of n. For large n, we would therefore apply some heuristic to find b instead. In analogy to the spacefilling curve heuristic for the travelling salesman problem ([1]), one could use a smooth spacefilling curve (such as the Hilbert curve) to sort the data beforehand, and then simply take b to be the observation preceding a. In each step, Z_{ka} and Z_{kb} are temporarily replaced by $Z'_{ka} := Z_{ka} - \delta/2$ and $Z''_{kb} := Z_{kb} + \delta/2$, giving a triple $(\mu, \sigma, \varrho)'_{kab}$. As before, the element-wise extreme values of the latter then give the publication bounds $(\underline{\mu}, \underline{\sigma}, \underline{\varrho})$ and $(\overline{\mu}, \overline{\sigma}, \overline{\varrho})$.

5 Outlook

The results and ideas presented in this paper indicate that confidentiality protection is possible when publishing results of multivariate analyses. For analyses which rely only on first and second moments of esentially unbounded and real-valued data, the algorithm described in section 3 can be used to ensure confidentiality. For other cases, including that of variables with boundary conditions, but probably including also other kinds of statistical measures to be published, the compensation algorithm suggested at the end of the preceding section might be used.

In order to assess the practical performance and efficiency of these algorithms, both thorough testing and additional theoretical work are needed. The former should involve various types of data and sample sizes, including dummy variables for categorical data. The latter could try to find bounds for the required publication imprecision and for the resulting error of dependent statistics such as test statistics, for instance. Even more important, what is the inference interval for an individual data cell when publishing other common statistical measures such as certain quantiles, moments of higher order, rank correlation coefficients, etc.?

Answers to these questions will help us further improve scientists' access to official but confidential microdata.

References

[1] Bartholdi J. J. III and Platzman L. K. (1988). *Heuristics based on space-filling curves for combinatorial problems in Euclidean space*. Management Science, **34**, 291 – 305.

[2] Menger Karl (1928). *Untersuchungen über allgemeine Metrik*. Math. Annalen **100**, 75 – 163.

Address: Statistisches Bundesamt, IT User Service / Statistical and Geographical Information Systems, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany

E-mail: `jobst.heitzig@destatis.de`

CLASSIFICATION AND OUTLIER IDENTIFICATION FOR THE GAIA MISSION

Christian Hennig

Key words: Support vector machine, star survey, quasars.

COMPSTAT 2004 section: Statistical data mining.

Abstract: The GAIA satellite is scheduled for launch in 2010. GAIA will observe spectral data of about 1 billion celestial objects. Part of the preparation of the GAIA mission is the choice of an efficient classification method to classify the observed objects automatically as stars, double stars, quasars or other objects. For this reason, there have been two blind testing experiments on simulated data. In this paper, the blind testing procedure is described as well as the results of a cross-validation experiment to choose a good classifier from a broad class of methods, comprising, e.g., the support vector machine, neural networks, nearest neighbor methods, classification trees and random forests. Because of a lack of information about their nature, no outliers ("other objects"-class) have been simulated. A new strategy to identify outliers based on only "clean" training data independent of the chosen classification method is proposed.

1 The GAIA mission and its classification tasks

When launched in 2010, the GAIA satellite will undertake a very detailed and extensive astrometric and photometric study of our Galaxy with the primary goal to determine its formation, composition and evolution. GAIA will observe some 1 billion stars, galaxies, quasars and solar system objects, and there are also numerous supplementary science projects ranging from exo-solar planets to fundamental physics. General information about GAIA can be found at `http://astro.estec.esa.nl/GAIA/`.

A main goal of the GAIA classification working group on identification, classification and parametrization (ICAP) is the development of a methodology to classify the observed objects into some general classes, the members of which have to be treated differently in the following astronomical investigations such as determination of astrophysical parameters. The primary information for classification purposes will come from a set of 5-15 medium band photometers (the system is not fixed yet), each generating a photon count number, see [2] for details. A basic problem is that any existing data on celestial objects deviates from the data GAIA will deliver because of the differing observation conditions. Therefore, information of existing data and models of GAIA's photometric instruments have been combined to simulate calibration data. To prevent the unintended use of possibly inadequate background information, the ICAP designed some blind testing experiments in which about ten scientists (among which I have been the only non-astronomer) or groups got training data (in the following referred to as "data A") with known classes ("star", "double star", "quasar" and "other") and test data ("data B") with unknown classes to classify with any method of choice. The data sets are described in Section 2. These experiments should not only assess the classification algorithms but also possible filter systems, about which a decision should be made in the near future.

The ICAP is also concerned with further tasks, namely more (and less clearly described) classes, spectral data with high dimensionality and determination of astrophysical parameters, see [2] and the ICAP homepage `http://www.mpia-hd.mpg.de/GAIA/`. But in the present paper, the focus is on the blind testing experiments, which will be described in Section 2. To find a good classifier, I compared a lot of methods by a simple cross-validation scheme. The methods and results are given in Section 3. I used the methods as implemented in the statistical freeware system R (`www.R-project.org/`) with almost no parameter tuning, so that the comparison reflects the requirements of a user with a limited amount of time not using particular expertise about any of the methods, as will often be the case in applied data mining. A comparison of my results applied to data B compared to the results of other participants along with a rough description of their methods can be found in [3], [4]. Because of multiple quality criteria and data sets, the experiments did not have a competition character, but my classification results have been almost always good and often the best.

The identification of objects that do not come from any known class (be it erroneous observations or interesting new types of objects) will be extremely important in the GAIA mission. A decision rule for this task is proposed in Section 4, which is based on the Mahalanobis distance to the nearest training object of the class to which the new object is classified. Unlike other methods for outlier identification in classification such as the atypicality index [1] and one-class support vector machine [6], this method is not associated to any particular classification method.

2 The data for blind testing

Up to now, there have been two cycles of blind testing experiments. In the first cycle, there have been only quasars and stars, in the second cycle there have also been double stars. I focus on the description of the second cycle. The data sets have been generated by a combination of observations of real objects and simulation. Spectral information for quasars and double stars has been selected randomly from catalogs of real objects, while the stars have been simulated to cover a grid of astrophysical parameter values of interest. A model of the photometric observation instruments of GAIA has been applied and additional noise (background noise and observation errors)

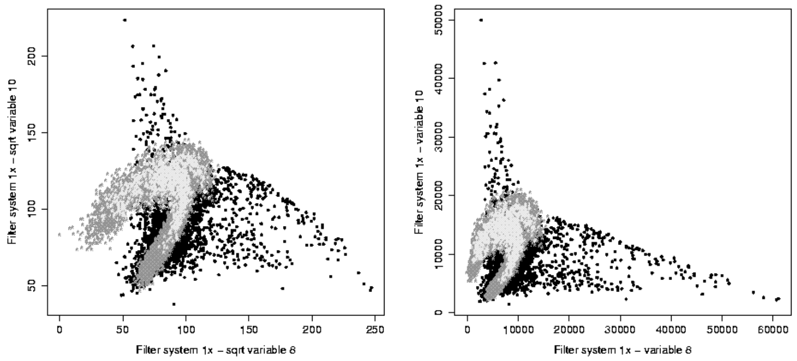


Figure 1: Variables 8 and 10 of data A, filter system 1X, magnitude $G = 19$. Left: photon counts, right: square root of photon counts. Quasars are black (plotted firstly, thus partly overplotted), stars are dark gray (plotted secondly), double stars are bright gray.

has been added to simulate the photon counts recorded by the filters. For more details see [5].

Since the simulated object brightness (magnitude) has been homogeneous in every data set, there have been eight data sets, namely four simulated filter systems applied to four different magnitudes with two filter systems for each magnitude. All data sets have been generated from the same basic objects, namely 20000 stars, 4000 quasars and 1000 double stars (except of the brightest magnitude, where no quasars have been generated) for data A and 41400 stars, 40000 quasars and 6240 double stars for data B (this distribution was not known to the testers). Data A has been provided with 20 different realizations of the observational error for each object and even error free, but after some experiments I decided only to use a single error version. The eight data sets have been analyzed separately, informations from different filter systems have not been combined. Note that the real distribution of observable stars:quasars:(double stars) is estimated as about 1:(8e-4):1, but the distribution used in the blind testing experiments was governed also by other reasons, namely the availability of reliable models and observed objects and scientific goals such as separating the quasars. In particular, there is no physical model to generate "other objects".

The filter systems generated 6-12 photon count variables. Dimension reduction did not improve the classification results, as could be expected from the favorable relation n/p.Inspection of all two-dimensional scatterplots and high-dimensional "rotation" with the grand tour (using the R-package `xgobi`) showed different multivariate nonlinear relationships between the variables for different classes and heterogeneous distributional shapes. As an example, a scatterplot of two variables is shown in Figure 1, left side. Typical features of this scatterplot are the large area occupied by some extreme quasars and the fact that the double stars cannot be separated from the single stars. Since all one-dimensional distributions are more or less skew (all classes taken together), I decided to transform all variables to square roots, which yielded better classification results than the raw data and than some log-transformed data for all but the transformation invariant classifiers, compare the right side of Figure 1.

3 Comparison of classification methods on data A

In this section I describe my comparisons to find a good classifier for data set B given data set A with 25000 observations. Other testers used either nearest neighbor methods or neural networks after some preprocessing using astronomical background information, while I used only the information in the data, see [3], [4].

Because of tight deadlines and relatively large data sets, I compared all methods by a simple split of the data set A into twice 12500 points, of which I used the first half as training set and the second half as test set. In some situations, this has been repeated up to 10 times to assess the variability of the results. I used the following classification algorithms for cycle 2:

- quadratic discriminant analysis ("QDA" in Table 1),
- 1-, 3- and 7-nearest neighbors (Euclidean distances; "1-NN, 7-NN"),
- classification trees from R-package `rpart` ("tree"),
- neural networks from R-package `nnet` ("NNet"),
- support vector machines from R-package `e1071` ("SVM"),
- random forests from R-package `randomForest` ("RFor").

Not all classifiers have been compared on all data sets; some methods which worked poorly in cycle 1 (boosting of stumps, multiple additive regression splines) have not been tested again in cycle 2, others only on a single data set. In most cases, I did not change any tuning parameters. Only for the support vector machine, the `cost`-parameter has been set to 50, and some tuning has been done with the neural networks, because the initial results have been very bad. Different transformations of the data (square root, $\log(x + 1)$, zero mean/unit covariance matrix, the robust analogue with the minimum covariance determinant method (MCD), separate standardization of all variables to zero mean/unit variance, zero median, unit MAD, also after taking square roots) have been compared together with some of the classifiers. The square root transformation has generally been optimal. The median/MAD-standardization has been optimal for all methods for the data in Table 1, and often, but not always, for other filter systems and magnitudes.

Star sample ($n_{true} = 10003$)			
method	size	% mis	% OK
SVMmm	10432	4.6	99.4
SVMmcd	10241	5.0	97.2
RFor	10513	5.3	99.5
1-NN	10228	5.7	96.4
NNet	10530	5.8	99.1
SVM-q	10543	6.0	99.1
QDA	9730	6.3	91.1
7-NN	10652	6.7	99.3
RFor-q	10720	6.9	99.8
tree	10779	9.3	97.8

Quasar sample ($n_{true} = 1993$)			
method	size	% mis	% OK
SVM-q	1737	0.1	87.1
RFor-q	1686	0.2	84.4
7-NN	1668	0.4	83.4
SVMmm	1878	0.5	93.7
RFor	1884	1.6	93.0
1-NN	1814	2.0	89.2
NNet	1915	3.4	92.8
SVMmcd	1990	5.0	94.8
QDA	1881	8.0	86.9
tree	1721	13.0	75.1

Double star sample ($n_{true} = 504$)			
method	size	% mis	% OK
RFor	103	18.4	16.7
RFor-q	94	20.2	14.9
SVMmm	187	25.1	27.8
7-NN	180	34.4	23.4
NNet	55	41.8	6.3
SVM-q	220	43.6	24.6
1-NN	458	70.7	26.6
SVMmcd	269	71.7	15.1
QDA	829	89.9	16.7
tree	0	0/0	0.0

Table 1: The "$\mathcal{C}$ sample" ($\mathcal{C}$ = star/quasar/double star) denotes all objects classified as $\mathcal{C}$ (from a 12500 points test subsample of data set A, filter system 1X, $G = 19$). The columns show the size of the $\mathcal{C}$ sample, the percentage of non-$\mathcal{C}$-objects classified as $\mathcal{C}$ with respect to the $\mathcal{C}$ sample and the percentage of $\mathcal{C}$-objects classified as $\mathcal{C}$ with respect to all true $\mathcal{C}$-objects. The tables are ordered according to "% mis". The n_{true} is the true number of objects from class $\mathcal{C}$.

Exemplary, in Table 1, the SVM results are given with the median/MAD ("SVMmm") and the worst (MCD; "SVMmcd") standardization.

To choose the best methods for application to data set B, I used the number of misclassifications as a target criterion. This led in all cases to the SVM with one exception in cycle 1, where 1-NN (with MCD transformation) has been optimal. No unique target criterion can be defined on astronomical grounds, because the classification result will be used in connection with different scientific goals. An important goal is to build samples of quasars and double stars with as small contamination as possible. For a given class $\mathcal{C}$, A. Brown [4] used an estimator of the ratio of the number of non-members of

$\mathcal{C}$ classified as $\mathcal{C}$ to the whole number of objects classified as $\mathcal{C}$, corrected for the real distribution of the classes (see Section 2), as a criterion to evaluate the second blind testing cycle. The statistics given in Table 1, namely the contamination percentage and the percentage of correctly classified $\mathcal{C}$-objects, are closely related to this criterion. It yields different optimality results for the three classes. In Table 1, two results are given where class weights have been chosen for the SVM ("SVM-q") and random forests ("RFor-q") to optimize Brown's measure for quasars. The resulting SVM with class weights $1 : 0.15 : 1$ was applied to data set B after the end of cycle 2 and yielded the optimal result classifying only one non-quasar as quasar while positively identifying about 86% of the quasars.

To interpret the results, the nonlinear structure of the problem and the strong interactions between variables are important. The distributions of quasars and non-quasars are very different, so that a good separation of these classes is in principle possible. It is clear that the QDA and the simple tree are not adequate to cope with these features. The problem is structured enough that improvements over the NN methods are possible. It may be possible to design better neural networks for this task then I did, but the tuning problem turned out to be much harder than for the SVM and for the forest. The SVM is able to locate well the nonlinear decision boundary between quasars and non-quasars. It loses its optimality for the double stars, which are not by any means properly separable from the stars. It seems that the forest, which does not directly fit a boundary, is better to find some interactions of variables that are at least a bit useful here.

4 Outlier identification from clean training data

Though no outliers have been generated, the blind testers have been encouraged to classify objects in an "other"-class. I intended to do this in a way which could be applied on real future GAIA observations independent of the classification algorithm that will be finally chosen. Characteristics of the present situation are that there is neither an outlier model, nor are there training outliers (apart from extreme but proper classifiable points), and that the distributional shapes of the classes are very different and do not satisfy simple assumptions such as being elliptical.

The proposed procedure is as follows:

1. Compute the covariance matrix of the training data in each class.

2. For each point in data set A, compute the Mahalanobis distance (w.r.t. the covariance matrix computed above) to the nearest neighbor of its own class in data set A.

3. For each point in data set B, take the class to which it is assigned by the chosen classification method and compute the Mahalanobis distance (as above) to the nearest neighbor of this class in data set A.

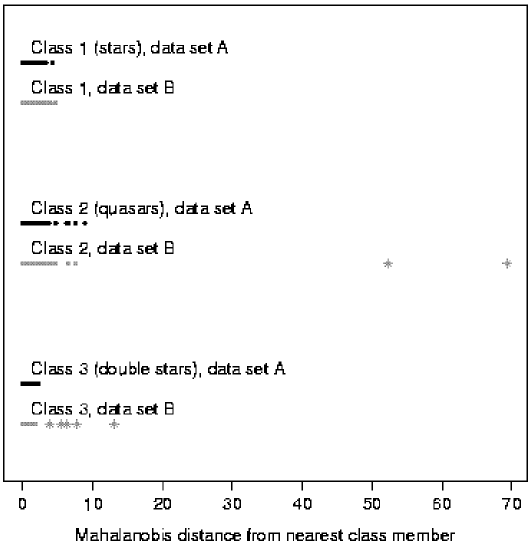


Figure 2: Mahalanobis distances to nearest data A neighbor of the class the point belongs to (data A) or is classified to (data B) for filter system 2B, $G = 20$. Stars denote points classified as outliers.

4. Compare the distributions generated by 2 and 3 separately for each class and mark extreme points in data set B as outliers.

Because of the different local variance structures of quasars and stars, it seems reasonable to perform an outlier detection class-wise, i.e., each point of data B is classified first, and then it is compared only to training points of the resulting class. This has the further advantage that points that are erroneously detected as outliers by this procedure have already been misclassified with high probability by the initial classifier. Nearest neighbors are used because of the hypothesis that true class members can occur in the neighborhood of each proper training point.

For the blind testing cycle 2, I carried out the comparison of distributions by graphical means, see Figure 2 (because of space restrictions, I do not show separate graphics for all classes). I marked outlier candidates only in three data sets out of eight (recall that there were no real outliers). The data set shown in Figure 2 contained by far the most and the most extreme candidates. For realtime use with GAIA, the astronomers should specify a tolerable percentage of false positives, and enough training data should be generated to compute the corresponding quantiles of the distance distributions, so that the outlier identification can be done automatically. Note that this assumes that the real within-class distributions can be simulated with arbitrarily many training points. While it is possible to generate any required number of points, the basic objects for generating quasars and double stars stem from a finite catalog, so that not the whole area of their distribution might be covered.

5 Conclusion

While lots of classification methods have been compared and some valuable insights have been gained, the GAIA blind testing experiments leave open some interesting problems. What are reasonable class proportions for data sets A and B of such experiments, given the mentioned conditions? How can outliers be simulated, and which loss function should be used to assess outlier identification procedures? Should different classification methods be used for different scientific goals and/or for different object magnitudes?

The recent observation models are not based on any real GAIA observations. Therefore it will be extremely important to assess the difference between the real observed objects and the models, and to adjust the classification methods after GAIA has been launched.

References

[1] Aitchison J., Habbema J., Kay J. (1977). *A critical comparison of two methods of statistical discrimination.* Applied Statistics **26**, 15 – 25.

[2] Bailer-Jones C.A.L. (2003). *On the classification and parametrization of GAIA data using pattern recognition methods.* GAIA Spectroscopy, Science and Technology, ASP Conference Series **298**, U. Munari (ed.), Astronomical Society of the Pacific, San Francisco, 199 – 208.

[3] Brown A. (2003a). *Results of the first cycle of blind testing.* GAIA document ICAP-AB-003.
www.mpia-hd.mpg.de/GAIA/ICAP_documents/ICAP-AB-003.pdf

[4] Brown A. (2003b). *Results of the second cycle of blind testing.* GAIA document ICAP-AB-004.
www.mpia-hd.mpg.de/GAIA/ICAP_documents/ICAP-AB-004.pdf

[5] Jordi C., Carrasco J. M., Figueras F., Bailer-Jones C. A. L. (2003). *Simulated GAIA photometry for blind testing cycle 2.* GAIA document UB-PWG-014.
gaia.am.ub.es/PWG/photo_sim/blindtesting_cycle2/ReadMe.UB-PWG-014

[6] Tax D. M. J., Duin R. P. W. (1999). *Support vector data description.* Pattern Recognition Letters **20**, 1191 – 1199.

Address: C. Hennig, Faculty of Mathematics - SPST, University of Hamburg, Bundesstr. 55, D-20146 Hamburg, Germany

E-mail: hennig@math.uni-hamburg.de

TESTING THE EQUALITY OF THE ODDS RATIO PARAMETERS IN THE K ORDERED 2×2 TABLES

C. Hirotsu, E. Ohta and S. Aoki

Key words: Cumulative chi-squared statistic, exact algorithm, isotonic inference, maximally selected chi-squared, odds ratio, risk analysis.

COMPSTAT 2004 section: Biostatistics.

Abstract: This paper deals with the use of the isotonic inference for the comparison of odds ratio parameters in K ordered 2×2 tables. The emphasis is on the algorithmic side of the problem enabling reliable calculation.

1 Introduction

There are a lot of papers for the isotonic inference on the K ordered normal means but rather a few papers dealing with the ordered alternative for the interaction effects especially in the discrete models. It is probably due to the computational difficulty in the well known restricted maximum likelihood approach related to those complicated cases. On the other hand Hirotsu [3], [4], [6], [8], [9] developed the cumulative chi-squared method for the ordered alternatives in normal means which is sufficiently powerful against a wide range of the simple ordered alternative and yet easily extended to more complicated cases including the interaction problems. The method has also been extended to the discrete models in Hirotsu [5], [7], [10] and Hirotsu et al. [11]. In this paper we extend the idea to comparing odds ratio parameters in K ordered 2×2 tables, namely to the isotonic inference on the 3-way interaction. Comparing odds ratios is very common in the epidemiology and the risk analysis, e.g. if we have 2×2 tables for the occurrence of the lung cancer vs the existence of the smoking habit stratified by the age groups then we are interested in testing the effects of age on the odds ratios between cancer and smoking.

We give the mathematical basis of the problem in the next section, and the approaches based on the cumulative chi-squared statistic and the maximally selected accumulated statistic are given in Sections 3 and 4, respectively. In particular an exact algorithm dealing with the distribution function of the maximally selected statistic is given in Section 5. Finally in Section 6 we show that the powers of these methods are satisfactory as compared with the restricted maximum likelihood method by Barmi [1] and a real example is given in Section 7.

2 Mathematical formulation

Let y_{ijk} be the observed frequency at the ith row and jth column of the kth 2×2 table, $i, j = 1, 2;\ k = 1, \cdots, K$. We assume a natural ordering in the K tables so that we are interested in testing the simple ordered alternative

$$H_1 : \eta_1 \leq \eta_2 \leq \cdots \leq \eta_K \tag{1}$$

in the log odds ratio parameters

$$\eta_k = \log \frac{p_{11k}p_{22k}}{p_{12k}p_{21k}},$$

where the p_{ijk} is the occurrence probability at the (i, j, k) cell. The related probability density function of the y_{ijk} is

$$f(\boldsymbol{y} \mid y_{ij.}, y_{i.k}, y_{.jk}) = C^{-1}(\boldsymbol{\eta}, y_{ij.}, y_{i.k}, y_{.jk}) \frac{\exp\{\sum_k y_{11k}\eta_k\}}{\prod_{i,j,k} y_{ijk}!}, \tag{2}$$

where $\boldsymbol{y}$ is the vector of y_{ijk} arranged in the dictionary order, $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_K)'$ and we can assume $\eta_K = 0$ without any loss of generality. Then the equality of the log odds ratio parameters is equivalent to the null hypothesis,

$$H_0 : \eta_1 = \eta_2 = \cdots = \eta_K = 0.$$

The density (2) is essentially the conditional distribution of $(y_{111}, \cdots, y_{11K})$ given all the two-way marginal totals $y_{ij.},\ y_{i.k},\ y_{.jk}$. Now according to the complete class lemma given in Hirotsu [5] the complete class of tests for the ordered alternative H_1 is given by all the tests that are increasing in every element of the accumulated efficient score vector $\boldsymbol{Y}$ evaluated at the null hypothesis $H_0; \boldsymbol{Y} = (Y_{111}, \ldots, Y_{11,K-1}),\ Y_{11k} = \sum_{l=1}^{k} y_{11l}$. Among these tests are the cumulative chi-squared statistic (χ^{*2}) based on the sum of squares in Y_k^*, the standardized version of Y_{11k}, which we develop in Section 3 and the maximal accumulated chi-statistic (max acc. χ) based on the maximally selected element of $(Y_1^*, \cdots, Y_{k-1}^*)$, which we develop in Section 4.

The χ^{*2} is useful essentially for the two-sided version H_2 of the ordered alternative H_1 (1) and has a very good chi-squared approximation for the p value calculation. The max acc. χ can be applied to both of H_1 and H_2 and has an exact algorithm for evaluating the p value if the contingency table is not too large. For a larger table it has also a very efficient algorithm based on the asymptotic normal distribution.

3 The cumulative chi-squared statistic χ^{*2}

The cumulative chi-squared statistic χ^{*2} is defined by

$$\chi^{*2} = \sum_{k=1}^{K-1} Y_k^{*2} \tag{3}$$

where Y_k^* is the standardization of Y_{11k} so as to have 0 mean and unit variance under H_0. In the following we derive the asymptotic null distribution of χ^{*2}. First the asymptotic null distribution of the y_{ijk} given $y_{ij.},\ y_{i.k},\ y_{.jk}$ is the normal with the mean vector $\boldsymbol{m}$ obtained by the well known iterative scaling procedure starting from all the elements m_{ijk} to be unity and adjusting the marginal totals $m_{ij.},\ m_{i.k},\ m_{.jk}$ to $y_{ij.},\ y_{i.k},\ y_{.jk}$ iteratively. The asymptotic variance of $\boldsymbol{y}$ is given by

$$\boldsymbol{V} = \boldsymbol{V}_0 - \boldsymbol{V}_0\boldsymbol{L}(\boldsymbol{L}'\boldsymbol{V}_0\boldsymbol{L})^{-1}\boldsymbol{L}'\boldsymbol{V}_0, \tag{4}$$

where $\boldsymbol{V}_0$ is a diagonal matrix with m_{ijk} as its $(i-1)2K+(j-1)K+k$th element and $\boldsymbol{L}$ a matrix for $\boldsymbol{L}'\boldsymbol{y}$ to form a set of sufficient statistics $y_{ij.},\ y_{i.k},\ y_{.jk}$ under H_0, $i=1,2,\ j=1,2,\ k=1,\cdots,K$, see [13] for details.

Based on $\boldsymbol{V}$ (4) the variance of Y_{11k} is calculated as

$$V_k = \mathrm{Var}(Y_{11k}) = \boldsymbol{c}_k'\boldsymbol{U}\boldsymbol{c}_k,$$

where $\boldsymbol{c}_k' = (1,\cdots,1,0,\cdots,0)$ is a vector with 1's as its first k elements and 0's as its last $K-k$ elements and an explicit form of $\boldsymbol{U}$ is given in (A.2). Then we have $\boldsymbol{Y}^* = (Y_1^*,\cdots,Y_{K-1}^*)' = \mathrm{diag}(V_k^{-1/2})\boldsymbol{C}'(y_{111}-m_{111},\cdots,y_{11K}-m_{11K})'$ with $\boldsymbol{C}' = (\boldsymbol{c}_1,\cdots,\boldsymbol{c}_{K-1})'$. Thus the asymptotic null distribution of $\boldsymbol{Y}^*$ is a normal with mean vector $\mathbf{0}$ and variance matrix

$$V(\boldsymbol{Y}^*) = \mathrm{diag}(V_k^{-1/2})\boldsymbol{C}'\boldsymbol{U}\boldsymbol{C}\mathrm{diag}(V_k^{-1/2}), \tag{5}$$

where $\mathrm{diag}(\delta_k)$ is a diagonal matrix with δ_k as its kth diagonal element. The statistic χ^{*2} (3) is a positive quadratic form in normal variables and is therefore well approximated by a constant times chi-squared variable $d\chi_\nu^2$ with d.f. ν, where the constants d and ν are obtained by adjusting first two moments.

4 Maximally selected accumulated χ statistic and its distribution function

The maximally selected accumulated χ statistic, max acc. χ, is defined by max acc. $\chi = \max_k Y_k^*$.

For calculating the p value of max acc. χ we have an exact and very elegant algorithm, which is based on the Markov property of the subsequent Y_k^*'s. To show it we first derive the distribution function of $\boldsymbol{Y}$ which is easily converted to that of $\boldsymbol{Y}^*$. A special attention for handling the distribution function of $\boldsymbol{Y}$ is needed since the amalgamation invariance does not hold for the no three-way interaction model unlike in the case of two-way interaction, see Darroch [2]. We therefore directly factorize the simultaneous null distribution of $\boldsymbol{y}$ with respect to the accumulated statistic $\boldsymbol{Y}$ as follows. The null distribution of $\boldsymbol{y}$ given the fixed marginal totals is given by (2) with $\boldsymbol{\eta} = \mathbf{0}$, where $C(\mathbf{0}, y_{ij.}, y_{i.k}, y_{.jk}) = \sum \prod_{i,j,k} (y_{ijk}!)^{-1}$ with the summation with respect to y_{ijk} subject to the given marginal totals. Then it is factorized into

$$G(\boldsymbol{Y}) = \left\{ \prod_{k=2}^{K} F(Y_{11,k-1}|Y_{11k}) \right\} \times g(Y_{11K}), \tag{6}$$

where

$$F(Y_{11,k-1}|Y_{11k}) = C_k^{-1} C_{k-1} \prod_{ij} \{(Y_{ijk} - Y_{ij,k-1})!\}^{-1}$$

with $Y_{ijk} = \sum_{l=1}^{k} y_{ijl}$ is the conditional distribution of $Y_{11,k-1}$ given Y_{11k} and $g(Y_{11K}) = C^{-1} C_K(Y_{11K})$ is actually a constant since $Y_{11K} = y_{11.}$. Further C_1 is defined as $C_1 = \prod_{ij} (y_{ij1}!)^{-1}$ and other C_k's are defined recursively by $C_k = \sum_{Y_{11,k-1}} C_{k-1} \prod_{ij} \{(Y_{ijk} - \ Y_{ij,k-1})!\}^{-1}$ so that C_k is a function of Y_{11k}. The form (6) proves the Markov property in Y_{11k}, which immediately shows the Markov property in Y_k^*.

5 Exact and asymptotic algorithms for calculating the p value of max acc. χ

Let a random variable Z denote the max acc. χ. Then the distribution function of Z is defined by

$$H(z) = \Pr\{Z \leq z\} = \Pr(Y_1^* \leq z, \cdots, Y_K^* \leq z | y_{ij.}, y_{i.k}, y_{.jk}),$$

where Y_K^* is defined to be 0 for convenience.

Define the conditional simultaneous distribution up to Y_k^*,

$$H_k(z|Y_k^*; y_{ij.}, y_{i.k}, y_{.jk}) = \Pr(Y_1^* \leq z, \cdots, Y_k^* \leq z | Y_k^*; y_{ij.}, y_{i.k}, y_{.jk}).$$

Then we have a recurrence formula for the distribution function $H(z) = H_K(z|Y_K^* \ ; \ y_{ij.}, y_{i.k}, y_{.jk})$ as in Lemma 1.

Lemma 1. $H_k(z|Y_k^* \ ; \ y_{ij.}, y_{i.k}, y_{.jk})$ is equal to

$$\begin{cases} \sum_{Y_{k-1}^*} H_{k-1}(z|Y_{k-1}^*; y_{ij.}, y_{i.k}, y_{.jk}) \Pr\{Y_{k-1}^*|Y_k^*\}, \text{ if } Y_k^* \leq z, \\ 0, \text{ otherwise for } k = 2, \cdots, K. \end{cases} \tag{7}$$

Proof is essentially the same as that of Lemma 2 of Hirotsu and Marumo [12] and omitted.

Starting from the initial function $H_1(z|Y_1^* \ ; \ y_{ij.}, y_{i.k}, y_{.jk}) = 1$ if $Y_1^* \leq z$ and 0 otherwise, we can apply the recurrence formula (7) until $H(z)$ is obtained at $k = K$. Actually we go back to Y_{11k} from Y_k^* for convenience in applying the formula.

For a larger table where the exact algorithm doesn't work well we already have the asymptotic normal distribution for $\boldsymbol{Y}^*$ with mean vector $\mathbf{0}$ and variance matrix $V(\boldsymbol{Y}^*)$ given in (5). Then the recurrence formula (7) works almost as it is just by altering the summation into integration and avoiding a time consuming multiple integration in calculating the p value.

6 Power comparisons

We compare the powers of χ^{*2} and max acc. χ with that of the likelihood ratio test (lrt) by Barmi [1]. We give in Table 1 only the results for the balanced cases 1.$m_{ijk} \equiv 25$(top), 2.$m_{ijk} \equiv 20$(middle) and 3.$m_{ijk} \equiv 15$(bottom).

The 5% points have been chosen by the chi-squared approximation for χ^{*2}, exactly for max acc. χ and by simulation for the lrt. For the unbalanced case there is a slight reduction of power but the relative efficiencies among those three methods are almost the same with the balanced case. From Table 1 we see that the cumulative chi-squared method keeps relatively high power against other two methods for a wide range of the ordered alternative. The max acc. χ and the lrt behave rather similarly and are sometimes better than χ^{*2} for the simple change point type hypothesis (3), (4) or (5). The figures for H_0 are actually the type I error rate and show that the chi-squared approximation for χ^{*2} is satisfactory. The exact algorithm is inevitably sometimes considerably conservative for too small a table.

7 Concluding remarks

The three methods are applied to the same data as treated by Barmi for testing the effects of age on the odds ratios between breathlessness and wheeze to give p values 0.0001 for χ^{*2}, 0.0004 for max acc. χ and 0.0002 for the lrt. From this and other examples, and also by the simulation for power comparisons we conclude that the three methods are rather similar but the cumulative chi-squared statistic is most stable against a wide range of the simple ordered alternative.

References

[1] El Barmi, H.(1997). *Testing for or against a trend in odds ratios.* Commun. Statist. Theory Math **26**, 1877 − 1891.

[2] Darroch, J. N.(1974). *Multiplicative and additive interaction in contingency tables.* Biometrika **61**, 207 − 214.

[3] Hirotsu C. (1978). *Ordered alternatives for interaction effects.* Biometrika **65**, 561 − 570.

[4] Hirotsu C. (1979). *The cumulative chi-squares method and a studentized maximal contrast method for testing an ordered alternative in a one-way analysis of variance model.* Rep. Statist. Appl. Res., JUSE **26**, 12 − 21.

[5] Hirotsu C. (1982). *Use of cumulative efficient scores for testing ordered alternatives in discrete models.* Biometrika **69**, 567 − 577.

[6] Hirotsu C. (1983a). *An approach to difining the pattern of interaction effects in a two-way layout.* Annals Inst. Statist. Math. **A35**, 77 − 90.

[7] Hirotsu C. (1983b). *Defining the pattern of association in two-way contingency tables.* Biometrika **70**, 579 − 590.

[8] Hirotsu C. (1986). *Cumulative chi-squared statistic as a tool for testing goodness of fit.* Biometrika **73**, 165 – 173.

[9] Hirotsu C. (1991). *An approach to comparing treatments based on repeated measures.* Biometrika **78**, 583 – 594.

[10] Hirotsu C. (1993). *Beyond analysis of variance techniques : Some applications in clinical trials.* Int. Statist. Rev. **61**, 183 – 201.

[11] Hirotsu C., Aoki, S., Inada, T. and Kitao, Y. (2001). *An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis.* Biometrics **57**, 769 – 778.

[12] Hirotsu C. and Marumo, K. (2002). *Change point analysis as a method for isotonic inference.* Scand. J. Statist. **29**, 125 – 138.

[13] Plackett, R. L. (1981). *The analysis of categorical data* 2^{nd} *ED.* Charles Griffin & Company Ltd., London.

Address: C. Hirotsu, E. Ohta, S. Aoki, Meisei Univ., Hino-shi,
Tokyo 191-8506, Japan

E-mail: hirotsu@ge.meisei-u.ac.jp

GROWTH CURVE APPROACH TO PROFILES OF ATMOSPHERIC RADIATION

Daniel Hlubinka

Key words: Growth curves, nonlinear regression, initial estimator, kernel smoothing.

COMPSTAT 2004 section: Applications.

Abstract: In the Czech Hydrometeorological Institute there exists an unique set of measurements consisting of the values of vertical atmospheric levels of beta and gamma radiation. Since theoretical model explaining the physical background of this process is not known we propose a statistical model for the phenomenon. As the data are naturally nonnegative we decided to base our statistical model on the cumulative sums of the data. These sums result in a growing sequence and hence the growth curve approach is justified. In this paper we discuss the choice of the model and, as we need to estimate many parameters, the problem of initial estimator of the parameters.

1 Profiles of atmospheric radiation

Since 1994 in the Czech Hydrometeorological Institute (CHMI) the balloons are used to lift once monthly a Geiger – Müller counter in order to get a measurement of gamma and beta radiation in the atmosphere. The balloon is released usually at 12:00 CET and can reach approximately the altitude 30 – 35 kilometers. The measured data contains information about the time since the release, beta and gamma counts since the last impuls and some meteorological observation related to the actual altitude of the balloon like temperature or pressure. This raw data are later recalculated for further analysis.

The final data obtained from each lift of the balloon contain the recalculated values of variables that were measured during the intervals of the length of ten seconds. Thus the information about the altitude, temperature, pressure, dew point, humidity, time and average number of beta and gamma counts per second in consecutive (ten second long) intervals is available. All these measurement and calculations are done in the CHMI.

As the balloon does not climb with a constant speed, namely the speed can differ for each release, we prefer to consider the altitute to be the explanatory variable instead of the time since the release. Our main goal is to find a relatively simple but well fitted parametric curve explaining the data using the altitude as the regressor. As can be seen from the figures the proper model must be highly nonlinear.

When speaking about a *measurement* we always mean one release of the balloon. This measurement consists of many *observations* in the 10s intervals. Typical measurements of the intensities of beta and gamma counts with respect to the altitude are shown in Figure 1. In the sequel following notation is used:

- $Y_i, i = 1, \dots, n$ denote values of the response variable, i.e. the per second average of the beta (gamma) counts in an i-th time interval.
- $t_i, i = 1, \dots, n$ denote time moments in which the measurements were obtained. We put $t_i = i, i = 0, 1, 2, \dots$ and interpret it as the time moment $10i$ seconds (since the release of the balloon).
- $X_i, i = 1, \dots, n$ denote values of the altitude in kilometers in time moments t_i.
- n denotes number of observations for a given measurement.

Using this notation, the data can be described by a model

$$Y_i = m\big(X_i\big) + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where $m(\cdot)$ denotes an unknown regression function describing the mean amount of beta (gamma) counts, and the "error term" ε_i, $i = 1, \dots, n$ is composed of two parts, i.e.

a. the errors of measurements;
b. natural variability of the measured quantity.

We are looking for an appropriate parametric function m.

2 Model selection

Although the "bell-shape" of the original data may suggest quite smooth model one typically has no idea of the form for the fitted function. We tried to find better starting point for our analysis. As the data are all nonnegative one can get the idea to look at the cumulative sum of the data. Obtaining $V_k = \sum_{i=1}^{k} Y_i$ we can see that V_k is an increasing sequence and it is quite natural to consider model

$$V_k = g(X_k) + \eta_k,$$

where g is some growth curve (see Seber and Wild, 1988 or Ratkowski, 1983), and η_k is an error term.

2.1 Richars curve

After some analysis we have decided to use so called Richards growth curve defined by

$$R(x; a, b, c, d) = a\left(1 + (b-1)\mathrm{e}^{-c(x-d)}\right)^{1/(1-b)}. \tag{2}$$

This four parameter curve is enough flexible to fit the data $(V_k, k = 1, \dots, n)$.

As the Richards curve is used for the cumulative data, for the original data it seems to be natural to use derivative of the curve. We have mentioned already that the altitudes X_i are not equidistant. It may cause problems when we just transfer the model from (V_i) to (Y_i) using its derivative. On the other hand although the distance $X_{i+1} - X_i$ is not constant, it ranges from 40 metres to 60 metres and does not change abruptly. Hence we may use the derivative as the model, however, the parameters of the fitted curve must be re-estimated.

Let us recall that the logistic, Gompertz or monomolecular models are submodels of the Richards curve.

2.2 Derivative of Richards curve

The derivative of the Richards curve $r(\cdot)$ is of the form

$$r(x; a, b, c, d) = ca \left(1 + (b-1)\mathrm{e}^{-c(x-d)}\right)^{b/(1-b)} \mathrm{e}^{-c(x-d)}, \tag{3}$$

hence we assume the model in the form

$$Y_i = ca \left(1 + (b-1)\mathrm{e}^{-c(X_i-d)}\right)^{b/(1-b)} + \varepsilon_i. \tag{4}$$

Our goal is to estimate the parameters a, b, c, d using the least sum of squares method

$$\min_{a,b,c,d} \sum_{i=1}^{n} \left[Y_i - ca \left(1 + (b-1)\mathrm{e}^{-c(X_i-d)}\right)^{b/(1-b)} \right]^2 .$$

One can see that this is quite nontrivial minimization and must be performed numerically. It is desirable to look for an initial estimator a_0, b_0, c_0, d_0 used for the numerical optimization. This estimator should be based on the data therefore we need to understand well the role of the parameters in the model.

After examination of the function r we may get the following role of parameters in the Richards model:

a. Parameter a is the area below (above) the function r provided $a > 0$ $(a < 0)$. It means that the original Richards curve attains values from the interval $[0, a]$.

b. Parameter b is a "shape" parameter. If $b \geq 1$, then r is well defined for all x. If $b < 1$, then it is defined for $x \geq d + \log(1-b)/c$ if $c > 0$ and for $x \leq d - \log(1-b)/c$ if $c < 0$. Moreover $R(d; a, b, c, d) = ab^{1/(1-b)}$ and hence the area below the function r up to the value d is a given proportion to the total integral of r.

c. Assume $a > 0$. If $c > 0$, then r is nonnegative while $c < 0$ implies that it is nonpositive. If $a < 0$, we can see the same behavior provided the inequalities for c are reversed. Finally, for $c = 0$ we get a constant function. Note that $r(d; a, b, c, d) = cab^{b/(1-b)}$ and hence $r(d)/R(d) = cb^{-1}$.

d. Parameter d informs us about the position of the global extreme of r, or in the language of R it is its point of inflection (provided it exists).

3 Initial estimators

3.1 Basic model

Having the function r and its parameters examined we can follow with the initial estimators based on the data. The following values can serve as the reasonably good starting point for the estimation of the parameters of r as follows form the pbservations a.−d. of the previous section.

$$\begin{aligned} d_0 &= \Big\{X_i : Y_i = \max_j Y_j\Big\} & c_0 &= \frac{Y(X_{d_0})}{a_0 b_0^{b_0/(1-b_0)}} \\ b_0 &: \textstyle\sum_{\{i:x_i\le d_0\}} Y_i\big(X_i - X_{i-1}\big) = a_0 b_0^{1/(1-b_0)} & a_0 &= \sum_{i=1}^{n} Y_i\big(X_i - X_{i-1}\big) \end{aligned} \tag{5}$$

However, there still exists a serious problem with the initial values of parameters due to the fact that our data does not decrease back to zero. Hence a_0 is strongly underestimated. We can estimate the missing area using simple linear regression fitted to the descending part of the data.

Moreover, there is another important point we have to keep in mind. It is the fact that standard Richards growth curve originates from zero. However, due to the existing small ground radioactivity measured on the earth this is not the case for our measurements. Therefore, to have a good fit close to the earth, it is recommendable to add a constant or a linear function to our model r. It means that we add a new pair of parameters, the course of the surface radioactivity.

3.2 Generalized model

The proposed nonlinear regression function now becomes

$$r^*(x;a,b,c,d,e,f) = \begin{cases} e + fx + r(x;a,b,c,d) & \text{if } x \ge d + \log(1-b)/c \\ e + fx & \text{if } x \le d + \log(1-b)/c. \end{cases} \tag{6}$$

if $b < 1$ or simply $r^*(x;a,b,c,d,e,f) = e + fx + r(x;a,b,c,d)$ if $b > 1$.

We must find an appropriate initial estimator again for the generalized model. One can notice quite high variability of the original data. In order to decrease its impact to the initial estimators we propose to use rather smoothed data. We replace the original data Y_i by the smoothed data Z_i calculated in the same altitudes using local polynomial regression (quadratic).

Let us suppose that

$$Z_i = \alpha_0 + \alpha_1 X_i$$

is the estimated regression lines for the low altitude. Obviously the initial estimators of e, f are

$$e_0 = \alpha_0; \ f_0 = \alpha_1.$$

Now we can find the position of the maximum of the smoothed data having on mind the regression line $e_0 + f_0x$.

Having the maximum, we will use all observations above d_0 to estimate the trend line

$$Z_i = \beta_0 + \beta_1 X_i$$

of the descending part of the data. The area between the two regression lines, for the high ald low altitudes will serve as an estimator of the unknown radioactivity.

The initial estimators b_0 and c_0 of the two shape factors are estimated in a similar way as before.

Let us summarize the initial estimators in order of calculation.

$$\begin{aligned}
e_0 &= \alpha_0 \\
f_0 &= \alpha_1 \\
m &= \min\{i : Z_i - e_0 - f_0 X_i = \max_{1\le k\le n}\{Z_k - e_0 - f_0 X_k\}\} \\
d_0 &= X_m \\
a_{01} &= \sum_{i=2}^{n} Z_i (X_i - X_{i-1}) \\
a_{02} &= (X_n - X_1)[e_0 + f_0(X_n - X_1)/2] \\
a_{03} &= \frac{1}{2}[(e_0 - \beta_0)/(\beta_1 - f_0) - X_n][\beta_0 - e_0 + (\beta_1 - f_0)X_n] \\
a_0 &= a_{01} - a_{02} + a_{03} \\
a_0 b_0^{1/(1-b_0)} &= \sum_{i=2}^{m} Z_i (X_i - X_{i-1}) - (X_m - X_1)[e_0 + f_0(X_m - X_1)/2] \\
&\Downarrow \\
&b_0 \\
c_0 &= \frac{Z_m - e_0 - f_0 X_m}{a_0 b_0^{b_0/(1-b_0)}}
\end{aligned}$$

The two auxiliary regressions together with the final estimator are illustrated on Figure 2.

4 Results

We illustrate the method on the data set consisting of approximately three years of measurement. All the data were cleaned such that obvious outliers were removed.

4.1 Regression line

First of all we have used local quadratic regression estimator to get smoothed estimation $Z(x)$ of the data $Y(x)$ at the altitude x. Based on the data Z we have calculated the initial estimator $(a_0, b_0, c_0, d_0, e_0, f_0)$. Finally, using MATLAB, we get the approximate least sum of squares estimator $(\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d}, \widehat{e}, \widehat{f})$ of the six parameters of our generalized model

$$\mathrm{E}\, Y(x) = r^*(x; a, b, c, d, e, f).$$

The performance of the procedure is illustrated on Figures 2 and 3.

4.2 Quantile regression

We were asked by the CHMI to provide confidence bounds for the radiation in given altitude. Therefore we have also calculated the nonlinear quantile regression. As we have calculated this estimators for several quantile levels it is possible to use as an initial estimator the set of parameters estimated for the closest quantile. It means that for quartile regression we have used $(\widehat{a}, \widehat{b}, \widehat{c}, \widehat{d}, \widehat{e}, \widehat{f})$ as the initial estimator. The parameters estimated for the 0.25 and 0.75 regression were used as an initial estimator for the 0.1 and 0.9 regression and so on. The results are illustrated on Figure 3.

5 Figures

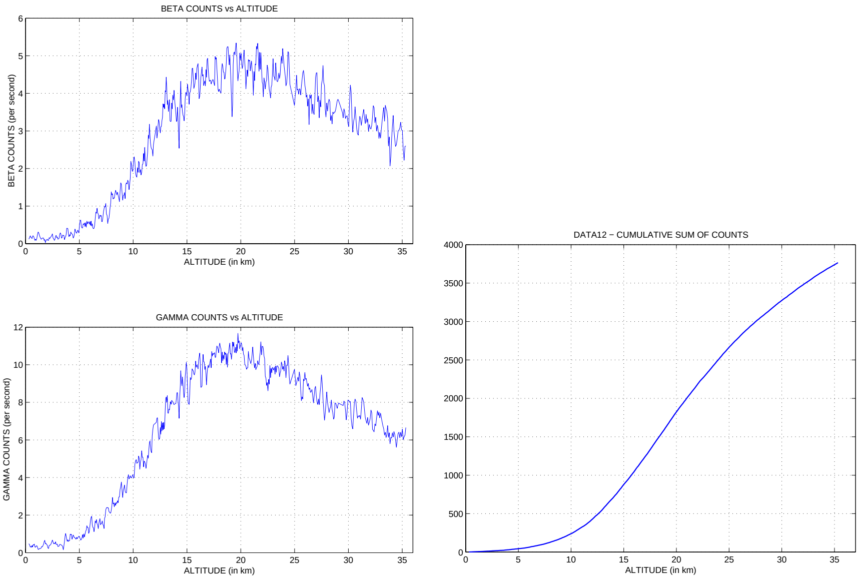


Figure 1: Typical measurement and its cumulative sums.

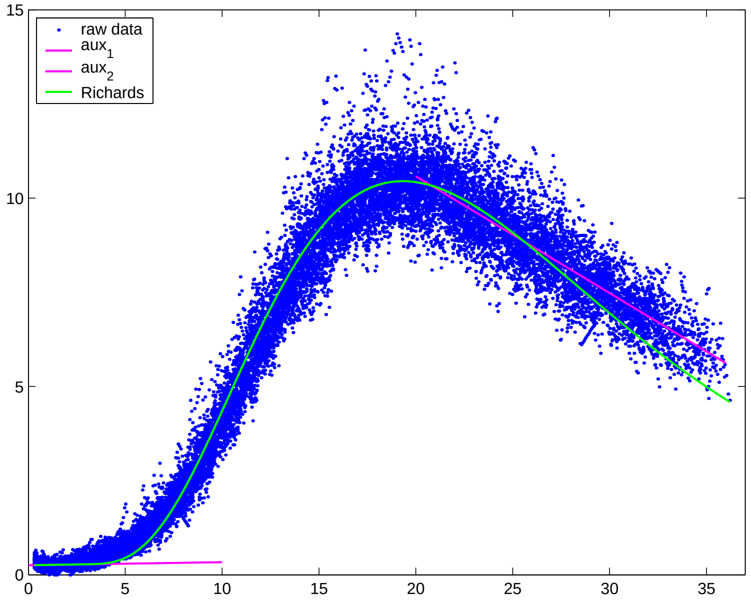


Figure 2: Fitted derivative of Richard curve and two auxiliary regressions.

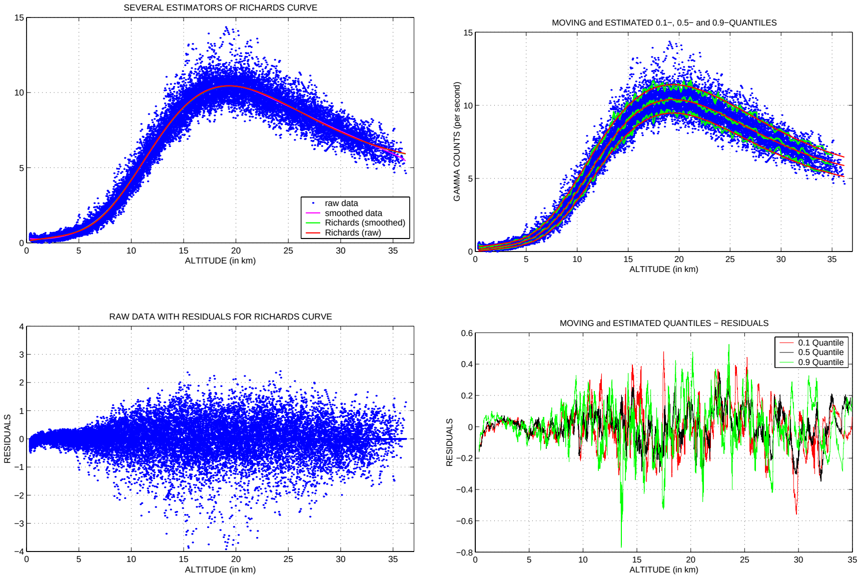


Figure 3: Fitted regression line and 0.1 and 0.9 quantile regression with residuals.

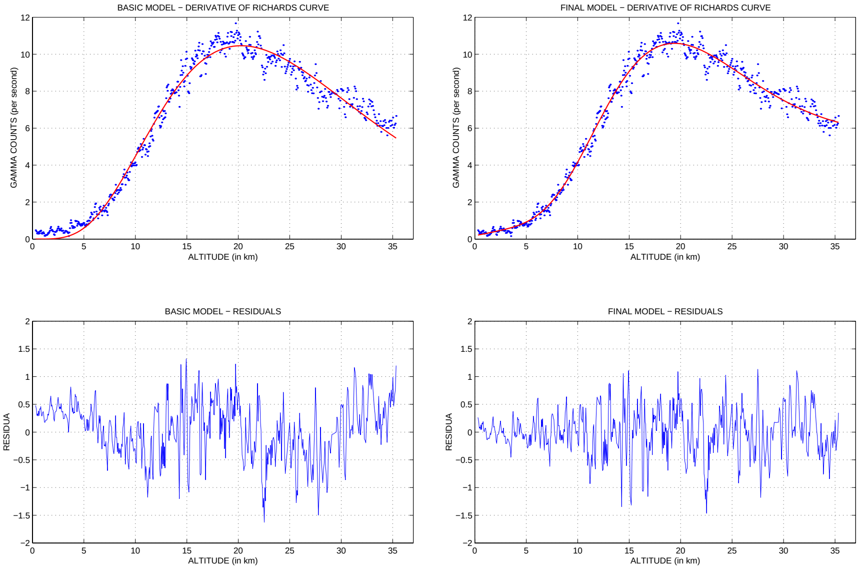


Figure 4: Comparison of the basic (left) and the generalized (right) model.

References

[1] Seber G.A.F., Wild C.J. (1989). *Nonlinear regression.* J. Wiley, Chichester.

[2] Schimek M.G. (2000). *Smoothing and regression.* J. Wiley, Chichester.

Acknowledgement: The analysis is supported by the postdoc project GAČR 210/03/P138 and by the project MSM 113200008 “Mathematical Methods in Stochastics” of the Czech Ministry of Education. Author would like to express his thanks to Prof. Antoch for valuable comments and help.

Address: D. Hlubinka, Univerzita Karlova, Matematicko-fyzikální fakulta,
Katedra pravděpodobnosti a statistiky, Sokolovská 83, 186 75 Praha 8,
Czechia

E-mail: `daniel.hlubinka@mff.cuni.cz`

COMPSTAT'2004 Symposium

CALIBRATED INTERPOLATED CONFIDENCE INTERVALS FOR POPULATION QUANTILES

Yvonne H.S. Ho

Key words: Bandwidth, calibration, interpolation, quantile, smoothed bootstrap.

COMPSTAT 2004 section: Nonparametrical statistics.

Abstract: Beran and Hall's [1] simple linear interpolation provides a very convenient approach to constructing nonparametric confidence intervals for population quantiles based on a random sample of size n. We show that the coverage error of the interpolated interval, which is of order $O(n^{-1})$, can be improved upon by calibrating the nominal coverage level. Three distinct methods of calibration are considered. The analytic and Monte Carlo methods succeed in reducing the order of coverage error to $O(n^{-3/2})$, while the smoothed bootstrap method reduces it further to $O(n^{-25/14})$. We provide guidelines for practical implementation of the calibration methods. Their performance is compared with the simple linear interpolated interval in a simulation study, which confirms superiority of the calibrated intervals.

1 Introduction

A nonparametric confidence interval for a population quantile is most naturally derived from the sample quantile. Due to binomial discreteness of its end points, the interval has precisely known one-sided coverage probability which cannot be rendered closer than $O(n^{-1/2})$ from the nominal coverage level in general [2]. The smoothed bootstrap and studentization of sample quantiles have the potential to reduce the coverage error: see Falk and Janas [3] and Janas [6]. Ho and Lee [4] extend and sharpen the above works by showing that coverage-calibrating the smoothed bootstrap percentile method reduces coverage error to $O(n^{-2/3})$ and $O(n^{-58/57})$ respectively. Coverage errors of bootstrap intervals can in priniciple be reduced progressively by iterating the bootstrap procedure further. However, such iterations require expensive computational input.

In the same context Beran and Hall [1] contemplate a somewhat different approach to correcting sample-quantile-based confidence intervals by means of simple linear interpolation of order statistics. Their intervals have coverage error of order $O(n^{-1})$ under any smooth distributions, which cannot be reduced further by higher-order interpolations. Compared with the bootstrap methods, the interpolated interval is computationally much more convenient to construct and is asymptotically more accurate than most bootstrap intervals.

We investigate the asymptotic effects on coverage error of calibrating the nominal coverage level of the simple linear interpolated confidence interval. Section 2 reviews Beran and Hall's simple linear interpolated intervals. Section 3 outlines the proposed methods and establishes the asymptotic results. Section 4 presents the simulation results. Section 5 concludes our findings.

2 Interpolated confidence interval

Let $\mathcal{X}$ be a random sample of size n from a univariate distribution function F with density f. For a fixed $q \in (0,1)$, we focus on the problem of constructing a nominal level α upper confidence interval for $F^{-1}(q)$, where $0 < \alpha < 1$.

Let B be a binomial (n,q) random variable and $B_r = \mathbb{P}(B \le r)$. Choose $r \in \{0,1,\ldots,n\}$ such that $B_{r-1} < \alpha \le B_r$. Put $\pi_\alpha = (\alpha - B_{r-1})/(B_r - B_{r-1})$ and $Q(\alpha) = (1-\pi_\alpha)X_{[r]} + \pi_\alpha X_{[r+1]}$, where $X_{[.]}$ is the order statistic. Beran and Hall's [1] simple linear interpolated upper confidence interval, of nominal level α, is defined to be $I_{BH}(\alpha) = (-\infty, Q(\alpha))$.

Assume that f is continuously differentiable in a neighbourhood of $F^{-1}(q)$ and that $f(F^{-1}(q)) > 0$. We prove that

$$\begin{aligned}&\mathbb{P}(F^{-1}(q) \in I_{BH}(\alpha))\\ &\quad = \quad \alpha + n^{-1}p_1(\alpha;q) + n^{-3/2}p_2\left(\alpha;q,f(F^{-1}(q)),f'(F^{-1}(q))\right) + O(n^{-2}),\end{aligned}$$

where the $p_j(\alpha;\cdots)$ are Lipschitz continuous in α and are smooth with respect to the arguments following α.

3 Nominal level calibration

In what follows, let β be the adjusted level in each of the methods.

3.1 Analytic method

Our first method, which we term the analytic method, simply calibrates α to $\beta = \alpha - n^{-1}p_1(\alpha;q)$, where $p_1(\alpha;q) = [q(1-q)]^{-1}\,\pi_\alpha(1-\pi_\alpha)u_\alpha\phi(u_\alpha)$, ϕ denotes the standard normal density and u_α its αth quantile. We then define the resulting calibrated interval to be $I_{AN}(\alpha)$.

3.2 Monte Carlo method

Ideally the exact adjusted nominal level β is the solution to the equation $\mathbb{P}(F^{-1}(q) \in I_{BH}(\beta)) = \alpha$. We exploit this relation to provide an approximation to the exact β by solving instead the equation $\mathbb{P}(G^{-1}(q) \in J_{BH}(\beta)) = \alpha$, where G is a completely specified reference distribution function and $J_{BH}(\beta)$ denotes the corresponding level β interpolated confidence interval for $G^{-1}(q)$. The solution for β is obtained by a Monte Carlo procedure as follows.

Generate M random samples, each of size n, from G. Denote by $Y^i_{[1]} \le \cdots \le Y^i_{[n]}$ the order statistics of the ith sample, $i = 1,\ldots,M$. For each i, choose $\hat{r}$ such that $Y^i_{[\hat{r}]} < G^{-1}(q) \le Y^i_{[\hat{r}+1]}$. Put $\hat{\pi} = (G^{-1}(q) - Y^i_{[\hat{r}]})/(Y^i_{[\hat{r}+1]} - Y^i_{[\hat{r}]})$ and $\beta^i = (1-\hat{\pi})B_{\hat{r}-1} + \hat{\pi}B_{\hat{r}}$. We then order the β^i as $\beta^{[1]} \le \cdots \le \beta^{[M]}$ and set $\beta = \beta^{[\ell]}$, where ℓ is the integer part of $M\alpha$. Define the calibrated interval to be $I_{MC}(\alpha) = I_{BH}(\beta)$. In the ideal case where $G = F$ and $M = \infty$, $I_{MC}(\alpha)$ has exact coverage probability.

3.3 Smoothed bootstrap method

Let F_n be the empirical distribution of $\mathcal{X}$ and $\hat{F}_{n,h}$ be its smoothed version with density $\hat{f}_{n,h} = \hat{F}'_{n,h}$, where $\hat{f}_{n,h}(t) = \frac{1}{nh}\sum_{i=1}^n k\left(\frac{t-X_i}{h}\right), t \in \mathbb{R}$, $h > 0$ is the smoothing bandwidth and k denotes the kernel function. The smoothed bootstrap calibration method resembles the Monte Carlo method with $G = \hat{F}_{n,h}$. We denote the resulting calibrated interval by $I_{BT}(\alpha)$.

3.4 Theory and remarks

We establish the following theroem in Ho and Lee [5].

Theorem 3.1. *Assume that F and G are two times continuously differentiable in some neighbourhoods of $F^{-1}(q)$ and $G^{-1}(q)$ respectively, and that $F'(F^{-1}(q)), G'(G^{-1}(q)) > 0$. Assume also that k is non-negative, symmetric about zero and $h \propto n^{-\Delta}$ for some $\Delta \in (0,1)$. Then we have, for $\alpha \in (0,1)$, that*

$$\begin{aligned}
\mathbb{P}(F^{-1}(q) \in I_{AN}(\alpha)) &= \alpha + O(n^{-3/2}), \\
\mathbb{P}(F^{-1}(q) \in I_{MC}(\alpha)) &= \alpha + O(n^{-3/2}), \qquad \textit{and} \\
\mathbb{P}(F^{-1}(q) \in I_{BT}(\alpha)) &= \alpha + O(n^{-3/2}h^2 + n^{-2}h^{-3/2}).
\end{aligned}$$

We see from Theorem 3.1 that both $I_{AN}(\alpha)$ and $I_{MC}(\alpha)$ outperform the uncalibrated $I_{BH}(\alpha)$ with a reduction in coverage error from $O(n^{-1})$ to $O(n^{-3/2})$. If we choose $h \propto n^{-1/7}$, the interval $I_{BT}(\alpha)$ is the most accurate with coverage error of order $O(n^{-25/14})$.

Note that the adjustment made to α using either the analytic or Monte Carlo method is deterministic in the sense that it is unaffected by the observed data. Simplicity of their operations makes these two methods computationally very attractive. In practice, the calibrated levels β can be tabulated in advance for different nominal levels α and for each combination of n and q. These tabulated values can be referred to for calibration regardless of the observed data in hand. We remark that for the Monte Carlo method, the calibration also depends on our choice of the reference distribution G.

4 Simulation study

To illustrate the Monte Carlo method we select three different distributions for G, N$(0,1)$, exp(1) and U$[0,1]$, take M = 1,000,000 and compute the adjustment $\beta - \alpha$ for $n = 20$ and for different α and q. Figure 1 shows the adjustments $\beta - \alpha$ against α along with those computed by the analytic method. Calibration using either method exhibits more or less similar oscillating patterns and is insensitive to the choice of G for the Monte Carlo method.

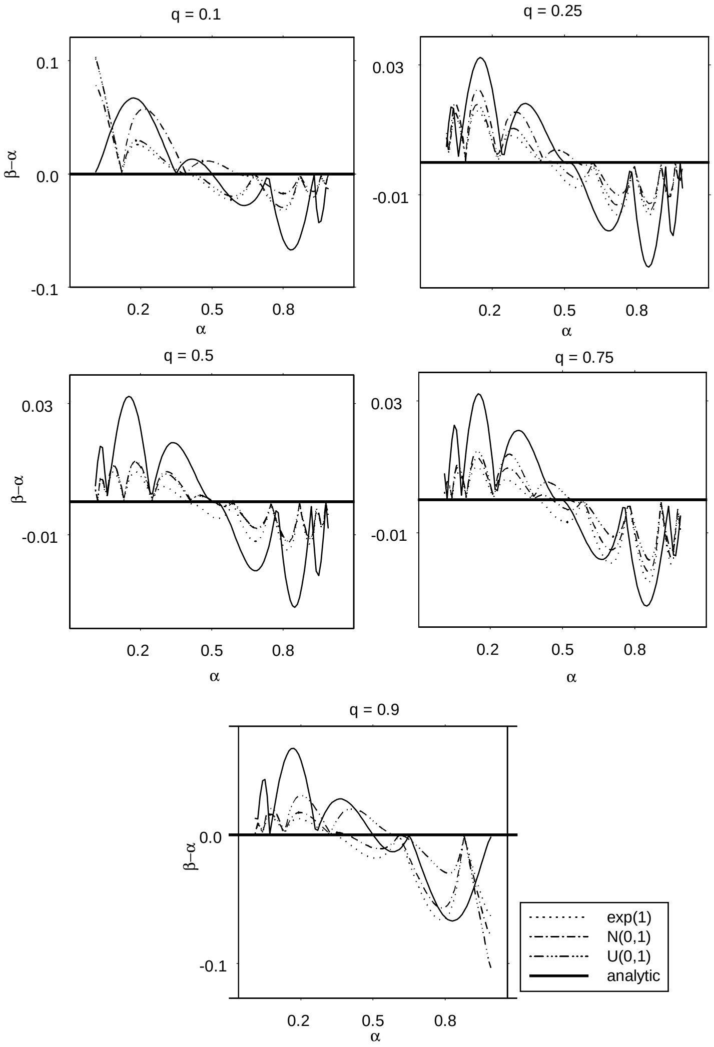


Figure 1: Additive adjustments $\beta - \alpha$ based on analytic and Monte Carlo methods, with G taken to be $\exp(1)$, $N(0, 1)$ and $U[0, 1]$ for the latter method, for $n = 10$ and $q = 0.1, 0.25, 0.5, 0.75, 0.9$.

A simulation study was also conducted for the same values of sample size and q and for $\alpha = 0.90, 0.91, \ldots, 0.99$. Exp(1) was considered as the underlying distribution F. Construction of $I_{BT}(\alpha)$ was based on $G = N(0,1)$. For both $I_{MC}(\alpha)$ and $I_{BT}(\alpha)$, we fixed M to be $1,000,000$. All coverage probabilities were estimated by averaging over 5,000 random samples drawn from F.

We see from Figure 2 that the absolute coverage errors of our methods are smaller than that of $I_{BH}(\alpha)$ by a noticeable margin and their magnitudes are often below 0.01. For each q, $I_{MC}(\alpha)$ and $I_{BT}(\alpha)$ have more stable performance over a wide range of α, while the coverage error of $I_{BH}(\alpha)$ typically succumbs to severe fluctuations. In general, although $I_{AN}(\alpha)$ is not as accurate as $I_{MC}(\alpha)$ or $I_{BT}(\alpha)$, its coverage error is remarkably smaller than that of $I_{BH}(\alpha)$. We also compare the means and the standard deviations of the intervals and find that the calibrated intervals produce shorter intervals and more stable endpoints. Table 1 gives the figures for $\alpha = 0.90$.

q	0.25	0.50	0.75
$\boldsymbol{I_{BH}}$			
mean	0.2052	0.3674	0.6676
var	0.0305	0.0863	0.3092
$\boldsymbol{I_{AN}}$			
mean	0.2028	0.3463	0.6588
var	0.0305	0.0842	0.3043
$\boldsymbol{I_{MC}}$			
mean	0.2041	0.3455	0.6454
var	0.0305	0.0842	0.2971
$\boldsymbol{I_{BT}}$			
mean	0.2046	0.3480	0.6420
var	0.0306	0.0846	0.2932

Table 1: Estimated means and standard deviations of 90% confidence intervals, I_{BH}, I_{AN}, I_{MC} and I_{BT}.

5 Conclusion

We have proposed three simple and computationally very attractive methods for calibrating Beran and Hall's [1] interpolated confidence intervals for pop-

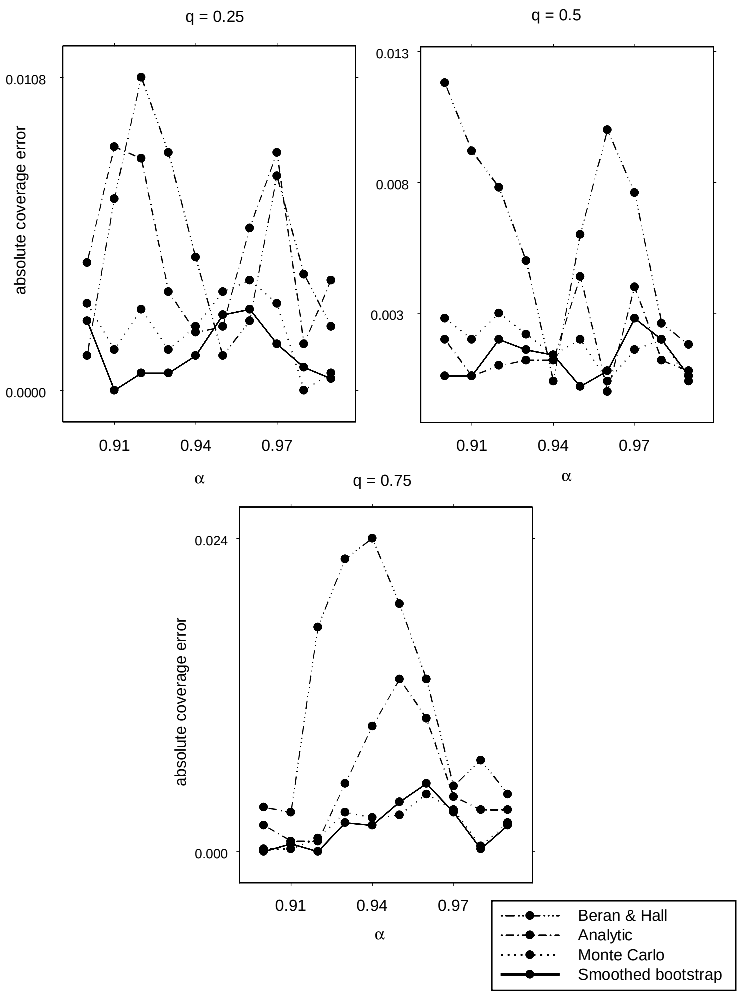


Figure 2: Absolute coverage errors of upper calibrated interpolated intervals for $F^{-1}(q)$, for nominal levels $0.9, 0.91, \ldots, 0.99$, under $exp(1)$.

ulation quantiles. The calibrated intervals are asymptotically more accurate than any existing intervals.

Of the three methods proposed, the analytic and Monte Carlo methods call for straightforward and deterministic coverage calibration, which is universal regardless of the data in hand. For each n and q, the calibrated levels can be pre-determined and tabulated in advance, to which practitioners can routinely refer for practical calibration. On the other hand, the smoothed bootstrap calibrates the nominal level adaptively and trades computational simplicity for increased coverage accuracy. The important issue of bandwidth selection can be dealt with in the same way as what is generally recommended for density and density derivative estimation problems. It seems from our simulation study that the ad hoc determination of h by the normal referencing rule gives quite satisfactory results.

References

[1] Beran R., Hall P. (1993). *Interpolated nonparametric prediction intervals and confidence intervals.* J. Roy. Statist. Soc. Ser. B **55**, 643 – 652.

[2] De Angelis D., Hall P., Young G. A. (1993). *A note on coverage error of bootstrap confidence intervals for quantiles.* Math. Proc. Camb. Phil. Soc. **114**, 517 – 531.

[3] Falk M., Janas D. (1992). *Edgeworth expansions for studentized and prepivoted sample quantile with applications to confidence intervals.* Statist. Probab. Lett. **14**, 13 – 24.

[4] Ho Y.H.S., Lee S.M.S. (2003a). *Iterated smoothed confidence intervals for population quantiles.* Research Report No. 352. Department of Statistics and Actuarial Science, The University of Hong Kong.

[5] Ho Y.H.S., Lee S.M.S. (2003b). *Calibrated interpolated confidence intervals for population quantiles.* Research Report No. 365. Department of Statistics and Actuarial Science, The University of Hong Kong.

[6] Janas D. (1993). *A smoothed bootstrap estimator for a studentized sample quantile.* Ann. Inst. Statist. Math. **45**, 317 – 329.

Acknowledgement: Thanks to my family for their generous support.

Address: Department of Statistics and Actuarial Science, The University of Hong Kong. Address: Rm 515, Meng Wah Complex, The Universiy of Hong, Pokfulam Road, Hong Kong.

E-mail: hohs@graduate.hku.hk

BAGGING SURVIVAL TREES FOR PROGNOSIS BASED ON GENE PROFILES

Thu M. Hoàng and Van L. Parsons

Key words: Survival forests; gene profiles; T-ALL; prognosis.

COMPSTAT 2004 section: Tree based method.

Abstract: Combinations of survival regression trees called survival forests (SF) have been proposed by Breiman to estimate survival functions and assess variable importance. We examine some operating characteristics of the method, add new options and apply SF to gene expression profiles data for predicting survival.

1 Background

In a broad view of the future of genomics research, Collins et al [6] identify three series of grand challenges in the spirit of the 23 problems Hilbert laid out at the turn of the twentieth century, and advocate the development of methods that catalyze the translation of genomic information into therapeutic advances. Indeed gene expression profiling is increasingly proposed as support to clinical decision with the hope of highly individualized and more effective prognosis, e.g. Alizadeh et al. [1] for diffuse large B-cell lymphoma and Ferrando et al. [7] for T-cell acute lymphoblastic leukemia (T-ALL). An issue of interest is to identify a group of *coregulated* genes related to survival. The usual approach proceeds with initial molecular classification of the disease by clustering then standard survival analysis of the disease classes thus identified.

The validity of such procedure of prognosis rests on the reliability of the clustering which in turn depends on the chosen algorithm. In section 2, using the T-ALL data of Ferrando et al. we show that different algorithms yield different sets of clusters. An alternative strategy is then to directly analyze survival using gene expression profiles as features. Survival time is a continuous univariate response endowed with the order on the real line but subject to censoring when the event (death or recurrence) has not yet been observed. Using ensemble estimation by survival forest (SF) and isotonized prediction we estimate ordered level sets of survival and *concurrently* identify genes that are determinant in defining these levels of survival. This method of mining microarrays data for prognosis is applied to T-ALL data .

[0] Research results and conclusions expressed in this paper are those of the authors and do not necessarily indicate concurrence by the National Center for Health Statistics.

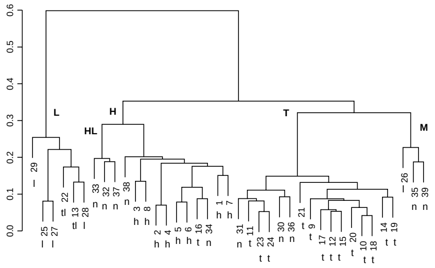


Figure 1: Average linkage hierarchical cluster of 39 T-ALL samples.

2 Revisiting a prognosis analysis of T-ALL data based on gene profiles

The data consisting of 39 T-ALL samples were analyzed with both DNA microarray (Affymetrix HU6800 with 7129 probe sets) for the global patterns of gene expression and RT-PCR (reverse transcriptase polymerase chain reaction) for expression of single genes. RT-PCR detected 27 samples with aberrant expression of one of the three oncogenes HOX11, LYL1 or TAL1, i.e., the "pure" cases identified as h, l or t cases, 2 expressing both LYL1 and TAL1, i.e., the mixed cases identified as tl cases, and 10 without detectable expression of these oncogenes identified as nc cases.

By permutation tests Ferrando et al. obtained 72 genes whose expression patterns best distinguished among h, t, l, and nc cases. Then using these genes they grouped the samples and identified 3 main tumor classes labelled H, L and T and 2 tumor subclasses, M and HL. With the average linkage agglomerative clustering algorithm and the 1-ρ distance, where ρ is the Pearson correlation, we obtained a tree similar to Ferrando et al.'s (Figure 1). Ferrando et al. claimed the clinical importance of the finding by showing significant difference in survival between the classes. Now with single linkage clustering the class H and the two subclasses M and HL disappear (Figure 2). Furthermore there is no evidence of hierarchical relationships in biological functions of the genes that have been modelled by the hierarchical clustering.

If the question of interest is the eventual influence of gene signature on treatment outcome then an alternative is to analyze survival as the response while considering gene expression profiles as covariates/features.

3 Survival forests

Breiman [5] suggested combinations of survival regression trees called survival forests (SF) for estimating survival functions and assessing variable importance. A survival tree is constructed using a bootstrap sample, and an ensemble of trees is created over many such samples. In view of prognosis,

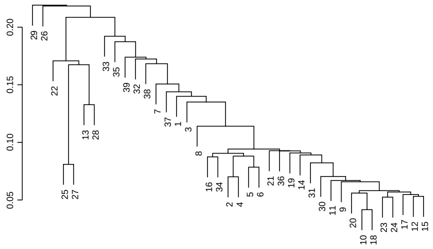


Figure 2: Single linkage hierarchical cluster of 39 T-ALL samples.

a new case can be run through the ensemble of trees to obtain a survival function, which in turn is used to get survival prediction. Given a survival time, $T(\mathbf{x})$, depending on a covariate vector $\mathbf{x}$, for a time t define the hazard, $h(t,\mathbf{x}) = P(T(\mathbf{x}) \in (t, t+dt)|T(\mathbf{x}) > t)/dt$, the cumulative hazard, $H(t,\mathbf{x}) = \int_0^t h(\tau,\mathbf{x})d\tau$ and the survival,

$$S(t,\mathbf{x}) = P(T(\mathbf{x}) \geq t) = \exp\left(-H(t,\mathbf{x})\right)$$

An experiment consists of N independent survival times subject to random independent right censoring. Denote the data by $(\mathbf{x}_i, t(\mathbf{x}_i), c_i), i = 1, 2, ..., N$ where $c_i = 1$ (0) if the survival is observed (censored). To grow a survival tree consider maximizing the loglikelihood $\ell = \sum c_i \log h(t_i, \mathbf{x}_i) - \sum \int_0^{t_i} h(\tau, \mathbf{x}_i)d\tau$ for fixed covariate effects and partitioning the product time-covariate space into L rectangles $r_l = I_l \otimes R_l$ where I_l is a time interval and R_l a rectangle in the covariate space, for $l = 1, 2, \ldots L$. These rectangles are the nodes of the tree. The hazard function is assumed to be stepwise constant $h(t,\mathbf{x}) = \exp\left(\sum_l 1_{r_l}(t,\mathbf{x})\alpha_l\right)$ where $1_{r_l}(t,\mathbf{x}) = 1$ if $(t,\mathbf{x}) \in r_l$, and 0 otherwise. Maximizing over the α_l yields

$$\alpha_l = \log \frac{D_l}{T_l} \quad \text{and} \quad \ell = \sum_l D_l \log \frac{D_l}{T_l} - \sum_l D_l$$

where $D_l = \sum_i c_i 1_{r_l}(t,\mathbf{x}_i)$ is the number of events, e.g., deaths, in the node l, and $T_l = \sum_i 1_{R_l}(\mathbf{x}_i) \cdot \text{length}[(0,t_i) \cap I_l]$.

The tree is grown so that each node l is split which most increases $\sum_l D_l \log \frac{D_l}{T_l}$. However, estimates based upon one tree may be unstable (e.g., see [10]). Recent experience in combining or integrating an ensemble of models computed by the same or different algorithms over multiple data subsets are shown to improve both efficiency and scalability by executing the estimation processes in parallel [3][11]. For classification and regression Breiman introduced random forests [4] that are committees of tree predictors in which randomization is injected to keep correlation low, particularly randomized selection of predictors at each internal node of each tree. For survival regression, Breiman proposed to use a bootstrap sample to create a survival tree and then average over many such trees, hence the name survival forest, to reduce variability and improve accuracy. However in SF there is no random selection of features.

A bootstrap sample of size N from the original N data points is selected and a survival tree grown until each terminal node has exactly one uncensored event. For each out-of-bag observation (*oob*), i.e., a case not used in the bootstrap sample, the covariate $\mathbf{x}_{\text{oob}}$ is put through this single tree to get an estimate of $S(t, \mathbf{x}_{\text{oob}})$. A bootstrap sample serves as a training set, and training over many bootstrap samples obtains the corresponding trees. The averaged survival function estimated for all the *oob* cases is de facto a test set estimate. If an independent test set is also available, for each new observation the covariate $\mathbf{x}_{new}$ can be put through all the trees to get an estimate of the individual survival curve $\hat{S}(t, \mathbf{x}_{new})$.

4 Tuning SF parameters. Assessing fit and prediction

While the concept of survival forests seems promising, the experience in applications is limited. Here, we focus upon the implementation of the SF algorithm. We add to the original algorithm the local random selection of features, measures of goodness of fit and goodness of prediction, and isotonic regression to estimate levels of survival. We use residuals to assess fit and the c-index and the Brier score to gauge quality of prediction,

For a single bootstrap sample, b, and out-of-bag $\mathbf{x}$ the survival tree will produce an estimator of the cumulative hazard, $\hat{H}_b(t, \mathbf{x})$, and the survival function $\hat{S}_b(t, \mathbf{x}) = \exp(-\hat{H}_b(t, \mathbf{x}))$ at selected time points. The aggregated estimators of S and H are $\hat{S}(t, \mathbf{x}) = \sum_{b \in B_{\mathbf{x}}} \hat{S}_b(t, \mathbf{x})/|B_{\mathbf{x}}|$ and $\hat{H}(t, \mathbf{x}) = \sum_{b \in B_{\mathbf{x}}} \hat{H}_b(t, \mathbf{x})/|B_{\mathbf{x}}|$, where for each $\mathbf{x}$, $B_{\mathbf{x}} = \{b : \mathbf{x} \text{ is an } oob \text{ case}\}$ and $|B_{\mathbf{x}}|$ = number in $B_{\mathbf{x}}$. Using only the *oob* cases reduces overfitting.

Probability p_s to choose between splitting along the covariate or the time. Breiman suggested splitting probabilities larger than 0.50. If $p_s = 1$, the covariates are ignored, optimization works on time alone and the estimators $\hat{S}(t, \mathbf{x})$ tend to a common curve. If $p_s = 0$ the tree growing optimizes on the covariates $\mathbf{x}$ alone, the curves $\hat{S}(t, \mathbf{x})$ tend to be diverse, and poor predictive properties observed. A robust interval for p_s can be found by experimentation.

Selecting locally m out of k covariates. Borrowing an idea implemented in random forests, we introduce local random selection of m out of the total number of covariates, k. The starting value recommended for this parameter, known as *m-try*, was $\log_2(k) + 1$. At each node left to be split along $\mathbf{x}$, the best split is chosen from the randomly selected predictors. This reduces the correlation among the trees, but maintains the strength of each tree's predictive capability.

Residuals. The property that the transformed survival time, $H(T(\mathbf{x}), \mathbf{x})$, has an exponential(1) distribution suggests treating the Cox-Snell residuals $(\hat{H}(t(\mathbf{x}_i), \mathbf{x}_i), c_i)$ as a censored sample from that distribution. We assess fit with the statistic $\sum_{i=1}^{N} \frac{1}{N} |\hat{H}_{\mathrm{NA}}(\hat{H}(t(\mathbf{x}_i), \mathbf{x}_i)) - \hat{H}(t(\mathbf{x}_i), \mathbf{x}_i)|$, where $\hat{H}_{\mathrm{NA}}$ is the Nelson-Aalen estimator. Martingale residuals, $c_i - \hat{H}(t(\mathbf{x}_i), \mathbf{x}_i)$, can also serve for fit assessment. Means of residuals are negative for $\hat{H}(t, \mathbf{x})$ too large and positive for $\hat{H}(t, \mathbf{x})$ too small. These residuals although ad-hoc help making bad fits stand out.

Harrell c-index. The c-index [9] is the proportion of predictions that are concordant out of all pairs of observation for which ordering of the survival times can be determined. This measure uses both noncensored and censored survival times. It ranges from 0 to 1, and equals 0.5 for the constant predictor. A c-index near 0.5 means that the model is not predictive while a c-index near 1 indicates that the model is highly predictive. From each curve several predictors may be derived, e.g., means or medians, and a c-index computed for each of them.

Brier integrated score. For $\mathbf{x}$ given $\hat{S}(t|\mathbf{x})$ is a prediction probability that the event $T(\mathbf{x}) \geq t$ occurs. To evaluate the quality of prediction adjusted for censoring, we use the Brier mean squared error scores (Graf et al. [8]) $B_s(t) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{\hat{G}(t_i)} \hat{S}(t|\mathbf{x}_i)^2 c_i 1_{t_i < t} + \frac{1}{\hat{G}(t)} [1 - \hat{S}(t|\mathbf{x}_i)]^2 1_{t_i \geq t} \right]$ where t_i are observed times, $t > 0$ and $\hat{G}(t)$ the Kaplan-Meier estimate of the distribution of the time to censor assumed free of $\mathbf{x}$. If all $c_i = 1$, then $\hat{G} \equiv 1$ and the above is the traditional mean square error. The integrated Brier score for global assessment is $B_I = \frac{1}{\max(t_i)} \int_0^{\max(t_i)} B_s(t) dt$.

5 Applying SF to T-ALL data

Observations and comments. When assessing the impact of the tuning parameters we focus mainly on the goodness of prediction measures, the c-index and integrated Brier score, B_I. Experimental runs on real and simulated data (not presented) show that for a fixed p_s, B_I stabilizes with mild fluctuations for *m-try* above some threshold, and for a fixed *m-try*, B_I is minimum at a certain value of p_s.

A node l is split to increase $D_l \log(D_l/T_l)$. Thus, a small value of T_l may lead to a large single bootstrap estimate of the cumulative hazard, and strong influence of few observations. If $\hat{H}(t, \mathbf{x})$ is too large at some t, then it may appear too short tailed. To reduce impact of sporadic spikes on the estimation, a node splitting constraint, e.g., $D_l/T_l <$ constant, can be added.

Frequently, different prediction functions of $\mathbf{x}$ give similar c-indexes, suggesting that the covariate vectors can be ordered to define prediction classes. When the predictions appear ordered correctly but out of line with the data some calibration on the prediction may be needed.

Kaplan-Meier estimator $\hat{S}_{KM}(t)$	Individual $\hat{S}(t)$	Average $\overline{\hat{S}}(t)$
0.218	0.192	0.222

Table 1: Brier scores for the survival curves.

gene	description
U51240	KIAA0085 gene. partial cds
HG3521-HT3715	Ras-Related Protein Rap1b
M33552	Lymphocyte-specific protein 1 (LSP1) mRNA
J04182	LAMP1 Lysosome-associated membrane protein 1
Y00796	ITGAL Integrin. alpha L
U43185	STAT5A Signal transducer and activator of transcription 5A
X72889	Transcriptional activator hSNF2a

Table 2: The seven most important genes for survival as determined by SF.

Fitting SF to survival with gene profiles as covariates. In the T-ALL data consisting of 39 observations and the reduced set of 72 genes, only 12 survival times are noncensored. To achieve a reasonable resampling convergence 10,000 independent bootstrap samples and resulting survival trees are generated. The survival curves are evaluated for goodness of prediction by the Brier integrated score, B_I. We determine that $p_s = .10$ and *m-try* =10 provide small B_I values (around 0.19). The small value of p_s is most likely the result of having only 9 distinct and rather tight event times, thus making the "optimal" SF close to a regression tree. When using the individual curves $\hat{S}(t, \mathbf{x})$ instead of the average curve, $\overline{\hat{S}}(t) = \frac{1}{39}\sum_{i=1}^{39} \hat{S}(t, \mathbf{x}_i)$, the rate of improvement is the ratio of the B_I's for respective curves. Its value of 0.86 indicates that covariates are important to survival prediction.

One of the difficulties incurred by the small sample resides in the adverse "regression to the mean" effect. If an extreme time is out-of-bag, then for a given bootstrap iteration, its prediction tree tends to be computed from non-extreme observations. For four patients whose recorded survival times are null, the survival curve are estimated with a higher degree of bias than for the others. The 13 highest times being censored the corresponding fitted curves all have heavy right tails. The probabilities of survival at the maximum observed time range from 0.31 to 0.81. For the prediction we use $E(T|\mathbf{x}) = \int S(t, \mathbf{x})dt$ and linear interpolation and extrapolation. To find ordered prediction levels we apply isotonic regression to smooth the bootstrap expectations into a smaller number of constants, i.e., the isotonized average.

Survival prediction and variable importance. The c-index is about 0.60 for the average survival curve, and .70 for the isotonized prediction. The latter is comparable with the value one of us obtained (c-index=.71 [10]) for a prognostic model in metastatic breast cancer using clinical, not molecular, features and a combination of a Bayesian neural network and a regression tree. It is also comparable to documented values of c-index in other clinical studies. The Brier scores given in Table 1 show that, compared to the Kaplan-Meier estimator $\hat{S}_{KM}(t)$, the average survival curve $\overline{\hat{S}}(t)$ displays similar predictive performance and the individual curve $\hat{S}(t)$ provides better performance, an improvement of 12% in B_I reduction.

The plot of predicted survival for 39 T-ALL samples in Figure 3 identifies clearly three classes of tumor whose isotonized survival times correspond to

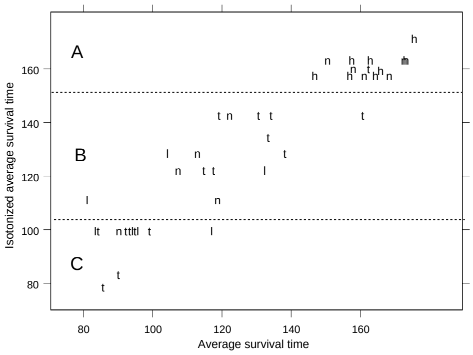


Figure 3: Isotonized survival prediction vs average survival time.

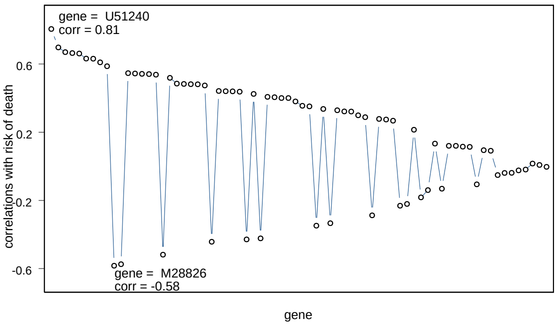


Figure 4: Variable/gene expression importance

the partition of time into three intervals defined by $0 \leq 102 \leq 150 \leq 200^{+}$ months, that is the classes of samples with poor (C), neutral (B) or good prognosis (A). It is remarquable that the class A is identical to Ferrando et al.'s cluster H (Figure 1) that contains all samples that express HOX11^{+}, and the like, and one t sample as detected by RT-PCR. The finding that the class A is clearly disjoint from the two other classes is coherent with the fact that the log-rank test for comparing survival in these classes was significant.

Class B contains twelve t samples and the like, and three l samples. Class C contains two l samples, the two mixed tl samples, and six t samples. The clustering of l samples within classes of t type is coherent with the fact that the survival of the l cases and the t cases are not different by the log-rank test. It may also just reflect the fact that genes do not necessarily act in a hierarchical ordered manner.

The set of covariates and curves $(\mathbf{x}, \hat{S}(t, \mathbf{x}))$ can be further analyzed to determine variable importance. Breiman suggested graphing, for each individual covariate component x_p, the estimator $\widehat{\mathrm{Corr}}(x_p, -\log(\hat{S}(t, \mathbf{x})))$ as a function of t . For the T-ALL data the lack of dispersion among the noncensored time points results in almost constant functions in time of these statistics. We use means over time of $\widehat{\mathrm{Corr}}(x_p, -\log(\hat{S}(t, \mathbf{x})))$, ranging from -0.58 to 0.81, as indicators of importance. Figure 4 exhibits their plot vs the index of the genes ranked in decreasing importance. The notable fact is that the distribution of these correlations is skewed. Their sign is more frequently positive than negative (over 4 times), and the absolute values of the positive correlations are on average larger than those of the negative ones, thus indicating that the list of genes is more related to risk of death than chance of survival. Seven genes exhibiting the highest absolute correlation with survival, i.e., seven most important genes for survival in T-cells acute lymphoblastic leukemia, are listed in Table 2.

References

[1] Alizadeh A.A., Elsen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Ran T., Yu X. (2000). *Distinct types of diffuse large B-cell lymphoma identiffed by gene expression profiling.* Nature **403**, 503-511.

[2] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984). *Classification and regression trees.* Wadsworth, Belmont, CA.

[3] Breiman L. (1996). *Stacked regressions.* Machine Learning **24**, 41 – 48.

[4] Breiman L. (2001). *Random Forests.* Machine Learning **45**, 5 – 32.

[5] Breiman L. (2002). *Wald III Lecture: Software for the masses.* Lecture notes available at http://stat www.berkeley.edu/users/breiman/wald2002-3.pdf

[6] Collins F.S., Green E.D., Guttmacher A.E. and Guyer M.S., on behalf of the US National Human Genome Research Institute* (2003).*A vision for the future of genomics research A blueprint for the genomic era.* Nature, **422**, 835 – 847.

[7] Ferrando A.A., Neuberg D.S., Staunton J., Loh M.L., Huard C., Raimondi S.C., Behm F.G., Pui C.H. Downing J.R., Gilliland D.G., Lander E.S., Golub T.R. and Look A.T. (2002). *Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia.* Cancer Cell, **1**, 75 – 87.

[8] Graf Z., Smchoor C., Sauerbrei W., and Schumacher M. (1999). *Assessment and comparison of prognosis classification for survival data.* Statistics in Medicine, **18**, 2529 – 2543.

[9] Harrell FE., Lee KL. and Mark DB.(1996). *Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Statistics in Medicine **15**, 361 – 387.

[10] Hoàng T., Trinh QA. and Asselain B. (2002). *Construction and validation of a prognostic model for metastatic breast cancer using Bayesian neural network and regression tree.* Intelligent Data Analysis in Medicine and Pharmacology, Workshop Notes, 37 – 43. `http://www.cs.uu.nl/ lucas/idamap2002/idamap2002-proc.pdf`

[11] LeBlanc M. and Tibshirani R.(1993). *Combining estimates in regression and classification.* Technical Report 9318, Department of Statistics, University of Toronto, Toronto, O.N.

Address: T.M. Hoàng, Université René Descartes, France
Van L. Parsons, National Center for Health Statistics, USA
E-mail: `hoang@biomedicale.univ-paris5.fr`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

WEB-BASED ANALYSIS SYSTEM IN DATA-ORIENTED STATISTICAL SYSTEM "DoSS@*d*"

Keisuke Honda, Yuichi Mori, Yoshiro Yamamoto and Hiyoshi Yadohisa

Key words: Web application, client/server system, XML, statistical education.

COMPSTAT 2004 section: E-statistics.

Abstract: A web-based statistical system, the "Data-oriented Statistical System (**DoSS@*d***), is presented. The system is a kind of database which stores real data sets and corresponding analysis stories with an online analysis system as an interactive tool. The system is implemented using technologies such as Java and XML. The online analysis system is illustrated in detail, which provides an easy-to-access statistical environment where users learn the skills of real-world data analysis via the Internet.

1 Introduction

The current broadband network environment has made the distribution of applications via the web easier than ever before. Interactive and dynamic client/server (C/S) systems utilizing new technologies such as Java and extensible markup language (XML) rather than older static systems such as hypertext markup language (HTML) have also provided universal accessibility, whereby users in many different environments can now use almost the same functions on the web as possible on standalone computers. These web-based systems also have the advantages of ease of maintenance and instant updates, as well as offering platform-independent applications using Java and XML.

One area where a web-based solution can be expected to be of significant benefit is data archival as an educational tool. Successful data analysis involves the development of problem solving ability through the collection of data, the selection of a suitable statistical method, analysis of the data, interpreting the results and finally making decisions based on the interpretation. The ability to make full use of statistical software is also desirable, involving the specification of parameters based on data attributes and operating the statistical package to obtain the desired information.

The number of statistical education programs focusing on the analysis of real data using a statistical package has increased, superseding text-based and method-oriented education. Many teaching strategies have been proposed based on educational trials, and many data sets suitable for educational use have been published on the Internet (e.g., Chance Database [3], Data and

Story Library [4], Statlib [8], Data Representation System [5] and MD*Base [6]). However, it remains difficult for users to find data sets suitable for their intended purpose and obtain documents that describe how to analyze the data. The compilation of good examples is therefore considered important for learning processes and procedures relevant to past analyses. The archiving of example data sets and the corresponding example analyses is also expected to facilitate in the learning of statistical packages in a computational environment, where users can follow the steps of the example computation as an instructional tool.

Recognizing the potential of such as system, the authors began development of a kind of databank on the Internet, in which data sets are classified by research subject and statistical method. This databank represents an online database of data sets and documentation describing the processes of the original analyses (we call this kind of documentation “analysis story”), and also incorporates an online analysis system that performs automatic analysis based on the analysis story (i.e., using the same parameters as ones in the original analysis). This web-based system therefore consists of two functions: a database of typical real-world data sets and analysis stories, and an analysis system with a graphical user interface (GUI) to allow data sets in the database to be analyzed online. This environment has been name the “Data-oriented Statistical System” or **DoSS**$^{@}$***d*** [2], where “@d” reinforces that the system is used for real data. When utilized for statistical education, teaching scenarios can be developed easily, giving the students the chance to learn various statistical techniques using real data sets as well as mastering statistical software using the online analysis function. Students can also perform there own analysis to confirm the results of the analysis story through the use of simple operations, and can easily examine the effect of using different parameters.

This paper presents an outline of **DoSS**$^{@}$***d*** and the details of prototype of the online analysis system implemented in **DoSS**$^{@}$***d*** as a C/S system in Java and XML using R and XploRe as statistical engines is also presented.

2 DoSS$^{@}$*d*

DoSS$^{@}$***d*** is located `http://mo161.soci.ous.ac.jp/@d/index.html` and consists of three subsystems, **DoDStat**$^{@}$***d*** (Data-oriented Database of Statistics), **DoAStat**$^{@}$***d*** (Data-oriented Analysis System of Statistics), **DoLStat**$^{@}$***d*** (Data-oriented Learning System of Statistics) (Figure 1).

2.1 DoDStat$^{@}$*d*

DoDStat$^{@}$***d*** is the database system of **DoSS**$^{@}$***d***. Each stored data set consists of a data description and the data body. The former is written in XML and describes attributes such as data name, case name, variable names, and variable types. The latter is provided in several formats, including tab-,

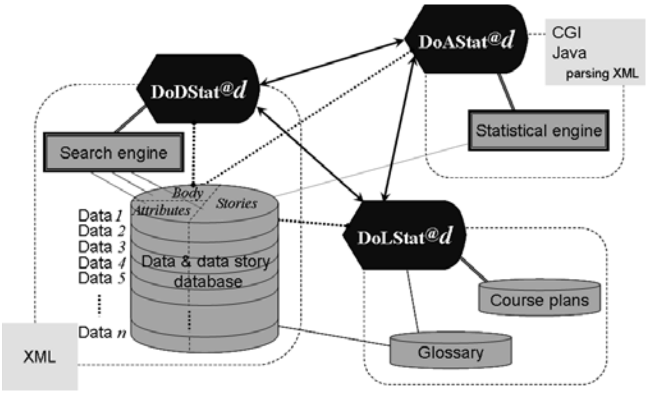


Figure 1: Stucture of **DoSS@*d***.

comma- and space-delimited values. **DoDStat@*d*** also stores analysis stories written in XML. The user is able to select an interesting or appropriate data set using a retrieval key such as research subject, statistical method and keyword.

2.2 DoAStat@*d*

DoAStat@*d* is a web-based application for the analysis of any data set stored in **DoDStat@*d***, as well as data sets stored on the local computer. Currently this system executes data analysis using R or XploRe Quantlet Server (XQS) as a statistical engine. The R Server-based system DoA_R communicates with the server by common gateway interface (CGI), and XQS-based system DoA_X is programmed in Java. **DoAStat@*d*** is described in more detail in the next section.

2.3 DoLStat@*d*

DoLStat@*d* is a learning system, in which a variety of learning courses such as “Statistics introductory course” and “Economics course” are provided based on analysis stories stored in **DoDStat@*d*** according to the study target.

3 DoAStat@*d*

In addition to allowing users to analyze data sets, **DoAStat@*d*** system also provides a function that allows the users to easily obtain the same results as descried in the analysis story of the data by automatically importing the parameters stored in the XML document of the analysis story. Two versions of **DoAStat@*d*** have been implemented; a CGI version using R Server (DoA_R), and a Java version using XQS (DoA_X). The latter is discussed in this paper. The architecture of **DoAStat@d** is outlined in Figure 4.

3.1 DoA_X

DoA_X communicates with XQS on the network using a Java communication interface called MD*Crypt [1,7]. This interface provides several GUIs for statistical techniques such as principal component analysis (PCA) and regression analysis, allowing users to operate the GUI without requiring knowledge of the analysis scripts (quantlet in XploRe) or methods. That is, this interface hides the original XploRe interface, XQC (XplpRe Quantlet Client), allowing users to perform data analysis without needing to study a new statistical language each time. Users select a data set stored in **DoDStat@d** and a statistical method to be applied in the top page of **DoAStat@d** (upper left of Figure 2). DoA_X starts automatically when the [Execute analysis] button is clicked, and reads the data from the server (lower right of Figure 2).

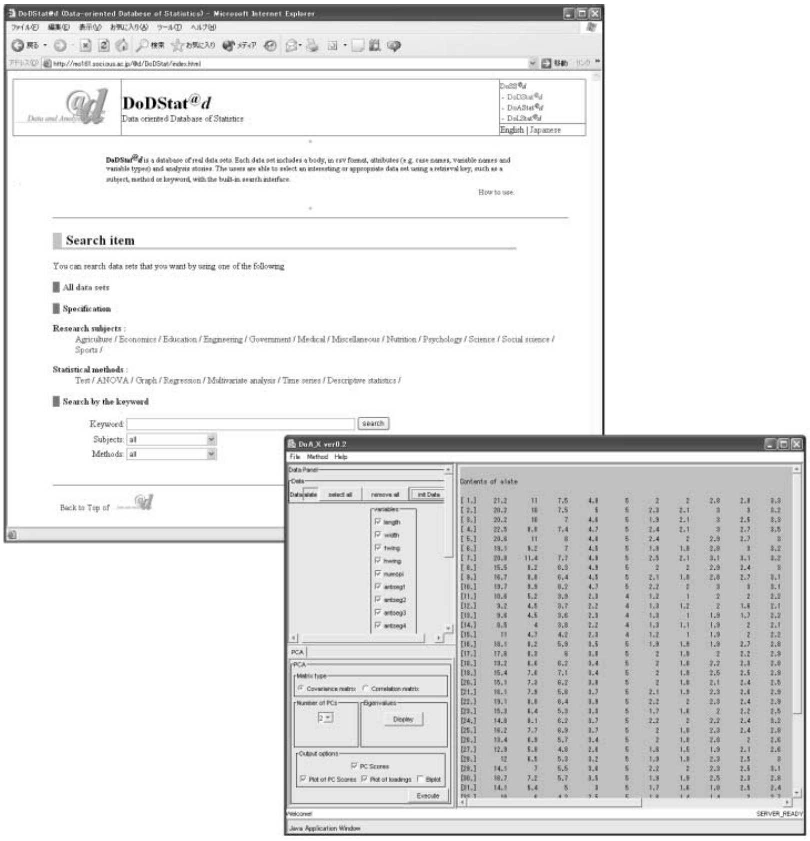
Users perform the analysis by selecting variables and specifying parameters in the same way as in ordinary statistical packages. DoA_X then returns the results of the analysis. This function can be used directly from the analysis story page (Figure 3). Clicking the [Analysis] button in the story page brings up the same GUI appears but with all initial parameters such as matrix type and number of components for PCA set based on the analysis story.

3.2 Architecture of DoA_X

DoA_X is written in Java and consists of an XML parser, a network handling section, and a GUI (Figure 4).

3.2.1 XML parser. This component parses two XML documents; the data and the analysis story. These XMLs are downloaded from the server and mapped as Java objects. Figure 5 shows an excerpt of an XML document containing an analysis story for “Physical measurements of alate adelges”. The parser reads text, abstracts “pca” and “yes” in the `method` tag and “2” and “cov” in the `option` tag, and then passes the result to the GUI (Figure 6).

3.2.2 Network. The web application DoSS_Server is installed as part of **DoSS@d** to realize network functionality as a Java servlet with middleware between DoA_X and the database. The role of DoSS_Server is to handle content (data and story XMLs) as elements in the database and offer the content to the client DoA_X according to user request. Thus, the software installed on the client machine (DoA_X) acts only as an interface. DoA_X behaves as if it hosts all the resources on the client machine, and performs analysis interactively. Dynamic use of content provides flexibly with respect to future changes of specification and increases in the volume of data. Communication with XQS through MD*Crypt is executed independently of DoSS_Server, re-



DoDStat@d (Data-oriented Database of Statistics) - Microsoft Internet Explorer

http://mo161.socious.ac.jp/@d/DoDStat/index.html

DoDStat@d

Data oriented Database of Statistics

DoSS@d
- DoDStat@d
- DoAStat@d
- DoLStat@d

English | Japanese

DoDStat@d is a database of real data sets. Each data set includes a body, in csv format, attributes (e.g. case names, variable names and variable types) and analysis stories. The users are able to select an interesting or appropriate data set using a retrieval key, such as a subject, method or keyword, with the built-in search interface.

How to use.

Search item

You can search data sets that you want by using one of the following

All data sets

Specification

Research subjects :
Agriculture / Economics / Education / Engineering / Government / Medical / Miscellaneous / Nutrition / Psychology / Science / Social science / Sports /

Statistical methods :
Test / ANOVA / Graph / Regression / Multivariate analysis / Time series / Descriptive statistics /

Search by the keyword

Keyword: search
Subjects: all
Methods: all

Back to Top of @d

DoA_X ver0.2

Data Panel — Data: slate

Variables: ☑ length ☑ width ☑ fwing ☑ hwing ☑ numspi ☑ antseg1 ☑ antseg2 ☑ antseg3 ☑ antseg4

PCA

Matrix type: ● Covariance matrix ○ Correlation matrix

Number of PCs: 2

Output options: ☑ PC Scores ☑ Plot of PC Scores ☑ Plot of loadings □ Biplot

```
Contents of slate

[ 1,]   21.2    11   7.5   4.8   5   2     2     2.9   2.8   3.3
[ 2,]   20.2    10   7.5     5   6   2.3   2.1     3     3   3.2
[ 3,]   20.2    10     7   4.6   5   1.9   2.1     3   2.5   3.3
[ 4,]   22.5   8.8   7.4   4.7   5   2.4   2.1     3   2.7   3.5
[ 5,]   20.6    11     8   4.8   5   2.4     2   2.9   2.7     3
[ 6,]   19.1   9.2     7   4.5   5   1.8   1.8   2.8     3   3.2
[ 7,]   20.8  11.4   7.7   4.9   5   2.5   2.1   3.1   3.1   3.2
[ 8,]   15.5   8.2   6.3   4.9   5     2     2   2.9   2.4     3
[ 9,]   16.7   8.8   6.4   4.5   5   2.1   1.8   2.8   2.7   3.1
[10,]   19.7   9.9   8.2   4.7   5   2.2     2     3     3   3.1
[11,]   10.6   5.2   3.9   2.3   4   1.2     1     2     2   2.2
[12,]    9.2   4.5   3.7   2.2   4   1.3   1.2     2   1.6   2.1
[13,]    9.6   4.5   3.6   2.3   4   1.3     1   1.9   1.7   2.2
[14,]    8.5     4   3.8   2.2   4   1.3   1.1   1.9     2   2.1
[15,]     11   4.7   4.2   2.3   4   1.2     1   1.9     2   2.2
[16,]   18.1   8.2   5.9   3.5   5   1.9   1.8   1.9   2.7   2.8
[17,]   17.8   8.3     6   3.8   5     2   1.9     2   2.2   2.9
[18,]   18.2   6.6   6.2   3.4   5     2   1.8   2.2   2.3   2.8
[19,]   15.4   7.6   7.1   3.4   5     2   1.8   2.5   2.5   2.9
[20,]   15.1   7.3   6.2   3.8   5     2   1.8   2.1   2.4   2.5
[21,]   16.1   7.9   5.8   3.7   5   2.1   1.9   2.3   2.6   2.9
[22,]   19.1   8.8   6.4   3.9   5   2.2     2   2.3   2.4   2.9
[23,]   15.3   6.4   5.3   3.3   5   1.7   1.6     2   2.2   2.5
[24,]   14.8   8.1   6.2   3.7   5   2.2     2   2.2   2.4   3.2
[25,]   16.2   7.7   6.9   3.7   5     2   1.8   2.3   2.4   2.8
[26,]   13.4   6.9   5.7   3.4   5     2   1.8   2.0     2   2.6
[27,]   12.9   5.8   4.8   2.6   5   1.8   1.5   1.9   2.1   2.6
[28,]     12   6.5   5.3   3.2   5   1.9   1.8   2.3   2.5     3
[29,]   14.1     7   5.5   3.4   5   2.2     2   2.3   2.5   3.1
[30,]   18.7   7.2   5.7   3.5   5   1.9   1.9   2.5   2.3   2.8
[31,]   14.1   6.4     5     3   5   1.7   1.6   1.8   2.5   2.4
```

Welcome!

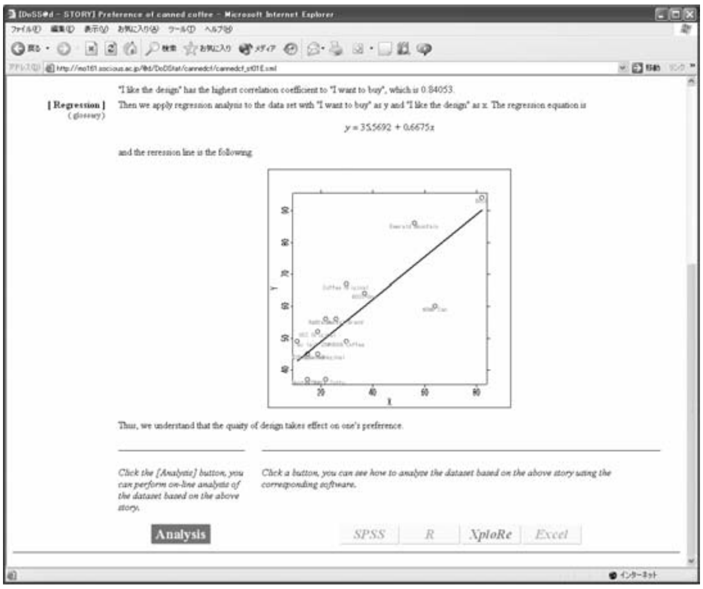
SERVER_READY

Java Application Window

Figure 2: Top page of **DoDStat@d** and GUI of DoA_X.

ceiving parameters from the GUI, transforming them to scripts and delivering them to XQS on the server.

3.2.3 GUI. DoA_X was developed as a Java application, and the GUI was constructed based on Swing as a standard component. As such, the GUI is a more intuitive and interactive user interface than web applications based on HTML. Java Web Start is utilized to deliver DoA_X to client machines. Using this system, DoA_X is automatically installed on the client machine when the [Execute analysis] button is first clicked. Subsequently, the software is preserved on the local machine and will not be re-downloaded until the original DoA_X has been updated. Thus, the software installed on the client machines can keep its latest version.


[DoSS@d – STORY] Preference of canned coffee – Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り メディア
アドレス(D) http://mo161.socious.ac.jp/@d/DoDStat/cannedcf/cannedcf_st01E.xml
移動 リンク
x
y
20
40
60
80
インターネット

"I like the design" has the highest correlation coefficient to "I want to buy", which is 0.84053.

[Regression] (glossary)

Then we apply regression analysis to the data set with "I want to buy" as y and "I like the design" as x. The regression equation is

$$y = 35.5692 + 0.6675x$$

and the reression line is the following.

Thus, we understand that the quaity of design takes effect on one's preference.

Click the [Analysis] button, you can perform on-line analysis of the dataset based on the above story.

Click a button, you can see how to analyze the dataset based on the above story using the corresponding software.

Analysis SPSS R XploRe Excel

Figure 3: Analysis story page.

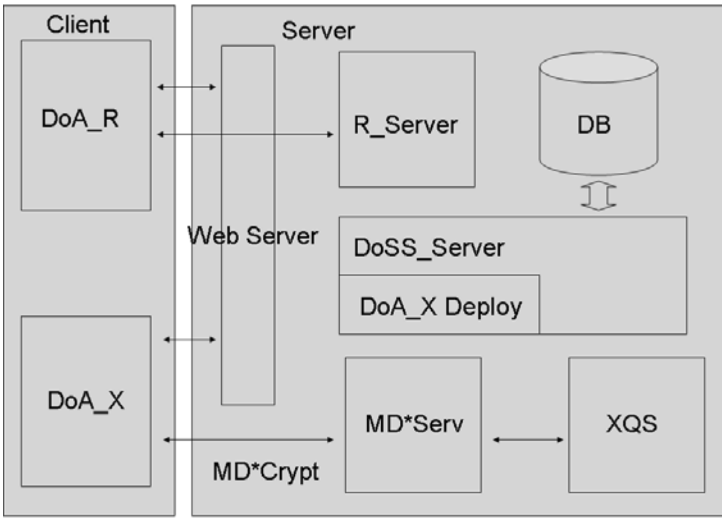
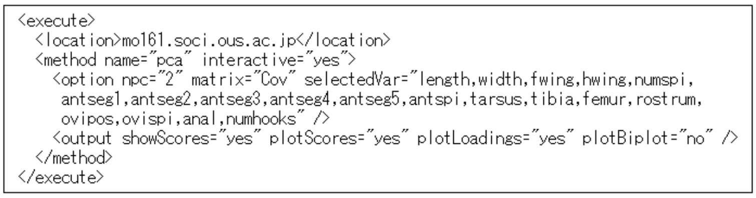
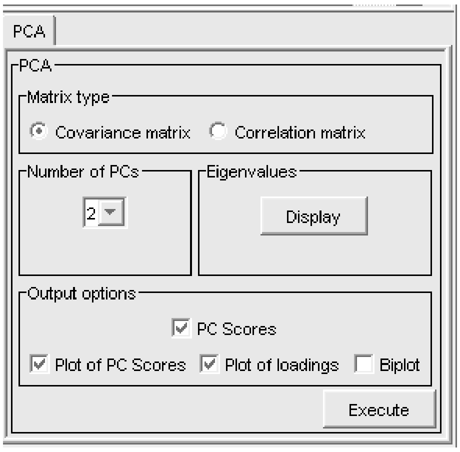


Figure 4: Architecture of **DoAStat@*d***.



```
<execute>
  <location>mo161.soci.ous.ac.jp</location>
  <method name="pca" interactive="yes">
    <option npc="2" matrix="Cov" selectedVar="length,width,fwing,hwing,numspi,
     antseg1,antseg2,antseg3,antseg4,antseg5,antspi,tarsus,tibia,femur,rostrum,
     ovipos,ovispi,anal,numhooks" />
    <output showScores="yes" plotScores="yes" plotLoadings="yes" plotBiplot="no" />
  </method>
</execute>
```

Figure 5: A part of Story XML (for PCA).



PCA

Matrix type: ● Covariance matrix ○ Correlation matrix

Number of PCs: 2

Eigenvalues: Display

Output options: ☑ PC Scores ☑ Plot of PC Scores ☑ Plot of loadings □ Biplot

Figure 6: Parameter specification (for PCA).

3.3 Characteristic of DoA_X

The DoA_X system allows users to perform immediate analysis of any data set stored in **DoDStat@*d*** using the built-in analysis system, and can also do so using pre-set parameters. Using the GUI, users can analyze data without needing to have knowledge of macros or script languages. There are also a number of advantages in constructing a web-based system, such as universal accessibility, ease of maintenance, instant updates, and automated installation. Finally, the developer can construct the desired system for statistics in a short time using the XploRe C/S system, which handles the TCP/IP process internally through MD*Crypt.

4 Concluding remarks

The authors have developed a web-based statistics system (mainly for educational use) that archives data sets and corresponding analysis stories and allows statistical analysis to be performed directly on these data sets on the web. Data sets are stored as XML documents in the database and are collected from various fields. The XML format allows the data to be categorized in terms of research subject and statistical method. The system also provided a user-friendly GUI for easy analysis.

Future development of the system will include refinement of the GUI design for DoA_X, increasing the number of methods that can be applied in **DoAStat@*d***, collecting more data sets and analysis stories in **DoDStat@*d***, and implementing a data-uploading/removing system and data-evaluation system. Furthermore, it will be necessary to re-design this system in order to deal with very large data set and computations considering efficient data transfer.

References

[1] Feuerhake J. (2002). *XQS/MD*Crypt as means of education and computation.* In: COMPSTAT 2002 Proceedings in Computational Statistics (Härdle, W., Rönz, B.), Heidelberg: Phisica-Verlag, 635 – 640.

[2] Mori Y., Yamamoto Y., Yadohisa H. (2003). *Data-oriented learning system of statistics based on analysis scenario/story (DoLStat).* Bulletin of the International Statistical Institute, 54th Session Proceedings Volume LX Two Books, Book 2, 74 – 77.

[3] Chance Database Home Page, `http://www.dartmouth.edu/~chance/`

[4] Data and Story Library (DASL), `http://lib.stat.cmu.edu/DASL/`

[5] Data Representation System (DRS), `http://www.sci.kagoshima-u.ac.jp/~drs/en/`

[6] MD*Base (Statistical methodology and interactive data analysis), `http://www.quantlet.org/mdbase/`

[7] MD*Tech, MD*Crypt, `http://www.md-crypt.com/`

[8] Statlib, `http://lib.stat.cmu.edu/`

Address: K. Honda, Graduate School, Okayama University of Science, 1-1 Ridai-cho, Okayama 700-0005, Japan
Y. Mori, Department of Socio-Information, Okayama University of Science, 1-1 Ridai-cho, Okayama 700-0005, Japan
Y. Yamamoto, Department of Mathematics, Tokai University, 1117 Kita-Kaname, Hiratsuka 259-1292, Japan
H. Yadohisa, Department of Mathematics and Computer Science, Kagoshima University, Kagoshima 890-0065, Japan

E-mail: `honda@soci.finfo.ous.ac.jp`, `mori@soci.ous.ac.jp`, `yoshiro@yamamoto.name`, `yado@sci.kagoshima-u.ac.jp`

THE INTERACTIVE EXERCISE TEXT BOOK

Karel Hrach

Key words: Statistical software, MS-Excel, teaching statistics.

COMPSTAT 2004 section: Teaching statistics.

1 The need for interactive exercise texts in statistics

In many tertiary institutions, university students are required to study statistics. Knowledge of theory as well as of practical skills in calculation are usually required to pass the examination. Many excellent textbooks have been written, many of which exist in electronic form (see references). As far as classical exercise books are concerned, one disadvantage is discernible and that is that the number of exercises they contain is always limited. The moment a reader attempts to calculate a similar type of exercise, he or she remains uncertain as to whether the result is correct. There exists the possibility to check such results via the application of more or less specialized statistical software, but not all readers can access or deal with it. In addition to this, such software does not always show the necessary steps involved in non-computer calculation thereby depriving the reader of the chance to understand the exact reason for the potentially incorrect result.

This is how the idea of preparing the simplest statistical software originated. The user is not required to do anything apart from enter the data together with some essential parameters. MS-Excel was chosen for this purpose, in combination with text in rtf format. Note that the reader is not required to be skilled in the use of MS-Excel. All the sheets are locked so that the reader only can enter the data and choose the required results. Unfortunately, the entire exercise book is only available in Czech language at present but an English one is in preparation.

Introduction
Basic descriptive statistics - frequencies, quantiles, moments
Probabilistic distributions - binomial, hypergeometric, Poisson, exponential, normal
Confidence intervals for probability, for mean, for variance
Testing hypothesis - one-sample, paired and two-sample t-tests, one-way ANOVA, chi-square test of goodness-of-fit and of independence
Simple linear regression and correlation, time-series description
Www links

Table 1: Contents of the main page.

2 Main page

The main page in rtf format (STA-MAIN.rtf) serves for the realized on-line statistical exercise book as the contents page (Table 1) and is recommended as a hypertext link from other www pages. It is located at the www address http://fse1.ujep.cz/MS_hrach.asp. It contains links to chapters covering a common syllabus of statistics and also leading to some other interesting www sources (statistical offices, statistical software, other on-line textbooks), some of which are in English.

Each chapter, written also in rtf format, summarizes briefly the main formulas and shows at least one solved example, followed by graphical link (small pie-chart) to the appropriate MS-Excel file. For example, graphical link from the chapter about analysis of variance (file STA-anova.rtf) leads to the file, where this type of test can be done (STA-anova.xls).

3 Excel files

Every prepared MS-Excel file offers the sheet ZADANI (setting) automatically. All the cells are locked here except for the yellow cells for entering data. Some files require the use of prepared macro (under the keyboard abbreviation Ctrl+T) but most of the results are immediately available on the appropriate sheets. If necessary, it is possible on the subsequent sheets to enter such parameters, as the level of significance alpha. The last sheet in each file is List1 containing only the help calculations. The sheets automatically check values by entering (e.g. if a variance is positive when the confidence interval for mean with the known variance is performed).

For example, file STA-anova.xls contains sheets ZADANI (setting), Momenty (with descriptive characteristics), Test (with the test results after entering alpha) and List1.

4 Final remarks

This interactive exercise book could be extremely useful for students, mainly those studying part-time, in the form of supporting material for the study of the basic topics of statistics. But it could also be used by teachers as a tool for correcting calculations, step by step. The advantage of its structure is that other chapters or language versions can simply be added.

References

[1] *Zaklady statistiky.*
http://www.quantlet.com/mdstat/scripts/mmcze/java/start.html

[2] *E-books, combining statistical methods and data.*
http://www.xplore-stat.de/ebooks/ebooks.html

[3] *Elektronicke ucebnice EuroMISE-centra.*
http://ucebnice.euromise.cz/index.php?conn=0&section=knihy

[4] *Ucebnice StatSoft*.
http://www.statsoft.cz/textbook/stathome.html

[5] *On-line ucebnice statistiky*.
http://badame.vse.cz/ucebnice.php

Acknowledgement: Preparation of the interactive statistical exercise book was in the year 2003 supported by the grant FRVS F5/2659 of the Czech Ministry of education.

Address: K. Hrach, VSEM, Herbenova 10, 40001 Ústi nad Labem, Czech Republic

E-mail: `hrach@vsem.cz`

COMPSTAT'2004 Symposium

BAYESIAN LIKE PROCEDURES FOR DETECTION OF CHANGES BASED ON EMPIRICAL CHARACTERISTIC FUNCTIONS

Marie Hušková and Simos G. Meintanis

Key words: Empirical characteristic functions, change point analysis, numerical algorithms.

COMPSTAT 2004 section: Statistical software, Nonparametric statistics.

Abstract: A class of Bayesian like procedures for detection of a change in the distribution of a sequence of independent observations based on empirical characteristic functions is developed and their limit properties are studied. Theoretical results are accompanied by a simulation study.

1 Introduction and procedures

Let $Y_1, \ldots, Y_n$ be independent random variables, Y_j has a distribution function $F_j, j = 1, \ldots, n$. We consider the testing problem

$$H_0 : F_1 = \ldots = F_n \tag{1}$$

against

$$H_1 : F_1 = \ldots = F_m \neq F_{m+1} = \ldots = F_n \qquad for \quad m < n, \tag{2}$$

where m, F_1 and F_n are unknown. The nonparametric type procedures considered in the literature are based either on empirical distribution functions, quantile functions or U-statistics, for survey of the recent results see, e.g., the book by Csőrgö and Horváth [3], Brodsky and Darkhovsky [2] and a survey paper Antoch et al. [1]. Hušková and Meintanis [4] developed max-type test procedures based on empirical characteristics functions, studied their limit behavior both under the null hypothesis and alternatives and conducted simulation study. They proposed a data based approximation of the critical values that lead to the tests with prescribed level α. The advantage of the procedures based on empirical characteristic functions is that they work under quite weak assumptions. They can be even applied for discrete distributions and heavy tailed distributions. However, the asymptotic distributions under H_0 of these procedures depend on the unknown distribution F_1.

In the present paper *Bayesian like test procedures* based on empirical characteristic functions for testing problem H_0 against H_1 are introduced and their properties are studied. Particularly, the following class of test statistics is studied:

$$T_{n,B}(q, w) = \sum_{k=1}^{n} q_{k,n} \frac{k(n-k)}{n} \int_{-\infty}^{\infty} |\phi_k(t) - \phi_k^0(t)|^2 w(t) dt, \tag{3}$$

where $q_{k,n}$, $k = 1, \dots, n$ represent prior, w is a nonnegative weight function, $\phi_k(t)$ and $\phi_k^0(t)$ are empirical characteristic functions based on $Y_1, \dots, Y_k$ and $Y_{k+1}, \dots, Y_n$, respectively, i.e.

$$\phi_k(t) = \frac{1}{k}\sum_{j=1}^{k} \exp\{itY_j\}, \quad \phi_k^0(t) = \frac{1}{n-k}\sum_{j=k+1}^{n} \exp\{itY_j\}, \quad k = 1, \dots, n. \tag{4}$$

The test statistics (3) can be alternatively expressed as

$$T_{n,B}(q,w) = \sum_{k=1}^{n} q_{k,n}(\beta)\frac{k(n-k)}{n}V_{k,n}(w), \tag{5}$$

with

$$V_{k,n}(w) = \frac{1}{k^2}\sum_{l,m=1}^{k} h_w(Y_l, Y_m) + \frac{1}{(n-k)^2}\sum_{l,m=k+1}^{n} h_w(Y_l, Y_m) \tag{6}$$

$$-\frac{2}{k(n-k)}\sum_{l=1}^{k}\sum_{m=k+1}^{n} h_w(Y_l, Y_m),$$

$$h_w(x,y) = \int_{-\infty}^{\infty} \cos((x-y)t)w(t)\,dt, \tag{7}$$

which is more appropriate for computations. We consider the prior $q_{k,n}$, $k = 1, \dots, n$ of the form

$$q_{k,n}(\beta) = \frac{1}{n}\Big(\frac{k(n-k)}{n^2}\Big)^{-\beta}, \quad k = 1, \dots, n-1, \tag{8}$$

where $\beta < 1$. From the Bayesian point of view it means that the prior distribution of the change point m has the form

$$P(m = k) = c_n q_{k,n}, \quad k = 1, \dots, n,$$

where c_n is determined in such a way that $\sum_{k=1}^{n} P(m = k) = 1$. Clearly, for $\beta > 0$ change point m close to 1 and n have quite high prior w.r.t. prior of a change in the middle while vice versa for $\beta < 0$. The below presented results hold true under more general assumptions on prior but we consider (8) just for simplicity. We let

$$T_{n,B}(\beta,w) = \sum_{k=1}^{n}\Big(\frac{k(n-k)}{n^2}\Big)^{-\beta}\frac{k(n-k)}{n}V_{k,n}(w). \tag{9}$$

The choice of the weight function w influences the limit behavior of the considered test statistic. The results presented below hold true for any nonnegative weight function w with the property

$$0 < \int w(t)dt < \infty. \tag{10}$$

Typical choices are either

$$w_a^{(1)}(t) = \exp(-a|t|), \quad t \in R^1,\, a > 0, \tag{11}$$

or

$$w_a^{(2)}(t) = \exp(-at^2), \quad t \in R^1,\, a > 0, \tag{12}$$

and then the corresponding functions h_w in (7) have the form

$$h_w^{(1)}(x-y) = 2a/(a^2 + (x-y)^2), \quad z \in R^1. \tag{13}$$

and

$$h_w^{(2)}(x-y) = \sqrt{\pi/a}\, \exp(-(x-y)^2/4a), \quad z \in R^1, \tag{14}$$

respectively. The respective test statistics are denoted by $T_{n,B}^{(1)}(\beta, a)$ and $T_{n,B}^{(2)}(\beta, a)$. The role of the weight parameter $a > 0$ is to control the rate of decay of the weight function. For more details see, e.g. Meintanis [6].

Large values of the test statistics indicate that the null hypothesis is not true. Hence the null hypothesis is rejected when the critical value is exceeded, where the critical value is determined in such a way that the test has level α. Usually, reasonable approximations for critical values are based either on the limit distribution under the null hypothesis or on resampling methods. Unfortunately, the limit distributions depend on the underlying distribution of the observations (see Theorem A below) and hence this approach does not provide proper approximations for critical values. Resampling methods provide good approximation when the data follow the null hypothesis or local alternatives, however at any case they lead to consistent tests.

Next, we describe the application of bootstrap without replacement. The bootstrap version $T_{n,B}(\beta, w, \boldsymbol{R})$ of $T_{n,B}(\beta, w)$ is defined by (3) or equivalently by (5) with $Y_1, \dots, Y_n$ is replaced by $Y_{R_1}, \dots, Y_{R_n}$, where $R_1, \dots, R_n$ is a random permutation of $1, \dots, n$ independent on $Y_1, \dots, Y_n$. The critical value $d_{n,\gamma}(\alpha, \boldsymbol{Y})$ is obtained as the $100(1-\alpha)\%$ quantile of the conditional distribution of $T_{n,B}(\beta, w, \boldsymbol{R})$ given $Y_1, \dots, Y_n$. The resulting tests then reject H_0 on the level α if

$$T_{n,B}(\beta, w) \geq d_{n,\gamma}(\alpha, \boldsymbol{Y}). \tag{15}$$

Since the conditional distribution function of $T_{n,B}(\beta, w, \boldsymbol{R})$ is discrete the exact level α need not be attained. However, it can be reached using classical randomization, see e.g. Lehmann [5]. Therefore this procedure leads to the test with the exact level α.

Section 2 contains asymptotic theoretical results. They imply that bootstrap without replacement provides approximations for critical values that lead to consistent test. Section 3 represents results of simulation study that support application of the bootstrap without replacement.

2 Asymptotic results

Here we formulate the assertion on asymptotic distribution of $T_{n,B}(\beta, w)$ under H_0 (Theorem A) and under alternatives (Theorem B) and asymptotic conditional distribution (given $\boldsymbol{Y}$) of $T_{n,B}(\beta, w, \boldsymbol{R})$ (Theorem C). They give pictures about behavior of the proposed procedures.

We consider two sets of assumptions on the observations $Y_1, \ldots, Y_n$:

(I) $Y_1, \ldots, Y_n$ be i.i.d. random variables with common distribution function F.

(II) $Y_1, \ldots, Y_m$ are i.i.d. with d.f. F_* and $Y_{m+1}, \ldots, Y_n$ are i.i.d. r.v.'s with d.f. F^*, where all Y_i's are independent and m and F_*, F^* satisfy

$$m = m_n = [\kappa n], \qquad \text{for some} \qquad 0 < \kappa < 1 \tag{16}$$

and

$$\Delta = E(h_w(Y_1, Y_2) - 2h_w(Y_1, Y_n) + h_w(Y_n, Y_{n-1})) \neq 0. \tag{17}$$

The set of assumptions (I) corresponds to H_0 while the set (II) expresses fixed alternatives. The quantity Δ can be alternatively expressed as

$$\Delta = \int\int h_w(x,y) d(F^*(x) - F_*(x)) d(F^*(y) - F_*(y)).$$

Now, we need some additional notation. We put

$$\tilde{h}_w(x,y) = h_w(x,y) - E_F(h_w(x, Y_s) - E_F(h_w(Y_r, y) - E_F h(Y_r, Y_s),\ r \neq s, \tag{18}$$

where the expectation E_F corresponds to the model (I). Hence

$$E_F(\tilde{h}_w(Y_r, Y_s))|Y_r) = E_F(\tilde{h}_w(Y_r, Y_s))|Y_s) = E_F \tilde{h}_w(Y_r, Y_s) = 0, \quad r \neq s. \tag{19}$$

and since the function $\tilde{h}_w(x,y)$ is symmetric in its arguments and

$$E_F\, \tilde{h}_w^2(Y_1, Y_2) = \int \widetilde{h_w}^2(x,y) dF(x) dF(y) < \infty \tag{20}$$

there exist orthogonal eigenfunctions $\{\psi_j(t),\ j = 1, 2, \ldots\}$ and eigenvalues $\{\lambda_j,\ j = 1, 2, \ldots\}$ such that (see, e.g., Serfling [7])

$$\lim_{K \to \infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big(\widetilde{h}_w(x,y) - \sum_{j=1}^{K} \lambda_j \psi_j(x) \psi_j(y) \Big)^2 dF(x) dF(y) = 0, \tag{21}$$

$$\int_{-\infty}^{\infty} \psi_j(x) \psi(x) dF(x) = \delta_{ij}, \quad i, j = 1, 2, \ldots, \tag{22}$$

and

$$E_F\,\widetilde{h}_w^2(Y_1,Y_2) = \int \widetilde{h}_w^2(x,y)dF(x)dF(y) = \sum_{j=1}^{\infty}\lambda_j^2, \tag{23}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

Similarly, as in Hušková and Meintanis [4] one can show:

Theorem A. Let $Y_1,\dots,Y_n$ satisfy (I). Let the prior $q_{k,n}$ have the form (8) with $\beta < 1$ and let w be symmetric nonnegative function satisfying (10). Then the limit behavior of $T_{n,B}\beta, w)$ is the same as that of

$$\int_0^1 (t(1-t))^{-\beta}dt\Big(\int w(z)dz - E_F\,h(Y_1,Y_2)\Big) \tag{24}$$
$$+\sum_{j=1}^{\infty}\lambda_j\int_0^1 (t(1-t))^{-\beta}\Big\{\frac{B_j^2(t)}{(1-t)t} - 1\Big\}dt,$$

where $\{B_j(t),\, t\in(0,1)\}$, $j = 1,2\dots$, are independent Brownian bridges.

Since the eigenvalues $\{\lambda_j\}_j$ depend on the underlying distribution function F which is unknown, so does the limit distributions of (24) therefore the limit distributions under H_0 do not provide useful approximations for the critical values.

Next, we state the assertion on the asymptotic behavior of $T_{n,B}(\beta,w)$ under fixed alternatives. By the results in Hušková and Meintanis [4] it can be shown that

Theorem B. Let $Y_1,\dots,Y_n$ satisfy (II). Let the prior $q_{k,n}$ have the form (8) with $\beta < 1$ and let w be symmetric nonnegative function satisfying (10). Then, as $n\to\infty$,

$$T_{n,B}(\beta,w) = \int_0^1 (t(1-t))^{-\beta}\Big((\int w(z)dz - E_{F_*}h(Y_1,Y_2))+ \tag{25}$$
$$n\Delta\int_0^1 t(1-t)\Big(((\frac{1-\gamma}{1-t})^2 I\{t\le\gamma\} + (\frac{\gamma}{t})^2 I\{t>\gamma\}\Big)dt(1+o_P(1)).$$

Towards the limit behavior of the bootstrap version of $T_{n,B}(\beta,w)$ we set

$$F_\kappa(x) = \kappa F_*(x) + (1-\kappa)F^*(x), \quad x\in R^1,$$

and $\{\lambda_j(\kappa)\}_j$ be a sequence of eigenvalues corresponding to the function

$$\widetilde{h}_w(x,y;\kappa) = h_w(x,y) - \int h_w(,y)dF_\kappa(t)$$
$$-\int h_w(x,z)dF_\kappa(z) + \int\int h_w(t,z)dF_\kappa(z)dF_\kappa(t).$$

Notice that

$$P(Y_{R_i} \le x) = \frac{m}{n} F_*(x) + \frac{n-m}{n} F^*(x) = F_\kappa(x) + O(n^{-1}), \quad x \in R^1.$$

Theorem C. Let $Y_1, \dots, Y_n$ satisfy (II). Let the prior $q_{k,n}$ has the form (8) with $\beta < 1$ and let w be symmetric nonnegative function satisfying (10). Then for all x, as $n \to \infty$,

$$P(T_{n,B}(\beta, w, \boldsymbol{R}) \le x | \boldsymbol{Y}) - P\Big(\max_{1 \le k < n} \Big(\frac{k(n-k)}{n^2} \Big)^\gamma \Big| \Big(\int w(t)dt - E_\kappa\, h(Z_1, Z_2) \Big) \tag{26}$$

$$+ \sum_{j=1}^{\infty} \lambda_j(\kappa) \left\{ \frac{B_j^2(k/n)}{(1-k/n)k/n} - 1 \right\} \Big| \le x \Big) \to^P 0, \quad x \in R^1,$$

where where Z_1, Z_2 are independent random variables with distribution function F_κ.

Combining Theorems A, B and C we can infer that
(i) under H_0 (I)

$$T_{n,B}(\beta, w) = O_P(1),$$

(ii) under the considered alternatives (II) as $n \to \infty$,

$$T_{n,B}(\beta, w) \to^P \infty,$$

(iii) under both H_0 (I) and the considered alternatives (II)

$$P(T_{n,B}(\beta, w, \boldsymbol{R}) = O_P(1).$$

Therefore the test with the critical region (15) is consistent.

3 Simulation results

For sample size $n = 100$, the behavior of the test statistics is assessed via a simulation experiment with 1000 replications. For each replication, the test statistic, say T_Y, is calculated based on the original sample $Y_1, \dots, Y_n$. Then 200 permutations $(R_1^{(j)}, \dots, R_n^{(j)})$, $j = 1, 2, \dots, 200$, of $(1, 2, \dots, n)$ are chosen at random from all $n!$ total number of permutations. For each $j = 1, 2, \dots, 200$, the test statistic, say T_j^*, is computed based on $Y_{R_1^{(j)}}, \dots, Y_{R_n^{(j)}}$. Let $T_{(j)}^*$, denote the j^{th} order statistic of T_j^*, $j = 1, 2, \dots, 200$. Then $T_{(190)}^*$ (resp. $T_{(180)}^*$) is defined as the 5% (resp. 10%) critical point, and H_0 is rejected at 5% (resp. 10%) nominal level when $T_Y > T_{(190)}^*$ (resp. $T_Y > T_{(180)}^*$).

In Table 1, power results for $a = 1$ are shown (percentage of rejection rounded to the nearest integer). These results correspond to the test statistics

	$\beta = 0$				$\beta = 0.5$			
	$m = n/2$		$m = n/4$		$m = n/2$		$m = n/4$	
$F_1 - F_n$	5%	10%	5%	10%	5%	10%	5%	10%
$W_{1.0} - W_{1.5}$	23	36	16	28	21	33	18	28
	17	31	15	28	14	27	16	26
$W_{1.0} - W_{2.0}$	64	77	49	64	57	72	48	33
	44	61	39	54	36	54	38	54
$W_{2.0} - W_{3.0}$	20	33	16	28	18	30	17	29
	9	16	9	17	8	16	10	18
$t_3 - t_3 + 1$	87	92	71	79	84	90	70	78
	91	96	73	83	88	93	71	82
$t_4 - t_4 + 1$	88	94	71	81	86	91	70	80
	93	95	77	87	90	94	75	85
$t_5 - t_5 + 1$	91	95	75	84	89	94	75	84
	94	96	80	89	92	95	78	89

Table 1: Rejection rate under H_1 corresponding to 5% and 10% nominal level for the test statistics $T^{(1)}_{n,B}(\beta, a)$ (top entry) and $T^{(2)}_{n,B}(\beta, a)$ (bottom entry) for $a = 1$.

$T^{(1)}_{n,B}(\beta, a)$ (top entry) and $T^{(2)}_{n,B}(\beta, a)$ (bottom entry), for $\beta = 0$ and $\beta = 0.5$. The distributions considered are Weibull (W_φ) distributions with shape parameter φ, and Student's–t (t_ν) distributions with ν degrees of freedom. In the Weibull distributions case, shape changes are detected, whereas in the case of t_ν distributions, H_1 corresponds to simple shifts in location (in Table 1, Student–t_ν distributions with unit location shifts are denoted by $t_\nu + 1$). The results indicate that the tests based on $T^{(1)}_{n,B}(\beta, 1)$ and $T^{(2)}_{n,B}(\beta, 1)$ are considerably powerful under most of the alternative situations considered. More extensive results of simulations including comparisons will be presented on COMPSTAT Meeting.

References

[1] Antoch J., Hušková M., Jarušková D. (2001). *Off-line quality control.* In: Multivariate Total Quality Control: Foundation and Recent Advances, Lauro N. C. et al. eds., Springer-Verlag, Heidelberg, 1 – 86.

[2] Brodsky B.E., Darkhovsky B.S. (2000). *Non-parametric statistical diagnosis* Kluwer Academic Publishers, Dordrecht.

[3] Csörgő M., Horváth L. (1997). *Limit theorems in change-point analysis.* J. Wiley, New York.

[4] Hušková M., Meintanis S. (2004). *Change point analysis based on empirical characteristic functions.* Submitted for publication.

[5] Lehmann E.L. (1991). *Testing statistical hypotheses.* Wadworth & Brooks/Cole, California.

[6] Meintanis S. (2003). *Permutation tests for homogeneity based on the empirical characteristic function.* Submitted for publication.

[7] Serfling R. (1980). *Approximation theorems of mathematical statistics.* J. Wiley, New York.

Acknowledgement: The work was of the first author supported by grants GAČR 201/03/0945 and MSM 113200008.

Address: Marie Hušková, Charles University of Prague, Department of Statistics, Sokolovská 83, CZ – 186 75 Praha 8, Czech Republic;
Simos G. Meintanis, National and Kapodistrian University of Athens, Department of Economics, 8 Pesmazoglou Street, 105 59 Athens, Greece;

E-mail: `marie.huskova@karlin.mff.cuni.cz, simosmei@econ.uoa.gr`

DEVELOPMENT OF THE EDUCATIONAL MATERIALS FOR STATISTICS USING WEB

Masaya Iizuka, Tomoyuki Tarumi, Kikuo Yanagi, Kaoru Fueda and Tomokazu Fujino

Key words: Statistical education, simulation, multimedia, contents.

COMPSTAT 2004 section: Teaching statistics.

Abstract: In recent times only a small number of students have studied statistics before entering universities. It is therefore important to create a new course for statistics in the compulsory school levels and senior high school. We have developed an educational material "CASE" (Computer Assisted Statistical Education) to teach statistics. CASE consists of animation, video and teaching materials with simulations. In order to teach the concept of statistics, it is desirable to show students statistics interactively. To that end, we decided to introduce simulation methods allowing an opportunity for students to experience a sort of real experiment for statistics. That is why we included simulation methods in these teaching materials. We have also introduced animation into these teaching materials to increase their appeal and effectiveness. In this paper, we explain our educational materials, CASE.

1 Introduction

First, let us touch upon the present situation in terms of statistics education in Japan. When we go through the Government Guidelines for Teaching in terms of statistics in Japan, we find that many classes for handling statistics are included in the Guidelines. For example, many classes to teach statistical graphs and means are recommended to be taught in elementary school levels and probability and statistics in high school, though no units for statistics are allocated in the curriculums for junior high school. In spite of allotting many classes for teaching probability and statistics in the Government Guidelines for Teaching, statistics is not presently a subject for entrance-examinations for universities. That is, there are only a small number of students would learn statistics in high school levels. Since teachers are given a few opportunities to teach statistics at school, it is important for teachers to obtain information on statistics as well as receive appropriate teaching materials for statistics.

In order to teach the concepts of statistics to the students or children, we must device the teaching methods can be used for all levels of curriculums from elementary schools to university's introductory courses of statistics. But, it is rather difficult to teach children/students the concepts of statistics only through textbooks.

We realized the need to develop educational methods using interactive software. IT is a very useful tool for students to study the concepts of statistics. There are two advantages to using this interactive software: first, from the aspect of learning effect, it is more attractive for students to use educational materials interactively rather than learning only through texts and formulas, and, secondly, it can be possible for students to experiment with statistical problems on a computer. Experiments are necessary for statistical education. However, it is difficult to perform statistical experiment in the class.

In our project, we regard "simulation" as a statistical experiment. Treating simulation as statistical experiment can help students better understand statistical methods. To this end, we have developed statistical educational software with simulation called "CASE" (Computer Assisted Statistical Education) which contains a variety of teaching materials for statistics. In the past, we developed suitable tools for teaching purposes, which have been used in various situations in statistical education [1], [2].

2 Teaching materials

There are three kinds of teaching materials in CASE. One is a library of teaching tools for statistics; these libraries consist of some interactive software or macros. We developed these software or macros by using some programming language including JAVA, macromedia Flash MX, R and LISP-STAT. By using these libraries, teachers can teach or demonstrate one of the experiments to support the explanation in the class.

The second is applications software for statistics. These applications are of two types: standalone application and web-based application. The third is a content which includes some teaching tools and animations. We describe these materials in the next section.

2.1 Library

We already have developed some teaching tools for statistics. The library consists of the following teaching tools;

a. Graph calculator
b. Interactive applications for linear function and quadratic function
c. Explanation of statistical graph
d. Statistical calculator for uni-variate data (the mean, the variance and the median)
e. Explanation of frequency table and histogram
f. Simulation for the circular constant
   (a) by uniform random number
   (b) by Buffon Needle

g. Probability distribution

   (a) Relationship between a probability density function (p.d.f.) and a cumulative distribution function (c.d.f.)
   (b) p.d.f. of $N(\mu, \sigma^2)$

h. Simulation for statistical estimation

i. Simulation for statistical test

   (a) significant level
   (b) power of test

j. Box-Cox power transformation

k. Demonstration of bootstrap

l. Sensitivity in simple linear regression

m. Relationship between a scattergram and a correlation coefficient

n. Constellation graph

o. Face chart

These teaching tools with simulations are useful as a means of statistical experiments because we consider that experiments are an important way to understand statistical method. By using these interactive tools, students can experience simulation as if they were participating in real experiments. From the aspect of an educational effect, students will learn statistical concepts easier than using only textbooks.

Suppose we have two different learning cases: One is that teachers demonstrate these teaching tools with explanation to teach students units of statistics in the class, and students can learn the concept of these units of statistics with interest. Given those interactive teaching tools, after school they can review what they learned on that day by using them at home. Another case is that students perform statistical experiments themselves in the class. Even those who do not understand the concept of statistics when the explanation is given by teachers in the class may understand what it means after they do experiment by themselves.

For teachers who do not know the methods of statistics teaching, we will provide the explanation how to use these teaching tools on our Web site, so they can access such information easily.

Moreover, we have created multimedia application tools for educational use which was programmed by Macromedia Flash MX using ActionScript. Using this, we developed teaching materials such as “the mean” designed for elementary school levels, “the relationship between a probability density function (p.d.f.) and a cumulative distribution function (c.d.f.)” and so on.

In addition, we developed the library of statistical analysis including constellation graphs and face charts, and also provided some data and a glossary for statistics. The following details our teaching tools categorized by programming techniques.

2.1.1 Teaching tools programmed by Java

Java is a very useful programming technique to develop applications for our purpose. The teaching tools we have developed can be provided as a part of the Web (what is called Java applet). If students or teachers have the web browser which can perform Java in their computer, then our teaching tools can be performed without choosing any platform of the computer. Furthermore we can provide those teaching tools as Java applications.

Figure 1 shows a Java applet of "relationship between a scattergram and a correlation coefficient". This data was automatically generated by random numbers, so generated data is different every time. On loading this applet, this tool generates a data which has correlation coefficient 0. We can see a correlation coefficient of this data at left-hand side in the bottom of the applet. In addition to that, a scattergram is interactively changed by moving the slider. This slider points out the value of correlation coefficient of this changed data. Then, students can understand the relationship between a correlation coefficient and a scatter gram visually.

Furthermore, students can click the refresh button; this tool generates new dataset which has correlation coefficient of zero.

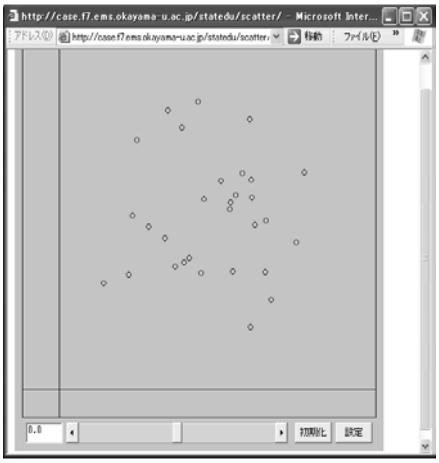

http://case.f7.ems.okayama-u.ac.jp/statedu/scatter/ - Microsoft Inter...
アドレス(D)
http://case.f7.ems.okayama-u.ac.jp/statedu/scatter/
移動
ファイル(F)
0.0
初期化
設定

Figure 1: Relationship between scattergram and correlation coefficient.

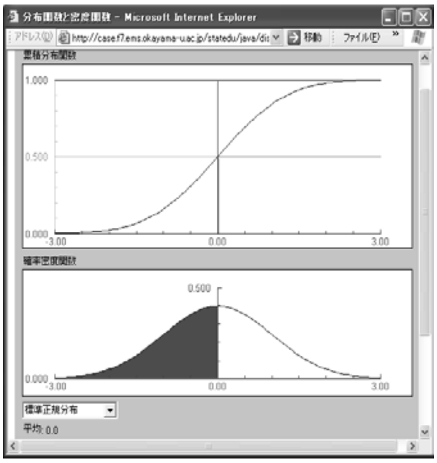
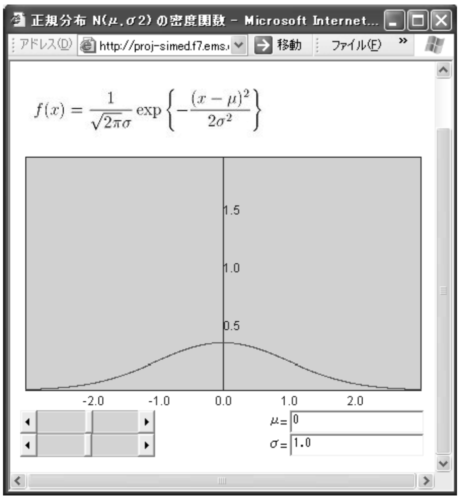


Figure 2: Relationship between p.d.f. and c.d.f.

Figure 2 is a screen shot as a teaching tool for learning of relation between c.d.f. and p.d.f.. We can specify some distributions at the left-hand side on the bottom of the applet, including standard normal distribution, χ^2 distribution, F distribution, t distribution, beta distribution, exponential distribution, uniform distribution, triangular distribution and trapezoid distribution. By using this applet, students will comprehend that the length of line of c.d.f. in the graph of c.d.f. equals to the area framed by p.d.f.

and x-axis in the graph of p.d.f.. It is important for students to learn this relationship.

Figure 3 is an applet which displays the p.d.f. of a normal distribution. This can be used for understanding the shape of the p.d.f. in normal distribution and the meaning of parameters. Students can learn about the following characteristics by moving two sliders below in this applet. When μ (the mean) changes, the graph of p.d.f. moves in parallel. Students learn that μ means a location parameter. In addition to that, when σ changes, the graph of p.d.f. are widen. Students learn that parameter σ means a spread parameter.


正規分布 N(μ,σ2) の密度関数 - Microsoft Internet...
アドレス(D)
http://proj-simed.f7.ems.
移動
ファイル(F)
1.5
1.0
0.5
-2.0
-1.0
0.0
1.0
2.0

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$\mu=$ 0

$\sigma=$ 1.0

Figure 3: p.d.f. of normal distibution.

2.1.2 Teaching tools programmed by Flash MX

Macromedia Flash MX[1] is a useful software to develop a teaching tool requiring a small amount of data, and it can program the software with an animation as interactive teaching materials. To operate this tool, a player is provided an add-in for browsers and various operating systems. Moreover, a stand-alone player is also provided.

Figure 4 shows one of the teaching materials for students to learn "the mean". This tool aims at telling students that the meaning of "the mean" is to 'flatten' the data. It is targeted at elementary school children. To attract students interesting, we have made this program as closer as possible to a game.

These are our teaching tools. As stated above, such teaching tools are

[1] http://www.macromedia.com/

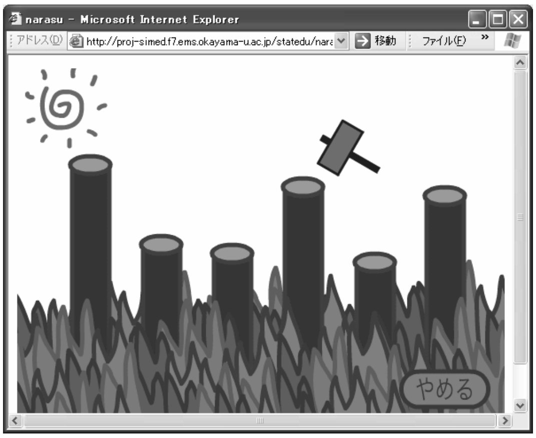

narasu - Microsoft Internet Explorer
アドレス(D)
http://proj-simed.f7.ems.okayama-u.ac.jp/statedu/nara
移動
ファイル(F)
やめる

Figure 4: Animation (for elementary school).

open to the public now. We will also provide these teaching tools through Java, as well as Flash MX, R, and Lisp-Stat functions.

2.2 Glossary of statistics

We make the interface of the "Glossary of Statistics" in the style of a dictionary. The layout of the "Glossary of Statistics" is such: an index term is placed in left-hand side, and description of the term is placed in right-hand side. We plan to link this Glossary of Statistics with Java applet and Flash MX materials so that students can study synthetically. Since these statistical terms are difficult for students, we recommend this "Glossary of Statistics" for tertiary students only.

3 Contents

The teaching tools introduced above are mainly used by way of explanation. Here, we introduce the "Contents" section, which includes animation, video and some of our teaching tools to be used in each lesson. For example, one lesson lasts 50 minutes: we allocate 15 minutes for that "contents", another 15 minutes for explanation of this unit and the last 15 minutes for a conclusion part. In the first 15 minutes, a teacher shows a "contents" page which covers that day's unit. For these contents we have prepared a simple animation story dealing with the unit for the day. By watching this "contents" and performing a simulation using teaching tools in these contents, students learn about the abstract of that day's lesson.

We are planning to develop the following contents for this year.

a. Graph contents (target of this content is schoolchildren or older)

Students will develop their knowledge of graphs including how to draw a graph, and how to understand what the graph shows in the classes of social studies and science. They will also learn the essential meaning of the mean by using the above teaching materials.

b. Probability contents (targeted junior high school students or older)
This content is created for students to learn about the concept of probability, and the difference between probability and the figures can be observed periodically.

c. Arrangement of data (target of this contents is high school students or older)
Students will learn how to read the data of the frequency table and how to draw a histogram from the frequency table. They will also learn about the relationship between two variables observed from the scattergram using the correlation coefficient.

3.1 Examples of contents

This is an example of contents dealing with probability. This content is designed for elementary school children. Figure 5 is a part of animation for the probability contents. At first this animation introduces the probability of precipitation which children are used to hearing about on the TV weather report. In order to keep students from being bored and to make them participate in the animation, they are asked to perform the simulation: they throw a coin in this animation. It is not possible for students to actually throw coins 100 times or more during the class hour, but they can perform it through the simulation by using these animation contents.

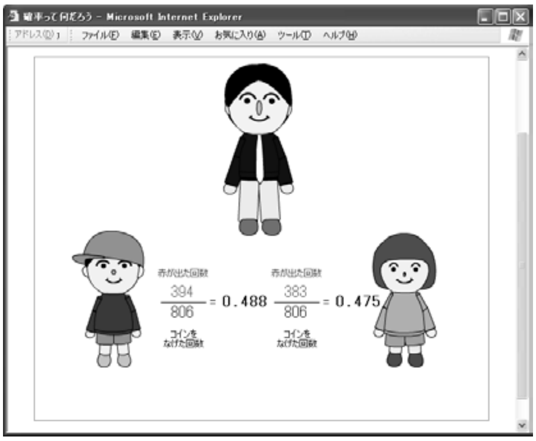


Figure 5: Probability contents.

4 Conclusion

We have contracted out the animation work of our materials. Now, animations for several units are ready to be used. Scenarios including instruction schemes and teaching plans for more units are currently being developed. This project is now proceeding to make more content and various materials in order to cover the units for statistics from elementary and secondary education levels to the liberal arts of the university.

We have uploaded the contents and materials developed on the following server.

http://case.f7.ems.okayama-u.ac.jp/
e-mail: case@f7.ems.okayama-u.ac.jp

This project is related to Scientific Research of Priority Area "Developmental Research for Science Education in the New Millennium". Please refer to the following URL:

http://risuka.ei.tohoku.ac.jp/rsroot/

References

[1] Tarumi T., Fueda K., Iizuka M., Yanagi K., Fujino T. (2003). *Software with simulation for statistical education.* Bulletin of the International Statistical Institute (54th Session Contributed Papers Book. Provided by CD-ROM).
URL: http://case.f7.ems.okayama-u.ac.jp/report/ISI2003.pdf

[2] Yanagi K., Tarumi T., Fueda K., Iizuka M., Fujino T. (2002), *Statistical software for educaiton with simulation.* Proceedings of the 4th ARS, 228 – 229.

Acknowledgement: This research is supported by Scientific Research of Priority Areas, "Developmental Research for Science Education in the New Millennium" in a Grant-in-Aid for Scientific Research from The Japanese Ministry of Education, Science, Sports and Culture. (Subject number 14022232 and 15020240)

Address: M. Iizuka, T. Tarumi, K. Fueda, Okayama University, 3-1-1,
Tsushima-naka, Okayama, 700-8530, Japan
K. Yanagi, Okayama University of Science, 1-1, Ridaicho, Okayama,
700-0005, Japan
T. Fujino, Fukuoka Women's University, 1-1, Kasumigaoka, Fukuoka,
813-8529, Japan

E-mail: masa@law.okayama-u.ac.jp, tarumi@ems.okayama-u.ac.jp,
yan@mis.ous.ac.jp, fueda@ems.okayama-u.ac.jp,
fujino@fwu.ac.jp

COMPSTAT'2004 Symposium

ON THE DEGREES OF FREEDOM IN RICHLY PARAMETERISED MODELS

Salvatore Ingrassia and Isabella Morlini

Key words: Richly parameterised models, small data sets.

COMPSTAT 2004 section: Neural networks.

Abstract: Using richly parameterised models for small datasets can be justified from a theoretical point of view according to some results due to Bartlett [1] which show that the generalization performance of a multi layer perceptron (MLP) depends more on the L_1 norm $\|\mathbf{c}\|_1$ of the weights between the hidden and the output layer rather than on the number of parameters in the model. In this paper we investigate the problem of measuring the generalization performance and the complexity of richly parameterised procedures and, drawing on linear model theory, we propose a different notion of degrees of freedom to neural networks and other projection tools. This notion is compatible with similar ideas long associated with smoothers based models (like projection pursuit regression) and can be interpreted using the projection theory of linear models and showing some geometrical properties of neural networks. Results in this study lead to corrections in some goodness-of-fit statistics like AIC, BIC/SBC: the number of degrees of freedom in these indexes are set equal to the dimension p of the projection space intrinsically found by the mapping function. An empirical study is presented in order to illustrate the behavior of the L_1 norm $\|\mathbf{c}\|_1$ and of the values of some selection model criteria, varying the value of p in a MLP.

1 Introduction

An important issue in statistical modeling is related to so called indirect measures or *virtual sensors*, concerning the prediction of variables that are quite expensive to measure (e.g. the viscosity or the concentration of certain chemical species, some mechanical features) using other variables like, for example, the temperature or the pressure. This problem usually involves some difficulties: the available data sets is small and the input-output relation to be estimated is non-linear. A third difficulty arises when there are many predictor variables but, since the linearity cannot be assumed, it is quite difficult to reduce the dimensionality of the input by choosing a good subset of predictors or suitable underlying features. When we are provided with a data set of N pairs $(\mathbf{x}_n, y_n)$ of m-dimensional input vectors $\mathbf{x}_n$ and scalar target values y_n and the size N of the available sample is small compared with the number of weights of the mapping function, the model is considered over-parameterised. Over parameterised models can be justified from a theoretical point of view according to some results due to Bartlett [1] showing that the generalization performance of an MLP depends more on the size of the weights rather than on the number of weights and, in particular, on the L_1 norm $\|\mathbf{c}\|_1$ of the weights between the hidden and the output layer. Bartlett's results suggest a deeper look at the roles of the parameters in a neural network and in similar richly parameterised models. The roles of these parameters are here interpreted drawing from the projection theory of linear models and by means of some geometrical properties shared by neural networks and statistical tools realizing similar mapping functions.

2 Geometrical properties of the sigmoidal functions

In this section we investigate some geometrical properties of a mapping function of the form:

$$f_p(\mathbf{x}) \quad = \quad \sum_{k=1}^{p} c_k \tau(\mathbf{a}_k' \mathbf{x}) \ . \tag{1}$$

Without loss of generality, we assume that the bias term is equal to zero and that the function $\tau(\cdot)$ is sigmoidal and analytic, that is, it can be represented by a power series, on some interval $(-r, r)$, where r may be $+\infty$. The hyperbolic tangent $\tau(z) = \tanh(z)$ or the logistic function $\tau(z) = (1 + \mathrm{e}^{-z})^{-1}$ are examples of analytic sigmoidal functions. We point out that the function f_p is a combination of certain transformations of the input data. In particular, f_p realizes: *i*) a non-linear projection from $\mathbb{R}^m$ to $\mathbb{R}^p$ given by the sigmoidal function τ, that is $\mathbf{x} \to \tau(\mathbf{a}_1'\mathbf{x}), \ldots, \tau(\mathbf{a}_p'\mathbf{x})$; *ii*) a linear transformation from $\mathbb{R}^p$ to $\mathbb{R}$ according to $c_1, \ldots, c_p$. The results in this paper are based on the following theorem, see e.g. Rudin (1966):

Theorem 2.1. Let g be analytic and not identically zero in the interval $(-r, r)$, with $r > 0$. Then the set of the zeroes of g in $(-r, r)$ is at most countable.

Let $\mathbf{x}_1 = (x_{11}, \ldots, x_{1m}), \ldots, \mathbf{x}_p = (x_{p1}, \ldots, x_{pm})$ be p points of $\mathbb{R}^m$, with $p > m$; evidently these points are linearly dependent as $p > m$. Let $\mathbf{A} = (a_{ij})$ be a $p \times m$ matrix with values in some hypercube $[-u, u]^{mp}$, for some $u > 0$; thus the points $\mathbf{A}\mathbf{x}_1, \ldots, \mathbf{A}\mathbf{x}_p$ are linearly dependent because they are obtained by a linear transformation acting on $\mathbf{x}_1, \ldots, \mathbf{x}_p$. For $u = 1/m$ the points $\tau(\mathbf{A}\mathbf{x}_1), \ldots, \tau(\mathbf{A}\mathbf{x}_p)$, where:

$$\begin{aligned} \tau(\mathbf{A}\mathbf{x}_i) \quad &= \quad \left(\tau(\sum_{j=1}^{m} a_{1j} x_{ij}), \ldots, \tau(\sum_{j=1}^{m} a_{pj} x_{ij}) \right) \\ &= \quad \big(\tau(\mathbf{a}_1'\mathbf{x}_i), \ldots, \tau(\mathbf{a}_p'\mathbf{x}_i)\big) \qquad\quad i = 1, \ldots, p \ . \end{aligned}$$

are linearly independent for almost all matrices $\mathbf{A} \in [-u, u]^{mp}$, according to the following theorem.

Theorem 2.2. *[5]*Let $\mathbf{x}_1, \ldots, \mathbf{x}_p$ be p distinct points in $(-r, r)^m$ with $\mathbf{x}_h \neq \mathbf{0}$ $(h = 1, \ldots, p)$ and $\mathbf{A} = (a_{ij}) \in [-u, u]^{mp}$ be a $p \times m$ matrix, with $u = 1/m$. Let τ be a sigmoidal analytic function on $(-r, r)$, with $r > 0$. Then the points $\tau(\mathbf{A}\mathbf{x}_1), \ldots, \tau(\mathbf{A}\mathbf{x}_p) \in \mathbb{R}^p$ are linearly independent for almost all matrix $\mathbf{A} = (a_{ij}) \in [-u, u]^{mp}$.

This result proves that, given $N > m$ points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$, the transformed points $\tau(\mathbf{A}\mathbf{x}_1), \ldots, \tau(\mathbf{A}\mathbf{x}_N)$ generate an *over-space* of dimension $p > m$ if the matrix $\mathbf{A}$ satisfies suitable conditions. In particular, the largest over-space is attained when $p = N$, that is when in a MLP the hidden layer has as many units as the number of points in the learning set. Moreover, it gains insight why neural networks have been proved to work well in presence of multicollinearity. On this topic De Veaux & Ungar [3] present a case-study in which the temperature of a flow is measured by six different devices at various places in a production process: even though the inputs are highly correlated, a better prediction of the response is gained using a weighted combination of all six predictors rather than choosing the single best measurement having the highest correlation with the response.

Next result generalizes Theorem 2.

Theorem 2.3. Let $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ be a given learning set of N i.i.d. realizations of $(\mathbf{X}, Y)$ and $f_p = \sum_{k=1}^{p} c_k \tau(\mathbf{a}'_k \mathbf{x})$. If $p = N$, then the learning error $\widehat{\mathcal{E}}(f_p, \mathcal{L}) = \sum_{(x_n, y_n) \in \mathcal{L}} (y_n - f_p(\mathbf{x}_n))^2$ is zero for almost all matrices $\mathbf{A} \in [-1/m, 1/m]^{mp}$.

Proof. Theorem 2 implies that the points $\tau(\mathbf{A}\mathbf{x}_1), \ldots, \tau(\mathbf{A}\mathbf{x}_N)$ are linearly independent for almost all matrices $\mathbf{A} \in [-1/m, 1/m]$ for $p \geq N$. In particular, if $p = N$ the system:

$$\begin{array}{lcccccl} c_1\tau(\mathbf{a}'_1\mathbf{x}_1) & + & \cdots & + & c_N\tau(\mathbf{a}'_N\mathbf{x}_1) & = & y_1 \\ \cdots & + & \cdots & + & \cdots & = & \vdots \\ c_1\tau(\mathbf{a}'_1\mathbf{x}_N) & + & \cdots & + & c_N\tau(\mathbf{a}'_N\mathbf{x}_N) & = & y_N \end{array} \tag{2}$$

has a unique solution. ■

The upper bound on p given above looks too large but it refers to the worst case. In neural modelling, given a learning set $\mathcal{L}$ of N sample data, the good question seems not to be "what is the largest network we can train by $\mathcal{L}$ (if any)?" but "what is a suitable size – namely, the dimension p of the space $\mathbb{R}^p$ – necessary for fitting the input-output unknown dependence $\phi = \mathbb{E}[Y|\mathbf{X}]$?". This dimension p depends on the geometry of the data and this explains why neural models may be successfully applied as virtual sensors when the predictors exhibit a high degree of multicollinearity. As a matter of fact, the hidden units break the multicollinearity and exploit the contribution of each single predictor. This is the reason why the optimal size p of the hidden layer is often greater than the number m of predictors.

3 On counting the degrees of freedom in linear projection models

Consider the standard regression model:

$$\mathbf{y} = \bar{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{y}$ is a vector of N observations of the dependent variable measured about its mean, $\bar{\mathbf{X}}$ is an $N \times m$ matrix whose (i, j)th element is the value of the jth predictor variable for the ith observation, again measured about its mean, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\boldsymbol{\varepsilon}$ is the vector of the error terms satisfying the usual assumption of independence and homoscedasticity. The values of the principal components (PC) for each observation are given by:

$$\mathbf{Z} = \bar{\mathbf{X}}\mathbf{A}$$

where the (i, k)th element of $\mathbf{Z}$ is the value of the kth PC for the ith observation and $\mathbf{A}$ is the $m \times m$ matrix whose kth column is the kth eigenvector of $\bar{\mathbf{X}}'\bar{\mathbf{X}}$. Because $\mathbf{A}$ is orthogonal, $\bar{\mathbf{X}}\boldsymbol{\beta}$ can be rewritten as $\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = \mathbf{A}'\boldsymbol{\beta}$. Equation (3) can therefore be written as:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{4}$$

which has simply replaced the predictor variables by theirs PCs in the regression model. Principal component regression can be defined as the model (4) or as the reduced model:

$$\mathbf{y} = \mathbf{Z}_p\boldsymbol{\gamma}_p + \boldsymbol{\varepsilon}_p \tag{5}$$

where $\boldsymbol{\gamma}_p$ is a vector of p elements which are a subset of $\boldsymbol{\gamma}$, $\mathbf{Z}_p$ is an $N \times p$ matrix whose columns are the corresponding subset of columns of $\mathbf{Z}$ and $\boldsymbol{\varepsilon}_p$ is the appropriate error term, see Section 8.1 in Jolliffe (1986). From an algebraic point of view, computing the PCs of the original predictor variables is a way to overcome the problem of multicollinearity between them, since the PCs are orthogonal. ¿From a geometrical point of view, if we compute the values of the PCs for each of the N observations, we project the N points in a m-dimensional hyperplane for which the sum of squares of perpendicular distances of the observations in the original space is minimized. The total number of estimated quantities in model (4), which results from defining the projection of the centered matrix $\bar{\mathbf{X}}$ in the space spanned by the m PCs and then estimating the m regression coefficients plus the bias term, is larger than the number m of variables involved in the model. In any case, model (5) is given $p+1$ degrees of freedom, that is, the number of retained PCs plus one. The function given by the PC regression model can be rewritten as:

$$f(\bar{\mathbf{x}}_n) \quad = \quad \sum_{k=0}^{m} \beta_k(\mathbf{a}_k'\bar{\mathbf{x}}_n) \tag{6}$$

where $\bar{\mathbf{x}}_n$ is the m dimensional vector of the original centered variables, the $\mathbf{a}'_k$ with $(k = 1, ..., p)$ are the m normalized eigenvectors of the matrix $\bar{\mathbf{X}}'\bar{\mathbf{X}}$ and β_k are the regression coefficients. The function in (6) is formally identical to the mapping function realized by the projection pursuit regression or by a network model with the identity function in the second layer. This analogy between the PC regression and the mapping function realized by an MLP with p hidden nodes $(m \leq p \leq N)$ and a linear transfer function in the second level, also arises in accordance with Theorem 4 in previous Section. As in PC regression, the first transformation – from the input layer to the hidden layer – is a geometrical one which projects the point into a new space of dimension $p \leq N$. This transformation is non-linear and its optimality is not well established: the important point here is that in the new space the points are linearly independent (e.g. Theorem 2). A difference between the PC regression and the mapping function realized by the MLP is, beside the type of the geometrical projection (linear in the first case, non-linear in the second case) and the maximum dimension of this space (equal to m for PCs and in the interval $[m, N]$ for the neural network), is that the estimates of the projection matrix and the regression parameters are faced in a two stage procedure for the PC regression while simultaneously for the network.

In neural network modeling the “dimension” of the model, i.e. the number p of neurons in the hidden, is often chosen according to some model selection criteria listed in the summary reports of many statistical software for data mining. The linkage between the generalization error and the empirical error, used in these selection model criteria, has been approached in Ingrassia and Morlini [6] on the basis on the *Vapnik-Chervonenkis* theory [10]. The seminal work on model selection is based on the parametric statistics literature and is quite vast but it must be noted that, although model selection techniques for parametric models have been widely used in the past 30 years, surprisingly little work has been done on the application of these techniques in a semi-parametric or non-parametric context. Such goodness-of-fit statistics are quite simple to compute and even if the underlying theory does not hold for neural networks, the rule among users is to consider them as crude estimates of the generalization error and thus to apply these methods to very complex models, regardless of their parametric framework. These criteria, here denoted by Π, are an extension of the maximum likelihood and have the following form:

$$\Pi \quad = \quad \widehat{\mathcal{E}}(f_K) + \mathcal{C}_K \tag{7}$$

where the term $\widehat{\mathcal{E}}(f_K)$ is the deviance of the model f_K and $\mathcal{C}_K$ is a complexity term representing a penalty which grows as the number K of degrees of freedom in the model increases: if the model f_K is too simple it will give a large value for the criterion because the residual training error is large; while a model f_K which is too complex will have a large value for the criterion because the complexity term is large. Typical indexes include the *Akaike Information Criterion* (AIC), the Schwarz Bayesian Information Criterion (BIC or SBC), the *Final Prediction Error* (FPE), the *Generalized Cross Validation Error* (GCV) and the well-established *Unbiased Estimate of the Variance* (UEV) (for a review of these indexes we refer to Ingrassia and Morlini [6]. The classical statistical parametric viewpoint in which the dimensionality of the model complexity is given by $K = W$, i.e. the number W of all parameters defining the mapping function, does not seem to apply to flexible non-parametric or semi-parametric models, in which the adaptive parameters are not on the same level and have different interpretations, as we have seen in previous sections. The assumption that the degrees of freedom of a neural model should be different than W has been remarked by many authors, see e.g. Hodges and Sargent [4], Ye [11]. In the present study we propose an easy correction to the selection model criteria. According to the analogy with the PC regression, for models of the form $f_p(\mathbf{x}) = \sum_{k=1}^{p} c_k \tau(\mathbf{a}'_k \mathbf{x} + b_k) + c_0$ we should consider $K = p + 1$ rather than $K = W = p(m + 2) + 1$, that is the dimension of the projection space plus one. In this case both the FPE and the UEV are never negative (as it may happen when $K = W = p(m+2)+1$).

4 A case study

In this section we present some numerical results in order to investigate the behavior of Bartlett's constant $\|\mathbf{c}\|_1$ and its relation with the learning error $\widehat{\mathcal{E}}(f_p, \mathcal{L})$ and the test error $\widehat{\mathcal{E}}(f_p, \mathcal{T})$. We consider the *polymer data set* modeled by De Veaux et al. [2] by means of a MLP with 18 hidden units. This dataset contains 61 observations with 10 predictors concerning measurements of controlled variables in a polymer process plant an a response concerning the output of the plant (data are from ftp.cis.upenn.edu in pub/ungar/chemdata). The data exhibit a quite large degree of multicollinearity, as shown by the variance inflation factor (VIF) of some predictors:

	X_1	X_3	X_4	X_5	X_6	X_7	X_8	X_9
VIF:	2.82	25.03	100.25	49.73	95.90	57.99	1.65	3.75

In general, VIFs larger than 10 imply serious computational problems for many statistical tools [8]. As in De Veaux et al. [2], we use 50 observations for the learning set and 11 for the test set. Here, however, we consider 100 different samples with different observations for the training and the test sets. We train networks with increasing numbers of hidden units from $p = 2$ to $p = 25$. For each p we train 1000 times the network varying the sample and the initial weights; we adopt either the weight decay or the early stopping regularization techniques. Then, we retained the 100 networks with the smallest test errors. The distribution of the learning error vs. the number p is plotted in Figure 1 a) using boxplots. In Figure 1 b) we plot the distribution of $\|\mathbf{c}\|_1$ vs. p using weight decay. Similar distributions are

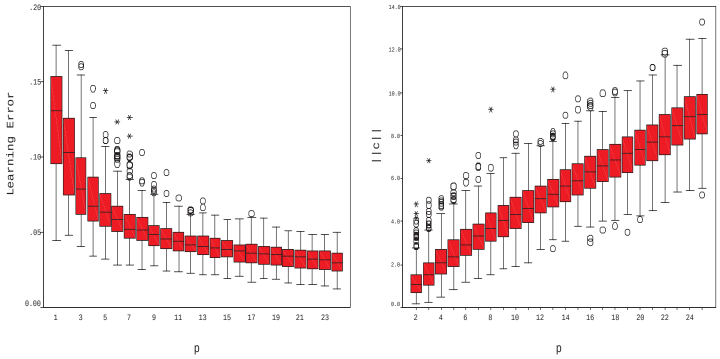


Figure 1: *Polymer data: distribution of the learning error and of* $\|\mathbf{c}\|_1$ *vs* p.

obtained with early stopping. Figure 1b) shows that there is a quite large number of models with different architectures (i.e. a different number p) having the same value of the constant $\|\mathbf{c}\|$ introduced in Bartlett [1]. In Table 1 for each group of the 100 best models are reported the mean values of the training error $\widehat{\mathcal{E}}(f_p;\mathcal{L})$, the test error $\widehat{\mathcal{E}}(f_p;\mathcal{T})$, the L_1 norm $\|\mathbf{c}\|_1$ and of the error complexity measures AIC, BIC/SBC, GCV and FPE computed with $K = p + 1$. The values p_* leading to the smallest mean values are

	AIC	BIC	GCV	FPE
p_*	9, 11	3, 5, 7	5, 7	7, 9, 11

and thus, on the basis of these statistics, models with $p_* = 7, 9, 11$ neurons in the hidden layer should be selected. In addition, we note that $p_* = 10$ is the value with the smallest absolute difference between the means of training and the test errors. This study confirms that $\|\mathbf{c}\|$ better describes the complexity of a model of the form (1) than the total number of parameters and is a more suitable characterization of the mapping function.

References

[1] Bartlett P.L. (1998). *The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network.* IEEE Transaction on Information Theory, **44**, n. 2, 525 – 536.

[2] De Veaux R.D., Schumi J., Schweinsberg J., Ungar L.H. (1998). *Prediction intervals for neural networks via nonlinear regression.* Technometrics **40**, (4), 273 – 282.

[3] De Veaux R.D., Ungar L.H. (1994). *Multicollinearity: a tale of two non parametric regressions.* In "Selecting Models from Data: AI and Statistics IV, (Eds. P. Cheeseman & R.W. Oldford)".

[4] Hodges J., Sargent D. (2001). *Counting degrees of freedom in hierarchical and other richly parameterised models.* Biometrika **88**, 367 – 379.

p	$\widehat{\mathcal{E}}(f_p;\mathcal{L})$	$\widehat{\mathcal{E}}(f_p;\mathcal{T})$	$\|\mathbf{c}\|_1$	K	BIC	AIC	FPE	GCV
2	0.0540	0.0275	1.513	3	1.2280	1.1133	3.0447	3.0557
3	0.0381	0.0273	1.982	4	0.9570	0.8040	2.2352	2.2496
4	0.0384	0.0286	2.535	5	1.0426	0.8514	2.3444	2.3681
5	0.0338	0.0282	2.966	6	0.9933	0.7638	2.1489	2.1803
6	0.0363	0.0294	3.467	7	1.1441	0.8764	2.4067	2.4548
7	0.0307	0.0296	3.776	8	1.0539	0.7480	2.1187	2.1743
8	0.0314	0.0313	4.517	9	1.1551	0.8109	2.2590	2.3346
9	0.0284	0.0306	4.913	10	1.1326	0.7502	2.1289	2.2176
10	0.0330	0.0323	5.569	11	1.3603	0.9397	2.5779	2.7091
11	0.0260	0.0312	6.033	12	1.2026	0.7437	2.1240	2.2538
12	0.0283	0.0351	6.420	13	1.3648	0.8676	2.4105	2.5853
13	0.0295	0.0353	7.271	14	1.4832	0.9478	2.6201	2.8430
14	0.0281	0.0362	7.353	15	1.5143	0.9407	2.6111	2.8693
15	0.0264	0.0352	8.030	16	1.5297	0.9178	2.5629	2.8553
16	0.0242	0.0367	8.605	17	1.5216	0.8715	2.4588	2.7802
17	0.0247	0.0378	9.003	18	1.6188	0.9305	2.6228	3.0133
18	0.0221	0.0375	9.537	19	1.5844	0.8579	2.4547	2.8689
19	0.0224	0.0394	9.946	20	1.6783	0.9135	2.6138	3.1117
20	0.0212	0.0379	10.671	21	1.6999	0.8969	2.5916	3.1466
21	0.0221	0.0393	11.235	22	1.8205	0.9792	2.8396	3.5213
22	0.0188	0.0382	11.614	23	1.7383	0.8588	2.5431	3.2256
23	0.0206	0.0398	12.060	24	1.9091	0.9913	2.9367	3.8159
24	0.0175	0.0413	12.594	25	1.8237	0.8677	2.6281	3.5042
25	0.0175	0.0425	12.755	26	1.9010	0.9068	2.7717	3.7989

Table 1: *Polymer data: summary statistics for different values of p.*

[5] Ingrassia S. (1999). *Geometrical aspects of discrimination by multilayer perceptrons.* Journal of Multivariate Analysis **68**, 226 – 234.

[6] Ingrassia S., Morlini I. (2002). *Neural network modeling for small data sets.* Submitted for publication.

[7] Jolliffe I.T. (1986). *Principal Component Analysis.* Springer-Verlag, N.Y.

[8] Morlini, I. (2002). *Facing multicollinearity in data mining.* Atti della XLI Riunione Scientifica della Societá Italiana di Statistica, Milano-Bicocca, 55 – 58.

[9] Rudin W. (1966). *Real and complex analysis.* Mc-Graw-Hill, New York.

[10] Vapnik V. (1998). *Statistical Learning Theory,* John Wiley & Sons, N.Y.

[11] Ye J. (1998). *On measuring and correcting the effects of data mining and model selection.* Journal of the American Statistical Association, **93**, 441, 120 – 131.

Address: S. Ingrassia, Dipartimento di Economia e Statistica, Università della Calabria, Arcavacata di Rende Italy

I. Morlini, Dipartimento di Economia, Università di Parma, Parma, Italy

E-mail: `s.ingrassia@unical.it, isabella.morlini@unipr.it`

COMPSTAT'2004 Symposium

AUTOMATIC RECOGNITION OF KEY-WORDS USING N-GRAMS

Radwan Jalam, Jean-Hugues Chauchat and Jean Dumais

Key words: n-gram, text representation, text categorization, text mining.

COMPSTAT 2004 section: Statistical data mining.

Abstract: Documentary research (e.g. bibliography for a thesis or a research project, technological monitoring, etc.) is often based on a few selected key-words. Yet, experience shows that other words may be characteristic of the topic of interest. Discovery and use of those in text searches often leads to relevant documents. The authors propose a fully automatic method to discover such key-words from a set of texts deemed characteristic of the topic of interest. The method relies on n-gram coding of the texts, identification of characteristic n-grams in a subset of the texts, and finally searching for words containing one or many of the characteristic n-grams. An example using 6709 Reuters news briefs sampled from the 10 most common classes. Discriminating one class from the others serves to simulate the topic-of-interest situation. Each discrimination yields a set of "candidate key-words". This is compared to searching of topic-related words. Using n-grams appears to be more efficient as specific roots show up in the n-grams; it is automatic and does not require prior linguistic analysis.

1 Introduction

Documentary research (e.g. bibliography for a thesis or a research project technological monitoring, etc.) is often based on a few selected key-words. Yet, experience shows that other words may be characteristic of the topic of interest. Discovery and use of those in text searches often leads to relevant documents. [8] propose a method for text analysis based on several correspondence analyses interweaved with human interventions. In this paper, the authors propose a fully automatic statistical method to discover such key-words from a set of texts deemed characteristic of the topic of interest. The method relies on n-gram coding of the texts.

Algorithm 1 Method for the selection of candidates key words

```
a. For each class $j$ find all n-grams in all texts of the learning set,
b. Create the cross table of the $N_{ij}$ occurrences of n-gram $i$ in class $j$,
c. Compute the corresponding relative frequencies $f_{ij}$: $f_{ij} = \frac{N_{ij}}{N}$
d. Compute $\chi^2_{ij}$, the contribution of cell $(ij)$ to the $\chi^2$ distance: $\chi^2_{ij} = \frac{\left(N_{ij} - \frac{N_{i.} \times N_{.j}}{N}\right)^2}{\frac{N_{i.} \times N_{.j}}{N}} = N \times \frac{(f_{ij} - (f_{i.} \times f_{.j}))^2}{f_{i.} \times f_{.j}}$
e. Compute $A_{ij} = \chi^2_{ij} \times sign(f_{ij} - (f_{i.} \times f_{.j}))$
f. Sort the $A_{ij}$ by decreasing order
g. For each class $j$ do
   (a) Create the list $\{gram_{lj}\}$ for $l = 1, ..., L$, of the $L$ first n-grams of the class
   (b) for each $gram_{lj}$ do
       i. find all words $(words_{jk})$ such that $gram_{lj} \subseteq words_{jk}$
       ii. count the number $nb_{words_{jk}}$ of repetitions of $word_{jk}$ in the class
   (c) For each $word_{jk}$ do
       i. extract all $gram_{word_{jk}}$ included in $word_{jk}$, their total is noted $nbGram_{word_{jk}}$
       ii. For each gram $gram_{word_{jk}}$ do
           if $(gram_{word_{jk}} \in \{gram_{lj}\})$ then $presenceGram_{word_{worjk}}$ ++
   (d) if $\frac{presenceGram_{word_{jk}}}{nbGram_{word_{jk}}} > threshold_1$ and $nb_{words_{jk}} > threshold_2$
       then $word_{jk} \in \{candidate\, key\, words\, for\, class\}$
```

Several papers have shown the effectiveness of n-grams as a means of representing texts for clustering (partition in homogeneous groups) or categorization (assigning a text to one or many categories from a given list); see [10], [7], [3], [4].

Why are n-grams so efficient at classifying texts? How can one move up from shape to substance? This paper will help understanding why n-grams are efficient. The passage from the form to the meaning of a text is restored, moving from n-grams characteristic of a class of texts to the words comprising them.

The operational result is the automatic unveiling of a list of statistically characteristic words, that is, candidate key-words from which the user can

The class	**Acquisition**	**Earn**	**Money-fx**	**Wheat**	**Trade**
Nb texts	1629	2841	528	209	362
The class	**Crude**	**Corn**	**Grain**	**Interest**	**Ship**
Nb texts	383	173	427	346	194

Table 1: Distribution of the number of texts among the 10 largest classes

select those to be retained. This work thus precedes that of [6] who, in pursuit of the same objective, had suggested an iterative method for the selection of words or of groups of words.

Steps needed in the detection of statistically characteristic words are described in section 2; two examples on rather large real life data sets follow; the final section indicates directions for future research. But first, let's review the principles of n-gram coding and its properties.

n-gram Coding A "n-gram" is a sequence of n consecutive characters. The set of n-grams (usually, n is set to 2, 3 or 4) that can be generated for a given document is basically the result of moving a window of n characters along the text. The window is moved one character at a time. Then, the number of occurrences of each n-gram is counted. For example, the phrase "The babysitter babysits the baby" can be represented by [the= 2, he_=2 , bab= 3, aby= 3, bys= 2, ysi=2 , sit=2 , itt= 1, tte= 1, ter= 1, er_= 1, r_b= 1, _ba= 3, its= 1, ts_= 1, s_t= 1, _th= 1, e_b= 2]. To simplify reading, the character "_" will be used to represent a blank.

Advantages of n-gram coding Techniques based on n-grams offer many advantages:

- Comparatively to other techniques, n-grams automatically capture the roots of the most frequent words [5]. There is no need for the identification of lexical roots (babysit, babysitter, babysitting, etc.).
- They work regardless of languages [4], contrary to word-based systems that require language-specific dictionaries (gender; singular-plural; conjugations; etc.). Moreover, n-grams do not require prior segmentation of the text into words; this is an interesting feature when dealing with languages where word limits are not clear, arabic language, for example.
- They are robust to spelling mistakes and distortions caused by optical text recognition. Scanned texts are often imperfect; for example, "chapters" could be recognized as "clapters". A word-based system might have some difficulty in recognizing "chapters" because of the erroneous spelling. Yet, a system using n-grams may still use the recognizable n-grams "apte", "pter", etc. [7] have shown that document search systems based on n-grams remained efficient in spite of a 30% distortion, a situation where no word-based system can operate correctly.

Acquisition class		**Crude class**	
The most significant 3-grams	**Extracted key-words**	**The most significant 3-grams**	**Extracted key-words**
acq cqu qui uis iti sit	acquisition	oil _oi il_ il,	oil oil,
acq cqu qui uir	acquire acquired acquiring acquiring	rud cru ude	crude
sha har are	share	bar arr rel els	barrels
sha har are reh hol lde eho old der	shareholder holders holding hold	cua uad dor ado	ecuador ecuadorean
com omp any pan	companies company;	bpd _bp pd_ pd,	bpd (baril par jour)
tak ake eov keo	takeover stake take	gas	gas
sto toc ock lde	stockholders	ene erg rgy nergy_	energy
mer erg	merger; merge	pet etr leu eum ole	petroleum
off ffe fer	offer offers offering offered	plo xpl lor ora	exploration
has pur has	purchase	sau aud udi di_	saudi
usa sai air	USAir	zue ezu nez uel	venezuela
buy	buy buys	bbl	bbl (barrel)
inv nve sto	investment investment investor	pip ipe pel	pipeline
scl	disclosed undisclosed	xxo exx	exxon
cyc ycl	cyclops	ref ner efi	refinery
sac	transaction		
com omp ple let	complete	ara rab iea	arabian arabia
fil	filing	cub bic ubi	cubic
oup	group	_ku kuw uwa wai	kuwait kuwaiti
tst	outstanding	ric ice ces	prices
twa	twa (Trans World Airline)	ope pec	opec (non-opec)

Table 2: The first few significant grams and corresponding key-words for *Acquisition* and *Crude* classes

- Finally, n-gram techniques do not have to discard stop words and do not require stemming. These processes improve word-based systems. For n-gram based systems, studies have shown that they don't improve after stemming and discarding stop words [9].

2 Identification of characteristic words

The key idea is to extract the n-grams typical to a class, then to retain the words containing the n-grams. The authors have written a Java program that looks for and counts n-grams among classes of texts, selects those that appear to be most typical of the classes, and then looks for words containing the typical n-grams and eliminates parasite words, see algorithm 1.

Identifying characteristic n-grams Before the complete algorithm is described, here are its principles. The key steps are: (i) identify all n-grams contained in all the texts of the learning set; (ii) build the cross table (text class×n-gram); (iii) compute (χ^2_{ij}) the cell contributions to the independence χ^2; (iv) for each class: identify characteristic n-grams (those significantly more frequent in some classes that in others); (v) search for words containing those n-grams.

While a number of statistics can be derived from the matrix of frequency (N_{ij}) of n-gram i in text class j, the χ^2_{ij} distance is often quoted as the most efficient in empirical comparisons [1], [11]. There are theoretical reasons for that; [2] showed that it is asymptotically equivalent to the "information gain".

In practice, the method suggested here yields a long list of words among which a number of parasite words, that is words, though otherwise uninteresting, happen to contain one of the characteristic n-grams. The objective, now, is to determine the list of "candidate key-words".

Filtering out parasite words In order to avoid "parasites", the process can be repeated backwards: for each word identified earlier, the n-grams it contains are examined and matched against the list of the n-grams characteristic for the class. If:

- the proportion of n-grams in the word matched to the list of n-grams characteristic of the class reaches some threshold, and
- the frequency of the word in the text also reaches some threshold,

then the word is considered to be a "candidate key word". If the word appears often on the text, it should be a candidate. If it occurs rarely and was selected because it contains one or two of the n-grams found in a common word, then it must be a parasite.

For example, a 3-gram like "*acq*" in the class of "Acquisition" will yield words characteristic of the class, like "acquisition" or "acquire"; but the 3-gram can also be found in uncharacteristic words, like "Jacques" or "racquets" that will be considered as parasites because they are rare in the class.

The most significant 3-grams
[oil] [_oi] [bpd] [rud] [_bp] [il_] [cru] [pd_] [bar] [rel] [cua] [arr] [etr] [uad] [gas] [...] [rgy] [eum] [leu] [xpl] [els] [sau] [dor] [pet] [ado] [zue] [ezu] [0_b] [nez] [uel] [ira] [aud] [ole] [di_] [bbl] [ec_] [pip] [lor] [cks] [l_p] [pd,] [tpu] [plo] [utp] [gy_] [..1] [odu] [cub] [rod] [ude] [kuw] [uwa] [pd.] [n_b] [i_a] [_cr] [pel] [iea] [rre] [bic] [_ir] [ice] [xxo] [exx] [raq] [bia] [ner] [udi] [/bb] [ara] [ipe] [rab] [ene] [pd)] [mex] [obr] [thq] [hqu] [s/b] [uak] [_op] [duc] [al-] [ref] [(bp] [e_o] [ubi] [fie] [ait] [pec] [tro] [fue] [uot] [_ie] [ora] [ls_] [dri] [quo] [fsh] [efi] [eia] [mob] [as_] [exa] [pdv] [vsa] [dvs] [wai] [_ku] [_ga] [f_o] [ric] [um_] [_bb] [ces] [ukm] [dez] [iel] [urk] [aqi] [try] [xic] [uct] [abi] [naz] [rol] [xac] [a_b] [erg] [eik] [_dr] [put] [prt] [qi_] [c_m] [kh_] [ian] [tex] [ikh] [aeg] [ia_] [ia'] [_pd] [aq_] [rs/] [wti] [_km] [noc] [ope] [ubr] [il,]
Extracted key-words
[oil.] [oil,] [oil] [oilfield] [bpd] [(bpd)] [bpd.] [bpd,] [crude] [crude,] [crudes] [crude.] [oil prices] [oil industry] [oil companies] [oil prices,] [oil prices.] [oil price] [oil and] [oil production] [bpd in] [barrel.] [barrel] [barrels] [barrels.] [barrel,] [barrels,] [reliance] [ecuador,] [ecuador] [ecuador's] [ecuadorean] [petroleum] [petrobras] [petroleos] [petroleum,] [gasoline] [gas] [energy] [exploration] [exploratory] [exploration.] [saudi] [saudis] [venezuela] [venezuela,] [venezuelan] [venezuela's] [fuel] [iraqi] [iraq] [iranian] [iran] [iran's] [saudi arabia] [dlrs/bbl.] [opec] [non-opec] [pipeline] [pipeline,] [stocks] [output] [products] [product] [production] [production.] [producing] [producer] [produce] [producers] [produced] [products.] [products,] [production,] [cubic] [kuwait] [kuwait,] [kuwaiti] [mln barrels] [mln bpd] [iea] [current] [prices] [prices.] [price] [prices,] [exxon] [arabia] [arabian] [arabia's] [arabia,] [refinery] [general] [refineries] [refiners] [including] [arab] [mexico] [earthquake,] [earthquake] [operating] [opec's] [operations] [open] [reduction] [refining] [crude oil] [the oil] [because of] [price of] [fields] [field] [expected] [quota] [quoted] [barrels per] [barrels of] [barrels a] [drill] [drilling] [offshore] [mobil] [was] [has been] [as] [texas] [as a] [texaco] [pdvsa] [of oil] [american] [sources] [industry] [ministry] [a barrel.] [a barrel] [sheikh] [drop] [canadian]

Table 3: Complete list of class specific n-grams and candidate key-words for Crude class.

3 An example

The proposed method should help the user to select documents of interest among an unstructured set of documents like the Web. First, the user must assemble a learning set, that is, provide two sub-sets of documents: one subset containing some documents of interest, and a second sub-set containing texts from the same source(s) but irrelevant to the task at hand.

As this work is user specific, the following example is general enough that the reader can easily understand and control, namely selecting news briefs on a given topic.

Reuters' indexed data The example uses Reuters press agency news briefs as a benchmark.

For the purpose of the example, the learning sub-set of 6709 news briefs from the 10 largest classes is derived from the "Apte version" comprising 7789 briefs [11]. Table 1 shows the size of each class.

Some results Results for Acquisition and Crude classes are displayed in Table 2. The method selected about 100 candidate key-words for each class; due to space constraint, only the first few significant grams and corresponding words are listed.

About the results using the Reuters collection For each class, the method proposes a list of class-specific candidate key-words, free of most parasites. The lists appear reasonable. Obviously, these results are in part due to the events of the time having generated the texts. The rule of "*all else being equal*" applies here as well.

The method is completely independent from the language of the text, since there is no need to remove spaces and punctuation marks.

No real improvement was noted from the different trials were conducted with either 1+2+3-grams or 4-grams did: results are quasi identical. This coincides with other authors' findings, for example [6, 3].

Two tables complete the exposition:

- The complete list of the class-specific n-grams for "crude" class are displayed in Table 3.
- A comparison of the result of four different text coding techniques using the proposed method is shown in Table 4:
  - Computation of χ^2_{ij} on the $(word_i \times class_j)$ cross table with elimination of spaces and punctuation,
  - Computation of χ^2_{ij} on the $(word_i \times class_j)$ cross table without elimination of spaces and punctuation,
  - Computation of χ^2_{ij} on the $(gram_i \times class_j)$ cross table with elimination of spaces and punctuation,
  - Computation of χ^2_{ij} on the $(gram_i \times class_j)$ cross table without elimination of spaces and punctuation.

It can be seen that the proposed method, based on n-grams without preprocessing, gives excellent results. It is hence completely independent from the language of the texts of interest.

4 Conclusion

The algorithm proposed in this paper is adept at extracting candidate key-words that are characteristic of a set of texts. An example is given on a set of 6709 Reuters news briefs organized in 10 classes (the 10 largest of the

text coding techniques	**Extracted key-words**
Extracted key-words **from complete words with** punctuations and spaces elimination	[oil] [bpd] [crude] [opec] [barrels] [barrel] [ecuador] [energy] [exploration] [petroleum] [prices] [gasoline] [gas] [refinery] [saudi] [saudis] [pipeline] [production]
Extracted key-words **from complete words without** elimination of the punctuations and spaces	[oil.] [oil,] [oil] [crude] [opec] [opec's] [non-opec] [barrels] [barrels.] [barrels,] [bpd] [(bpd)] [bpd.] [bpd,] [energy] [petroleum] [ecuador,] [ecuador] [ecuador's] [exploration] [gasoline] [gas] [refinery] [saudi] [saudis] [prices] [prices.] [prices,] [barrel.] [barrel] [barrel,] [cubic] [production] [production,] [output] [stocks] [drilling] [pipeline] [pipeline,] [today] [day] [days] [yesterday] [iea] [arabia] [arabian] [natural] [venezuela] [venezuelan] [texaco] [petrobras] [api] [herrington] [mobil] [exxon] [offshore] [iranian] [feet] [15.8] [quota] [refining] [reserves] [kuwait] [wells] [fuel] [fields] [industry] [field] [iraqi] [minister] [spot] [demand] [price] [lukman] [santos] [producing] [iraq] [shell] [sources] [texas] [rigs] [research] [sea] [iran] [greece] [gulf]
Extracted key-words from **n-grams with** elimination of the punctuations and spaces	[oil] [bpd] [bp] [crude] [crudes][oil prices] [oil industry] [oil stocks] [oil companies][oil minister] [oil company] [oil price] [oil and] [oil production] [bpd in] [barrel] [barrels] [ecuador] [ecuadorean] [petroleum] [petrobras] [petroleos] [petro-canada] [gasoline] [gas] [energy] [exploration] [exploratory] [levels] [saudi] [saudis] [venezuela] [venezuelan] [fuel] [iraq] [iranian] [iran] [000 barrels] [000 bpd] [saudi arabia] [bbl] [pipeline] [stocks] [output] [products] [product] [production] [producing] [producer] [produce] [producers] [produced] [cubic] [kuwait] [kuwaiti] [iea] [mln barrels] [mln bpd] [current]
Extracted key-words **from n-grams without** elimination of the punctuations and spaces	[oil.] [oil,] [oil] [oilfield] [bpd] [(bpd)] [bpd.] [bpd,] [crude] [crude,] [crudes] [crude.] [oil prices] [oil industry] [oil companies] [oil prices,] [oil prices.] [oil price] [oil and] [oil production] [bpd in] [barrel.] [barrel] [barrels] [barrels.] [barrel,] [barrels,] [reliance] [ecuador,] [ecuador] [ecuador's] [ecuadorean] [petroleum] [petrobras] [petroleos] [petroleum,] [gasoline] [gas] [energy] [exploration] [exploratory] [exploration.] [saudi] [saudis] [venezuela] [venezuela,] [venezuelan] [venezuela's] [fuel] [iraqi] [iraq] [iranian] [iran] [iran's] [saudi arabia] [dlrs/bbl.] [opec] [non-opec] [pipeline] [pipeline,] [stocks] [output] [products] [product] [production] [production.] [producing] [producer] [produce] [producers] [produced] [products.] [products,] [production,] [cubic] [kuwait] [kuwait,] [kuwaiti] [mln barrels] [mln bpd] [iea] [current] [prices] [prices.] [price] [prices,] [exxon] [arabia] [arabian] [arabia's] [arabia,] ...

Table 4: Comparisons of four techniques on the class ”crude”.

Reuters collection). The proposed method gives good results as it selects key-words sharing characteristic n-grams. A method to remove parasite words, based on the proportion of significant n-grams they contain, is also described. The method is efficient although it operates on raw texts, without any prior linguistic analysis.

References

[1] Aas K., Eikvil L. (1999). *Text categorization: a survey.* Technical report, Norwegian Computing Center.

[2] Benzecri J.P. (1976). *L'Analyse des Donn'ees*, volume 2. Dunod, Paris.

[3] Cavnar W.B., Trenkle J.M. (1994). *N-gram-based text categorization.* In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 161 – 175, Las Vegas, US.

[4] Dunning T. (1994). *Statistical identification of languages.* Technical Report MCCS 94-273, Computing Research Laboratory.

[5] Grefenstette G. (1995). *Comparing mTwo language identification schemes.* In Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95), Rome, Italy.

[6] Lelu A., Hallab M. (2000). *Consultation 'floue' de grandes listes de formes lexicales simples et compos'ees: un outil pr'eparatoire pour l'analyse de grands corpus textuels.* In Rajman M. and Chappelier J.C., editors, JADT'2000, **1**, 317 – 324, Lausanne.

[7] Miller E., Shen D., Liu J., Nicholas C. (1999). *Performance and scalability of a large-scale N-gram based information retrieval system.* Journal of Digital Information, **1**(5).

[8] Morin A. *Intensive use of correspondance analysis for information retrieval.* In ITI'04. To appear.

[9] Sahami M. (1999). *Using machine learning to improve information access.* PhD thesis, Computer Science Department, Stanford University.

[10] Teytaud O., Jalam R.(2001). *Kernel based text categorization.* In Proceeding of IJCNN-01, 12th International Joint Conference on Neural Networks, Washington, US,2001. HEEEComputer Society Press, Los Alamitos, US.

[11] Yang Y. (1999). *An evaluation of statistical approaches to text categorization.* Information Retrieval, **1** (1/2), 69 – 90.

Address: R. Jalam, J.-H. Chauchat, J. Dumais, ERIC Laboratory, Lyon 2 University 5, av. Pierre Mends-France F-69676 Bron - France

E-mail: {rjalam, chauchat}@univ-lyon2.fr; jean.dumais@statcan.ca

MODELLING TIME OF UNEMPLOYMENT VIA LOG-LOCATION-SCALE MODEL

Eva Jarošová, Ivana Malá, Miroslav Esser, Jan Popelka

Key words: Survival analysis, regression methods, interval censored data, log-location-scale model, smoothing splines.

COMPSTAT 2004 section: Applications.

Abstract: Factors influencing time of unemployment in the Czech Republic are analysed by means of parametric regression models. The sample of unemployed was obtained from the Labour Force Sample Survey organised by the Czech Statistical Office. The time of unemployment is modelled in dependence on age, sex, education and region and the character of dependence on age is examined via smoothing splines. S-Plus 4.5 was used to fit the model.

1 Introduction

The analysis described in our paper is a part of the project "Analysis of factors influencing time to reemployment in the Czech Republic" supported by IGA[1]. The aim of the project is modelling of the length of unemployment depending on demographic, social and other characteristics. Data describing the unemployed are gathered either at Labour Offices or by the Labour Force Sample Survey (LFSS) organised by the Czech Statistical Office (CZSO). The main differences between these two sources are

a. The Labour Office register includes all (registered) unemployed whereas the unemployed in LFSS come from households chosen at random.

b. When data from Labour Offices are used, a cohort of unemployed, who had been registered at a given time point, can be followed for some time interval. The CZSO survey is carried out quarterly and the information about how long the individual has been unemployed is recorded. The sample of households is partly renewed so that 20% of them are missing in the following sample.

c. Labour Office registers contain right-censored data relating to the time of unemployment. Censoring occurs due to the fact that some subjects are not employed during the follow-up period, the other data are non-censored. In LFSS an interviewed respondent does not have to give the length of unemployment exactly, and his or her response concerning duration of the search of work is then categorized by means of intervals, such as up to 1 month, from 1 to 3 months, from 3 to 6 month etc.

[1] The paper was written with the financial support of Internal Grant Agency of the University of Economics, grant No IG 410043.

Data from LFSS are right or interval censored, non-censored data are not recorded in the database.

The aim of this study was to examine whether LFSS data about unemployed can be used for modelling, and if they can, to find a suitable model and quantify differences among the regions in the Czech Republic, between males and females and among education levels. We wanted to investigate the form of dependence on age as a part of our model.

Two main approaches to regression modelling of censored data can be used. In the proportional hazards Cox model the hazard function is directly modelled and explanatory variables affect it multiplicatively. The advantage is that no special form of distribution for time is required, the disadvantage consists in the fact that interval censored data cannot be incorporated. This drawback can be got round by transforming the problem and using the general linear model with the binomial distribution and complementary the log-log link. In this approach the censoring variable is treated as the dependent variable, but the censoring intervals must not overlap. This requirement is not fully met in LFSS data.

In the log-location-scale models explanatory variables affect the time scale. The advantage of these models (also called accelerated failure time models) is the possibility to apply them both on right-censored and interval-censored data. Because only censored observations are present in our sample, we cannot perform the usual residuals diagnostics and must rely on comparison of models by Akaike information criterion (AIC). The form of continuous variables in the model can advantageously be explored by means of smoothing splines. The general additive model may be used in connection with the semiparametric Cox model [4]. As the Cox model cannot be applied on interval censored data, some modifications of data must be done.

2 Subjects and data collection

The sample involves 1455 unemployed found out in LFSS in the first quarter of 2003 who continued in the next survey in the second quarter of 2003. Only those younger than 60 and unemployed less than 4 years were included into the analysis. There are 628 males and 827 females in the sample. Characteristics of the unemployed are given in Tables 2 to 4. Counts of unemployed according to the regions, sex, education and age groups correspond to the percentages of unemployed reported by CZSO in this period and the sample can be considered as representative.

Interval (months)	Count	
	Interval-censored	Right-censored
0 - 1	0	7
1 - 3	0	11
3 - 8	6	208
6 - 14	10	361
12 - 20	181	199
14 - 20	86	0
18 - 26	1	122
24 - 50	1	262

Table 1: Intervals for the length of unemployment.

Region	Count
Prague	65
Central Bohemia	105
Southwest	159
Northwest	208
Northeast	203
Southeast	224
Central Moravia	207
N.Moravia, Silesia	284

Table 2: Frequency table for regions.

Education	Count
Basic	307
Secondary without GCE	718
Secondary with GCE	386
Tertiary	44

Table 3: Frequency table for education levels.

Age	Count
15 - 19	83
20 - 29	486
30 - 39	339
40 - 49	280
50 - 59	267

Table 4: Frequency table for age groups.

In our sample 285 observations represent interval-censored data, the other data are right-censored, the interval distribution of the time of unemployment is in Table 1[2]. The variables are described as follows:

AGET	age in 2003 - 36	(36 is the median age in the sample)
SEX	sex	(1, male; 2, female)
EDU	education	(4 levels, see Table 3)
REG	region	(8 levels, see Table 2)

As for factors SEX, EDU, and REG, the model parameters correspond to the individual factor levels excluding the first one.

3 Methods

3.1 A parametric regression model

From the reason given above log-location-scale model is used. In this approach a suitable probabilistic distribution is fitted to the logarithm of the

[2] Right censoring relates to the lower limit of intervals, e.g. 11 individuals from our sample have been unemployed for more than one month.

time of unemployment T and four independent variables are included. The model parameters are estimated by the maximum likelihood method.

The regression model is considered in the form

$$ln(T) = \mu + \mathbf{x}'\boldsymbol{\beta} + \sigma W,$$

where μ is a constant, σ is a scale parameter, $\boldsymbol{\beta}$ is a vector of regression parameters and $\mathbf{x}$ is a vector of explanatory variables. A suitable form of random term W distribution is selected by means of AIC[3].

The survival function $P(T > t) = 1 - F(t)$ and the hazard function $f(t)/(1 - F(t))$ of the baseline distribution (for an individual with $\mathbf{x} = \mathbf{0}$) will be denoted $S_0(t)$ and $h_0(t)$, respectively. In our parametrization it corresponds to the distribution of the time of unemployment of a 36-year old man from Prague with basic education. For an individual with values of independent variables $\mathbf{x}$ the survival function $S(t, \mathbf{x})$ and the hazard function $h(t, \mathbf{x})$ are determined by the formulas

$$S(t, \mathbf{x}) = S_0(te^{-\mathbf{x}'\boldsymbol{\beta}}) \qquad \text{and} \qquad h(t, \mathbf{x}) = e^{-\mathbf{x}'\boldsymbol{\beta}}h_0(te^{-\mathbf{x}'\boldsymbol{\beta}})$$

The acceleration factor exp(-$\mathbf{x}'\boldsymbol{\beta}$) describes the change in the time scale from the baseline. The survival time is accelerated if $\mathbf{x}'\boldsymbol{\beta}$ is negative and decelerated if $\mathbf{x}'\boldsymbol{\beta}$ is positive. It means that the time of unemployment tends to be shorter if $\mathbf{x}'\boldsymbol{\beta}$ is negative and longer for positive values of $\mathbf{x}'\boldsymbol{\beta}$.

3.2 Exploring the functional form of continuous covariates

The use of splines helps to find a suitable form of the continuous covariate in the model. Fitting the parametric regression model for survival data with regression or natural splines included is easy in S-Plus. But knots of these splines, being placed evenly between the extremes of the covariate, need not necessarily correspond to the real change-points in the relationship. Smoothing splines are preferable. The use of smoothing splines requires the general additive model to be implemented. It can be applied on right censored data after the suitable transformation of the Cox proportional hazards model [3].

The hazard function from the Cox model can be expressed in the form

$$h(t, \mathbf{x}) = h_0(t)\text{exp}\left(\sum_{j=1}^{p} f_j(\mathbf{x}_j)\right) = h_0(t)\text{exp}(\eta(\mathbf{x}))$$

where $f_j (j = 1, 2, ..., p)$ are spline functions, $\eta(\mathbf{x}) = \sum_{j=1}^{p} f_j(\mathbf{x}_j)$. It follows that

$$\text{-ln}S(t) = H_0(t)\text{exp}\left(\eta(\mathbf{x})\right),$$

where $H_0(t) = \int_{-\infty}^{t} h_0(u)du$ is the cumulative baseline hazard function.

[3] AIC $= -2\ln\hat{L} + 2q$, where q is number of parameters.

The log-likelihood has the form

$$\begin{aligned} L &= \sum_{i=1}^{n} \{c_i \left[\ln h_0(t_i) + \eta(\mathbf{x}_i)\right] - H_0(t_i)\exp(\eta(\mathbf{x}_i))\} = \\ &= \sum_{i=1}^{n} \left\{ c_i \left[\ln H_0(t_i) + \eta(\mathbf{x}_i)\right] - H_0(t_i)\exp(\eta(\mathbf{x}_i)) + c_i \ln\left(\frac{h_0(t_i)}{H_0(t_i)}\right) \right\}. \end{aligned}$$

By replacing $H_0(t_i)\exp(\eta(\mathbf{x}_i))$ by μ_i we get

$$L = \sum_{i=1}^{n} (c_i \ln \mu_i - \mu_i) + \sum_{i=1}^{n} c_i \ln\left(\frac{h_0(t_i)}{H_0(t_i)}\right).$$

The first term corresponds to the kernel of the likelihood function for n independent Poisson variables c_i with means μ_i while the second term does not include unknown functions f_j. The generalized additive model with the Poisson distribution and the logarithmic link function can be used with a fixed intercept (offset) $\ln H_0(t_i)$ included in the predictor. The cumulative baseline hazard function $H_0(t)$ can reasonably be estimated by fitting the Cox model based on untransformed data.

To make use of the Cox model connected with the general additive model and to find an approximate form of the age term, we replaced interval limits of interval-censored data by the intervals midpoints and regarded them as non-censored in this part of analysis. After fitting the general additive model we suggested some suitable parametric forms of dependence based on the plot of the fitted term vs age. Then parametric regression models with various forms of the age term were fitted and AIC was used to select the most suitable one.

S-Plus 4.5 was used in the analysis.

4 Results

First the dependence on age was explored without any assumption about the distribution of T. The plot of the fitted term for the variable AGET of the Poisson generalized additive model with an offset is presented in Figure 1. It can be seen that there is some decrease of the log hazard at the young age and another downturn comes at the age somewhere around 48. The log hazard is approximately constant in the middle part. It implies that the curve could satisfactorily be replaced by a piecewise linear regression function with the ages of 32 and 48 as the change points or by a cubic polynomial.

We considered various underlying probability distributions together with different forms of the age dependence. AIC was computed for each model. Some of the results are summarized in Tables 5 and 6.

Form of AGET	**AIC**
linear	1245.1
quadratic polynomial	1242.6
cubic polynomial	1242.3
piecewise linear	1238.5

Table 5: Comparison of alternative forms of age
(log-logistic distribution).

Distribution of T	**AIC**
Weibull	1422.0
generalized gamma	1242.0
log-normal	1240.0
log-logistic	1238.5

Table 6: Comparison of alternative distributions of T
(piecewise linear model).

The model with log-logistic distribution of T and the piecewise linear age dependence was selected based on AIC. Results of the final fitted model are in Table 7.

Term	**Coefficient**	**Term**	**Coefficient**
SEX	0.074*	REG2	0.095
EDU2	-0.142**	REG3	-0.006
EDU3	-0.180***	REG4	0.160*
EDU4	-0.145	REG5	-0.004
AGET1	0.020***	REG6	-0.004
AGET2	-0.024**	REG7	0.090
AGET3	0.019	REG8	0.210*
μ	3.300	σ	0.200

Table 7: The estimated values of μ, σ and the coefficients $\boldsymbol{\beta}$, with statistical significance given (* P<0.05, ** P<0.01, *** P<0.001).

The coefficients AGET1, AGET2 and AGET3 correspond to the linear effect of age in periods up to 32, 32-48, after 48.

The log-logistic distribution of T corresponds to the standard logistic distribution of the error term W. The parameters μ and σ in the model and parameters α and λ of the log-logistic distribution are related by $\alpha = 1/\sigma$ and $\lambda = \exp(-\mu/\sigma)$. It means that $\alpha = 5$ and $\lambda = 6.8 \cdot 10^{-8}$. For $\alpha > 1$, the hazard rate of the log-logistic distribution is increasing to the maximum at $t = [(\alpha - 1)/\lambda]^{1/\alpha}$, i.e. at 35.8 months according to our model, and it is decreasing to 0 at infinity. The median of the baseline distribution (for $\mathbf{x} = \mathbf{0}$) is given by $\sqrt[\alpha]{1/\lambda}$ and equals to 27.1 months. The median of T for an individual with the vector of explanatory variables $\mathbf{x}$ is a product of the baseline distribution median and $\exp(\mathbf{x}'\boldsymbol{\beta})$. E.g. the estimated median of the time of unemployment for a 36-year-old man from the Central Bohemia with secondary education with GCE is 22.6 months (95percent confidence interval (19.4, 26.3)). For a woman with the same values of other independent variables the estimate of median is exp(0.074) times greater, that means 24.3 months.

The regression coefficients associated with different factor levels enable to compare various groups of individuals corresponding to these levels. There is a significant difference between men and women, among unemployed with basic, secondary and secondary with GCE education.

The log-logistic distribution has a property of proportional odds in the form

$$\frac{S(t,\mathbf{x})}{1-S(t,\mathbf{x})} = \frac{S_0(te^{-\mathbf{x}'\boldsymbol{\beta}})}{1-S_0(te^{-\mathbf{x}'\boldsymbol{\beta}})} = \exp(-\mathbf{x}'\frac{\boldsymbol{\beta}}{\sigma})\frac{S_0(t)}{1-S_0(t)}$$

In addition to medians, this formula can be used to compare chances of various individuals. It is the reason why the log-logistic distribution is often preferred to the log-normal distribution although the AIC values are usually similar.

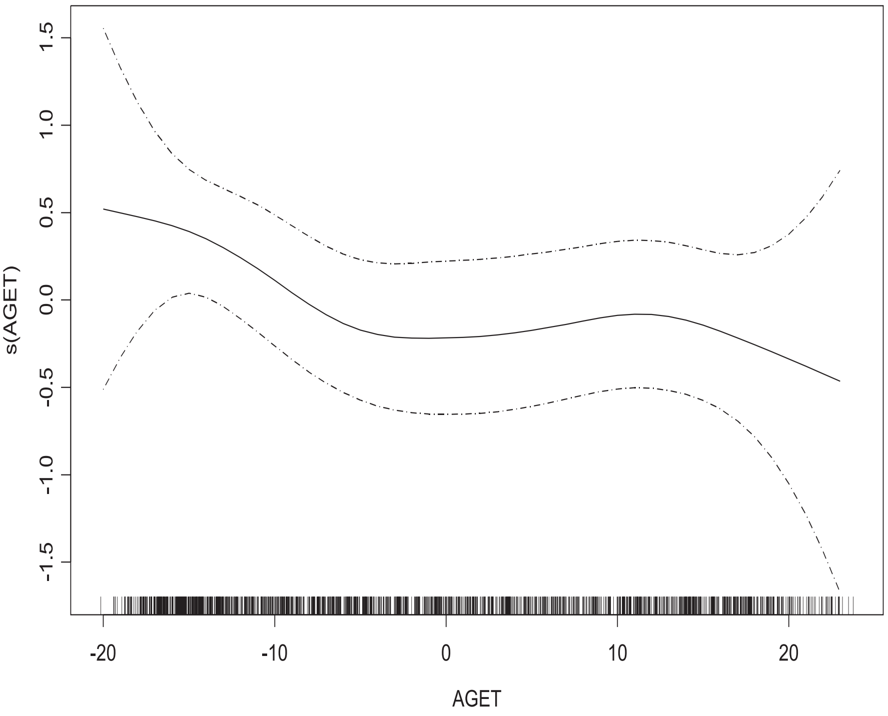


Figure 1: Plot of the fitted term for the variable AGET of the Poisson generalized additive model with an offset.

5 Conclusion

The log-logistic regression model without interactions and with the piecewise linear dependence on age was selected as the best one. The dependence on education and sex was confirmed and the corresponding coefficients are similar as in [2]. As for the dependence on age, the course for lower ages is rather surprising. While the "risk" of the employment was increasing for age up to 22 in [2], what was understandable, now the "risk" is decreasing up to 32. The further trend is similar in both studies, the drop after 48 seems to be regular.

We are convinced that our model can be used at least for purposes of comparison. The analysis could be improved if the precise times of unemployment (if available) were provided by CZSO. They are recorded in the questionnaire during the LFSS interview but they are not saved in the database. Additional explanatory variables included in the CZSO survey (such as disability and experience) could be incorporated into the model in frame of next investigation.

References

[1] Hosmer D.W., Lemeshow S. (1999). *Applied survival analysis.* J.Wiley&Sons.

[2] Jarosova E., Mala I., Esser M. (2004). *Analysis of factors influencing time of unemployment.* Aplimat 2004, Bratislava 04.02.2004 - 06.02.2004, Slovak University of Technology, sp. 525 - 530.

[3] McCullagh P., Nelder J.A. (1989). *Generalized linear models.* Chapman&Hall, 2.ed.

[4] Therneau T.M., Grambsch P.M.(2000). *Modeling survival data.* Extending the Cox Model, Springer-Verlag

Address: E. Jarošová, Department of Statistics and Probability, University of Economics, Prague W. Churchilla sq. 4, 130 67 Prague, Czech Republic

E-mail: `jarosova@vse.cz`

SEMIPARAMETRIC BAYESIAN ANALYSIS OF EPO PATENT OPPOSITION

Alexander Jerak and Stefan Wagner

Key words: Bayesian semiparametric binary regression, Bayesian P–splines, Markov Chain Monte Carlo.

COMPSTAT 2004 section: Bayesian methods.

1 Introduction

The analysis of patent data has a long tradition in economic research. Recent empirical work focuses on the patent system and especially the analysis of patent opposition attracted the interest of numerous researches, which consider this legal action as a mechanism to ensure a certain level of quality of issued patents. Considering the procedure at the European Patent Office (EPO), any third party can oppose a patent by filing an opposition within nine months after the grant decision and substantiating it by presenting evidence that one or more of the patentability criteria aren't satisfied by the protected invention.

The common methodology in the analysis of the incidence of such procedures is to model the probability of the occurrence of the discrete event 'opposition or not' with a linear predictor depending on a variety of patent indicators, see e.g. [3]. Among the most prominent indicators found are the number of citations received by younger patents (*forward citations*), the number of claims stated in the patent (*claims*) and the number of states in which an innovation seeks patent protection (*designated states*), which are all measures of a patent's value and do increase the probability of a patent being opposed. Additionally, measures of patent breadth as well as information on the filing strategy are usually included, which are not of primary interest in our analysis.

In this paper, we apply a semiparametric approach described in [1] and [2] to analyze the determinants and the effects of patent oppositions in Europe, in which linear effects $x'\beta$ of metrical covariates x are replaced by smooth regression functions $f(x)$. Within a Bayesian framework we apply Markov Chain Monte Carlo methods (MCMC) for estimation purposes. The methodology presented is implemented in *BayesX*, a software package for Bayesian generalized additive regression based on MCMC techniques available free of charge at `http://www.stat.uni-muenchen.de/~lang`. In order to characterize the benefits from applying semiparametric models we compare our specification to the results of a simple linear probit model employed by [3] using their dataset on EPO patents from the biotechnology/pharmaceutical and semiconductor/computer software sector.

2 Bayesian semiparametric binary regression

2.1 Structural assumptions

Consider regression situations, where observations (y_i, z_i), $i = 1, \dots, n$, on a binary response y and covariates z are given, which can be divided into metrical covariates $x_1, \dots, x_p$ and categorical covariates $w_1, \dots, w_q$. In case of the widely used probit model, the responses y_i, given the covariates, are binomially distributed with the probability of success $\pi_i = P(y_i = 1|z_i)$ being modeled as $\pi_i = \Phi(\eta_i)$. Here, η_i is the predictor that models the influence of the covariates on the probability π_i.

An alternative way of obtaining a probit model, which is very useful for Bayesian inference, is to express binary regression models in terms of latent utilities, see e.g. [2]. Introducing the metric latent utilities $U_i = \eta_i + \epsilon_i$ with i.i.d. errors ϵ_i, we define $y_i = 1$ if $U_i > 0$ and $y_i = 0$ if $U_i < 0$. Then, the assumption $\epsilon_i \sim N(0,1)$ yields the well known probit model.

Concerning the form of the predictor and the type of the influence of metrical covariates $x_1, \dots, x_p$ the following three settings will be distinguished for the rest of the paper, with $x_i = (x_{i1}, \dots, x_{ip})'$ and $w_i = (1, w_{i1}, \dots, w_{iq})'$.

M_1: The effects of the metrical covariates are incorporated into the model by additive linear terms $x_1'\beta_1, \dots, x_p'\beta_p$. The predictor is then given by

$$\eta_i^{(1)} = \sum_{j=1}^{p} x_{ij}\beta_j + w_i'\gamma \tag{1}$$

with the unknown regression parameters $\theta = (\beta_1, \dots, \beta_p, \gamma)$.

M_2: A simple way to allow for non–linearities in the effects of metrical covariates used in [3] is to categorize some or all $x_1, \dots, x_p$ and then construct a set of dummy variables $\tilde{x}_j, j = 1, \dots, p$. The linear terms in (1) are then replaced by $\tilde{x}_j'\tilde{\beta}_j$ with $\tilde{\beta}_j = (\tilde{\beta}_{j1}, \dots, \tilde{\beta}_{jr_j})'$ and the predictor can be defined by

$$\eta_i^{(2)} = \sum_{j=1}^{p} \tilde{x}_{ij}'\tilde{\beta}_j + w_i'\gamma \tag{2}$$

with $\tilde{x}_{ij} = (\tilde{x}_{i1}, \dots, \tilde{x}_{ip})'$ and $\theta = (\tilde{\beta}_1, \dots, \tilde{\beta}_p, \gamma)$.

M_3: An alternative, more flexible and data-driven method for modeling non–linear effects of metrical covariates is to incorporate them additively into the predictor by using smooth regression functions $f_j(x_j)$ instead of the linear terms in (1) and (2). This leads to the semiparametric additive predictor

$$\eta_i^{(3)} = \sum_{j=1}^{p} f_j(x_{ij}) + w_i'\gamma \tag{3}$$

with the unknown parameters $\theta = (f_1(x_1), \dots, f_p(x_p), \gamma)$ and $f_j(x_j)$ representing a vector of function evaluations.

2.2 Bayesian inference via Markov Chain Monte Carlo

Prior assumptions: In a Bayesian approach unknown functions $f_1, \dots, f_p$ and parameters $\beta = (\beta_1, \dots, \beta_p)$, $\tilde{\beta} = (\tilde{\beta}_1, \dots, \tilde{\beta}_p)$, γ of fixed effects are considered as random variables and have to be supplemented by appropriate prior distributions. For more details about the priors in the given context see [1] and [2]

In the absence of any prior knowledge a typical assumption for the parameters of the fixed effects is to use independent diffuse priors, i.e. $p(\beta) \propto const$, $p(\tilde{\beta}) \propto const$ and $p(\gamma) \propto const$.

For the unknown regression functions f_j, we use a P–splines approach formulated in a Bayesian setting by [1]. In a P–splines approach it is assumed that the unknown functions f_j can be approximated by linear combinations

$$f_j(x_j) = \sum_{k=1}^{m_j} \delta_{jk} B_{jk}(x_j)$$

of $m_j = l_j + r_j$ linearly independent B–spline basis functions B_{jr} of degree l_j defined on a set of r_j equally spaced knots $x_{j,min} = \xi_{j0} < \ldots < \xi_{jr_j} = x_{j,max}$. The basis functions can be regarded to have compact local support in the sense that they are nonzero only on a domain spanned by the $l_j + 2$ knots, whereas the B–spline coefficients $\delta_j = (\delta_{j1}, \dots, \delta_{jm_j})'$ act as weights assigned to each single basis function.

To ensure both enough flexibility and sufficient smoothness of the fitted curves, a relatively large number of knots is used, but, in order to prevent overfitting, adjacent B–spline coefficients are penalized with differences of order d. In a frequentist approach this leads to penalized likelihood estimation with roughness penalties, where the trade off between flexibility and smoothness is controlled by additional smoothing parameters λ_j.

In a Bayesian setting, the difference penalties are replaced by their stochastic analogues, i.e. random walks of order d. For simplicity, we will restrict to $d = 2$, which corresponds to a second order random walk

$$\delta_{jk} = 2\delta_{j,k-1} - \delta_{j,k-2} + u_{jk}$$

for adjacent B–splines coefficients δ_{jk} with Gaussian errors $u_{jk} \sim N(0, \tau_j^2)$ and diffuse priors $p(\delta_{j1})$ and $p(\delta_{j2}) \propto const$ for initial values. The amount of smoothness is controlled by the error variances τ_j^2, which are related to the smoothness parameters λ_j by $\lambda_j = (\tau_j^2)^{-1}$. Thus, larger (smaller) values for the variances lead to rougher (smoother) estimates for the regression function. The joint prior of the B–splines coefficients δ_j is Gaussian and can easily be computed as

$$\delta_j | \tau_j^2 \propto \exp\left(-\frac{1}{2\tau_j^2} \delta_j' K_j \delta_j \right)$$

with a penalty matrix $K_j = D'D$, where D is a second order difference matrix.

For a fully Bayesian analysis, variance or smoothness parameters τ_j^2 are also considered to be unknown and estimated simultaneously with the unknown regression parameters. Therefore, hyperpriors are assigned to them in a second stage of the hierarchy by assuming highly dispersed inverse gamma distributions $\tau_j^2 \sim IG(a_j, b_j)$ with known hyperparameters a_j and b_j. A common choice for the hyperparameters is $a_j = 1$ and $b_j = 0.0005$ leading to an almost diffuse prior for τ_j^2. Note, that these prior assumptions for the smoothness parameters are a major advantage over a classical frequentist approach, where smoothness parameters usually have to be specified by hand or a complex grid search algorithm has to be performed.

Posterior analysis: For reasons of brevity, we will focus only on some key results given in [1] and [2]. For a thorough treatment of MCMC in general refer, for example, to [4].

Bayesian inference is based on the posterior and is carried out using recent MCMC simulation techniques. Let θ denote the vector of all unknown parameters in the model. Then, under usual conditional independence assumptions, the posteriors augmented by the latent variables for the three approaches described in Section 2.1 are given by

$$\begin{aligned}
M_1: \quad & p(\theta|Y) && \propto && p(Y|U) \cdot p(U|\eta) \cdot p(\beta) \cdot p(\gamma) \\
M_2: \quad & p(\theta|Y) && \propto && p(Y|U) \cdot p(U|\eta) \cdot p(\tilde{\beta}) \cdot p(\gamma) \\
M_3: \quad & p(\theta|Y) && \propto && p(Y|U) \cdot p(U|\eta) \cdot \prod_{j=1}^{p} \{p(\delta_j|\tau_j^2) p(\tau_j^2)\} \cdot p(\gamma)
\end{aligned}$$

Because the direct maximization of all three posterior distributions is not possible, MCMC methods have to be applied in order to be able to estimate the unknown parameters β, $\tilde{\beta}$, γ, δ_j and τ_j^2, which make use of the full conditionals, i.e. the distribution of a certain parameter block given all the other parameters.

The full conditionals for the fixed effects parameters β, $\tilde{\beta}$ and γ as well as for the parameter vectors $\delta_1, \ldots, \delta_p$ are multivariate Gaussian. For the variance components τ_j^2 the full conditionals are inverse gamma distributions. Finally, it can be shown that the full conditionals of the latent variables U are truncated normals, subject to the constrains $U_t > 0$ if $y_t = 1$ and $U_t < 0$ if $y_t = 0$.

Thus, a Gibbs sampler can be used for MCMC simulation, drawing successively from the full conditionals for the latent variables U, for the fixed effects parameters β, $\tilde{\beta}$ and γ, for the B–splines coefficients δ_j and for the variances τ_j^2. Running this Gibbs sampler yields random samples from the marginal distributions of the regression parameters β, $\tilde{\beta}$, γ, δ_j and τ_j^2, from which Bayesian point estimates like posterior means or posterior medians and credible regions based on suitable quantiles can be calculated.

3 Analysis of patent opposition at the EPO

In this section we reinvestigate a dataset of approximately 4800 patents from the biotechnology/pharmaceutical and semiconductor/computer software sectors granted by the EPO between 1980 and 1997, which has previously been analyzed by [3]. The aim is to model the probability that an opposition against a granted patent was filed, which is coded by $y_i = 1$. We use only the significant covariates found by [3] and restrict the following presentation to the metrical covariates

- x_1: Grant year
- x_2: Number of EPO forward citations
- x_3: Number of designated states
- x_4: Number of EPO claims

Concerning the categorical covariates w and the obtained results see [6].

3.1 Results

As a first step for modeling the probability of an opposition given the covariates, we use a simple linear model M_1 with the predictor

$$\eta_i^{(1)} = \sum_{j=1}^{4} x_{ij}\beta_j + w_i'\gamma$$

The results are given in Table 1 (a). Obviously, the probability of an opposition decreases over time, whereas we find an increase in the probability due to higher numbers of EPO forward citations, EPO claims and designated states, which is in line with the previous findings mentioned in Section 1. Finally, the computed 95 % credible regions for the estimated parameters given in Table 1 (b) and (c) show that all effects are significant on the 5 % error level.

Covariate	(a)	(b)	(c)
Intercept	-0.4355	-0.6043	-0.2737
x_1	-0.0483	-0.0596	-0.0376
x_2	0.0916	0.0738	0.1102
x_3	0.0489	0.0366	0.0625
x_4	0.0134	0.0087	0.0182

Table 1: EPO patent opposition. Results for model M_1. (a) Posterior mean estimate of regression parameter. (b) Lower value of 95 % credible region. (c) Upper value of 95 %credible region.

As an extension of this fully linear model, we applied the approach M_2 used by [3] and our semiparametric approach M_3 with smooth regression functions $f_1(x_1), \dots, f_4(x_4)$. The predictors can then be defined by

$$\begin{aligned}\eta_i^{(2)} &= \sum_{j=1}^{4} \tilde{x}_{ij}'\tilde{\beta}_j + w_i'\gamma \\ \eta_i^{(3)} &= \sum_{j=1}^{4} f_j(x_{ij}) + w_i'\gamma\end{aligned}$$

with the dummy vectors $\tilde{x}_{ij}$ based on the categories given in [3]. Slightly differing from [3], we only used 9 biannual categories for the grant year.

Figure 1 displays the estimated effects of the metrical covariates for both M_2 and M_3. Note that the effects have been centered appropriately to ensure identifiability and comparability. Roughly speaking, the results for the metrical covariates are similar to the ones obtained from M_1, but it is obvious that especially the effects for the number of designated states depicted in Figure 1 (b) and the number of EPO forward citations depicted in Figure 1 (c) are clearly non–linear and that the dummy effects obtained from M_2 are very raw approximations of the true underlying dependency structure represented by the smooth effects in M_3. Additionally, Figure 1 (d) shows, that especially for the number of a patent's claims, the categorization used by [3] is not chosen very well in putting all patents with more than 15 EPO claims into one category with a constant effect.

The significance of the smooth effects in M_3 is supported by the pointwise 95 % credible regions also depicted in Figure 1, which are clearly different from zero for most values of the corresponding covariate.

3.2 Model evaluation

To give a more formal rationale for the benefits in using our semiparametric approach, we compared the three approaches M_1, M_2, M_3 in terms of the deviance information criterion (DIC). The DIC is a Bayesian analogue to the Akaike information criterion (AIC) penalizing the fit of a model measured by the deviance with the complexity of a model represented by the effective number of model parameters. Following [5] it can be defined by $DIC = D(\bar{\theta})+2p_D$, where $D(\bar{\theta})$ is the deviance of the model evaluated at the posterior mean estimate $\bar{\theta}$ and p_D is the effective number of model parameters. The results are given in Table 2 and show, that the DIC is clearly minimized by our semiparametric approach M_3.

Additionally, we did compare our three models in terms of their accuracy ratios, a widely used performance measure based on receiver operating characteristic (ROC) curves. In our context of patent oppositions, the construction of a ROC curve can be shortly summarized as follows: Given the observed values of our binary response variable Y and based on the estimated probabilities $\hat{\pi} = P(Y = 1)$ for a patent being opposed, the hit rates $H(c) = P(\hat{\pi} \geq c|Y = 1)$ and false alarm rates $F(c) = P(\hat{\pi} \geq c|Y = 0)$ are es-

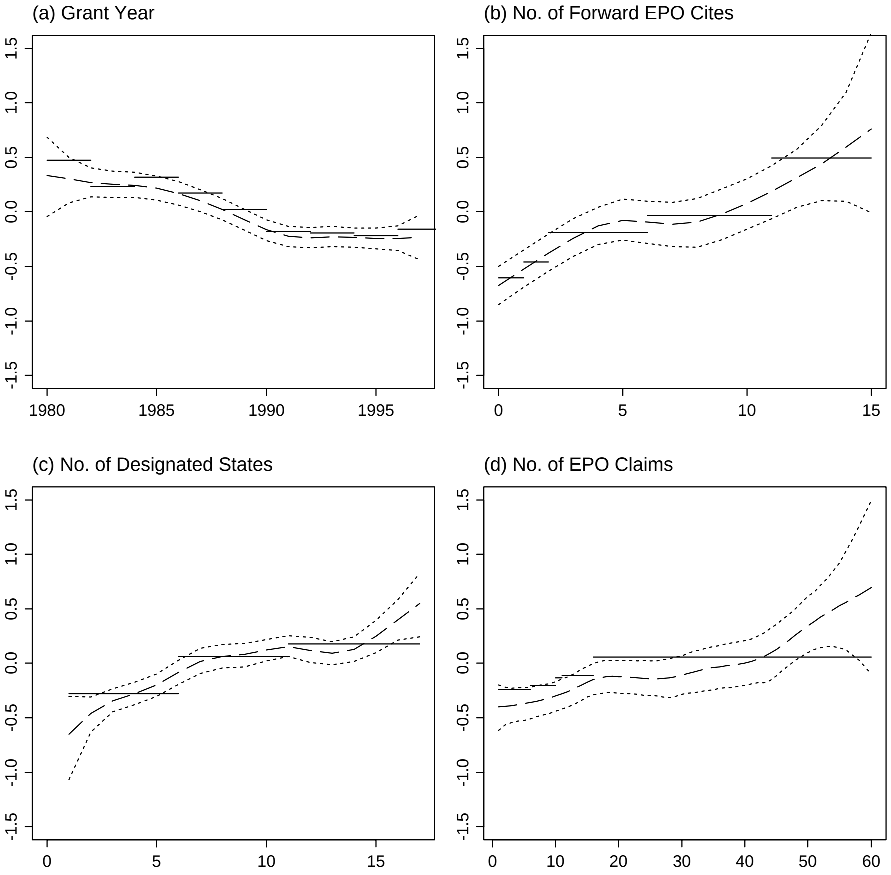


Figure 1: EPO patent opposition. Results for effect of (a) grant year, (b) number of forward EPO cites, (c) number of designated states, (d) number of EPO claims. Shown is for model M_2 the posterior mean (—) of the corresponding dummy effects, for model M_3 the posterior mean (- - -) of the corresponding regression function within 95 % credible regions $(\cdots)$.

timated for a set of ordered threshold values $c = \{c_1, c_2, \ldots, c_S\}$, $0 \leq c_j \leq 1$, and plotted in a squared box of length one, with $F(c)$ on the abscissa and $H(c)$ on the ordinate. Typical global performance measures based on the ROC curve are the area under the curve (AUC) or the accuracy ratio (AR) defined by $AR = 2 \cdot AUC - 1$, with higher values indicating a better performance of the classification system. For a more detailed description and key literature concerning ROC curves please refer to [6].

The results for the accuracy ratios of the three approaches considered in this paper are given in Table 2 and show, that the highest value is obtained by our semiparametric model M_3, which can therefore be considered as the superior approach in terms of both DIC and AR.

	Deviance	pD	DIC	AR
M_1	5680.64	12.05	5704.74	0.4736
M_2	5671.33	25.81	5722.95	0.4754
M_3	5628.84	31.32	5691.48	0.4874

Table 2: EPO patent opposition. Comparison for models M_1 - M_3 of deviance, effective number of parameters (pD), deviance information criterion (DIC) and accuracy ration (AR) for receiver operating characteristic curve.

4 Future work

One focus for future research could be a segmentation routine detecting similarities in patent/opposition characteristics independent of prespecified technology or geographical classifications based on an extension of Bayesian additive mixed models. Additionally, the application of Bayesian semiparametric models for multicategorical response and for survival analysis might be useful in the analysis of the outcome or the duration of the opposition procedure.

References

[1] Brezger A., Lang S.(2003). *Generalized structured additive regression based on Bayesian P–splines.* Discussion Paper 321, SFB 386, University of Munich.

[2] Fahrmeir L., Lang S.(2001). *Bayesian semiparametric regression analysis of multicategorical time–space data.* Annals of the Institute of Statistical Mathematics **53**, 10 − 20.

[3] Graham S., Hall B., Harhoff D., Mowery D.(2002). *Post-issue patent "quality control": A comparative study of US patent reexaminations and European patent oppositions.* Working Paper 8807, NBER.

[4] Green P. J.(1999). *A primer on Markov Chain Monte Carlo.* In O. E. Barndorff-Nielsen, D. R. Cox, C. Klžppelberg (editors), *Complex Stochastic Systems*, 1 − 62. Chapman and Hall, London.

[5] Hennerfeind A., Brezger A., Fahrmeir L.(2003). *Geoadditive survival models.* Discussion Paper 333, SFB 386, University of Munich.

[6] Jerak A., Wagner S.(2003). *Estimating probabilities of patent oppositions in a Bayesian semiparametric regression framework.* Discussion Paper 323, SFB 386, University of Munich.

Acknowledgement: This research was supported by the German National Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures".

Address: A. Jerak, S. Wagner, Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany

E-mail: `jerak@stat.uni-muenchen.de`

COMPSTAT'2004 Symposium

MODELLING THE PROBABILITY OF REJECTION IN A QUALIFICATION TEST BASED ON PROCESS DATA

I. Juutilainen and J. Röning

Key words: Statistical computing, heteroscedastic linear model, dispersion modelling, process data analysis, acceptance criteria, probability prediction, conditional distribution function.

COMPSTAT 2004 section: Applications.

Abstract: We present a method for using process data measured from an industrial production process to predict the probability of rejection in a qualification test of the characteristics of the end product. In the proposed approach, the mean and the dispersion are modelled jointly using a heteroscedastic regression model. The estimation of probabilities is based on the predicted mean, the predicted deviation and an estimate of the cumulative distribution function of the standardised residuals. In a case study, a prediction model is developed for the probability of fulfilling the strength requirements of a steel plate in a tensile test.

1 Introduction

Predictive modelling aims to find out how the conditional distribution of the response depends on the explanatory variables. In the usual approach, the conditional distribution is derived from the distributional assumption and the predicted mean. A more general approach is to model also the variance and to approximate the conditional distribution using both the estimated variance and the estimated mean. Further, the most general and the most complex approach is to model the whole conditional distribution function as a function of the explanatory variables.

In an industrial process of production, millions of measurements on the process can be made daily. The process data surely contain information that would be very useful in process and product planning or process control. One obvious possibility is to utilise prediction models estimated from the process data. Sometimes, not only the location but also the shape of the conditional distribution function depends strongly on the process settings, and a pure mean model may not give a sufficient idea of the conditional distribution. Joint modelling of mean and dispersion extends remarkably the model framework of conditional distribution functions. However, joint modelling of mean and dispersion has been rarely applied to process data. Process data have several special features that must be taken into account in modelling: The numbers of observations and variables are large, and the observations are clustered in the variable space. Thus, the modelling methods used must be able to handle satisfactorily high dimensionality and large data sets.

In a qualification test, the properties of the end product are measured and compared to pre-defined acceptance criteria. Rejections in qualification tests are very expensive for the company. A model for predicting the probability of rejection may be a valuable tool in optimising the process settings. As an example, we examine the tensile testing of steel plates in chapter 5. To the best of our knowledge, methods for rejection probability prediction using joint modelling of the mean and dispersion have not been specifically described previously, although rejection probabilities have been predicted.

This work illustrates the use of joint modelling of mean and dispersion for predicting the probability of rejection in a qualification test. The method is simple: The distribution of standardised residuals is assumed to be independent of the process variables, and the probability is predicted directly using the estimated conditional cumulative distribution function. In chapter 2, joint modelling of the mean and dispersion is reviewed. In chapter 3, the idea for probability prediction is presented. In chapter 4, a class of cumulative distribution functions for distributions resembling the normal distribution is presented. The discussion is mainly based on the frequently rational assumption that the response has a distribution resembling the normal distribution. The models for mean and variance are estimated using the normal distribution assumption, but probabilities are predicted based on a cumulative distribution function of standardised residuals, which is estimated from the data.

2 Joint modelling of mean and dispersion

Prediction accuracy is determined by following components: the prediction model, the variance of the response and the amount of uncertainty about the explanatory variables. The model tends to be less accurate in sparse data areas because of the greater model variance. Dispersion effects are due to the explanatory variables with a direct effect on the residual variance. In this paper, dispersion effects are assumed to be the main cause of the heterogeneity of variance. The residual variance may also depend on the mean. The variation between the values used in prediction and the real values of the explanatory variables decreases the prediction accuracy. For example in product planning, the upcoming values of the explanatory process variables may not be exactly known at the planning phase.

Dispersion modelling has been justified based on its intrinsic interest, the gain of efficiency for mean model estimation and the construction of confidence intervals. Carroll and Ruppert [1] reviewed the methods of dispersion modelling. The purpose of heteroscedastic models has been joint modelling of mean and dispersion. The most widely discussed parametric methods, such as heteroscedastic linear regression, are included in the framework of double-generalised linear models [7]. Non-parametric methods have also been widely used [8]. In quantile regression, the conditional quantile function is modelled. The quantile function is the inverse of the cumulative distribution function, and thus quantile regression is equivalent to modelling the conditional distribution function. Several nonparametric and parametric methods have been proposed for quantile regression [9]. Heteroscedastic models have been widely used for the analysis of industrial quality improvement experiments [4], [8], but seldom for analysing process data. Variance estimation is simpler when variance and mean are independent. A common method to make mean and variance independent is to transform the response [1].

The heteroscedastic linear regression model is written as

$$\begin{aligned} y_j &\sim N(\mu_j, \sigma_j^2) \\ \mu_j &= x_j^{\mathrm{T}} \beta \\ \sigma_j^2 &= g(z_j^{\mathrm{T}} \tau). \end{aligned} \tag{1}$$

Here x_j and z_j are the observations of the explanatory variables and y_j is the jth observation of the response. The vectors x_j and z_j can have common variables. In the literature, the inverse link function $g(x) = \exp(x)$ is commonly used. The estimation of the heteroscedastic linear model can be based on the result that the squared error term is gamma-distributed when the response is normally distributed: $\varepsilon \sim N(0, \sigma^2) \Rightarrow \varepsilon^2 \sim \mathrm{Gamma}(\sigma^2, 2)$. Residual maximum likelihood estimation (REML) has been considered the best method for estimating the parameters of model (1). Smyth et al. [6] described an iterative method for approximating REML. The method is fast and simple, and thus suitable for process data-based modelling. In each iteration, the mean model is estimated by minimising the weighted sum of squares with weights $1/\widehat{\sigma}_j^2$ from the previous variance model estimation. The variance model is estimated by maximising the gamma model log-likelihood related to the model with an expectation value σ_j^2, response $(y_j - \widehat{y}_j)^2/(1-h_j)$ and prior weights $(1-h_j)$ from the mean model estimation. Here, h_j denotes the leverage.

Model (1) predicts the error variance. However, the expected squared prediction error for a new observation $\kappa_j^2 = \mathrm{E}\ (y_j - \widehat{y}_j)^2$ is the sum of error variance and model variance $\mathrm{E}\ (y_j - \mathrm{E}\ y_j)^2 + \mathrm{E}\ (\widehat{y}_j - \mathrm{E}\ y_j)^2$. Thus, an estimate of model variance should be added to the prediction of the variance model (1) to estimate the squared prediction error without bias. In linear regression, the model variance is estimated with $h_j\, \widehat{\sigma}^2$, where $h_j = x_j^{\mathrm{T}} \mathrm{Cov}\ (\widehat{\beta}) x_j$.

Let x denote the realised values of the process variables and x^* the respective planning values, i.e. the values expected on the basis of the process settings applied. If the model (1) is built on x, the predictions are conditional on the assumption that the explanatory variables are known. In that case, the model explains how the response depends on the process variables, which is often more attractive than to explain how the response depends on the planning values. However, only x^* may be available for prediction. The variance increment due to the variation between x and x^* is suggested to be taken into account using the error propagation formula for independent errors, see (2), [4].

As a summary, we propose to estimate the expected squared prediction error of a new observation j with the sum of the variance components

$$\widehat{\kappa}_j^2 = \widehat{\sigma}_j^2 + h_j\,\widehat{\sigma}_j^2 + \sum_i \Big(\frac{\partial\mu_j(\widehat{\beta},x)}{\partial x_{(i)}}\Big)^2_{(x=x_j)} \widehat{\sigma}^2_{x_{(i)}}. \qquad (2)$$

Here, the term $\widehat{\sigma}_j^2 = g(z_j^{\mathrm{T}}\widehat{\tau})$ is the predicted error variance, the term $h_j\,\widehat{\sigma}_j^2$ is the estimate of the model variance, and $\widehat{\sigma}^2_{x_{(i)}}$ is an estimate of the variance of the difference between the planning values and the realised values of the explanatory variable $x_{(i)}$.

3 Probability prediction using heteroscedastic regression

The basic assumption of our method for probability prediction is that the standardised residual $\widehat{\varepsilon}_j$ has a distribution F, which is independent of the explanatory variables.

$$\begin{aligned} \widehat{\varepsilon}_j &= \frac{(y_j - \widehat{y}_j)}{\widehat{\kappa}_j} \\ \widehat{\varepsilon}_j &\sim F \quad \forall j. \end{aligned} \qquad (3)$$

The assumption holds approximately in many cases, and our method is therefore rational. Let l_j and u_j be the minimum and maximum requirements for y_j, and let $A_j = [l_j, u_j]$ denote the acceptance interval of the response. The interesting quantities are the rejection and acceptance probabilities

$$\begin{aligned} P_r &= P(y_j \notin A_j) \\ P_a = 1 - P_r &= P(y_j \in A_j). \end{aligned} \qquad (4)$$

The rejection probability is predicted using the predicted mean, the predicted deviation and the distributional assumption:

$$P_r = F\big(\frac{l_j - \widehat{\mu}_j}{\widehat{\kappa}_j}\big) + \big[1 - F\big(\frac{u_j - \widehat{\mu}_j}{\widehat{\kappa}_j}\big)\big]. \qquad (5)$$

Assumption (3) naturally does not always hold, and its justification should be checked before the method is used. However, the proposed probability prediction method is simple enough to be applied in industrial practice.

Non-parametric methods have more problems with high dimensionality, and we suggest that heteroscedastic linear models are suitable for process data-based dispersion modelling. The selection of the explanatory variables can be carried out stepwise: First, the mean model is specified and the dispersion model is specified using the squared residuals from the mean model and deviance related to gamma distribution. [3].

4 A class of approximations for cumulative distribution functions

In model (1), the error term was assumed to be normally distributed. In practice, there are often small deviations from the normal distribution. In the estimation of model (1), small discrepancies have only a minor impact, but in the prediction of rejection probability (5), the impact may be moderate. Thus, a class of distributions resembling the normal distribution is presented as an alternative for the distribution function F in model (3), and the heteroscedastic model (1) is estimated using the normal distribution assumption.

The empirical distribution function of the standardised residuals is the starting point for distribution function estimation. The aim is to find a simple distribution function that fits well to the empirical distribution. Generally, there should be a formulation to which many empirical distributions fit. The following piecewise formulation for the cumulative distribution function is proposed

$$\begin{aligned} F(t) &= C_l e^{-t\alpha_l}, \text{when} \quad t < l \\ F(t) &= \Phi(\frac{t-\mu}{\sigma_l}), \text{when} \quad t \in [l, \mu] \\ F(t) &= \Phi(\frac{t-\mu}{\sigma_u}), \text{when} \quad t \in [\mu, u] \\ F(t) &= 1 - C_u e^{-t\alpha_u}, \text{when} \quad t > u. \end{aligned} \tag{6}$$

Both tails of the distribution are modelled using the exponential function proposed by Hill [2]. The middle part of the distribution is approximated by using two normal distributions, and $\Phi(t)$ denotes the cumulative distribution function of the standard normal distribution. Our cumulative distribution is therefore described by the parameters $C_l, C_u, \alpha_l > 0, \alpha_u > 0, \mu, \sigma_l > 0$, $\sigma_u > 0$ l and u. For the purpose of probability prediction, it is enough to find parameter values that fit well to the empirical distribution function. As an addition to function (6), a short transitional zone is used around the changing points l and u to make the distribution function continuous and monotone. In the transitional zone, the estimated cumulative distribution function can be, for example, the weighted sum of Hill's approximation and the normal approximation.

5 Application: Strength of steel plates

Steel plates are manufactured in a rolling mill. Strength is the most characteristic mechanical property of steel. The strength of steel plates is controlled by alloying and thermomechanical treatments applied during the process of production. Every steel plate product has its own specifications, including acceptance criteria for strength.

The data were measured at a Rautaruukki steel plate mill. A total of 45 variables were measured, and the number of observations was about 90000. Tensile strength and yield strength were modelled separately using heteroscedastic linear models (1), with about 30 explanatory variables in the mean model and 10 explanatory variables in the variance model. The number of model terms was much higher because of the product terms of the explanatory variables. Square function $\sigma_j^2 = (z_j^{\mathrm{T}}\tau)^2$ was used as the link function of variance because the link $\exp(z_j^{\mathrm{T}}\tau)$ did not fit well to the data. The variance of strength depends clearly on the process settings (Table 1).

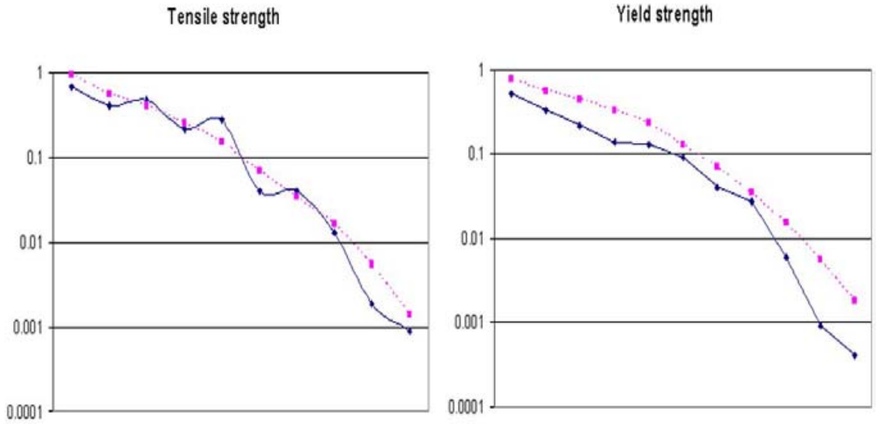
	Yield strength		Tensile strength	
	Mean	Deviation	Mean	Deviation
5% quantile	279	9.2	400	6.2
Median	364	13.2	511	7.9
95 % quantile	441	21.2	553	16.5

Table 1. Descriptive statistics for the predicted mean and for the predicted standard deviation in the data.

The probability of rejection in a single tensile test was modelled using the proposed method (1-6). It seemed that the empirical distribution function of standardised residuals was satisfactory independent of the process variables and the predicted variance. The prediction accuracy of the models was verified in an independent test data set, which was not used for estimation. When a normal distribution was used instead of the proposed cumulative distribution approximation (6), the accuracy of probability prediction decreased only slightly in the test data set. The heteroscedastic linear models were estimated using the real values of process variables, although for some variables, only the planning values can be used in prediction. The error propagation of variances (2) did not improve the probability prediction in the test data. When the estimated model variance was taken into account in the probability prediction, as in (2), the tensile strength rejections were predicted significantly better (the log-likelihood difference was 14), but the prediction of yield strength rejections did not improve. So, in this application, a simplified model $\widehat{\kappa}_j^2 = \widehat{\sigma}_j^2$ was very competitive with model (5) in probability prediction.

The predicted rejection probabilities were consistent with the realised tensile strength rejections (Figure 1a). For yield strength, there were fewer rejections than predicted by the model (Figure 1b). The reason for this seems to arise from bias in the estimated heteroscedastic linear model: The average residual differs very significantly from zero for the plates with a high predicted rejection probability.

The estimated prediction models for strength, deviation of strength and rejection probability were implemented in an Excel-based application with a user interface for choosing the process settings. The application has been used for choosing the process settings for steel plate products. As hundreds of products are manufactured, the definition of settings separately for each product is not an easy task. The users of the application have considered it a very useful tool for the purpose. The efficiency of planning has increased, and it is believed that the number of strength disqualifications can be decreased.



Figures 1a and 1b. The observed proportion of rejections (solid line) in the classes of predicted rejection probability (dashed line) in the test data set for tensile strength (1a) and yield strength (1b).

6 Discussion and conclusion

In this work, a method for estimating rejection probabilities in a qualification test is proposed. The method is based on fitting heteroscedastic models to process data measured from a normal production process. The study shows that information about the variability of variance can be successfully extracted from process data.

In a test application, the method was successfully used for product planning in a steel plate mill. For that problem, the method could have been simplified, but in other applications, the possibility to take into account the imprecision of the explanatory variables, the deviations from normality and the model variance may be useful.

The method is based on the assumption that the distribution of standardised error terms is independent of the explanatory variables. The presented method is suitable for situations where this basic assumption holds sufficiently. Often, however, this assumption may not be realistic, but simplifications are needed to obtain an applicable model for rejection probability prediction.

High dimensionality is a problem for many modelling techniques. Heteroscedastic regression models can handle satisfactorily the dimensionality problem. Thus, a model for the conditional distribution function based on the predicted mean, the predicted variance and the distributional assumptions is a rational approach for process data-based modelling. The problem of process data-based analysis is that process data do not contain much new information. The process is adjusted to settings that are known to be workable. To improve the process the engineers need information about novel settings, which would yield even better results. One way to gain new information is to carry out designed experiments, but process data surely also contain new and useful information. Industrial processes are intentionally developed, which means that the process is changing. As a consequence, the need for model updates is obvious when the models are based on process data. This holds especially for dispersion models.

References

[1] Carroll R.J., Ruppert D. (1988). *Transformation and weighting in regression.* Chapman and Hall, New York.

[2] Hill B.M. (1975). *A simple general approach to inference about the tail of a distribution.* Annals of Statistics **3**, 1163 – 1174.

[3] Juutilainen I., Röning J. (2003). *Heteroscedastic linear models for analysing process data.* WSEAS Transactions on Mathematics **2** 179 – 187.

[4] Myers R.H., Khuri A., Vining G. (1992). *Response suface alternatives to the Taguchi robust parameter design approach.* The American Statistician **46**, 131 – 139.

[5] Ruppert D., Wand M. P., Holst U., Hössjer O. (1997). *Local polynomial variance-function estimation.* Technometrics **39**, 262 – 273.

[6] Smyth G.K., Huele A.V., Verbyla A.P. (2001). *Exact and approximate REML for heteroscedastic regression.* Statistical modelling **1**, 161 – 175.

[7] Smyth G.K., Verbyla A.P. (1999). *Adjusted likelihood methods for modelling dispersion in generalized linear models.* Environmetrics **10**, 696 – 709.

[8] Vining G., Bohn L.L. (1998). *Response surfaces for the mean and variance using a nonparametric approach.* Journal of Quality Technology **30**, 282 – 291.

[9] Yu K., Lu Z., Stander J. (2003). *Quantile regression: applications and current research areas.* Journal of the Royal Statistical Society: Series D **52**, 331 – 350.

Acknowledgement: Thanks to Rautaruukki Steel Plate Mill for practical and financial support

Address: I. Juutilainen, J. Röning, Computer Engineering Laboratory, University of Oulu, PO BOX 4500, FIN-90014, Finland

E-mail: `ilmari.juutilainenr@ee.oulu.fi`

COMPSTAT'2004 Symposium

ESTIMATING $ED50$ USING THE UP-AND-DOWN METHOD

Ene Käärik and Andres Sell

Key words: Up-and-down method, small samples, estimators for ED_{50}, logistic distribution.

COMPSTAT 2004 section: Biostatistics.

Abstract: This paper studies up-and-down method for estimating ED50.

1 Introduction

In some experiments it is not possible to make more than one observation on a given specimen. Once a test has been made the specimen is altered so that the result cannot be obtained from a second test. The technique for obtaining sensitivity data has been developed and used in explosives research in 1943. Because of the specific properties, the procedure is usually called the 'up and down' method [5].

Researchers often deal with continuous variables which cannot be measured in practice. Usually it can be assumed that the probability of positive response increases monotonically with stimulus level. The rule followed is that if the response at the current stimulus level is positive then the next observation is made at some fixed distance d below this level, otherwise it is made at d above.

The up-and-down method is suggested for getting data in order to estimate the median of latent response. Dixon [6] gave approximate maximum likelihood estimates for the parameters of the (normal) response curve, and approximate formulas of the standard errors of these estimates. He pointed out, that using formulas of the asymptotic variance, the up-and-down method was 30 to 40 percent more efficient for estimating median of latent response than the usual probit analysis method.

2 Getting data using the 'up-and-down' method

Suppose that the stimulus is a dosage and the individuals either responds (have positive response) or not, depending on the level of the dosage.

Let x_j $(j = 0, 1, 2, \ldots, n)$ be the dosage used and suppose the starting level in an up-and-down experiment is x_0. Dixon [6] assumed that the latent response is normally distributed with mean μ and variance σ^2, and suggested equal spacing d between doses that approximately equals to σ $(d/\sigma \approx 1)$. Then the first level of the test is $x_1 = x_0 - d$, if the response y is positive (usually denoted $y(x_0) = 1$) or $x_1 = x_0 + d$, if the response y is negative (usually denoted $y(x_0) = 0$).

So we perform the first trial at some dosage level and make trials sequentially. A series of trials is carried out following the rule of an increase in dose in case of negative response and decrease in dose in case of positive response.

In case of a small sample, the result very much depends on starting value x_0, which usually requires some prior knowledge. The first test should be performed at the level as near as possible to the expected dose of 50% positive response (ED_{50}, *median effective dose*).

3 Estimates of ED_{50}

Estimation of the median of the latent response is one of the important statistical problems in bioassay. The well-known sequential method for estimating is the up-and-down rule.

Several formulas for estimating the *median effective dose* ED_{50} have been proposed by Dixon [6] and Brownlee, Hodgs and Rosenblatt [1] based on the maximum likelihood approach. Hsi [7] proposed a new sampling procedures based on multiple samples. Tsutakawa [13] studied the up-and-down procedure and proved that maximum likelihood estimators are asymptotically unbiased and have an asymptotic normal distribution.

The estimator proposed by Dixon [6] for a small sample tests is

$$\widehat{ED_{50}} = x_f + k \cdot d, \tag{1}$$

where x_f is the final level used in testing and k is obtained from special table (see [6], p. 968) based on the maximum likelihood analysis for each possible configuration of responses assuming normal distribution. The sample sizes for Dixon's k-table vary from 2 to 6.

The following formula is recommended for sample sizes $n > 6$:

$$\widehat{ED_{50}} = \frac{\sum x_i}{n} + \frac{d}{n}(A + C), \tag{2}$$

where the mean of the test level is corrected by a factor, which depends on the constants A and C given in the Table 2 (see [6], p. 970) and assigned by test series and numbers of positive and negative response in the final trial.

Brownlee *et al* [1] indicated that the up-and-down strategy provides efficient estimation of median responses even for small samples when the uniform spacing of dose levels d is between $2\sigma/3$ and $3\sigma/2$. They noted also that, in this case, the median response estimator is relatively robust to d and starting level x_0.

An alternative estimator was suggested by Wetherill [15] and modified by Choi [4] based only on the peaks and troughs of the response series. Let $t_1, t_2, \ldots$ be the dose levels at the turning points. Define the following points

$$w_i = \begin{cases} t_i + d/2, & \text{if } \ t_i \text{ is a trough,} \\ t_i - d/2, & \text{if } t_i \text{ is a peak.} \end{cases}$$

Then the estimator (Choi, 1971) based on k modified turning points is

$$\widetilde{ED_{50}} = \sum_{i=2}^{k} \frac{w_i}{k-1}. \tag{3}$$

Choi [3] has shown that the bias of $\widetilde{ED_{50}}$ is smaller that $\widehat{ED_{50}}$.

Suggested standard error of Dixon's ED_{50} estimator [6] is approximately equal to

$$\sigma\sqrt{\frac{2}{n}}. \tag{4}$$

Little [10] preferred corresponding standard error approximately equal to be $\sigma\sqrt{\frac{2}{n+1}}$. Kershawa [8], Tsutakawa [13] and Choi [4] presented asymptotic confidence interval for ED_{50}.

An additional serious problem occurs when looking into estimation the variance σ of underlying tolerance distribution. The difficulty of estimation of σ arises from the fact that observations from an up-and-down experiment are not independent. So, the variation among test is related to the individual variation and a complex method for estimation is necessary.

In principle we can test the null hypothesis that σ lies in appointed (fixed) interval, using some nonparametric method like run-test for a single observation series.

Brownlee et al ([1]) required that the parameter σ is known within rough limits and assumed that estimates of the median (or the mean) of the response are not sensitive to errors in the guessed value of σ.

Williams ([16]) assumed the logistic tolerance distribution and constructed the confidence interval from the likelihood ratio criterion instead of maximum likelihood method.

Choi ([3]) focused on serious problems of estimation of the variance σ and have proposed a new method for constructing the confidence interval for the ED_{50} of the normal tolerance distribution. They presented an approximation of the covariance matrix for turning points.

Chao and Fuh ([2]) developed a more appropriate method to estimate σ that takes into the account the special dependent data structure, the authors have also presented two bootstrap procedures for estimating confidence interval.

Kershawa [9] studied small sample properties of several estimators of ED_{50} and by his research no estimator seems to have particular advantages compared to others.

4 Logit analysis

One of the appropriate methods for analyzing the data achieved by the up-and-down procedure is *logit* analysis. It is natural to use the following latent variable model here. Suppose there is an unobservable continuous random variable Y^* (*latent* response) such that binary random variable Y takes the value one if and only if Y^* exceeds an certain thresholds θ. In biometrics the latent variable Y^* is often the influence of the dose and Y is the outcome. The positive outcome occurs only when the latent response exceeds the threshold, so we can write the probability p of a positive outcome as $p = P(Y = 1) = P(Y^* > \theta)$. Suppose now, that the outcome depends on a covariate X. To model this dependence we can write $Y^* = X^T\beta + U$, where β is the vector of coefficients and U is the error term with the distribution function $F(u)$. The obvious choice of an error distribution is the normal, and then we get *probit* model, which has already been used before Dixon's up-and-down method. An alternative to the normal is the logistic distribution that has the advantage of a closed form expression, and the inverse transformation of which is *logit*.

We focused on estimating ED_{50} as a median point of the latent response curve assuming logistic distribution. Let the quantity (dosage) X be the level at which the observation is taken, so $p(X) = P(Y(X) = 1)$. During the experiment there are several quantities x_i and responses y_i. Let p_i denote the probability of positive response at the ith level. According to logistic distributions, the probability of positive response is expressed in following way

$$p_i = [1 + \exp(-(\alpha + \beta x_i))]^{-1}. \tag{5}$$

and the *logit*-model is the following

$$\ln \frac{p_i}{1 - p_i} = \alpha + \beta x_i$$

Modifications based on logistic model were proposed, for example, by Wetherill *et al* [15] and Wu [14].

Little [10] has compared logistic and normal distribution approaches and has indicated, that the normal distribution provides the smaller mean squared error of median response only for very small sample sizes and relatively bad starting point. For relatively good starting points (close to the median), the logistic distribution provides slightly smaller mean squared error of median response.

The main disadvantages of logit model compared to the up-and-down method are as follows.

- The estimates of the parameters in the logit model may be poor, especially when the sample size is small.

- The up-and-down estimates are easy to compute and exist always, whereas the logit estimate is computed iteratively and may not exist.

5 Clinical trial

5.1 Data

The aim of the study was to determine the minimum effective local anaesthetic dose (MLAD) of isobaric levobupivacaine and ropivacaine administered via a spinal catheter for hip replacement surgery. MLAD is defined as minimum effective dose of a local anaesthetic at which 50% of individuals will satisfy the certain anaesthetic criteria (have positive response value), what means, in this case, $MLAD = ED_{50}$.

We used the up-and-down allocation technique to assess the minimum effective local anaesthetic dose of levobupivacaine and ropivacaine.

41 patients were randomly allocated to one of the two spinal solutions in a double-blind manner. There are 2 treatment groups: Ropi-group ($n = 20$) and Levo-group ($n = 21$). Spacing between dosage level was 1 mg and starting dose for Ropi-group was 14 mg and for Levo-group 12 mg.

The dose of local anaesthetic was determined by the response of the previous patient: the effective dose resulted in a 1 mg decrease in the dose of levobupivacaine or ropivacaine, an ineffective dose resulted in a 1 mg increase. Both treatment groups are homogeneous with respect to age, weight, height and sex of individuals. There were no side effects on response, and no difference between two treatment groups in median expected values were caused by treatments.

Testing normality of dose we got $p > 0.15$ in both treatment group.

5.2 Results

Three methods were used to calculate the ED_{50}: (2) for $\widehat{ED_{50}}$ by Dixon, (3) for $\widetilde{ED_{50}}$ by Wetherill and logit analysis (used only the peaks and troughs and inverse transformation from (5) to get estimate $ED_{50}*$). For Dixon's and Wetherill's estimates the SAS/IML program and for logit model the SAS procedure *Probit* and *Logistic* were used. The results are given in Table 1.

Group	$\widehat{ED_{50}}$	$\widetilde{ED_{50}}$	$ED_{50}*$
Levo	11.7	11.7	11.4
Ropi	12.8	13	12.7

Table 1: Estimates of ED_{50}.

The results of different estimators are similar, what confirms their reliability.

For estimating the standard error of ED_{50} the Dixon's approximate formula (4) was used. The reason of choosing Dixon's confidence intervals was the simplicity of the calculations. Based on empirical knowledge we assumed σ of underlying tolerance distribution to be equal one.

We got the 95% confidence interval (11.13, 12.37) in Levo-group and (12.16, 13.44) in Ropi-group. As the calculated confidence limits overlapped, we could not prove statistically significant difference between the minimum effective local anaesthetic dose (MLAD) of two treatments.

6 Concluding remarks

The up-and-down method allows to get relatively good estimates for median expected dose ED_{50} in the case of small and very small sample sizes.

This method is simple to implement but it can cause serious problems with bias of ED_{50} estimator, depending on the starting level of dosage and assumed value of σ, which reflects the variability of the response across doses.

The *logit* model for estimating ED_{50} may be preferred because of existence standard statistical software.

The observations in up-and-down experiments can be handled as correlated observations, when the next level depends on the previous result. Such experimental plan reduces the dispersion of observations.

Most of the estimators of ED_{50} require uniform spacing between stimulus levels. This may be an obstruction to get good estimates, especially when the starting point and estimated σ are not good. It seems to be reasonable to change spacing during the experiment using currently the information from the obtained results, but here we need additional simulation studies.

References

[1] Brownlee K.A., Hodges J.L., Rosenblatt Jr. M. (1953). *The up-and-down method with small samples.* JASA **48**, 262 – 277.

[2] Chao M.T., Fuh C.D. (2001). *Bootstrap methods for the up and down test on pyrotechnics sensitivity analysis.* Statistica Sinica **11**, 1 – 21.

[3] Choi S.C. (1971). *An investigation of Wetherill's method of estimation for the up-and-down experiment.* Biometrics **27**, 961 – 970.

[4] Choi C.S. (1990). *Interval estimation of the LD_{50} and up-and-down experiment.* Biometrics **46**, 485 – 492.

[5] Dixon W.J., Mood A.M. (1948). *A method for obtaining and analyzing sensitivity data.* JASA **43**, 109 – 126.

[6] Dixon W.J. (1965). *The up-and-down method for small samples.* JASA **68**, 967 – 978.

[7] Hsi B.P. (1969). *The multiple sample up-and-down method in bioassay.* JASA **64**, 147 – 162.

[8] Kershaw C.D. (1985). *Asymptotic properties of $\bar{w}$, an estimator of the ED50 suggested for use in up-and-down experiments in bio-assay.* Annals of Statistics **13**, 85 – 94.

[9] Kershaw C.D. (1987). *A comparison of estimators of the ED_{50} in up-and-down experiments.* Journal of Statistical Computation and Simulation **27**, 175 – 184.

[10] Little R.E. (1974). *A mean square error comparison of certain median response estimates for the up-and-down method with small samples.* JASA **69**, 202 – 206.

[11] Little R.E. (1974). *The up-and-down method for small samples with extreme value response distributions.* JASA **69**, 803 – 806.

[12] Little R.E. (1975). *The up-and-down method for small samples with two specimens "in series".* JASA **70**, 846 – 851.

[13] Tsutakawa R. K. (1967). *Asymptotic properties of the block up-and-down method in bio-assay.* The Annals of Mathematical Statistics **38**, 1822 – 1828.

[14] Wu C.F.J. (1985). *Efficient sequential designs with binary data.* JASA **80**, 974 – 984.

[15] Wetherill G.B., Chen H., Vasudeva R.B. (1966). *Sequential estimation of quantal response curves: a new method of estimation.* Biometrika **53**, 493 – 454.

[16] Williams D.A. (1986). *Interval estimation of the median lethal dose.* Biometrics **42**, 641 – 645.

Acknowledgement: The authors are grateful to Prof. Ene-Margit Tiit from Institute of Mathematical Statistics, University of Tartu, to Prof. Riku Aantaa, to Prof. Klaus Olkkola from Department of Anaesthesiology of Turku University Hospital and to statistician Hans Helenius from Department of Biostatistics of Turku University for helpful hints and suggestions.

Address: E. Käärik, Institute of Mathematical Statistics, University of Tartu, J. Liivi 2-516, Tartu 50409 , Estonia
A. Sell, Anaesthesiology and Intensive Care Clinic of Tartu University Hospital, L. Puusepa 8, Tartu 51014, Estonia

E-mail: Ene.Kaarik@ut.ee, Andres.Sell@kliinikum.ee

DURBIN-WATSON TEST FOR LEAST WEIGHTED SQUARES

Jan Kalina

Key words: Durbin-Watson test, weighted least squares, robust regression, least weighted squares.

COMPSTAT 2004 section: Robustness.

Abstract: This paper studies the Durbin-Watson statistic for residuals of least weighted squares regression. We stress it is possible to compute the exact p-value of an asymptotic equivalent of the statistic. This can be also approximated by the statistic for ordinary least squares. To show the contrast with weighted least squares, we start by describing the exact Durbin-Watson test for this context.

1 Introduction and notation

In the whole paper we consider the regression model

$$Y_t = \beta_1 X_{t1} + \cdots + \beta_r X_{tr} + e_t, \quad t = 1, \ldots, n, \tag{1}$$

which can be rewritten in the usual matrix notation as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. An intercept may (but does not have to) be included in the model. We assume that data are observed as a time series in equidistant time intervals.

Now we can introduce some general notation, which will be used throughout the whole paper.

- R denotes the set of all real numbers.
- An indicator of a random event A is denoted by $I[A]$.
- $\mathcal{I}_n$ denotes a unit matrix of size $n \times n$.
- $\overset{\mathcal{D}}{=}$ denotes equality in distribution.

2 Durbin-Watson test for weighted regression

First we describe the Durbin-Watson test for weighted least squares and show how to search for the exact p-value. We have forsaken the approach of Durbin and Watson [2], who derive lower and upper bounds for the critical value. In fact the *exact* p-value can be estimated by simulations. Moreover, we do not restrict ourselves to the limiting assumptions of an intercept in (1).

For the model (1), we assume that the vector of disturbances (errors) $\mathbf{e}$ follows the multivariate normal distribution $\mathbf{e} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathbf{W}^{-1})$, where $0 < \sigma^2 < \infty$, $\mathbf{W}^{-1}$ is the inverse of the diagonal matrix $\mathbf{W} = \mathsf{Diag}\{w_1, \ldots, w_n\}$

and $w_1, \ldots, w_n$ are *positive* weights. The (classical) weighted least squares estimator is equal to

$$\mathbf{b}_W = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}.$$

Multiplying (1) by $\sqrt{w_t}$ for each $t = 1, \ldots, n$ leads to an equivalent model

$$\sqrt{w_t}\, Y_t = \sqrt{w_t}\, \beta_1 X_{t1} + \cdots + \sqrt{w_t}\, \beta_r X_{tr} + \sqrt{w_t}\, e_t, \quad t = 1, \ldots, n, \qquad (2)$$

with the same parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^T$. Because $\sqrt{w_t}\, e_t \sim \mathsf{N}(0, \sigma^2)$ for each t, ordinary least squares can be used to estimate $\beta_1, \ldots, \beta_r$. For ordinary least squares (OLS), we will denote the estimator of $\boldsymbol{\beta}$ by $\mathbf{b} = (b_1, \ldots, b_r)^T$ and its residuals by

$$u_t = Y_t - b_1 X_{t1} - \cdots - b_r X_{tr}, \quad t = 1, \ldots, n.$$

But the Durbin-Watson test for weighted regression should be based on residuals of (2), namely $\mathbf{u}^* = (u_1^*, \ldots, u_n^*)^T$, where $u_t^* = \sqrt{w_t}\, u_t$ for each t. For the Durbin-Watson test statistic

$$d = \frac{\sum_{t=2}^n (u_t^* - u_{t-1}^*)^2}{\sum_{t=1}^n u_t^{*2}},$$

it is necessary to determine the probability distribution under the null hypothesis H_0 : disturbances are independent.

We denote $\mathbf{M} = \mathcal{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, $\mathbf{V} = \mathsf{Diag}\{\sqrt{w_1}, \sqrt{w_2}, \ldots, \sqrt{w_n}\}$ and

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix},$$

where blank spaces represent zeros.

Thanks to the assumptions, it holds under H_0 that

$$d = \frac{\mathbf{u}^{*T}\mathbf{A}\mathbf{u}^*}{\mathbf{u}^{*T}\mathbf{u}^*} = \frac{\mathbf{u}^T\mathbf{V}\mathbf{A}\mathbf{V}\mathbf{u}}{\mathbf{u}^T\mathbf{W}\mathbf{u}} \overset{\mathcal{D}}{=} \frac{\mathbf{e}^T\mathbf{M}\mathbf{V}\mathbf{A}\mathbf{V}\mathbf{M}\mathbf{e}}{\mathbf{e}^T\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{e}} \overset{\mathcal{D}}{=} \frac{\mathbf{E}^T\mathbf{V}^{-1}\mathbf{M}\mathbf{V}\mathbf{A}\mathbf{V}\mathbf{M}\mathbf{V}^{-1}\mathbf{E}}{\mathbf{E}^T\mathbf{V}^{-1}\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{V}^{-1}\mathbf{E}},$$

where $\mathbf{E} = (E_1, \ldots, E_n)^T \sim \mathsf{N}(\mathbf{0}, \mathcal{I}_n)$. The exact p-value of the test is equal to the probability

$$\mathsf{P}\left[\frac{\mathbf{E}^T\mathbf{V}^{-1}\mathbf{M}\mathbf{V}\mathbf{A}\mathbf{V}\mathbf{M}\mathbf{V}^{-1}\mathbf{E}}{\mathbf{E}^T\mathbf{V}^{-1}\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{V}^{-1}\mathbf{E}} \leq d\right],$$

where d is the value of the statistic computed from given data. This (theoretical) probability can be estimated by its empirical counterpart (relative frequency of the event) by random generating of $\mathbf{E}$.

Results of this section hold also for ordinary least squares, which is a special case of weighted regression.

3 Least weighted squares

This section defines the least weighted squares regression and takes over Plát's [5] asymptotic representation, which will be used in Section 4.

The least weighted squares estimator is a robust regression method with a high breakdown point, proposed by Víšek [7]. For the definition we need the following notation and the concept of a weight function.

For the model (1), let us consider (any) estimate $\mathbf{b} = (b_1, \dots, b_r)^T \in \mathsf{R}^r$ of the parameter $\boldsymbol{\beta}$. By

$$u_t(\mathbf{b}) = Y_t - b_1 X_{t1} - b_2 X_{t2} - \cdots - b_r X_{tr}, \quad t = 1, \dots, n$$

we denote the residual corresponding to the t-th observation. Let us order the squared residuals

$$u^2_{(1)}(\mathbf{b}) \leq u^2_{(2)}(\mathbf{b}) \leq \cdots \leq u^2_{(n)}(\mathbf{b}).$$

Definition 1: Let the function $\psi : [0,1] \to [0,1]$ be nonincreasing and continuous on $[0,1], \psi(0) = 1$ and $\psi(1) = 0$. Moreover, we assume that both one-sided derivatives of ψ exist in all points of $(0,1)$, are bounded by a common constant and we assume the existence of a finite left derivative in 0 and right in point 1. Then the function ψ is called a *weight function*, numbers

$$v_\ell = \left[\psi\left(\frac{\ell - 1}{n}\right) - \psi\left(\frac{\ell}{n}\right)\right], \quad \ell = 1, 2, \dots, n,$$

are called *weights* defined by the weight function ψ.

Definition 2: Let $\mathcal{K} \in \mathsf{R}^r$ denote a compact set, let the true regression parameter $\boldsymbol{\beta}^0$ fulfill $\boldsymbol{\beta}^0 \in \mathcal{K}^0$, where $\mathcal{K}^0$ denotes the interior of $\mathcal{K}$, let ψ be a weight function. Then

$$\mathbf{b}_{LWS} = \arg\min_{\mathbf{b} \in \mathcal{K}} \sum_{k=1}^{n} \psi\left(\frac{k-1}{n}\right) u^2_{(k)}(\mathbf{b})$$

is called the *least weighted squares* (LWS) estimator of $\boldsymbol{\beta}$.

Remark: Plát [4] explains the equivalence of the definition 2 with

$$\mathbf{b}_{LWS} = \arg\min_{\mathbf{b} \in \mathcal{K}} \sum_{k=1}^{n} w_k u^2_{(k)}(\mathbf{b}),$$

which explains the difference from weighted least squares.

Computational aspects of the LWS are studied by Čížek [1], who proposed a procedure that adaptively chooses the weights and thus controls the balance between robustness and efficiency of the estimator. Kalina [3] gives an argument that an available algorithm (analogy of an algorithm for least trimmed squares) gives a tight approximation to the correct value of the estimate.

Assumptions $\mathcal{A}$:

- $\{e_t\}_{t=1}^{\infty}$ is a sequence of independent random variables, where for each t it holds $e_t \sim \mathsf{N}(0, \sigma^2)$ with $0 < \sigma^2 < \infty$
- $\{\mathbf{X}_t\}_{t=1}^{\infty} = \{\mathbf{X}_{t1}, \ldots, \mathbf{X}_{tr}\}_{t=1}^{\infty}$ is a sequence of fixed (nonrandom) vectors, which satisfy

$$\sum_{t=1}^{n} \|\mathbf{X}_t\| = \mathcal{O}(n) \quad \text{as } n \to \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbf{X}_t \mathbf{X}_t^T = \mathbf{Q},$$

where $\mathbf{Q}$ is a regular matrix of size $r \times r$.

We introduce the notation $g(.)$ for the density of $\mathsf{N}(0, \sigma^2)$ distribution and z_α for its quantile in the form

$$z_\alpha = \sigma \cdot \Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \quad \alpha \in (0,1).$$

Theorem 1: Let Assumptions A hold, let $\mathbf{b}_{LWS}$ be defined according to definition 2, let the weights $w_1, \ldots, w_n$ be defined by some weight function ψ for every integer n. Then $\sqrt{n}\,(\mathbf{b}_{LWS} - \boldsymbol{\beta}^0) = \mathcal{O}_P(1)$ as $n \to \infty$ and

$$\sqrt{n}\,(\mathbf{b}_{LWS} - \boldsymbol{\beta}^0) = \frac{\gamma}{\sqrt{n}} \mathbf{Q}^{-1} \sum_{\ell=1}^{n} v_l \sum_{t=1}^{n} e_t \mathbf{X}_t I\left[e_t^2 \leq z^2_{1-\frac{\ell}{n}}\right] + \boldsymbol{\eta},$$

where coordinates of $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)^T$ are of order $o_P(1)$ and

$$\frac{1}{\gamma} = -\sum_{\ell=1}^{n} v_\ell \left[\frac{\ell}{n} - 2 z_{1-\frac{\ell}{n}} g(z_{1-\frac{\ell}{n}})\right].$$

Proof: a consequence of Plát [5], who considers weaker assumptions.

4 Durbin-Watson test for least weighted squares

We propose a test of independence of disturbances for the least weighted squares regression. We use the asymptotic representation to study the asymptotic distribution of the test statistic under the null hypothesis.

Assumptions $\mathcal{A}$ are assumed in the whole section. Replacing $\mathbf{Q}^{-1}$ by $n(\mathbf{X}^T\mathbf{X})^{-1}$ in the statement of Theorem 1, residuals of LWS regression can be expressed as $\tilde{\mathbf{u}} = \mathbf{Y} - \mathbf{X}\mathbf{b}_{LWS} =$

$$= \mathbf{e} - \gamma \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \sum_{\ell=1}^{n} v_\ell \sum_{t=1}^{n} e_t \mathbf{X}_t I\left[e_t^2 \leq z^2_{1-\frac{\ell}{n}}\right] - \frac{1}{\sqrt{n}} \mathbf{X}\boldsymbol{\eta}. \tag{3}$$

Denoting $\boldsymbol{\tau} = -\frac{1}{\sqrt{n}}\mathbf{X}\boldsymbol{\eta}$ and $\boldsymbol{\kappa}$ by $\boldsymbol{\kappa} = \tilde{\mathbf{u}} - \boldsymbol{\tau}$, the Durbin-Watson statistic for LWS residuals equals

$$d_{LWS} = \frac{\sum_{t=2}^{n}(\tilde{u}_t - \tilde{u}_{t-1})^2}{\sum_{t=1}^{n}\tilde{u}_t^2} = \frac{\tilde{\mathbf{u}}^T\mathbf{A}\tilde{\mathbf{u}}}{\tilde{\mathbf{u}}^T\tilde{\mathbf{u}}} = \frac{\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa} + 2\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\tau} + \boldsymbol{\tau}^T\mathbf{A}\boldsymbol{\tau}}{\boldsymbol{\kappa}^T\boldsymbol{\kappa} + 2\boldsymbol{\kappa}^T\boldsymbol{\tau} + \boldsymbol{\tau}^T\boldsymbol{\tau}}. \quad (4)$$

We now examine the order of various terms under the hypothesis of independence. The vector $\mathbf{e}$ has coordinates of order $\mathcal{O}_P(1)$. Because $\sqrt{n}(\mathbf{b}_{LWS} - \boldsymbol{\beta}^0) = \mathcal{O}_P(1)$, it follows $\mathbf{b}_{LWS} - \boldsymbol{\beta}^0 = \mathcal{O}_P(n^{-1/2})$, so the second term of (3) has order $\mathcal{O}_P(n^{-1/2})$ and the third $o_P(n^{-1/2})$. Thus we have proven the asymptotic equivalence of d_{LWS} and $\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa}/\boldsymbol{\kappa}^T\boldsymbol{\kappa}$ under the hypothesis of independence, because other terms in (4) are negligible in probability with respect to the leading terms.

Theorem 2: Under the hypothesis of independence, $\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa}/\boldsymbol{\kappa}^T\boldsymbol{\kappa}$ is invariant with respect to σ^2.

Proof: Scale-invariance of γ can be verified easily. From the scale-equivariance of $\boldsymbol{\kappa}$ follows the scale-invariance of the ratio.

The (exact) p-value of the asymptotic test is equal to the probability

$$\mathsf{P}\left[\frac{\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa}}{\boldsymbol{\kappa}^T\boldsymbol{\kappa}} \leq d_{LWS}\right],$$

where d_{LWS} is the value of the statistic computed from given data. This probability can be estimated by random generating of vectors $(E_1, \ldots, E_n)^T \sim \mathsf{N}(\mathbf{0}, \mathcal{I}_n)$.

We now approach to approximating the asymptotic distribution of the test statistic. Let us denote $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, $\boldsymbol{\varphi} = \mathbf{M}\mathbf{e}$,

$$\boldsymbol{\Psi}(\mathbf{e}) = \left(e_1\sum_{\ell=1}^{n} v_\ell I\left[e_1^2 \leq z^2_{1-\frac{\ell}{n}}\right], \ \ldots \ , e_n\sum_{\ell=1}^{n} v_\ell I\left[e_n^2 \leq z^2_{1-\frac{\ell}{n}}\right]\right)^T$$

and

$$\boldsymbol{\phi} = \mathbf{H}\mathbf{e} - \gamma\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Psi}(\mathbf{e}) = \mathbf{H}\left[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})\right].$$

It holds $\boldsymbol{\kappa} = \boldsymbol{\varphi} + \boldsymbol{\phi}$ and $\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa}/\boldsymbol{\kappa}^T\boldsymbol{\kappa} =$

$$= \frac{\mathbf{e}^T\mathbf{MAMe} + 2\mathbf{e}^T\mathbf{MAH}[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})] + [\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})]^T\mathbf{HAH}[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})]}{\mathbf{e}^T\mathbf{Me} + [\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})]^T\mathbf{H}[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})]}, \quad (5)$$

just like in Víšek [8], [9], who used this for the least trimmed squares and M-estimators. Here the steps can be repeated, the only difference is the definition of $\boldsymbol{\Psi}(\mathbf{e})$.

There exists an orthogonal matrix $\mathbf{L}$ of size $n \times n$ such that $\mathbf{L}^T\mathbf{M}\mathbf{L} = \mathbf{D}$, where $\mathbf{D}$ is the diagonal matrix with eigenvalues of $\mathbf{M}$ as diagonal elements. Without loss of generality, in this context we always assume eigenvalues in a nondecreasing order. Let us partition the matrix $\mathbf{L}^T\mathbf{A}\mathbf{L}$ as

$$\begin{pmatrix} \mathbf{B}_1 & \mathbf{B}_3 \\ \mathbf{B}_2 & \mathbf{B}_4 \end{pmatrix}$$

so that $\mathbf{B}_1$ has size $n - r \times n - r$ and $\mathbf{B}_4$ has size $r \times r$. Let $\mathbf{N}_1$ and $\mathbf{N}_2$ be orthogonal matrices diagonalizing $\mathbf{B}_1$ and $\mathbf{B}_4$ respectively, which means

$$\mathbf{N}_1^T\mathbf{B}_1\mathbf{N}_1 = \mathsf{Diag}\{\nu_1, \nu_2, \ldots, \nu_{n-r}\}$$

and
$$\mathbf{N}_2^T\mathbf{B}_4\mathbf{N}_2 = \mathsf{Diag}\{\nu_{n-r+1}, \nu_{n-r+2}, \ldots, \nu_n\},$$

where $\nu_1 \leq \nu_2 \leq \cdots \leq \nu_{n-r}$ are eigenvalues of $\mathbf{B}_1$ and $\nu_{n-r+1} \leq \cdots \leq \nu_n$ are eigenvalues of $\mathbf{B}_4$. Let $\mathbf{N}$ denote

$$\mathbf{N} = \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 \end{pmatrix},$$

let us put

$$\tilde{\boldsymbol{\xi}} = \mathbf{N}^T\mathbf{L}^T\mathbf{e},\ \boldsymbol{\vartheta} = \mathbf{L}^T\mathbf{A}\mathbf{e},\ \tilde{\boldsymbol{\vartheta}} = \mathbf{L}^T\left[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})\right],\ \boldsymbol{\zeta} = \mathbf{N}^T\mathbf{L}^T\left[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})\right].$$

The notation for coordinates of these vectors will be obvious, for example $\tilde{\boldsymbol{\xi}} = (\xi_1, \ldots, \xi_n)^T$.

We can express (5) in an alternative way, using the new notation. Víšek [8] has done this, but we do not agree with one of his expressions. We formulate this as a lemma.

Lemma 1: Using the notation introduced above, it holds

$$\mathbf{e}^T\mathbf{M}\mathbf{A}\mathbf{M}\left[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})\right] = \sum_{t=n-r+1}^{n} \vartheta_t\tilde{\vartheta}_t - \sum_{t=n-r+1}^{n} \nu_t\tilde{\xi}_t\zeta_t.$$

Proof: From

$$\mathbf{N}^T\mathbf{L}^T\mathbf{H}\mathbf{L}\mathbf{N} = \mathbf{N}^T\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_r \end{pmatrix}\mathbf{N} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_r \end{pmatrix},$$

it follows

$$\mathbf{N}^T\mathbf{L}^T\mathbf{H}\mathbf{A}\mathbf{H}\mathbf{L}\mathbf{N} = (\mathbf{N}^T\mathbf{L}^T\mathbf{H}\mathbf{L}\mathbf{N})\mathbf{N}^T\mathbf{L}^T\mathbf{A}\mathbf{L}\mathbf{N}(\mathbf{N}^T\mathbf{L}^T\mathbf{H}\mathbf{L}\mathbf{N}) =$$

$$= \mathsf{Diag}\{0, 0, \ldots, 0, \nu_{n-r+1}, \nu_{n-r+2}, \ldots, \nu_n\},$$

so we arrive at

$$\mathbf{e}^T\mathbf{M}\mathbf{A}\mathbf{H}\left[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})\right] = \mathbf{e}^T\mathbf{A}\mathbf{H}\left[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})\right] - \mathbf{e}^T\mathbf{H}\mathbf{A}\mathbf{H}\left[\mathbf{e} - \gamma\boldsymbol{\Psi}(\mathbf{e})\right],$$

which leads to the statement of the lemma.

Lemma 2: $\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa}/\boldsymbol{\kappa}^T\boldsymbol{\kappa} =$

$$= \frac{\sum_{t=1}^{n-r}\nu_t\tilde{\xi}_t^2 + 2\sum_{t=n-r+1}^{n}\vartheta_t\tilde{\vartheta}_t - 2\sum_{t=n-r+1}^{n}\nu_t\tilde{\xi}_t\zeta_t + \sum_{t=n-r+1}^{n}\nu_t\zeta_t^2}{\sum_{t=1}^{n-r}\tilde{\xi}_t^2 + \sum_{t=n-r+1}^{n}\zeta_t^2}. \tag{6}$$

Proof: The steps from (5) to the right hand of (6) can be found in Víšek [8], [9], but lemma 1 should be applied.

Theorem 3: Under the hypothesis of independence, $\frac{\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa}}{\boldsymbol{\kappa}^T\boldsymbol{\kappa}}$ is asymptotically equivalent with

$$\frac{\sum_{t=1}^{n-r}\nu_t\tilde{\xi}_t^2}{\sum_{t=1}^{n-r}\tilde{\xi}_t^2}.$$

Proof: From the lemma of Durbin and Watson [2], it follows that numbers $\nu_1,\ldots,\nu_{n-r}$ are bounded between 0 and 4. All coordinates of $\tilde{\boldsymbol{\xi}}$ are bounded in probability uniformly with respect to n, so that for each $\epsilon > 0$ there is $K > 0$ such that for each integer n

$$\mathsf{P}\left[\left|\tilde{\xi}_t\right| > K\right] < \epsilon \quad \text{for each } t = 1,\ldots,n.$$

We get

$$n^{-1/2}\sum_{t=1}^{n-r}\nu_t\tilde{\xi}_t^2 = \mathcal{O}_P(1) \ \text{ as } n\to\infty \quad \text{and} \quad \sum_{t=1}^{n-r}\nu_t\tilde{\xi}_t^2 \xrightarrow{\mathsf{P}} \infty$$

from Lindeberg-Feller central limit theorem; see for example Serfling [6, Chapter 1.9]. The other sums in the numerator contain for each n only r elements, so they are bounded in probability and negligible with respect to the first term. Similar reasoning for the denominator leads to the asymptotic equivalence.

Theorem 4: Under the hypothesis of independence, d_{LWS} is asymptotically equivalent with $\mathbf{e}^T\mathbf{MAMe}/\mathbf{e}^T\mathbf{Me}$.

Proof: follows immediately from the asymptotic equivalence of d_{LWS} and $\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa}/\boldsymbol{\kappa}^T\boldsymbol{\kappa}$, from (6) and Theorem 3.

$\mathbf{e}^T\mathbf{MAMe}/\mathbf{e}^T\mathbf{Me}$ is exactly the Durbin-Watson statistic for OLS (see Section 2). Tables of lower and upper bounds for critical values can be used, as well as simulations for approximations the exact p-value.

Víšek [8] inspected the magnitude of particular sums of (6) and compared $\mathbf{e}^T\mathbf{MAMe}/\mathbf{e}^T\mathbf{Me}$ with $\boldsymbol{\kappa}^T\mathbf{A}\boldsymbol{\kappa}/\boldsymbol{\kappa}^T\boldsymbol{\kappa}$ for the least trimmed squares. Such approximation turns out to be good already for moderate sample sizes. Unfortunately it does *not* give any evidence about the relation between the classical Durbin-Watson statistic and the statistic computed from LTS residuals. Anyway, $\boldsymbol{\kappa}^T \mathbf{A} \boldsymbol{\kappa} / \boldsymbol{\kappa}^T \boldsymbol{\kappa}$ can be approximated by simulations for *both* least trimmed squares and least weighted squares.

References

[1] Čížek P. (2001). *Essays on robust estimation in econometrics.* Dissertation, CERGE UK, Praha.

[2] Durbin J., Watson G.S. (1950, 1951). *Testing for serial correlation in least squares regression I, II.* Biometrika **37, 38**, 409 – 428, 159 – 178.

[3] Kalina J. (2003). *Autocorrelated disturbances of robust regression.* In Fournier B. et al. (eds.): Proceedings EYSM 2003. Ovronnaz, Switzerland. ISBN 3-908152-17-8.

[4] Plát P. (2004). *Konzistence odhadu metodou nejmenších vážených čtverců.* Submitted to Bulletin of the Czech Econometric Society. (Consistency of the least weighted squares estimator. In Czech.)

[5] Plát P. (2004). *Odhad metodou nejmenších vážených čtverců. $\sqrt{n}$-konzistence a asymptotická reprezentace.* Submitted to Bulletin of the Czech Econometric Society. (The least weighted squares estimator. $\sqrt{n}$-consistency and asymptotic normality. In Czech.)

[6] Serfling R.J. (1980). *Approximation theorems of mathematical statistics.* Wiley, New York.

[7] Víšek J.Á. (2001). *Regression with high breakdown point.* Proceedings of Robust 2000 (at Nečtiny), 324-356. JČMF and Česká statistická společnost (Czech Statistical Society).

[8] Víšek J.Á. (2001). *Durbin-Watson statistic for the least trimmed squares.* Bulletin of the Czech Econometric Society **14**, 1 – 40.

[9] Víšek J.Á. (2004). *Durbin-Watson statistic in robust regression.* To appear in Probability and mathematical statistics.

Acknowledgement: This work has been supported by the grant GA ČR 402/03/0084 of the Grant Agency of the Czech Republic.

Address: J. Kalina, Department of Statistics, Charles University, Prague, Czech Republic & University Duisburg–Essen, Fachbereich 06, D-45117 Essen, Germany

E-mail: `kalina@stat-math.uni-essen.de`

THE EXPECTED EFFECTIVE RETIREMENT AGE AND THE AGE OF RETIREMENT

Jari Kannisto

Key words: Expected effective retirement age, expectancy, average effective retirement age, pension.

COMPSTAT 2004 section: Model selection.

Abstract: To follow up retirement the Finnish Centre for Pensions has developed a new indicator, the expected effective retirement age. In this paper we will refer to this concept as the expectancy. It meets the targets set for an good indicator: it reacts to the retirement risk immediately and in the right direction, and it is independent of the age structure of the population.

One of the main aims of the Finnish pension policy is to postpone effective retirement. To measure changes in the actual effective retirement age, better tools than before (average and median age) are needed. These tools are used to monitor the development and to support decision-making.

Due to age structure the average effective retirement age increases even if the retirement risk for each age group would not change. For instance the average effective retirement age will increase in the period 2002 - 2010 by about one year (cf. figure 1). This change is primarily due to the post-war babyboomers nearing retirement age.

1 The new indicator is based on life expectancy

When designing the indicator for the effective retirement age the aim was an indicator based on the retirement risk which would react immediately and in the right direction to changes in the retirement risk and only in the retirement risk and not in the number of new pensions. Such an indicator would be independent of the age structure of the population.

The indicator is comparable to the calculation of life expectancy. Life expectancy means the number of years that a person would live if the mortality rate for each age group stays unchanged.

Correspondingly the expectancy describes the average effective retirement age for a person of a specific age on the condition that the age specific retirement risk and the mortality rates do not change. The effect of mortality on the expectancy is slight (cf. figure 2).

The expectancy can be calculated for a person at any age, but The Finnish Centre for Pensions calculates it for both 25-year-olds and 50-year-olds. The expectancy for a 25-year-old reflects the whole population which is insured for earnings-related pension benefits, since at that age participation in working life begins to stabilise. It is very rare to retire earlier. In figure 3 the expectancy for every age between 25 and 65 has been calculated.

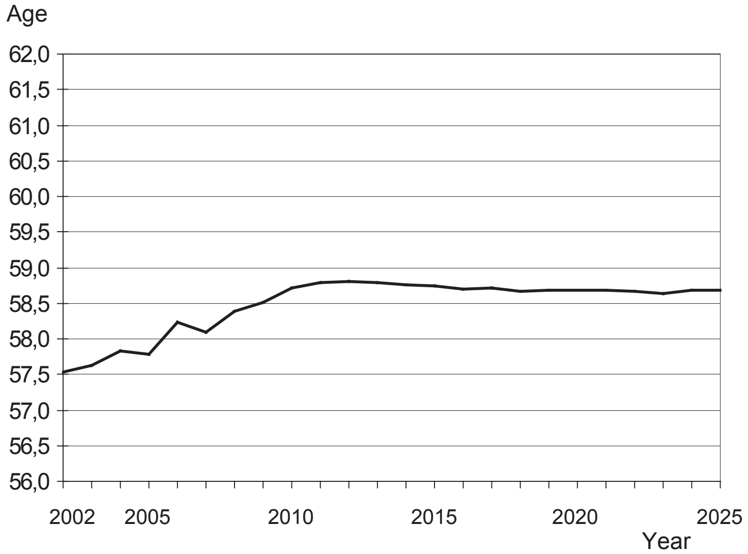


Figure 1: Effect of the age structure on the average effective retirement age of those who have retired from the private sector.

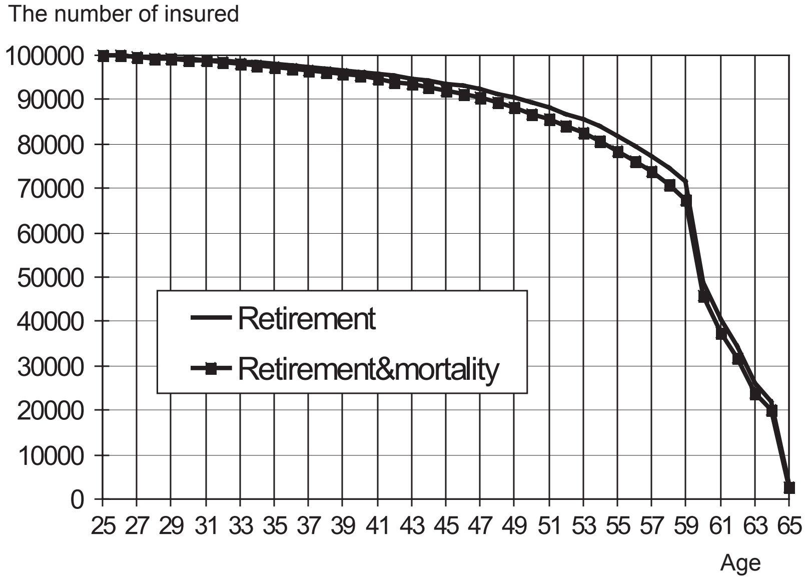


Figure 2: The theoretical number of insured 100,000 persons by age when calculating the expectancy in 2003.

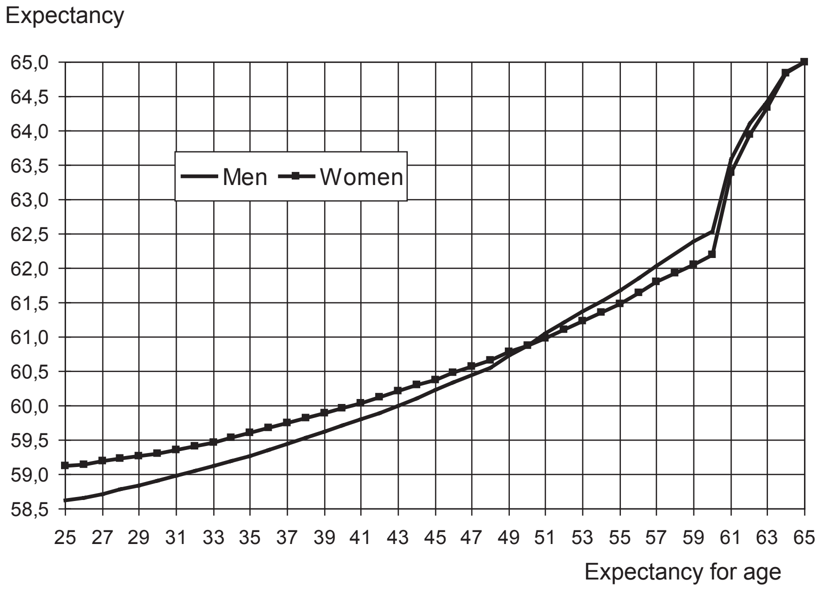


Figure 3: The expected effective retirement age (i.e. the expectancy) in 2003.

Of those retiring, about 15 percent are less than 50 year old. Illnesses and injuries in this group often prevent working. Calculating the expectancy for 50-year-olds is needed because willingness to retire can be influenced by pension policy.

2 Definition and mode of calculating the expectancy

The expectancy is calculated in that way that first the mortality risk and the retirement risk of those insured in the year of observation are calculated for each age group. Using these proportions, the number of persons of a group of insured of a chosen size and of a certain age (for instance 100,000 25-year-olds) who would retire within a year is calculated. The number of insured remaining at a one year higher age is obtained by subtracting from the original number those who have retired and the number of deceased calculated from the mortality rates. Continuing in this way age by age until the old-age retirement age, the calculated numbers of those retiring are obtained for each age group (cf. figure 2). The average age calculated from these assumed retirements is the expectancy (i.e. the expected effective retirement age).

Equation:
The proportion of persons retiring at age j is obtained from the equation

$$A_j = e_j \prod_{k=m}^{j-1} (1 - e_k - y_k) \tag{1}$$

and the expectancy is the weighted average of ages:

$$E_m = \left(\sum_{j=m}^{65} jA_j \right) / \sum_{j=m}^{65} A_j$$

e_j = retirement risk at age j
y_j = mortality risk at age j
m = chosen age of exit

The old-age retirement age in Finland is 65 year. The Finnish statutory earnings-related pension scheme is divided into the private and the public sector. Of insured more than a fourth works in the public sector.

3 The expectancy meets the requirements that were set for it

Several requirements were set for this new indicator:

a. The indicator should react in a correct way to changes in the retirement risk

   - It decreases when the retirement risk increases in some younger age group under 65 and increases when the retirement risk decreases in these age groups

b. The indicator may react only to changes in the retirement risk

   - It must not be affected by changes in population i.e. such as the age structure of the population

c. The indicator should react immediately to changes in the retirement risk

   - When the calculations are based on the number of new pensions, the indicator reacts immediately to changes in the retirement risk. If the calculations were made on the basis of the number of retired, the changes would be seen only slowly in the results.

d. The adequate statistics should be available

   - In Finland the Finnish Centre for Pensions maintains a central register of all pensions and employment contracts, which makes the calculations of the retirement risks possible.

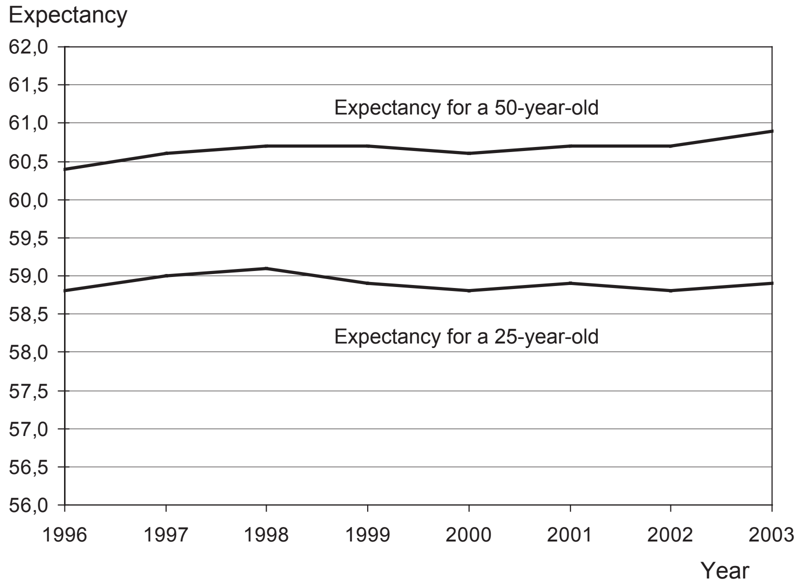


Figure 4: The expected effective retirement age (i.e. the expectancy) in 1996-2003.

These aforementioned basic criteria the expectancy does meet well indeed! On the other hand, a further criterion is for instance international comparability. In many countries it is probably difficult to obtain the data. We know that Rikstrygdeverket in Norway calculates the effective retirement age in a corresponding way, i.e. using a formula based on life expectancy. However, the calculation formula is not quite the same. It is not very useful to calculate the expectancy for a small population, because the number of new pensions in each age group in the population should reflect the probability of retirement. This criterion already requires such a large population that for instance calcu-lating the expectancy for the personnel of a single company is not useful (an exception maybe the largest companies in Europe).

4 Results

As regards all retired persons the expectancy has been calculated from 1996 (see figure 4). Between 1996 and 2003 the changes are small. Instead figure 5 shows the development of the expectancy for the private sector from 1983. It shows that the expectancy reacts to changes as wanted. The changes can be explained by legislative changes. The figure also shows how the average effective retirement age may react in an inappropriate way in certain situations. One example of this is the year 1986. At that time, flexible retirement programmes were introduced in Finland. As a result an uncommonly large

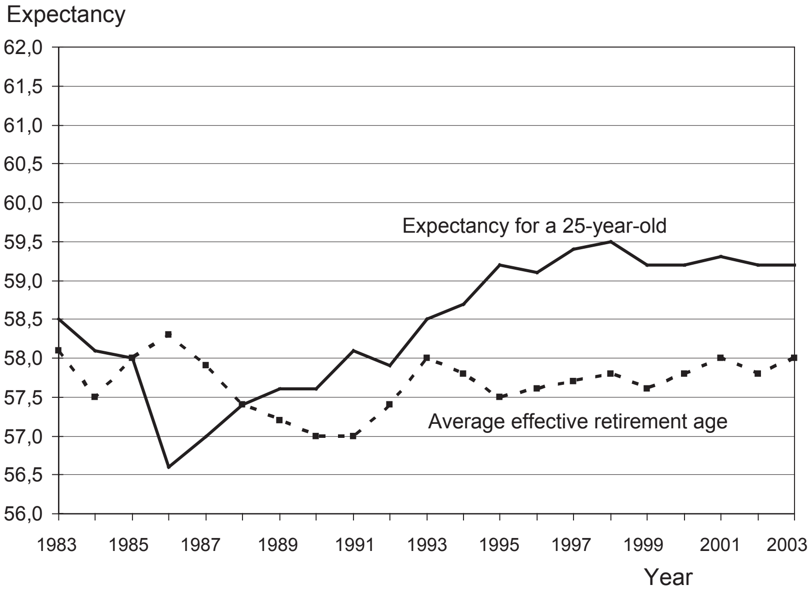


Figure 5: The expected effective retirement age (i.e. the expectancy) in 1983-2003 in the private sector.

number of persons aged 55-64 retired on these new types of pension. Although the early exit was most popular in 1986, the average effective retirement age was at the highest level and the expectancy was at the lowest level at the same time.

References

[1] Kannisto, Klaavo, Rantala, Uusitalo. (2003). *Missä iässä eläkkeelle?* Raportti työeläkkeelle siirtymisen iästä ja sen mittaamisesta (Retirement at which age? Report on the age of retirement on an earnings-related pension and its measurement) (Reports of the Finnish Centre for Pensions 32/2003)

Address: J. Kannisto, Statistical Department Finnish Centre for Pensions 00065 Elaketurvakeskus, Finland

E-mail: `jari.kannisto@etk.fi`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

TOTAL VARIATION PENALTY IN IMAGE WARPING: SOME COMPARISONS WITH BOOKSTEIN ROUGHNESS PENALTY

Stanislav Katina and Ivan Mizera

COMPSTAT 2004 section: Image analysis.

Abstract: We give some comparisons between parametric regression based on elastic splines and nonparametric regression based on plastic splines on total variation penalty.

1 Introduction

The objective of this paper is to compare Bookstein penalty and total variation penalty in the sense of allocation problem.

Consider two k landmark configuration matrices $\mathbf{X}_{k\times d} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)^T$ and $\mathbf{Y}_{k\times d} = (\mathbf{y}_1, \ldots, \mathbf{y}_k)^T$ in $\mathcal{R}^d$ (in the paper $d = 2$), both the matrices of Bookstein coordinates and we wish to deform $\mathbf{X}$ into $\mathbf{Y}$ (the source into the target), where $\mathbf{x}_j = \left[x_j^{(1)}, x_j^{(2)}\right]$, $\mathbf{y}_j = \left[y_j^{(1)}, y_j^{(2)}\right]$, $j = 1, \ldots, k$; $\mathbf{x}_j, \mathbf{y}_j \in \mathcal{R}^2$. We seek $\mathbf{f}$ to fit the dependence of $\mathbf{Y}$ on $\mathbf{X}$, where $\mathbf{f}$ is obtained as a minimizer of

$$S_{pen}(\mathbf{f}) = \sum_{j=1}^{k} (\mathbf{y}_j - \mathbf{f}(\mathbf{x}_j))^2 + \lambda J(\mathbf{f}). \tag{1}$$

The penalized estimator $\widehat{\mathbf{f}}$ is defined to be the minimizer of the functional (1) over the class of twice-differentiable continuous functions $\mathbf{f}$, with absolutely continuous first derivative and square integrable second derivative. The first term is traditionally called *(in)fidelity*, since it measures the overall lack of fit of $\mathbf{f}(\mathbf{x}_j)$ to $\mathbf{y}_j$. The second term is called *penalty* controlled by the *regularization parameter* λ (see Green and Silverman [6].)

1.1 Elastic splines

In univariate setting, given any twice-differentiable function f defined on $\langle a, b\rangle$, and a smoothing parameter $\lambda > 0$, define the penalized sum of squares [19]

$$S_{pen}(f) = \sum_{j=1}^{k} (y_j - f(x_j))^2 + \lambda \int_a^b (f'')^2\, dx. \tag{2}$$

The penalized least square estimator $\widehat{f}$ is defined to be the minimizer of the functional (2) over the class of twice-differentiable continuous functions f, with absolutely continuous first derivative and square integrable second derivative. The addition of the roughness penalty term $\lambda \int (f'')^2$ ensures that the cost (2) of a particular curve is determined not only by its goodness-of-fit to the data as quantified by the residual sum of squares $\sum (y_j - f(x_j))^2$ but also by its roughness $\lambda \int (f'')^2$. The smoothing parameter λ represents the "rate of exchange" between residual error and local variation and gives the amount in terms of summed square residual error that corresponds to one unit of integrated squared second derivative. For given the value of λ, minimizing (2) gives the best compromise between smoothness and goodness-of-fit. If λ is large, then the main component in (2) is the roughness penalty term and the minimizer $\widehat{f}$ displays very little curvature. In the limiting case as λ tends to infinity, the term $\int (f'')^2$ is forced to zero and the curve $\widehat{f}$ approaches the linear regression fit. On the other hand, if λ is relatively small then the main contribution to (2) is the residual sum of squares and the curve estimator $\widehat{f}$ tracks the data closely. In the limit case, as λ tends to zero, $\widehat{f}$ approaches the data closely.

In the same manner as the cubic splines interpolant, the thin-plate spline interpolant can be generalized to define the thin-plate smoothing spline as a unique function f which minimizes, for some positive parameter, λ the expression

$$S_{pen}(f) = \sum_{j=1}^{k} (y_j - f(\mathbf{x}_j))^2 + \lambda \int_{\mathcal{R}^2} \left[\begin{array}{c} \left(\frac{\partial^2 f}{\partial x^{(1)} \partial x^{(1)}}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x^{(1)} \partial x^{(2)}}\right)^2 \\ + \left(\frac{\partial^2 f}{\partial x^{(2)} \partial x^{(2)}}\right)^2 \end{array} \right] dx^{(1)} dx^{(2)} \tag{3}$$

As in the case of the cubic splines smoothing function, the first term of (3) measures fidelity to the data and is equal to zero if f interpolates the data exactly. The second term measures smoothness and is equal to zero if f is a plane, corresponding to a thin steel sheet with no deformation. The parameter λ controls the trade-off between the desire for fidelity to the data and desire for smoothness, with larger values producing smoother (but less faithful) estimates [18]. The primary reason for the popularity of the thin-plate splines as a function estimation is the fact that there exists an efficient algorithm for computation (Sibson and Stone, 1991). The general form of the thin-plate spline is

$$f(\mathbf{x}) = c + a_1 x^{(1)} + a_2 x^{(2)} + \sum_{j=1}^{k} w_j \phi_j(\mathbf{x}),$$

where $\phi_j(\mathbf{x}) = \phi(\mathbf{x} - \mathbf{x}_j) = \|\mathbf{x}\|^2 \log\left(\|\mathbf{x}\|^2\right)$, $j = 1, 2, \ldots, k$, if $\|\mathbf{x}\| > 0$, is known as a *radial (nodal) basis function* (see [8]. The function ϕ is not defined at the origin, it can be extended by continuity to have value $\phi_j(\mathbf{x}) = 0$, if $\|\mathbf{x}\| = 0$. Every thin-plate spline is a direct sum of a linear functions and a finite number of translations of the radially symmetric function ϕ. Harder and Desmarais [7], Duchon [4], Dyn, Levin and Rippa [5], Jackson [8] and Powell [17] suggest some other choices of ϕ.

Penalty $J(\mathbf{f})$ in (1) can be considered as a natural extension of the easily interpretable one dimensional prototype $\int (f'')^2$. The unnatural feature is the fact that the penalty is evaluated over all of $\mathcal{R}^2$ instead of over a more realistic bounded domain Ω. The latter alternative was considered by Green and Silverman [6] who coined the name *finite-window thin-plate splines (elastic splines)*. The name elastic splines comes from known physical model underlying the whole setting, in which the penalty is interpreted as the potential energy of a displacement, from the horizontal position, of an elastic thin metal plate, the displacement that mimics the form of fitted function interpolating the data points. The displacement should be small, one may say infinitesimal, thus rather in the form $\varepsilon\mathbf{f}$ than $\mathbf{f}$. So, $J(\mathbf{f})$ measures the overall roughness or "wiggliness" of $\mathbf{f}$; it measures rapid variation in $\mathbf{f}$ and departure from local linearity or flatness. $J(\mathbf{f})$ will be large if the function of $\mathbf{f}$ exhibits high local curvature, because this will result in large second derivatives.

The material aspects of the plate are rather limiting with respect to its physical reality - that is, one may abstract from its third dimension; it is elastic, hence it does not deform, only bend, and it is a plate, not a membrane, which means it is stiff - its behaviour is rather that of steel than that of gum [13]. The splines arise as a solution of (1) with the penalty (Bookstein [1], [2]; see also Dryden and Mardia [3]; Katina [9], [10])

$$J(\mathbf{f}) = \sum_{d=1}^{2} \int\int_{\mathcal{R}^2} \left[\begin{array}{c} \left(\frac{\partial^2 f_d}{\partial x^{(1)} \partial x^{(1)}}\right)^2 + 2\left(\frac{\partial^2 f_d}{\partial x^{(1)} \partial x^{(2)}}\right)^2 \\ + \left(\frac{\partial^2 f_d}{\partial x^{(2)} \partial x^{(2)}}\right)^2 \end{array} \right] dx^{(1)} dx^{(2)}, \tag{4}$$

where $\mathbf{f}(\mathbf{x}) = \mathbf{c} + \mathbf{A}\mathbf{x} + \mathbf{W}^T\mathbf{s}(\mathbf{x}) + \varepsilon$, $\mathbf{s}(\mathbf{x})_{k\times 1} = [\phi_1(\mathbf{x}), \ldots, \phi_k(\mathbf{x})]^T$. In matrix form

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{S} + \lambda\mathbf{I}_k & \mathbf{1}_k & \mathbf{X} \\ \mathbf{1}_k^T & \mathbf{0} & \mathbf{0} \\ \mathbf{X}^T & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{W} \\ \mathbf{c}^T \\ \mathbf{A}^T \end{pmatrix} + \varepsilon, \tag{5}$$

where $\mathbf{1}_k^T\mathbf{W} = \mathbf{0}, \mathbf{X}^T\mathbf{W} = \mathbf{0}$, $k \geq 3$. If $\lambda = 0$, $\mathbf{f}$ is interpolation spline, if $\lambda > 0$, $\mathbf{f}$ is smoothing spline. Penalty (4) in model (5) leads to $J(\mathbf{f}) = \frac{1}{8\pi} tr\left(\mathbf{W}^T\mathbf{S}\mathbf{W}\right)$. $J(\mathbf{f})$ is zero if $\mathbf{f}$ is linear (the computed spline is $\mathbf{f}(\mathbf{x}) = \mathbf{c} + \mathbf{A}\mathbf{x}$).

1.2 Plastic splines

Plastic splines arise as a solution of another instance of the regularization scheme, where penalty is $J(\mathbf{f}, \boldsymbol{\Omega}) = \int\int_{\boldsymbol{\Omega}} \left\|\nabla_{\mathbf{f}}^2\right\| dx^{(1)} dx^{(2)}$, the right side being a definition for smooth functions. This is subsequently extended to all functions whose gradient has bounded variation. It is more appropriate to refer to $J(\mathbf{f}, \boldsymbol{\Omega})$ as to family of penalties with necessity of choosing a matrix norm. Plastic penalties are always considered over bounded $\mathbf{\Omega}$, and obvious choice of the matrix norm is the l_2 Hilbert - Schmidt norm $\|\cdot\|_2$. It establishes parallelism between elastic and plastic penalties by

$$J^*\left(\mathbf{f}\right)=\int\int_{\mathcal{R}^2}\sqrt{\begin{array}{c}\sum_{d=1}^{2}\left(\frac{\partial^2 f_d}{\partial x^{(1)}\partial x^{(1)}}\right)^2\\ +2\left(\frac{\partial^2 f_d}{\partial x^{(1)}\partial x^{(2)}}\right)^2+\left(\frac{\partial^2 f_d}{\partial x^{(2)}\partial x^{(2)}}\right)^2\end{array}}\,dx^{(1)}dx^{(2)}. \qquad (6)$$

The penalty (6) was introduced by Koenker and Mizera [14]; see also Katina [11]. Any plastic penalty can be considered as a natural extension of the one-dimensional penalty equal to the total variation of the derivative, that is $J\left(f\right)=\int\left|f''\left(x\right)\right|dx=TV\ \left(f'\right)$, where $TV\left(\cdot\right)$ is total variation for smooth functions [15]. Plastic splines can be viewed as an l_1 alternative to the l_2 elastic ones and can be interpreted as the deformation energy, the work done by the stress in the course of deformation, of the plate displacement mimicking the interpolated shape [13]. The theory in which this interpretation is possible, the deformation theory of a perfectly rigid-plastic body, is a kind of a limit case of various other physical approaches to plasticity. Another link, connected to the mathematical expression of the penalty, gives the so-called *total-variation based denoising* of Rudin, Osher and Fatemi [20], motivated by a desire to recover edges, extrema, and other sharp features, while not penalizing smoothness. This leads (in an extreme case) to piecewise linear functions, but only in $\mathcal{R}$.

Plastic splines are computed via penalized triogram algorithm, which can be interpreted as the *Lagrange finite-element method*. The solutions are approximated by functions piecewise-linear on the triangular tessellation of the domain $\mathbf{\Omega}$, whose vertices encompass all covariate points (in $\mathcal{R}^2$, for Rudin, Osher and Fatemi [20] penalty is not known explicit solution and that is the reason to use approximation by finite-element method). All plastic penalties yields to the same result [11]

$$J\left(\mathbf{f},\mathbf{\Omega},\|\cdot\|\right)=c\sum_m\left\|\nabla^+_{\mathbf{f}_{e_m}}-\nabla^-_{\mathbf{f}_{e_m}}\right\|\|e_m\|\,, \qquad (7)$$

where m runs over all the interior edges of the triangulation, $\left\|\nabla^+_{\mathbf{f}_{e_m}}-\nabla^-_{\mathbf{f}_{e_m}}\right\|$ is the Euclidean length of the difference between gradients of $\mathbf{f}$ on the triangles adjacent to e_m, $\|e_m\|$ is the Euclidean length of the edge e_m. Model formulation is $\mathbf{Y}=\mathbf{f}(\mathbf{X})+\varepsilon$, where $\mathbf{f}:\mathcal{R}^2\longmapsto\mathcal{R}^2$. The linearization of the model is $\mathbf{Y}^*=\mathbf{f}(\mathbf{X}^*_\lambda)+\varepsilon$, where $\mathbf{f}^*(\mathbf{X}^*_\lambda)=\mathbf{X}^*_\lambda\beta$, $\mathbf{Y}^*=\left[\mathbf{Y}\vdots\mathbf{0}\right]^T$, $\mathbf{X}^*_\lambda=\left[\mathbf{X}_F\vdots\lambda\mathbf{X}_P\right]^T$, $\mathbf{X}_F$ is fidelity part and $\mathbf{X}_P$ is penalty part, $\lambda>0$ is smoothing parameter.

2 Example

2.1 Data and methods

A total of 117 specimens of *Gymnocephalus cernuus* (GC) and 89 specimens of *G. baloni* (GB) were collected from various locations in Slovak stretch of the River Danube between the village of Radvan nad Dunajom and the Devinske side-arm (rkm 1749 to 1880). The material was preserved in 4% formaldehyde solution. Standard length (SL, distance of landmarks 10 and 7), as well as 17 anatomically and morphologically specified landmarks (Figure 1), in the two species were examined using the *Impor PRO 32* software [16]. The measurements were acquired using image analysis based on digital photographs taken by the *Nikon CoolPix* 5000 camera.

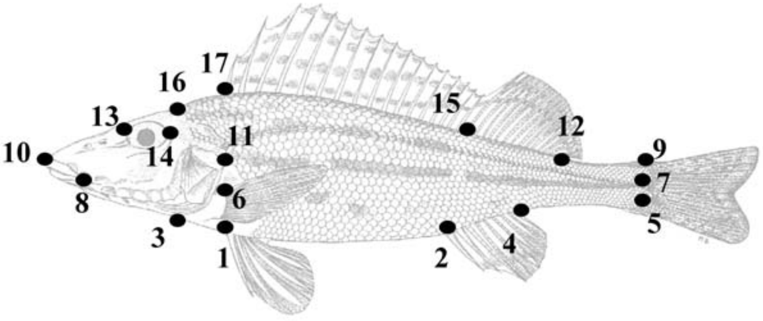


Figure 1:

The characters measured were transformed to Bookstein's coordinates, referred to SL. For the purpose of Bookstein penalty and total variation penalty comparison, only the specimens with higher SL (determined as an adult), $n_{GC} = 34$ $(SL > 80mm)$, $n_{GB} = 39$ $(SL > 100mm)$, were chosen. From both groups random samples ($n = 25$, function `sample()` in R) were used. Remaining part of the specimens were used as control groups.

The penalized model is used in the form

$$Model\ 1 : \mathbf{B}_i = \mathbf{f}\left({}_{25}\overline{\mathbf{B}}_g\right) + \varepsilon \mapsto \mathbf{B}_i^* = {}_{25}\overline{\mathbf{B}}_g^* \beta_{gi} + \varepsilon, \tag{8}$$

where $i = 1, 2, \ldots, n_{GB} + n_{GC}, g = GC,\ GB$, where ${}_{25}\overline{\mathbf{B}}_g$ are Bookstein mean shapes (matrices 17×2, modified by Mardia - Dryden bias constant) of two random samples, $\mathbf{B}_i$ are 17×2 matrices of Bookstein coordinates, $\mathbf{B}_i^*{=}\left[\mathbf{B}_i \vdots 0\right]^T$, ${}_{25}\overline{\mathbf{B}}_g^* = \left[\overline{\mathbf{B}}_{Fg} \vdots \lambda\overline{\mathbf{B}}_{Pg}\right]^T$, $\overline{\mathbf{B}}_{Fg}$ is fidelity part and $\overline{\mathbf{B}}_{Pg}$ is penalty part, $\lambda > 0$ is smoothing parameter. In the first approach,total variation penalty (7) and interpolation in the model (8) were used, where λ was close to zero. In the second approach, smoothing was used, with λ simulated from $I = (0, 1\rangle$ by 0.01. As a base of calculus was used *Delaunay triangulation.* We approximate over mentioned 17 landmarks.

For the purpose of comparison, Bookstein parametric model (5) was used. Matrix form of the model is

$$Model\ 2: \begin{pmatrix} \mathbf{B}_i \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{S}+\lambda\mathbf{I} & \mathbf{1} & {}_{25}\overline{\mathbf{B}}_g \\ \mathbf{1}^T & \mathbf{0} & \mathbf{0} \\ {}_{25}\overline{\mathbf{B}}_g^T & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{W} \\ \mathbf{c}^T \\ \mathbf{A}^T \end{pmatrix} + \varepsilon,$$

for $i = 1, 2, \ldots, n_{GB} + n_{GC}$, $g = GC$, GB, $\lambda = 0.000001$. Corresponding Bookstein penalty from (4)

$$J(\mathbf{f}) = \frac{1}{8\pi} trace\left(\widehat{\mathbf{W}}^T \mathbf{S} \widehat{\mathbf{W}}\right) = \frac{1}{8\pi} trace(\widehat{\mathbf{W}}^T \widehat{\mathbf{B}}_i).$$

For each type of the model (*Model 1* and *2*), the result is $(n_{GB} + n_{GC})$ $\times 2$ matrix $\mathbf{P}_{tv,\lambda}$ of total variation penalties, resp. $\mathbf{P}_B$ of Bookstein penalties with the first column for *Model 1* and the second column for *Model 2.* The best λ was found on the base of minimal number of misclassifications in linear discriminant analysis (LDA). In LDA were specimens classified using $\mathbf{P}_{tv,\lambda}$, resp. $\mathbf{P}_B$ to the groups GC and GB.

One can plot $\lambda \in I$ against penalty (7) or $\left\|\mathbf{B} - \widehat{\mathbf{B}}_\lambda\right\| + \lambda J_\lambda\left(\widehat{\mathbf{B}}_\lambda\right)$. Other approach is to plot $\lambda \in I$ against number of misclassifications.

2.2 Results

In *Model 2* were found 5 misclassifications (Figure 2), in *Model 1* (interpolation, $\lambda = 0.000001$) were found 18 misclassifications and in *Model 1* (smoothing with optimal $\lambda = 0.26$) were found 6 misclassifications (Figure 3). The line (Figure 2 and 3) is optimal classifier for two groups (GC, GB) found by two variables - the first and the second column of $\mathbf{P}_{tv,\lambda}$, resp. $\mathbf{P}_B$.

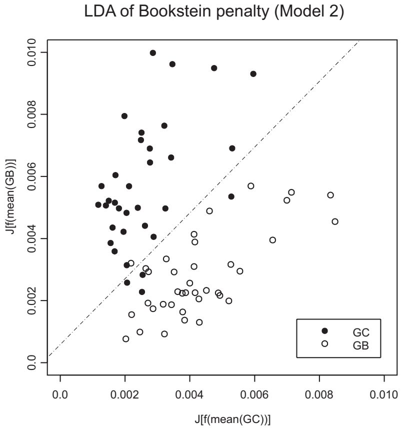


Figure 2:

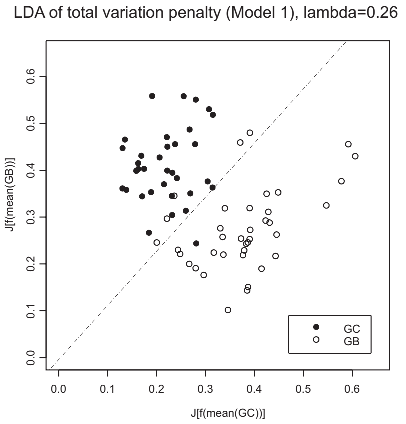


Figure 3:

Note: Figures - x-axis contains penalties from the models $\mathbf{B}_i = \mathbf{f}\left({}_{25}\overline{\mathbf{B}}_{GC}\right) + \varepsilon, i = 1, 2, \ldots, n_{GB} + n_{GC}$ (the first column of $\mathbf{P}_{tv,\lambda}$, resp. $\mathbf{P}_B$) and y-axis contains penalties from the models $\mathbf{B}_i = \mathbf{f}\left({}_{25}\overline{\mathbf{B}}_{GB}\right) + \varepsilon, i = 1, 2, \ldots, n_{GB} + n_{GC}$ (the second column of $\mathbf{P}_{tv,\lambda}$, resp. $\mathbf{P}_B$). Line represents linear discriminant function.

3 Conclusion

Each of mentioned methods has its own specific properties and areas of applications, we revealed only some of them - in biological sciences - in the comparison of two fish species. We used plastic and elastic penalties in allocation problem to reduce dimensions and classify objects into two groups. In the example, nonparametric regression model based on plastic splines (smoothing with $\lambda = 0.26$) is as good as parametric regression model based on elastic splines. The objective - to isolate two groups - overlaps with mathematical attributes of plastic splines - to isolate local extrema, spikes, sharp edges and similar phenomena in the data.

References

[1] Bookstein F.L. (1989). *Principal warps: thin-plate splines and the decomposition of deformation.* IEEE Transactions of Pattern and Machine Intelligence **11**, (6), 567 – 582.

[2] Bookstein F.L. (1991). *Morphometric tools for landmark data.* Geometry and Biology (Orange Book). Cambridge University Press, New York.

[3] Dryden I.L., Mardia K.V. (1999). *Statistical shape analysis.* John Willey, New York.

[4] Duchon J. (1977). *Spline minimising rotation-invariant seminorms in Sobolev space.* In: Constructive Theory of Functions of Several Variables, Lecture Notes in Math. 571, Schempp W., Zeller K. (eds), Springer, Berlin, 85 – 100.

[5] Dyn N., Levin D., Rippa S. (1986). *Numerical procedures for global surface fitting on scattered data by radial functions.* SIAM. Journal of Sci.Stat.Comp. **7**, 639 – 659.

[6] Green P.J., Silverman B.W. (2000). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman-Hall, New York.

[7] Harder R.L., Desmarais R.N. (1972). *Interpolation using surface splines.* Journal of Aircraft **9**, 189 – 191.

[8] Jackson I.R.H. (1989). *Radial basis functions: a survey and new results. DAMTP 1988/NA16.* In: The Mathematics of Surfaces III, Handscomb D.C., Ed., Oxford University Press, 115 – 133.

[9] Katina S. (2002). *Multivariate shape analysis: implementation of some methods to S-PLUS, application to biological sciences.* PhD. minimal thesis, Comenius University, Bratislava.

[10] Katina S. (2003). *Application of allometric shape analysis to fish growth with geometrical notes.* Tatra Mountains Mathematical Publication **26**, 323 – 336.

[11] Katina S. (2004). *Total variation penalty in image warping and selected multivariate methods in Shape Analysis.* PhD. thesis, Comenius University, Bratislava.

[12] Katina S., Mizera I. (2003). *Total variation penalty in image warping.* (draft)

[13] Koenker R., Mizera I. (2002). *Elastic and plastic splines: some experimental comparisons.* Statistical Data Analysis based on the L1-norm and Related Methods Y. Dodge (ed.), Birkhäuser, Basel, 405 – 414.

[14] Koenker R., Mizera I. (2003). *Penalized triograms: total variation regularization for bivariate smoothing.* (Journal of the Royal Statistical Society, Series B, in press).

[15] Koenker R., Ng P., Portnoy S. (1994). *Quantile smoothing splines.* Biometrika **81**, 673 – 680.

[16] Kovac V., Katina S. (2003). *Ontogenetic patterns and interspecific variability in external morphology of three sympatric Gymnocephalus species.* Proceedings from the Percis III Conference, Medison, Wisconsin, USA, 25 – 26.

[17] Powell M.J.D. (1990). *The theory of radial basis functions.* Approximation in 1990. DAMTP 1990/NA11. Cambridge.

[18] Ramsay T. (1999). *A bivariate finite element smoothing spline applied to image registration.* A thesis submitted to the Dept. of Mathematics and Statistics in conformity with the requirements for degree of Doctor of Philosophy, Queenss University Kingston.

[19] Ramsay J.O., Silverman B.W. (1997). *Functional data analysis.* Springer, New York.

[20] Rudin L.I., Osher S., Fatemi E. (1992) *Nonlinear total variation based noise removal algorithms.* Physica D **60**, 259 – 268.

Acknowledgement: The research was partially supported by the Slovak Scientific Grant Agency VEGA, projects No 1/7154/20 and 1/0272/03.

Address: S. Katina, Department of Probability and Mathematical Statistics,
Faculty of Mathematics, Physics and CS, Comenius University,
842 48 Bratislava, Slovakia
I. Mizera, Department of Mathematical and Statistical Sciences, Faculty of Science, University of Alberta, T6G 2G1, Edmonton, Alberta, Canada

E-mail: `katina@fmph.uniba.sk, mizera@stat.ualberta.ca`

COMPSTAT'2004 Symposium

FUNCTIONAL DATA ANALYSIS OF THE DYNAMICS OF YIELD CURVES

Y. Kawasaki and T. Ando

Key words: Yield curves, surface smoothing, penalized likelihood.

COMPSTAT 2004 section: Functional data analysis.

Abstract: This paper mainly concerns the prediction of next business day's yield curves and hence bond prices, given the past bond data up to today. We propose a forecast method for yield curves based on functional data analysis. At first, yield surface is estimated over several days, hence the past data is once reduced to a function of time and maturity. Yield curves for prediction are constructed by a continuously weighted average of the yield surface. The prediction accuracy of FDA approach is compared to that of the naïve method where the yield curves of a day obtained from cross-sectional regression is simply extrapolated to the next day.

1 Introduction

There have been a number of studies attempting to establish a suitable technique for estimating the term structure of interest rates from a cross-section of coupon bond prices. Under the assumption that the price of a bond is equal to the present value of its future coupon payments and redemption, McCulloch [8] regressed cash flows on a set of basis functions to estimate discount functions. Once the discount function $\delta(t)$ is estimated, the zero-coupon yield $\eta(t)$ and the forward rate $f(t)$ can be obtained by transformations of the discount function by such relationships as $f(t) = -\delta'(t)/\delta(t)$ and $\eta(t) = -\ln(\delta(t))/t$. See for example Anderson et al. [2] for the derivations of these relationships. Though we estimate the discount function only in this paper, we often use the generic term 'yield curve(s)' because these curves essentially contain the same information.

The approach adopted by McCulloch [8], [9] was followed by several related studies which try to refine this approach. To increase the stability of the estimated yield curves, some researchers are concerned with the choice of basis functions when defining a spline function, while others question how to place knots efficiently. However, what is essential is that the instability of the estimated yield curves is caused by the ill-posed nature of the regression spline, rather than by the inappropriate choice of the basis function. By ill-posed it is meant that a model may be over-parameterized compared to the amount of sample information. From this perspective, smoothing spline is an indispensable tool for yield curve estimation. Fisher, Nychka and Zervos [5] is one of the early works frequently referred in this literature. In this respect, we consistently employ penalized likelihood approach throughout this paper.

The main contribution of this paper is to propose a scheme for forecasting next business day's yield curves (and hence bond prices) based on functional data analysis (FDA), see Ramsay and Silverman [10]. It is a natural extension of smoothing spline approach for static estimation of yield curves. Moreover, establishing a dynamic model based on the function approximation of data is easier than working on raw discrete data because the classes of actually traded bonds vary from day to day. Hence, it makes difficult to consider a multivariate time series model with respect to individual bonds.

This paper is organized as follows. The section 2 presents a FDA methodology to estimate yield curves dynamically. In the first step of our method, yield *surface* is estimated from the bond data of past several days. In other words, the past data is once reduced to a function of time and maturity. Yield curves for prediction are constructed by regressing bond price data of the latest business day on the past yield surface, or a continuously weighted average of yield surface. In the section 3, the prediction accuracy of FDA approach is compared to that of the naïve method where the yield curves of a day is simply extrapolated to the next day. The section 4 concludes this paper.

2 Prediction via yield surface

Consider a set of n_i bonds traded on i-th day, and suppose that we have stocked bond data for $r+2$ days, hence i runs from 1 to $r+2$ ($i = 1, \ldots, r+2$). For i-th day, the yield curves can be estimated by the method explained in the previous section. Now we extend this scheme so that the yield curves of i-th day should be expressed as a weighted sum of past yield curves. In the following scheme, we determine the shape of yield curves on $(r+1)$-th day based on the *yield surface* estimated from the past r-days data, and use the estimated $(r+1)$-th day's yield curve to predict the bond prices of $(r+2)$-th day.

2.1 Estimation of yield surface

Let $p_{i\alpha}$ be the price of bond $i\alpha$ on i-th day, $c_{i\alpha}$ be its coupon payment, which is paid at time $t_1^{i\alpha}, \ldots, t_{L_{i\alpha}}^{i\alpha}$, let $R_{i\alpha}$ be the redemption payment, and let $L_{i\alpha}$ be the number of remaining payments. Following the theory of bond pricing [8], we assume that the price of a bond (plus accrued interest $a_{i\alpha}$) is equal to the present value of its future coupon payments and the redemption, i.e.,

$$p_{i\alpha} + a_{i\alpha} = c_{i\alpha} \sum_{k=1}^{L_{i\alpha}} \delta(t_k^{i\alpha}, i) + R_{i\alpha} \delta(t_{L_{i\alpha}}^{i\alpha}, i) + \varepsilon_{i\alpha}, \tag{1}$$

$$\alpha = 1, \ldots, n_i, i = 1, \ldots, r,$$

where $\delta(\cdot, i)$ is the discount function for i-th day, $\varepsilon_{i\alpha}$ are independent and normally distributed with mean of zero and variance σ^2.

We give a basis approximation for $\delta(t,i)$, and expressed it as a linear combination of a set of m underlying basis functions for each i as follows.

$$\delta(t,i) = 1 + \sum_{j=1}^{r}\sum_{k=1}^{m} w_{jk}\phi_{jk}(t,i) = 1 + \boldsymbol{w}'\boldsymbol{\phi}(t,i). \tag{2}$$

Basis function methods are quite common in the literature of FDA, see Chapter 3.2 of Ramsay and Silverman [10]. In this paper, we choose Gaussian radial basis function (RBF) for $\phi_{jk}(t,i)$, or $\boldsymbol{\phi}(t,i)$ is an mr-dimensional vector constructed from a set of basis functions in the following form,

$$\phi_{jk}(t,i) = \exp\left\{-\frac{(t-c_{jk1})^2}{2s_1^2} - \frac{(i-c_{jk2})^2}{2s_2^2}\right\},\ k = 1,\dots,m; j = 1,\dots,r,$$

and $\boldsymbol{w}$ is an unknown parameter vector to be estimated.

The location of each radial basis function $\boldsymbol{c}_{jk} = (c_{jk1}, c_{jk2})$ is determined as follows. As for the time axis, the choice of c_{jk2} is quite simple as $c_{jk2} = i$. As for c_{jk1} $(j = 1,\dots,r)$, they should be given so that they divide the interval $[0,T]$ into equal length subintervals. Here T denotes the longest maturity where the redemption payment (of some bond) occurs.

It follows from equations (1) and (2) that the bond price model based on a linear combination of basis functions is as follows.

$$f(y_{i\alpha}|\boldsymbol{t}_{i\alpha};\boldsymbol{w},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_{i\alpha}-\boldsymbol{c}'_{i\alpha}\Phi_{i\alpha}\boldsymbol{w})^2}{2\sigma^2}\right\}, \tag{3}$$

where $\boldsymbol{t}_{i\alpha} = (t_1^{i\alpha},\dots,t_{L_{i\alpha}}^{i\alpha})'$ is the vector of the points of time at which payments occur, $y_{i\alpha} = p_{i\alpha} + a_{i\alpha} - L_{i\alpha}c_{i\alpha} - R_{i\alpha}$, $\Phi_{i\alpha} = (\boldsymbol{\phi}(t_1^{i\alpha}),\dots,\boldsymbol{\phi}(t_{L_{i\alpha}-1}^{i\alpha}), \boldsymbol{\phi}(t_{L_{i\alpha}}^{i\alpha}))'$ and $\boldsymbol{c}_{i\alpha} = (c_{i\alpha},\dots,c_{i\alpha},c_{i\alpha}+R_{i\alpha})'$, respectively. Then one might try to maximize the log-likelihood function over r days, namely

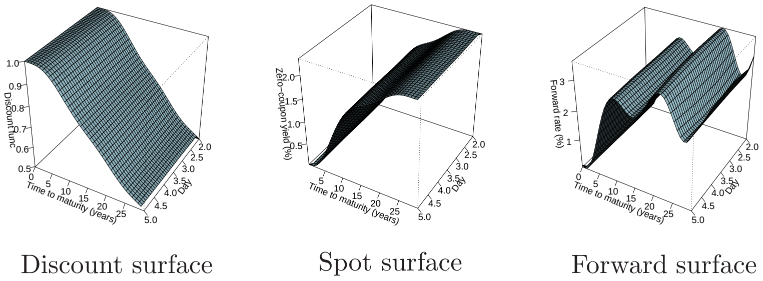
$$\ell(\boldsymbol{w},\sigma^2) = \sum_{i=1}^{r}\sum_{\alpha=1}^{n_i}\log f(y_{i\alpha}|\boldsymbol{t}_{i\alpha};\boldsymbol{w},\sigma^2)$$

to obtain estimates for the unknown parameters.

2.2 Penalized likelihood approach

In practice, however, the maximum likelihood method does not yield satisfactory results because the parameter estimates tend to be unstable and lead to overfitting. To avoid this, a penalty term on the smoothness of the unknown coefficients is introduced into the log-likelihood. This is sometimes referred to as the roughness penalty or regularization approach, see Chapter 4 of Ramsay and Silverman [10]. Specifically, we maximize

$$\ell_\lambda(\boldsymbol{w},\sigma^2) = \sum_{i=1}^{r}\sum_{\alpha=1}^{n_i}\log f(y_{i\alpha}|\boldsymbol{t}_{i\alpha};\boldsymbol{w},\sigma^2) - \frac{n\lambda}{2}\boldsymbol{w}'\boldsymbol{w}, \tag{4}$$



Discount surface Spot surface Forward surface

Figure 1: Yield surfaces estimated from Aug. 2 to 5, 2003.

where λ is the smoothing parameter controlling the smoothness of the discount function, and $n = \sum_{i=1}^{r} n_i$.

Given λ, s_1^2, s_2^2, and m, the unknown parameters $\boldsymbol{w}$ and σ^2 can be obtained as the solution of $\partial l_\lambda(\boldsymbol{w}, \sigma^2)/\partial \boldsymbol{w} = 0$ and $\partial l_\lambda(\boldsymbol{w}, \sigma^2)/\partial \sigma^2 = 0$. Now let us define the $n_i \times m$ matrix B_i for i-th day by $B_i = (\Phi'_{i1}\boldsymbol{c}_{i1}, \ldots, \Phi'_{i,n_i}\boldsymbol{c}_{i,n_i})'$, and use it to define the $n \times m$ matrix $B = (B'_1, \ldots, B'_r)'$. Then the maximum penalized likelihood estimates of $\boldsymbol{w}$ and σ^2 in the bond price model (3) are explicitly provided by

$$\hat{\boldsymbol{w}} = \left(B'B + n\lambda\sigma^2 I\right)^{-1} B'\boldsymbol{y}, \qquad \hat{\sigma}^2 = \sum_{i=1}^{r}\sum_{\alpha=1}^{n_i} \frac{1}{n_i}\left\{y_{i\alpha} - \boldsymbol{c}'_{i\alpha}\Phi_{i\alpha}\hat{\boldsymbol{w}}\right\}^2. \tag{5}$$

To complete our scheme, we need some criterion to choose λ, s_1^2, s_2^2 and m. In this paper, we use AIC_M, the modification of AIC [1]. AIC_M is defined by replacing the bias correction term by the trace of smoother matrix $H = B(B'B + n\lambda\sigma^2 I)^{-1}B'$, hence by

$$\text{AIC}_M = -2\ell(\hat{\boldsymbol{w}}, \hat{\sigma}^2) + 2\text{tr}(H). \tag{6}$$

For AIC_M, see Hastie and Tibshirani [6], Eilers and Marx [4], Konishi et al. [7]. Three panels in Figure 1 shows the estimated yield surfaces using the data of Japanese governmental bond traded from August 2, 2003 to August 5, 2003. Once the discount surface (the analytic form of the left panel) is estimated, other surfaces can be easily derived.

2.3 Regression analysis for a functional response

Based on the estimated discount surface, we predict the yield curve of $(r+1)$-th day. We assume that the discount function for $(r+1)$-th day can be derived as the average of (2) weighted by $\beta(u) = \sum_{k=1}^{z} \theta_k \phi_k(u)$, where $\phi_k(u)$ is one-dimensional Gaussian RBF with dispersion parameter s_3^2. The location of each radial basis function is given by dividing the interval $[1, r]$ into equal length subintervals.

Note that the weight function $\beta(u)$ is a continuous function of u, and the weight curve is assumed to be common to all the maturity t;

$$d(t) = \int_1^r \hat{\delta}(t,u)\beta(u)du. \tag{7}$$

Using this discount function, the bond equation for $(r+1)$-th day is defined as

$$p_{r+1,\alpha} + a_{r+1,\alpha} = c_{r+1,\alpha} \sum_{k=1}^{L_{r+1,\alpha}} d(t_k^{r+1,\alpha}) + R_{r+1,\alpha} d(t_{L_{r+1,\alpha}}^{r+1,\alpha}) + \varepsilon_{r+1,\alpha}, \tag{8}$$

$$\alpha = 1, \dots, n_{r+1},$$

and the likelihood takes the form of

$$f(y_{r+1,\alpha}|\boldsymbol{t}_{r+1,\alpha}; \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_{r+1,\alpha} - \boldsymbol{c}'_{r+1,\alpha}\Phi_{r+1,\alpha}\boldsymbol{\theta})^2}{2\sigma^2} \right\}, \tag{9}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_z)$. Same as in section 2.2, this naturally leads to the maximization of penalized likelihood,

$$\ell_{\tilde{\lambda}}(\boldsymbol{\theta}, \sigma^2) = \sum_{\alpha=1}^{n_{r+1}} \log f(y_{r+1,\alpha}|\boldsymbol{t}_{r+1,\alpha}; \boldsymbol{\theta}, \sigma^2) - \frac{n_{r+1}\tilde{\lambda}}{2}\boldsymbol{\theta}'\boldsymbol{\theta},$$

Unknown parameters are to be estimated by maximizing $\ell_{\tilde{\lambda}}$. The smoothing parameter $\tilde{\lambda}$, the dispersion of RBF s_3^2 and the number of basis z should be chosen so as to minimize the AIC$_M$ criterion.

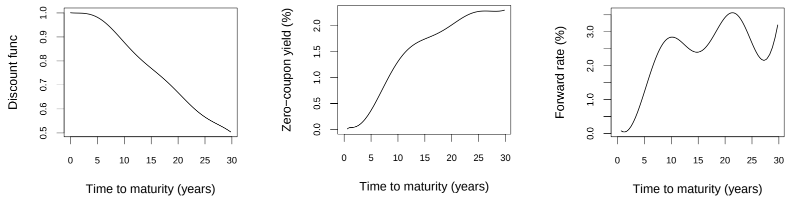
This final step is eventually the cross-sectional regression with regularization. What distinguishes our approach from a cross-section data based framework of yield curve estimation is that the discount function $d(\cdot)$ in (8) is estimated based on the continuous surface $(\hat{\delta}(t,u))$ from equation (7). Figure 2 shows the estimated yield curves by the FDA approach presented here.

3 Empirical analysis

3.1 Smoothing spline for cross-section data

As a competitive forecasting scheme, we briefly review the smoothing spline approach for a cross-sectional bond data. The situation is much simpler than that explained in section 2.1. We only use the set of n_{r+1} bonds traded on $(r+1)$-th day. The discount function $\delta(t)$ is to be estimated only on this $(r+1)$-th day's data. Therefore we have the following bond equation,

$$p_{r+1,\alpha} + a_{r+1,\alpha} = c_{r+1,\alpha} \sum_{k=1}^{L_{r+1,\alpha}} \delta(t_k^{r+1,\alpha}) + R_{i\alpha}\delta(t_{L_{r+1,\alpha}}^{r+1,\alpha}) + \varepsilon_{r+1,\alpha}, \tag{10}$$



Discount function Spot rate Forward rate

Figure 2: Yield curves of Aug. 6, 2003 estimated by FDA approach.

$$\alpha = 1, \dots, n_{r+1}$$

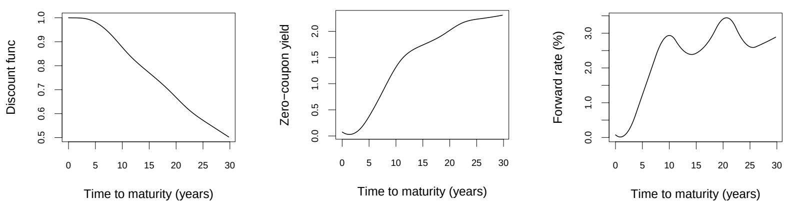
where $\delta(t)$ is the discount function, $\varepsilon_{r+1,\alpha}$ are independent and normally distributed with mean of zero and variance σ^2. The most significant difference between (8) and (10) is that in the former the discount function $d(\cdot)$ is estimated from the continuous surface $\hat{\delta}(t,u)$ extracted from discrete data but in the latter splines or other basis functions are placed on $\delta(t)$, and $\delta(t)$ for r-th day and $(r+1)$-th day are estimated independently.

We will not repeat the form of basis expansion approximation, and do not explicitly mention the probability density and likelihood function because the argument is almost parallel to those of section 2.3. In our experiment, we have chosen the natural cubic spline specification proposed by McCulloch [9]. As a penalty term on the roughness of spline coefficients, we put the second order restriction $\sum_{j=2}^{m}(\Delta^2 w_j)^2$ where $w_1, \dots, w_m$ denotes the coefficients to be estimated, and $\Delta w_k = w_k - w_{k-1}$ is the difference operator. As for the determination of smoothing parameter and the number of basis functions, we use generalized cross-validation (GCV, Craven and Wahba [3]) because it is the most common choice in this literature.

Three panels in Figure 3 are the estimated yield curves for August 6, 2003, and only the data observed on that day is used to estimate these curves. Apparently, there seems to be no significant difference between the discount functions in Figure 2 and 3, but the shapes of forward rate are different to some extent. In the next subsection, we examine the predictability of these yield curves through empirical forecasts of the bond price on August 7, 2003.

3.2 Comparison of prediction accuracy

So far we presented two different methods for estimating yield curves of JGB on August 6, 2003. One is a FDA approach proposed in section 2, and the other is well-known McCulloch's framework with roughness penalty. Here, the forecasts from former method will be called 'FDA forecasts' and the 'naïve forecasts'. Goodness of prediction is compared empirically through



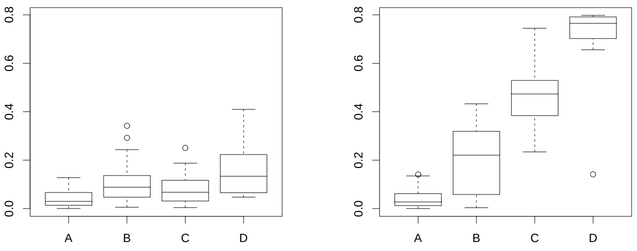
Discount function Spot rate Forward rate

Figure 3: Yield curves of Aug. 6, 2003 estimated by McCulloch's cubic spline with roughness penalty.

the pricing errors from the out-of-sample forecast for the price of traded bonds on August 7, 2003.

The maturity interval up to 30 years is divided into four subintervals, $A = [0, 5]$, $B = (5, 10]$, $C = (10, 20]$ and $D = (20, 30]$. Every bond will be classified into either subinterval according to the duration by its redemption payment. On August 7, 2003, 232 bonds were traded, and the sample size of each subinterval is 127, 59, 39, 7 respectively. Pricing errors are measured by the absolute deviation from the true (or realized) bond price, and these errors are averaged within four disjoint intervals.

Two box plots in Figure 4 clearly exhibit that the forecasts by FDA approach is better than those by simple-minded extrapolation of previous day's yield curve in the sense of mean absolute pricing errors, especially at longer maturities.



FDA forecasts Naïve forecasts

Figure 4: Box-plot of absolute pricing errors for Aug. 7, 2003.

4 Conclusion

A FDA approach which enables dynamic modelling of yield curves is presented. After reducing the past bond data to a function of time and of maturity, the model is constructed by a continuously weighted average of the yield surface in an integral form. Comparison of prediction accuracy shows FDA approach is better than the simple extrapolation of a static yield curve. As a future work, it is interest to investigate some theoretical results (asymptotic properties of the estimators) about the procedure proposed in this paper.

References

[1] Akaike H. (1973). *Information theory and an extension of the maximum likelihood principle.* 2nd Inter. Symp. on Information Theory Petrov B.N. and Csaki F.(eds), Akademiai Kiado, Budapest, 267 − 281.

[2] Anderson N., Breedon F., Deacon M., Derry A., Murphy G. (1996). *Estimating and interpreting the yield curve.* John Wiley and Sons, Chichester.

[3] Craven P., Wahba G. (1979). *Smoothing noisy data with spline functions.* Numerishe Mathmatik **31**, 377 − 403.

[4] Eilers P.H., Marx B.D. (1996). *Flexible smoothing with B-splines and penalties.* Statistical Science **11**, 89 − 102.

[5] Fisher M.E, Nychka D., Zervos D. (1995). *Fitting the term structure of interest rates with smoothing splines.* Federal Reserve Bank Finance and Economics Discussion Paper 95-1.

[6] Hastie T.J., Tibshirani R.J. (1990). *Generalized additive models.* Chapman and Hall/CRC, Boca Raton.

[7] Konishi S., Ando T., Imoto S. (2003). *Bayesian information criteria and smoothing parameter selection in radial basis function networks.* ISM Research Memorandum No. 846 (to appear Biometrika, 2004).

[8] McCulloch J.H. (1971). *Measuring the term structure of interest rates.* Journal of Business **44**, 19 − 31.

[9] McCulloch J.H. (1975). *The tax-adjusted yield curve.* Journal of Finance **30**, 811 − 830.

[10] Ramsay J.O., Silverman B.W. (1997). *Functional data analysis*, Springer,New York.

Acknowledgement: This research was supported by Grant-in-Aid for Scientific Research sponsored by JSPS (14208025, 14380127) and by MEXT (14780177).

Address: Y. Kawasaki, The Institute of Statistical Mathematics,
4–6–7 Minami-Azabu, Minato-ku, Tokyo 106–8569, Japan

T. Ando, The Institute of Medical Science, The University of Tokyo, 4–6–1 Shirokanedai, Minato-ku, Tokyo 108–8639, Japan

E-mail: `kawasaki@ism.ac.jp & ando@ims.u-tokyo.ac.jp`

COMPSTAT'2004 Symposium

ON ORDERING OF SPLITS, GRAY CODE, AND SOME MISSING REFERENCES

Jan Klaschka

Key words: Gray codes, classification and regression trees.

COMPSTAT 2004 section: Algorithms.

Abstract: In the COMPSTAT'98 paper [8] Klaschka and Mola proposed a combinatorial "trick" and its use in tree-growing algorithms. Only later the "trick", supposed novel, revealed itself as an independently reinvented old idea. The present paper shows a formerly overlooked context of related works.

1 Introduction

This paper follows up the COMPSTAT'98 paper by Klaschka and Mola [8]. In the 1998 paper we dealt with an economical way of calculating, in the tree-based methods, all the $2^{n-1}-1$ values of splitting criterion for the splits based on an n-valued categorical variable. We proposed a specific ordering of splits suitable for recalculating one value from another.

The core of the paper was a combinatorial idea: The $(n-1)$-tuples of 0's and 1's coding the splits can be ordered in a sequence so that any two successive elements differ in exactly one position. The 1998 work is recapitulated in Section 2.

Only later I learned that the same combinatorial idea had already been utilized as soon as in the sixties in the all possible subsets regression, as recalled in Section 3. Moreover, even in the sixties the idea was not quite novel, as explained in Section 4, together with a brief review of so called combinatorial Gray codes and their applications.

Not only authors of [8] were unaware of the existence of previous works: Concluding remarks in Section 5 mention some other missing references.

2 Ordering of splits: COMPSTAT'98 recap

In COMPSTAT'98 paper [8] Klaschka and Mola introduced a sequence of integers $o_0, o_1, o_2, \ldots$ whose definition can be restated as follows:

(D1) $o_0 = 0$.

(D2) Given the first 2^n elements ($n \geq 0$), list them in the reverse order, increase each by 2^n, and join the list to the sequence as $o_{2^n}, \ldots, o_{2^{n+1}-1}$.

The sequence possesses the following properties:

(P1) For every $n \geq 0$, the first 2^n elements of the sequence are a permutation of $0, 1, \ldots, 2^n - 1$.

(P2) For every $i \geq 0$, elements o_i and o_{i+1} differ by a power of 2, i.e. their binary representations differ in exactly one digit.

The sequence is the core of an economical algorithm of splitting criterion calculation in tree-based methods, proposed in [8]. A binary split based on a categorical predictor X with values $x_1, \dots, x_n$ can be identified with a partition of set $\{x_1, \dots, x_n\}$ into an unordered pair of nonempty disjoint subsets.

For various splitting criteria it is reasonable that manipulating the raw data is confined to calculation of so called *auxiliary statistics* (see [9]) $\alpha(x_1)$, $\dots, \alpha(x_n)$ assigned to the individual values of X. From these, analogous auxiliary statistics for the subsets of $\{x_1, \dots, x_n\}$ can be constructed. The splitting criterion value for a split can then be obtained as a function of the pair of auxiliary statistics for two subsets.

The $2^{n-1}-1$ splits based od X can be coded with numbers $1, \dots, 2^{n-1}-1$: The code assigned to a split is $\sum_{i \in L} 2^{i-1}$, where L is the set of such indices i between 1 and $n-1$ that x_i and x_n belong to different subsets.

When the splits based on X are ordered so that their codes are o_1, $\dots, o_{2^{n-1}-1}$, then the neighbours in the sequence can be obtained from one another by moving a single element x_j from one subset to the other. This allows, when the splits are evaluated in the given order, to obtain all the necessary auxiliary statistics for subsets by stepwise recalculations corresponding to inclusion or exclusion of a single element. The number of operations with the auxiliary statistics is thus minimized, and as a result, the best split based on X is found economically. (For details, see [8].)

The combinatorial idea has a more intuitive geometric interpretation. The nonnegative integers up to $2^n - 1$ can be identified with the n-tuples of digits of their binary representations, and these, in turn (through cartesian coordinates), with vertices of n-dimensional unique cube. Then, a numeric sequence of length 2^n fulfilling (P1) and (P2) "translates" into such a path along (some) edges of an n-dimensional cube that visits exactly once each vertex. The concrete path given by (D1) and (D2) can then be obtained as follows. (This is, in fact, the way the combinatorial idea was presented during the COMPSTAT'98 lecture.)

Let the *floor* and *ceiling* of the cube consist of all the vertices with n-th coordinate equal to zero and one, resp. The way of identification of the floor and ceiling with $(n-1)$-dimensional cubes (by deleting the n-th coordinate) is obvious.

The definition of the numeric sequence may then be "translated":

(C1) For $n = 0$, visit the only vertex of the cube.

(C2) If $n > 0$, follow the path corresponding to dimension $n-1$ on the floor, then jump to the ceiling, and finally repeat on the ceiling, what has been done on the floor, but in the reverse order.

As an example, the path for $n = 3$ is shown at Fig. 1.

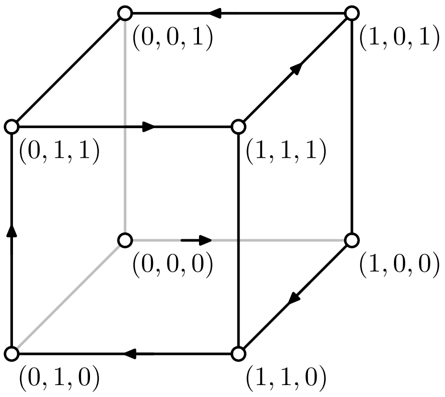


Figure 1: Path along edges of cube given by (C1), (C2) for $n = 3$. The path starts in $(0, 0, 0)$, and ends in $(0, 0, 1)$.

3 All possible subsets regression

Besides binary splits, there is another structure whose elements can be naturally identified with vertices of an n-dimensional cube: The set of all subsets of an n-element set. Each position represents one element, 1 coding presence, and 0 absence of the respective element in the subset. Two vertices are then joined with an edge if and only if the two corresponding subsets differ by a single element.

Thus, applying the combinatorial "trick" outlined in Section 2 in another area, namely in the search for the best subset of regressors in multiple linear regression, seemed to be a promising idea. Evaluation of 2^n subsets in the order given by $o_0, \ldots, o_{2^n-1}$ would consist of stepwise recalculation of regression models, where at each step one variable would be entered or rejected.

However, such work proved unnecessary due to papers by Garside [4] and Schatzoff et al. [11] from the sixties. Both papers came with exactly the same combinatorial idea as [8]. Moreover, the paper [11] contains a figure resembling Fig. 1 (the path through vertices of 4-dimensional cube is shown using two 3-dimensional cubes), and a table of subsets, somewhat similar to Tab. 1 of splits in [8]. (Not much pleasant finding. By no means I intend to accuse myself of plagiarism, and I hope that the reader still trusts me that the core idea in [8] was just independently reinvented.)

Note that works [4] and [11] do not represent the state of art in the field of all possible subsets regression nowadays. They have been surpassed by even more efficient algorithms by Furnival [2] and Furnival and Wilson [3].

4 Gray code(s)

Papers [4] and [11] have not remained the only surprise following COMPSTAT'98.

After the conference, we often discussed possible extensions and applications of the main combinatorial idea of [8] with Jaromír Antoch. Once he asked a student to construct a sequence with properties (P1), (P2) as homework. The young lady came back with solution (D1), (D2). Asked to explain how she had found out the sequence, she said: "It is the Gray code. My boyfriend has learned it at the Technical University."

With the right key word, it was much easier to search for related works...

The Gray code, or *binary reflected Gray code* is named after Frank Gray, who took out a patent for it in 1953 [6]. Diaconis and Holmes in [1] and some other papers give reference [5] which suggests that Gray proposed the code even much earlier, as soon as in 1939. The reference, however, seems to be wrong[1].

The Gray code, the sequence given by (D1) and (D2), is an important instance of a broader concept – so called *combinatorial Gray codes.* According to [10] (an extensive review including over 150 references), *"the term combinatorial Gray code ... is ... used to refer to any method for generating combinatorial objects so that successive objects differ in some prespecified, usually small, way."*

As an example, all permutations of set $\{1, 2, \ldots, n\}$, ordered so that successive ones differ only by the swap of one pair of adjacent elements, may be mentioned. Another example is such a sequence of all spanning trees of a graph that consecutive trees differ only by a single edge. (For more examples, as well as references related to those given here, see [10]).

Combinatorial Gray codes have a number of application fields, including circuit testing, signal encoding, data compression, hashing, information storage and retrieval, computing the permanent, and statistics, too. (For a more complete list and references, see [10].) Among the combinatorial Gray codes appearing in applications, the binary reflected Gray code seems to be the most frequent.

Paper [1], referred to in [10], deals with various ways of using Gray codes (including, but not not only, the binary reflected Gray code) in statistics. Emphasis is put, besides other topics, on permutation tests and exhaustive calculation of bootstrap distributions.

[1] Standard library services have not found the article in the given journal volume. Moreover, the list of all Gray's works published in the Bell System Technical Journal, provided kindly by Diane Lambert from the Statistics and Data Mining Research of Bell Laboratories in Murray Hill, does not include such a paper.

5 Missing references and Gray code reinventing

In this section, the term *Gray code* will mean just the binary reflected Gray code. (The more general topic of combinatorial Gray codes will be left.)

Gray code is an old, simple, beautiful and extensively useful idea. It seems to possess a high potential of being repeatedly reinvented, especially in statistics, since it appears to be widely unknown in the statistical community.

Note that not only the authors of [8] were ignorant of the existence of the Gray code when reinventing it in 1998. Neither the referee, nor the audience of the COMPSTAT lecture (though apparently not apathetic) could have been aware of previous related works. (One of the most active COMPSTAT'98 discussants urged me, in an informal conversation, to publish the result in a journal as soon as possible. Then, as he thought, there was a good chance that the algorithm would bear my name.)

Note, too, that the authors of both [4] and [11] came independently to the same combinatorial principle, and none of them was aware of the fact that it had already been given the name Gray code. The key word is missing in both of [4] and [11]. Thus, no wonder that references [4] and [11] are missing, in turn, in [10]. The fact that [4] and [11] remain unmentioned in [1], too, might have (besides obsolence) a similar reason.

By the way, Gray was not the first to invent the Gray code. According to [7], the Gray code was used by Emile Baudot's telegraph, awarded a gold medal at the Universal Exposition in Paris in 1878.

I believe that the history of Gray code must, in reality, be even much longer. At the same time, I hope that many colleagues, ignoring this paper, will enjoy their own Gray code reinventions.

References

[1] Diaconis P. and Holmes S. (1994). *Gray codes for randomization procedures.* Statistics and Computing **4**, 287 – 302.

[2] Furnival G. M. (1971). *All possible regressions with less computation.* Technometrics **13**, 403 – 408.

[3] Furnival G. M. and Wilson R. W. (1974). *Regressions by leaps and bounds.* Technometrics **16**, 499 – 511.

[4] Garside M. J.(1965). *The best sub-set in multiple linear regression.* Applied Statistics, Journal of the Royal Statistical Society, Series C **14**, 196 – 200.

[5] Gray F. (1939). *Coding for data transmission.* Technical report, Bell System Technical Journal.

[6] Gray F. (1953). *Pulse code communications.* U.S. Patent 2632058.

[7] Heath F. G. (1972). *Origins of the binary code.* Scientific American **227**, 76 – 83.

[8] Klaschka J. and Mola F. (1998). *Minimization of computational cost in tree-based methods by a proper ordering of splits.* COMPSTAT 1998, Proceedings in Computational Statistics (eds. R. Payne and P. Green), Physica-Verlag, Heidelberg, 359 – 364.

[9] Klaschka J., Siciliano R. and Antoch J. (1998). *Computational enhancements in tree-growing methods.* Advances in Data Science and Classification, Proceedings of IFCS'98 (eds. A. Rizzi, M. Vichi and H.-H. Bock), Springer, Berlin, 295 – 302.

[10] Savage C. (1997). *A survey of combinatorial Gray codes.* SIAM Review **39**, 605 – 629.

[11] Schatzoff M., Tsao R. and Fienberg S. (1968). *Efficient calculation of all possible regressions.* Technometrics **10**, 769 – 779.

Acknowledgement: This work was supported by grants 201/02/1456 from the Czech Science Foundation and ME 701 from the Ministry of Education, Youth and Sports of the Czech Republic. I am grateful to Diane Lambert for her help with bibliographical search (see a footnote in Section 4).

Address: J. Klaschka, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic

E-mail: klaschka@cs.cas.cz

COMPSTAT'2004 Symposium

Q&A - VARIABLE MULTIPLE CHOICE EXERCISES WITH COMMENTED ANSWERS

Siegbert Klinke

Key words: e-teaching, web-based teaching, multiple choice exercises, MM-Stat, E-Stat.

COMPSTAT 2004 section: Teaching statistics.

Abstract: This contribution is devoted to the author's experience with the construction and the use of two e-teaching software packages, i.e. MM-Stat and E-Stat.

1 Introduction

Currently at our university we have two developments: first the number of students in economics grows every year; second the teaching hours for the exercise classes in statistics are slowly reduced due to general financial cuts.

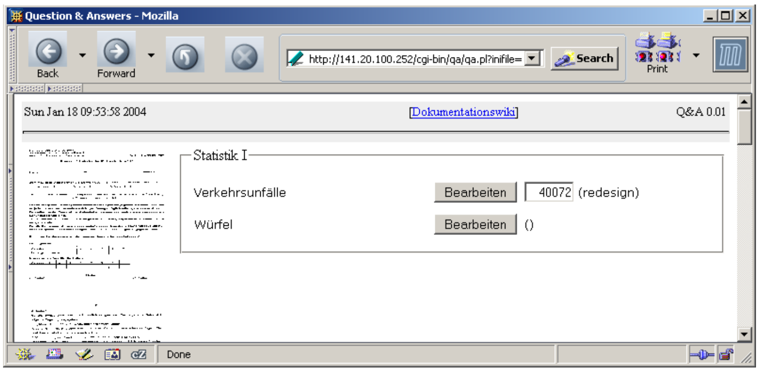
Because of examination regulations we have to offer the written exam twice per semester break. Every student must have the possibility to take the second exam if he fails the first one. We have decided to use "Multiple-Choice"-type questions. This type of exam requires an intensive preparation, but allows for fast correction. Since it allows the student easily to crib, we usually prepare two to three different versions of the exam.

In the institute we were involved in the development of two e-teaching software packages: MM-Stat by Härdle and Rönz [2] and E-Stat described in Cramer, Cramer and Kamps [3]. Both packages include exercises for the students. MM-Stat only involves true/false multiple-choice questions that are presented in random order. Interactivity is in E-Stat restricted to the use of applets. Since exercises require interactivity a teacher has to use applets or has to link to external web pages.

2 Q&A software

The software we develop should be able to handle two types of exercises: "simple" and "variable" exercises. In both types of exercises we present to the student a problem with multiple answers where he can select no, one or several answers (see Figures 2 and 3). When we prepare different versions for an exam then we usually change the numerical values given in an exercise. Another possibility, the permutation of the order of the exercises, leads in our experience to a correlation to the average number of points for one version. For example putting the difficult exercises (with more points) at the beginning of the exam leads (statistical not significant) to a larger average number of points in the exam.

In Figure 1 you see beside the exercise “Verkehrsunfälle” (mode estimation) a seed number. Every time you reload the page in the browser you will get another random seed; but the user can also enter himself a random seed. The random seed determines which numerical values in the exercise will be used; the same random seed produces the same exercise.



http://141.20.100.252/cgi-bin/qa/qa.pl?inifile=

Sun Jan 18 09:53:58 2004 [Dokumentationswiki] Q&A 0.01

Statistik I

Verkehrsunfälle 40072 (redesign)

Würfel ()

Figure 1: A “simple” and “variable” exercise. Note the random seed that determines the numerical values used in the exercise.

3 Simple exercise

Let us consider now the “simple” exercise shown in Figure 2: *Assume you throw two dices. What is the probability that one dice shows a* 2 *and the other a* 3 *under the condition that the sum of both dices should be* 5 *?*

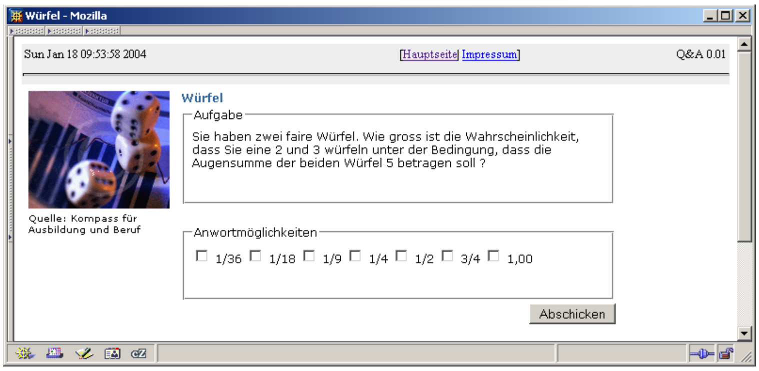
The students typically make two mistakes. They do not recognize that they have

(1) to compute a conditional probability and
(2) to take into account that there are two elementary events.

Depending on the answers the student chooses, we can trace back what kind of mistakes they most probably have done. We categorized the answers of 215 students in seven classes

(a) (1) wrong, (2) wrong	23%,
(b) (1) wrong, (2) correct	39%,
(c) (1) correct, (2) wrong	6%,
(d) (1) correct, (2) correct	16%,
(e) not answered at all	1%,
(f) an answer, which belongs to the other version	5%,
(g) other answers	10%.

Thus we have prepared four different web pages as commented answers to the exercise in Figure 2, one for each of the solutions (a)–(d).



Sun Jan 18 09:53:58 2004

Würfel

Aufgabe

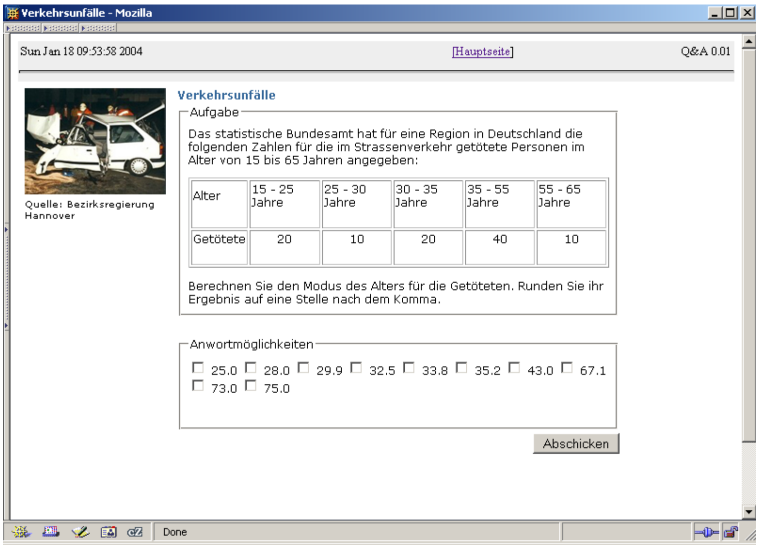
Sie haben zwei faire Würfel. Wie gross ist die Wahrscheinlichkeit, dass Sie eine 2 und 3 würfeln unter der Bedingung, dass die Augensumme der beiden Würfel 5 betragen soll ?

Quelle: Kompass für Ausbildung und Beruf

Anwortmöglichkeiten

□ 1/36 □ 1/18 □ 1/9 □ 1/4 □ 1/2 □ 3/4 □ 1,00

Figure 2: A "simple" exercise: Throwing two dices.



Sun Jan 18 09:53:58 2004

Verkehrsunfälle

Aufgabe

Das statistische Bundesamt hat für eine Region in Deutschland die folgenden Zahlen für die im Strassenverkehr getötete Personen im Alter von 15 bis 65 Jahren angegeben:

Alter	15 - 25 Jahre	25 - 30 Jahre	30 - 35 Jahre	35 - 55 Jahre	55 - 65 Jahre
Getötete	20	10	20	40	10

Berechnen Sie den Modus des Alters für die Getöteten. Runden Sie ihr Ergebnis auf eine Stelle nach dem Komma.

Quelle: Bezirksregierung Hannover

Anwortmöglichkeiten

□ 25.0 □ 28.0 □ 29.9 □ 32.5 □ 33.8 □ 35.2 □ 43.0 □ 67.1
□ 73.0 □ 75.0

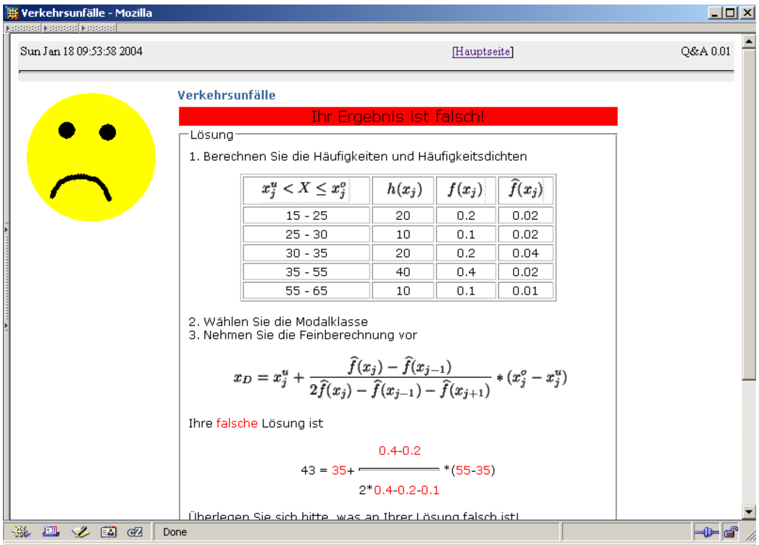
Figure 3: A "variable" exercise: Estimating the mode for grouped data (seed=40072).

4 Variable exercises

Figure 3 shows an example for a variable exercise: *Assumed the data given in the table what is the mode of the grouped data?* In the exercise the seed determines the group widths and the frequencies for each group. We can generate approximately 4.500 different exercises with five groups.

The main problem for the students is to recognize that the group widths are different for each group and to take that into account for the computation. In a written exam only 32% of 194 students found the correct solution.

Figure 4 shows the answer that the student obtains when he chooses wrong answer. In the commented answer we do not say the student directly what is wrong. We just give a hint, e.g. by red colored text, where he is wrong.



Verkehrsunfälle - Mozilla

Sun Jan 18 09:53:58 2004 [Hauptseite] Q&A 0.01

Verkehrsunfälle

Ihr Ergebnis ist falsch!

Lösung

1. Berechnen Sie die Häufigkeiten und Häufigkeitsdichten

$x_j^u < X \leq x_j^o$	$h(x_j)$	$f(x_j)$	$\widehat{f}(x_j)$
15 - 25	20	0.2	0.02
25 - 30	10	0.1	0.02
30 - 35	20	0.2	0.04
35 - 55	40	0.4	0.02
55 - 65	10	0.1	0.01

2. Wählen Sie die Modalklasse
3. Nehmen Sie die Feinberechnung vor

$$x_D = x_j^u + \frac{\widehat{f}(x_j) - \widehat{f}(x_{j-1})}{2\widehat{f}(x_j) - \widehat{f}(x_{j-1}) - \widehat{f}(x_{j+1})} * (x_j^o - x_j^u)$$

Ihre falsche Lösung ist

$$43 = 35 + \frac{0.4-0.2}{2*0.4-0.2-0.1} * (55-35)$$

Überlegen Sie sich bitte, was an Ihrer Lösung falsch ist!

Done

Figure 4: A “variable” exercise: Commented answer to a wrong solution to the mode estimation of grouped data.

5 Implementation

For implementing our software a lot of techniques are available: JavaScript, Java Applets, Java Server Pages, CGI with Perl- or PHP-scripts, PHP with a MySQL database etc. For example Java Applets are used in the JUMBO project [4] and PHP with MySQL by Bartels [1]. We wanted a robust and easy-to-install solution, thus we decided for CGI with Perl scripts. Figure 5 shows the complete structure of the Q&A software. It consists of three Perl scripts for building the web pages from text files and displaying them and one C program to analyze the answer of the student.

The generation of an exercise for the Q&A system should be as easy as possible for a teacher. For a simple exercise the teacher's contribution consists only in creating HTML files and editing text files. The contents of the text files, which are similar to MS Window's INI files, are entered into template HTML files and then are displayed to the student. Implementing the “two dices” exercise required a little bit less than 1,5 hours.

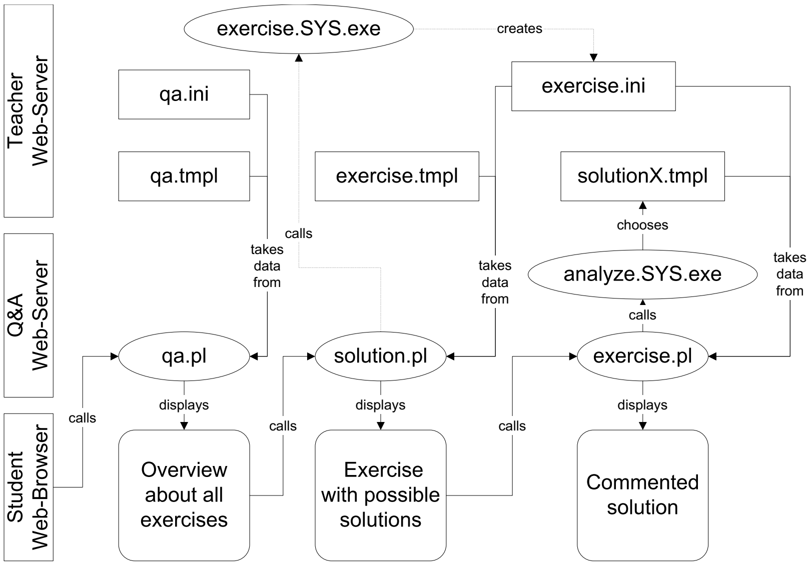


Figure 5: The structure of the Q&A software separated by teacher and software provided parts.

For variable exercises it is necessary to write a program (see in Figure 5 `exercise.SYS.exe`) that generates the appropriate text files. We wrote our programs in Perl and C, but it can be written in any programming language as long as it is able to create executable programs for the different operating systems (currently: MS Windows and Linux). For further details, see the documentation Wiki at `http://141.20.100.252/qa/`.

With Indigoperl [5] we have a Perl environment with an integrated Apache Web server for MS Windows available. Thus we can offer the students to download the whole system rather than to do the exercises online.

6 Conclusion and outlook

This work showed us that the construction of good exercises is a very difficult task. Especially when we try to figure out what kind of mistakes the students could possibly make. Currently (January 2004) the first evaluation period has started. We plan to bring more exercises into the system. Open is the question if we should integrate true/false multi-choice questions as in MM-Stat.

For access to `http://141.20.100.252/qa/` please contact me via e-mail. Since we are create exercises for undergraduate students, they are only in German available.

References

[1] Bartels K. (2002). *e-stat: Automatic evaluation of online exercises.* In Proceedings in Computational Statistics, W. Härdle, B. Rönz, (Hrsg.), Physica-Verlag, Heidelberg, 315 – 320.

[2] Härdle W., Rönz B. (2003). *Statistik - Wissenschaftliche Datenanalyse leicht gemacht (multilingual edition,* formerly MM-Stat*).* `http://www.mhsg.de`.

[3] Cramer E., Cramer K., Kamps U. (2002). *e-stat: A web-based learning environment in applied statistics.* In Proceedings in Computational Statistics, W. Härdle, B. Rönz (eds.), Physica-Verlag, Heidelberg, 309 – 314. `http://www.e-stat.de`.

[4] Köpcke W., Heinecke A. (1998) *Die JAVA-unterstützte Münsteraner Biometrie Oberfläche (JUMBO).* In Computer Based Training in der Medizin, M. Adler, J.W. Dietrich, M.F. Holzer, M.R. Fischer (eds.), Shaker Verlag, Aachen.

[5] IndigoStar software (2002). *IndigoPerl 5.6 build 09*, `http://www.indigostar.com/indigoperl.htm`

Acknowledgement: The research and the development of the software is supported by the Multimedia-Förderprogram 2003 “Neue Medien in Lehre und Forschung” of the Humboldt-Universität zu Berlin and the Multimedia Lehr- und Lernzentrum.

Address: S. Klinke, Humboldt-Universität zu Berlin, School of Business and Economics, Institute for Statistics and Econometrics, Spandauer Strasse 1, D-10178 Berlin, Germany

E-mail: `sigbert@wiwi.hu-berlin.de`

USE OF FOURIER TRANSFORMATION FOR KERNEL SMOOTHING

Jan Koláček

Key words: Bandwidth selection, Fourier transform, kernel estimation, nonparametric regression.

COMPSTAT 2004 section: Nonparametrical statistics.

Abstract: The problem of bandwidth selection for non-parametric kernel regression is considered. We will follow the Nadaraya - Watson and local linear estimators especially. The circular design is assumed in this work to avoid the difficulties caused by boundary effects. Most of bandwidth selectors are based on the residual sum of squares (RSS). It is often observed in simulation studies that these selectors are biased toward undersmoothing. This leads to consideration of a procedure which stabilizes the RSS by modifying the periodogram of the observations. Simulation studies suggest that the proposed selector has preferable properties than the classical one.

1 Basic terms and definitions

Consider a standard regression model of the form

$$Y_t = m(x_t) + \varepsilon_t, \qquad t = 0, \dots, T-1, \quad T \in \mathbb{N},$$

where m is an unknown regression function, x_t are design points, Y_t are measurements and ε_t are independent random variables for which

$$E(\varepsilon_t) = 0, \qquad var(\varepsilon_t) = \sigma^2 > 0, \qquad t = 0, \dots, T-1.$$

The aim of kernel smoothing is to find suitable approximation $\hat{m}$ of an unknown function m.

In next we will assume

1. The design points x_t are equidistantly distributed on the interval $[0,1)$, that is $x_t = t/T, \ t = 0, \dots T-1$.
2. We use a "cyclic design", that is, suppose $m(x)$ is a smooth periodic function and the estimate is obtained by applying the kernel on the extended series $\tilde{Y}_t$, where $\tilde{Y}_{t+kT} = Y_t$ for $k \in \mathbb{Z}$. Similarly $x_t = t/T$, $t \in \mathbb{Z}$.

$Lip[a,b]$ denotes the class of continuous functions satisfying the inequality

$$|g(x) - g(y)| \leq L \cdot |x - y|, \quad \forall x, y \in [a,b], L > 0, L \text{ is a constant.}$$

Definition. Let κ be a nonnegative even integer and assume $\kappa \geq 2$. The function $K \in Lip[-1,1]$, support$(K) = [-1,1]$, satisfying the following conditions

$$(i) \qquad K(-1) = K(1) = 0$$

$$(ii) \qquad \int\limits_{-1}^{1} x^j K(x)dx = \begin{cases} 0, & 0 < j < \kappa \\ 1, & j = 0 \\ \beta_\kappa \neq 0, & j = \kappa, \end{cases}$$

is called a *kernel* of order κ and a class of all these kernels is marked $S_{0\kappa}$.

These kernels are used for an estimation of the regression function (see [4]). Let $K \in S_{0\kappa}$, set $K_h(.) = \frac{1}{h}K(\frac{.}{h})$, $h \in (0, 1)$. A parameter h is called a *bandwidth*.

2 Kernel estimation of the regression function

Commonly used non-parametric methods for estimating $m(x)$ are the kernel estimators

1. **Nadaraya - Watson estimators** (Nadaraya & Watson 1964)

$$\hat{m}_{NW}(x; h) = \frac{\sum\limits_{k=-T}^{2T-1} K_h(x_k - x)\tilde{Y}_k}{\sum\limits_{k=-T}^{2T-1} K_h(x_k - x)}$$

2. **Local linear estimators** (Stone 1977, Cleveland 1979)

$$\hat{m}_{LL}(x; h) = \frac{1}{T} \sum_{k=-T}^{2T-1} \frac{\{\hat{s}_2(x; h) - \hat{s}_1(x; h)(x_k - x)\}K_h(x_k - x)\tilde{Y}_k}{\hat{s}_2(x; h)\hat{s}_0(x; h) - \hat{s}_1(x; h)^2}$$

where

$$\hat{s}_r(x; h) = \frac{1}{T} \sum_{k=-T}^{2T-1} (x_k - x)^r K_h(x_k - x)$$

In the cyclic design, the kernel estimators can be generally expressed as

$$\hat{m}(x; h) = \sum_{k=-T}^{2T-1} W_k^{(j)}(x)\tilde{Y}_k,$$

where the weights $W_k^{(j)}(x)$, $j \in \{NW, LL\}$ correspond to the weights of estimators $\hat{m}_{NW}$, $\hat{m}_{LL}$.

From the assumption of the circular model results the interesting fact, that the weights of Nadaraya-Watson and local linear estimators are identical at design points, that is

$$W_k^{(LL)}(x_t) = W_k^{(NW)}(x_t),$$

for $k \in \{-T, -T-1, \ldots, 2T-1\},\ t \in \{0, 1, \ldots, T-1\}$, so in next, we will write only $W_k(x_t)$ without upper index.

Let $K \in S_{0\kappa}, h \in (0,1), t \in \{0, \ldots, T-1\}$. Then the sum $\sum\limits_{k=-T}^{2T-1} K_h(x_k - x_t) = \sum\limits_{k=-T+1}^{T-1} K_h(x_k)$ is independent on t. Set $C := \sum\limits_{k=-T+1}^{T-1} K_h(x_k)$. We can simply write the value of weight functions at design points $x_t,\ t = 0, \ldots, T-1$

$$W_k(x_t) = \frac{1}{C} K_h(x_k - x_t).$$

3 Choosing the smoothing parameter h

The optimal bandwidth h minimizes the average mean squared error

$$(AMSE) \qquad R_T(h) = \frac{1}{T} E \sum_{t=0}^{T-1} (m(x_t) - \hat{m}(x_t; h))^2$$

There are many estimators of this error function, which are asymptotically equivalent and asymptotically unbiased (see [2]). However, in simulation studies, it is often observed that most of selectors are biased toward undersmoothing and give smaller bandwidths more frequently than predicted by asymptotic results. Most of bandwidth selectors are based on the residual sum of squares

$$(RSS) \qquad RSS_T(h) = \frac{1}{T} \sum_{t=0}^{T-1} [Y_t - \hat{m}(x_t; h)]^2.$$

For example Rice (see [3]) considered

$$\hat{R}_T(h) = RSS_T(h) - \hat{\sigma}^2 + \frac{2K_h(0)\hat{\sigma}^2}{TC}, \tag{1}$$

where $\hat{\sigma}^2$ is an estimate of σ^2

$$\hat{\sigma}^2 = \frac{1}{2T-2} \sum_{t=1}^{T-1} (Y_t - Y_{t-1})^2.$$

The estimate $\hat{h}_{opt}$ of optimal bandwidth is defined as

$$\hat{h}_{opt} = \arg\min \hat{R}_T(h).$$

One of the classical bandwidth selectors is the *leave-out method*. This method is based on regression smoothers in which *one*, say the kth, observation is left out

$$\hat{m}_k(x_k; h) = \sum_{i \neq k} W_i(x_k) Y_i.$$

With these modified smoothers, the selector has the form

$$CV(h) = \frac{1}{n}\sum_{t=0}^{T-1}[\hat{m}_t(x_t;h) - Y_t]^2. \tag{2}$$

The function $CV(h)$ is commonly called a *cross-validation function.*

4 Fourier transformation

Let $S_t = m(x_t)$. The periodogram of the series $Y_t, t = 0, \ldots T-1$, is defined by I_{Y_λ}

$$I_{Y_\lambda} = |Y_\lambda^-|^2/2\pi T,$$

where

$$Y_\lambda^- = \sum_{k=0}^{T-1} Y_k e^{-\frac{i2\pi k\lambda}{T}}$$

is the finite Fourier transform of the series Y_t. This transformation we denote $Y^- = DFT^-(Y)$.

The periodograms and Fourier transforms of the series ε_t and S_t are defined similarly. Under mild conditions, the periodogram ordinates I_{ε_j} on the Fourier frequencies $\frac{2\pi j}{T}$, for $j = 1, \ldots, N = \left[\frac{T-1}{2}\right]$, are approximately independently and exponentially distributed with means $\frac{\sigma^2}{2\pi}$. Here $[x]$ means the greatest integer which is less or equal to x.

In the next the Parseval's identity

$$\sum_{t=0}^{T-1}|x_t|^2 = \frac{1}{T}\sum_{t=0}^{T-1}|x_t^-|^2, \qquad \mathbf{x} \in \mathbb{C}^T, \quad \mathbf{x}^- = DFT^-(\mathbf{x}) \tag{3}$$

will be useful.

Definition Let $\boldsymbol{x} = (x_0, \ldots, x_{T-1}), \boldsymbol{y} = (y_0, \ldots, y_{T-1}) \in \mathbb{C}^T$;

$$z_t = \sum_{k=0}^{T-1} x_{<t-k>_T} y_k,$$

where $< t-k >_T$ marks $(t-k) \mod T$. Then $\boldsymbol{z} = (z_0, \ldots, z_{T-1})$ is called *the discrete cyclic convolution* of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$; we write $\boldsymbol{z} = \boldsymbol{x} \star \boldsymbol{y}$.

Note Let $\boldsymbol{x}, \mathbf{y} \in \mathbb{C}^{\mathbf{T}}$, then $(\boldsymbol{x} \star \boldsymbol{y})^- = \boldsymbol{x}^- \cdot \boldsymbol{y}^-$, where $\boldsymbol{x}^- \cdot \boldsymbol{y}^-$ means the scalar product of vectors $\boldsymbol{x}^-$,$\boldsymbol{y}^-$.

Let's define a vector $\mathbf{w} := [w_0, w_1, \ldots, w_{T-1}]$, where

$$w_t = W_0(x_t - 1) + W_0(x_t) + W_0(x_t + 1).$$

Let $h \in (0,1)$, $K \in S_{0\kappa}, t \in \{0, \ldots, T-1\}$. Then we can write $\hat{m}(x_t;h)$ as a discrete cyclic convolution of vectors $\mathbf{w}$ and Y.

$$\hat{m}(x_t;h) = \sum_{k=0}^{T-1} w_{<t-k>_T} Y_k. \tag{4}$$

Theorem Let $h \in (0,1)$, $K \in S_{0\kappa}$. Then

$$RSS_T(h) = \frac{4\pi}{T}\sum_{t=1}^{N} I_{Y_t}\left\{1 - w_t^-\right\}^2, \tag{5}$$

where $w_t^- = \sum\limits_{k=-T+1}^{T-1} W_0^{(j)}(x_k)e^{-\frac{i2\pi kt}{T}}$ is the finite Fourier transform of $\mathbf{w}$.

Proof

$$\begin{aligned} RSS_T(h) &= \frac{1}{T}\sum_{t=0}^{T-1}|Y_t - \hat{m}(x_t;h)|^2 \overset{(4)}{=} \frac{1}{T}\sum_{t=0}^{T-1}\left|Y_t - \sum_{k=0}^{T-1} w_{<t-k>_T}Y_k\right|^2 \\ &= \frac{1}{T}\sum_{t=0}^{T-1}|Y_t - (\mathbf{w}\star Y)_t|^2 \overset{(3)}{=} \frac{2}{T^2}\sum_{t=1}^{N}\left|Y_t^- - w_t^-Y_t^-\right|^2 \\ &= \frac{4\pi}{T}\sum_{t=1}^{N} I_{Y_t}\left\{1 - w_t^-\right\}^2. \end{aligned}$$

From (1) and (5) we arrive at the equivalent expression for $\hat{R}_T(h)$

$$\hat{R}_T(h) = \frac{4\pi}{T}\sum_{t=1}^{N} I_{Y_t}\{1 - w_t^-\}^2 - \hat{\sigma}^2 + \frac{2\hat{\sigma}^2}{CT}K_h(0). \tag{6}$$

Similarly,

$$R_T(h) = \frac{4\pi}{T}\sum_{t=1}^{N}\left\{I_{S_t} + \frac{\sigma^2}{2\pi}\right\}\{1 - w_t^-\}^2 - \sigma^2 + \frac{2\sigma^2}{CT}K_h(0). \tag{7}$$

5 The proposed selector

Let $D(h) = \hat{R}_T(h) - R_T(h)$. From previous expressions we obtain

$$D(h) = \frac{4\pi}{T}\sum_{t=1}^{N}\left\{I_{Y_t} - I_{S_t} - \frac{\sigma^2}{2\pi}\right\}\{1 - w_t^-\}^2. \tag{8}$$

The periodogram ordinates I_{S_t} decrease rapidly for smooth $m(x)$. So I_{Y_t} do not contain much information about I_{S_t} at high frequencies (for the rigorous proof see [3]). This leads to the consideration of the procedure described below. The main idea is to modify RSS to make it less variable. We find the first index J such that $I_{Y_J} < c\hat{\sigma}^2/2\pi$ for some constant $c > 1$, where $\hat{\sigma}^2$ is an estimate of σ^2. The constant c sets a threshold. In our experience, setting $1 < c < 3$ yields good results.

The modified residual sum of squares is defined by

$$MRSS_T(h) = \frac{2\pi}{T}\sum_{t=0}^{T-1}\tilde{I}_{Y_t}\{1 - W_t\}^2,$$

where

$$\tilde{I}_{Y_t} = \begin{cases} I_{Y_t}, & t < J \\ \hat{\sigma}^2/2\pi, & t \geq J, \end{cases}$$

(see Fig.1, Fig.2).

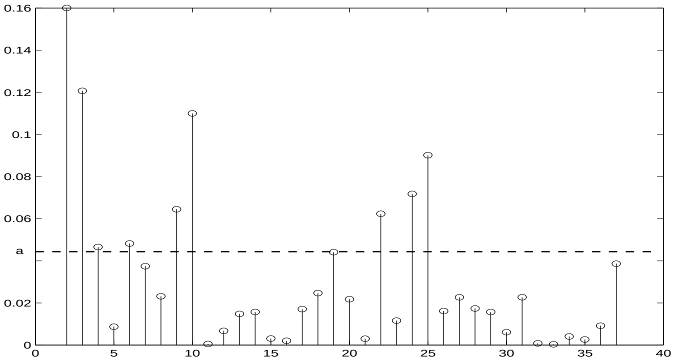


Figure 1: The periodogram ordinates I_{Y_t} as a function of t, $a = 2\frac{\hat{\sigma}^2}{2\pi}$.

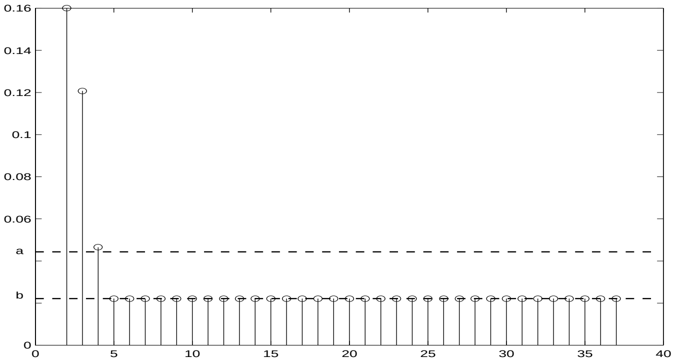


Figure 2: The modified periodogram ordinates $\tilde{I}_{Y_t}$ as a function of t, $a = 2\frac{\hat{\sigma}^2}{2\pi}$, $b = \frac{\hat{\sigma}^2}{2\pi}$.

Thus, the proposed selector is

$$\tilde{R}_T(h) = MRSS_T(h) - \hat{\sigma}^2 + \frac{2\hat{\sigma}^2}{TC} K_h(0) \tag{9}$$

and the new estimate of optimal bandwidth

$$\hat{\hat{h}}_{opt} = \arg\min \tilde{R}_T(h).$$

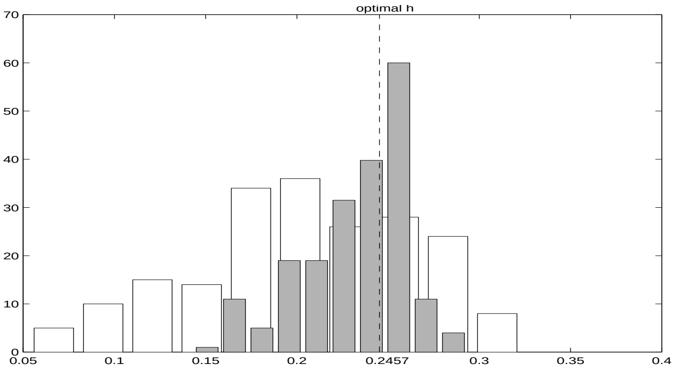


Figure 3: The histogram of results of all 200 experiments obtained by Rice's selector (white) and by our proposed selector (grey).

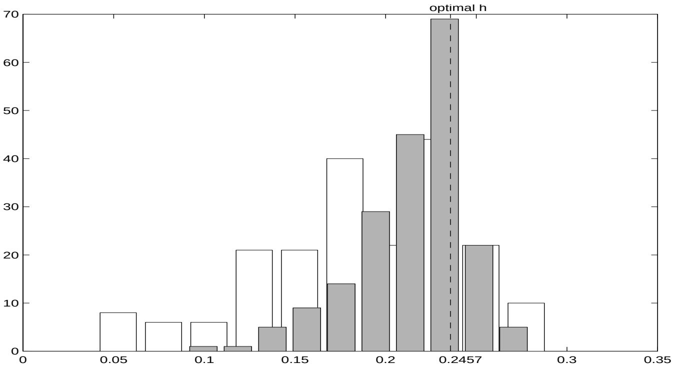


Figure 4: The histogram of results of all 200 experiments obtained by cross-validation method (white) and by our proposed selector (grey).

Set $\tilde{D}(h) = \tilde{R}_T(h) - R_T(h)$. From (7) and (9) we have

$$\tilde{D}(h) = \frac{4\pi}{T}\sum_{t=1}^{J-1}\left\{I_{Y_t} - I_{S_t} - \frac{\sigma^2}{2\pi}\right\}\{1-w_t^-\}^2 - \frac{4\pi}{T}\sum_{t=J}^{N} I_{S_t}\{1-w_t^-\}^2. \qquad (10)$$

Since I_{S_t} decays rapidly for smooth $m(x)$, the second term in (10) is negligible. By comparison of (8) and (10) it can be concluded that our proposed selector has preferable properties than the classical one.

6 A simulation study

We carried out a small simulation study to compare the performance of the bandwidth estimates. The observations, Y_t, for $t = 0, \ldots, T = 100$, were obtained by adding independent Gaussian random variables with mean zero and variance $\sigma^2 = 0.2$ to the function

$$m(x) = \cos(9x - 7) - (3 + x^{12})/6 + 8^{x-1}.$$

The theoretical optimal bandwidth (see [4]) is $h_{opt} = 0.2457$ for the kernel of order (0,4)

$$K(x) = \begin{cases} \frac{15}{16}(\frac{7}{2}x^4 - 5x^2 + \frac{3}{2}), & |x| \le 1 \\ 0, & |x| > 1. \end{cases}$$

Two hundred series were generated. Table 1 summarizes the sample means and the sample standard deviations of bandwidth estimates, $E(\hat{h})$ is the average of all 200 values and $std(\hat{h})$ is their standard deviation.

Selector	$E(\hat{h})$	$std(\hat{h})$
CV	0.1853	0.0546
$\hat{R}$	0.1985	0.0614
$\tilde{R}$	0.2317	0.0291

Table 1: Summary of sample means and standard deviations of bandwidth estimates.

7 Conclusion

This method was proposed by Chiu (1990) (see [1]) for a special class of symmetric probability density functions from S_{02} and for Pristley-Chao estimators. In this paper, this approach has been generalized for kernels from the class $S_{0\kappa}$, κ even and for Nadaraya-Watson and local linear estimators.

References

[1] Chiu S.T. (1990). *Why bandwidth selectors tend to choose smaller bandwidths, and a remedy.* Biometrika **77**, 222 − 226.

[2] Härdle W. (1990). *Applied nonparametric regression.* Cambridge University Press, Cambridge.

[3] Rice J. (1984). *Bandwidth choice for nonparametric regression.* The Annals of Statistics **12**, 1215 − 1230.

[4] Wand M.P., Jones M.C. (1995). *Kernel Smoothing.* Chapman & Hall, London.

Acknowledgement: Supported by the CEZ: J07/98: 14310001

Address: Jan Koláček, PhD student, Department of Applied mathematics, Faculty of Science, Janáčkovo nám. 2a, 602 00 Brno, Czech Republic

E-mail: `kolacek@math.muni.cz`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

RANK ESTIMATORS FOR THE TIME OF A CHANGE IN CENSORED DATA

Lenka Komárková

Key words: Rank estimators, change-point, censored data, limit behavior.

COMPSTAT 2004 section: Nonparametrical statistics.

Abstract: This article concerns estimation of a change-point in the model with randomly censored data. A class of rank based estimators of the change-point corresponding to the class of weighted log-rank statistics is used and its limit behavior is presented. We demonstrate the usefulness of this technique on the Stanford Heart Transplant Data and we do also simulation study.

1 Model and notation

We introduce the basic notation concerned random censorship model. For more detailed information see e.g. Kalbfleisch and Prentice [5]. Typically, $X_{1n}^0, \dots, X_{nn}^0$ is a sequence of independent nonnegative random variables (*the lifetimes* or *the survival times*), where the index i of X_{in}^0 corresponds to the chronological order in which the subject of interest (e.g. patient) has entered the study. The patient can be withdrawn from the study due to many reasons, e.g. an accidental death, a migration of human population or limited time of the study. More precisely, the lifetimes can be censored from the right by independent random variables $C_{1n}, \dots, C_{nn}$, the so-called *censoring times*. In other words, instead of survival times $X_{1n}^0, \dots, X_{nn}^0$ we observe pairs $(X_{1n}, \Delta_{1n}), \dots, (X_{nn}, \Delta_{nn})$ only, where

$$X_{jn} = \min(X_{jn}^0, C_{jn}), \quad \Delta_{jn} = I(X_{jn}^0 \le C_{jn}) = \begin{cases} 1, & \text{if } X_{jn} \text{ is uncensored,} \\ 0, & \text{if } X_{jn} \text{ is censored.} \end{cases}$$

We assume the lifetimes and the censoring times are independent variables. Particularly, their distributions need not be the same over the complete observation period due to e.g. medical development. We suppose that the survival variables $X_{1n}^0, \dots, X_{mn}^0$ and $X_{m+1\,n}^0, \dots, X_{nn}^0$ have the common distribution functions F_{1n} and F_{2n}, respectively, $F_{1n} \ne F_{2n}$, and the censoring variables $C_{1n}, \dots, C_{m_c n}$ and $C_{m_c+1\,n}, \dots, C_{nn}$ have the common distribution function G_{1n} and G_{2n}, respectively, $G_{1n} \ne G_{2n}$. The parameters m and m_c are unknown and the distribution functions F_{1n}, F_{2n} and G_{1n}, G_{2n} are supposed to be absolutely continuous but unknown. We call such model *the random censoring model with the change-point* m (RCM). Therefore we are primarily interested in estimating of the unknown time m, when the distribution of the lifetimes changed.

In case of no censoring there has been published a number of papers concerning estimators of the change-point m, e.g. Antoch et al [1]. Detection of a change point and related problems can be also found in Csörgő and Horváth [2]. In case of censored data there have been published only a few papers. We mention the work of Gombay and Liu [3] who based their detection of change point on a generalization of Wilcoxon's rank statistic.

We would like to detect and estimate the change-point m, when m_c is nuisance parameter. Our estimator of the change-point m is motivated by the test statistic developed by Hušková and Neuhaus [4]. Particularly, they considered testing problem $H_0 : m = n$ (no change in distribution of the lifetimes) against $H_1 : m < n$ (one change in distribution of the lifetimes).

Suppose for a moment that m is known, then we get a two-sample problem for the random censorship. Thus, along the lines of a two-sample rank test for randomly censored data, the test procedure is based on the weighted log-rank type test statistics

$$L_k(\tau_0) = \frac{|S_k(\tau_0)|}{\sqrt{nV_k(\tau_0)}}, \quad k = 1, \dots, n-1, \tag{1}$$

with

$$S_k(\tau_0) = \sum_{j=1}^{k} \left(\int_0^{\tau_0} w_n(t) \, \mathrm{d}N_j(t) - \int_0^{\tau_0} w_n(t) \, \frac{Y_j(t)}{\sum_{j=1}^{n} Y_j(t)} \, \mathrm{d}\Big(\sum_{j=1}^{n} N_j(t) \Big) \right),$$

where $Y_j(t) = I(X_{jn} \geq t)$ and $N_j(t) = \Delta_{jn} \, I(X_{jn} \leq t)$. The variables $V_k(\tau_0)$ are related to the variances of $S_k(\tau_0)$, $k = 1, \dots, n-1$, and $w_n(t)$ is a weight function and they will be specified later. The value τ_0 is chosen as a positive number fulfilling

$$0 < \tau_0 < \tau = \sup\{t; \lim_{n\to\infty} F_{in}(t) < 1, \ \lim_{n\to\infty} G_{in}(t) < 1, \ i = 1, 2\}.$$

Limit properties of $L_k(\tau_0)$ for $\min(k, n-k) \to \infty$ were studied by Neuhaus [7].

But m is unknown, so the change in distribution of the survival variables can occur in an arbitrary time-point $k = 1, \dots, n-1$. Applying *the union-intersection principle* (for more details see e.g. Csörgő and Horváth [2]), we reject H_0 if at least one of $L_k(\tau_0)$, $k = 1, \dots, n-1$, takes large value. This leads to the maximum-type test statistic and the rejection region

$$T_n(\tau_0) = \max_{1\leq k<n} L_k(\tau_0) \geq \frac{-\log(-\log(\sqrt{1-\alpha})) + d_2(\log n)}{d_1(\log n)}, \tag{2}$$

where $d_1(t) = \sqrt{2 \log t}$ and $d_2(t) = 2\log t + \frac{1}{2} \log\log t - \frac{1}{2} \log \pi$.

Next, we consider the local alternative $H_1 : F_{1n} \neq F_{2n}$, where the differences $F_{1n}(t) - F_{2n}(t)$ tend to 0 in a certain way. In this case, we propose the estimator of the change-point m as the point k, where the statistics $L_k(\tau_0)$ takes its maximum, i.e.

$$\hat{m}(\tau_0) = \min\left\{ k : \max_{1\leq j<n} L_j(\tau_0) = L_k(\tau_0) \right\} = \operatorname*{argmax}_{1\leq k<n} L_k(\tau_0) \tag{3}$$

and we will study its limit properties in next two sections.

2 Consistency of estimator

We suppose that the weights $w_n(X_{jn}, \Delta_{jn}) \geq 0$ fulfill, as $n \to \infty$,

$$\sup_{0\leq t\leq \tau_0} |w_n(t) - w(t)| = O_{\mathrm{P}}\left(\sqrt{\frac{\log\log n}{n}}\right), \tag{4}$$

where w is a continuous nonrandom function on $[0, \tau_0]$. The property poses the class of commonly used weights given by

$$w_n(t) = (\hat{S}_n(t-))^{\rho}\left(\frac{\sum_{j=1}^n Y_j(t)}{n}\right)^{\kappa} I\Big(\sum_{j=1}^n Y_j(t) > 0\Big), \tag{5}$$

where $\rho, \kappa \geq 0$ and $\hat{S}_n(t-) = \prod_{i:X_i<t}\left(1 - \frac{\Delta_i}{Y(X_i)}\right)$ is the left-continuous Kaplan–Meier estimate of the survival function.

$$V_k(\tau_0) = \frac{1}{n}\int_0^{\tau_0} w_n^2(t)\frac{\sum_{j=1}^k Y_j(t)\ \sum_{j=k+1}^n Y_j(t)}{\left(\sum_{j=1}^n Y_j(t)\right)^2}\,\mathrm{d}\Big(\sum_{j=1}^n N_j(t)\Big) + v_k,$$

are appropriate estimator for the variance of $S_k(\tau_0)$, $k = 1, \ldots, n-1$, and v_k ensure that $V_k(\tau_0)$ are bounded away from 0 and have e.g. the following form

$$v_k = \frac{k(n-k)}{n^2}\,(I(k \leq \log\log n) + I(k \geq n - \log\log n)).$$

Further, we assume the following:

(S.1) there exists $0 < \gamma < 1$ such that $m = \lfloor \gamma n \rfloor$;

(S.2) there exists a distribution function $G(t)$ such that $\lim_{n\to\infty}\sup_{0\leq t\leq\tau_0}|G_{in}(t) - G(t)| = 0, \quad i = 1, 2;$

(S.3) there exists a hazard function $\lambda_F(t) = -\frac{\mathrm{d}\log(1-F(t))}{\mathrm{d}t}$ such that

$$\lim_{n\to\infty}\int_0^{\tau_0}|\lambda_{F_{in}}(t) - \lambda_F(t)|\,\mathrm{d}t = 0, \quad i = 1, 2; \tag{6}$$

(S.4) for $A_n(\tau_0) = \int_0^{\tau_0} w(t)(1-F(t))(1-G(t))\,(\lambda_{F_{1n}}(t) - \lambda_{F_{2n}}(t))\,\mathrm{d}t$ we have, as $n \to \infty$,

$$\frac{nA_n^2(\tau_0)}{\log\log n} \to \infty;$$

(S.5) $I(\tau_0) = \int_0^{\tau_0} w^2(t)(1-G(t))\,\mathrm{d}F(t) > 0.$

The assumption (S.2) expresses "closeness" of G_{1n} and G_{2n} and (S.3) "closeness" of F_{1n} and F_{2n}. More precisely, the term $\lambda_{F_{1n}}(t) - \lambda_{F_{2n}}(t)$, which is the difference of the hazard functions for the lifetimes before and after the change-point $m = m(n)$, reflects the discrepancy between the distribution functions F_{1n} and F_{2n}. The assumption (6) entails that $\lim_{n\to\infty} |A_n(\tau_0)| = 0$ and moreover, $\lim_{n\to\infty} \sup_{0\le t\le\tau_0} |F_{in}(t) - F(t)| = 0$, $i = 1, 2$. Hence, the considered alternative ($|A_n(\tau_0)| \approx \sqrt{\frac{\log\log n}{n}}$) is local but not contiguous according to (S.4). The assumption (S.5) is a technical condition ensuring that $\frac{n^2}{k(n-k)} V_k(\tau_0)$ are asymptotically bounded away from zero.

Theorem 2.1. *Let (4) and the assumptions (S.1)–(S.5) be satisfied. Then we have, as $n \to \infty$,*

$$|\hat{m}(\tau_0) - m| = O_{\mathrm{P}}(A_n^{-2}(\tau_0)), \quad i.e. \ \left|\frac{\hat{m}(\tau_0)}{n} - \gamma\right| = O_{\mathrm{P}}(n^{-1}A_n^{-2}(\tau_0)),$$

where $A_n(\tau_0)$ is defined in (S.3).
If $m = n$ and $1 \le m_c \le n$, $F_{1n} = F$ and $G_{in} = G_i$, $i = 1, 2$, then we have for an arbitrary $\varepsilon \in (0, 1/2)$, as $n \to \infty$,

$$\mathrm{P}(\hat{m}(\tau_0) < n\varepsilon) \to \frac{1}{2} \quad and \quad \mathrm{P}(\hat{m}(\tau_0) > n(1-\varepsilon)) \to \frac{1}{2}.$$

Proof. The proof can be found in Komárková [6]. □

It means that under the local alternative H_1 $\hat{m}(\tau_0)/n$ is a consistent estimator of γ, where $m = \lfloor n\gamma \rfloor$ and moreover, under the null hypothesis H_0 the limit distribution of $\hat{m}(\tau_0)$ has two peaks at the beginning and at the end of the observation period.

3 Application and simulation

Example The survival times of patients in the Standford Heart Transplantation Program have been studied extensively, for a description see Kalbfleisch and Prentice [5]. In this case X_{in}^0 denotes survival time (time to death in days) from admission to study and the index i corresponds to the order of the patient acceptance date for transplantation. The patients entered the study randomly between 1967 and 1974. We would like to know if the behavior of patients have changed due to some reasons, e.g. medical development, over the duration of the study from January 1, 1967, to April 1, 1974, and if yes, we would like to estimate when it had happened. Notice that ties occur in the data, so we use the method of randomization to treat this problem.

We use the three types of weights considered in (5)

- *log-rank* type (LR) with $\rho = 0$, $\kappa = 0$;
- *Gehan–Wilcoxon* type (GW) with $\rho = 0$, $\kappa = 1$;
- *Prentice–Wilcoxon* type (PW) with $\rho = 1$, $\kappa = 0$.

Table 1 contains the results of the test for detection of change point and the corresponding estimator $\hat{m}(\tau_0)$. The values of the test statistic $T_n(\tau_0)$ for the three types of weights are summarized in the left part of the mentioned table. In Figure 1, we see a plot of the statistics $L_k(\tau_0)$, $k = 1, \dots, 195$ used to compute $T_n(\tau_0)$.

	test		estimate
w_n	$T_n(\tau_0)$	p-value	$\hat{m}(\tau_0)$
LR	5.006	0.005	7
GW	4.634	0.009	7
PW	4.662	0.009	7

Table 1: The results of the test.

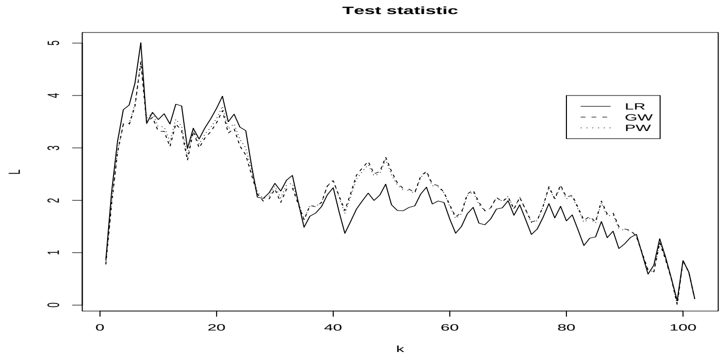


Figure 1: The statistics $L_k(\tau_0)$ for tree types of weights (LR, GW, PW).

The test for the considered weights gives p-values smaller than $\alpha = 0.05$, that is why we reject the hypothesis of no-change in survival times for the significance level $\alpha = 5\%$ and we use our estimator for determination of m. The change-point estimator is $\hat{m}(\tau_0) = 7$ which corresponds to 0.7 of a year since the opening date of the study.

Simulation Study We have prepared a small simulation study in statistical software *R* illustrating properties of the estimators $\hat{m}(\tau_0)$. Suppose RCM with $m = \lfloor n\gamma \rfloor$.

We proceed with $n = 200$ as follows:

1. The survival times $X^0_{1n}, \dots, X^0_{nn}$ are simulated using the chosen combination of parameters $X^0_{in} = \delta_n \, I(i > \lfloor n\gamma \rfloor) + \varepsilon_i$ for $i = 1, \dots, n$ (we use $\delta_n = 0; 0.5; 1$, $\gamma = 0.25; 0.5$, $\varepsilon_i \sim F$, $F = \text{Exp}(1)$ or $\text{LN}(0, 1)$).
2. The censoring times $C_{1n}, \dots, C_{nn}$ are simulated using the chosen combination of parameters $C_{in} = \delta_{C,n} \, I(i > \lfloor n\eta \rfloor) + \epsilon_i$ for $i = 1, \dots, n$ (we use $\delta_{C,n} = 1; 2$, $\eta = 0.25; 0.5$, $\epsilon_i \sim G$, $G = F$).

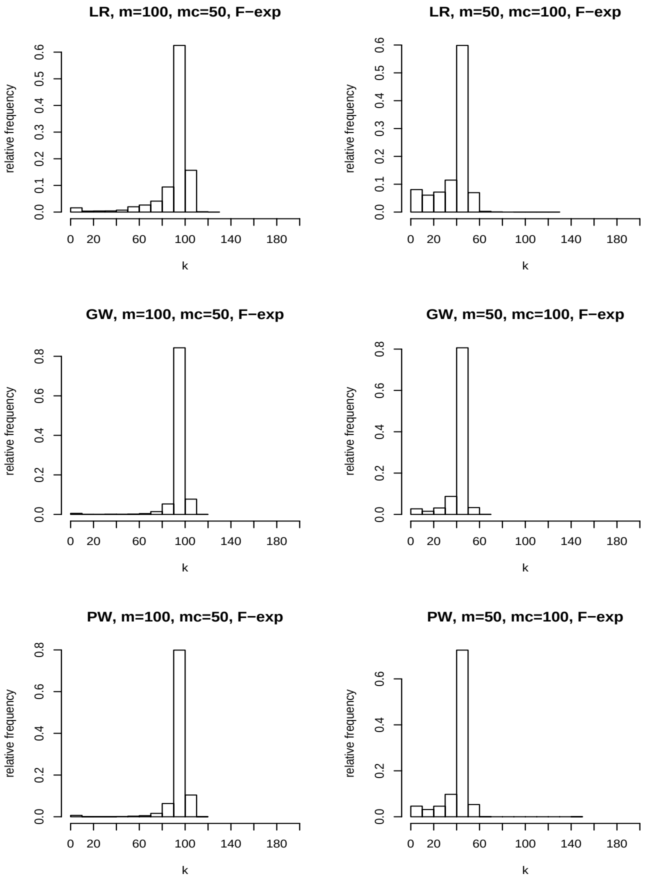


Figure 2: The histogram of $\hat{m}(\tau_0)$ for the parameters $\gamma = 0.5$ or 0.25 and $\eta = 0.25$ or 0.5 and the size of the change $\delta_n = \delta_{C,n} = 1$ and type errors $F = G \sim \text{Exp}(1)$ and all the tree types of weights (LR, GW, PW).

3. The pairs of observations $(X_{1n}, \Delta_{1n}), \ldots, (X_{nn}, \Delta_{nn})$ are computed.
4. The estimator $\hat{m}(\tau_0)$ is calculated and its value stored.
5. The steps $(1) - (4)$ are repeated 10^4 times.
6. The histogram with relative frequency of $\hat{m}(\tau_0)$ is drawn.

Notice that for the observations X_{in}'s before $\min(m_c, m)$ is the expected proportion of censoring 50% due to $F = G$. We can see in Figure 2 and 3 that the peak of the histograms of $\hat{m}(\tau_0)$ is in the neighborhood of m. Particularly, in Figure 2 the histograms for the exponential type errors and for the various choice of weights and two cross position of m_c and m are demonstrated and we can observe that the difference among them is practically negligible, only the hight of peak is a bit less evident for the weights $w_n(t) = 1$. If we focus on

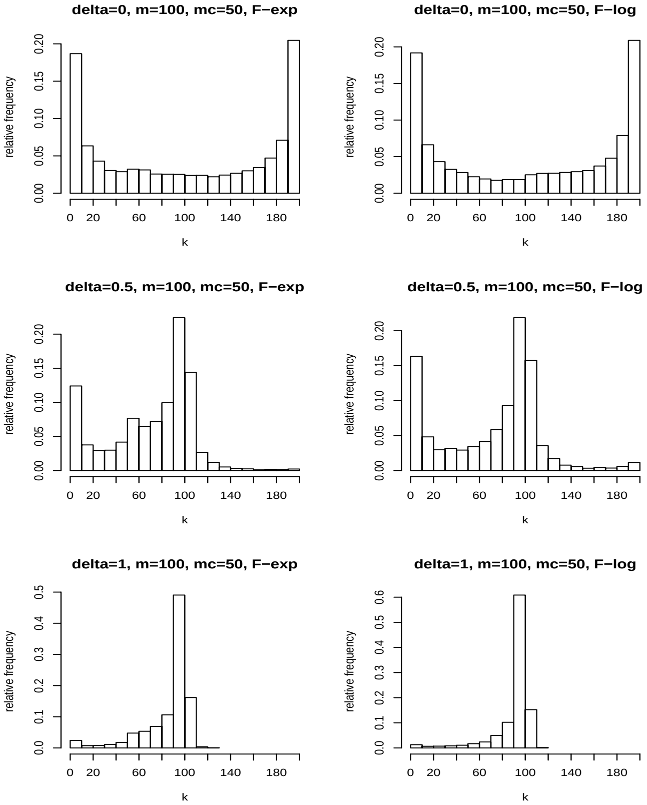


Figure 3: The histogram of $\hat{m}(\tau_0)$ for the parameters $\gamma = 0.5$ and $\eta = 0.25$ and the size of the change $\delta_n = 0; 0.5; 1$ and $\delta_{C,n} = 2$ and type errors $F = G \sim$ Exp(1) or LN (0,1) and only log-rank type weights (LR).

the cross locations of m_c and m, we ensure that the hight of the peak is a bit more evident for the situation $m_c < m$ against the situation $m < m_c$ because of lower proportion of censoring after the m_c, so in the neighborhood of m in the first case is our data more often completely observable. Comparing the frequency histograms of $\hat{m}(\tau_0)$ in Figure 3 we notice, first, that the peak is more evident for larger value of the change amount δ_n in the survival times. Second, for $\delta_n = 0$ we can see that the histogram has two main peaks which corresponds to the assertion in the second part of Theorem 2.1 and finally, we observe that the relative frequency of $\hat{m}(\tau_0)$ is not almost influenced by underlying distribution F of error types. The asymmetry with negative skewness of the simulated histogram is caused by the definition of the estimator $\hat{m}(\tau_0)$ in (3). If we replace in this definition minimum by maximum the curves will be positively skewed in the neighborhood of m.

References

[1] Antoch, J., Hušková, M., Veraverbeke, N. (1995). *Change-Point Problem and Bootstrap.* Journal of Nonparametrics Statistics **5**, 123 – 144.

[2] Csörgő, M., Horváth, L. (1997). *Limit Theorems in Change-Point Analysis.* J. Wiley, New York.

[3] Gombay, E., Liu, S. (2000). *A Nonparametric Test for Change in Randomly Censored Data.* The Canadian Journal of Statistics **21**, 113 – 121.

[4] Hušková, M., Neuhaus, G. (2004). *Change Point Analysis for Censored Data.* Journal of Statistical Planning and Inference, accepted for publication.

[5] Kalbfleisch, J. D., Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* J. Wiley, New York.

[6] Komárková, L. (2004). *Change Point Problem for Censored Data.* Doctoral Thesis, Charles University in Prague, Department of Statistics, in preparation.

[7] Neuhaus, G. (1993). *Conditional Rank Tests for the Two-Sample Problem Under Random Censorship.* The Annals of Statistics **21**, 1760 – 1779.

Acknowledgement: The work was supported by grants GAČR 201/03/0945 and MSM 113200008.

Address: L. Komárková, Charles University, Department of Mathematical Statistics, Sokolovská 83, Praha, Czech Republic; University of Economics, Department of Information Management, Jarošovská 1117/II, Jindřichův Hradec, Czech Republic

E-mail: `koblizle@fm.vse.cz`

COMPSTAT'2004 Symposium

CRITICAL VALUES FOR CHANGES IN SEQUENTIAL REGRESSION MODELS

Alena Koubková

Key words: Statistical computing, numerical algorithms, monitoring changes, regression model.

COMPSTAT 2004 section: Statistical software, Sequential analysis.

Abstract: We assume sequentially coming data following some linear regression model, which can change over time. A training data from the past, where no change occurs, are available to check whether the model changes. Horváth et al. [4] developed a CUSUM type procedure based on L_2-residuals. They also studied the delay between the real change and its detection. Here we confirm their estimation by some more simulations and additional to that we propose the procedure based on L_1-residuals.

1 Description of the problem

Structural stability is an important problem in many different fields such as economy, physics, medicine etc. We assume that training data from the past without any structural change are available to estimate an initial model. Typically, the observations arrive sequentially and after each new observation we decide whether the data obtained so far indicate a change in the model or not. In the positive case we stop observations. As in [4], we focus here on a particular situation, where the data follow the linear regression model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta}_i + \epsilon_i, \quad 1 \le i < \infty, \tag{1}$$

where $\boldsymbol{\beta}_i$, $1 \le i < \infty$ are regression parameters such that

$$\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_m,$$

i.e., the first m observations represent the training (historical) data, when the model does not change. The sequence $\mathbf{X}_i$, $1 \le i < \infty$ consists of independent observable vectors and the sequence ϵ_i, $1 \le i < \infty$ represents the random errors. Only changes in regression parameter $\boldsymbol{\beta}$ are considered.

Our problem can be formulated as hypothesis testing problem, where the null hypothesis corresponds to the model without any change

$$H_0 : \boldsymbol{\beta}_i = \boldsymbol{\beta}_0, \quad 1 \le i < \infty$$

and the alternative hypothesis describes that some change occurs

H_A : there exists $k^* \ge 1$ such that

$$\boldsymbol{\beta}_i = \boldsymbol{\beta}_0,\ 1 \le i < m + k^*, \quad \boldsymbol{\beta}_i = \boldsymbol{\beta}_*,\ m + k^* \le i < \infty, \quad \boldsymbol{\beta}_0 \ne \boldsymbol{\beta}_*.$$

Both, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_*$ are unknown.

The considered test procedure can be described in terms of stopping times (rejection rules) $\tau(m)$ defined as

$$\tau(m) = \begin{cases} \inf\{k \geq 1 : Q(m,k) \geq c_m(\alpha) g(m,k,\gamma)\} \\ \infty \text{ if } Q(m,k) < c_m(\alpha) g(m,k,\gamma) \text{ for all } k = 1, 2, \dots \end{cases}$$

where $Q(m,k)$ is a test statistic based on $Y_1, \dots, Y_{m+k}$, $g(m,k,\gamma)$ is a stopping boundary function and $c_m(\alpha)$ is a constant chosen such that the asymptotic level of the test is α.

We consider the following class of stopping boundary functions

$$g(m,k,\gamma) = \sqrt{m} \left(\frac{m+k}{m}\right) \left(\frac{k}{m+k}\right)^{\gamma},$$

where γ is a tuning constant from the interval $\gamma \in [0, \min\{1 - \nu_2^{-1}, 1/2\})$. The parameter ν_2 is specified in assumption 3. in next section. The function $g(m,k,\gamma)$ is increasing and concave in k and its exact shape changes depending on γ. The larger γ is, the faster the stopping boundary function increases in neighborhood of zero and the slower it increases latter. Notice, when multiplying by appropriate constants d_1 and d_2, it can happen that the functions $d_1 g(m,k,\gamma_1)$ and $d_2 g(m,k,\gamma_2)$ cross each other.

Two following conditions need to be satisfied.

$$\lim_{m\to\infty} P_{H_A}[\tau(m) < \infty] = 1, \tag{2}$$

$$\lim_{m\to\infty} P_{H_0}[\tau(m) < \infty] = \alpha. \tag{3}$$

The former one means that we discover a change if it occurs with probability tending to 1. The later one means that the false alarm has probability α, which determines the constants $c_m(\alpha)$.

The similar problem is dealt in [2].

2 CUSUM statistic based on L_2-residuals

In this section we assume that the random sequences ϵ_i, $1 \leq i < \infty$ and $\mathbf{X}_i^T$, $1 \leq i < \infty$ satisfy the following conditions.

I. ϵ_i, $1 \leq i < \infty$ are independent identically distributed random variables with $\mathsf{E}\,\epsilon_1 = 0$, $0 < \mathsf{Var}\,\epsilon_1 = \sigma^2 < \infty$ and $\mathsf{E}\,|\epsilon_1|^{\nu_1} < \infty$ for some $\nu_1 > 2$,

II. $\mathbf{X}_i^T$, $1 \leq i < \infty$ are independent p-dimensional random vectors of the form $\mathbf{X}_i^T = (1, X_{2i}, \dots, X_{pi})$, and the sets $\{\epsilon_i, 1 \leq i < \infty\}$ and $\{\mathbf{X}_i, 1 \leq i < \infty\}$ are independent,

III. there exists $\nu_2 > 1$ such that

$$\mathsf{E}\,|X_{ij}|^{\nu_2} < \infty, \quad j = 1, \dots, p, \quad 1 \leq i < \infty.$$

For the considered problem (H_0, H_A), Horváth et al. [4] proposed standardized CUSUM (cumulative sums) test procedure based on L_2-residuals

$$\widehat{\epsilon}_i = Y_i - \mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_m, \quad i = m+1, m+2, \dots,$$

where $\widehat{\boldsymbol{\beta}}_m$ is the least square estimator of $\boldsymbol{\beta}$ based on the training data of size m, i.e.

$$\widehat{\beta}_m = \left(\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T \right)^{-1} \sum_{j=1}^{n} \mathbf{X}_j Y_j.$$

The related standardized CUSUM test statistic is defined as

$$\widehat{Q}(m,k) = \frac{1}{\hat{\sigma}_m} \sum_{i=m+1}^{m+k} \widehat{\epsilon}_i, \quad k = 1, 2, \dots.$$

The estimate of the variance $\hat{\sigma}_m^2$ is calculated from the historical period as

$$\hat{\sigma}_m^2 = \frac{1}{m-p} \sum_{i=1}^{m} (Y_i - \mathbf{X}_i^T \widehat{\boldsymbol{\beta}}_m)^2.$$

Horváth et al. [4] showed that under the considered assumptions and under the null hypothesis

$$\lim_{m\to\infty} P\left[\sup_{1\le k<\infty} \frac{\left|\sum_{i=m+1}^{m+k} \widehat{\epsilon}_i\right|}{\hat{\sigma}_m \sqrt{m} \left(\frac{m+k}{m}\right) \left(\frac{k}{m+k}\right)^{\gamma}} \le c \right] = P\left[\sup_{0\le t\le 1} \frac{|W(t)|}{t^{\gamma}} \le c \right] \quad (4)$$

holds for all $c > 0$, where $\{W(t), 0 \le t < \infty\}$ denotes a Wiener process. Therefore, there exist $c(\alpha)$ such that the relation (3) holds and they can be determined from

$$P\left[\sup_{0\le t\le 1} \frac{|W(t)|}{t^{\gamma}} \le c(\alpha) \right] = \alpha,$$

as the authors did. Clearly $c_m(\alpha) \to c(\alpha)$.

Notice that the explicit form of the limit probability is known only for $\gamma = 0$, see e.g. Billingsley [1] and it is equal to

$$P[\sup_{0\le t\le 1} |W(t)| \le c] = 1 - \frac{4}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^k}{2k+1} \exp\left\{ -\frac{\pi^2(2k+1)}{8c^2} \right\}.$$

For $\gamma \in (0, \min\{\nu, 1/2\})$ the explicit form of the limit distribution is not known and the simulations are needed. By Csörgő and Horváth [3] we only have

$$\limsup_{t\to 0} \frac{|W(t)|}{t^{\gamma}} < \infty, \quad \text{a.s.}, \quad \gamma \in (0, \min\{1 - \nu_2^{-1}, 1/2\}).$$

When the alternative is true, Horváth et al. [4] proved that, as $m \to \infty$,

$$\sup_{1 \le k < \infty} \frac{\left| \sum_{i=m+1}^{m+k} \widehat{\epsilon}_i \right|}{\hat{\sigma}_m \sqrt{m} \left(\frac{m+k}{m}\right) \left(\frac{k}{m+k}\right)^{\gamma}} \xrightarrow{P} \infty, \tag{5}$$

which implies that the relation (2) holds.

3 CUSUM statistic based on L_1-residuals

In this section we work with different assumptions on the model (1). Particulary, instead of condition I from section 2, we assume

I'. ϵ_i, $1 \le i < \infty$ are independent, identically distributed random variables with distribution function F symmetric around zero, such that its second derivative exists and $F'(0) > 0$.

The corresponding CUSUM statistic is then based on L_1-residuals

$$\widetilde{\epsilon}_i = \operatorname{sign}(Y_i - \mathbf{X}_i^T \widetilde{\boldsymbol{\beta}}_m), \quad i = m+1, m+2, \ldots,$$

where $\widetilde{\boldsymbol{\beta}}_m$ is an L_1-estimator of the regression parameter $\boldsymbol{\beta}$ based on the training data of size m, i.e. a solution of the optimization problem

$$\widetilde{\boldsymbol{\beta}}_m = \arg\min_{\mathbf{b}} \sum_{i=1}^{m} |Y_i - \mathbf{X}_i^T \mathbf{b}|.$$

Since the estimate of the variance of $\widetilde{\epsilon}_i$, $1 \le i < \infty$ is $\hat{\sigma} = 1$, the standardized CUSUM statistic is defined as

$$\widetilde{Q}(m,k) = \sum_{i=m+1}^{m+k} \widetilde{\epsilon}_i.$$

Using the stopping boundary function $g(m,k,\gamma)$, we get the same limit distribution of $\sup_{1 \le k < \infty} \widetilde{Q}(m,k)/g(m,k,\gamma)$ under H_0 as in (4). The relation (5), representing the behavior under the alternative also holds. So all the conclusions made in section 2. remain true also here. The proofs will be published separately.

4 Simulations

A simulation study with two goals was performed. At first, the critical values $c_m(\alpha)$ were simulated for a location model under the null hypothesis. Secondly, the distribution of delay between the real change-point and its detection under various alternatives were simulated. All the simulations were done for the standardized CUSUM statistics based on L_2-residuals as well as for the statistics based on L_1-residuals.

We consider the location model

$$Y_i = \mu_i + \epsilon_i, \quad 1 \leq i < \infty \tag{6}$$

with the null and alternative hypotheses

$$\begin{aligned} &H_0 : \mu_i = \mu_0, \quad 1 \leq i < \infty, \\ &H_A : \text{there exists } k^* \geq 1 \text{ such that} \\ &\qquad \mu_i = \mu_0,\ 1 \leq i < m + k^*, \quad \mu_i = \mu_0 + \delta,\ m + k^* \leq i < \infty, \quad \delta \neq 0. \end{aligned}$$

Five values of γ uniformly covering the interval $[0, 1/2)$ were chosen: 0.05, 0.15, 0.25, 0.35, and 0.45. Also several values for the size of the historical data, where no change occurs, were used. These are $m = 10, 100, 500$ and 1000. The selected value of the expectation μ_0 was 5. Three distributions of the error terms ϵ_i were considered: $N(0, 1)$, Laplace and t_4. The approximations were conducted for the most common values of α: 0.1, 0.05, 0.025, and 0.01.

Several different alternative hypothesis were considered. They consist of two values for δ: 1 and 2, and of four values for the change-point: $k^* = 10, 100, 500$ and 1000. The series under the alternative hypotheses were generated only for three values of γ: 0.05, 0.25, and 0.45 and only for the case, where the error terms are from $N(0, 1)$ distribution.

Series of length $n = 100m$, but not less than 10 000, were generated 10 000 times under the null hypothesis and 2 500 times under the alternative.

In the simulations under the null hypothesis we proceed as follow.

1. $Y_1, \dots, Y_n$ were generated to follow the model (6), with $\mu_i = \mu_0$ $1 \leq i \leq n$,
2. $\widehat{Q}(m, k)$ and $\widetilde{Q}(m, k)$, for $k = 1, \dots, n - m$ were calculated,
3. $g(m, k, \gamma)$ for $k = 1, \dots, n - m$ were calculated,
4. the maxima of $\widehat{Q}(m, k)/g(m, k, \gamma)$ and $\widetilde{Q}(m, k)/g(m, k, \gamma)$ were determined and stored,
5. the steps (1) to (4) were repeated 10 000 times and the quantiles of such series were used as the approximations of $c_m(\alpha)$.

The procedure for simulation of the alternative hypothesis was.

1. $Y_1, \dots, Y_n$ were generated to follow the model (6), with $\mu_i = \mu_0$ $1 \leq i < m + k^*$ and $\mu_i = \mu_0 + \delta$, $m + k^* \leq i \leq n$,
2. $\widehat{Q}(m, k)/g(m, k, \gamma)$, and $\widetilde{Q}(m, k)/g(m, k, \gamma)$ were calculated,
3. the first time points, where the ratios exceed the critical value $c_m(\alpha)$ were stored,
4. the steps (1) to (3) were repeated 2 500 times and the summary and histograms of such series were used for estimation of the distribution of the delay.

The results of the simulations are presented in the following tables and figures. In Tables 1 and 2, there are the simulated critical values based on L_2- and L_1-residuals, respectively.

ϵ_i		$N(0,1)$			Laplace			t_4		
m	$\gamma \setminus \alpha$	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
10	0.05	2.237	2.659	3.611	2.333	2.754	3.692	2.302	2.795	3.978
	0.15	2.313	2.748	3.817	2.429	2.853	3.868	2.446	2.912	4.103
	0.25	2.402	2.789	3.801	2.555	3.006	4.138	2.579	3.093	4.280
	0.35	2.476	2.883	3.838	2.682	3.215	4.416	2.739	3.274	4.643
	0.45	2.677	3.137	4.199	2.942	3.532	5.000	2.909	3.496	5.226
100	0.05	1.980	2.241	2.830	1.988	2.270	2.835	2.007	2.262	2.823
	0.15	2.032	2.308	2.886	2.054	2.349	2.886	2.087	2.375	2.953
	0.25	2.122	2.421	3.025	2.159	2.449	3.002	2.197	2.509	3.149
	0.35	2.237	2.506	3.046	2.251	2.555	3.151	2.309	2.618	3.337
	0.45	2.478	2.724	3.305	2.549	2.881	3.575	2.608	2.951	3.708
500	0.05	1.992	2.272	2.782	1.966	2.255	2.784	1.968	2.251	2.888
	0.15	2.017	2.286	2.876	2.029	2.299	2.822	2.029	2.314	2.855
	0.25	2.112	2.383	2.935	2.097	2.370	2.919	2.151	2.458	2.952
	0.35	2.221	2.474	3.036	2.229	2.48	3.012	2.270	2.549	3.145
	0.45	2.503	2.759	3.252	2.565	2.841	3.340	2.588	2.899	3.716
1000	0.05	1.972	2.250	2.865	1.973	2.254	2.771	2.002	2.246	2.718
	0.15	2.024	2.298	2.842	2.022	2.297	2.856	2.035	2.307	2.827
	0.25	2.107	2.386	2.985	2.080	2.370	2.964	2.130	2.428	2.959
	0.35	2.233	2.486	3.055	2.247	2.485	3.055	2.251	2.523	3.087
	0.45	2.510	2.759	3.232	2.559	2.825	3.373	2.581	2.863	3.513

Table 1: Approximated critical values based on L_2-residuals.

ϵ_i		$N(0,1)$			Laplace			t_4		
m	$\gamma \setminus \alpha$	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
10	0.05	1.711	1.917	2.252	1.735	1.930	2.267	1.733	1.920	2.228
	0.15	1.763	1.958	2.287	1.774	1.963	2.277	1.758	1.929	2.271
	0.25	1.807	2.000	2.306	1.818	2.007	2.300	1.806	1.987	2.300
	0.35	1.867	2.051	2.343	1.866	2.032	2.352	1.873	2.051	2.365
	0.45	1.941	2.105	2.388	1.941	2.101	2.394	1.941	2.100	2.412
100	0.05	1.921	2.181	2.680	1.922	2.178	2.692	1.940	2.218	2.705
	0.15	2.018	2.291	2.831	1.992	2.226	2.738	1.976	2.240	2.767
	0.25	2.059	2.310	2.816	2.068	2.320	2.818	2.063	2.316	2.836
	0.35	2.171	2.415	2.940	2.157	2.402	2.885	2.181	2.416	2.936
	0.45	2.345	2.580	3.063	2.381	2.614	3.153	2.364	2.591	3.060
500	0.05	1.951	2.224	2.798	1.971	2.247	2.744	1.983	2.254	2.827
	0.15	2.017	2.298	2.819	2.018	2.279	2.823	2.021	2.301	2.850
	0.25	2.077	2.345	2.854	2.091	2.362	2.921	2.069	2.344	2.898
	0.35	2.226	2.485	3.008	2.184	2.439	2.936	2.210	2.475	2.980
	0.45	2.458	2.699	3.208	2.449	2.712	3.205	2.474	2.724	3.247
1000	0.05	1.968	2.228	2.765	1.991	2.269	2.795	1.986	2.265	2.817
	0.15	2.033	2.290	2.829	2.032	2.302	2.896	2.012	2.284	2.781
	0.25	2.116	2.377	2.901	2.089	2.365	2.863	2.088	2.353	2.936
	0.35	2.241	2.517	3.054	2.208	2.470	2.987	2.220	2.474	3.003
	0.45	2.459	2.713	3.275	2.470	2.737	3.252	2.471	2.731	3.218

Table 2: Approximated critical values based on L_1-residuals.

Horváth et al. [4] presented the approximated critical values based on the simulations of the limit distribution. We obtained the approximations by simulation from the location model (6), which converge to the limit ones as $m \to \infty$. From $m = 500$ these values become stable. The convergence is faster for the CUSUM's based on L_2-residual than for those based on L_1-residuals. In both cases, the fastest convergence is observed with normal error distribu-

CUSUM based on L_2-residuals										
	$\gamma = 0.05$					$\gamma = 0.45$				
k	min	1^{st} Q	med	3^{th} Q	max	min	1^{st} Q	med	3^{th} Q	max
10	17	32.75	38	46	108	1	20	26	32	87
100	55	138	155	174	300	1	136	154	174	296
500	68	604	662	729	1232	1	630	700	776	1420
1000	43	1187	1299	1421	2019	1	1252	1377	1525	2498
CUSUM based on L_1-residuals										
	$\gamma = 0.05$					$\gamma = 0.45$				
k	min	1^{st} Q	med	3^{th} Q	max	min	1^{st} Q	med	3^{th} Q	max
10	28	47	54	67	167	10	30	35	47	193
100	47	165	188	219	423	10	158	188	223	450
500	49	679	772.5	889	1544	10	712	820	944	1758
1000	47	1311	1492	1700	2751	10	1415	1626	1879	3416

Table 3: Summary of simulated delay distribution.

tion. Slowest it is with t_4 error distribution when using L_2-residuals and with Laplace error distribution when using L_1-residuals. For the CUSUM's based on L_2-residuals, the approximations $c(\alpha)$ presented in Horváth et al. [4] are a little bit underestimated, whereas for CUSUM's based on L_1-residuals they are overestimated.

The values $c_m(\alpha)$ were checked in the simulations under different alternatives and the distribution of the delay, between the real change-point and the time of its detection, was estimated. Since the results from all the tried alternatives show similar trend, only selected outputs are presented here. Table 3 contains summary of the estimated change-points in the 2 500 simulations with the alternative $\delta = 1$. They were obtained for $m = 100$, $\alpha = 0.05$, $\gamma = 0.05$ (the smallest one), $\gamma = 0.45$ (the largest one) and different change-points.

For smaller m the estimated change-points will be farther from the real ones, whereas for the larger m they will be closer. There is also a clear pattern when changing the tuning parameter γ. If the change-point occurs soon after the monitoring started ($k = 10$), the delay is shortest for the largest values of γ. On the other hand, when the change occurs late ($k \geq 500$), the shortest delay was obtained for γ as small as possible. This could be expected from the shape of the stopping boundary function $g(m, k, \gamma)$. In the alternative with $\delta = 2$, the change is detected about two times faster.

For the normal data, the CUSUM's based on L_2-residuals detect the change-point earlier than the CUSUM's based on L_1-residuals.

The graphical representation of these results is given in Figure 1. The first three plots correspond to $\gamma = 0.05$ (in figures denoted as g), the second three to $\gamma = 0.45$ and all are obtained based on the L_1-residuals. For both γ's, there are two histograms showing the estimated distribution of the change-point detection, when the change occurs after 10 and 100 observations, respectively. The last plot shows the changes of $\widetilde{Q}(m, k)/g(m, k, \gamma)$ when the change occurs after 100 observations (depicted by a dot).

The statistical software R(1.5.1) was used for the calculations.

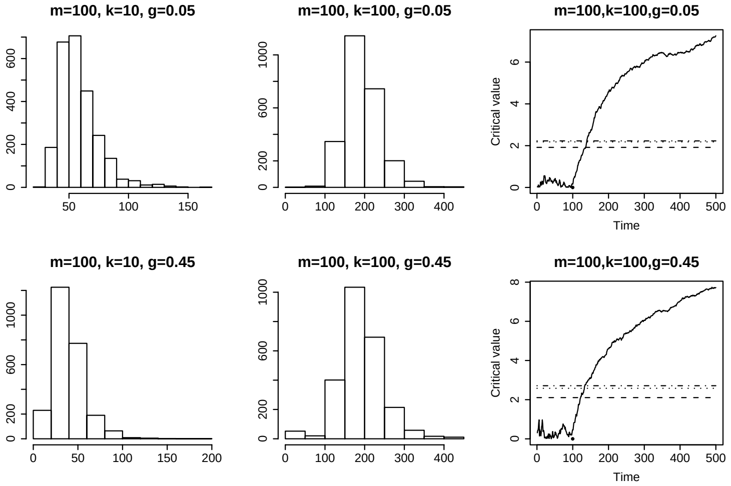


Figure 1: Delay of change-point detection.

5 Conclusion

Horváth et al. [4] proposed a procedure for detection of a change in a regression model, when the training data without any change are available. They propose to use L_2-residuals and their limit distribution, as the size of the training data tends to infinity. In this paper we studied how fast is the convergence to the limit distribution. Moreover we propose a procedure based on L_1-residuals.

References

[1] Billingsley P. (1977). *Convergence in Probability Measures.* John Willey & Sons, New York

[2] Chu C.-S. J., Stinchcombe M. and White H. (1996). *Monitoring Structural Change.* Econometrica **64**, 1045-1065

[3] Csörgő M. and Horváth L. (1993). *Weighted Approximations in Probability and Statistics.* John Willey & Sons, Chichester.

[4] Horváth L., Hušková M., Kokoszka P. and Steinebach J. (2004). *Monitoring Changes in Linear Models.* Accepted for publication in Journal of Statistical Planning and Inference

Acknowledgement: I want to thank to prof. Marie Hušková for her ideas and advices to this work. The work was supported by grant GAČR 201/03/0945 and MSM 1132 00008.

Address: A. Koubková, Department of Statistics, Charles University, Sokolovská 83, CZ – 186 00, Czech Republic

E-mail: `aja@matfyz.cz`

COMPSTAT'2004 Symposium

MULTIPLE TEST PROCEDURES WITH MULTIPLE WEIGHTS

Siegfried Kropf and L.A. Hothorn

Key words: Multiple test procedure, multipe weights, highdimensional data, familywise error in the strong sense.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: Recently, Westfall, Kropf and Finos [10] proposed a multiple test procedure for the parallel consideration of the variables in multivariate data which is based on weights and can be tuned by a free parameter η. Unfortunately, η has to be chosen in advance. To avoid a heavy loss in power due to an inappropriate choice of this parameter, two procedures for multiple weights are considered here.

1 Multiple test problem and related standard procedures

In biological and medical research, one often has experiments where many variables are measured on the individuals and the questions of interest are not primarily of a multivariate nature. Instead, tests are directed to the single variables and the problem of multiple testing occurs to avoid a high rate of false positive findings. A typical example is the analysis of gene expression data. With the so-called microarrays, the activity of several thousands of genes can be measured simultaneously. One wants to find out which genes change their expression values under specified conditions. For example, Eszlinger and coauthors [3] studied the differences in gene expression in autonomously functioning thyroid nodules (AFTNs) without known mutations with respect to the surrounding tissue. They used the U95Av2 Affymetrix GeneChip with 12,625 genes. The formal data (sample size, $n = 15$, and number of variables, $p = 12,625$) of this study have been used as a template for the simulation experiments in this paper.

The multiple test problem can be characterized in the following way. Let

$$\mathbf{X} = (x_{ji}) = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

be the data matrix consisting of n independent multivariate normal sample vectors $\mathbf{x}_j \sim N_p(\mu, \mathbf{\Sigma})$ with expectation vector $\mu = (\mu_1, \ldots, \mu_p)'$ and arbitrary covariance matrix $\mathbf{\Sigma}$ $(j = 1, \ldots, n)$. We want to test if the expectation is zero. However, we are not primarily interested in the test of the global null hypothesis $H: \ \mu = \mathbf{0}$ but in the local hypotheses $H_i: \ \mu_i = 0$ $(i = 1, \ldots, p)$.

In the example, the sample vectors are the differences of the log-transformed expression vectors from the nodule and from the surrounding tissue for each patient. The local null hypothesis H_i means that the expression of the ith gene is not changed in the nodule.

The log-transformation of the expression values has actually two effects. It produces approximately normal distributions and it brings the variances of the (transformed) expression values of the p genes to a similar size. That will be utilized later on in the procedures.

We want to guarantee the familywise type I error for the p tests of the local hypotheses H_i $(i = 1, \dots, p)$ in the strong sense. That means that in all p tests any type I error occurs with a prespecified probability α at most.

Standard procedures for multiple testing have problems with such extremely high dimensions p as in the gene expression analyses, particularly when the sample sizes are small as is usual here. In the Bonferroni-Holm procedure [4], we have very small critical levels for the first steps. The same is true for the usual modifications of this procedure. Tests with a priori ordered hypotheses [1] are not applicable because our biological knowledge is usually insufficient. The permutation procedure of Westfall and Young [11] is recommended , e.g., in [2], for the application in microarray analyses. With small samples, however, the permutation procedure is lacking power, too.

2 Procedures utilizing the similarity of the variances

In order to include an additional source of information, one can use a very simple procedure [5], called **Procedure I** here:

- Sort the p variables for decreasing values of the corresponding sums of squares $w_i = \sum_{j=1}^{n} x_{ji}^2$.
- In the obtained order, carry out usual one-sample t tests, each at the unadjusted level α.
- Reject the corresponding local hypotheses as long as all tests yield significance. Stop when the first nonsignificant result occurs and accept the remaining local hypotheses.

The proof that the procedure holds the familywise type I error in the strong sense (given in [5]) utilizes multivariate theorems developed by Läuter et al. [7] and is valid for an arbitrary covariance structure. The power, however, depends on the real degree of heteroscedasticity of the p variables. This can be seen from the decomposition $\sum_{j=1}^{n} x_{ji}^2 = \sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2 + n\bar{x}_i^2$, where $\bar{x}_i$ denotes the variablewise means. When the variances of the variables are approximately equal, then the first term on the right hand side of the equation should be similar for all variables and the order of the variables determined from the w_i is essentially determined by the absloute means of the variables. Hence, variables with large deviations from the null hypothesis will be in the front positions and have better chances to be detected in the second step. In case of very heterogeneous variables, however, those variables with the largest variances will be in early positions even if they have no mean effect. Then the procedure would stop early.

The additional information of the sums of squares is used here in a very strict manner which makes the procedure sensible to disturbances, just like testing with a priori ordered hypotheses. Therefore, Westfall et al. [10] proposed another procedure which uses the w_i for weighting the variables (called **Procedure II** here):

- Determine the unadjusted p-values p_i from the one-sample t tests and the sums of squares w_i for all p variables.
- With a fixed $\eta \geq 0$, calculate weights $g_i = w_i^\eta$ and weighted p-values $q_i = p_i/g_i$ $(i = 1, \ldots, p)$.
- Sort the variables for increasing weighted p-values $q_{(1)} \leq \cdots \leq q_{(p)}$ and denote the corresponding weights by $g_{(1)}, g_{(2)}, \ldots, g_{(p)}$.
- In this order, reject $H_{(j)}$ as long as

  $$q_{(j)} \leq \frac{\alpha}{\sum_{i=j}^{p} g_{(i)}} \, .$$

  Stop at the first non-significant result and accept the remaining hypotheses.

As shown in [10], Procedure II holds the familywise type I error in the strong sense, too. A very interesting property of this procedure is that the two limiting values for η correspond to known procedures: $\eta = 0$ yields the usual Bonferroni-Holm procedure, and the procedure converges to Procedure I for $\eta \to \infty$. Thus, it is possible to tune the procedure in dependence on the expected gain in power by utilizing the sums of squares w_i.

Test versions for the two-sample case and corresponding nonparametric procedures are treated in [10] and [6], repectively. The considerations of the subsequent sections can be transferred to these versions in a straightforward manner.

3 Choice of the free parameter η

As already mentioned, Procedure II holds the familywise type I error in the strong sense for each value of η. The power, however, strongly depends on η, which just gives the possibility for a large gain in power compared to other procedures. Simulation experiments in [10] and [6] have shown that the optimal choice of η essentially depends on the sample size. The larger the sample size, the smaller η should be. It seems that the advantage of Procedure II is lost for sample sizes above 50. Simulation experiments with the same parameters n and p as in the planned real data may help to find an appropriate value for η.

There are also other potentially influential parameters, which are not known in advance. Figures 1 and 2 show the results of some simulation

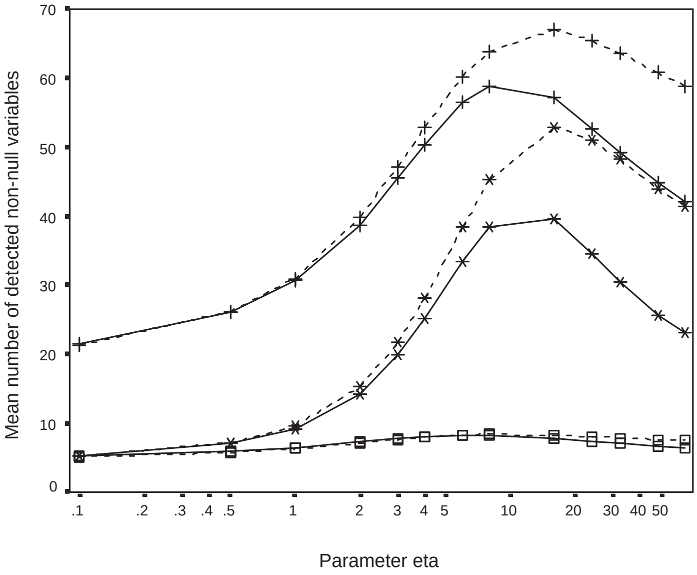


Figure 1: Results of simulation experiments for Procedure II (cf. text) with normally distributed samples of size 15 and 1,000 replications per parameter set: solid lines: uncorrelated data, dashed lines: pairwise correlation coefficients all 0.5; star symbols: 100 non-nullvariables from a total of 12,625, plus symbols: 100 out of 1,000, squares: 10 out of 100.

experiments. The sample size is 15 in all series, corresponding to the real data in [3]. The prespecified multiple error level is $\alpha = 0.05$ in all tests. The parameter η varies in these experiments in discrete steps (powers of 2, supplemented by 0 and a few intermediate values) over a wide range.

In Figure 1, we start with 12,625 independent normal variables (solid line, stars). From these, 100 have expectation 1.155, the others have expectation 0, all have the common variance 1. The figure shows the mean number of detected non-null variables (from the 100) in dependence from η. With the considered grid for η, the maximum of 39.6 is attained for $\eta = 16$. In a second series (dashed line, stars), the data have been generated with a pairwise correlation coefficient of 0.5 between all variables; the other parameters are the same. The maximum power is also found for $\eta = 16$, but now 52.9 of the 100 variables are detected on average. The two lines with the plus symbol result from simulation experiments with 1000 variables of which 100 have the expectation 1.155 as before and the others zero. As the 100 non-null variables are hidden here in a smaller total number of variables, the two power curves for uncorrelated and correlated data are above the previous ones with maximal values of 58.8 (uncorrelated) and 67.0 (correlated). With

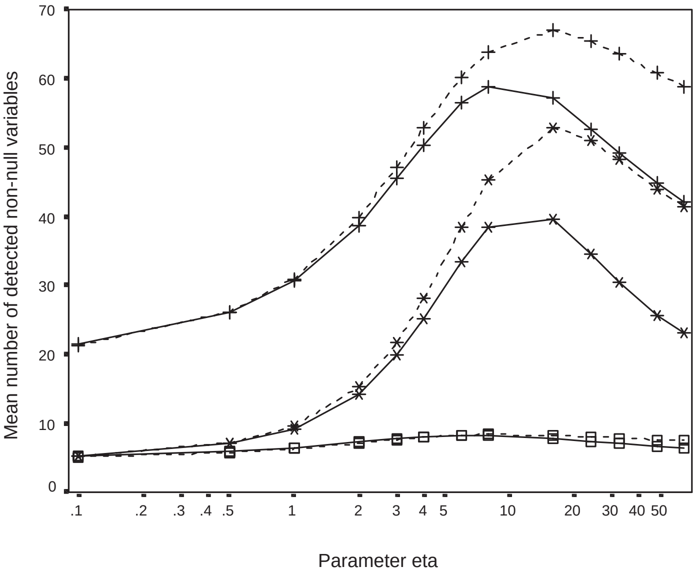


Figure 2: Results of simulation experiments for Procedure II (cf. text) with normally distributed samples of size 15, 100 non-null variables out of a total of 12,625 variables and 1,000 replications per parameter set: solid lines: uncorrelated data, dashed lines: pairwise correlation coefficients all 0.5; star symbols: all standard deviations equal to 1, plus symbols: standard deviations vary between 0.61 to 1.65, squares: standard deviations vary between 0.37 and 2.72.

the given η-grid, these maxima are again attained for $\eta = 16$ in the one case and for the next smallest value $\eta = 8$ in the other case. Finally, the lines with the square symbols give the results for a total of 100 variables of which 10 have the expectation 1.155 (the others zero). Here, the maximum number of detected non-null variables is found near $\eta = 8$. But in this case, the differences in the mean detected numbers for varying η are much smaller. Thus, summarizing the results of Figure 1, one can state that the correlation between the variables and also the total number of variables as well as the number of non-null variables has an obvious influence on the power of the procedures but not so much on the optimal η-value.

In Figure 2, the lines with the star symbol represent the same simulation experiments as in Figure 1. The lines with the plus symbol and with the squares now display the results for corresponding analyses, where the variances of the 12,625 variables are no longer identical. In the lines with the plus symbol, the standard deviations vary in a range from 0.61 to 1.65, in those with the squares from 0.37 to 2.72. Again, the curves for the uncorrelated and the correlated data do not differ too much. However, the influence of heteroscedasticity is dramatic. As discussed before in the introduction of Procedure II, now also the choice of η is strongly affected. A value of 16 that was quite well for homogeneous variances, results in a very low power for the configurations with heterogeneous variances.

Thus, in case of unknown degree of heterogeneity of the variances of the variables, it may be difficult to choose a suitable η. Choosing it in dependence from the data may violate the type I error (according to the present state of the theory).

4 Consideration of multiple weights

One obvious question that arises is whether one could consider multiple weights based on different values of η. Of course, the familywise type I error in the strong sense has to be kept. We will test two simple strategies here. The first one uses the Bonferroni principle, the second one the Simes test [9]. In both cases, we first have to select a set of suitable η values covering the suspected region of potentially useful values. Let the selected values be given by $\eta_1, \dots, \eta_m$.

Then, for each η_k $(k = 1, \dots, m)$, the p-values p_{ik} for the variables x_i $(i = 1, \dots, p)$ according to Procedure II are determined. For this, we carry out the first three steps of Procedure II as decribed above. The fourth step is modified into

- In this order, determine the "multiple p-values" by calculating the values $q_{(j)} \cdot \sum_{i=j}^{p} g_{(i)}$ and replacing each of them by the maximum of this value and the preceding one.

In the Bonferroni version, a variable x_i is considered as significant if at least one of the p-values $p_{i1}, \dots, p_{im}$ is less than or equal to α/m. For the Simes version, the m p-values $p_{i1}, \dots, p_{im}$ are sorted in increasing manner, such that $p_{i(1)} \leq \dots \leq p_{i(m)}$, and the variable x_i is considered as significant if $p_{i(k)} \leq \alpha \frac{k}{m}$ for at least one $k = 1, \dots m$. It it obvious that each significant result in the Bonferroni procedure is also significant in the Simes procedure, but not vice versa. In the special case that all m p-values for different η are identical, the Simes procedure would detect a variable as significant if this p-value is less than or equal to α, i.e., we would not have to pay an extra price for the artificial multiplicity. The Bonferroni procedure should be conservative in any case because we spend a risk for committing any type I error of α/m for each η_k cumulating to a total risk of α at most. For the Simes procedure, exact results are given only for independent p-values, but simulation investigations by Simes himselves [9] and by Samuel-Cahn [8] have shown a conservative behaviour also for positively correlated p-values and in two-sided tests also for negatively correlated p-values. In our case, we consider jointly the p-values obtained with different weights for the same variable. Therefore, positive correlations between the m p-values can be expected. With the two procedures (and particularly with Bonferroni-Holm), m should not be chosen too large in order to restrict the influence of the α-adjustment.

In order to check the properties of these two procedures, we apply them to the same simulated data as before for Figure 2. Assume that out prior knowledge on the heterogeneity of the variances was very weak and that all three situations used in Figure 2 might be realistic. Then, the optimal η might be suspected in the range of 1 to 16. Therefore, we choose $m = 3$ and $\eta_1 = 1$, $\eta_2 = 4$ and $\eta_3 = 16$.

The familywise type I error was kept in all simulation series. The proportion of data sets with at least one significant result for a variable with real expectation zero was always distinctly below 0.05 (actually, below 0.04), where – as expected – the Bonferroni version was slightly more conservative.

The mean numbers of detected non-null variables (out of 100) in the 1,000 replications are shown in the table below for both procedures, together with the results of Procedure II for the optimal choice of η within the given grid.

std.dev.	1.00		0.61–1.65		0.37–2.72	
correlation	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.0$	$\rho = 0.5$
Bonferroni	31.7	42.0	12.4	14.0	23.1	22.9
Simes	32.3	42.4	13.3	15.1	23.1	22.9
P. II, best η	39.6	52.9	17.4	18.6	28.7	28.7

Both procedures do not differ very much in their results. Of course, the Simes procedure has slight advantages. Both procedures detect distinctly less non-null variables than the original Procedure II with the optimal choice of η. But – as can also be seen from Figure 2 – the loss in power is much smaller than it would be with an inappropriate choice of η. Therefore, both methods can be recommended in case of uncertain prior information on the data. As expected, a repeated simulation series (not presented here) with the enhanced set of η values over the same range, $m = 5$, $\eta_1 = 1$, $\eta_2 = 2$, $\eta_3 = 4$, $\eta_4 = 8$, and $\eta_5 = 16$, gave distinctly lower power values for the Bonferroni version, but only slightly deteriorated values for the Simes procedure.

Possibly, other methods could better exhaust the prespecified familywise error level and could support larger number m of multiple weights in this twofold multiple test problem (multiple variables and multiple η). The application of permutation techniques may help. However, one has to prevent then, that – in optimizing the treatment of the problem of multiple η – the error control for the multiplicity in the variables is lost.

References

[1] Bauer P., Röhmel J., Maurer W., Hothorn L.A. (1998). *Testing strategies in multiple-dose experiments including active control.* Statistics in Medicine **17**, 2133–2146.

[2] Dudoit S., Shaffer J.P., Boldrick J.C. (2003). *Multiple hypothesis testing in microarray experiments.* Statistical Science **18**, 407 – 418.

[3] Eszlinger M., Krohn K., Frenzel R., Kropf S., Tönjes A., Paschke R. (2004). *Gene expression analysis reveals evidence for interaction of the TGF-β signaling cascada in autonomously functioning thyroid nodules.* Oncogene **23**, 795 – 804.

[4] Holm S. (1979). *A simple sequentially rejective multiple test procedure.* Scandinavian Journal of Statistics **6**, 65 – 70.

[5] Kropf S., Läuter J. (2002). *Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression analysis.* Biometrical Journal **44**, 789 – 800.

[6] Kropf S., Läuter J., Eszlinger M., Krohn K., Paschke R. (2004). *Nonparametric Multiple test procedures with data-driven order of hypotheses and with weighted hypotheses.* Journal of Statistical Planning and Inference, in press.

[7] Läuter J., Glimm E., Kropf S. (1998). *Multivariate tests based on left-spherically distributed linear scores.* Annals of Statistics **26**, 1972 – 1988. Correction: Annals of Statistics **27**, 1441.

[8] Samuel-Cahn E. (1996). *Is the Simes improved Bonferroni procedure conservative?* Biometrika **83**, 928 – 933.

[9] Simes R.J. (1986). *An improved Bonferroni correction for multiple tests of significance.* Biometrika **73**, 751 – 754.

[10] Westfall P.H., Kropf S., Finos L. (2004). *Weighted FWE-controlling methods in high-dimensional situations.* Accepted for: Benjamini Y., Bretz F., Sarkar S.K. (eds.). Recent developments in multiple comparison procedures, IML Lecture Notes and Monograph Series.

[11] Westfall P.H., Young S.S. (1993). *Resampling based multiple testing.* John Wiley & Sons, New York.

Acknowledgement: The authors thank the referee for useful hints.

Address: S. Kropf, Institute of Biometry and Medical Informatics, Otto von Guericke University, Leipziger Str. 44, 39120 Magdeburg, Germany
L.A. Hothorn, Teaching Unit Bioinformatics, University of Hannover, Herrenhäuser Str. 2, 30419 Hannover, Germany

E-mail: `siegfried.kropf@medizin.uni-magdeburg.de`
`hothorn@bioinf.uni-hannover.de`

COMPSTAT'2004 Symposium

CLUSTERING OF TRANSACTION DATA

Lukáš Křečan and Petr Volf

Key words: Model based clustering, mixture distribution, EM algorithm.
COMPSTAT 2004 section: Clustering.

Abstract: The present contribution deals with the problem of model based clustering of data from card payment terminals. The computation is performed via the EM algorithm, BIC is used in order to determine optimal number of clusters. The influence of data components on the final clustering is analyzed additionally with the aid of classification tree.

1 Introduction

The objective of the cluster analysis consists in retrieval the internal structure of observed data and, consequently, in estimation of the structure of data source. That is why the clustering algorithms are often used as basic tools in the data mining problems. In the present paper we deal with the cluster analysis of transaction data. More precisely, the data are the output information from several credit card payment terminals, during a set of days, so that our problem can also be formulated as the problem of determining the profiles of individual terminals and of finding the criteria how to discriminate between them, or how to detect a set of transactions quite different from all profiles. Naturally, such a solution could be a valuable tool also for an online detection of atypical transactions, provided the profiles are well defined and the procedure inspects the transaction processes on the continuous time basis.

Here, however, we shall work with the discretized time case. The data consist of n one day records of payments. Every day record of each terminal is divided into m intervals, for each interval we have the number of payments and also the sum of paid money in that interval. Hence, the data can be written as a matrix with n rows and $2 \times m$ columns, m columns contain the counts (numbers of payments), other m columns contain the sums of payments.

As a rule, we also know from what terminals the data were obtained, hence we know the label (the number of terminal) corresponding to each data row, but, as it is not clear that the profiles found in data will correspond exactly to different terminals, this information will be used just for the comparison and reference.

2 Model based clustering

The first objective of cluster analysis consists in the selection of an appropriate clustering method giving reasonable results. After a set of experiments, we decided to use so called model based clustering, for its theoretical tractability an relatively good practical performance. In the model based clustering scheme, it is assumed that the data are generated by a mixture of underlying distributions. Each component of the mixture represents a different subpopulation or cluster. We assume that the population of observations consists of G different subpopulations and that the density or probability of K-dimensional observation $\bar{x}$ from the $j-$th subpopulation is $f_j(\bar{x};\theta)$ for some unknown vector of parameters θ. Hence, in the beginning, the number G and probabilities $f_j(\bar{x};\theta)$ should be selected. Further, as in real cases there always exist data lying evidently out of the region of attraction of all natural clusters (even here we can call them 'outliers'), it has also sense to consider a group $j = 0$ of them - such a group is easily defined from objects having their probabilities of occurrence in all other clusters less than a properly selected p_0 (see also [1]).

One of the most difficult problems is that of selection of optimal number of clusters. There exists a set of criteria which can be applied to the selection of optimal complexity of model, for instance AIC, criteria developed from Bayes approach (BIC, Bayes factor, also contemporary MCMC methodology is able to deal with number of model units as with one of model parameters), Gap method etc., see [3], [4]. However, each of the methods still has certain 'degrees of freedom' for an analyst's own decision, no consistent theory exists up to now. The model based clustering can adapt these criteria, too, due the fact that, from one point of view, it is assumed that the data are the result of sampling from a mixture distribution of probability (of a random variable $\boldsymbol{X}$), namely

$$f(\bar{x}) = \sum_{j=0}^{G} q_j \cdot f_j(\bar{x};\theta), \tag{1}$$

where $\{q_j, j = 0, 1, ..., G\}$ is a discrete probability distribution on $\{0, 1, ..., G\}$ and $f_0 = p_0$ is the uniform probability of the component 'zero'. Theoretically, it should be selected proportional to $1/A$, A denoting the area of the domain of values of $\boldsymbol{X}$.

On the other hand, the clustering itself is based on the search for an optimal set of identifying labels $\boldsymbol{z} = (z_1, \ldots, z_n)$, where $z_i = j$ if the i-th data vector $\bar{x}_i$ comes from j-th subpopulation (cluster). Namely, given observations $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_n)$, parameters θ and labels $\boldsymbol{z}$ should be chosen as maximizers of the likelihood

$$L(\theta, \boldsymbol{z}) = \prod_{i=1}^{n} f_{z_i}(\bar{x}_i;\theta). \tag{2}$$

2.1 Model

The data of our problem consisted of $n = 140$ rows of $K = 8$ dimensional observations (each day was divided to four time intervals of six hours). The nature of the data was such that the most popular Gaussian form of the model could not be applied, at least not directly. The number of payments is a discrete integer variable describing the number of events, hence the Poisson distribution is the natural choice of its model. Let us denote x_{il} the number of payments in i-th observation day and l-th time interval. Then the corresponding probability component equals the probability of value x_{il} provided the observation i belongs to cluster j, namely

$$P_c^l(z_i = j|\lambda_{jl}) = \frac{\lambda_{jl}^{x_{il}}}{x_{il}!} e^{-\lambda_{jl}}, \tag{3}$$

where λ_{jl} denotes the intensity of corresponding Poisson distribution.

As regards the amounts transacted, it could seem that for them (or, even better, for their averages) the normal distribution would be a convenient choice (also due the central limit theorem). However, as numbers of payments were not too large, it turned out that the normal distribution was applicable rather for logarithms of averaged sums transacted in followed time intervals (it means that we assume log-normal distribution for averaged amounts of collected money). Hence, if we denote y_{il} the logarithm of average amount of money paid in i-th observation and l-th interval, we assume that the corresponding component of probability equals to the normal density of the value y_{il} provided the observation i belongs to the cluster j,

$$P_a^l(z_i = j|\mu_{jl}, \sigma_{jl}) = \frac{1}{\sqrt{2\pi}\sigma_{jl}} \exp\left(-\frac{1}{2\sigma_{jl}^2}(y_{il} - \mu_{jl})^2\right), \tag{4}$$

where μ_{jl} is the mean and σ_{jl} is the standard deviation of distribution.

Under the assumption of independence between number of payments and average amount and also independence between intervals, we obtain the following components of mixture model (1) and likelihood (2).

$$f_j(\bar{x}_i; \theta) = \prod_{l=1}^{4} P_c^l(z_i = j|\lambda_{jl}) P_a^l(z_i = j|\mu_{jl}, \sigma_{jl}), \tag{5}$$

where $\bar{x} = (x, y)$ and $\theta = (\lambda, \mu, \sigma)$.

It is seen that we consider also the estimation of variances of Gaussian components from the data, though in purely Gaussian models such an approach is not recommended. It is well known that the clustering algorithms then tend to overestimate some variances and underestimate others. In our case it seems that the presence of non-Gaussian parts of distribution prevents such a degradation.

2.2 Probability of empty interval

When examining the previous model we encountered one problem. In some instances (so that with positive probability) there is no payment at some terminals through certain time interval. In order to cover such a case, the continuous distribution of logarithms of averaged amounts should be amended by a point probability of such a case (for which the log of averaged payments is not defined). It could be done in the following way: Let us denote $\bar{p}_{jl}$ the probability that in cluster j and interval l there is no payment. We then alter the probability (4) to the form

$$P_a^l(z_i = j|\mu_{jl}, \sigma_{jl}, \bar{p}_{jl}) = \tag{6}$$
$$= \begin{cases} (1-\bar{p}_{jl})\frac{1}{\sqrt{2\pi}\sigma_{jl}} \exp\left(-\frac{1}{2\sigma_{jl}^2}(y_{il}-\mu_{jl})^2\right), \text{ for } x_{il} \neq 0 \\ \bar{p}_{jl}, \text{ when } x_{il} = 0 \end{cases}$$

Consequently, we also have to change the probability (3) to

$$P_c^l(z_i = j|\lambda_{jl}) = \begin{cases} 1, \text{ when } \lambda_{jl} = 0 \text{ and } x_{il} = 0 \\ 0, \text{ when } \lambda_{jl} = 0 \text{ and } x_{il} \neq 0 \\ \frac{\lambda_{jl}^{x_{il}}}{x_{il}!} e^{-\lambda_{jl}} \text{ otherwise} \end{cases} \tag{7}$$

Now, the model is more complete. We also included the cluster of outliers, with its uniform probability p_0 of each result. When checking the number of parameters of our version of the mixture model (1) (i.e. the model explaining the data, describing their 'generator'), for each j=1,2,...,G we have four sets of parameters $\mu, \sigma, \lambda, \bar{p}$, then the weight parameters q_j ($q_0 = 1 - \sum_1^G q_j$). We do not count parameter p_0, because it is selected in advance. Therefore, the model has together $G \times 17$ free parameters. This number will later be used in the BIC criterion of selection of optimal number of clusters.

3 Procedure of computation, (C)EM algorithm

The objective is to find the maximum of the likelihood function (2), when the distributions entering it are given by (5) with (6) and (7). A direct way to solution is complicated, that is why we have utilized, as it is often recommended in literature, iterative EM algorithm. More precisely, the version adapted for the problem of classification, i.e. computing the labels z_i. The weights q_j of mixture distribution (1) are obtained from the final solution and are not used during computations (see [3]). The algorithm consists of two regularly repeated steps:

3.1 E-step

The first one, E-step, updates the classification of observations. Using the current values of parameters we calculate for each $i \in \{1 \dots n\}$ $P_c^l(z_i = j|\lambda_{jl})$ and $P_a^l(z_i = j|\mu_{jl}, \sigma_{jl}, \bar{p}_{jl})$ according to the equations (6) and (7) for all $j \in \{1 \dots G\}$ and $l \in \{1 \dots 4\}$. Then we set

$$z_i = \text{argmax}_{j \in \{1 \dots G\}} \left(\prod_{l=1}^{4} P_a^l(z_i = j|\mu_{jl}, \sigma_{jl}, \bar{p}_{jl}) P_c^l(z_i = j|\lambda_{jl}) \right),$$

$z_i = 0$ when the maximum above is less than p_0.

3.2 M-step

This step updates the values of parameters, given actual classification of observations. First, we count the current cluster sizes

$$c_j = \sum_{i=1}^{n} \boldsymbol{I}_{\{z_i=j\}},$$

then, by means of standard ML estimation, we estimate the probabilities of empty intervals

$$\bar{p}_{jl} = \frac{1}{c_j} \sum_{i=1}^{n} \boldsymbol{I}_{\{x_{il}=0\}} \boldsymbol{I}_{\{z_i=j\}},$$

the means and variances of normal distributions

$$\mu_{jl} = \frac{1}{c_j(1-\bar{p}_{jl})} \sum_{i=1}^{n} y_{il} \boldsymbol{I}_{\{x_{il}\neq 0\}} \boldsymbol{I}_{\{z_i=j\}},$$

$$\sigma_{jl}^2 = \frac{1}{c_j(1-\bar{p}_{jl})} \sum_{i=1}^{n} (y_{il}-\mu_{jl})^2 \boldsymbol{I}_{\{x_{il}\neq 0\}} \boldsymbol{I}_{\{z_i=j\}},$$

as well as the intensities of Poisson distributions

$$\lambda_{jl} = \frac{1}{c_j} \sum_{i=1}^{n} x_{il} \boldsymbol{I}_{\{z_i=j\}}.$$

In the case that $\bar{p}_{jl} = 1$ we do not calculate estimates of μ_{jl} and σ_{jl}. They are not defined and not needed in the following E-step since in such a situation the corresponding distribution (5) is degenerated.

3.3 Initial conditions

The EM algorithm proceeds by repetition of both steps until convergence – in our case until there are no changes of classification. The algorithm may start both with its E-step or M-step. The only difference is in initial conditions. It is important to start from values that are not too far from the optimal solution. In an opposite case there is a danger that the EM algorithm will tend to a local maximum instead to the global one.

When we start with the E-step we need some initial estimation of the distribution parameters $\bar{p}_{jl}$, μ_{jl}, σ_{jl} and λ_{jl}. When we start with the M-step, we have to estimate somehow the initial classification z_i. From this point of view, it seems that the second way is easier. We may find an initial division by means of some other (simpler, ad hoc) clustering method. For example by means of the hierarchical clustering based just on the sums of collected money.

In order to increase the chance that the procedure will really lead to the global maximum, it is recommended to start the algorithm repeatedly from different (even randomly selected) initial values. Such an approach is possible, naturally, if the computation time is not long.

G:	3	4	5	6	7
-loglik.	916.17	823.39	765.45	722.17	687.20
BIC	2084.4	1982.8	1950.9	1948.4	1962.5

Table 1: Resulting log-likelihoods and BIC criteria.

G=4	1	2	3	4	5	6
0	1	2	2	3	4	4
1	20	0	0	1	2	2
2	0	19	32	0	0	6
3	0	9	8	0	0	4
4	0	1	0	17	2	1
G=5	1	2	3	4	5	6
0	1	2	1	3	4	5
1	20	0	0	1	2	2
2	0	19	13	0	0	5
3	0	5	20	0	0	2
4	0	1	0	17	2	1
5	0	4	8	0	0	2
G=6	1	2	3	4	5	6
0	1	4	1	2	4	2
1	20	0	0	1	2	2
2	0	20	9	0	0	5
3	0	4	14	0	0	2
4	0	1	0	17	2	1
5	0	2	8	1	0	5
6	0	0	10	0	0	0

Table 2: Overview of results.

4 Numerical results

The procedure described in the previous section has been used several times, for different (fixed during one computation) numbers of clusters G. At each case the additional group of outliers has been considered, too, the critical probability level was set to $p_0 = 1 \cdot 10^{-5}$. The criterion of successful classification was the value of likelihood (2). Table 1 shows a part of best results. To each result, the value of BIC criterion has been computed, too, as an indication of optimal number of clusters. We utilized the standard formula penalizing minus log-likelihood by the number of model parameters (we have already said that it equaled $d = 17 \times G$), namely BIC$= -2 \cdot loglik + \log n \cdot d$. The values in Table 1 show that, from such a point of view, the optimal G

Cluster	1	2	3	4	5	6
λ_1	0	0.0500	0	0.0476	1.3750	0
μ_1	NaN	3.4657	NaN	6.9518	6.7015	NaN
σ_1	Inf	1.0000	Inf	1.0000	0.4781	Inf
λ_2	3.2000	4.3500	2.5294	0	3.9375	5.1000
μ_2	8.4973	6.6069	6.5022	NaN	6.6481	6.7669
σ_2	0.8310	0.3490	0.2070	Inf	0.2185	0.0837
λ_3	5.0800	5.7500	2.8235	2.0952	3.0625	7.6000
μ_3	8.4594	6.4230	6.4272	7.0348	6.5346	6.6207
σ_3	1.1308	0.1668	0.4307	1.1402	0.1534	0.2147
λ_4	0.0400	2.0000	3.1471	4.8571	2.5625	6.9000
μ_4	8.9872	6.1142	6.6668	7.7714	6.5468	6.4918
σ_4	1.0000	0.5112	0.3748	0.8417	0.3831	0.2096

Table 3: Estimated parameters of distributions in clusters.

with minimal BIC was still $G = 6$, though with practically the same value of BIC as $G = 5$.

Other details of the result are displayed in two following tables. Table 2 offers the comparison of cluster analysis with actual labels of data - i.e. actual numbers of terminals (1 to 6) from which the data were collected, and for cases $G = 4, 5, 6$. Each column corresponds to one real terminal, while the rows denote the clusters to which the data vectors $\bar{x}$ were assigned. The table shows that terminals 1, 2, and 4 had quite specific profiles, data from terminal 3 fell to at least 3 different groups, while the data from terminals 5 and 6 did not show any clear characteristic profiles.

Table 3 then displays the final estimates of parameters of six clusters (i.e. the case with $G = 6$), namely triplets of parameters λ, μ, σ for four time intervals. All computations were made with the aid of R-language codes, both own and modified from libraries. The data we analyzed are available on *http://siprint.utia.cas.cz/public/income/volf/data_survival/proc_d6.txt.*

5 Conclusion

We have presented a problem of model based cluster analysis with a model that was the mixture of Gauss and Poisson distributions, though sometimes degenerated to one point. From this point of view, the procedure utilized was more-less a standard one, though with some interesting features, connected actually with the subject of application. The goal was the retrieval the set of profiles of data sources, in our case the payment terminals. The sense of determining the profiles of normal behavior of processes consists in that we are then able to compare them with the actual data and detect the changes. In this context, the outliers are not mere nuisance data, on the contrary, in certain situations they can be the data of a particular interest (e.g. in a fraud detection problem). That is also why the method initialized by [1] was selected. Naturally, a similar analysis is valuable not only in the area of financial processes analysis. Hence, it has also a sense to study the form of departures, of changes, and to find their relation to certain real phenomena. That is why we checked additionally the results of cluster analysis with the help of classification tree in order to find out the influence of components of data to final grouping. The analysis revealed that the most discriminating, from this point of view, was the amount of payments in the third part of the day, then the amounts and also frequencies of payments in the second interval. As regards the outliers, i.e. the data classified to group zero, as it has been expected, the most crucial were relatively rare too high payments in all time intervals, but also two cases with high frequencies of medium amounts.

References

[1] Banfield J.D., Raftery A.E. (1993). *Model-based Gaussian and non-Gaussian clustering*. Biometrics **49**, 803 – 821.

[2] Böhning D., Seidel W. (Guest Editors) (2003). *Recent developments in mixture models*. Comp. Statistics & Data Analysis **41**, No. 3 – 4.

[3] Fraley C.F., Raftery A.E. (1998). *How many clusters? Which clustering method? Answers via model-based cluster analysis*. Technical Report No. 329, Department of Statistics, University of Washington.

[4] Hastie T., Tibshirani R., Friedman J. (2001). *The elements of statistical learning*. Springer, New York.

Acknowledgement: The research has partly been supported by the grant of GA ČR No. 201/02/0049.

Address: L. Křečan, Faculty of Nuclear Sciences and Physical Engineering, ČVUT, Praha, Czech Republic
P. Volf, TU Liberec, Hálkova 6, 461 17 Liberec, Czech Republic
E-mail: lukas@krecan.net, petr.volf@vslib.cz

COMPSTAT'2004 Symposium

CONSISTENT ESTIMATION OF AN ELLIPSOID WITH KNOWN CENTER

Alexander Kukush, Ivan Markovsky and Sabine Van Huffel

Key words: Adjusted least squares, ellipsoid fitting, measurement error model.

COMPSTAT 2004 section: Applications.

Abstract: A consistent estimator is derived for the parameter of an implicit quadratic measurement error model. The true model describes a centered ellipsoid and the true values of the measurement points are on the boundary of the ellipsoid. The estimation procedure is applicable for ellipsoid fitting with known center. Due to the nonlinearity of the model the ordinary least squares and the orthogonal regression estimators are inconsistent. The proposed consistent estimator is derived by correcting the ordinary least squares cost function. The correction is explicitly given in terms of the measurement error variance. If the variance is unknown an estimator for this parameter is proposed as well. The measurement errors are assumed to form an i.i.d. sequence of normal random variables. The derived estimator is asymptotically normal.

1 Introduction

A parameter estimation problem occurs when the relation among some observed variables $x_1, \dots, x_n$ is described by a parameterized model. The parameters identify a unique model in a given *model class*, and the problem is to choose a model from the model class, given a set of observations $\{x^{(l)}\}_{l=1}^m$, where $x^{(l)} := [x_1^{(l)} \cdots x_n^{(l)}]^\top$ is the l-th observed vector of variables. The model is selected according to certain performance criteria, specified later.

We consider a *quadratic form model*, $x^\top A x = 1$, $A > 0$ relating the variables $x := [x_1 \cdots x_n]^\top$. The parameter of the model is the positive definite matrix A. The set of points $\mathcal{E}(A) := \{\, x \in \mathbb{R}^n : x^\top A x = 1 \,\}$ satisfying the quadratic form model is a centered hyper ellipsoid.

The vector of variables x is observed with additive error $\tilde{x} = [\tilde{x}_1 \cdots \tilde{x}_n]^\top$ and the error is described stochastically. The true value $\bar{x} = [\bar{x}_1 \cdots \bar{x}_n]^\top$ of the measured variables is assumed to satisfy the model for some unknown true value $\bar{A} > 0$ of the parameter. This assumption defines a *true model* in the model class. Models in which the variables are measured with additive noise $x = \bar{x} + \tilde{x}$ are called *measurement error models*.

The quadratic form model is linear in the parameters, so that the linear least squares technique can be applied. This corresponds to estimation criterion: $\min_A \sum_{l=1}^m (x^{(l)T} A x^{(l)} - 1)^2$. We will call the resulting estimator the *ordinary least squares* (OLS) estimator, in order to distinguish it from the adjusted least squares estimator, introduced later. The presence of measurement errors in all the covariates makes the OLS estimator biased, see, *e.g.*, [1].

Another approach for estimating the parameter of the quadratic from model from the observed data points is the *orthogonal regression* estimation. Let $\operatorname{dist}(x, \mathcal{E})$ be the Euclidean distance from the point x to the set $\mathcal{E}$. The orthogonal regression estimator is defined as a global solution of the following optimization problem: $\min_A \sum_{l=1}^{m} \operatorname{dist}(x^{(l)}, \mathcal{E}(A))^2$. The nonlinearity of the model with respect to the measurements, implies the inconsistency of this estimator as well, see [8] and the discussion in [3, p. 250].

We assume that the measurement errors $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(m)}$ are centered, independent among the measurement, and normally distributed, $\tilde{x}^{(l)} \sim \mathrm{N}(0, \bar{\sigma}^2 I)$ for all l, with noise variance $\bar{\sigma}^2 I$. We consider both cases, when $\bar{\sigma}^2$ is given, and when $\bar{\sigma}^2$ is unknown. The stochastic description of the measurement errors can be viewed as a model with parameter σ^2 (a nuisance parameter of the model).

Using the noise model assumptions, we apply an adjustment procedure that takes into account the quadratic structure of the model and corrects the OLS estimate appropriately. The resulting estimator, called an *adjusted least squares* (ALS) estimator, is consistent.

A nice feature of the ALS estimator is that its computation, as the computation of the OLS estimator requires solving a linear system of equations. If $\bar{\sigma}^2$ is a priori known, we give the corrected system in terms of $\bar{\sigma}^2$. If however, $\bar{\sigma}^2$ is unknown, then it has to be estimated together with the model parameters. We propose a consistent procedure to estimate the unknown measurement error variance.

The theory presented in the paper is generalized for the ellipsoid estimation problem with *unknown* center in [5]. The computational aspect of the ellipsoid estimation method with unknown center is treated in [7]. Due to space limitation the proofs of the statements are not given. They can be found in the technical report [4], available from

```
ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/markovsky/
```

2 Model $x^\top A x = 1$

Let $x^{(l)} \in \mathbb{R}^{n\times 1}$, $l = 1, \ldots, m$, represent observations of boundary points of an ellipsoid. The model is

$$x^{(l)T} A x^{(l)} \approx 1, \quad \text{for} \quad l = 1, \ldots, m, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is a positive semidefinite symmetric matrix to be estimated. We suppose that the vectors $x^{(l)}$ are observed with errors,

$$x^{(l)} = \bar{x}^{(l)} + \tilde{x}^{(l)}, \quad \text{for} \quad l = 1, \ldots, m, \tag{2}$$

and that there exists a symmetric positive semidefinite matrix $\bar{A} \in \mathbb{R}^{n\times n}$, such that $\bar{x}^{(l)T}\bar{A}\bar{x}^{(l)} = 1$, for $l = 1,\dots,m$. The matrix $\bar{A}$ is the true value of the parameter A in the model (1). The vectors $\bar{x}^{(l)}$ are the true values of the measurements $x^{(l)}$, and $\tilde{x}^{(l)}$ represent the measurement errors.

We derive a strongly consistent estimator for $\bar{A}$, based on the approach proposed in [6]. We suppose that the error terms $\{\tilde{x}^{(l)}, l \geq 1\}$ form an i.i.d. sequence and the distribution of $\tilde{x}^{(l)}$, denoted as $\mathcal{L}(\tilde{x}^{(l)})$, is normal $N(0,\sigma^2)$, with $\sigma^2 > 0$.

3 The adjusted least squares estimator

We start with the OLS objective function

$$q_{\mathrm{ls}}(A,x) := (x^\top Ax - 1)^2, \qquad A \in \mathbb{S}, \quad x \in \mathbb{R}^{n\times 1},$$

where $\mathbb{S}$ denotes the space of the $n\times n$ symmetric matrices. In the space of the square matrices $\mathbb{R}^{n\times n}$, we introduce a scalar product $\langle T,S\rangle := \mathrm{tr}(TS^\top)$, for $T,S \in \mathbb{R}^{n\times n}$. The derivative $\partial q_{\mathrm{ls}}/\partial A$ is a linear functional on $\mathbb{R}^{n\times n}$. It acts on $H \in \mathbb{R}^{n\times n}$ according to the rule

$$\frac{1}{2}\frac{\partial q_{\mathrm{ls}}}{\partial A}(H) = (x^\top Ax - 1)x^\top Hx = \ \langle (x^\top Ax-1)xx^\top,\ H\rangle. \tag{3}$$

We can identify the derivative $\partial q_{\mathrm{ls}}/\partial A$ with a matrix, which represents it in the equality (3). Thus

$$\frac{1}{2}\frac{\partial q_{\mathrm{ls}}}{\partial A} = (x^\top Ax)xx^\top - xx^\top := \psi_{\mathrm{ls}}(A,x).$$

In this section, we suppose that σ^2 is known, therefore the measurement error distribution is known.

We are looking for a corrected score function ψ, such that

$$\mathbf{E}\,\psi(A,\bar{x}+\tilde{x}) = \psi_{\mathrm{ls}}(A,\bar{x}), \qquad \text{for all } \bar{x} \in \mathbb{R}^{n\times 1}. \tag{4}$$

Here $\mathcal{L}(\tilde{x}) = \mathcal{L}(\tilde{x}^{(l)}) = N(0,\sigma^2)$. We have

$$\psi_{\mathrm{ls}}(A,x) = \psi_{\mathrm{ls},1}(A,x) - \psi_{\mathrm{ls},2}(A,x), \tag{5}$$

with $\psi_{\mathrm{ls},1}(A,x) := (x^\top Ax)xx^\top$, $\psi_{\mathrm{ls},2}(A,x) := xx^\top$. Using (5), we split the problem (4) into two problems. Thus we look for the auxiliary functions $\psi_s(A,x)$, $s = 1,2$, such that

$$\mathbf{E}\,\psi_s(A,\bar{x}+\tilde{x}) = \psi_{\mathrm{ls},s}(A,\bar{x}), \quad \text{for} \quad s = 1,2. \tag{6}$$

Then $\psi_2(A,x) = xx^\top - \sigma^2 I_n$, where I_n is the $n\times n$ identity matrix. We have for the (p,q)-th entry of the matrix $\psi_{\mathrm{ls},1}$, $[\psi_{\mathrm{ls},1}(A,x)]_{pq} = \sum_{i,j=1}^n a_{ij}x_ix_jx_px_q$.

Let f_{ijpq} be such a polynomial function of x, of order 4, that

$$\mathbf{E}\, f_{ijpq}(\bar{x}+\tilde{x}) = \bar{x}_i\bar{x}_j\bar{x}_p\bar{x}_q, \qquad \text{for all } \bar{x}\in\mathbb{R}^{n\times 1}. \tag{7}$$

Then the (p,q)-th entry of ψ_1 equals $[\psi_1(A,x)]_{pq} = \sum_{i,j=1}^{n} a_{ij} f_{ijpq}(x)$.

Therefore the (p,q)-th entry of ψ is

$$[\psi(A,x)]_{pq} = \sum_{i,j=1}^{n} a_{ij} f_{ijpq}(x) - x_p x_q + \sigma^2\delta_{pq}, \tag{8}$$

where δ_{pq} is the Kronecker symbol. Now, consider the function (7). Let t_s be such a polynomial that $\mathbf{E}\, t_s(\xi+\sigma N(0,1)) = \xi^s$, $s = 1,2,\ldots$ Then, see, *e.g.*, [2], $t_1(\xi)=\xi$, $t_2(\xi)=\xi^2-\sigma^2$, $t_3(\xi)=\xi^3-3\xi\sigma^2$, $t_4(\xi)=\xi^4-6\xi^2\sigma^2+3\sigma^4$. We have the following cases.

a) All i, j, p, q are different. Then $f_{ijpq}(x) = x_i x_j x_p x_q$.

b) $i=j=p$, $q\neq i$ (with permutations). Then $f_{iiiq}(x) = x_q t_3(x_i)$.

c) $i=j=p=q$. Then $f_{iiii}(x) = t_4(x_i)$.

d) $i=j$, $p=q$, $i\neq p$. Then $f_{iipp}(x) = t_2(x_i)t_2(x_p)$.

e) $i=j$ and i, p, q are different. Then $f_{iipq}(x) = x_p x_q t_2(x_i)$.

The preliminary adjusted least squares (ALS) estimator $\hat{A}_1$ is defined from (8) by the observations (2) as a solution of the set of equations

$$\sum_{i,j=1}^{n} a_{ij}\cdot\frac{1}{m}\sum_{l=1}^{m} f_{ijpq}(x^{(l)}) = \frac{1}{m}\sum_{l=1}^{m} x_p^{(l)}x_q^{(l)} - \sigma^2\delta_{pq}, \qquad 1\le p\le q\le n. \tag{9}$$

Since $a_{ij}=a_{ji}$, the values a_{ij}, $1\le i\le j\le n$ identify the matrix $A\in\mathbb{S}$. Here $x_p^{(l)}$ is the p-th coordinate of the vector $x^{(l)}\in\mathbb{R}^{n\times 1}$. Later on we will show that under appropriate conditions the system of equations (9) has a unique solution $(\hat{a}_{ij},\ 1\le i\le j\le n)$, for all $m\ge\bar{m}(\omega)$, a.s. The estimator $\hat{A}_1$ is defined as a random symmetric matrix, of which the entries coincide with the unique solution of (9), provided (9) actually has a unique solution. At the second stage, the ALS estimator $\hat{A}$ is defined as the projection of $\hat{A}_1$ on the subspace of the symmetric positive semidefinite matrices.

4 Consistency

We need two assumptions.

(i) There exist a constant $c_1 > 0$ and a number $\bar{m}$, such that for each $m\ge\bar{m}$ and for each $H\in\mathbb{S}$, $\frac{1}{m}\sum_{l=1}^{m}\left(\bar{x}^{(l)T}H\bar{x}^{(l)}\right)^2 \ge c_1||H||_F^2$, where $||\bullet||_F$ denotes the Frobenius norm.

(ii) There exists a constant $c_2 > 0$ and a number $\gamma \in [0, 1)$, such that $\frac{1}{m}\sum_{l=1}^{m} ||\bar{x}^{(l)}||^6 \leq c_2 m^{\gamma}$, for all $m \geq 1$, where $||\bullet||$ denotes the Euclidean norm.

The assumption (i) is a contrast condition, see the discussion in [6]. A necessary, but not sufficient condition for (i) is that $\mathrm{span}\{\bar{x}^{(l)}\bar{x}^{(l)T},\ l \geq 1\} = \mathbb{S}$. Assumption (i) means that the true points $\bar{x}^{(l)}$, $l \geq 1$ are nicely dispersed on the boundary of the ellipsoid.

The assumption (ii) is a restriction from above. The $\bar{x}^{(l)}$'s are not allowed to grow quickly. This condition will be used in the corresponding strong law of large numbers. If $\bar{A}$ is positive definite then the ellipsoid is a bounded set, and (ii) holds with $\gamma = 0$.

Theorem 4.1 (Strong consistency). *Assume that conditions (i) and (ii) hold. Then* $||\hat{A} - \bar{A}||_F \to 0$, *as* $m \to \infty$, *a.s.*

5 Asymptotic normality

Again we suppose that σ^2 is known, and under stronger assumptions derive the asymptotic normality of the estimator $\hat{A}_1$. If $\bar{A}$ is positive definite, this implies the asymptotic normality of $\hat{A}$. First we strengthen the assumptions (i) and (ii).

1. The sequence of operators $L_m A := \frac{1}{m}\sum_{l=1}^{m} \left(\bar{x}^{(l)T} A \bar{x}^{(l)}\right)\bar{x}^{(l)}\bar{x}^{(l)T}$ converges to an operator L, which is positive definite on $\mathbb{S}$.

2. The sequence $\{\bar{x}^{(l)},\ l \geq 1\}$ is bounded.

Assumption (1.) implies assumption (i), and assumption (2.) implies assumption (ii) with $\gamma = 0$. Recall that assumption (2) holds if $\bar{A}$ is positive definite. The next assumption is about the stability of empirical moments of $\bar{x}^{(l)}$ up to the sixth order.

3. For each $s = 1, 2, \ldots, 6$ and for each $\{i_1, \ldots, i_s\} \subset \{1, \ldots, n\}$, there exists a finite limit $\min_{m\to\infty} \frac{1}{m}\sum_{l=1}^{m} \bar{x}_{i_1}^{(l)}\bar{x}_{i_2}^{(l)} \cdots \bar{x}_{i_s}^{(l)}$.

Theorem 5.1 (Asymptotic normality).

a) Assume that the conditions (1.) to (3.) hold. Then $\sqrt{m}\cdot\mathrm{vec}(\hat{A}_1 - \bar{A}) \xrightarrow{d} N(0, \Sigma)$, *where* Σ *is a positive semidefinite covariance matrix.*

b) If $\bar{A}$ *is positive definite then the conditions (1.) and (3.) imply the convergence* $\sqrt{m}\cdot\mathrm{vec}(\hat{A} - \bar{A}) \xrightarrow{d} N(0, \Sigma)$.

6 The case of unknown noise variance

We start with the following observation: under σ^2 known and using the ALS estimator, we can re-estimate σ^2.

Lemma 6.1. *Assume that conditions (i) and (ii) hold and, in addition, that* $\frac{1}{m}\sum_{l=1}^{m}||\bar{x}^{(l)}||^2 = \mathrm{O}(1)$. *Then the ALS estimator* $\hat{A}$ *satisfies the relation*

$$\varphi(\hat{A}) := \frac{1}{\mathrm{tr}(\hat{A}}) \cdot \frac{1}{m}\sum_{l=1}^{m}\left(x^{(l)T}\hat{A}x^{(l)} - 1\right) \to \sigma^2, \quad \textit{as} \quad m \to \infty, \ \textit{a.s.} \tag{10}$$

Lemma (6.1) shows that if σ^2 is indeed the true value of the variance and based on this number we construct $\hat{A}$, then $\varphi(\hat{A})$ is close to σ^2. We propose the following empirical criterion: select σ^2 to minimize $|\varphi(\hat{A}_{\sigma^2}) - \sigma^2|$, where $\hat{A}_{\sigma^2}$ is the ALS estimator, determined by taking the number σ^2 as the true value of the variance. This gives the following iterative algorithm:

```
1. Given $\{x^{(l)}\}_{l=1}^{m}$ and a convergence tolerance $\varepsilon$.
2. Select an initial guess $\sigma^2(0) > 0$ for $\bar{\sigma}^2$, e.g., $\sigma^2(0) = \frac{\lambda}{nm}\sum_{l=1}^{m}||x^{(l)}||^2$,
   for some $0 < \lambda < 1/2$ and let $k := 0$.
3. until $|\sigma^2(k-1) - \sigma^2(k)| \geq \varepsilon$ do
    3.1. Calculate $\hat{A}_{\sigma^2(k)}$.
    3.2. Let $\sigma^2(k+1) := \max\{0, \varphi(\hat{A}_{\sigma^2(k)})\}$ and $k := k+1$.
4. If $|\sigma^2(k-1) - \sigma^2(k)| \geq \varepsilon$, then we select $\hat{\sigma}^2$, as $\sigma^2(s)$ for which
       $|\sigma^2(s-1) - \sigma^2(s)| = \min_{1 \leq k \leq N} |\sigma^2(k-1) - \sigma^2(k)|.$
   Otherwise we select $\hat{\sigma}^2 = \sigma^2(k)$.
5. Return the noise variance estimate $\hat{\sigma}^2$ and the parameter estimate $\hat{A}_{\hat{\sigma}^2}$.
```

7 Simulation examples

In this section, we show simulation examples, comparing the ALS estimators with the OLS estimator, and the orthogonal regression (OR) estimator. In all experiments, the true value of the parameter is

$$\bar{A} = \begin{bmatrix} 1 & -0.75 \\ -0.75 & 1 \end{bmatrix}.$$

First, we consider an estimation with sample size $m = 100$ data points and noise with standard deviation $\sigma = 0.3$. The true values of the data points are equidistantly distributed on the boundary of the ellipsoid. For a particular noise realization, the relative errors of estimation $e := ||\hat{A} - \bar{A}||_F / ||\bar{A}||_F$, for the different estimates are: $e_{\mathrm{ols}} = 0.4966$, $e_{\mathrm{als1}} = 0.0534$, $e_{\mathrm{als2}} = 0.0350$, $e_{\mathrm{orth}} = 0.0357$, where "ALS1" stands for the ALS estimation with known noise variance and "ALS2" stands for the ALS estimation with unknown noise variance. Figure (1), left, shows the data points used for the estimation and Figure (1), right, shows the estimated ellipses and the true ellipse. Note that the ALS2 estimator is better than the ALS1 estimator. This relation also holds when the relative error is averaged over many repetitions of the experiment.

Figure (2) shows the averaged relative error of estimation $\bar{e} = \frac{1}{N}\sum_{k=1}^{N} e^{(k)}$ as a function of the sample size m. Here $\hat{A}^{(k)}$ is the estimate obtained on the k-th experiment and in total $N = 500$ repetitions of the estimation, with different noise realizations, are used. The noise standard deviation is $\sigma = 0.1$.

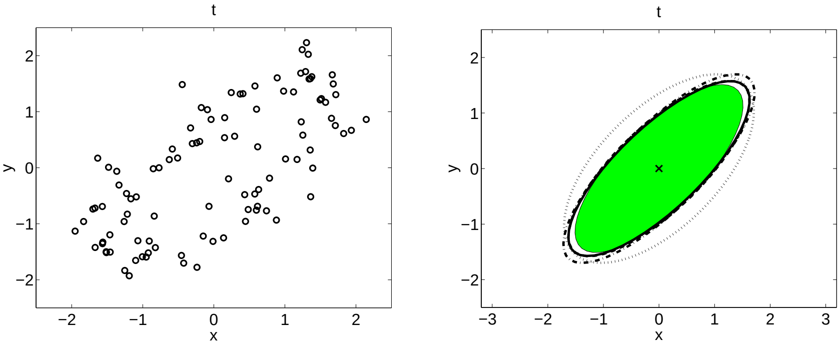


Figure 1: True ellipse—shaded, OLS—dotted, ALS1—solid, ALS2—dashed (coincides with ALS1), OR—dashed-dotted.

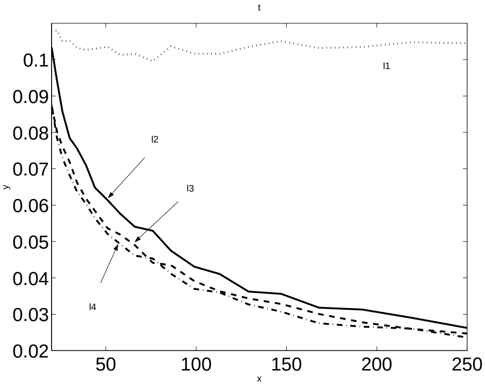


Figure 2: Relative error of estimation as a function of the sample size.

8 Conclusion

We considered a problem of estimating an ellipsoid with known center from observations of points on its boundary. We assumed that the measurement errors are normal. In the case of known variance, we constructed the ALS estimator, starting with the OLS objective function and using the adjustment for the measurement errors. Under appropriate conditions, we proved the strong consistency and asymptotic normality of the estimator. Though we did not present an approximate asymptotic covariance matrix, it can be obtained using the empirical mixed moments of the data up to the sixth order. In the case of unknown variance, we proposed an iterative procedure that estimates simultaneously the noise variance and the model parameters.

References

[1] Carroll R.J., Ruppert D., Stefanski L.A. (1995). *Measurement error in nonlinear models.* Number 63 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

[2] Cheng C., Schneeweiss H. (1998). *Polynomial regression with errors in the variables.* J. R. Statistical Society B **60**, 189 – 199.

[3] Fuller W. A. (1987). *Measurement error models.* New York: Wiley.

[4] Kukush A., Markovsky I., Van Huffel S. (2002). *Consistent estimation of an ellipsoid with known center from observations of points on its boundary.* TR#02–119, Dept. EE, K.U. Leuven.

[5] Kukush A., Markovsky I,. Van Huffel S. (2004). *Consistent estimation in an implicit quadratic measurement error model.* To appear in Comp. Stat. & Data Anal.

[6] Kukush A., Zwanzig S. (2001). *On adaptive minimum contrast estimators in the implicit nonlinear functional regression models.* Ukr. Math. Journal **53(9)**, 1204 – 1209.

[7] Markovsky I., Kukush A., Van Huffel S. (2004). *Consistent least squares fitting of ellipsoids.* To appear in Numerische Mathematik.

[8] Neyman J., Scott E. L. (1948). *Consistent estimates based on partially consistent observations.* Econometrica **16(1)**, 1 – 32.

Acknowledgement: I. Markovsky is supported by a K.U.Leuven doctoral scholarship. Our research is supported by: Research Council KUL: GOA-Mefisto 666, Flemish Government: FWO grants G.0078.01 (structured matrices), G.0270.02 (nonlinear Lp approximation), research communities (ICCoS, ANMMM); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006)).

Address: A. Kukush, I. Markovsky, S. Van Huffel, K.U.Leuven, ESAT-SCD,
Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium
A. Kukush is on a leave from National Taras Shevchenko University,
Vladimirskaya st. 64, 01601, Kiev, Ukraine

E-mail: alexander_kukush@univ.kiev.ua,
ivan.markovsky@esat.kuleuven.ac.be,
sabine.vanhuffel@esat.kuleuven.ac.be

LEARNING FROM DATA AS AN INVERSE PROBLEM

Věra Kůrková

Key words: Learning from data, generalization, minimization of empirical error, regularization, kernel methods.

COMPSTAT 2004 section: Neural networks and machine learning.

Abstract: In this paper, we reformulate the problem of minimization of an empirical error functional as a linear inverse problem by introducing a suitable operator. We describe properties of this operator (compactness, representation of its adjoint) and apply theory of continuous linear inverse problems in the domain of infinite dimensional Hilbert spaces. We describe relationship between a pseudosolution and regularized solutions for variable regularization parameters and analyze improvements of stability that can be obtained by regularization in terms of condition numbers of Gram matrices and size of data samples.

1 Introduction

The goal of supervised learning is to adjust parameters of a neural network so that it approximates with a sufficient accuracy a functional relationship between inputs and outputs. Typically, such a relationship is not known analytically. Instead, a training set is given consisting of a sample of input/output pairs $z = \{(x_i, y_i) \in \mathcal{R}^d \times \mathcal{R}, i = 1, \dots, m\}$. So the task of learning is to find a function from a hypothesis set formed by functions computable by a given class of neural networks that approximates the sample of *empirical data*. A similar task of finding a function fitting to astronomical data was solved by Gauss and Legendre by the *least square method*, i.e., minimization of the sum of squares of errors. The least square method became popular in statistics and engineering and was also used in many neural network learning algorithms such as backpropagation.

The problem of finding a function from a given parameterized family fitting to empirical data belongs to a wider class of *inverse problems* of determining unknown causes (such as shapes of functions, forces, shapes of distributions) from known consequences (empirical data). Inverse problems are fundamental in various domains of applied science such as medical diagnostics (tomography), seismology and meteorological forecasting. The dependence of consequences on causes is usually modelled by an operator, the simplest type of which is linear. For finite dimensional case, the theory of linear inverse problems is based on *Moore-Penrose pseudoinverse* of a matrix. Pseidoinverse method was generalized to infinite dimensional Hilbert spaces [5], [2] and combined with *regularization* introduced by Tikhonov and Arsenin [14] to develop a theory describing properties of least-squares pseudosolutions, their stability and their relationship to regularized solutions [2]. Modelling of generalization based on Tikhonov's regularization was introduced by Poggio and Girosi [12]. Later Girosi [4] considered regularization in the domain of a special class of Hilbert spaces, called *reproducing kernel Hilbert spaces* (RKHS), the norms on which can play a role of measures of various types of oscillations and thus enable to model a variety of *conceptual data*, which has to be added to the empirical ones to guarantee generalization capability. RKHS, defined by Aronszajn [1], were introduced into interpolation of data by Parzen [11] and Wahba [15]. For a survey of applications of RKHS to learning see, e.g., [3], [13].

2 Minimization of empirical error as an inverse problem

Let Ω be a nonempty set, m a positive integer and $z = \{(x_i, y_i) \in \Omega \times \mathcal{R},\ i = 1, \dots, m\}$ be a sample of pairs of data. A standard approach to learning from data used , e.g., in backpropagation is based on minimization of the *empirical error* functional defined as $\mathcal{E}_{z,V}(f) = \frac{1}{m}\sum_{i=1}^{m} V(f(x_i), y_i)$, where $V : \mathcal{R}^2 \to [0, \infty)$ satisfying $V(y, y) = 0$ for all $y \in \mathcal{R}$ is a *loss function* that measures how much is lost when $f(x)$ is computed instead of y. The most common loss function is the *square loss* $V(f(x), y) = (f(x) - y)^2$. To simplify notation, we denote by $\mathcal{E}_z$ the empirical error functional with the square loss function, i.e., $\mathcal{E}_z(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$.

Using a standard terminology from the theory of optimization we denote by (M, Φ) the problem of minimization of a functional Φ over a set M, which is called a *hypothesis set*. Every $f^o \in M$ such that $\Phi(f^o) = \min_{f \in M} \Phi(f)$ is called a *solution* of the problem (M, Φ). We denote by $argmin(M, \Phi) = \{f^o \in M : \Phi(f^o) = \min_{f \in M} \Phi(f)\}$ the set of all solutions of (M, Φ).

The problem of minimization of the empirical error functional can be studied in the framework of theory of inverse problems. Given an operator $A : (X, \|.\|_X) \to (Y, \|.\|_Y)$ between Banach spaces, an *inverse problem* defined by A is to find for $g \in Y$ some $f \in X$ such that $A(f) = g$ [2]. An inverse problem is called *linear* when A is a linear operator. Elements of X are called *solutions* and elements of Y *data*. When Y is finite dimensional, the inverse problem is called a problem with *discrete data*.

If for every $g \in Y$ there exists a unique solution $f \in X$, then the inverse problem is called *well-posed*. So for a well-posed inverse problem, there exists a unique inverse operator $A^{-1} : Y \to X$. When A is continuous, then by the Banach open map theorem A^{-1} is continuous, too. Even a continuous dependence of solutions on data does not always guarantee robustness against a noise. As a measure of stability of solutions of an inverse problem is used the *condition number* defined for a well-posed problem given by an operator A as $cond(A) = \|A\|\,\|A^{-1}\|$.

Often, inverse problems are ill-posed or ill-conditioned. When a solution does not exist, one can search for best approximate solution f^o, called a *pseudosolution*, defined by $\|A(f^o) - g\|_Y = \min_{f \in X} \|A(f) - g\|_Y$ and a *nor-*

mal pseudosolution f^+, which is a pseudosolution of the minimal norm, i.e., $\|f^+\|_X = \min\{\|f^o\|_X : f^o \in S(g)\}$, where $S(g)$ is the set of all psedosolutions of the inverse problem given by an operator A and data g. When for every $g \in Y$ there exists a normal pseudosolution f^+, then a *pseudoinverse operator* $A^+ : Y \to X$ can be defined as $A^+(g) = f^+$. Similarly as in the case of well-posed problems, the *condition number* of an operator A with a pseudoinverse A^+ is defined as $cond(A) = \|A\|\,\|A^+\|$.

For X and Y finite dimensional, the pseudosolution can be described in terms of Moore-Penrose pseudoinverse of the matrix corresponding to the operator A. The concept of Moore-Penrose pseudoinversion has been extended to the case of linear continuous operators between Hilbert spaces [5]. To take advantage of the theory of generalized inversion in Hilbert spaces, we express as an inverse problem the problem $(X, \mathcal{E}_z)$ of minimization of the empirical error $\mathcal{E}_z$ over a Hilbert space X of functions on some set Ω. Let $z = (x, y)$, where $x = (x_1, \ldots, x_m) \in \Omega^m$ and $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$, be a sample defining the empirical error functional $\mathcal{E}_z$. Consider an operator $L_x : X \to \mathcal{R}^m$ defined as $L_x(f) = \left(\frac{f(x_1)}{\sqrt{m}}, \ldots, \frac{f(x_m)}{\sqrt{m}}\right)$. Then $\mathcal{E}_z$ can be represented as $\mathcal{E}_z = \left\|L_x - \frac{y}{\sqrt{m}}\right\|_2^2$, where $\|.\|_2$ denotes the l_2-norm on $\mathcal{R}^m$. Similarly, $\langle . \rangle_2$ denotes the inner product on $\mathcal{R}^m$, while $\|.\|_X$ and $\langle . \rangle_X$ denote the norm and the inner product, resp., on X.

Thus the problem of minimization of $\mathcal{E}_z$ over X is equivalent to the problem of finding a pseudosolution of the inverse problem given by the operator L_x for the data $\frac{y}{\sqrt{m}}$. As the range of the operator L_x is finite dimensional, this problem belongs to the class of problems with discrete data. When $(X, \|.\|_X)$ is chosen in such a way that L_x is continuous, we can apply the following theorem summarizing properties of the pseudosolution of a continuous linear operator stated in [2, pp. 56-60] and in [5, pp.37-46].

For any operator $A : X \to Y$, we denote by $N(A) = \{f \in X : A(f) = 0\}$ its *null space*, by $R(A) = \{g \in Y : (\exists f \in X)(A(f) = g)\}$ its *range*, by $\pi_R : Y \to R(A)$ the projection of Y onto $R(A)$ and if A has an adjoint A^*, by $\pi_N : Y \to N(A^*)$ the projection of Y onto the null space of A^*. For any $g \in Y$, we denote $S(g) = \{f^o \in X : \|A(f^o) - g\|_Y = \min_{f \in X} \|A(f) - g\|_Y\}$.

Theorem 2.1. *Let X, Y be Hilbert spaces, $A : X \to Y$ be a continuous linear operator with a closed range, then:*
(i) A has an adjoint A^*;
(ii) $R(A)$ is closed and $N(A^*) \oplus R(A) = Y$;
(iii) there exists a unique continuous linear operator $A^+ : Y \to X$ such that for every $g \in Y$, $A^+(g) \in S(g)$, $\|A^+(g)\|_X = \min_{f^o \in S(g)} \|f^o\|_X$ and $S(g) = \{A^+(g) + f : f \in N(A)\}$;
(iv) for every $g \in Y$, $AA^+(g) = \pi_R(g)$;
(v) $A^+ = (A^*A)^+A^* = A^*(AA^*)^+$.

3 Minimization of empirical errors over reproducing kernel Hilbert spaces

To apply Theorem 2.1 to learning from data we need to find proper hypothesis spaces (formed by functions defined on some sets Ω), on which the operators L_x are continuous for all $x = (x_1, \dots, x_m) \in \Omega^m$. $(\mathcal{L}_2(\Omega), \|.\|_{\mathcal{L}_2})$ cannot be used as such a hypothesis space as its elements are not pointwise defined functions. But even the subspace of the space of continuous functions $\mathcal{C}(\Omega)$ containing functions with finite $\mathcal{L}_2$-norms is not suitable as some L_x might not be continuous on this space. For example, for $\Omega = \mathcal{R}^d$, L_0 defined as $L_0(f) = f(0)$ is not bounded and hence it cannot be continuous (L_0 maps the sequence $\left\{n^d e^{-(\frac{\|x\|}{n})^2}\right\}$ of functions with $\mathcal{L}_2$-norms equal to 1 to an unbounded sequence of real numbers). But there exits a large class of Hilbert spaces, on which operators L_x are continuous. Moreover, norms on spaces from this class can play roles of measures of various types of oscillations of input/output mappings.

A reproducing kernel Hilbert space RKHS is a Hilbert space formed by functions defined on a nonempty set Ω such that for every $x \in \Omega$ the evaluation functional $\mathcal{C}_x$, defined for any f in the Hilbert space as $\mathcal{C}_x(f) = f(x)$, is bounded [1], [3]. RKHS can be elegantly characterized in terms of *kernels*, which are *symmetric positive semidefinite functions* $K : \Omega \times \Omega \to \mathcal{R}$, i.e., functions satisfying for all m, all $(w_1, \dots, w_m) \in \mathcal{R}^m$, and all $(x_1, \dots, x_m) \in \Omega^m$, $\sum_{i,j=1}^m w_i\, w_j\, K(x_i, x_j) \geq 0$. A kernel is *positive definite* if $\sum_{i,j=1}^m w_i w_j K(x_i, x_j) = 0$ for any distinct $x_1, \dots, x_m$ implies that for all $i = 1, \dots, m$, $w_i = 0$.

To every RKSH one can associate a unique kernel $K : \Omega \times \Omega \to \mathcal{R}$ such that for every f in the RKHS and $x \in \Omega$, $f(x) = \langle f, K_x \rangle_K$, where $K_x : \Omega \to \mathcal{R}$ is defined as $K_x(y) = K(x, y)$ for all $y \in \Omega$. On the other hand, every kernel $K : \Omega \times \Omega \to \mathcal{R}$ generates a RKHS denoted by $(\mathcal{H}_K(\Omega), \|.\|_K)$, which is defined as the completion of the linear span of the set of functions $\{K_x : x \in \Omega\}$ with the inner product defined by $\langle K_x, K_y \rangle_K = K(x, y)$ [1].

For a kernel $K : \Omega \times \Omega \to \mathcal{R}$, a positive integer m and a vector $x = (x_1, \dots, x_m)$, by $\mathcal{K}[x]$ is denoted the $m \times m$ matrix defined as $\mathcal{K}[x]_{i,j} = K(x_i, x_j)$, which is called the *Gram matrix of the kernel* K *with respect to the vector* x.

The following theorem describes properties of inverse problems defined by operators L_x on RKHSs.

Proposition 3.1. *Let* $K : \Omega \times \Omega \to \mathcal{R}$ *be a kernel,* m *be a positive integer, and* $z = (x, y)$*, where* $x = (x_1, \dots, x_m) \in \Omega^m, y = (y_1, \dots, y_m) \in \mathcal{R}^m$*, then:*
(i) $L_x : \mathcal{H}_K(\Omega) \to \mathcal{R}^m$ *is a Lipschitz continuous compact linear operator with a closed range;*
(ii) the adjoint operator ${L_x}^* : \mathcal{R}^m \to \mathcal{H}_K(\Omega)$ *is compact and satisfies for every* $u \in \mathcal{R}^m$, ${L_x}^*(u) = \frac{1}{\sqrt{m}} \sum_{i=1}^m u_i K_{x_i}$*;*

(iii) $R(L_x)$ *is closed and* $N(L_x^*) \oplus R(L_x) = \mathcal{R}^m$ *and when* K *is positive definite, then* $N(L_x{}^*) = \{0\}$ *and* $R(L_x) = \mathcal{R}^m$*;*
(iv) $L_x L_x^* : \mathcal{R}^m \to \mathcal{R}^m$ *can be represented by the matrix* $\frac{1}{m}\mathcal{K}[x]$*;*
(v) there exists a continuous linear pseudoinverse operator $L_x^+ : \mathcal{R}^m \to \mathcal{H}_K(\Omega)$ *such that for every* $u \in \mathcal{R}^m$, $L_x L_x^+(u) = \pi_R(u)$ *and when* K *is positive definite, then* $L_x L_x^+(u) = u$*;*
(vi) $L_x^+ = (L_x^* L_x)^+ L_x^* = L_x^*(L_x L_x^*)^+$.

The next theorem states properties of the solutions of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_z)$.

Theorem 3.1. *Let* $K : \Omega \times \Omega \to \mathcal{R}$ *be a kernel,* m *be a positive integer and* $z = (x, y)$*, where* $x = (x_1, \ldots, x_m) \in \Omega^m$*,* $x_1, \ldots, x_m$ *are distinct and* $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$*, then:*
(i) $L_x^+(\frac{y}{\sqrt{m}}) \in$ *argmin*$(\mathcal{H}_K(\Omega), \mathcal{E}_z)$*, for every* $f^o \in$ *argmin*$(\mathcal{H}_K(\Omega), \mathcal{E}_z)$*,* $\|L_x^+(\frac{y}{\sqrt{m}})\|_K \le \|f^o\|_K$ *and argmin*$(\mathcal{H}_K(\Omega), \mathcal{E}_z) = L_x^+(\frac{y}{\sqrt{m}}) + N(L_x)$*;*
(ii) for every $f^o \in$ *argmin*$(\mathcal{H}_K(\Omega), \mathcal{E}_z)$*,* $L_x(f^o) = \pi_R(\frac{y}{\sqrt{m}})$ *and when* K *is positive definite,* $L_x(f^o) = \frac{y}{\sqrt{m}}$*;*
(iv) $\min_{f \in \mathcal{H}_K(\Omega)} = \frac{1}{m}\|\pi_R(y) - y\|_2^2$ *and when* K *is positive definite, then* $\min_{f \in \mathcal{H}_K(\Omega)} \mathcal{E}_z(f) = 0$*;*
(v) $L_x^+(\frac{y}{\sqrt{m}}) = \sum_{i=1}^m c_i K_{x_i}$*, where* $c = (c_1, \ldots, c_m) = \mathcal{K}[x]^+ y$*;*
(vi) for every $f^o \in$ *argmin*$(\mathcal{H}_K(\Omega), \mathcal{E}_z)$*,* $\sum_{i=1}^m f^o(x_i) K_{x_i} = \sum_{i=1}^m y_i K_{x_i}$ *and when* K *is positive definite, then* f^o *interpolates the data* z*, i.e.,* $f^o(x_i) = y_i$ *for all* $i = 1, \ldots, m$*.*

So for every kernel K and every sample of empirical data z, there exists a solution of the problem of minimization of the empirical error functional $\mathcal{E}_z$ over the space $\mathcal{H}_K(\Omega)$. The set of such solutions *argmin*$(\mathcal{H}_K(\Omega), \mathcal{E}_z)$ is a closed convex set of the form $\sum_{i=1}^m c_i K_{x_i} + N(L_x)$, where $c = \mathcal{K}[x]^+ y$ and $N(L_x)$ is the null space of the operator L_x. Minimum of $\mathcal{E}_z$ over $\mathcal{H}_K(\Omega)$ is equal to $\frac{1}{m}\|\pi_R(y) - y\|_2^2$, where π_R is the projection of $\mathcal{R}^m$ onto $R(L_x)$. For K positive definite, the solution interpolates the data and minimum of $\mathcal{E}_z$ over $\mathcal{H}_K(\Omega)$ is equal to zero.

Stability of the solution $\sum_{i=1}^m c_i K_{x_i}$ with respect to a small perturbation of the vector of output data y depends on the condition number of the matrix $\mathcal{K}[x]$. The solution is robust against noise only when the condition number is close to 1. For Gaussian kernels G_ρ, upper bounds on such condition numbers growing with the dimension d of the input data and the product ρq^2, where q is the separation radius of the input data x (which is defined as $q = \frac{1}{2}\min\{\|x_i - x_j\|_2 : i, j = 1, \ldots, m, i \ne j\}$), but independent on the size m of the data sample, were derived in [10].

4 Learning with generalization as a regularized inverse problem

The function $\sum_{i=1}^m c_i K_{x_i}$ with $c = \mathcal{K}[x]^+ y$ guarantees the best fit to the sample of data z that can be achieved using functions from the space $\mathcal{H}_K(\Omega)$. By choosing as a hypothesis space a RKHS, we impose a condition on oscillations of potential solutions. The type of such a condition can be illustrated on convolution kernels $K : \mathcal{R}^d \times \mathcal{R}^d \to \mathcal{R}$ satisfying $K(u,v) = k(u-v)$ for some $k : \mathcal{R} \to \mathcal{R}$ with positive Fourier transform $\tilde{k}$. For such kernels $\|f\|_K^2 = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{R}^d} \frac{\tilde{f}(\omega)^2}{\tilde{k}(\omega)} d\omega$ [4].

The restriction on potential solutions can be further strengthened by penalizing the size of the norm $\|.\|_K$ of the solution. This approach to constraining solutions of ill-posed inverse problems has been developed in 1960th by several authors. It is called *Tikhonov's regularization* due to Tikhonov's unifying formulation [14]. Tikhonov's regularization replaces the problem of minimization of the functional $\|A(.) - g\|_Y^2$ over X with minimization of $\|A(.) - g\|_Y^2 + \gamma \|.\|_X^2$, where the *regularization parameter* γ plays the role of a trade-off between fitting to empirical and conceptual data. The following theorem summarizes properties of solutions of regularized inverse problems stated in [2, pp.68-70] and [5, pp.74-76]. By I is denoted the identity operator $I : \mathcal{R}^d \to \mathcal{R}^d$ and by $\mathcal{I}$ the corresponding $m \times m$ matrix.

Theorem 4.1. *Let X, Y be Hilbert spaces, $A : X \to Y$ be a continuous linear operator with a closed range, then:*
(i) for every $\gamma > 0$, there exists a unique operator $A^\gamma : Y \to X$ such that for every $g \in Y$, $\{A^\gamma(g)\} = argmin(X, \|A(.) - g\|_Y^2 + \gamma \|.\|_X^2)$;
(ii) for every $\gamma > 0$, $A^\gamma = (A^*A + \gamma I)^{-1} A^* = A^*(AA^* + \gamma I)^{-1}$;
(iii) for every $g \in Y$, $e_g : (0,\infty) \to (0,\infty)$ defined as $e_g(\gamma) = \|AA^\gamma(g) - g\|_Y$ is strictly increasing, $\lim_{\gamma \to 0} e_g(\gamma) = \|\pi_N(g)\|_Y$ and $\lim_{\gamma \to \infty} e_g(\gamma) = \|g\|_Y$;
(iv) for every $g \in Y$, $E_g : (0,\infty) \to (0,\infty)$ defined as $E_g(\gamma) = \|A^\gamma(g)\|_X$ is strictly decreasing, $\lim_{\gamma \to 0} E_g(\gamma) = \|A^+(g)\|_X$ and $\lim_{\gamma \to \infty} E_g(\gamma) = 0$.

So even if the original inverse problem is ill-posed (it does not have a unique solution), for every $\gamma > 0$, the regularized problem has a unique solution. This is due to uniform convexity of the functional $\|.\|_Y^2$ (see, e.g., [9]). With γ going to zero, the solutions of regularized problem converge to the pseudosolution with the minimal norm $A^+(g)$. The next theorem describes properties of regularized solutions, their relationship to the pseudosolution and improvement of stability achievable using regularization for the problem of minimization of $\mathcal{E}_z$ over a RKHS.

Theorem 4.2. *Let $K : \Omega \times \Omega$ be a kernel, m be a positive integer, $z = (x, y)$, where $x = (x_1, \ldots, x_m) \in \Omega^m$, $x_1, \ldots, x_m$ are distinct, $y = (y_1, \ldots, y_m) \in \mathcal{R}^m$ and $\gamma > 0$, then:*
(i) there exists a unique solution f^γ of the problem $(\mathcal{H}_K(\Omega), \mathcal{E}_z + \gamma \|.\|_K^2)$;
(ii) $f^\gamma = \sum_{i=1}^m c_i K_{x_i}$, where $c = (\mathcal{K}[x] + \gamma m \mathcal{I})^{-1} y$;

(iii) $e : (0, \infty) \to [0, \infty)$ *defined as* $e(\gamma) = \mathcal{E}_z(f^\gamma)$ *is strictly increasing,* $\lim_{\gamma\to\infty} e(\gamma) = \frac{1}{\sqrt{m}}\|y\|_2$ *and* $\lim_{\gamma\to 0} e(\gamma) = \|\pi_R(y) - y\|_2$*, which is equal to* 0 *for* K *positive definite;*
(iv) $E : (0, \infty) \to [0, \infty)$ *defined as* $E(\gamma) = \|f^\gamma\|_K$ *is strictly decreasing,* $\lim_{\gamma\to 0} E(\gamma) = \| \sum_{i=1}^m a_i K_{x_i}\|_K$*, where* $a = \mathcal{K}[x]^+ y$*, and* $\lim_{\gamma\to\infty} E(\gamma) = 0$*;*
(v) when K *is positive definite, then* $cond(\mathcal{K}[x]+\gamma m\mathcal{I}) = 1+\frac{(cond(\mathcal{K}[x])-1)\lambda_{\min}}{\lambda_{\min}+\gamma m}$*, where* $\lambda_{\min}$ *is the minimal eigenvalue of* $\mathcal{K}[x]$*.*

Theorem 4.2 (ii) shows that the Representer Theorem [13], [3], [15] on learning from data in RKHS is a special case of a more general result from theory of regularization of inverse problems in Hilbert spaces. Note that some direct proofs of the Representer Theorem such as the one in [13] use the same argument based on annihilation of all directional derivatives as the proof of Theorem 4.1(ii) [5, pp.74-75], [2, pp.68-69].

Moreover, Theorem 4.2(v) shows how much ill-conditioning of the problem of minimization of $\mathcal{E}_z$ over a RKHS can be improved by regularization. As $\lim_{\gamma m\to\infty}(1 + \frac{cond(\mathcal{K}[x]-1)\lambda_{\min}}{\lambda_{\min}+\gamma m}) = 1$, for sufficiently large γm, the condition number of the matrix $\mathcal{K}[x] + \gamma m\mathcal{I}$ is close to 1. The size of γ is limited by requirements of fitting to the sample of empirical data z, while the size m of the sample can be enlarged. Thus for a sufficiently large m, regularization improves stability of the solution.

5 Discussion

Using theory of generalized inversion in Hilbert spaces, we have described solutions of the learning task modelled as the least square problem in the domain of reproducing kernel Hilbert spaces. Such spaces can be used to model radial-basis networks with various types of radial function with fixed width. Practical applications of formulas for computing pseudosolution and regularized solutions given in Theorems 3.1(v) and 4.2(ii) are limited by computational efficiency of iterative methods for solving systems of linear equations $c = \mathcal{K}[x]^+y$ and $c = (\mathcal{K}[x] + \gamma m\mathcal{I})^{-1}y$ and by the condition numbers of the matrices $\mathcal{K}[x]$ and $\mathcal{K}[x] + \gamma m\mathcal{I}$. We have shown how regularization improves properties of solutions of the learning task: it guarantees uniqueness and might improve stability when the size of the sample of data, their separation radius and the kernel defining the hypothesis space are properly chosen.

The requirement of continuity of the operator L_x do not allow to extend our results to the space of continuous functions on $\mathcal{R}^d$ with finite $\mathcal{L}_2$-norms. Another limiting factor is strong dependence of theory of pseudoinversion on Hilbert space setting. Thus most of our results apply only to empirical error with the square loss function, while, e.g., in the case of absolute value loss, only much weaker results holding for inverse problems with range in $\mathcal{R}^m$ with l_1-norm can be used.

References

[1] Aronszajn N. (1950). *Theory of reproducing kernels.* Transactions of AMS **68**, 33 – 404.

[2] Bertero M. (1989). *Linear inverse and ill-posed problems.* Advances in Electronics and Electron Physics **75**, 1 – 120.

[3] Cucker F., Smale S. (2001). *On the mathematical foundations of learning.* Bulletin of AMS **39**, 1 – 49.

[4] Girosi F. (1998). *An equivalence between sparse approximation and support vector machines.* Neural Computation **10**, 1455 – 1480.

[5] Groetch C. W. (1977). *Generalized inverses of linear operators.* Dekker, New York.

[6] Kůrková V. (2003). *High-dimensional approximation by neural networks.* Chapter 4 in, Advances in Learning Theory: Methods, Models and Applications, J. Stuykens et al., (ed.), 69 – 88. IOS Press, Amsterdam.

[7] Kůrková V. (2004). *Supervised learning as an inverse problem.* Research Report ICS-2004-906, Institute of Computer Science, Prague.

[8] Kůrková V., Sanguineti, M. (2004). *Error estimates for approximate optimization by the extended Ritz method.* SIAM Journal on Optimization (to appear).

[9] Kůrková V., Sanguineti M. (2003). *Learning with generalization capability by kernel methods with bounded complexity.* Research Report ICS-2003-901, submitted to Journal of Complexity.

[10] Narcowich F. J., Sivakumar N., Ward J. D. (1994). *On condition numbers associated with radial-function interpolation.* Journal of Mathematical Analysis and Applications **186**, 457 – 485.

[11] Parzen E. (1966). *An approach to time series analysis.* Annals of Math. Statistics **32**, 951 – 989.

[12] Poggio T., Girosi F. (1990). *Networks for approximation and learning.* Proceedings IEEE **78**, 1481 – 1497.

[13] Poggio T., Smale S. (2003). *The mathematics of learning: dealing with data.* Notices of the AMS **50**, 536 – 544.

[14] Tikhonov A. N., Arsenin V. Y. (1977). *Solutions of ill-posed problems.* W.H. Winston, Washington, D.C.

[15] Wahba G. (1990). *Splines models for observational data.* SIAM, Philadelphia.

Acknowledgement: This work was partially supported by GA ČR grant 201/02/0428. The author thanks to P. C. Kainen from Georgetown University and M. Sanguineti from Universita di Genova for fruitful comments and discussions.

Address: V. Kůrková, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic

E-mail: `vera@cs.cas.cz`

DATA AUGMENTATION ALGORITHM FOR GRAPHICAL MODELS WITH MISSING DATA

Masahiro Kuroda

Key words: Data augmentation algorithm, local computation, graphical models, computational efficiency, missing data.

COMPSTAT 2004 section: MCMC.

Abstract: In this paper, we discuss an efficient Bayesian computational method when observed data are incomplete in discrete graphical models. The data augmentation (DA) algorithm of Tanner and Wong [8] is applied to finding the posterior distribution. Utilizing the idea of local computation, it is possible to improve the DA algorithm. We propose a local computation DA (LC-DA) algorithm and evaluate its computational efficiency.

1 Introduction

A graphical model is characterized by conditional independence relationships among variables of a statistical model. Graphical models are broadly used in various fields to describe complex statistical models and to specify the multivariate distributions, see Whittaker [9] and Edwards [4].

For a large graphical model, it is rare to obtain complete observed data. When observed data are incomplete, it is extremely difficult to obtain the exact posterior distribution for a graphical model and the calculation may take a long time when observed data are moderately large. To overcome this computational difficulty, various algorithms related to graph structures are proposed, see Cowell et al. [1]. In this paper, we apply the data augmentation (DA) algorithm of Tanner and Wong [8] to approximating the posterior distribution of a graphical model. Then,

incorporating the idea of local computation into the DA algorithm, it is possible to reduce the computational effort. We propose a local computation DA (LC-DA) algorithm and evaluate its computational efficiency.

In Section 2, we show the exact Bayesian computation to find the posterior distribution for a discrete graphical model with missing data. In Section 3, instead of doing the infeasible exact computation, we give the DA algorithm to approximate the posterior distribution. In Section 4, we present the LC-DA algorithm. Section 5 discusses the efficiency of the LC-DA algorithm from the viewpoint of computational complexity.

2 Graphical model with missing data and exact Bayesian computation

Let V denote the set of vertices in a graph and $X_V = \{X_i \mid i \in V\}$ be the set of discrete random variables. Associated with each vertex $i \in V$, a random variable X_i takes values in a sample space Ω_i. For a subset $A \subseteq V$, we write X_A for $\{X_i \mid i \in A\}$ and $\Omega_A = \prod_{i \in A} \Omega_i$. Let the joint probability of X_V denote

$$p_V(x_V) = \Pr(X_V = x_V),$$

for every $x_V \in \Omega_V = \prod_{i \in V} \Omega_i$ and let $\theta_V = \{p_V(x_V) \mid x_V \in \Omega_V\}$. The marginal probability of X_A for $A \subset V$ can be written as

$$p_A(x_A) = \Pr(X_A = x_A) = \sum_{x_{V \setminus A}} p_V(x_V),$$

for every $x_A \in \Omega_A = \prod_{i \in A} \Omega_i$ and $\theta_A = \{p_A(x_A) \mid x_A \in \Omega_A\}$. The symbol "\" denotes the operator of a difference set. The conditional probability of X_A given $X_B = x_B$ is defined as

$$p_{A|B}(x_A|x_B) = \Pr(X_A = x_A \mid X_B = x_B) = p_{A \cup B}(x_{A \cup B})/p_B(x_B),$$

providing $A \cup B \subset V$ and $A \cap B = \emptyset$ where $\emptyset$ denotes the empty set, and also $\theta_{A|B} = \{p_{A|B}(x_A|x_B) \mid x_A \in \Omega_A, x_B \in \Omega_B\}$.

In this paper, we assume that the graph of X_V has the *global independence* that, for a triplet (A, B, C) of mutually disjoint subsets of V and $V = A \cup B \cup C$, each vertex of A is separated from each vertex of B given the subset C. Then, under the global independence structure, X_A is independent of X_B given X_C. Thus we have

$$p_V(x_V) = p_C(x_C) p_{A|C}(x_A|x_C) p_{B|C}(x_B|x_C),$$

so that $\{\theta_C, \theta_{A|C}, \theta_{B|C}\}$ are mutually independent. Suppose that observed data can be classified into three groups such that one is complete data and the others are incomplete data with X_B and X_A missing. The observed data patterns are indicated by $T = \{t_0, t_1, t_2\} = \{V, A \cup C, B \cup C\}$. In addition, we assume missingness at random in the sense of Rubin [7]. Each of observed data is denoted by $n^0 = \{n_{t_0}(x_{t_0}) \mid x_{t_0} \in \Omega_{t_0}\}$, $n^1 = \{n_{t_1}(x_{t_1}) \mid x_{t_1} \in \Omega_{t_1}\}$ and $n^2 = \{n_{t_2}(x_{t_2}) \mid x_{t_2} \in \Omega_{t_2}\}$. The sizes of the incomplete data n^1 and n^2 are considerably larger than the size of the complete data n^0. Assuming that observed data $n = (n^0, n^1, n^2)$ have a multinomial distribution with θ_V, the likelihood $L(n|\theta_V)$ is given by

$$\begin{aligned} L(n|\theta_V) &= f(n^0|\theta_{t_0}) f(n^1|\theta_{t_1}) f(n^2|\theta_{t_2}) \\ &\propto \prod_{0 \le i \le 2} \left\{ \prod_{x_{t_i} \in \Omega_{t_i}} p_{t_i}(x_{t_i})^{n_{t_i}(x_{t_i})} \right\}. \end{aligned} \tag{1}$$

For the multinomial model, we assume that the prior distribution of θ_V is a Dirichlet distribution which has the density function

$$\pi(\theta_V|\alpha_V) \propto \prod_{x_V \in \Omega_V} p_V(x_V)^{\alpha_V(x_V)-1}, \tag{2}$$

where $\alpha_V = \{\alpha_V(x_V) \mid x_V \in \Omega_V\}$ is a hyper-parameter. Then, according to the mutually independence relationships among $\{\theta_C, \theta_{A|C}, \theta_{B|C}\}$, it is possible to factorize $\pi(\theta_V|\alpha_V)$ into

$$\pi(\theta_V|\alpha_V) = \pi(\theta_C|\alpha_C)\pi(\theta_{A|C}|\alpha_{AC})\pi(\theta_{B|C}|\alpha_{BC}), \tag{3}$$

where $\alpha_C = \{\alpha_C(x_C) \mid x_C \in \Omega_C\}$, $\alpha_{AC} = \{\alpha_{AC}(x_{A\cup C}) \mid x_{A\cup C} \in \Omega_{A\cup C}\}$ and $\alpha_{BC} = \{\alpha_{BC}(x_{B\cup C}) \mid x_{B\cup C} \in \Omega_{B\cup C}\}$. The Dirichlet prior distribution (3) describes conditional independence of prior distributions and is called hyper Dirichlet prior distribution by Dawid and Lauritzen [3].

From the equations (1) and (2), we can obtain the mixture posterior distribution with the density

$$\begin{aligned} &\pi(\theta_V \mid n) \\ &\quad\propto \quad L(n|\theta_V) \times \pi(\theta_V \mid \alpha_V) \\ &\quad\propto \quad \sum_{\Omega(n^1)} \binom{n_{t_1}(x_{t_1})}{\{\tilde{n}_{t_1}(x_V)\}} \sum_{\Omega(n^2)} \binom{n_{t_2}(x_{t_2})}{\{\tilde{n}_{t_2}(x_V)\}} \prod_{x_V \in \Omega_V} p_V(x_V)^{\tilde{\alpha}_V(x_V)-1}, \end{aligned} \tag{4}$$

where, for $i = 1, 2$,

$$\sum_{\Omega(n^i)} \binom{n_{t_i}(x_{t_i})}{\{\tilde{n}_{t_i}(x_V)\}} = \prod_{x_{t_i} \in \Omega_{t_i}} \sum_{\Omega(n_{t_i}(x_{t_i}))} \binom{n_{t_i}(x_{t_i})}{\{\tilde{n}_{t_i}(x_V)\}}$$

and $\sum_{\Omega(n_{t_i}(x_{t_i}))}$ denotes the sum over all possible $\tilde{n}_{t_i}(x_V)$ for all $x_V \in \Omega_V$ under the conditions $\tilde{n}_{t_i}(x_V) \geq 0$ and $\sum_{x_{V\setminus t_i}} \tilde{n}_{t_i}(x_V) = n_{t_i}(x_{t_i})$, and

$$\tilde{\alpha}_V(x_V) = \alpha_V(x_V) + n_{t_0}(x_V) + \tilde{n}_{t_1}(x_V) + \tilde{n}_{t_2}(x_V).$$

Because of combinational explosion, the posterior density (4) has a very complicated function. Therefore, it is extremely difficult to calculate exactly $\pi(\theta_V|n)$ and these computation may take a long time when the observed data are moderately large.

Instead of performing the infeasible Bayesian computation, we use the DA algorithm which imputes incomplete data and finds $\pi(\theta_V|n)$ using the Monte Carlo method.

3 DA algorithm to graphical models

The DA algorithm is a type of Markov chain Monte Carlo. In the case that the incomplete-data posterior density is complicated as the posterior distribution (4) and the complete-data posterior is relative easy to handle and draw from, the DA algorithm is very suitable. Each iteration of the DA algorithm consists of an *Imputation*-step and a *Posterior*-step.

For this case, the exact posterior distribution $\pi(\theta_C|n)$ can be obtained without any iterations, since the complete marginal data for X_C are calculated from n. Then the DA algorithm for the graphical model is given by the following iterative scheme:

Initialization: Set an initial distribution $\pi^{(0)}(\theta_V|n) = \pi(\theta_V|\alpha_V)$.

Imputation-step: Repeat the following steps for $l = 1, \dots, L$ to obtain the augmented data $\tilde{n} = (n^0, \tilde{n}^1, \tilde{n}^2)$, where

$$\begin{aligned}
\tilde{n}^1 &= \{\tilde{n}_{t_1}(x_V) \mid x_V \in \Omega_V, \sum_{x_B \in \Omega_B} \tilde{n}_{t_1}(x_V) = n_{t_1}(x_{t_1}), \tilde{n}_{t_1}(x_V) \geq 0\}, \\
\tilde{n}^2 &= \{\tilde{n}_{t_2}(x_V) \mid x_V \in \Omega_V, \sum_{x_A \in \Omega_A} \tilde{n}_{t_2}(x_V) = n_{t_2}(x_{t_2}), \tilde{n}_{t_2}(x_V) \geq 0\}.
\end{aligned}$$

1. Generate θ_V^* from the current approximation $\pi^{(t-1)}(\theta_V|n)$.
2. Generate $\tilde{n}^{1(l)}$ and $\tilde{n}^{2(l)}$ from the predictive multinomial distributions $f(\tilde{n}^1|\theta^*_{B|t_1}, n^1)$ and $f(\tilde{n}^2|\theta^*_{A|t_2}, n^2)$, where

$$\begin{aligned}
\theta^*_{B|t_1} &= \{p^*_{B|t_1}(x_B|x_{t_1}) \mid x_B \in \Omega_B, x_{t_1} \in \Omega_{t_1}\}, \\
\theta^*_{A|t_2} &= \{p^*_{A|t_2}(x_A|x_{t_2}) \mid x_A \in \Omega_A, x_{t_2} \in \Omega_{t_2}\}.
\end{aligned}$$

Posterior-step: Update $\pi^{(t)}(\theta_V|n)$ given $\{\tilde{n}^{(l)} \mid 1 \leq l \leq L\}$ using the Monte Carlo method:

$$\pi^{(t)}(\theta_V|n) = \frac{1}{L}\sum_{l=1}^{L} \pi(\theta_C|n)\pi(\theta_{A|C}|\tilde{n}^{(l)})\pi(\theta_{B|C}|\tilde{n}^{(l)}).$$

Then

$$\begin{aligned}
\pi(\theta_{A|C}|\tilde{n}^{(l)}) &\propto \prod_{x_C \in \Omega_C} \prod_{x_A \in \Omega_A} p_{A|C}(x_A|x_C)^{\tilde{\alpha}^{(l)}_{AC}(x_{A\cup C})-1}, \\
\pi(\theta_{B|C}|\tilde{n}^{(l)}) &\propto \prod_{x_C \in \Omega_C} \prod_{x_B \in \Omega_B} p_{B|C}(x_B|x_C)^{\tilde{\alpha}^{(l)}_{BC}(x_{B\cup C})-1},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\alpha}^{(l)}_{AC}(x_{A\cup C}) &= \alpha_{AC}(x_{A\cup C}) + n_{t_0}(x_{A\cup C}) + n_{t_1}(x_{t_1}) + \sum_{x_B \in \Omega_B} \tilde{n}^{(l)}_{t_2}(x_V), \\
\tilde{\alpha}^{(l)}_{BC}(x_{B\cup C}) &= \alpha_{BC}(x_{B\cup C}) + n_{t_0}(x_{B\cup C}) + n_{t_2}(x_{t_2}) + \sum_{x_A \in \Omega_A} \tilde{n}^{(l)}_{t_1}(x_V).
\end{aligned}$$

Until the approximations $\pi^{(t)}(\theta_{A|C}|n)$ and $\pi^{(t)}(\theta_{B|C}|n)$ converge to stationary distributions, the *Imputation*- and *Posterior*-steps are alternated repeatedly. Achieving convergence of the DA algorithm, the true posterior distribution $\pi(\theta_V|n)$ can be found.

In the practical implementation of the DA algorithm, the selection of the number of imputation (L) is perhaps more crucial. When the proportion of missing data in incomplete data is high and the size of the incomplete data is large, L must be considerably large. However, it is difficult to determine L on theoretical bases. In order to assess the convergence, diagnostic techniques are applied to output from the DA iteration. Cowles and Carlin [2] provide the comparative reviews of many convergence diagnostic techniques.

4 Application of LC-DA algorithm

In this section, we present the LC-DA algorithm. The important property of the LC-DA algorithm is that the DA algorithm is applied to each of factorized posterior distributions according to a graph structure and each posterior distribution is computed independently. Then it is possible to reduce the computational efforts from the viewpoint of computational complexity.

We denote the marginal data for $X_{A\cup C}$ and $X_{B\cup C}$ as $n_{AC} = (n^0_{AC}, n^1, n^2_C)$ and $n_{BC} = (n^0_{BC}, n^1_C, n^2)$, where

$$\begin{aligned} n^0_{AC} &= \{n_{t_0}(x_{A\cup C}) \mid x_{A\cup C} \in \Omega_{A\cup C}\},\ n^0_{BC} = \{n_{t_0}(x_{B\cup C}) \mid x_{B\cup C} \in \Omega_{B\cup C}\}, \\ n^1_C &= \{n_{t_1}(x_C) \mid x_C \in \Omega_C\},\quad n^2_C = \{n_{t_2}(x_C) \mid x_C \in \Omega_C\}. \end{aligned}$$

With local computation to find $\pi(\theta_V|n)$, we can obtain the following theorem.

Theorem 4.1. *Suppose that C separates A and B in V. If $C \subseteq t$ for all $t \in T$, then the calculation of the posterior distributions of $\theta_{A|C}$ and $\theta_{B|C}$ can be done independently.*

Theorem 1 guarantees that the DA algorithm can execute separately to obtain the posterior distributions of $\theta_{A|C}$ and $\theta_{B|C}$ given n_{AC} and n_{BC}. The condition of $C \subseteq t$ for all $t \in T$ is called "lossless decomposition" by Geng and Li [5]. Then the LC-DA algorithm realizes the computation according to the following iterative scheme:

The DA iteration of $\pi(\theta_{A|C}|n_{AC})$

Initialization: Set an initial distribution $\pi^{(0)}(\theta_{A|C}|n_{AC}) = \pi(\theta_{A|C}|\alpha_{AC})$.

Imputation-step: Repeat the following steps for $l = 1, \ldots, L$ to impute $\tilde{n}^2_C$, where

$$\begin{aligned} \tilde{n}^2_C &= \{\tilde{n}_{t_2}(x_{A\cup C}) \mid x_{A\cup C} \in \Omega_{A\cup C}, \sum_{x_A} \tilde{n}_{t_2}(x_{A\cup C}) = n_{t_2}(x_C), \\ &\qquad \tilde{n}_{t_2}(x_{A\cup C}) \geq 0\}. \end{aligned}$$

1. Generate $\theta^*_{A|C}$ from the current approximation $\pi^{(t-1)}(\theta_{A|C}|n_{AC})$.
2. Generate the imputed data $\tilde{n}_C^{2(l)}$ from the predictive multinomial distribution $f(\tilde{n}_C^2|\theta^*_{A|C}, n_C^2)$.

Posterior-step: Update $\pi^{(t-1)}(\theta_{A|C}|n_{AC})$ by the Monte Carlo method:

$$\pi^{(t)}(\theta_{A|C}|n_{AC}) = \frac{1}{L}\sum_{l=1}^{L} \pi(\theta_{A|C} \mid \tilde{\alpha}_{AC}^{(l)}),$$

where

$$\tilde{\alpha}_{AC}^{(l)}(x_{A\cup C}) = \alpha_{AC}(x_{A\cup C}) + n_{t_0}(x_{A\cup C}) + n_{t_1}(x_{A\cup C}) + \tilde{n}_{t_2}^{(l)}(x_{A\cup C}).$$

The DA iteration of $\pi(\theta_{B|C}|n_{BC})$

The DA algorithm to obtain $\pi(\theta_{B|C}|n_{BC})$ are similar to the DA iteration of $\pi(\theta_{A|C}|n_{AC})$: The *Imputation*-step generates $\{\tilde{n}_C^{1(l)} \mid 1 \le l \le L\}$, where $\tilde{n}_C^1 = \{\tilde{n}_{t_1}(x_{B\cup C}) \mid x_{B\cup C} \in \Omega_{B\cup C}, \sum_{x_B} \tilde{n}_{t_1}(x_{B\cup C}) = n_{t_1}(x_C), \tilde{n}_{t_1}(x_{B\cup C}) \ge 0\}$.The *Posterior*-step finds $\pi^{(t)}(\theta_{B|C}|n_{BC})$ using n_{BC}^0, n^2 and $\{\tilde{n}_C^{1(l)} \mid 1 \le l \le L\}$.

When each of the approximations $\pi^{(t)}(\theta_{A|C}|n_{AC})$ and $\pi^{(t)}(\theta_{B|C}|n_{BC})$ converges to a stationary distribution, the true posterior distribution $\pi(\theta_V|n)$ can be calculated.

5 Computational efficiency of LC-DA algorithm

We now evaluate the computational efficiency of the LC-DA algorithm from the viewpoint of computational complexity. Here we introduce two quantities:

- $||\Omega_V||$ = the number of all possible values in Ω_V
- $||\Omega_A||$ = the number of all possible values in Ω_A where $A \subset V$

As for the space complexity, the amount of the storage space required by the DA algorithm is $||\Omega_V||$. Alternatively, in the LC-DA algorithm, it can not exceed $\max(||\Omega_{A\cup C}||, ||\Omega_{B\cup C}||)$. Next consider the time complexity under the worst-case assumption. The time complexity of the DA algorithm can be expressed by $O(||\Omega_V||)$. The implementation of the LC-DA algorithm can be done in $O(\max(||\Omega_{A\cup C}||, ||\Omega_{B\cup C}||))$.

The LC-DA algorithm is more efficient than the DA algorithm from both aspects of the space and time complexities and then can reduce the computational efforts.

Finally, we briefly describe the convergence speed of the LC-DA algorithm. The LC-DA algorithm is regarded as the collapsed Gibbs sampler of Liu [6]. Then, according to Liu's [6] result, the convergence speed of the LC-DA algorithm is faster than the speed of the DA algorithm. We shall investigate its convergence speed in detail.

Appendix

Proof of Theorem 1

Since $C \subseteq t$ for all $t \in T$, we have $t \cap C = C$ and

$$
\begin{aligned}
&\prod_{x_t\in\Omega_t} p_t(x_t)^{n_t(x_t)} \\
&= \left\{\prod_{x_C\in\Omega_C} p_C(x_C)^{n_t(x_t)}\right\} \\
&\quad \times \left\{\prod_{x_{t\cap(A\cup C)}\in\Omega_{t\cap(A\cup C)}} \sum_{\Omega(n_t(x_{t\cap(A\cup C)}))} \binom{n_t(x_{t\cap(A\cup C)})}{\{\tilde{n}_t(x_{A\cup C})\}} \right. \\
&\quad \left. \prod_{x_C\in\Omega_C}\prod_{x_A\in\Omega_A} p_{A|C}(x_A|x_C)^{\tilde{n}_t(x_{A\cup C})}\right\} \\
&\quad \times \left\{\prod_{x_{t\cap(B\cup C)}\in\Omega_{t\cap(B\cup C)}} \sum_{\Omega(n_t(x_{t\cap(B\cup C)}))} \binom{n_t(x_{t\cap(B\cup C)})}{\{\tilde{n}_t(x_{B\cup C})\}} \right. \\
&\quad \left. \prod_{x_C\in\Omega_C}\prod_{x_B\in\Omega_B} p_{B|C}(x_B|x_C)^{\tilde{n}_t(x_{B\cup C})}\right\} \\
&= \left\{\prod_{x_C\in\Omega_C} p_C(x_C)^{n_t(x_t)}\right\} \\
&\quad \times \left\{\sum_{\Omega_{t\cap(A\cup C)}(n_t)} \binom{n_t(x_{t\cap(A\cup C)})}{\{\tilde{n}_t(x_{A\cup C})\}} \prod_{x_C\in\Omega_C}\prod_{x_A\in\Omega_A} p_{A|C}(x_A|x_C)^{\tilde{n}_t(x_{A\cup C})}\right\} \\
&\quad \times \left\{\sum_{\Omega_{t\cap(B\cup C)}(n_t)} \binom{n_t(x_{t\cap(B\cup C)})}{\{\tilde{n}_t(x_{B\cup C})\}} \prod_{x_C\in\Omega_C}\prod_{x_B\in\Omega_B} p_{B|C}(x_B|x_C)^{\tilde{n}_t(x_{B\cup C})}\right\}.
\end{aligned}
$$

Then for any $s \subset V$, $t \cap s = s$, it holds $n_t(x_s) = \tilde{n}_t(x_s)$ and

$$\sum_{\Omega_{t\cap s}(n_t)} \binom{n_t(x_{t\cap s})}{\{\tilde{n}_t(x_s)\}} = 1.$$

Therefore it is possible to factorize the likelihood (1) as follows:

$$L(\theta_V|n) = L(\theta_C|n_C)L(\theta_{A|C}|n_{AC})L(\theta_{B|C}|n_{BC}). \tag{5}$$

From the prior distribution (3) and the likelihood (5), we can obtain the posterior distribution

$$
\begin{aligned}
\pi(\theta_V|n) &\propto \{L(\theta_C|n_C)\pi(\theta_C|\alpha_C)\} \times \left\{L(\theta_{A|C}|n_{AC})\pi(\theta_{A|C}|\alpha_{AC})\right\} \\
&\quad \times \left\{L(\theta_{B|C}|n_{BC})\pi(\theta_{B|C}|\alpha_{BC})\right\} \\
&= \pi(\theta_C|n_C)\pi(\theta_{A|C}|n_{AC})\pi(\theta_{B|C}|n_{BC}).
\end{aligned}
$$

Since it also holds the mutual independence among the posterior distributions, we can compute each posterior distribution independently. ∎

References

[1] Cowell R.G., Dawid A.P., Lauritzen S.L., Spiegelhalter D.J. (1999). *Probabilistic networks and expert systems.* Springer-Verlag, New York.

[2] Cowles M.K., Carlin D.B. (1996). *Markov chain Monte Carlo convergence diagnostics: A comparative review.* Journal of the American Statistical Association **91**, 883 – 904.

[3] Dawid A.P., Lauritzen S.L. (1993). *Hyper Dirichlet laws in the statistical analysis of decomposable graphical models.* The Annals of Statistics **21**, 1272 – 1317.

[4] Edwards D. (2000). *Introduction to graphical modeling.* Second edition. Springer-Verlag, New York.

[5] Geng Z., Li K. (2003). *Factorization of posteriors and partial imputation algorithm for graphical models with missing data.* Statistics & Probability Letters **64**, 369 – 379.

[6] Liu J. (1994). *The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem.* Journal of the American Statistical Association **89**, 958 – 966.

[7] Rubin D.B. (1976). *Inference and missing data.* Biometrika **63**, 581 – 592.

[8] Tanner M.A., Wong W.H. (1987). *The calculation of posterior distributions by data augmentation.* Journal of the American Statistical Association **82**, 528 – 540.

[9] Whittaker J. (1990). *Graphical models in applied multivariate statistics.* John Wiley & Sons, Chichester.

Acknowledgement: The author would like to thank the editor and an anonymous referee for providing valuable comments on the paper.

Address: M. Kuroda, Department of Socio-Information, Okayama University of Science, 1-1 Ridai-cho, Okayama 700-0005, Japan

E-mail: `kuroda@soci.ous.ac.jp`

PRINCIPAL VARIABLE ANALYSIS

Aziz Lazraq and Robert Cléroux

Key words: Principal component analysis, principal variable analysis .

COMPSTAT 2004 section: Multivariate analysis.

Abstract: In this paper we discuss an inferential method or PVA and compare it with two methods (non inferential) that exist in the literature.

1 Introduction

Principal component analysis (PCA) is a statistical method having well known optimal properties. It reduces the size of a problem by replacing a vector of variables $Y = (Y_1, Y_2, \ldots, Y_p)'$ by a vector of principal components $X = (X_1, X_2, \ldots, X_Q)'$ with $q \ll p$ where each X_i is a linear function of all the Y_j's. PCA is useful in practical situations when it is possible to give a meaning to the principal components. However this is not always possible and one naturally looks for other methods of size reduction. The existence, in most PCA computer programs of several types of axis rotation emphasis this point.

The problem we consider in this paper is the following: reduce the size of the problem by replacing the vector of variables $Y = (Y_1, Y_2, \ldots, Y)_p)'$ by a sub-vector of Y denoted by $X : q \times 1$ with $q \ll p$ without loosing too much information in the sense of total variance of Y. The variables of X are called principal variables by McCabe [5]. Of course, since each principal variable is a particular (trivial) linear function of $Y_1, Y_2, \ldots, Y_p$ it is clear that principal components will be better than principal variables. If, say, only two principal components are sufficient to explain well the vector Y, more than two principal variables will be necessary to do as well. But it may turn out that only 3 or 4 principal variables are needed. And given the problem of interpreting the principal components, on may want to chooses, in such circumstances, to perform a principal variable analysis (PVA) in place of a PCA.

2 The problem

Let $Y : p \times 1$ be a multivariate normal random vector with positive definite covariance matrix Σ and assume for the moment that $\Sigma <$ is known. We look for a sub-vector X of Y which can explain Y well with respect to a given criterion. McCabe [5] suggests several criteria:

(i) maximize $\sum_{i=1}^{m} \rho_i^2$, where Σ_{XX} is the covariance matrix of the selected variables, $\Sigma_{YY \cdot X}$ is the conditional covariance matrix of the variables not selected given those selected, the ρ_i''s are the canonical correlations between the variables selected ans those not selected. McCabe [5] gives an algorithm (not inferential) for criterion (ii) which consists of completely all the possible variable selections.

(ii) maximize $|\Sigma_{XX}|$ or minimize $|\Sigma_{YY\cdot X}|$,

(iii) minimize $\mathrm{tr}(\Sigma_{YY\cdot X}|$

(iv) minimize $\parallel \Sigma_{YY\cdot X} \parallel^2$ and

Here we deal with the second criterion. It can also be written as maximize $\sum_{i=1}^{p} \sigma_{ii} R^2\ (Y_i, X)$ where $\sigma_{ii} = \mathrm{var}(Y_i)$, the i^{th} component of Y, and where $R^2(Y_i, X)$ is the multiple correlation coefficient squared between Y_i and the selected variables (see also Okamoto [6]). This last form of criterion (iii) is the numerator of the redundancy index

$$\rho I = \frac{\sum_{i=1}^{P} \sigma_{ii} R^2(Y_i, X)}{\sum_{i=1}^{p} \sigma_{ii}} = \frac{\mathrm{tr}(\Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY})}{\mathrm{tr}(\Sigma_{YY})}$$

where $\Sigma_{YX} = \Sigma'_{XY}$ is the covariance matrix between Y and X and $\Sigma_{YY} = \Sigma$ is the covariance matrix of Y. Thus criterion (iii) is equivalent to maximizing ρI since its denominator is constant. When Σ is unknown we work with its classical estimator S

$$\hat{\rho I} = RI = \frac{\mathrm{tr}(S_{YX} S_{XX}^{-1} S_{XY})}{\mathrm{tr}(S_{YY})}$$

which is the fraction of the total variance of Y explained by the multivariate regression of Y on X, the selected variables.

In Lazraq and Cléroux [4] inferential algorithms (forward, backword and stopwise) have been given for the problem of multivariate linear regression which is equivalent to maximizing RI. We use the variable selection algorithms for the regression of Y on X (a subset of Y) and find the significant subset X of principal variables.

Another principal variable selection method (non inferential) has been used in the literature Gonzalez et al. [2] used a best subset selection type algorithm with maximizing

$$RV_{reg} = \Big(\frac{\mathrm{tr}(S_{YX} S_{XX}^{-1} S_{XY})^2}{\mathrm{tr}(S_{YY})^2}\Big)^{1/2}.$$

RV_{reg} has been introduced by Robert and Escoufier (1976) and is also related to multivariate linear regression of Y or X.

3 Examples

Example 1 Consider the Iris Versicolor data with $n = 50$ and $p = 4$ (se a principal component analysis on the S matrix yields the following results: the first PC explains 78.1% of the total variance, the first two explain 89.7% and the first three 98.4%.

McCabe [5] using criteria (ii) with complete enumeration of subsets and multivariate regressions to compute the fraction of total variance explained gets the following : variable Y_1 explains 69.0% of total variance, Y_1 and Y_3 explain 87.3% and variables Y_1, Y_2, Y_3 explain 98.2%. Using Lazraq and Cléroux (1988) algorithms (the variables have been transformed to have means zero) we obtain the result of Table 1. We see that only two principal variables, Y_1 and Y_3, are significant, that they explain 87.2% of the total variance of Y and that they perform almost as well as the first two PC.

Forward			Backward			Stepwise		
var in	RI	p-value	var out	RI	p-value	var	RI	p-value
1	.690	0	4	.982	1	1 in	.690	0
3	.873	0	2	.873	.18	3 in	.873	0
2	.982	.18	3	.690	0			
4	1	1	1	0	0			

Table 1: PVA of Iris Versicolor data.

Example 2 Consider the tobacco data (the last six chemical data only) with $n = 25$ and $p = 6$ (see Anderson and Baucroft [1, p. 205]). A principal component analysis on the S matrix yields : the first PC explains 55.0% of total variance, the first two explain 87.2% and the first three 97.2%.

The results of the algorithms of Lazraq and Cléroux [4] on centered data are shown in Table 3. It is seen that three principal variables are significant at the 10% level, that they explain 94.8% of the total variance of Y and that they perform almost as well as the first three principal components. We see also that four principal variables are more efficient than three principal components and much easier to interpret.

Forward			Backward			Stepwise		
var in	RI	p-value	var out	RI	p-value	var	RI	p-value
2	.545	0	4	.999	1	2 in	.545	0
5	.861	0	6	.993	.21	5 in	.861	0
1	.948	.09	3	.948	.16	1 in	.948	0.9
3	.993	.16	1	.861	.09			
6	.999	.21	5	.545	0			
4	1	1	2	0	0			

Table 2: PVA of Tobacco data.

Example 3 Consider the country data with $n = 49$ and $p = 7$ (see Gunst and Mason [3]). Gonzalez et al [2] using best subset selection and maximizing the criterion RV_{reg} obtained $BSS(1) = \{6\}$, $BSS(2) = \{4, 6\}$, $BSS(3) = \{1, 4, 6\}$ and $BSS(4) = \{2, 5, 6, 7\}$ where BSS(j) is the best subset with j variables.

The results of Lazraq and Cléroux are shown in Table 3 for the standardized data. It is seen that only three principal variables, 4, 6 and 7, are significant. Here again, like in regression in general, the variables selected by a best subset selection procedure and by a stepwise procedure do not always coincide.

Forward			Stepwise		
var in	RI	p-value	var out	RI	p-value
6	.382	0	6 in	.382	0
4	.652	0	4 in	.652	0
7	.798	0	7 in	.798	0
2	.879	.18			
1	.939	.51			
3	.992	.16			
5	1	1			

Table 3: PVA of country data.

4 Conclusion

In this paper we discussed principal variable analysis (AVA). The method of AVA can give results that are comparable to those of PCA. Since PCA explains, for any given size of X, the maximum total variance of Y, AVA cannot be optimal. However we have seen in some examples that it can be near optimal and that it avoids the problem of giving a meaning to principal components.

References

[1] Anderson R.L., Bancroft, T.A. (1952). *Statistical Theory in Research.* McGraw Hill, New York.

[2] Gonzalez P.L., Cléroux R., Rioux B. (1990). *Selecting the Best Subset of Variables in Principal Component Analysis.* Physica-Verlag Heidelberg for IASC, 115–120.

[3] Gunst R.F., Mason R.L. (1980), *Regression analysis and its applications.* Marcel Dekker Inc., New York.

[4] Lazraq A., Cléroux R. (1988). *Un algorithme pas à pas de sélection de variables en régression linéaire multivarié.* Stat. Anal. Donn. **13**, 39 – 58.

[5] McCabe G.P. (1984). *Principal variables.* Technometrics **26**, 137 – 144.

[6] Okamoto M. (1969). *Optimality of principal components.* In Multivariate Analysis III, ed. P.R. Krishuoiah, Academic Press, New York, 673 – 685.

Address: A. Lazraq, l'Ecole Nationale de l'Industrie Minérale, Rabat
R. Cléroux, Dépt. de mathématique et de statistique, Univ. de Montréal

E-mail: `lazraq@enim.ac.ma, cleroux@dms.umontreal.ca`

COMPSTAT'2004 Symposium

GENEGOBI: VISUAL DATA ANALYSIS TOOLS FOR MICROARRAY DATA

Eun-Kyung Lee, Dianne Cook, Heike Hofmann, Eve Wurtele, Dongshin Kim, Jihong Kim and Hogeun An

Key words: Data visualization, gene expression data, GGobi, metabolic networks, multivariate data analysis.

COMPSTAT 2004 section: Statistical software.

Abstract: GeneGobi is software for the exploratory analysis of microarray data and metabolic networks. It helps biologists analyze the connections between microarray data and metabolic pathways interactively. GeneGobi provides a "user-friendly" interface for biologists working on microarrays and metabolic networks to the analysis and graphical tools available R and GGobi.

1 Introduction

Microarray experiments generate huge data sets. Current software provides only limited ways to look at microarray data, and there is no available software that combines the visualization and analysis of metabolic networks. GeneGobi is developed for exploratory analysis of microarray data and metabolic networks. It helps biologists explore patterns in gene expression data in association with the regulatory and metabolic pathways. This software is originally designed for analysis of data collected on the plant Arabidopsis and to analyze the connections between microarray data and the metabolic pathways of Arabidopsis. However it can be used generally for gene expression data analysis connected to network diagrams.

2 GeneGobi

GeneGobi is a visual data analysis tool for microarray data and metabolic networks. It is based on R and GGobi. R is open source statistical analysis software and there are many contributed statistical analysis packages in R. Also packages for gene expression data analysis are available (e.g. Bioconductor).

GGobi is an interactive and dynamic data visualization system for multivariate data. It is able to be used from R using the RGGobi and RGtk packages. GGobi allows the user to explore multivariate data using bar charts, scatterplots, 3D rotating plots and higher dimensional rotations (unique to GGobi), and profile plots. The user can interact with each plot by brushing points and lines using different symbols and line types, and these actions simultaneously change elements of other plots as appropriate. The linking between plots is fast and can be sophisticated, using selected columns in the data as the key to graph elements in different plots. This is different from virtually all other packages. Elements of the plots can also be identified with user-provided labels, or by selected columns in the data.

GGobi can also display metabolic networks. Users can read in a layout from another package such as FCModeler (`http://www.eng.iastate.edu/~julied/research/fcmodeler`), and interact with the network by brushing nodes and edges and identifying nodes or edges. Elements of the network can be linked to other types of data displayed in other plots. For example, where we have identified the LocusID of the gene responsible for a particular entity in the network this provides a key to link the node to gene expression data.

There are a couple of drawbacks in R and GGobi. To analyze gene expression data or develop new methods, we need to write code in R and it is not easy for a naive user. Another one comes from huge data. Usually gene expression data has thousands of genes. In stand-alone ggobi with thousands of genes, it is not easy to brush or identify a few specific genes. This can be controlled with RGGobi.

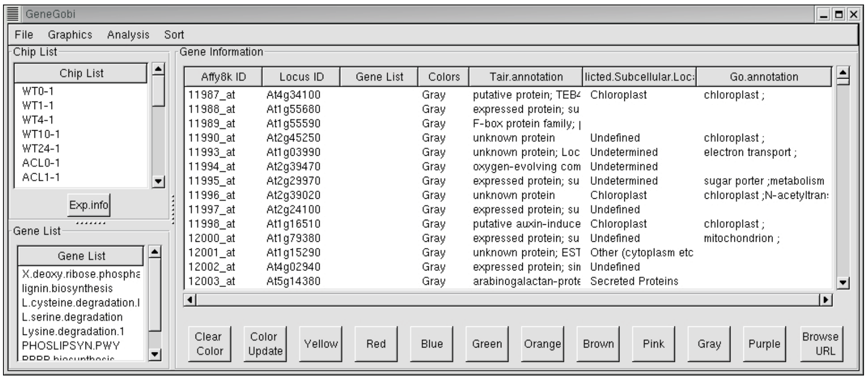
The main purpose of GeneGobi is to help biologist explore patterns in gene expression data and the regulatory and metabolic pathway. It helps analyze the connections between microarray data and metabolic pathway visually and interactively. It also can combine statistical analysis results with interactive plots for more sophisticated analysis.

GeneGobi provides a user-interface to both GGobi and R.It adds a spreadsheet to more information about each gene, links to literature from the Web, menus of analysis and visualization options and an interface to lists of selected genes created from expert knowledge or previous analyses. We can brush selected genes from this spread sheet using color buttons.

We give an overview of GeneGobi by summarizing main GUI, menu bar and buttons. For demonstration, we use gene expression data from an experiment using Arabidopsis. This experiment uses two different genotype (WT and ACL) at 5 different time points. We also considered a hypothetical jasmonic acid network to show how gene expression data analysis and network analysis are combined with GeneGobi.

2.1 Main GUI

Main GeneGobi GUI has a lot of features. **Gene Information** provides a lot of information about genes. It needs to be provided by the user. In this **Gene Information**,the user can select genes that she wants to analyze. **Chip List** provides a list of experiments. We can choose interesting experiments for future analysis. **Exp.info** button shows the experimental information, such as genotype, time, treatments, etc. for each experiment. This information is used for deciding on useful comparisons and guiding the statistical analysis. **Gene List** shows a list of clusters that can be known from biological sources or constructed by the user. When one list is selected, genes on this list are changed to “Red” and the other genes are changed to “Gray” in GGobi plot.



Chip List

WT0-1
WT1-1
WT4-1
WT10-1
WT24-1
ACL0-1
ACL1-1

Gene List

X.deoxy.ribose.phospha
lignin.biosynthesis
L.cysteine.degradation.I
L.serine.degradation
Lysine.degradation.1
PHOSLIPSYN.PWY
PPPP biosynthesis

Affy8k ID	Locus ID	Gene List	Colors	Tair.annotation	licted.Subcellular.Loc:	Go.annotation
11987_at	At4g34100		Gray	putative protein; TEB4	Chloroplast	chloroplast ;
11988_at	At1g55680		Gray	expressed protein; su		
11989_at	At1g55590		Gray	F-box protein family;		
11990_at	At2g45250		Gray	unknown protein	Undefined	chloroplast ;
11993_at	At1g03990		Gray	unknown protein; Loc	Undetermined	electron transport ;
11994_at	At2g39470		Gray	oxygen-evolving com	Undetermined	
11995_at	At2g29970		Gray	expressed protein; su	Undetermined	sugar porter ;metabolism
11996_at	At2g39020		Gray	unknown protein	Chloroplast	chloroplast ;N-acetyltran:
11997_at	At2g24100		Gray	expressed protein; su	Undefined	
11998_at	At1g16510		Gray	putative auxin-induce	Chloroplast	chloroplast ;
12000_at	At1g79380		Gray	expressed protein; su	Undefined	mitochondrion ;
12001_at	At1g15290		Gray	unknown protein; EST	Other (cytoplasm etc	
12002_at	At4g02940		Gray	expressed protein; sin	Undefined	
12003_at	At5g14380		Gray	arabinogalactan-prote	Secreted Proteins	

Figure 1: Main Window of GeneGobi: The menu bar is at the top. The chip List(experiment name) and Gene List(user defined) is in the left side. All the information about genes are in the right side. There are also color control buttons at the bottom of the right side.

In the Gene Information spreadsheet, genes in this cluster are shifted to the top list and the name of this list is shown in the Gene List column.

2.2 Menu bar

2.2.1 File

- Read Files: read files that are needed for data analysis
  - Gene Expression: read the gene expression data. It should be the normalized data in text mode (tab delimited or comma separated).
  - Gene Information: read the gene information file that contains affy id, locus id, tair.annotation, go.annotation, etc. At least, you need to have the locus id.
  - Network: uses SBML (System Biologist Markup Language)
  - Gene List: read the user defined lists. It should have list name and names of the genes in text mode.
  - Experimental Information: read the experimental design file. It contains genotype, replicate, time, treatments, etc.
- Exit: quit GeneGobi

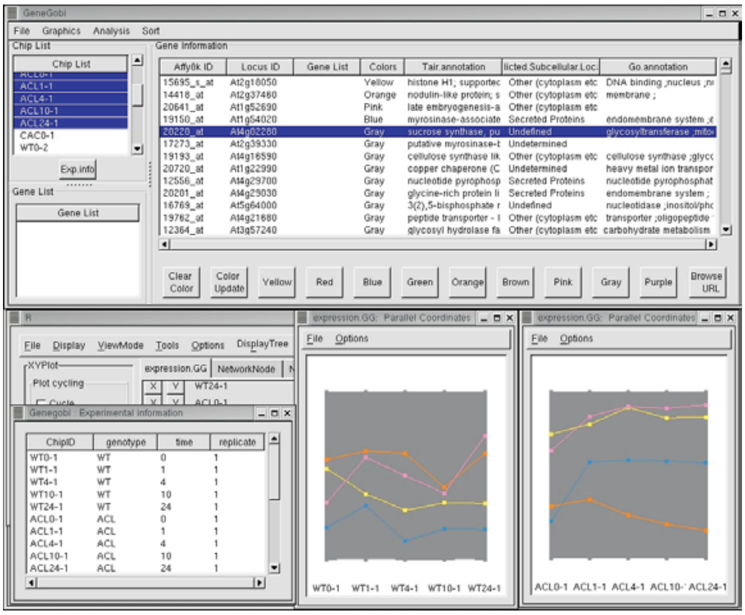
2.2.2 Graphics

- Open GGobi: Before you choose this menu item, you need to have loaded the data (at least the gene expression and the gene information data).

  - with Gene Expression: Only if you want to analyze gene expression data alone.
  - with Gene Expression + Network: if you have some metabolic network information along with the expression data.

2.2.3 Analysis

- Compare Treatments: You need more than 1 chip to use this tool.
  - Correlation: Calculate the correlation between two genotypes. In the Gene Information spreadsheet, genes are sorted by this correlation in descending order.
  - Covariance: Calculate the covariance between two genotypes. In the Gene Information spreadsheet, genes are sorted by this covariance in descending order.



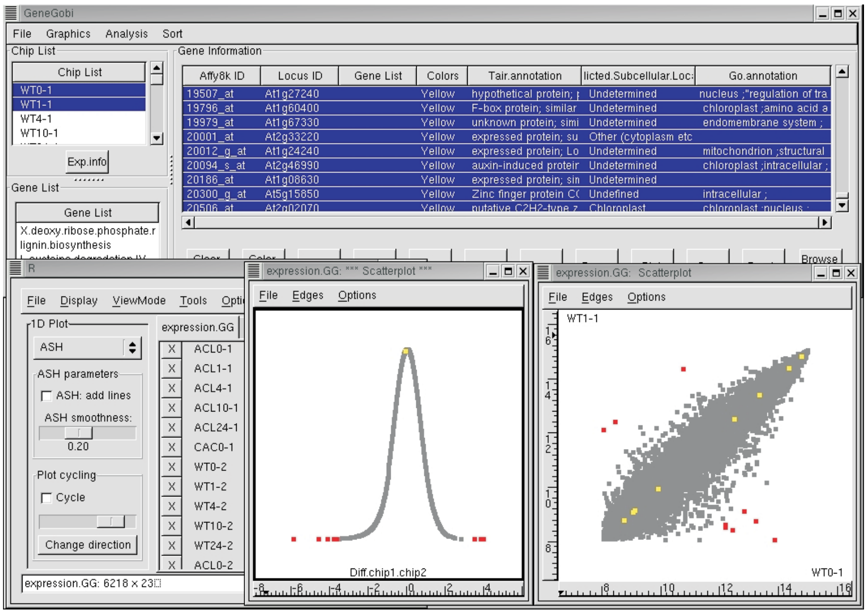
Affy8k ID	Locus ID	Gene List	Colors	Tair.annotation	licted.Subcellular.Loc.	Go.annotation
15695_s_at	At2g18050		Yellow	histone H1; supportec	Other (cytoplasm etc	DNA binding ;nucleus ;nı
14418_at	At2g37460		Orange	nodulin-like protein; s	Other (cytoplasm etc	membrane ;
20641_at	At1g52690		Pink	late embryogenesis-a	Other (cytoplasm etc	
19150_at	At1g54020		Blue	myrosinase-associate	Secreted Proteins	endomembrane system ;e
20220_at	At4g02280		Gray	sucrose synthase, pu	Undefined	glycosyltransferase ;mito
17273_at	At2g39330		Gray	putative myrosinase-t	Undetermined	
19193_at	At4g16590		Gray	cellulose synthase lik	Other (cytoplasm etc	cellulose synthase ;glycc
20720_at	At1g22990		Gray	copper chaperone (C	Undetermined	heavy metal ion transpor
12556_at	At4g29700		Gray	nucleotide pyrophosp	Secreted Proteins	nucleotide pyrophosphat
20201_at	At4g29030		Gray	glycine-rich protein li	Secreted Proteins	endomembrane system ;
16769_at	At5g64000		Gray	3(2),5-bisphosphate r	Undefined	nucleotidase ;inositol/phc
19762_at	At4g21680		Gray	peptide transporter - I	Other (cytoplasm etc	transporter ;oligopeptide
12364_at	At3g57240		Gray	glycosyl hydrolase fa	Other (cytoplasm etc	carbohydrate metabolism

ChipID	genotype	time	replicate
WT0-1	WT	0	1
WT1-1	WT	1	1
WT4-1	WT	4	1
WT10-1	WT	10	1
WT24-1	WT	24	1
ACL0-1	ACL	0	1
ACL1-1	ACL	1	1
ACL4-1	ACL	4	1
ACL10-1	ACL	10	1
ACL24-1	ACL	24	1

Figure 2: How to use the "Compare Genotype" menu: 1) Choose the first replicates of two genotypes, WT and ACL from the Chip List. 2) Choose Compare Genotype → Covariance 3) gene 15695_s_at(yellow) is the most different gene between WT and ACL according to this covariance. It has higher expression for ACL on the 5 time points than WT.

- Find Interesting Genes: Before you choose this menu item, you need to select two chips from Chip List. After calculating the statistics, a new variable containing these values is added to GGobi and the genes in the Gene Information spreadsheet are sorted by this calculated measure.

- Difference: Fit $Y = X$ for two selected experiments from the Chip List and calculate the residuals. Genes are sorted by the absolute values of these residuals in descending order. This measure is used for the log-scaled data to represent fold changes.
- Regression: Fit a simple linear regression model($Y = a \cdot X + b$) for the two selected experiments from the Chip List and calculate the residuals. Genes are sorted by the absolute values of these residuals in descending order.
- Angle: Calculate the angle from Y=X line for two selected experiments from Chip List. Genes are sorted by these angles in descending order. This measure is used for the raw data (without log transformation) to represent fold changes.
- Mahalanobis: Calculate the Mahalanobis distance from the means of two selected experiments from Chip List. $d_M(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{x}})^T \Sigma^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}})$. Genes are sorted by these distances in descending order. Here, Σ is estimated from the two selected experiments from the Chip List.)



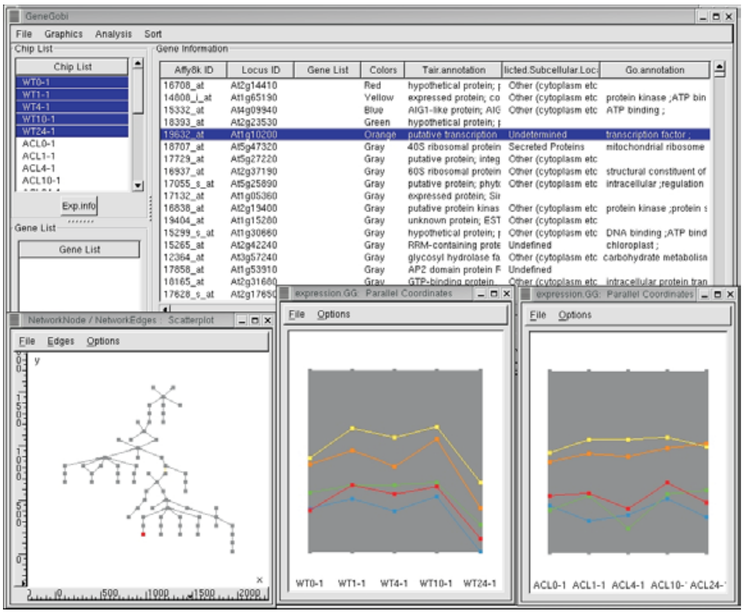
Affy8k ID	Locus ID	Gene List	Colors	Tair.annotation	licted.Subcellular.Loc:	Go.annotation
19507_at	At1g27240		Yellow	hypothetical protein; ɪ	Undetermined	nucleus ;"regulation of tra
19796_at	At1g60400		Yellow	F-box protein; similar	Undetermined	chloroplast ;amino acid a
19979_at	At1g67330		Yellow	unknown protein; simi	Undetermined	endomembrane system ;
20001_at	At2g33220		Yellow	expressed protein; su	Other (cytoplasm etc	
20012_g_at	At1g24240		Yellow	expressed protein; Lo	Undetermined	mitochondrion ;structural
20094_s_at	At2g46990		Yellow	auxin-induced proteir	Undetermined	chloroplast ;intracellular ;
20186_at	At1g08630		Yellow	expressed protein; sin	Undetermined	
20300_g_at	At5g15850		Yellow	Zinc finger protein C(	Undefined	intracellular ;
20506_at	At2g02070		Yellow	putative C2H2-type z	Chloroplast	chloroplast ;nucleus ;

ASH smoothness: 0.20

expression.GG: 6218 × 23

Figure 3: How to use “Find Interesting Genes” menu: 1) Choose two chips from Chip List, WT-0, and WT-1. 2) Choose Find Interesting Genes → Difference 3) genes with red colors are most different genes between two chips and genes with yellow colors are most similar genes between two chips.

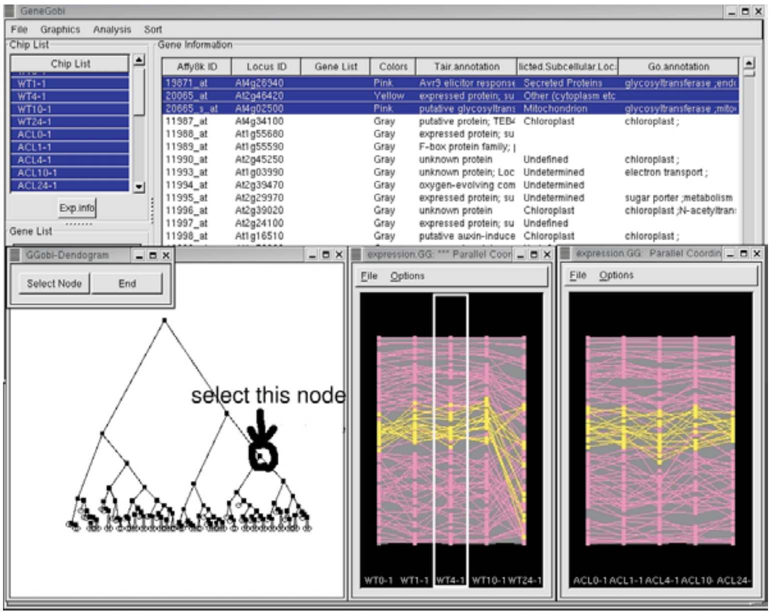
- Find Similar Genes: Before you choose this menu, you need to select one gene from the Gene Information spreadsheet and select chips that you want to consider. After calculating the statistics, the values are added to GGobi as a new variable and the genes in the Gene Information spreadsheet are sorted by this calculated statistic.
  - Euclidean: Calculate the Euclidean distance from the selected gene($\boldsymbol{x}^*$). Genes that are the closest will be the most similar.
  - Corr: Calculate the correlation distance from the selected gene. Genes that have a similar pattern will be the most similar.
  - Zerocorr: Calculate the zero correlation distance from the selected gene. Genes that have a similar pattern with respect to a baseline of 0 will be the most similar.



Affy8k ID	Locus ID	Gene List	Colors	Tair.annotation	licted.Subcellular.Loc	Go.annotation
16708_at	At2g14410		Red	hypothetical protein; p	Other (cytoplasm etc	
14808_i_at	At1g65190		Yellow	expressed protein; co	Other (cytoplasm etc	protein kinase ;ATP bin
15332_at	At4g09940		Blue	AIG1-like protein; AIG	Other (cytoplasm etc	ATP binding ;
18393_at	At2g23530		Green	hypothetical protein; p		
19632_at	At1g10200		Orange	putative transcription	Undetermined	transcription factor ;
18707_at	At5g47320		Gray	40S ribosomal protein	Secreted Proteins	mitochondrial ribosome
17729_at	At5g27220		Gray	putative protein; integ	Other (cytoplasm etc	
16937_at	At2g37190		Gray	60S ribosomal protein	Other (cytoplasm etc	structural constituent of
17055_s_at	At5g25890		Gray	putative protein; phyt	Other (cytoplasm etc	intracellular ;regulation
17132_at	At1g05360		Gray	expressed protein; Sir		
16838_at	At2g19400		Gray	putative protein kinas	Other (cytoplasm etc	protein kinase ;protein s
19404_at	At1g15280		Gray	unknown protein; EST	Other (cytoplasm etc	
15299_s_at	At1g30660		Gray	hypothetical protein; p	Other (cytoplasm etc	DNA binding ;ATP bind
15265_at	At2g42240		Gray	RRM-containing prote	Undefined	chloroplast ;
12364_at	At3g57240		Gray	glycosyl hydrolase fa	Other (cytoplasm etc	carbohydrate metabolism
17858_at	At1g53910		Gray	AP2 domain protein F	Undefined	
18165_at	At2g31680		Gray	GTP-binding protein	Other (cytoplasm etc	intracellular protein tran
17628_s_at	At2g1765					

Figure 4: How to use “Find Similar Genes” menu: The plots at the bottom include a view of a hypothetical jasmonic acid network and two profile plots of gene expression data.

- Find Clusters:
  - hclust: First select a subset of data (selected genes and selected chips), and use hierarchical clustering to cluster the genes. An interactive dendogram is drawn and the user can select nodes at the dendogram to highlight the cases in the cluster.



Affy8k ID	Locus ID	Gene List	Colors	Tair.annotation	licted.Subcellular.Loc.	Go.annotation
19871_at	At4g26940		Pink	Avr9 elicitor response	Secreted Proteins	glycosyltransferase ;endo
20065_at	At2g46420		Yellow	expressed protein; su	Other (cytoplasm etc	
20665_s_at	At4g02500		Pink	putative glycosyltrans	Mitochondrion	glycosyltransferase ;mito
11987_at	At4g34100		Gray	putative protein; TEB	Chloroplast	chloroplast ;
11988_at	At1g55680		Gray	expressed protein; su		
11989_at	At1g55590		Gray	F-box protein family;		
11990_at	At2g45250		Gray	unknown protein	Undefined	chloroplast ;
11993_at	At1g03990		Gray	unknown protein; Loc	Undetermined	electron transport ;
11994_at	At2g39470		Gray	oxygen-evolving com	Undetermined	
11995_at	At2g29970		Gray	expressed protein; su	Undetermined	sugar porter ;metabolism
11996_at	At2g39020		Gray	unknown protein	Chloroplast	chloroplast ;N-acetyltran
11997_at	At2g24100		Gray	expressed protein; su	Undefined	
11998_at	At1g16510		Gray	putative auxin-induce	Chloroplast	chloroplast ;

Figure 5: How to use the "Find Cluster" menu: Use the "select node" button to change the dendogram to interactive mode. When a node is selected the cases underneath the node (the cluster) are colored yellow and the cases corresponding to this cluster are highlighted yellow in other plots.

2.2.4 Sort menu : sort genes in the Gene Information spreadsheet by AffyID or LocusID or GeneList or Colors.

2.3 Buttons

- Clear Color: change all colors in GGobi plot to "Gray" which is used as a base color
- Color Update: when user changes the colors in GGobi directly using "BRUSH", use this button to update color information in the gene information spreadsheet.
- Yellow/Red/Blue/$\cdots$/Purple: change colors for selected genes in the gene information spreadsheet.
- Browse URL: for selected genes, link to literature from the web

3 Future directions

GeneGobi provides a user-interface to both GGobi and R. It adds a spreadsheet to more information about each gene, links to literature from the Web, menus of analysis and visualization options and an interface to lists of selected genes created from expert knowledge or previous analyses. For the biologist accustomed to point-and-click interfaces, it is intended to provide an easier entry into the analysis tools of R. The expert user has the full functionality of R available. For it to be accessible to biologists it needs to be available on the Windows and Mac operating systems. It is currently available only to linux users, but when the components are compiled for the other operating systems, it will be available and easy to install on the other systems.

The next directions of the software are: (1) Test the ease of use for biologists, (2) Interface with the graph manipulation software in R for metabolic network handling, (3) Expand the data types handled to include proteomics and metabolomics data. We aim for GeneGobi to be a new platform for exploring relationships between genes, proteins and metabolites, and for it to be freely available and easy to use by biologists. More information about GeneGobi is available at `http://www.public.iastate.edu/∼dicook/GeneGobi/GeneGobi.html`.

References

[1] Parmigiani G., Garrett E.S., Irizarry R.A., Zeger S.L. (2003). *The analysis of gene expression data*. Springer, New York, NY.

[2] Swayne D.F., Temple Lang D., Buja A., Cook D. (2002). *GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization*. Journal of Computational Statistics and Data Analysis **43**, 423 − 444.

Acknowledgement: This work has been supported in part by grant from the National Science Foundation Arabidopsis 2010 DB1-0209809.

Address: E.-K. Lee, D. Cook, H. Hofmann, E. Wurtele, D. Kim, J. Kim, H. An, Iowa State University, Ames, IA, USA

E-mail: `kyung@iastate.edu`

EXPLORING THE STRUCTURE OF MIXTURE MODEL COMPONENTS

Friderich Leisch

Key words: Finite mixture models, regression, cluster analysis, visualization.

COMPSTAT 2004 section: Clustering, Statistical software.

Abstract: Model-based cluster analysis and latent class regression are popular methods for grouping observations into unobserved segments. In many applications it is of great interest to the practitioner to assess the relationships between those segments, especially which segments are close to each other and which are markedly different from the rest. We present several new tools for the R statistical computing environment that allow the user to visually explore the component structure of arbitrary mixture models and do computations using a graph representation of the model.

1 Introduction

Finite mixture models have been used for more than 100 years, but have seen a real boost in popularity over the last decade due to the tremendous increase in available computing power. The areas of application of mixture models range from biology and medicine to physics, economics and marketing. On the one hand these models can be applied to data where observations originate from various groups and the group affiliations are not known, and on the other hand to provide approximations for multi-modal distributions [4], [12], [9].

In the 1990s finite mixture models have been extended by mixing standard linear regression models as well as generalized linear models [14]. An important area of application of mixture models and also of these extensions are in market segmentation [15], where finite mixture models replace more traditional cluster analysis and cluster-wise regression techniques as state of the art. Finite mixture models with a fixed number of components are usually estimated with the EM algorithm within a maximum likelihood framework [2] and with MCMC sampling [3] within a Bayesian framework.

The R environment for statistical computing [10] features several packages for finite mixture models, including `mclust` for mixtures of multivariate Gaussian distributions [6, 5], `fpc` for mixtures of linear regression models [7] and `mmlcr` for mixed-mode latent class regression [1]. All of those primarily target one or more special cases of mixture models. Package `flexmix` implements an extensible framework for mixture modelling where users can easily create new models by supplying their own M-step for the EM algorithm [8].

Efficient estimation of mixture models has received a lot of attention over the last years, however model diagnostics and general visualization techniques are scarcely available. E.g., the confidence ellipses commonly used to visualize low-dimensional Gaussians cannot be used for regression models. In this paper we present several new tools implemented in `flexmix` that can be used to graphically explore the structure of the components of any finite mixture model. Of special interest in all applications where mixtures are used to group observations is which components are overlapping or "close" to each other. If the mixture model is used for market segmentation it is important to know for the practitioner which clusters are distinct market niches and which clusters are parts of larger consumer groups.

2 The posterior class probabilities

Consider finite mixture models with K components of form

$$h(y|x,w) = \sum_{k=1}^{K} \pi_k f(y|x,\theta_k) \tag{1}$$

$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1$$

where y is a (possibly multivariate) dependent variable with conditional density h, x is a vector of independent variables, π_k is the prior probability of component k, and θ_k is the component specific parameter vector for the density function f.

If f is a normal density with component-specific mean $\beta_k' x$ and variance σ_k^2, we have $\theta_k = (\beta_k', \sigma_k^2)'$ and Equation (1) describes a mixture of standard linear regression models, also called latent class regression. A special case is $x \equiv 1$, which gives a mixture of Gaussians without a regression part. If f is a member of the exponential family, we get a mixture of generalized linear models (GLMs).

The posterior probability that observation (x, y) belongs to class j is given by

$$\P(j|x,y) = \frac{\pi_j f_j(y|x,\theta_j)}{\sum_k \pi_k f_k(y|x,\theta_k)}$$

Histograms or rootograms of the posterior class probabilities can be used to assess the cluster structure [11], this is now the default plot method for `"flexmix"` objects. Rootograms are very similar to histograms, the only difference is that the height of the bars correspond to square roots of counts rather than the counts themselves, hence low counts are more visible and peaks less emphasized.

Usually in each component a lot of observations have posteriors close to zero, resulting in a high count for the corresponing bin in the rootogram which obscures the information in the other bins. To avoid this problem, all probabilities with a posterior below a threshold are ignored (we use 0.0001). A peak at probability 1 indicates that a mixture component is well seperated from the other components, while no peak at 1 and/or significant mass in the middle of the unit interval indicates overlap with other components.

As example we use a 2-component mixture of Poisson regression models with one independent variable, parameters $\theta_1 = (2, -0.2)'$ and $\theta_2 = (1, 0.1)'$, and the exponential link function. Hence, given x the response y in group k has a Poisson distribution with mean $\exp((1, x) \cdot \theta_k)$. A sample with 100 observations in each group is shown in Figure 1. For data stored in an R data frame `mydata` the mixture model can be estimated using the commands

```
R> model1 = flexmix(y ~ x, data = mydata, k = 2,
+     model = FLXglm(family = "poisson"))

Classification: weighted
  10 Log-likelihood:    -458.3680
  20 Log-likelihood:    -458.1333
  24 Log-likelihood:    -458.1307
converged
```

The estimated parameters are

```
     (Intercept)      x
[1,]       1.922 -0.181
[2,]       0.997  0.106
```

which is close to the true parameters. The corresponding clusters can be seen in the right panel of Figure 1.

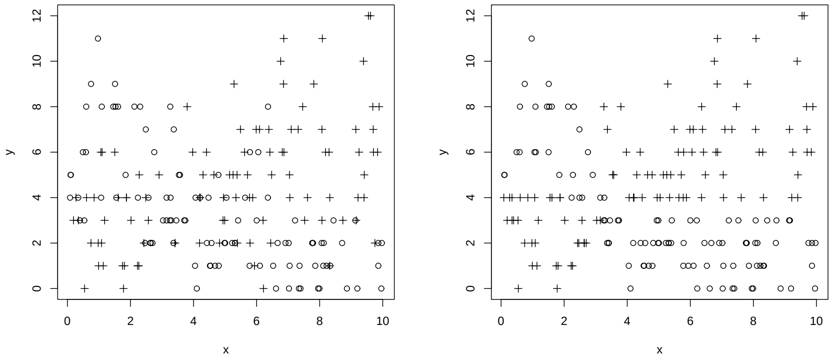


Figure 1: Poisson regression mixture with 2 components: true groups (left) and groups found by `model1` (right).

Issuing the command `plot(model1)` gives the rootograms shown in the left panel of Figure 2. The obvious overlap between the clusters is easily identified, the posteriors have almost a uniform distribution over the interval $[0, 1]$.

Now assume that instead of 200 independent observations we have 2 measurements each from 100 persons and that column `id` of `mydata` contains a factor identifying the 100 persons. If we use the additional information the EM algorithm needs only half the number of iterations to converge:

```
R> model2 = flexmix(y ~ x | id, data = mydata, k = 2,
+     model = FLXglm(family = "poisson"))
```

```
Classification: weighted
  10 Log-likelihood:    -889.0594
  13 Log-likelihood:    -889.0556
converged
```

The `model2` parameter estimates

```
     (Intercept)      x
[1,]        1.96 -0.201
[2,]        1.04  0.101
```

are only slightly better than for `model1`, but now we can assign the observations with more confidence into the two classes as the posteriors are shifted towards 0 and 1 (middle panel of Figure 2). If we have 4 repeated measurements from 50 persons, this effect is of course even much more pronounced (right panel of Figure 2) and there are only very few observations with posteriors close to 0.5.

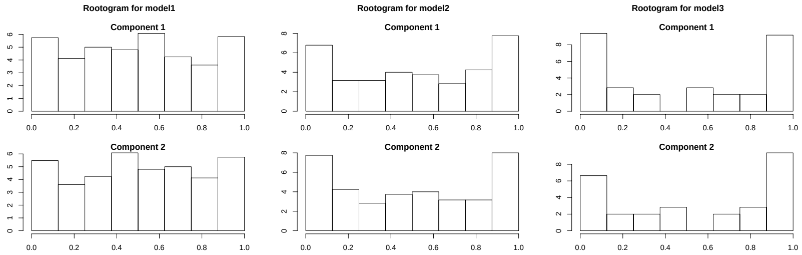


Figure 2: Rootograms for models with no repeated measurements (left), 2 (middle) and 4 (right) measurements per person.

3 Kullbach-Leibler divergence between component

Histograms or rootograms of posteriors visualize with how much confidence observations are assigned to clusters, but can not help to identify relationships between clusters in case of more than 2 components. Consider the smiley data from R package `mlbench` shown in Figure 3. Although only the "eyes" are really Gaussian, we can use model-based clustering with Gaussians to approximate the multimodal density (similar to a density estimate using a Gaussian kernel).

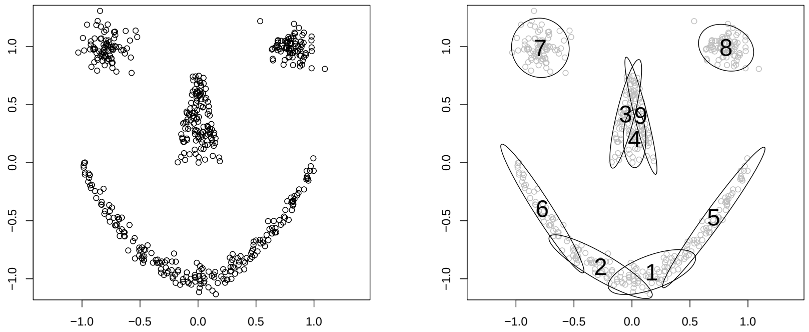


Figure 3: The smiley data (right) and a 9 component partition.

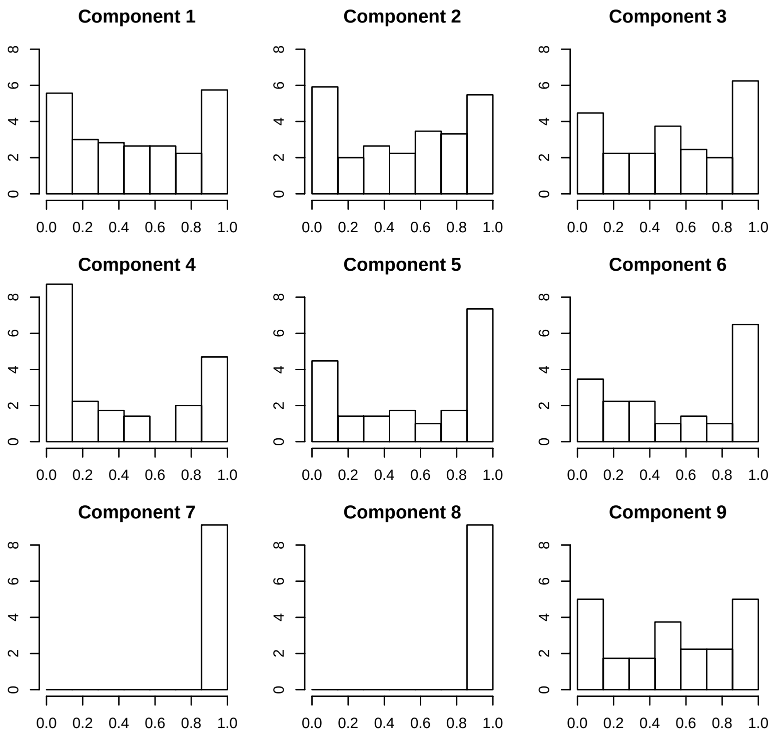


Figure 4: Rootograms of the 9 components for the smiley data.

The corrsponding rootograms are shown in Figure 4. Only components 7 and 8 (the "eyes") do not overlap with any other cluster, all others have a lot of posteriors much smaller than 1. One possibility to explore which pairs of clusters overlap would be to use the brushing facilities of interactive histograms as provided by the R package `iplots` [13]. Another way is to compute pairwise distances between the clusters. The most common distance measure for two distributions with densities f and g is the Kullbach-Leibler (KL) divergence

$$KL(f,g) = \int f(x)\,(\log f(x) - \log g(x))\,dx$$

(for discrete distributions the integral is replaced by a sum), which cannot be solved analytically in many cases. However, the KL divergence between mixture components k and l can be estimated as

$$\begin{aligned} KL(f_k, f_l) &\approx \sum_{n=1}^{N} p_{nk}\,(\log p_{nk} - \log p_{nl}) \\ p_{nk} &= \pi_k f(y_n|x_n, \theta_k) \end{aligned}$$

Evaluating the sum is numerically problematic because most p_{nk} are almost zero. To get a stable estimate we remove all terms in the sum involving densities below a threshold of $\epsilon = 0.01$, which results in the KL divergence matrix

```
     1   2   3   4   5   6   7   8   9
1    0  14   .   .  26   .   .   .   .
2   28   0   .   .   .  18   .   .   .
3    .   .   0 139   .   .   .   .  41
4    .   .  20   0   .   .   .   .  15
5   18   .   .   .   0   .   .   .   .
6    .  19   .   .   .   0   .   .   .
7    .   .   .   .   .   .   0   .   .
8    .   .   .   .   .   .   .   0   .
9    .   .  18 109   .   .   .   .   0
```

for the smiley data (rounded to integers). Dot entries correspond to components where the regions with densities larger than ϵ do not overlap.

The KL divergence matrix can be represented by a directed graph with one node for each component as shown in Figure 5. Overlapping clusters correspond to connected nodes and modes of the mixture density to cliques of the graph. For our 2-dimensional example data without covariates a natural positioning of the graph nodes are the centers of the clusters. For higher-dimensional data or regression mixtures this is not possible and we have to restrict ourselves to general graph layout algorithms. Sammon mapping of the KL divergences results in the right panel of Figure 5, where especially the linear structure of the "mouth" is preserved correctly. Of course all unconnected components of the graph are placed randomly with respect to each other (and could as well be projected seperately).

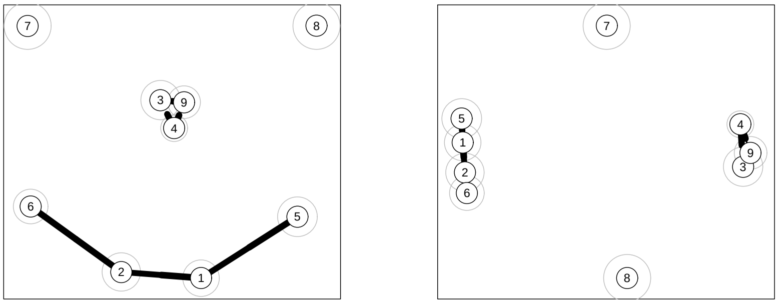


Figure 5: Graph corresponding to the KL divergences: node positions according to cluster centers (left) and Sammon mapping (right). The circles around the nodes are proportional to the cluster sizes.

4 Summary

Finite mixture models have become increasingly popular in many domains of applications, yet diagnostic tools for fitted models (and especially corresponding software) are much less developed. We have developed several new tools available in the R package `flexmix` which allow the user to explore the relationships between components of fitted mixture models.

All methods presented in this paper work off the densities or posterior probabilities of the observations and thus do not depend on the dimensionality of the input space. While we have used simple 2-dimensional examples to demonstrate the techniques, they can easily be used on high-dimensional data sets or models with complicated covariate structures.

As a next step we will integrate the graph representation of the model into more interactive visualization systems such that the user can easily explore the distribution of background variables. E.g., if each mixture component corresponds to a market segment, clicking on a node in the graph could show the distribution of sales and sociodemographic data of the consumers in the respective segment.

References

[1] Buyske S. (2003). *R Package* `mmlcr`*: mixed-mode latent class regression.* version 1.3.2.

[2] Dempster A., Laird N., Rubin D. (1977). *Maximum likelihood from incomplete data via the EM-alogrithm.* Journal of the Royal Statistical Society, B **39**, 1 – 38.

[3] Diebolt J., Robert C.P. (1994). *Estimation of finite mixture distributions through Bayesian sampling.* Journal of the Royal Statistical Society, Series B **56**, 363 – 375.

[4] Everitt B.S., Hand D.J. (1981). *Finite mixture distributions.* London: Chapman and Hall.

[5] Fraley C., Raftery A.E. (2002). *MCLUST: Software for model-based clustering, discriminant analysis and density estimation.* Technical Report 415, Department of Statistics, University of Washington, Seattle, WA, USA.

[6] Fraley C., Raftery A.E. (2002). *Model-based clustering, discriminant analysis and density estimation.* Journal of the American Statistical Association **97**, 611 – 631.

[7] Hennig C. (2000). *Identifiability of models for clusterwise linear regression.* Journal of Classification **17**, 273 – 296.

[8] Leisch F. (2003). FlexMix: *A general framework for finite mixture models and latent class regression in R.* Report 86, SFB "Adaptive Information Systems and Modeling in Economics and Management Science".

[9] McLachlan G., Peel D. (2000). *Finite mixture models.* John Wiley and Sons Inc.

[10] R Development Core Team. (2003). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

[11] Tantrum J., Murua A., Stuetzle W. (2003) *Assessment and pruning of hierarchical model based clustering.* In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 197 – 205. ACM Press, New York, NY, USA.

[12] Titterington D., Smith A., Makov U. (1985). *Statistical analysis of finite mixture distributions.* Chichester: Wiley.

[13] Urbanek S., Theus M. (2003). *iPlots — high interaction graphics for R.* In Hornik K., Leisch F., Zeileis A., (eds), Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

[14] Wedel M., DeSarbo W.S. (1995). newblock *A mixture likelihood approach for generalized linear models.* Journal of Classification **12**, 21 – 55.

[15] Wedel M., Kamakura W.A. (2001). *Market segmentation - conceptual and methodological foundations.* Kluwer Academic Publishers, Boston, MA, USA, 2 edition.

Address: F. Leisch, Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Wien, Austria

E-mail: `Friedrich.Leisch@ci.tuwien.ac.at`

TESTING NONLINEAR COINTEGRATION

Jin-Lung Lin and Clive W.J. Granger

Key words: Nonlinear cointegration, unit root, local time, KPSS test.

COMPSTAT 2004 section: Time series analysis.

Abstract: This paper proposes a framework for defining and testing nonlinear cointegration. We make use of Park and Phillips's asymptotic analysis and propose a residual-based test. For the given specific functional form, we first run the nonlinear least squares to estimate the underlying parameters. Under the null of nonlinear cointegration, these NLS estimates converge to the true values generally at a rate faster than $\sqrt{T}$ and the estimated residuals behave similarly to the true residuals, as is in the case of linear cointegration. Then, by applying the KPSS test to estimated residual which has I(0) as null hypothesis will determine if there exists nonlinear cointegration. Simulation results support the proposed methods as power of the test converges to one quickly when sample size increases. We also examine the notion of spurious co-trend especially when the function is polynomial with even power.

1 Introduction

While cointegration is a powerful tool for the analysis of long run economic equilibrium, it is build upon linear models. It is widely believed that relationships between important economic series are nonlinear but each series is persistent. A simple example is:

$$y_t = a_0 x_t^2 + u_t$$

where $x_t = x_{t-1} + \epsilon_t + \gamma\epsilon_{t-1}, u_t$ and ϵ_t are two white noise processes independent of each other. x_t is clearly a linearly integrated IMA(1) process and y_t shares similar properties of an I(1) process. y_t and x_t are not linearly cointegrated but y_t and x_t^2 are. In other words, y_t and x_t are *nonlinearly cointegrated* in the sense that there exists an instantaneous nonlinear transformation, f, such that y_t and $f(x_t)$ are *linearly cointegrated*.

This paper proposes a framework for defining and testing nonlinear cointegration. We make use of Park and Phillips's asymptotic analysis [10], [11], as well as Chang, Park and Phillips [1] and propose a residual-based test. For some limited range of functional forms, we first run the nonlinear least squares (hereafter, NLS) to estimate the underlying parameters. Under the null of nonlinear cointegration, these NLS estimates converge to the true values generally at a rate faster than $\sqrt{T}$ and the estimated residuals behave similarly to the true residuals as is in the case of linear cointegration. Then, by applying the KPSS test [9] to estimated residual which has I(0) as null hypothesis will determine if there exists nonlinear cointegration.

In additional to this introduction, Section 2 reviews properties of nonlinear transformations of integrated processes. The definition of and testing nonlinear cointegration are discussed in Section 3. Simulation results are summarized in Section 4 and Section 5 concludes.

2 Nonlinear transformation of integrated processes

Let $\{y_t, x_t\}$ be generated by

$$y_t = f(x_t) + u_t, \quad x_t \quad = \quad x_{t-1} + v_t, \quad v_t = \sum_{k=1}^{\infty} \phi_k \epsilon_{t-k},$$

where u_t, ϵ_t are two $i.i.d.'s$ with mean 0 and $\sum_{i=1}^{\infty} |\phi_k| < \infty$ and f is nonlinear. As was pointed out by Granger and Hallman [6], and Ermini and Granger [3], the properties of the transformed process, y_t, depend crucially on the types of nonlinear functions. It is genuinely impossible to give an exhaustive list of nonlinear functions and hence we only focus upon asymptotic homogeneous functions. A function f is asymptotic homogeneous provided

$$f(\lambda x, \theta) = \kappa(\lambda, \theta) H(x, \lambda, \theta) + R(x, \lambda, \theta)$$

where $R(x, \lambda, \theta)$ is of order smaller than $\kappa(\lambda, \theta)$. κ, is called the asymptotic order and H the limit homogeneous function. Note that $\kappa(\lambda, \theta) = \lambda$ is a special case arising from linear transformation. Functions like polynomial, logistic, and Box-Cox transformation are all such examples. See Park and Phillips [10] for details.

3 Defining and testing nonlinear cointegration

The notion of I(0) serves as the building block of linear cointegration theory but there are no consensus about defining I(0) for nonlinear process. Karlsen, Myklebust and Tjostheim [8] allows error term to be nonlinear transformation of Markov chain satisfying some mixing conditions. While this approach admits more general class of process, it is somewhat difficult to check the mixing conditions and implement parametric data generation process in practice. Thus, we define general I(0) loosely as the process such that functional central limit theorem are applicable. To be precise, u_t is general I(0) if $\frac{1}{\sqrt{T}\sigma} \sum_{t=1}^{[Tr]} u_t \rightarrow_d U(r)$ where σ is some constant, $[x]$ denotes the integer part of x, and $U(r)$ is a Brownian motion.

Since differencing is a linear operator and integration order is a concept for linear processes, it is difficult to generalize this definition to nonlinear processes. An ideal substitute would be a measure of nonlinear persistence, which can be applied to a large class of nonlinear models and can be easily computed for real data. While there are some works along this line, for examples, Gourierous and Jasiak [4] proposed a transformed autocorrelogram to measure temporal dependence of a transformed series, Granger and Lin [3] investigated mutual information coefficient, not many useful results are available so far. To take advantage of the asymptotic theory proposed by Park and Phillips, we propose to define nonlinear cointegration in terms of the order of asymptotically homogeneous functions. It is worth noting that any pair of I(1) series are nonlinearly cointegrated if no restriction is imposed upon f.

Definition: nonlinear cointegration
Let x_t be a linearly integrated process. y_t and x_t are called *nonlinearly cointegrated* with asymptotic function f provided $u_t = y_t - f(x_t)$ has order smaller than those of y and $f(x)$. For example, let $y_t = x_t^2 + x_t$ and then y_t and x_t are nonlinearly cointegrated with $f(x) = x^2$ since $y_t - f(x_t) = x_t$ which has order $\kappa(\lambda) = \lambda$ smaller than $\kappa(\lambda) = \lambda^2$, the asymptotic order of y_t and $f(x_t)$.

Definition: full nonlinear cointegration:
Let x_t be a linearly integrated process which is said to be *fully nonlinearly cointegrated* with y_t with asymptotic homogeneous function f if $u_t = y_t - f(x_t)$ is a general I(0) process.

Now, let us turn to hypothesis testing. For linear processes with either I(1) or I(0), no cointegration is often specified as null hypothesis and cointegration as alternative when using the *ADF* or *Z* tests. This can be justified that for AR models, cointegration implies singularity of covariance matrix that leads to inconsistent test. See Phillips and Quliaris [12] for details. However, this is inappropriate for nonlinear transformation of integrated processes. To illustrate, let x_t be I(1) series and $y_t = g(x_t) = x_t^3 + u_t$ where g is asymptotic homogeneous function with order $\kappa_g(\lambda) = \lambda^3$. Suppose one wants to test if y_t and x_t are nonlinearly cointegrated with $f(x) = ax^2$. Running the regression $y_t = ax_t^2 + u_t$ would result in $a = 0$ and $u_t = y_t$ under the null of no-cointegration. u_t is not linearly I(1) process and ADF or Z-test are not applicable. On the other hand, if the null hypothesis is specified as cointegration, KPSS-test would give the right distribution under the null hypothesis and power approaching one as sample size grows under the alternative. Another example is as below. Let x_{1t}, x_{2t} be two independent random walks and $y_t = x_{1t}^2 + u_t$, where u_t is white noise. Suppose the model is misspecified to include the wrong regressor, x_{2t} and Thus, the nonlinear cointegration regression becomes $y_t = \beta x_{2t}^2 + v_t$. Obviously, true $\beta = 0, v_t = y_t$. Thus, under the null of no cointegration, residual is no longer linear I(1) process. On the other hand, If y_t, x_t are indeed cointegrated, then we can employ KPSS test where the null is specified as cointegration. The test statistics will have power against the alternative of no cointegration where residuals generally follows a nonlinear process with asymptotic order greater than 1.

How to test if y_t, and x_t are nonlinear cointegrated with specific functional form $f(x, \theta)$? A possible strategy is as below:

1. run NLS to obtain $\hat{\theta}$ and $\hat{u}_t = y_t - f(x_t, \hat{\theta})$,
2. apply KPSS to $\hat{u}_t$

If KPSS test reject the I(0) null hypothesis, then reject the null of nonlinear cointegration hypothesis and vice versus.

It is worth noting that the nonlinear least square estimate is consistent for the nonlinear cointegration even when residual is not I(0), say, I(1) process. For examples, Let $x_t = x_{t-1} + \epsilon_t$ be a random walk and $y_t = a_0 x_t^2 + a_1 x_t + u_t$, where u_t is white noise for simplicity. Then y_t is also nonlinearly cointegrated with $f(x) = a_0 x_t^2$. The NLS for a_0 from regressing y_t on x_t^2 is consistent.

4 Simulation results

4.1 Model setup and experiment design

To assess the performance of the proposed test statistics, several simulation experiments are carried out. Two models are chosen for simulating time series: square function and Box-Cox function.

1. Model 1: $y_t = a_0 x_{1t}^2 + u_t, \quad a_0 = 1$
2. Model 2: $y_t = \frac{((|x_{1t}|)^{\theta} - 1)}{\theta}, \quad \theta = 0.1$

where $x_{1t} = x_{1t-1} + \epsilon_{1t} + \gamma_1 \epsilon_{1t-1}$, $x_{2t} = x_{2t-1} + \epsilon_{2t} + \gamma_2 \epsilon_{2t-1}$, $\gamma_1 = 0.2, \gamma_2 = 0.5$, ϵ_{1t}, ϵ_{2t} are two independently identically distributed random variables with mean 0 and standard deviation 0.1. For simplicity, u_t is drawn independent of $\epsilon_{1t}, \epsilon_{2t}$ with mean 0 and standard deviation, 0.1. See Chang, Park and Phillips [1] for details. As for power property, we examine three alternative models: (1) I(1) residual, (2) wrong functional form, and (3)wrong regressor. Sample size are respectively 100, 250, 500, and 1000 with number of replications set to be 1000. Algorithm, OPTMUM in GAUSS is used for nonlinear least square estimation. Simulation results are reported in Tables 1 and 2. Each panel contains mean and standard deviation of parameter estimates computed from 1000 replications (not the mean of theoretical deviation) are reported and so are rejection ratio of KPSS level and KPSS trend test.

4.2 Size

From Table 1, we observe that parameters of correctly specified models are well estimated with small bias and small deviation. As expected, as sample size grows, parameter estimate for the square model converges fast than that for the Box-Cox model. Both KPSS level and KPSS trend test perform well but the size distortion does not vanish as sample size increases. The result of benchmark model suggest that size distortion problem is not unique to nonlinear models.

4.3 Power

Table 2 summarizes the results for all two models with I(1) residuals. As is noted above, NLS estimate is still consistent for square model, which can be seen from first panel of this table. However, the standard deviation of

Model 1: $y_t = ax_t^2 + u_t, \quad a = 1$				
nobs	100	250	500	1000
$\hat{a}$	0.99996354	1.0001454	1.0000993	1.0000147
std	0.035664536	0.0078832890	0.0028966686	0.0011690319
KPSS level	0.031000000	0.035000000	0.030000000	0.033000000
KPSS trend	0.049000000	0.044000000	0.040000000	0.043000000
Model 2: $y_t = (\lvert x\rvert^\theta - 1)/\theta) + u_t, \quad \theta = 0.1$				
nobs	100	250	500	1000
$\hat{\theta}$	0.10022939	0.099700836	0.099891893	0.10014613
std	0.00906639	0.0060835815	0.0043199875	0.0026622429
KPSS level	0.03400000	0.046000000	0.046000000	0.056000000
KPSS trend	0.03900000	0.047000000	0.047000000	0.035000000
Benchmark: $y_t = u_t$				
nobs	100	250	500	1000
KPSS level	0.045000000	0.045000000	0.046000000	0.062000000
KPSS trend	0.054000000	0.052000000	0.046000000	0.038000000

Table 1: NLS estimates and size of KPSS test at 5%.

Power for I(1) Residuals						
Model 1: $y_t = ax_t^2 + U_t, U_t = \sum_{i=1}^t u_i, \quad a = 1$						
nobs	100	250	500	1000		
$\hat{a}$	0.95003881	0.96045263	0.99685236	0.97685468		
std	1.5294337	0.87039486	0.64718767	0.44820097		
KPSS level	0.63200000	0.86500000	0.96700000	0.99500000		
KPSS trend	0.77000000	0.93100000	0.99500000	0.99900000		
Model 2: $y_t = (\lvert x\rvert^\theta - 1)/\theta + U_t, U_t = \sum_{i=1}^t u_i, \theta = 0.1$						
nobs	100	250	500	1000		
$\hat{\theta}$	21.478836	6.4544065	0.78929601	0.64282239		
std	308.30909	134.03587	1.8076907	1.1604284		
KPSS level	0.80800000	0.92300000	0.98300000	0.99400000		
KPSS trend	0.77400000	0.94900000	0.98700000	0.99800000		
Benchmark : $y_t = U_t, U_t = \sum_{i=1}^t u_i$						
obs	100	250	500	1000		
KPSS level	0.82800000	0.94300000	0.99600000	0.99800000		
KPSS trend	0.81100000	0.97200000	0.99900000	1.0000000		
Power for different alternative						
Wrong functional Form: True: $y_t = ax_t^3 + u_t$, Estimate $y_t = bx_t^2 + u_t, \quad a = 1$						
nobs	100	250	500	1000		
$\hat{b}$	-0.0087281	0.017079	0.10096	-1.0868		
std	4.0753	6.1816	11.268	20.583		
KPSS for level	0.67200	0.87800	0.97300	0.99100		
KPSS for trend	0.69300	0.92500	0.99600	1.0000		
Wrong Regressors I: True: $y_t = ax_{1t} + u_t$, Estimate $y_t = bx_{2t} + u_t, \quad a = 1$						
nobs	100	250	500	1000	5000	10000
$\hat{b}$	-0.023211	-0.0021045	0.018929	0.049858	0.0045379	0.0011213
std	0.73661	0.74575	0.76449	0.7366	0.73491	0.73856
KPSS level	0.61800	0.87700	0.95500	0.99300	1.0000	1.0000
KPSS trend	0.77900	0.95300	0.99800	1.0000	1.0000	1.0000
Wrong Regressors II: True: $y_t = ax_{1t}^2 + u_t$, Estimate $y_t = bx_{2t}^2 + u_t, \quad a = 1$						
nobs	100	250	500	1000	5000	10000
$\hat{b}$	0.76007	0.89129	0.90559	0.81706	0.90476	0.81768
std	1.2998	1.5945	1.6541	1.3934	1.7099	1.4316
KPSS level	0.47000	0.77500	0.92800	0.97700	1.0000	1.0000
KPSS trend	0.66700	0.89500	0.98700	0.99500	1.0000	1.0000
Wrong Regressors III: True: $y_t = ax_{1t}^3 + u_t$, Estimate $y_t = bx_{2t}^3 + u_t, \quad a = 1$						
nobs	100	250	500	1000	5000	10000
$\hat{b}$	-0.28042	-0.074693	-0.076194	0.097472	-0.017838	0.023927
std	3.6577	3.5300	3.6212	3.3666	3.3784	4.0437
KPSS level	0.45200	0.77900	0.93900	0.98000	1.0000	1.0000
KPSS trend	0.64400	0.88300	0.98300	0.99500	1.0000	1.0000
Wrong Regressors IV: True: $y_t = ax_{1t}^4 + u_t$, Estimate $y_t = bx_{2t}^4 + u_t, \quad a = 1$						
nobs	100	250	500	1000	5000	10000
$\hat{b}$	2.4837	3.3251	2.5240	2.3810*	2.3850	2.1889
std	14.613	23.838	10.826	10.741	13.292*	10.920
KPSS level	0.39600	0.70800	0.88700	0.94200*	0.99700	0.99900
KPSS trend	0.52400	0.83600	0.95500	0.99100*	1.0000	1.0000

Table 2: Power of KPSS test.

parameter estimate is much larger than that in Table 1. As for Box-Cox and Logistic models, I(1) residuals results in inconsistent estimate, as is seen from panels 2 and 3 of Table 2. Power of KPSS test converges to one rapidly.

From the bottom panel of Table 2, we find that wrong functional form give rise to irregular parameter estimate as sample grows. Quite surprisingly, for the case of wrong regressor II, where $y_t = f(x_{1t}) = ax_{1t}^2 + u_t$, $a = 1$ but a wrong regressor x_{2t}, another random walk independent of x_{1t}, parameter estimate, $\hat{a}$ fluctuates relatively stable between 0.8 and 0.9 but bias does not shrink as sample size grows. To investigate if spurious co-trend is generated from square function, we further simulate the cases with linear, cubic and quartic functions and summarize the results in the lower panels of Table 2. From the table, we observe that the parameter estimate center around 0 for odd power but the standard deviation does not converge to 0. For the quartic case, the parameter centers around 2.5 with standard deviation fluctuating between 10 and 20. However, for all cases, power of KPSS level and KPSS tests converge to one as sample size increase.

To gain intuition about the simulation results above, we shall perform a simple asymptotic analysis. Let $\hat{b}_i, i = 1, \cdots, 4$ denote the NLS estimates of b_i in linear, quadratic, cubic and quartic models respectively. Note that in these simple cases, NLS reduces to OLS. $E(\hat{b}_1) = E(\hat{b}_3) = 0$, $E(\hat{b}_2) \neq 0$, $E(\hat{b}_4) \neq 0$ all have finite variances. In other words, for all the wrong regressor cases, asymptotic consistency does not hold and the parameter estimates without any normalization is a random variable. The mean of the estimates with even power are nonzero and might accidently give rise to close estimate to true parameters.

To summarize, KPSS level and trend test perform well in our simple simulation experiments and can be used to test for existence of nonlinear cointegration.

5 Conclusions

This paper has provided a framework for defining and testing nonlinear cointegration. We distinguish nonlinear cointegration and full nonlinear cointegration. Using the asymptotic results of Park and Phillips's, we propose a residual-based test where the null hypothesis has to be existence of nonlinear cointegration. For the given specific functional form, we run the nonlinear least squares to estimate the underlying parameters and then apply the KPSS test to the estimated residual. If the residuals behave like I(0), then we shall reject the null hypothesis of nonlinear cointegration. Simulation results support the proposed method.

This paper should be considered as an exploratory analysis into this difficult topic and there remain many questions to be investigated in the future. For examples, how to test nonlinear cointegration where order is reduced but residual still remains as persistent? Is there a link between classes of functions?

References

[1] Chang Y., Park J.Y., Phillips P.C.B. (2001). *Nonlinear econometric models with cointegrated and deterministically trending regressors.* Econometric Journal **4**, 1 – 36. Forthcoming Journal of Econometrics.

[2] Escribano A., Mira S.(1996). *Non-linear cointegration and non-linear error-correction models.* Working Paper, Department of Economics, Carlos III University of Madrid.

[3] Ermini L., Granger C.W.J.(1993). *Some generalizations on the algebra of I(1) processes.* Journal of Econometrics **58**, 369 – 384.

[4] Gourieroux C., Jasiak J.(1999). *Nonlinear persistence and copersistence.* Working Paper, CREST.

[5] Granger C.W.J. (1995). *Modelling nonlinear relationships between extended memory variables.* Econometrica **63**, 265 – 279.

[6] Granger C.W.J., Hallman J.(1991). *Nonlinear transformations of integrated time series.* Journal of Time Series Analysis **12**, 207 – 224.

[7] Granger C.W.J., Inoue T., Morin N.(1997). *Nonlinear stochastic trends.* Journal of Econometrics **81**, 65 – 92.

[8] Karlsen H A., Bruns J. and Tjostheim D.(2004). *Nonparametric estimation in nonlinear cointegration type model.* Working paper, Department of Mathematics, University of Bergen.

[9] Kwiatkowski D., Phillips P.C.B., Schmidt P., Shin Y.(1992). *Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?*. Journal of Econometrics **54**, 159 – 178.

[10] Park J.Y., Phillips P.C.B.(1999). *Asymptotics for nonlinear transformations of integrated time series.* Econometric Theory **15**, 269 – 298.

[11] Park J.Y. and Phillips P.C.B.(2001). *Nonlinear regression with integrated time series.* Econometrica **69**, 117 – 161.

[12] Phillips P.C.B. and Quliaris S.(1990). *Asymptotic properties of residual based tests for cointegration.* Econometrica, **58**, 165 – 193.

Acknowledgement: The first author would like to thank the financial support of National Science Council, Taiwan under the project NSC 91-2415-H-001-016.

Address: J.-L. Lin, Institute of Economics, Academia Sinica, No. 128, Sec. 2, Academia Rd, Nankang, Taipei, TAIWAN 11529

E-mail: `jlin@sinica.edu.tw`

CLUSTERING OF LARGE NUMBER OF STOCK MARKET TRADING RULES

Piotr Lipinski

Key words: Clustering, self-organizing maps, data mining, stock market financial time series, stock market trading rules.

COMPSTAT 2004 section: Applications, Clustering, Neural networks, Statistical data mining.

Abstract: This paper addresses the problem of clustering in large sets discussed in the context of financial time series. The goal is to divide stock market trading rules into several classes so that all the trading rules within the same class lead to similar trading decisions in the same stock market conditions. It is achieved using Kohonen self-organizing maps and the K-means algorithm. Several validity indices are used to validate and assess the clustering. Experiments were carried out on 350 stock market trading rules observed over a period of 1300 time instants.

1 Introduction

This paper addresses the problem of clustering in large sets discussed in the context of financial time series. Input data include tables of over one thousand rows (data vectors) and several hundreds of columns (variables). Each data vector is composed of trading decisions (observations) evaluated according to some so-called stock market trading rules (traits) at a given time instant.

The goal is to divide stock market trading rules into several classes so that all the trading rules within the same class lead to similar trading decisions in the same stock market conditions. In other words, the goal is to group variables on the basis of the observations gathered.

Before the actual clustering starts, dependencies among variables are investigated using principal component analysis and data dimensionality is reduced [6]. Next, variables are grouped using Kohonen self-organizing maps [7]. Finally, the K-means algorithm is applied to merge certain groups defined by the codebook vectors derived. Several validity indices, such as the Dunn index [3], the Davies-Bouldin index [2], the silhouette index [12], the C index [5] as well as the Goodman-Kruskal index [4], are used to assess and validate the clustering.

The issues discussed in this paper are an extension of my previous investigations on dependency mining in large set of stock market trading rules [10]. The problem considered is also related to my research on stock market decision support expert systems based on artificial intelligence, especially on evolutionary algorithms [8], [9].

This paper is structured in the following manner: Section 2 introduces the concept of a stock market trading rule and presents financial data from the Paris Stock Exchange. Section 3 discusses data preprocessing based on principal component analysis. Section 4 describes initial clustering using SOM. Section 5 explains how the groups defined by the codebook vectors derived are merged. Section 6 discusses some experiments carried out on real-life data from the Paris Stock Exchange. Finally, Section 7 concludes the paper.

2 Data description

Financial analysts and stock market traders observe quotations of stock prices with the aim to sell an item if it tends to lose value, to buy an item if it tends to gain value, and to take no action in the remaining cases. They often assume that the future stock market state can be, more or less accurately, predicted on the basis of past observations. Naturally, there are many various methods of financial data analysis widespread [1], [11], [14], which attempt to detect trends and discover contexts leading to the occurrence of particular events such as falls and rises in stock prices. Using these methods financial analysts and stock market traders make trading decisions.

In order to formalize financial analysis methods, the concept of *a stock market trading rule* may be introduced. Let K_t denote information available at time t, herein referred to as *a knowledge*, which, for instance, may represent historical price quotations. The stock market trading rule is a function f, which evaluates a trading decision $f(K_t) \in \{0.0 = \text{sell}, 0.5 = \text{do nothing}, 1.0 = \text{buy}\}$ on the basis of the knowledge K_t available at time t. Naturally, the function f may be defined in a variety of ways.

Let d denote a number of considered trading rules f_1, f_2, $\ldots$, f_d. For a given stock, at a given time instant t all these d functions may be evaluated producing a data vector $x_t = (x_{t1}, x_{t2}, \ldots, x_{td})$. This is a vector composed of values 0.0, 0.5 and 1.0 appropriately. Taking N consecutive time instants in such a way a data matrix X of size $N \times d$ is obtained. The i-th column X_i of the matrix X corresponds to the i-th variable (trading rule), and the j-th row x_j corresponds to the observation (results of trading rules) at the j-th instant of the time period.

This paper concerns a large set of trading rules observed over long time periods. Evaluation is performed on a set of 350 trading rules computed on five data sets. Each data set consists of a financial time series from the Paris Stock Exchange, which includes daily price quotations of a given stock over a period of about 1300 time instants from January 2, 1998 until May 12, 2003.

Details on input data are presented in Table 1. The first column, d, denotes the number of trading rules, and the second column, N, denotes the length of the time period. Although there is 350 trading rules, d can be less than 350, because, for each stock, columns with constant values in all the N time instants are removed from input data. The time period is the same for all experiments. However, the number of observations N can also differ, because for some stocks no price quotations were recorded on some specific days.

Stock	d	N
AXA	348	1302
Credit Lyonnaise	350	1290
Peugeot	348	1292
Renault	348	1292
Sodexho	348	1290

Table 1: Input data summary.

The large size of data, variables as well as observations, poses a major problem for many currently available methods.

3 Data preprocessing

Data preprocessing focuses on investigating dependencies among trading rules and reducing data dimensionality [10]. The method of dependency detection and dimensionality reduction is based on principal component analysis [6].

The analysis is performed using the correlation matrix $R = \{r_{ij}\}$, $i, j = 1, 2, \ldots, d$, calculated for the data matrix X after subtracting off the mean from its data vectors. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ denote eigenvalues and $v_1, v_2, \ldots, v_d$ corresponding eigenvectors of the matrix R. Since the matrix $V = [v_1 \; v_2 \; \ldots \; v_d]$ is orthogonal, the data matrix X may be transformed to a matrix $Y = XV$ and reproduced later as $X = YV'$. Columns $Y_1, Y_2, \ldots, Y_d$ of the matrix Y describe new variables that are uncorrelated and $\text{cov}(Y) = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$. However, the new variables may not return 0.0, 0.5 and 1.0, thus they may not correspond to any trading rules. Although the meaning of new variables is not immediate, they can be treated as a kind of financial indicators.

Since $X = YV'$, each original variable X_i may be reproduced as a linear combination of variables $Y_1, Y_2, \ldots, Y_d$ with linear coefficients defined by the vector v_i. However, my previous investigations showed that all the 350 trading rules considered in this paper are strongly correlated and all of them may be effectively reproduced by 150 first principal components [10], i.e. each original variable X_i may be reproduced as a linear combination of variables $Y_1, Y_2, \ldots, Y_{150}$, while variables $Y_{151}, Y_{152}, \ldots, Y_{350}$ are redundant.

Data preprocessing enables not only to reduce data dimension, i.e. to describe each data vector x_j by 150 variables $Y_1, Y_2, \ldots, Y_{150}$ instead of 350 variables $X_1, X_2, \ldots, X_{350}$, but also to characterize each variable X_i by 150 linear coefficients defined by 150 first coordinates of the vector v_i.

4 Self-organizing maps

After data preprocessing, each variable X_i corresponding to the trading rule f_i is characterized by 150 coefficients, on the basis of which the variables are grouped. Input data for clustering methods consists of d initial variables (new observations) described by 150 coefficients (new traits).

Trading rules are clustered using Kohonen self-organizing maps. The SOM consists of a number K of 150-dimensional vectors, referred to as codebook vectors, and a topology structure, such as a regular two-dimensional hexagonal lattice, which defines a neighborhood relation over codebook vectors. Codebook vectors are connected to adjacent ones by this neighborhood relation forming an elastic net embedded in the data space. During the training process, the SOM is adjusting to input data so that it folds onto the cloud of input data vectors.

Initially, codebook vectors are randomly placed on the plane defined by the two first principal components of the input data matrix. In the training process, an input data vector is randomly chosen from input data. Distances between this vector and all the codebook vectors are evaluated and the closest codebook vector, so-called the best matching unit, is chosen. Next, codebook vectors are updated. The BMU and its neighbors are moved closer to the input data vector. This process is repeated until the SOM fits into input data well enough or a given number of iterations is exceeded.

Finally, the SOM should be properly spread on input data vectors. Each input data vector is assigned to the nearest codebook vector. Thus, input data vectors assigned to the same codebook vector form a group.

In experiments, SOM with 88 codebook vectors and a hexagonal lattice sized 8×11 is generally used, but other structures with a number K of codebook vectors ranging from 40 to 160 were also studied.

5 Clustering of codebook vectors

Codebook vectors define a first partition of trading rules. In order to improve the clustering further, codebook vectors are grouped in a few classes, which introduces a second partition of trading rules. Naturally, the second partition is a generalization of the first one.

First, the K-means algorithm is used to divide codebook vectors into k classes, for $k = 1, 2, \ldots, K$ in turn. Next, a given validity index is evaluated for each clustering. Finally, the clustering with the optimal value of the validity index is chosen.

Several validity indices, such as the Dunn index [3], the Davies-Bouldin index [2], the silhouette index [12], the C index [5] as well as the Goodman-Kruskal index [4], are considered. It is worthy noticing that the computing time necessary to evaluate some of these indices for such a large data is very long (for instance dozens of minutes for the Goodman-Kruskal index).

6 Experiments

Experiments were performed on five financial time series from the Paris Stock Exchange described in Table 1. Constructing and training of SOM were carried out using the SOMToolbox, a free Matlab package [13].

Table 2 shows a summary of the initial clustering generated by SOM for each of the stocks considered. The second column contains the number of classes defined by the codebook vectors derived. Next columns contain values of the Dunn index, the Davies-Bouldin index, the silhouette index, the C index and the Goodman-Kruskal index respectively. It is worthy noticing that the number of classes may be less than the number of codebook vectors, because no input data vectors may be assigned to some codebook vectors.

Stock	Classes	D	D-B	s	C	G-K
AXA	83	0.74	3.98	0.43	1.12	0.83
Credit Lyonnaise	88	0.82	2.03	0.56	0.69	0.67
Peugeot	82	0.67	3.11	0.32	1.13	0.81
Renault	78	0.86	2.41	0.68	1.42	0.74
Sodexho	76	0.77	1.97	0.49	1.91	0.69

Table 2: Summary of variable clustering using SOM (the first partition).

Table 3 presents a similar summary related to the clustering of codebook vectors. This summary concerns only codebook vectors, not the input data vectors themselves. One can see that validity index values seems to be significantly better in Table 3 than in Table 2, because the clustering concerns a much smaller number of vectors (only 88 codebook vectors instead of 350 input data vectors).

Stock	Classes	D	D-B	s	C	G-K
AXA	23	0.92	0.68	0.91	0.59	0.91
Credit Lyonnaise	22	0.96	0.73	0.86	0.59	0.97
Peugeot	21	0.82	0.61	0.92	0.64	0.93
Renault	29	0.92	0.74	0.82	0.76	0.92
Sodexho	21	0.94	0.67	0.89	0.81	0.89

Table 3: Summary of codebook vectors clustering.

Each clustering of codebook vectors implies a corresponding clustering of input data vectors and consequently the second partition of stock market trading rules. Table 4 shows a summary of the final clustering obtained by merging certain groups defined by codebook vectors. It is easy to see that although validity index values juxtaposed in Table 4 are worse than in Table 3, the final clustering is better than the initial one described in Table 1. Thus, merging certain groups defined by codebook vectors improved the initial clustering generated by SOM.

Finally, Figure 1 presents an example of clustering - the partition of stock market trading rules into 21 classes evaluated on the data set concerning Peugeot using the Davis-Bouldin index. Each hexagon represents a map unit of the SOM corresponding to a codebook vector. A label inside the hexagon denote a number of trading rules assigned to the given codebook vector.

Stock	Classes	D	D-B	s	C	G-K
AXA	23	0.77	0.98	0.68	1.63	0.91
Credit Lyonnaise	22	0.86	1.01	0.76	0.99	0.96
Peugeot	21	0.91	0.89	0.73	1.21	0.93
Renault	29	0.83	1.71	0.58	1.42	0.82
Sodexho	21	0.82	1.07	0.69	0.91	0.84

Table 4: Summary of variables clustering using K-means (the second partition).

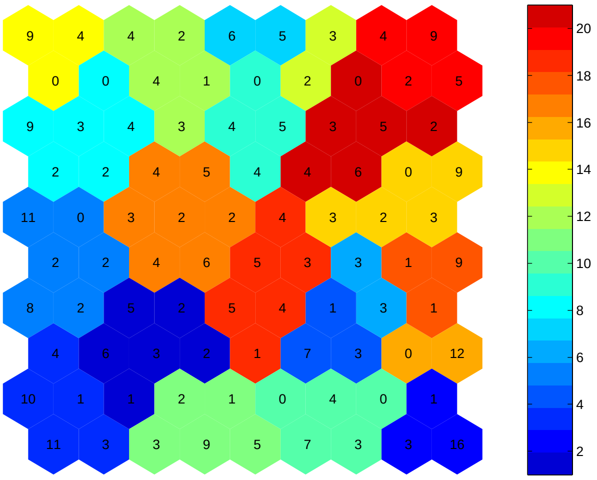


Figure 1: An example of clustering.

Hexagons are colored according to the codebook vector clustering obtained by the K-means algorithm.

Experiments show that, although no perfect clustering is obtained, the approach is relatively efficient.

All these experiments were computed in acceptable time as a result of efficient data preprocessing. Otherwise, if each trading rule was characterized by all the 350 coefficients instead of 150 coefficients after dimensionality reduction, the process of clustering would be much more time-consuming.

Although the initial partition generated by SOM usually detects some similarities between trading rules, the codebook clustering and the second partition significantly improve the quality of results.

7 Conclusions

In this paper, clustering of large number of stock market trading rules were presented. Data preprocessing significantly reduced data dimensionality facilitating and shortening computations. Consequently, each of 350 trading rules was characterized by 150 coefficients evaluated on the basis of the results of the trading rule observed over a period of about 1300 time instants. According to these coefficients, stock market trading rules were divided into several classes so that all the trading rules within the same class led to similar trading decisions in the same stock market conditions. The number of classes depended on the clustering validity index chosen. However, results were similar in some cases.

The approach proposed in this paper may be used in a few manners. For a given set of trading rules, clustering may be carried out on large sets of observations over long time periods to detect general dependencies among these trading rules. However, especially in real-time expert systems, general dependencies sometimes are insufficient. Thus, clustering may be carried out on sets of observations over shorter time periods to detect temporary relations between these trading rules over the specific time period, which may not hold in general.

The large number of trading rules is often a bottleneck for expert systems, which construct expertise based on selected trading rules. Clustering trading rules may enable optimization and a notable decrease in computing time, because experts might be discovered in two steps: the system might build a rough expert composed of selected representatives of some trading rule classes and then it might tune the rough expert by replacing each representative with an other trading rule selected from the class. The approach proposed may be applied in stock market decision support expert systems, such as described in [8], [9].

References

[1] Colby W., Meyers T. (1990). *The encyclopedia of technical market indicators.* Down Jones-Irwin.

[2] Davies D. L., Bouldin D. W. (1979). *A cluster separation measure.* IEEE Transactions on Pattern Recognition and Machine Intelligence, 224 – 227.

[3] Dunn J. C. (1974). *Well separated clusters and optimal fuzzy partitions.* Journal of Cybernetics, 95 – 104.

[4] Goodman L., Kruskal W. (1954). *Measures of associations for cross-validations.* Journal of American Statistical Associacion, 732 – 764.

[5] Hubert L., Schultz J. (1976). *Quadratic assignment as a general data-analysis strategy.* British Journal of Mathematical and Statistical Psychologie, 190 – 241.

[6] Jolliffe I. T. (1986). *Principal component analysis.* Springer.

[7] Kohonen T. (1997). *Self-organizing maps.* Series in Information Science, Springer.

[8] Korczak J., Lipinski P., Roger P. (2002). *Evolution strategy in portfolio optimization.* Artificial Evolution, P. Collet (ed.), Lecture Notes in Computer Science **2310**, Springer, 156 – 167.

[9] Korczak J., Roger P. (2002). *Stock timing using genetic algorithms.* Applied Stochastic Models in Business and Industry, 121 – 134.

[10] Lipinski P. (2003). *Dependency mining in large sets of stock market trading rules.* Proceedings of 10th International Multi-Conference on Advanced Computer Systems, Technical University of Szczecin, Szczecin, Poland, 2003 (electronic edition, to appear also in a Kluwer edition).

[11] Murphy J. (1998).*Technical analysis of the financial markets.* NUIF.

[12] Rousseeuw P. J. (1987). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 53 – 65.

[13] Vesanto J., Himberg J., Alhoniemi E. Parhankangas J. *SOM Toolbox for Matlab 5.* SOM Toolbox Team, Helsinki University of Technology, Finland (http://www.cis.hut.fi/projects/somtoolbox).

[14] Weigend A. S., Gershenfeld N. A. (1993). *Time series prediction: forecasting the future and understanding the past.* Addison-Wesley.

Address: P. Lipinski, Institute of Computer Science, University of Wroclaw, Wroclaw, Poland
LSIIT, CNRS, Universite Louis Pasteur, Strasbourg, France
E-mail: lipinski@ii.uni.wroc.pl

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

OPTIMAL SEPARATION PROJECTION

Karsten Luebke and Claus Weihs

Key words: Projection pursuit, visualization, LDA, simulated annealing.

COMPSTAT 2004 section: Dimension reduction.

Abstract: In this work we propose a new projection algorithm that can be used for dimension reduction in a classification problem with more than two classes. The projection aims at making the smallest distance between the means of two classes as large as possible. By this the error rate of a Linear Discriminant Analysis in the projected space can be reduced compared to the use of the discriminant coordinates.

1 Introduction

Dimension reduction of high dimensional data is helpful in statistical work. For example it can be useful for visualization of data or to avoid problems connected with overfitting and unstable estimates. Linear combinations of the original data are a good candidate for a projection because they are easy to handle and can sometimes be interpreted in terms of the subject matter at hand. Such a dimension reduction may be used as a preprocessing step before a further analysis of the data. Often this further analysis includes the ubiquitous challenge of classification. A vast amount of methods have been developed for this problem. One of the oldest methods in the statistical literature is the Linear Discriminant Analysis (LDA) developed by R.A. Fisher in 1936. Despite its age, the (relative) simple classification method of LDA does perform well even in situations, where the underlying premises like normal distributed data are not met (see for example [2, page 89]). Furthermore LDA also integrates an intrinsic dimension reduction method: It projects the data on so-called "discriminant" or "canonical" coordinates by a maximization of the between-class variance (of the explanatory variables) relative to the within-class variance in the projected space.

Motivated by the idea behind Support Vector Machines (SVM) where a separating hyperplane is constructed which creates the biggest margin between two classes we look for linear projections of the explanatory variables that maximize the minimum distance between any two class means. After such a projection an ordinary LDA is performed. By the proposed method we hope to combine two goals of a discriminatory analysis: We want to improve the allocation of objects by avoiding overfitting or unstable estimates. And second we want to maximally distinguish (separate) the groups for a better descriptive analysis.

This paper is organized as follows: In section 2 we introduce the basic notation as well as LDA and the projection used therein. In section 3 the new "Maximum-Minimum-Separation Projection" (MMSP) method is explained and related to other projection methods. In section 4 the performance of the new method is compared to LDA in a real-life econometric example. The paper is concluded in section 5.

2 Linear discriminant analysis

In this paper we use the following notation:

- n number of observations.
- n_i number of observations in class i.
- k number of different classes (groups, $k > 2$) .
- p number of variables.
- $X \in \mathbb{R}^{n \times p}$ matrix of predictor variables.
- $\bar{x}$ estimator for overall mean vector.
- $\bar{x}_i$ estimator for mean vector of class i.
- $B = \frac{1}{k-1} \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$ estimator for between-group covariance matrix.
- $W = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{x_j \in \text{class}_i} (x_j - \bar{x}_i)(x_j - \bar{x}_i)'$ estimator for within-group covariance matrix.
- $G \in \mathbb{R}^{p \times s}, \quad s \leq p$ projection matrix of X.

For simplicity we assume that the a-priori probabilities of the classes are equal.

2.1 Linear discriminant projection

Linear Discriminant Analysis is a statistical method for classification. In LDA the classification is based on the calculation of the posteriori probabilities of the different classes of a new observation. The class with the highest posteriori probability is chosen. To calculate the posteriori probability it is assumed that the data comes from a multivariate normal distribution where the classes share a common covariance matrix but have different mean vectors. So in a LDA the mean vector of each class and the common covariance matrix have to be estimated. Allocation is done with an argmax rule (1):

$$C(x) = \arg\max_i \{x' W^{-1} \bar{x}_i - \frac{1}{2} \bar{x}_i' W^{-1} \bar{x}_i\}. \tag{1}$$

It is well known ([4, page 89]) that discrimination is the same in the original space as in terms of the $r = \min(k-1, p)$-canonical (discriminant) variables. These canonical variables $Z = XG$ are linear projections of the original variables where the projection is found by subsequently maximizing the ratio

$$\frac{\gamma_j' B \gamma_j}{\gamma_j' W \gamma_j}, \tag{2}$$

($\gamma \in I\!R^p$) under the constraint

$$G'WG = I_r, \tag{3}$$

with $G = (\gamma_1, \ldots, \gamma_r)$. So for visualization or a dimension reduction to a more parsimonious model the data can be linear transformed to

$$Z_{(s)} = XG_{(s)}, \tag{4}$$

where $G_{(s)} = (\gamma_1, \ldots, \gamma_s), s \leq r$. If $s < r$, only the first s discriminant coordinates are used. The case $s = 2$ is interesting as then a 2 dimensional visualization of the data is possible which may lead to a better understanding of the data. If $s = r$, a LDA of the projected data $Z_{(r)}$ will lead to the same result as a LDA of X.

3 Maximum-minimum-separation projection

In order to improve the separation between the classes we propose a different criterion to find a projection $G_{(s)}$. In contrast to (2) which maximizes the variance of the class means and hence tends to make the large distances from the overall mean as large as possible ([4, page 93]) we aim at making the smallest difference between any two classes as large as possible. In case of $k = 2$ classes the two different optimization criteria match each other so we are only interested in the case of more then two classes. Heuristically maximizing the minimum distance may reduce the expected error rate for future observations as errors are more likely to occur when classes are close to each other. Under the assumption that the projected data follows a multivariate normal distribution with a common covariance matrix within each class the means are the sufficient statistic for the separation of the data – when the covariance matrix is assumed to be known or fixed. So to achieve the best margin (as in SVM) the minimum distance between any two classes must be maximized. Formally G is found as

$$G = \arg\max_G \mathit{mindist}(G), \tag{5}$$

with

$$\mathit{mindist}(G) := \min ||\bar{x}_i'G - \bar{x}_j'G||^2, \quad i, j = 1, \ldots, k, i \neq j. \tag{6}$$

The derivative of (6) need not to exist at points where the (means of) classes which are closest to each other switch. These points may also lead to local maxima in the objective function $\mathit{mindist}(\cdot)$. To illustrate this consider the following simple example: We have three classes in two dimensions, where the three class means are given by $\mu_1 = (0,0)', \mu_2 = (1,0)', \mu_3 = (0,1)'$ and the common covariance matrix is $\Sigma = W = I_2$. Then the one-dimensional projection matrix G equals a column vector. In order to fulfill the side condition (3) all feasible vectors can be represented by $(\sin(\alpha), \cos(\alpha))'$. Figure 1 shows $mindist(\alpha)$ for $-\pi \leq \alpha \leq \pi$. It can be seen that there are local maxima and that the derivative at the optimal points does not exist.

To overcome these problems we use Simulated Annealing (SA, see for example [7]) of vectorized projection matrices G for global maximization of (6). SA turned out to be flexible and successful in quite a lot optimization problems. To achieve comparable results to the use of discriminant coordinates the side-condition (3) is used in MMSP as well. Note that the side-condition implies that the euclidian distance in (6) is also the Mahalanobis distance in the projected space. After the SA algorithm generated a trial point (corresponding to a matrix H) this point is adopted to fulfill the constrain (3): Let $W = V'V$ be the Cholesky decomposition of the within-group covariance matrix. Than let $G = HR^{-1}$ with $VH = QR$, the QR decomposition of VH. This is done prior to the calculation of the objective function (6) of the adopted trial point G. The vectorized projection matrix of a LDA is a starting vector for the simulated annealing process.

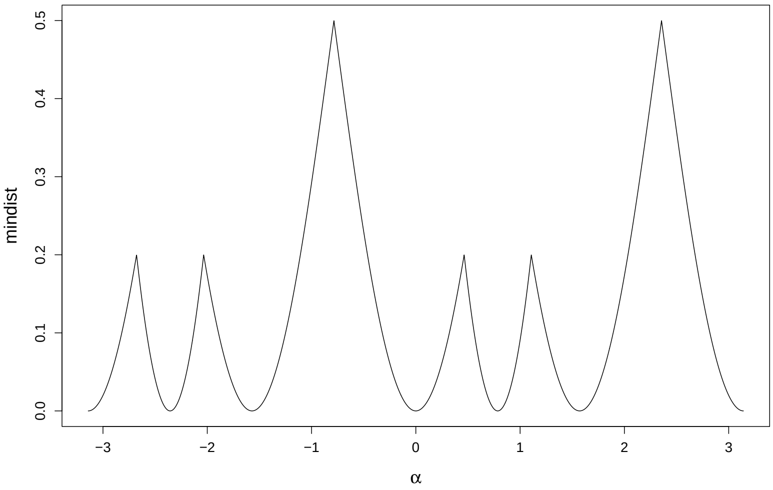


Figure 1: Two dimensional example of $mindist(\cdot)$.

The implemented algorithm is a stochastic version of the well-known Nelder-Mead Simplex method and based on [5]. The Nelder-Mead (or downhill simplex) method is transforming a simplex of $m + 1$ points in an m dimensional problem. The functional values are calculated and the worst point is reflected through the opposite face of the simplex. If this trial point is best a further expansion of the new simplex is tested. If the function value is worse than the second highest point the simplex is contracted. If no improvement at all is found the simplex is shrunk towards the best point. The implementation of Simulating Annealing can be summarized as follows:

1. Build $(rp+1)\times(rp)$ start simplex by the vectorized projection matrix of LDA and I_{rp}. The start temperature was chosen as $t_0 = r^2$.
2. Subtract from the function values $f = mindist$ of the points in the simplex (see (6)) a random number so that $f_{temp} = f - t|log(u)|$, where u is uniformly distributed over $(0, 1)$.
3. According to the Nelder-Mead transition function generate a trial point using the temporary function values f_{temp}.
4. Adapt the trial point to the side conditions (3) by a QR decomposition.
5. Accept the new trial point according to Nelder-Mead with the function value $f_{temp}(trial) = f(trial) + t|log(u)|$ of the trial point. So a better trial point is always accepted and a worse trial point is accepted with a certain probability.
6. Repeat step 2-5 250 times. Reduce the temperature to $t_{new} = 0.8 \cdot t_{old}$.
7. Repeat step 2-6 150 times.

Classification is done with a Linear Discriminant Analysis in the projected space, that is of $Z = XG_{(s)}$ with $G_{(s)}$ found by MMS projection. By the central limit theorem Z, as it is a linear combination of X, tends to be normally distributed. So the assumption of normal distributed data is not so critical in the projected space. A further advantage is that the projection also works in situations when the covariance matrices of the explanatory variables are singular, as the matrix inversion included in LDA (see (1)) is done in the (lower dimensional) projection space.

3.1 Related work

Different approaches exists for a dimension reduction of high dimensional data prior to a LDA. For example a Principal Component Analysis (PCA) can carried out beforehand. Here it is hoped that the dimensions with high variation in the explanatory variables X also carry most of the information for separation of the different classes. But it turned out that such a PCA is of limited value for the discriminant analysis ([4, page 197]). More recently [1] propose the use of an Partial Least Squares Analysis (PLS) as a preprocessing step. This is motivated by the relation between LDA and a Canonical Correlation Analysis (CCA) and the close relation between CCA and PLS. [1] claims that PLS will perform better then PCA. Both methods can be compared to MMSP as they are heuristically motivated but the optimality criterion is different for PLS and MMSP (covariance vs. margin).

A different approach is proposed by [6] with their Minimum Error Classification (MEC1). There a projection is looked for which directly minimizes the error-rate of a LDA or QDA estimated by bootstrap methods, assuming normality in the projected space. The projection is also found by Simulated Annealing. A problem with the MEC1 procedure of [6] is that as the error-rate is discrete quite a lot of projection matrices lead to the same error-rate.

This causes the optimization to be more difficult. As it uses bootstrap methods it is much more computational time demanding than the new MMSP procedure.

4 Real world example

In the following the classification performance of MMSP is compared to LDA in a real world problem.

The data set consists of 13 economic variables with 157 quarterly observations from 1955/4 to 1994/4 (see [3]) of the German business cycle. The German business cycle is classified in a four phase scheme: upswing (label: 1), upper turning point (label: 2), downswing (label: 3) and lower turning point (label: 4). There were 6 complete cycles in the time period.

The prediction ability was tested by the leave-one-cycle out validation: One cycle was left out as a validation set, the other 5 cycles are used to train the method and then the misclassification rate was estimated on the validation set.

It is shown in [8] that in general LDA is among the best classifiers for this classification task. For example it clearly outperforms a SVM. Despite the fact that the observed group sizes vary the a-priori group probabilities are set equal for the analysis.

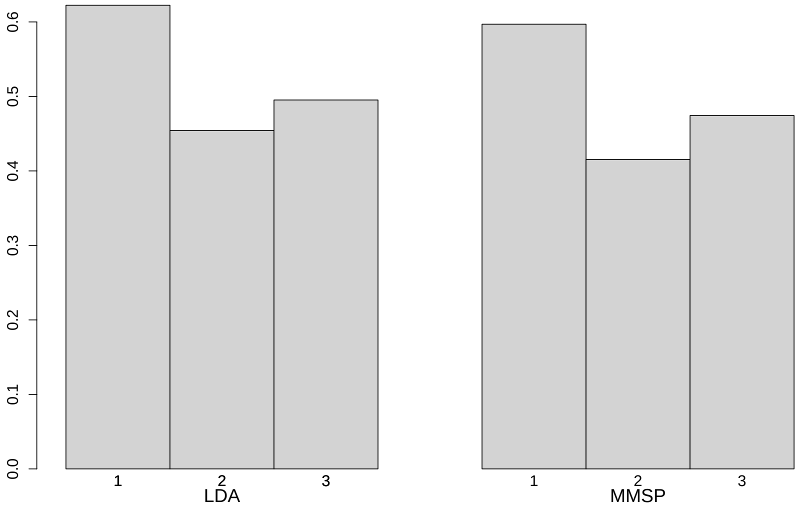


Figure 2: Error rates of LDA and MMSP for s =1, 2 and 3.

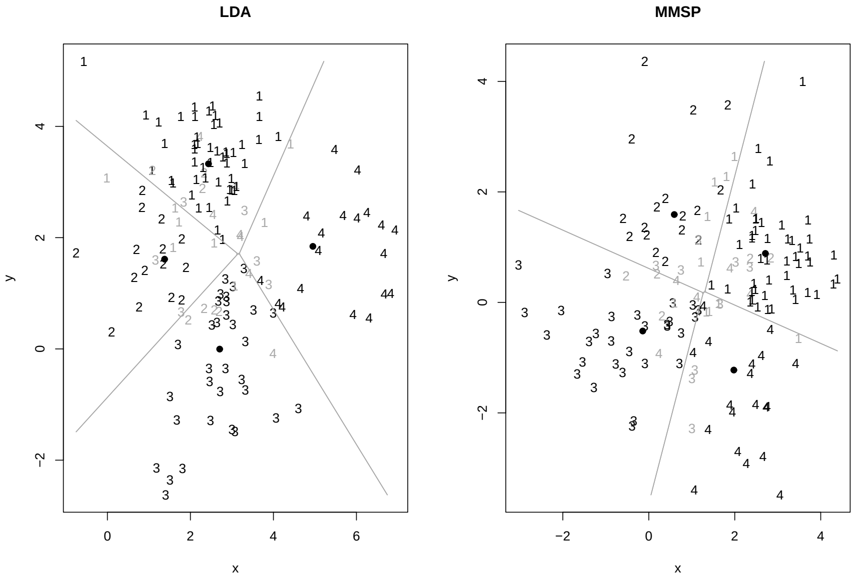


Figure 3: 2 Dimensional Partitions of LDA and MMSP.

The test error-rates of a LDA in the projected space where the projection is found by a LDA and MMSP is illustrated for the different projected dimensions in figure 2.

It can be seen in figure 2 that the dimension reduction by MMSP is outperforming LDA for every possible dimension of the solution space. Both LDA (0.45) and MMSP (0.41) have the lowest leave-one-cycle out error rate when the data is projected into a two-dimensional space. The error-rate improvement of MMSP is almost 10% of the error rate of the LDA.

In figure 3 the partitions of the whole data set are shown for LDA and MMSP (observations which are misclassified are plotted in grey color, the group means are marked by a solid circle).

Figure 3 shows that the separation of the classes upswing (1) and upper turning point (2) which are closest to each other has improved. The minimum distance between two class means is 2.01 with the discriminant coordinates found by LDA and 2.23 with the projection of MMSP. On the other hand the training error rate of the LDA is better (0.19 vs. 0.21).
The confusion matrices (see table 1) show the true classes versus the predicted classes of the projected data.

The misclassification performance improved mainly in the prediction of the upswing (class 1, 36 vs 42 true predictions). A surprising fact is that this improvement seems to be caused by less wrong predictions of the lower

true	Proj. LDA				Proj. MMSP			
	1	2	3	4	1	2	3	4
1	36	5	2	14	42	5	1	9
2	5	8	4	5	10	8	4	0
3	6	2	17	13	7	1	19	11
4	4	1	4	17	5	0	6	15

Table 1: Confusion matrices of LDA after 2-dim Projections.

turning point (4) when the upswing is the true class (9 vs 14) and not in the separation of the two closest classes (1 and 2).

5 Conclusion, outlook

The new Maximum-Minimum-Separation Projection turned out to be quite successful for dimension reduction in a classification problem. In the econometric example the test error-rate of a LDA after the MMS projection was significantly better then the classical LDA error-rate. This good result should be checked by a simulation study and on other data-sets.
Also a mathematical proof of the improved performance is part of the future research on this topic. The procedure may be extended to the case of unequal covariance matrices within each class (Quadratic Discriminant Analysis).

References

[1] Barker M., Rayens W. (2003). *Partial least squares for discrimination.* Journal of Chemometrics **17**, 166 – 173.

[2] Hastie T., Tibshirani R., Friedman J. (2001). *The elements of statistical learning.* Springer.

[3] Heilemann U., Münch H.J. (1996). *West german business cycles 1963-1994: A multivariate discriminant analysis.* CIRET-Conference in Singapore, CIRET-Studien 50.

[4] McLachlan G.J. (1992). *Discriminant analysis and statistical pattern recognition.* John Wiley & Sons.

[5] Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. (1992). *Numerical recipes in C.* 2nd ed., Cambridge University Press.

[6] Röhl M. C., Weihs C., Theis W. (2002). *Direct minimization of error rates in multivariate classification.* Computational Statistics **17**, 29 – 4.

[7] Salamon P., Sibani P., Frost R. (2002). *Facts, conjectures and improvement for simulated annealing*. Monographs on Mathematical Modeling and Computation. SIAM.

[8] Weihs C., Garczarek U.(2002). *Stability of multivariate representation of business cycles over time*. Technical Report 20, Sonderforschungsbereich 475, Universität Dortmund.

Acknowledgement: This work has been supported by the Collaborative Research Center 'Reduction of Complexity for Multivariate Data Structures' of the German Research Foundation (DFG).

Address: K. Luebke, C. Weihs, Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany

E-mail: `luebke@statistik.uni-dortmund.de`

COMPSTAT'2004 Symposium

TREE AND LOCAL COMPUTATION WITH THE MULTIPROPORTIONAL ESTIMATION PROBLEM

Francesco M. Malvestuto

Key words: Acyclic hypergraphs, iterative proportional fitting, minimum cross-entropy extension.

COMPSTAT 2004 section: Tree based methods.

Abstract: In this paper the biproportional estimating problem consisting in getting an estimate of an unknown two-way frequency distribution with given marginals is studied assuming that the proportions are proportional to a given two-way frequency distribution.

1 Introduction

The biproportional estimation problem [5] consists in getting an estimate $P(i,j)$ of an unknown two-way frequency distribution with given marginals, $p_1(i)$ and $p_2(j)$, assuming that $P(i,j)$ is proportional to a given two-way frequency distribution $q(i,j)$. An analogous problem arises in the economic analysis of inter-industry transactions [1]. In its multiproportional version, we are given a consistent system of distributions, say $\mathbf{S} = \{p_1, \cdots, p_n\}$, and a prior distribution q. Each p_i is defined on the state space of a set X_i of dicrete variables, and q is defined on the state space of a set W of discrete variables such that, if $V = \cup_{i=1,\cdots,n} X_i$, then V and W may overlap. We assume that every discrete variable has a finite set of values; moreover, for every susbset X of V, by $size(X)$ we denote the number of states of X. We want an estimate P of the (unknown) true probability distribution over $V \cup W$ that is 'multiproportional' to q. More precisely, the probability distribution P is a solution of the following constraint system, where $P[X_i]$ denotes the marginal of P with respect to X_i.

MULTIPROPORTIONAL ESTIMATION MODEL (MEM)

Marginality: $P[X_i] = p_i (i = 1, \cdots, n)$

Proportionality: There exist real-valued functions $F_1, \cdots, F_n$, where F_i is defined on the state space of X_i, such that $P(v, w) = F_1(x_1) \cdots F_n(x_n) q(w)$.

Under suitable hypotheses, we now state, a solution of the MEM not only exists but also is unique.

Let $Z = V \cap W$ and $\pi = \frac{q[Z]}{size(V-Z)}$. A distribution p over V such that $p[X_i] = p_i(i = 1, \cdots, n)$ is called an *extension* of $\mathbf{S}$; moreover, an extension p of $\mathbf{S}$ is a π-*extension* of $\mathbf{S}$ if $p(v) = 0$ for every state v with $\pi(v) = 0$. If there exists at least one extension (π-extension, respectively) of $\mathbf{S}$, we say that $\mathbf{S}$ is consistent (π-*consistent*, respectively). What it is requested for the existence of a solution of the MEM is that $\mathbf{S}$ be π-consistent. If this is the case, then the MEM has a unique solution which can be written as

$$P = p\frac{q}{q[Z]} \tag{1}$$

where p is the π-extension of $\mathbf{S}$ for which there exist functions $F_1, \cdots, F_n$ such that the following factorization

$$p(v) = F_1(x_1) \cdots F_n(x_n) q[Z](z) \tag{2}$$

holds for every state v of V with $p(v) \neq 0$. Equivalently, p is the *minimum-cross-entropy* (MCE, for short) π-extension of $\mathbf{S}$, that is, the π-extension of $\mathbf{S}$ that minimizes the cross-entropy $I(p, \pi)$ (or I-divercence or Kullback-Leibler distance). It should be noted that, if Z is empty or Z is contained in some X_i, then $I(p, \pi) = -H(p)$+const, where $H(p)$ is the *(Shannon) entropy* of p, so that p reduces to the *maximum-entropy* (ME, for short) extension of $\mathbf{S}$ and the factor $q[Z]$ in (2) can be omitted. From a computational point of view, the MCE π-extension and the ME extension of $\mathbf{S}$ can be found using the Iterative Proportional Fitting (IPF, for short) procedure [4], which is the multiproportional generalization of the Deming-Stephan algorithm. Each one of two extensions of $\mathbf{S}$ is the limiting distribution of a sequence $p^{(0)}, p^{(1)}, \cdots, p^{(r)}, \ldots$ where $p^{(r)}$ is obtained by fitting $p^{(r-1)}$ to the distribution p_i, for $r = hn + i$, $h \geq 0$ and $1 \leq i \leq n$. Finally, $p^{(0)}$ is set to π for the MCE π-extension of $\mathbf{S}$, and to $\frac{1}{size(V)}$ for the ME extension of $\mathbf{S}$. After computing the distribution p with the IPF procedure, the solution P of the MEM can be easily computed using (1). Needless to say, the dominant computational effort is given by the IPF procedure and each of its iteration cycles requires a number of arithmetic operations which is proportional to $n\ size(V) + \sum_{i=1}^{n} size(X_i)$ [2]. When p is the ME extension of $\mathbf{S}$, an efficient implementation of the IPF procedure based on tree and local computation can be found [2]. Therefore, if Z is empty or Z is contained in some X_i, then the solution P of the MEM can be efficiently found. In this paper we present an efficient implementation of the IPF procedure for computing the MCE π-extension of $\mathbf{S}$ for an arbitrary $Z, \oslash \subseteq Z \subseteq V$. To achieve this, we interpret the set family $\mathbf{H} = \{X_1, \cdots, X_n\}$, we call the *scheme* of $\mathbf{S}$, as a hypergraph with vertex set V. Some basic notions of hypergraph theory will be recalled in the next section.

2 Hypergraphs

A *hypergraph* [3] with vertex set V is a nonempty collection $\mathbf{H}$ of nonempty subsets of V, which are called *edges* of $\mathbf{H}$. $\mathbf{H}$ is *simple* if no edge is a subset of another edge. Given a subset A of V, the *subhypergraph* of $\mathbf{H}$ *induced* by A, denoted by $\mathbf{H}[A]$, is the simple hypergraph whose edges has the maximal sets in the family $A \cap X : X \in H$. A *path* is a sequence of edges that are pairwise nondisjoint. Two vertices of $\mathbf{H}$ are *connected* if they belong to two edges of $\mathbf{H}$ that are the ends of a path. The hypergraph $\mathbf{H}$ is connected if every two vertices are connected. The *components* of $\mathbf{H}$ are the subhypergraphs of $\mathbf{H}$ induced by its maximal sets of connected vertices. Two vertices of $\mathbf{H}$ are *separated* by a set B of vertices if they are in two components of $\mathbf{H} - B = \mathbf{H}[V - B]$. A *partial edge* of $\mathbf{H}$ is a nonempty set of vertices that is contained in some edge of $\mathbf{H}$. A partial edge B of $\mathbf{H}$ is a *divider* if there are two vertices of $\mathbf{H}$ that are separated by B but by no proper subset of B. By $\Delta(H)$ we denote the set of dividers of $\mathbf{H}$. A *cover* of $\mathbf{H}$ is a hypergraph C with vertex set V such that each edge of $\mathbf{H}$ is a partial edge of C. The *two-section* of $\mathbf{H}$, denoted by $[\mathbf{H}]_2$, is the simple graph with node set V where two nodes are adjacent if and only if they appear together in some edge of $\mathbf{H}$. A hypergraph $\mathbf{H}$ is *acyclic* if each clique (i.e., each nonempty set of pairwise adjacent nodes) of $[\mathbf{H}]_2$ is a partial edge of $\mathbf{H}$, and $[\mathbf{H}]_2$ is a chordal graph. If this is the case and $\mathbf{H}$ is simple, then $\mathbf{H}$ coincides with the *clique hypergraph* of $[\mathbf{H}]_2$, which has as its edges the maximal cliques of $[\mathbf{H}]_2$. Several other equivalent definitions of acyclicity exist [3]. Given a hypergraph $\mathbf{H}$, the problem of finding a minimum-cost acyclic cover of $\mathbf{H}$ is NP-hard. However, there exist efficient algorithms which find an acyclic cover F of $\mathbf{H}$ such that $F = H$ if and only if $\mathbf{H}$ is acyclic. These algorithms, called *zero-fill-in algorithms* [9], work with the two-section of $\mathbf{H}$ as follows: if $[\mathbf{H}]_2$ is chordal, then the output will be the clique hypergraph of $[\mathbf{H}]_2$; otherwise, the output will be the clique hypergraph of the chordal graph resulting by adding "fill-in" arcs to $[\mathbf{H}]_2$ which make it chordal. Acyclic covers of $\mathbf{H}$ obtained with this procedure will be referred to as *fill-in covers* of $\mathbf{H}$. A special acyclic cover of $\mathbf{H}$ is based on the notion of the compaction of $\mathbf{H}$ [2], we now recall. Two vertices of $\mathbf{H}$ are *tightly connected* if they are separated by no partial edge. Sets of pairwise tightly connected vertices of $\mathbf{H}$ are called *compacts*. Of course, each edge is a compact of $\mathbf{H}$. The *compact components* of $\mathbf{H}$ are the subhypergraphs of $\mathbf{H}$ induced by maximal compacts, and the *compaction* of $\mathbf{H}$ is the cover of $\mathbf{H}$ whose edges are its maximal compacts. The compaction of $\mathbf{H}$ has a number of nice properties: the compaction of $\mathbf{H}$ is acyclic, $\mathbf{H}$ is acyclic if and only if $\mathbf{H}$ coincides with its compaction, the dividers of the compaction of $\mathbf{H}$ are exactly the dividers of $\mathbf{H}$, and the compaction of $\mathbf{H}$ can be computed in polynomial time.

3 Computing the ME extension

In Section 2 we proved that in some cases the distribution p in formula (1) coincides with the ME extension of **S**. The following is an example.

Example 1. Consider a distribution system $S = p_1, \cdots, p_7$ over $V = \mathsf{ABCDEF}$ where $p_1, \cdots, p_7$ are over AB, AE, BC, BE, CF, DE, EF, respectively. Let q be a distribution over $W = \mathsf{BEGH}$. By (1), the solution P of the MEM can be expressed as $P(\mathsf{abcdefgh}) = p(\mathsf{abcdef})\frac{q(\mathsf{bdegh})}{q[\mathsf{BDE}](\mathsf{bde})}$ where p is the ME extension of **S** since $V \cap W = \mathsf{BE}$ is an edge of the scheme of **S**.

In this section we review the implementation of the IPF procedure given in [2] for computing the ME extension of **S**, which profits of tree and local computation. To this end, we now recall some definitions and results. Without loss of generality, we assume that the scheme **H** of **S** is a simple and connected hypergraph. A probability distribution p over V is *decomposable* by an acyclic hypergraph C on V if

$$p = \frac{\prod_{A \in C} p[\mathsf{A}]}{\prod_{B \in \Delta(C)} (p[\mathsf{B}])^{k_B}} \tag{3}$$

where k_B is the so-called adjusted replication numbers. The marginals $p[A]$, $A \in \mathbf{C}$, and $p[B]$, $B \in \Delta(\mathbf{C})$, will be referred to as the **C**-*components* of p.

Proposition 3.1. [8]. *Let* **S** *be a consistent system of distributions with scheme* **H**. *For every acyclic cover* **C** *of* **H**, *the ME extension of* **S** *is decomposable by* **C**.

Proposition 3.1 leads to an efficient method for computing the ME extension p of **S**: one first computes the **C**-components of p for some acyclic cover **C** of **H** and, then, applies formula (3). Let us distinguish two cases depending on whether **H** is or is not acyclic. In the former case, p is decomposable by **H** and the **H**-components of p are either explicitly or implicitly given in **S**; for example, if B is a divider of **H**, then the **H**-component $p[B]$ can be obtained by marginalizing p_i with respect to B, where i is chosen in such a way that X_i is a minimum-size edge of **H** that contains B. In the latter case, if **C** is an acyclic cover of **H**, then, as proven in [6], each distribution $p^{(r)}$, $r \geq 0$, involved in the IPF procedure is decomposable by **C**; moreover, the **C**-components of $p^{(r+1)}$ can be obtained from the **C**-components of $p^{(r)}$ with tree computation [2]. So, when the convergence is attained, the **C**-components of p are available. Such an implementation of the IPF procedure will be referred to as the *tree-IPF procedure* and it is not difficult to see [2] that computational cost of each iterative cycle of the tree-IPF procedure is proportional to $n \cdot \sum_{A \in C} size(A) + \sum_{i=1}^{n} size(X_i)$ so that, whenever $\sum_{A \in C} size(A) < size(V)$, the tree-IPF procedure is more efficient than the "brute-force" implementation of the IPF procedure. Note that the best choice for $\mathbf{C}$ is a minimum-size acyclic cover of $\mathbf{H}$; unfortunately, this problem is NP-hard in the general case: e.g., see [6]. However, with an appropriate choice of $\mathbf{C}$ local computation is viable and the computational costs can be reduced in a remarkable way. The key-notion for local computation is the 'collapsibility' of $\mathbf{S}$ we now recall. Let $\mathbf{S} = p_1, \cdots, p_n$ be a consistent distribution system with scheme $\mathbf{H} = \{X_1, \cdots, X_n\}$. If A is a subset of the vertex set of $\mathbf{H}$, the *subsystem* of $\mathbf{S}$ *induced* by A, denoted by $\mathbf{S}[A]$, is the (consistent) system of distributions with scheme $\mathbf{H}[A]$, where each distribution is obtained by marginalizing some distribution in $\mathbf{S}$. The distribution system $\mathbf{S}$ is *collapsible onto* A if the marginal of the ME extension of $\mathbf{S}$ with respect to A is the ME extension of the subsystem $\mathbf{S}[A]$.

Proposition 3.2. [8]. *The compaction of* $\mathbf{H}$ *is the minimal (with respect to covering) acyclic cover* $\mathbf{C}$ *of* $\mathbf{H}$ *such that every consistent distribution system* $\mathbf{S}$ *with scheme* $\mathbf{H}$ *is collapsible onto each edge of* $\mathbf{C}$.

Let p be the ME extension of $\mathbf{S}$. By Proposition 3.1, p is decomposable by the compaction $\mathbf{C}$ of $\mathbf{H}$ and, by Proposition 3.2, the marginal of p with respect to each edge A of the compaction of $\mathbf{H}$ can be computed "locally" by applying the tree-IPF procedure to $S[A]$ with $p^{(0)} = \frac{1}{size(A)}$. Moreover, the remaining $\mathbf{C}$-components $p[B],\ B \in \Delta(C)$, can be easily computed by exploiting the property that each divider of $\mathbf{C}$ is also a divider of $\mathbf{H}$. To sum up, the distribution p can be computed using the following algorithm.

ALGORITHM 1

```
(1) For each edge $A$ of $\mathbf{C}$
    (1.1) if $A$ is an edge of $\mathbf{H}$, say $X_i$, set $p[A] = p_i$; otherwise,
    (1.2) find a fill-in cover $F$ of the compact component $\mathbf{H}[A]$ of $\mathbf{H}$;
    (1.3) find the $F$-components of $p[A]$ by applying the tree-IPF procedure
          to $S[A]$ with $p^{(0)} = \frac{1}{size(A)}$;
    (1.4) set $p[A] = \frac{\prod_{E \in F} p[E]}{\prod_{D \in \Delta(F)} (p[D])^{k_D}}$.
(2) For each divider $B$ of $\mathbf{C}$, find a minimum-size edge $X_i$ of $\mathbf{H}$ containing
    $B$ and set $p[B]$ to the marginal of the distribution $p_i$ with respect to
    $B$.
(3) Compute $p$ using formula (3).
```

Example 1 (continued). The scheme of $\mathbf{S}$ is the connected hypergraph $\mathbf{H} =$ {AB, AE, BC, BE, CF, DE, EF}. The compaction of $\mathbf{H}$ is $\mathbf{C} =$ {ABE, BCEF, DE} and $\Delta(\mathbf{C}) =$ {BE, E}. Let us apply Algorithm 1.

(1) The edge ABE of $\mathbf{C}$ is not a partial edge of $\mathbf{H}$. The fill-in cover of the

compact component $\mathbf{H}[\mathsf{ABE}] = \{\mathsf{AB}, \mathsf{AE}, \mathsf{BE}\}$ of $\mathbf{H}$ is $\mathbf{F} = \mathbf{H}[\mathsf{ABE}]$. So, $p[\mathsf{ABE}]$ has one $\mathbf{F}$-component, namely, the whole distribution $p[\mathsf{ABE}]$, which is computed with the tree-IPF procedure. The edge BCEF of $\mathbf{C}$ is not a partial edge of $\mathbf{H}$. A fill-in cover of the compact component $\mathbf{H}[\mathsf{BCEF}] = \{\mathsf{BC}, \mathsf{BE}, \mathsf{CF}, \mathsf{EF}\}$ of $\mathbf{H}$ is $\mathsf{F} = \{\mathsf{BCE}, \mathsf{CEF}\}$. So, $p[\mathsf{BCEF}]$ has three $\mathbf{F}$-components, namely: $p[\mathsf{BCE}]$, $p[\mathsf{CEF}]$ and $p[\mathsf{CE}]$ which are computed with the tree-IPF procedure. After doing that, one has $p[\mathsf{BCEF}] = \frac{p[\mathsf{BCE}]\cdot p[\mathsf{CEF}]}{p[\mathsf{CE}]}$. The edge DE of $\mathbf{C}$ is also an edge of $\mathbf{H}$. So, $p[\mathsf{DE}] = p_6$.
(2) The divider BE of $\mathbf{C}$ is also an edge of $\mathbf{H}$. So, $p[\mathsf{BE}] = p_4$. The divider E of $\mathbf{C}$ is contained in the two edges BE and DE of $\mathbf{H}$. Assuming $size(\mathsf{B}) \leq size(\mathsf{D})$, we take $p[\mathsf{E}] = p_4[\mathsf{E}]$.
(3) The ME extension of $\mathbf{S}$ is computed as $p = \frac{p[\mathsf{ABE}]\cdot p[\mathsf{BCEF}]\cdot p[\mathsf{DE}]}{p[\mathsf{BE}]\cdot p[\mathsf{E}]}$.

4 Computing the π-MCE extension

In this section we first give a tree-IPF procedure for computing the π-MCE extension of $\mathbf{S}$ and, next, show how local computation can be introduced into the tree-IPF procedure in the case $\pi = \frac{q[Z]}{size(V-Z)}$. In order to obtain a tree-IPF procedure for computing the π-MCE, we need the following result which generalizes the above-mentioned result of [6] for an arbirary distribution π.

Theorem 4.1. *Let π be an arbitrary distribution over V and let $\mathbf{S}$ be a π-consistent distribution system over V with scheme $\mathbf{H}$. Let $\mathbf{C}$ be an acyclic cover of $\mathbf{H}$. If π is decomposable by $\mathbf{C}$, then the π-MCE extension of $\mathbf{S}$ is also decomposable $\mathbf{C}$.*

By Theorem 4.1, the π-MCE extension p of $\mathbf{S}$ is also the ME extension of its marginals $p[A], A \in C$, so that the $\mathbf{C}$-components of p can be computed using a tree-IPF procedure with zero-order approximation π. The point is that, for an arbitrary distribution π, the existence of a nontrivial acyclic cover of $\mathbf{H}$ by which π is decomposable is under question. However, in the MEM one has $\pi = \frac{q[Z]}{size(V-Z)}$ so that π is definitely decomposable by any acyclic cover $\mathbf{C}$ of the hypergraph $\mathbf{H}' = \mathbf{H} + \{Z\}$.

Example 2. Consider again the distribution system $\mathbf{S}$ of Example 1, but now q is a distribution over $W = \mathsf{BDEGH}$. By (1), the solution P of the MEM can be expressed as $P(\mathsf{abcdefgh}) = p(\mathsf{abcdef})\frac{q(\mathsf{bdegh})}{q[\mathsf{BDE}](\mathsf{bde})}$ where p is the π-ME extension of $\mathbf{S}$ with $\pi(\mathsf{abcdef}) = \frac{q[\mathsf{BDE}](\mathsf{bde})}{size(\mathsf{ACF})}$. In this case $\mathbf{H}' = \mathbf{H} + \{Z\} = \{\mathsf{AB}, \mathsf{AE}, \mathsf{BC}, \mathsf{BE}, \mathsf{CF}, \mathsf{DE}, \mathsf{EF}\} + \{\mathsf{BDE}\} = \{\mathsf{AB}, \mathsf{AE}, \mathsf{BC}, \mathsf{BDE}, \mathsf{CF}, \mathsf{DE}, \mathsf{EF}\}$. Consider the acyclic cover $\mathbf{C} = \{\mathsf{ABE}, \mathsf{BCEF}, \mathsf{BDE}\}$ of $\mathbf{H}'$, which has $\Delta(\mathbf{C}) = \{\mathsf{BE}\}$. Of course, the distribution $\pi = \frac{q[\mathsf{BDE}]}{size(\mathsf{ACF})}$ is decomposable by $\mathbf{C}$. So, by Theorem 3.1, the π-MCE extension p of $\mathbf{S}$ is decomposable by $\mathbf{C}$; that is, $p = \frac{p[\mathsf{ABE}]\cdot p[\mathsf{BCEF}]\cdot p[\mathsf{BDE}]}{(p[\mathsf{BE}])^2}$. Moreover, the $\mathbf{C}$-components of p can be computed with the tree-IPF procedure from the $\mathbf{C}$-components of π, namely, $\pi[\mathsf{ABE}] = \frac{q[\mathsf{BE}]}{size(A)}$, $\pi[\mathsf{BCEF}] = \frac{q[\mathsf{BE}]}{size(\mathsf{CF})}$, $\pi[\mathsf{BDE}] = q[\mathsf{BDE}]$ and $\pi[\mathsf{BE}] = q[\mathsf{BE}]$.

It is worth observing that, if Z is empty or a partial edge of $\mathbf{H}$, then $\mathbf{H}' = \mathbf{H}$ and p reduces to the ME of $\mathbf{S}$. In what follows, we assume that this is not the case. We now show that by taking $\mathbf{C}$ to be the compaction of $\mathbf{H}$' we can profit of local computation. First of all, note that, since Z is an edge of $\mathbf{H}'$, there is at least one edge of $\mathbf{C}$ that contains Z. By Theorem 3.1, p is the ME extension of the distribution system $\mathbf{S}' = \mathbf{S} \cup \{p[Z]\}$ with scheme $\mathbf{H}$' and, by Proposition 3.2, for each edge A of $\mathbf{C}$, $p[A]$ equals the ME extension of the induced distribution system $\mathbf{S}'[A] = \mathbf{S}[A] \cup \{p[A \cap Z]\}$. We say that an edge A of $\mathbf{C}$ is *sensitive* (to Z) if $A \cap Z \neq \emptyset$ and $A \cap Z$ is not a partial edge of $\mathbf{H}$, and that a divider B of $\mathbf{C}$ is *sensitive* (to Z) if each edge A of $\mathbf{C}$ containing B is sensitive. Consider now an edge A of $\mathbf{C}$ that is not sensitive, that is, $A \cap Z$ is either empty or a partial edge of $\mathbf{H}$. Then $\mathbf{S}'[A] = \mathbf{S}[A]$ and $p[A]$ is the ME extension of $\mathbf{S}[A]$ so that $p[A]$ can be computed as in Algorithm 1 (see Step 1). Analogously, if a divider B of $\mathbf{C}$ is not sensitive to Z, then $p[B]$ can be computed as in Algorithm 1 (see Step 2). At this point what remains to compute are the sensitive $\mathbf{C}$-components of p. Let $\mathbf{C}^*$ be the acyclic hypergraph whose edges are the sensitive edges of $\mathbf{C}$ and let A^* be vertex set of $\mathbf{C}^*$, that is, A^* is the union of the sensitive edges of $\mathbf{C}$. Of course, $p[A^*]$ is decomposable by $\mathbf{C}^*$ and its $\mathbf{C}^*$-components can be computed by applying the tree-IPF procedure to $\mathbf{S}[A^*]$ with zero-order approximation $\pi[A^*]$. After doing that, p can be computed from its $\mathbf{C}$-components using formula (3).

Example 2 (continued). First of all, note that the above acyclic cover $\mathbf{C} = \{\mathsf{ABE}, \mathsf{BCEF}, \mathsf{BDE}\}$ of $\mathbf{H}'$ is exactly the compaction of $\mathbf{H}'$. The edges ABE and BCEF of $\mathbf{C}$ as well as the divider BE of $\mathbf{C}$ are not sensitive so that the corresponding $\mathbf{C}$-components of p will be computed as in Example 1. Now $\mathbf{C}^* = \{\mathsf{BDE}\}$ and the $\mathbf{C}$-component $p[\mathsf{BDE}]$ is computed by applying the tree-IPF procedure to $\mathbf{S}[\mathsf{BDE}] = \{p_4, p_6\}$ with $\pi[\mathsf{BDE}] = q[\mathsf{BDE}]$.

References

[1] Bacharach M. (1970). *Biproportional matrices input-output change.* Cambridge, University Press.

[2] Badsberg J.-H. and Malvestuto F.M. (2001). *An implementation of the iterative proportional fitting procedure by propagation trees.* Computational Statistics & Data Analysis 37, 297-322.

[3] Beeri C., Fagin R., Maier D., and Yannakakis M. (1983). *On the desirability of acyclic database schemes.* J. ACM. 30, 479-513.

[4] Bishop Y.M.M., Fienberg S.E., and Holland P.W. (1975). *Discrete multivariate analysis.* MIT Press, Cambridge.

[5] Deming W.E. and Stephan F.F. (1940). *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known.* Ann. of Math. Statistics 11, 427-444.

[6] Jirousek R. and Preucil S. (1995). *On the effective implementation of the iterative proportional fitting procedure.* Computational Statistics & Data Analysis 19, 177-189.

[7] Malvestuto F.M. (2001). *A hypergraph-theoretic analysis of collapsibility for extended log-linear models.* Statistics & Computing 11, 155-169.

[8] Malvestuto F.M. and Moscarini M. (2000) *Decomposition of a hypergraph by partial-edge separators.* Theoretical Computer Science 237, 57-79.

[9] Tarjan R. E. and Yannakakis M. (1984). *Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce hypergraphs.* SIAM J. on Computing 13, 566-579.

Acknowledgement: This work was funded by the Italian MIUR under the project PRIN 2003: "WEB-based management and representation of spatial and geographic data".

Address: F.M. Malvestuto, Computer Science Department, "La Sapienza" University of Rome, Italy

E-mail: malvestuto@di.uniroma1.it

COMPSTAT'2004 Symposium

BOOTSTRAP TEST FOR THE EQUALITY OF NONPARAMETRIC REGRESSION CURVES UNDER DEPENDENCE

W.G. Manteiga and J.M. Vilar-Fernández

Key words: Hypothesis testing, regression models, nonparametric estimators, dependent data, bootstrap.

COMPSTAT 2004 section: Smoothing.

Abstract: The main objective of this paper is to study testing the equality of k regression curves under the assumption of the dependence.

1 Introduction

The comparison of several regression curves is an important problem of statistical inference. In many cases of practical interest, the objective consists in comparing regression functions of a response variable Y observed in two or more groups on an explanatory variable which is an adjustable parameter, for instance, time. In this paper we consider k regression models in fixed design given by

$$Y_{l,t} = m_l\left(x_t\right) + \varepsilon_{l,t},\ \ l = 1, \ldots, k \text{ and } t = 1, \ldots, n.$$

The points of the design are taken evenly spaced, that is, $x_t = t/n$, for $t = 1, \ldots, n$. The processes of random errors $\{\varepsilon_{l,t}\}$ are independent among themselves and each follows an ARMA(p_l, q_l) type dependence structure. The problem of testing the equality of k regression curves under the dependence assumption of the observations is therefore the main concern of this work. This is, to test the hypothesis

$$H_0:\ m_1 = \ldots = m_k \quad \text{versus} \quad H_1: m_l \neq m_j \text{ for some} \quad l, j \in \{1, \ldots, k\}.$$

So, we wish to test the equality of tendencies of k time series. In this problem it is important to take into account the existence of correlation among the errors. Ignoring this fact affects the power of the equality test used.

The problem of testing the equality of k regression functions by using nonparametric techniques has been broadly studied. Some relevant papers are King et al. [2], Dette and Neumeyer [1], Scheike [3], Hall et al. (1997) and Koul and Schick (1997), among others.

In a recent paper, Vilar-Fernández and González-Manteiga [4] studied the problem of checking the equality of k regression functions with dependent errors in a general context. The basic idea is using as test statistic a functional distance between nonparametric estimators of the regression functions. In particular, they proposed that H_0 was tested using the statistic

$$\hat{Q}_n = \sum_{l=2}^{k} \left(\sum_{s=1}^{l-1} \left(\int \left(\hat{m}_l(x) - \hat{m}_s(x) \right)^2 \omega_{l,s}\left(x\right) dx \right) \right), \tag{1}$$

where $\{\omega_{l,s}\left(x\right)\}$ are weight functions defined on the support of the design variables $C = [0,1]$. Without loss of generality, it is assumed that $m_l(x)$ is defined on $[0,1]$, and $\hat{m}_l$ is the nonparametric estimator of m_l, given by $\hat{m}_l(x) = \sum_{i=1}^{n} W_{l,i}\left(x\right) Y_{l,i}$, where $W_{l,i}\left(x\right) = W_{l,i}\left(x, h_n\right), \; i = 1, \ldots, n$, $l = 1, \ldots, k$ are Gasser-Müller weights with smoothing parameter $h = h_n$. Basically, $\hat{Q}_n$ is a consistent estimator of $Q = \sum_{l=2}^{k} \sum_{s=1}^{k-1} \int \left(m_l - m_s\right)^2 \omega_{l,s}$. Therefore, the null hypothesis is false if and only if $Q > 0$.

The asymptotic normality of the test statistic $\hat{Q}_n$ was obtained under general conditions. Vilar-Fernández and González-Manteiga [4] obtained the following result

$$\sqrt{n^2 h} \left(\hat{Q}_n - B_Q \right) \longrightarrow N\left(0, \sigma_Q^2\right) \quad \text{in distribution}, \tag{2}$$

$$B_Q = \frac{1}{n h_n} C_K \sum_{l=1}^{k} \left(\Gamma_l \sum_{s=1, s \neq l}^{k} \int \omega_{l,s} \right)$$

$C_K = \int K^2, \quad \Gamma_l = \sum_{j=-\infty}^{\infty} \nu_l\left(j\right) = E(\epsilon_{l,t}\epsilon_{l,t+j}), \quad l = 1, \ldots, k$, and

$$\sigma_Q^2 = 2k^2 \left(\int \left(K * K\right)^2 \right) \left(\begin{array}{c} \sum_{l=1}^{k} \Gamma_l^2 \int \left(\sum_{s=1, s \neq l}^{k} \omega_{l,s} \right)^2 \\ + \sum_{l=1}^{k} \sum_{s=1, s \neq l}^{k} \Gamma_l \Gamma_s \int \omega_{l,s}^2 \end{array} \right)$$

Now, using the limit distribution (2), the hypothesis H_0 is rejected at a significance level of α when

$$\sqrt{n^2 h} \left(\hat{Q}_n - \hat{B}_Q \right) > z_\alpha \hat{\sigma}_Q, \tag{3}$$

where z_α is such that $\Phi\left(z_\alpha\right) = 1 - \alpha$, with Φ the distribution function of the standard normal. $\hat{B}_Q$ and $\hat{\sigma}_Q^2$ are consistent estimators of B_Q and σ_Q^2, respectively.

From (3) it can be deduced that the test based on $\hat{Q}_n$ detects alternatives converging to the null hypothesis at a rate $\left(n\sqrt{h}\right)^{-1/2}$. So, the asymptotic approximation to the normal is very slow, for instance, for $h \approx n^{-\frac{1}{5}}$ the convergence speeds obtained are of the order of $n^{-\frac{1}{10}}$. For this reason, in similar problems, several authors have proposed resampling procedures to obtain the sampling distribution of the test statistics. In this work we present a bootstrap procedure as an alternative to the plug-in approximation given in (2) for obtaining the critical test points. Therefore, the distribution of $\hat{Q}_n$ is approximated by resampling, unlike the normal approximation.

In the second section we describe in detail the resampling method and present the most significant result that prove the validity and consistency of the proposed bootstrap. In the third section we present a small comparative simulation study between the test based on the asymptotic distribution and that obtained with the bootstrap algorithm.

2 Bootstrap test

In this paper we consider k regression models in fixed design given by

$$Y_{l,t} = m_l\left(x_t\right) + \varepsilon_{l,t},\ \ l = 1, \ldots, k \text{ and } t = 1, \ldots, n. \tag{4}$$

Without loss of generality, it is assumed that $m(x)$ is defined in $[0, 1]$. The points of the design are taken evenly spaced, that is, $x_t = t/n$, for $t = 1, \ldots, n$. The processes of random errors $\{\varepsilon_{l,t}\}$ are independent among themselves and each follows an ARMA(p_l, q_l) type dependence structure, i.e.

$$\varepsilon_{l,t} = \sum_{i=1}^{p_l} \phi_{l,i}\, \varepsilon_{l,t-i} + e_{l,t} + \sum_{j=1}^{q_l} \vartheta_{l,j}\, e_{l,t-j},\ \text{ with } t \in Z \text{ and } l = 1, \ldots, k, \tag{5}$$

where $\{e_{l,t},\, t \in Z\}$ is a sequence of independent random variables with zero mean, finite variance $\sigma^2_{l,e}$ and distribution function $F_{l,e}$. In addition, the series $\{\varepsilon_{l,t}\}$, $l = 1, \ldots, n$ is assumed to be stationary and invertible.

These regression models often arise by analyzing economical data, growth curves and, in general, whenever the observations are sequentially gathered in time. So, we wish to test the equality of tendencies of k time series.

The resampling mechanism consists of a simple and direct resampling of the original observations, taking into account that the processes of errors have an ARMA dependence structure. The algorithm follows the next steps.

Step 1. The test statistic $\hat{Q}_n$ is computed from the initial sample given by $\{(x_t, Y_{l,t}) : t = 1, \ldots, n;\ l = 1, \ldots, k\}$.

Step 2. Under the null hypothesis, nonparametric residuals $\hat{\varepsilon}_{l,t}$ are obtained by means of

$$\hat{\varepsilon}_{l,t} = Y_{l,t} - \hat{m}_{l,g_l}(x_t), \quad t = 1, \ldots, n;\ l = 1, \ldots, k,$$

where $\hat{m}_{l,g_l}(x)$ is the nonparametric estimator of the regression function computed from the sample $\{(x_t, Y_{l,t}) : t = 1, \ldots, n\}$ with auxiliary bandwidth g_l.

Step 3. A bootstrap sample of the residuals estimated in Step 2 is drawn as follows:

S3.A Estimates $\left(\hat{\phi}_l, \hat{\vartheta}_l\right)$, $l = 1, \ldots, k$, of the parameter vectors associated with the ARMA structure of the errors are constructed on the basis of the residuals estimated $\hat{\varepsilon}_{l,t}$.

S3.B Since the autoregressive representation of the error processes is invertible, estimates $\{\hat{e}_{l,t},\, t > r_l = \max(p_l, q_l)\}$ of the noise of the ARMA models can be obtained using $\{\hat{\varepsilon}_{l,t}\}$ and $\left(\hat{\phi}_l, \hat{\vartheta}_l\right)$, $l = 1, \ldots, k$. The estimated noise series is then centered as $\tilde{e}_{l,t} = \hat{e}_{l,t} - \hat{e}_{l,\cdot}$, for $t > r_l$, where $\hat{e}_{l,\cdot} = \frac{1}{n-r_l} \sum_{t=r_l+1}^{n} \hat{e}_{l,t}$, for $l = 1, \ldots, k$.

S3.C The empirical distribution of $\tilde{e}_{l,t}$, $\hat{F}_l(x)$, is derived for $l = 1, \ldots, k$.

S3.D From each $\hat{F}_l$, $l = 1, \ldots, k$, a sample of independent and identically distributed random variables $\left\{e^*_{l,-M}, \ldots, e^*_{l,-1}, e^*_{l,0}, e^*_{l,1}, \ldots, e^*_{l,n_0}\right\}$ is drawn, with $M > 0$.
The sequence $\left\{e^\star_{l,t}\right\}_{t=-M}^{n}$ is then used together with $\left(\hat{\phi}_l, \hat{\vartheta}_l\right)$ to generate a bootstrap sample of the error $\left\{\varepsilon^\star_{l,t}\right\}_{t=1}^{n}$, for $l = 1, \ldots, k$.

Step 4. A bootstrap sample $\{(x_t, Y^*_{l,t}) : t = 1, \ldots, n;\ l = 1, \ldots, k\}$ is obtained, making

$$Y^*_{l,t} = \hat{m}_{\cdot,g}(x_t) + \hat{\varepsilon}^*_{l,t}, \quad t = 1, \ldots, n;\ l = 1, \ldots, k,$$

where $\hat{m}_{\cdot,g}(x) = \frac{1}{k}\sum_{l=1}^{k} \hat{m}_{l,g}(x)$ is the nonparametric estimator of the regression computed from the total combined sample with auxiliary bandwidth g. The test statistic $\hat{Q}^*_n$ is now computed with this bootstrap sample.

Step 5. Steps S3.D and S4 are repeated a large number of times, say T, so that a sequence $\{\hat{Q}^*_{n,1}, \ldots, \hat{Q}^*_{n,T}\}$ is obtained. A bootstrap critical region of a significance level α is then given as

$$\hat{Q}_n > \hat{Q}^*_{n,([(1-\alpha)T])}, \tag{6}$$

where $[\cdot]$ represents the integer part and $\{\hat{Q}^*_{n,(i)}\}_{i=1}^{T}$ is the sample $\{\hat{Q}^*_{n,i}\}_{i=1}^{T}$ arranged in increasing order of magnitude.

Theorem 1 shows the validity of this resampling method. We suppose that the following assumptions are satisfied:

A.1. The regression functions m_l, $l = 1, \ldots, k$, are twice differentiable in $(0, 1)$ and their derivatives are bounded. The weight functions $\omega_{l,s}$ have a compact support contained in $(0, 1)$ and are differentiable with bounded derivative.

A.2. $E\left(|e_{l,t}|^{4+2\delta}\right) < \infty, \quad l = 1, \ldots, k,$ for some $\delta > 2$.

A.3. The kurtosis of the processes $e_{l,t}$ are zero for $l = 1, \ldots, k$.

A.4. The process of errors $\{\varepsilon_{l,t}\}$ follows an ARMA(p_l, q_l) structure as indicated in (5) for $l = 1, \ldots, k$. It is important to observe that although the design points $\{x_i\}$ approach each other when $n \to \infty$, the correlation between the errors only depends on the difference of the indices.

A.5. The smoothing weights used are Gasser-Müller type.

A.6. The kernel function K is continuously differentiable with compact support, and the smoothing parameter $h = h_n$ satisfies

$$nh_n^{3/2} \to \infty \quad \text{and} \quad n^{\frac{\delta+2}{2\delta+2}} h_n \to 0 \quad as \quad n \to \infty.$$

Theorem 1. *Assume that assumptions A.1-A.6 are satisfied and $n \to \infty$. Under the null hypothesis we have*

$$\sqrt{n^2 h}\,\left(\hat{Q}_n^* - B_{Q^*}\right) \xrightarrow{d^\star} N(0, \sigma_Q^2) \qquad \text{in probability,} \tag{7}$$

$$\begin{aligned} B_{Q^*} &= \frac{1}{nh_n} C_K \sum_{l=1}^{k} \left(\Gamma_l^* \sum_{s=1, s\neq l}^{k} \int \omega_{l,s} \right), \\ \Gamma_l^* &= \sum_{j=-\infty}^{\infty} \nu_l^* (j) = E^*(\varepsilon_{l,t}^* \varepsilon_{l,t+j}^*), \qquad l = 1, \ldots, k. \end{aligned}$$

By $\xrightarrow{d^\star}$ we denote the convergence in distribution under the resampling conditioned to the original sample. The same way, in what follows we denote by E^* and σ^* the mean and the variance, respectively, under the resampling.

Remark. Theorem 1 can be expanded to the case of k regression functions, where every curve follows a model of the type

$$Y_{l,i} = m_l\,(x_{l,i}) + \varepsilon_{l,i}, \quad i = 1, \ldots, n_l, \quad l = 1, \ldots, k, \tag{8}$$

and the design points follow different positive densities $\{f_l\}_{l=1}^k$ defined in $[0, 1]$. Hence the design points $\{x_{l,i}\}_{i=1}^{n_l}$ satisfy

$$\int_0^{x_{l,i}} f_l\,(t)\,dt = \frac{i}{n_l}, \quad i = 1, \ldots, n_l, \quad l = 1, \ldots, k.$$

About the sample size of each sample we assume the following: let $n = \sum_{l=1}^k n_l$ denote the total sample size, then

$$\frac{n_l}{n} = \pi_l + O\left(\frac{1}{n}\right), \quad \text{with} \quad \pi_l \in (0, 1), \quad l = 1, \ldots, k.$$

In this case, the bias and the variance of $\hat{Q}_n^*$ are

$$\begin{aligned} B_{Q^*} &= \frac{1}{nh_n} C_K \sum_{l=1}^{k} \left(\Gamma_l^* \sum_{s=1, s\neq j}^{k} \int \frac{\omega_{l,s}}{\pi_s f_s} \right) \\ \sigma_Q^2 &= 2 \left(\int (K * K)^2 \right) \left(\begin{array}{c} \sum_{l=1}^k \Gamma_l^2 \int \frac{\left(\sum_{s=1, s\neq l}^k \omega_{l,s}\right)^2}{\pi_s^2 f_s^2} \\ + \sum_{l=1}^k \sum_{s=1, s\neq l}^k \Gamma_l \Gamma_s \int \frac{\omega_{l,s}^2}{\pi_s f_s \pi_l f_l} \end{array} \right). \end{aligned}$$

Remark. The proof of Theorem 1 is too long, so to save space it is omitted. The complete proof is available from the authors.

3 Simulation study

In this section we present a simulation study in order to compare the two tests of equality of regression curves based on the statistical test $\hat{Q}_n$:

Test A is the plug-in version of the test which uses the asymptotic distribution given in (2).

Test B is the naive bootstrap given in (6).

Samples $\{(x_t, Y_{1,t}, Y_{2,t})\}_{t=1}^{n}$, of size $n = 100$, were simulated following the regression model given in (4), with $k = 2$ and $x_t = t/n$, $t = 1, \ldots, n$. The error processes were designed to follow the same AR(1) model: $\varepsilon_{l,t} = \phi\varepsilon_{l,t-1}+e_{l,t}$, $t \in Z$ and $l = 1, 2$,where $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ have the same distribution function, $N(0, \sigma^2)$, $\sigma^2 = 0.5$. In our study the regression functions $m_1(x) = 5\,(x - 0.5x)^2$ and $m_2(x) = m_1(x) + \Delta(x)$ were considered (under the null hypothesis $\Delta(x) = 0$) . In a first study a total of 500 trials were carried out with $\phi = 0.80$. Each one consisted in obtaining an initial sample and computing the values of the statistics $\hat{Q}_n$. Then, a bootstrap resample of size $T = 500$ was obtained using the bootstrap procedure exposed. Now, using the tests A and B, critical regions of significance levels $\alpha = 0.05$ and 0.10 were determined and the corresponding associated p-values were approximated. The bandwidth $g_l = 0.250$ was used to obtain the nonparametric residuals and to estimate Γ_l, and $h = 0.100$ was used to obtain $\hat{Q}_n$. These bandwidths were empirically chosen.

The results of our study are summarized in Table 1. In particular, the simulated rejection percentages of the proposed tests with level 10% and 5% are shown in Table 1 together with the average of the set of p-values obtained in the 500 trials. After simple inspection of Table 1, one can conclude that test B presents better performance than test A, although both tests obtain acceptable results.

A simulation study is carried out to study the influence of the smoothing parameter, h_n, and the behavior of the two tests as a function of h. We have simulated samples with $n = 100$, $\Delta(x) = 0$ (null hypothesis) and variable ρ. For each situation we have 50 trials, varying the smoothing parameter, h, from $h_{min} = 0.010$ to $h_{max} = 0.450$ in 0.010 steps. For each h the corresponding 50 $p-$values obtained using tests A and B were averaged. Let $A(h)$ and $B(h)$ denote the functions assigning to each h the average of the $p-$values obtained using test A and test B, respectively. In Figure 1 we represent these two functions for $\phi = 0.8$.

Again, in Figure 1, test B shows a better behaviour, $B(h)$ is closer to $Y = 0.5$ than $A(h)$. The small influence of the bandwidth in the computation of $B(h)$ is also observed.

Finally, we also carried out a study that confirms the better behaviour observed with the bootstrap test (test B) with respect to the normal test

	$\alpha = 0.10$		$\alpha = 0.05$		Mean $p-$value	
$\Delta(x)$	*Test A*	*Test B*	*Test A*	*Test B*	*Test A*	*Test B*
0	6.60	10.60	4.80	5.80	0.5763	0.4975
$0'5$	22.60	34.20	18.20	23.80	0.4102	0.2860
$1'0$	69.80	91.40	70.80	74.20	0.1204	0.0586
$1'5$	93.60	99.00	90.40	96.40	0.0200	0.0066
$0'5x$	9.80	17.20	8.20	9.60	0.5457	0.4283
$1'0x$	24.60	38.00	20.80	25.00	0.3832	0.2625
$1'5x$	48.60	66.00	39.40	53.20	0.2172	0.1212
$0'5\sin(2\pi x)$	12.60	19.00	8.60	22.40	0.5011	0.3836
$\sin(2\pi x)$	27.20	45.40	20.80	30.40	0.3189	0.1948

Table 1: Mean and standard deviation of the critical values of tests A and B and simulated rejection probabilities for levels $\alpha = 0.10$ and 0.05, with $\phi = 0.80$.

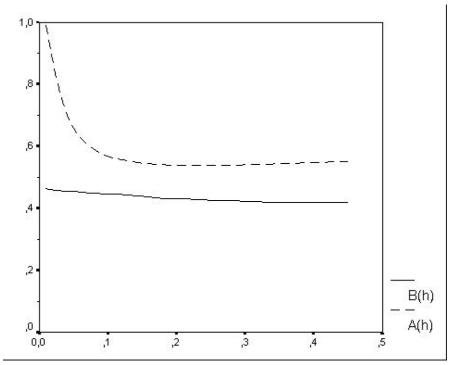


Figure 1: Graphs of $A(h)$ and $B(h)$ for $\phi = 0.8$.

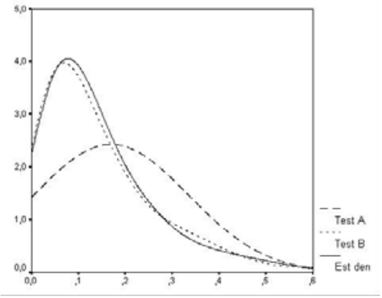


Figure 2: Graphs of limit density and desities of $\hat{Q}_n$.

(test A). We simulated 1000 samples of model with $m_1(x) = m_2(x)$ and $\phi=$ 0.80, and calculated the values of $\hat{Q}_n$. From these values, the density function of $\hat{Q}_n$ is estimated using the Rosenblatt-Parzen with bandwidth 0.05. From each of the simulated samples we obtained a resample using the algorithm described above and computed $\hat{Q}_n^*$. Now, we estimate the associated density function. The graphs of these two estimated densities and the theoretical density are represented in Figure 2, where it can be observed that the density of the bootstrap resamples follow the density of $\hat{Q}_n$ better than the theoretical density since the latter has a greater variance.

To study the influence of the smoothing parameter h_n on tests A and B, we computed the curves $A(h)$ and $B(h)$ when the regression function $m_2(x) = m_1(x) + \Delta(x)$, $\Delta(x) = 0, 0.5, 1.0$ and 1.5, with $\phi = 0.80$. Graphs of these functions are shown in Figure 3.

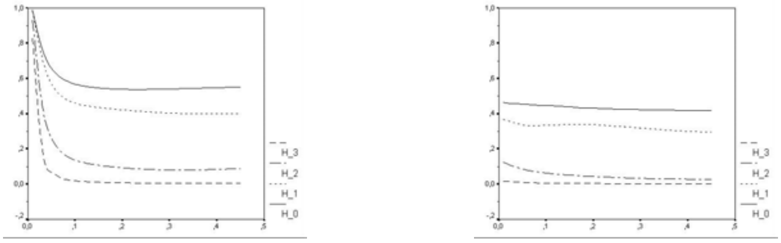


Figure 3: Graphs of $A(h)$ (left) and $B(h)$ (right) under H_0 and under the alternative hypothesis ($\Delta = 0.5, 1.0, 1.5$) with $\phi = 0.80$.

References

[1] Dette H., Neumeyer N. (2001). *Nonparametric analysis of covariance.* Ann. Stat. **29** (5), 1361 – 1400.

[2] King E.C., Hart J.D., Wehrly T.E. (1991). *Testing the equality of two regression curves using linear smoothers.* Stat. Probab. Lett., **12** (3), 239 – 247.

[3] Scheike T.H. (2000). *Comparison of nonparametric regression functions through their cumulatives.* Stat. Probab. Lett. **46**, 21–32.

[4] Vilar-Fernández J.M., González-Manteiga W. (2004). *Nonparametric comparison of curves with dependent errors.* To appear in Statistics.

Acknowledgement: This work has been partially supported by the DGICYT Spanish Grants BFM 2002-002665 and BFM2002-03213 (European FEDER support included) and XUGA Grants PGIDT01PXI10505PR, PGIDIT03PXI C20702PN, PGIDIT03PXIC10505PN.

Address: W.G. Manteiga, J.M. Vilar-Fernández, Department of Statistics, University of Santiago, 15706, SPAIN and Department of Mathematics, University of A Coruña, 15071, Spain

E-mail: `wences@zmat.usc.es, eijvilar@udc.es`

DO WE ALL COUNT THE SAME WAY?

Luboš Marek

Key words: Comparison of time series procedures, statistical software, Statistica, SPPS, SAS, SCA, Statgraphics, EViews, NCSS.

COMPSTAT 2004 section: Statistical software.

Abstract: This paper compares procedures used for time series analysis in different statistical software packages. Two basic procedures are compared: classical decomposition method, and ARIMA models. For the first procedure we compare the first three values of a seasonally adjusted time series; for the second we compare parameter estimates in an ARIMA model.

1 Introduction

The purpose of this paper is to compare procedures used for time series analysis in different statistical software packages. Time series analysis is carried out by many statisticians who publish many results and outcomes every year. However, if you check the calculations, you often get results not exactly the same as the published ones. The reasons may vary; for example different software may have been used. In this paper we will make an attempt at a comparison of procedures called the same name in different software packages. The comparison will be based on the same source data and, of course, will be carried out under the same conditions. Our goal is to find out whether procedures called or marked by the same name in different software packages are the same (i.e., give the same results) or if only their names are identical. Similar subject matter already was processed in articles [3].

We consider the following eight statistical software packages:

- SPSS for Windows Release 10.1.0
- Statistica Release 6
- Statgraphics Plus for Windows 3.1 (referred to as SGWIN below)
- Statgraphics Plus, Release 7.0
- NCSS 97
- SAS Release 8.02
- EViews 4.1
- SCA Statistical System

This list is not to be considered exhaustive, and there are many instances of other software that merited inclusion on it. Our comparison is based on what we have - in other words, we compare software packages that are legally available and used at the University of Economics in Prague. Another aspect we do realise is that we test two different classes of software – general statistical packets (Statgraphics, NCSS, Statistica, SPSS) on the one hand

and specialised software for time series analysis (EViews, SCA) on the other hand. The SAS package holds a special designation - it “can do everything.”

For purposes of the comparison we will only consider two of the basic tasks in time series analysis. We evaluate and compare the procedures and their results. Namely, the following two tasks are considered

- classical decomposition
- Box-Jenkins methodology.

It could be claimed that some different or additional methods should have been compared. However, if we wanted to put forth a comprehensive comparison of methods for time series analysis, we would need much more space than a few pages. Let us view this paper as an attempt at contemplating whether it is in principle possible to get the same results when different software is used, accepting the restriction implied by the fact that the “comparability” is only tested on two procedures.

Further it should be pointed out that the goal of this paper is not to provide a detailed treatise of calculation methods, theory and formulas on which the respective procedures are programmed. Any such treatise would go beyond the scope of this paper.

2 Classical decomposition

This is a basic procedure contained in all packages we consider. The packages differ in the range of procedures offered and, as shown below, by the results. Each package gives an option of pre-transforming the data before running the procedure. Let us briefly comment on each package.

SPSS

It offers both types of models – additive and multiplicative. Before the Seasonal Decomposition procedure is run, the “date” variable must be defined; otherwise the procedure cannot be performed at all.

Statistica

Statistica offers the Moving Averages and X11-ARIMA methods. You can choose between the additive and multiplicative models. There is a wide range of additional options according to which the analysis can be carried out.

SGWIN, Statgraphics Plus

It offers both types of models – additive and multiplicative. Seasonal adjustment of data makes use of Centred Moving Averages. A calculation of season indices is also offered.

NCSS

This package does not offer a choice between additive and multiplicative models; the multiplicative one is automatically applied. Seasonal adjustment of data makes use of Centred Moving Averages.

SAS

SAS uses the X-11 ARIMA method for seasonal decomposition; a choice between additive and multiplicative models can be made within the framework of this method. There is a very wide range of parameter and output options.

EViews

EViews provides a choice of four procedures for seasonal decomposition:

- Census X12
- Census X11
- Moving Averages
- Tramo/Seats

The latter method is only offered by EViews among all the packages we tested.

Software	Seasonally adjusted series		
	1st value	2nd value	3rd value
SPSS	3.81096	3.79465	3.32092
Statistica[1]	3.81096	3.79465	3.32092
Statistica[2]	3.80088	3.77009	3.32225
Statistica[3]	3.64216	3.28032	3.14534
SGWIN	3.68105	3.80067	3.36578
Statgraphics Plus	3.68105	3.80067	3.36578
NCSS[4]	3.62120	3.72264	3.36098
SAS	3.72600	3.70800	3.30500
EViews[5]	3.35592	3.46498	3.06850
EViews[6]	3.46824	3.49890	3.32720
EViews[7]	3.71975	3.70335	3.29641
EViews[8]	2.81500	2.67200	2.75500

Table 1: The first three seasonally adjusted values of the Sales series.

[1] the Centred Moving Averages option unchecked
[2] the Centred Moving Averages option checked
[3] X11 method
[4] it does not provide the values as a direct result of calculation, the user must complete it
[5] Moving Averages method
[6] Tramo/Seats
[7] Census X11 like X11-ARIMA
[8] Census X12

Table 1 shows the first three (seasonally adjusted) values of the Sales series. This time series is a demonstration data file from Statgraphics software. It is a time series recording monthly sales of sect (brand undisclosed) within a certain territory in 7 years. A graph of the Sales series is shown in the Figure. The graph shows that it is a distinctively seasonal series governed by a multiplicative model.

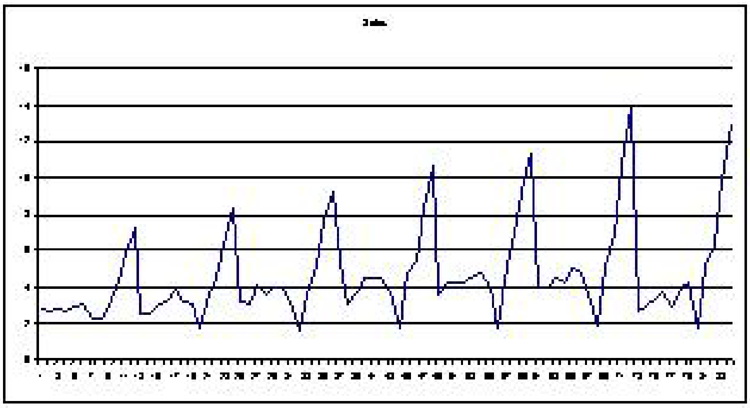


Figure 1: Sales series.

The series was processed an all the packages under assessment using the same procedures for seasonal decomposition or seasonal adjustment (the terminology varies between the packages). All methods of seasonal adjustment available in each package were applied. There are several observations implied by the Table:

- SGWIN and Statgraphics Plus give the same results - this was expected since the packages are identical; the former is a version for Windows, the latter for DOS;
- SPSS and Statistica give the same results if the option Centred Moving Averages remains unchecked in the latter; if the option is checked, the results differ;
- various forms of seasonal adjustment within the same package provide different results (that is natural since they represent different methods);
- the methods named the same in different packages give different results (this is more difficult to understand than the preceding observation: an application of centred moving averages is very simple and the procedures are not only identically named but their theoretical descriptions in the relevant package manuals are also the same).

3 ARIMA model

Let us now compare procedures for building ARIMA models in all packages under assessment. Here we had to face a problem of very many theoretical approaches. They are, of course, all based on the classical reference [1] but options of the procedures offered by the packages vary substantially. We tried to set the parameters and options so that the procedures actually applied were as similar as possible to each other in all packages. But the mentioned problem indicates that even if all parameters of the calculations are set identically, the results will not be identical in all packages.

For illustration we chose a time series recording the CZK/USD exchange rate over a rather long period – from January 1991 to December 2003. Hence the total number of data entries is 156. The source of the data was the website of the Czech National Bank [4]. Let us view the time-series graph:

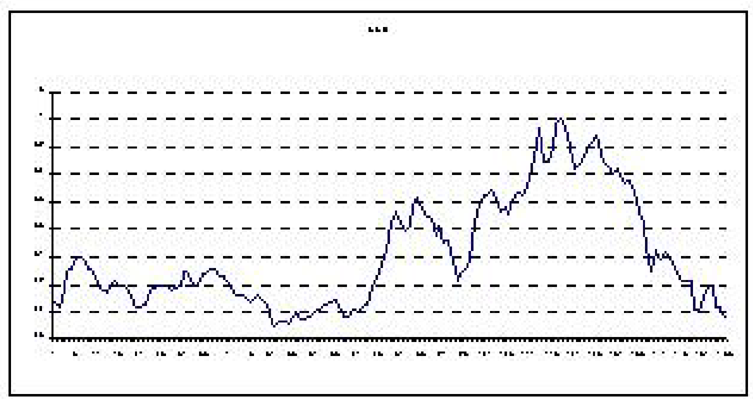


Figure 2: USD series.

It is clear at first sight that the classical decomposition cannot be applied to the series since it does not show any systematic components in the relevant period. That is why we used an ARIMA model. During the processing it became clear that differences had to be taken to achieve stationary distribution. In the end we identified a simple ARIMA model in the following general form:

$$(1 - \varphi_1 B)X_t = \varepsilon_t$$

where

$$X_t = (1 - B)\mathrm{USD}_t$$

and B is the backward-shift operator defined by

$$B(X_t) = X_{t-1} \,.$$

Let us see the estimated model in each package:

Software	Parameter Estimate	Residual Standard Error
SPSS	0.284058	0.784363
Statistica	0.285449	0.784417
Statgraphics Plus	0.284500	0.784417
SGWIN	0.284496	0.784417
NCSS	0.285908	0.784362
SAS	0.284050	0.784366
EViews	0.285449	0.783744
SCA	0.285400	0.783744

We can see that we have not obtained exactly identical models, but they are all very near to each other. The parameter estimates differ on the third decimal position, which supports trust in the methods used.

We can formulate a few general remarks in other text:

- All packages under assessment offer a procedure that builds ARIMA models.
- The level of processing and available options substantially vary among the packages; let us recall once more that we are testing two classes of software packages - general statistical packets (Statgraphics, NCSS, Statistica, SPSS) on the one hand and specialised software for time series analysis (EViews, SCA) on the other hand. The SAS package holds a special designation - it “can do everything”.
- There is a price to pay for a wide range of sophisticated options(EViews, SCA and SAS), namely, the use and control of the software is rather demanding. You cannot just sit down to them and get results in five minutes, which is quite easily done within the general statistical software packages. You have to understand the instructions, learn the syntax of procedures and study the manuals in detail. Application of the sophisticated procedures is not very far from programming.
- Even use of the same package can provide different results by varying the parameters; it is critical to know what the meaning of this or that parameter is and what we are actually computing (deep theoretical knowledge is a prerequisite for that).
- In several packages we can find very useful help that explains the underlying theory on which the procedures' programming was based. Sometimes the help is in electronic form, sometimes you have to read it in a printed manual. Very good manual has SCA [2].

4 Conclusions

Both procedures confirm that isn't software like software. In other words, we get different results depending on the package we choose (regardless of the same or different name or title). On the other hand, the differences are not substantial, and the results we get are more or less comparable. What is a must is to study the manual thoroughly to be sure what we are actually computing.

References

[1] Box G.E.P., Jenkins G.M., Reinsel G.C. (1994). *Time series analysis, forecasting and control.* Third edition. Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

[2] Forecasting and Time Series Analysis using the SCA Statistical System. Volume 1. Scientific Associates Corp. Oak Brook, Illinois, USA, 1991.

[3] Marek L. (1999). *Počítáme všichni stejně?* Výpočtová statistika, Bratislava, Slovenská statistická a demografická spoločnost, ISBN 80-88946-03-4.

[4] http://www.cnb.cz/

Address: L. Marek, Department of Statistics and Probability, W. Churchilla sq. 4, University of Economics, Prague 130 67 Prague Czech Republic

E-mail: marek@vse.cz

COMPSTAT'2004 Symposium

BEHAVIOUR OF THE LEAST WEIGHTED SQUARES ESTIMATOR FOR DATA WITH CORRELATED REGRESSORS

Libor Mašíček

Key words: Robust regression, the least trimmed squares, the least weighted squares.

COMPSTAT 2004 section: Robustness.

Abstract: The paper deals with least weighted squares estimator which is robust and generalizes classical least trimmed squares. We provide conditions under which this estimator is consistent and the condition for regressors is discussed in detail. The behaviour of both estimators for linearly dependent subgroup of regressors is presented on numerical example.

1 Introduction

Let us consider the following regression model

$$Y_i = X_i^T \beta_0 + Z_i \qquad \text{for } i = 1, \dots, n \tag{1}$$

where $X_i = (X_{i1}, \dots, X_{ip})^T$ is the $p \times 1$ column vector of random explanatory variables, β_0 is the $p \times 1$ column vector of unknown regression coefficients and Z_i are the random fluctuations with $\mathrm{E}Z_i = 0$. Moreover, the sequence of random vectors $X_1, \dots, X_n$ is independent and identically distributed (IID), the sequence of random variables $Z_1, \dots, Z_n$ is IID and the sequences are mutually independent.

For any $\beta \in R^p$ denote the i-th residual as

$$r_i(\beta) := (Y_i - X_i^T \beta) = Z_i - X_i^T(\beta - \beta_0) \tag{2}$$

and the h-th order statistics of the squared residuals by $r_{(h)}^2(\beta)$, i.e.

$$0 \le r_{(1)}^2(\beta) \le r_{(2)}^2(\beta) \le \dots \le r_{(n)}^2(\beta). \tag{3}$$

Now we can define *the least weighted squares estimator* (LWS) as

$$\hat{\beta}_n^{LWS} := \arg\min_{\beta \in R^p} \sum_{h=1}^{n} w_h r_{(h)}^2(\beta), \tag{4}$$

where $w_1, \dots, w_n$ are given weights. Typically we choose

$$w_h := w\left(\frac{h-1}{n}\right) \qquad \text{for } h = 1, \dots, n \tag{5}$$

where $w : [0,1] \to R$ is a given *weight function* which is nonincreasing (i.e. observations with larger residuals have smaller weight).

Notice that the order of words is important, i.e. the LWS estimator differs from *the classic weighted least squares* (WLS). For the former the weights are assigned to observations implicitly by the estimator itself while for the latter the weights are generated by an external rule.

The least weighted squares estimator was developed by Víšek (see [7] and [8]) and it generalizes classical *least trimmed squares* (LTS) proposed by Rousseeuw (see [4] or [5])

$$\hat{\beta}_n^{LTS} := \arg\min_{\beta \in R^p} \sum_{h=1}^{\lceil n\overline{\alpha} \rceil} r_{(h)}^2(\beta), \tag{6}$$

where $\overline{\alpha} \in (0,1)$ is a given constant ($\lceil x \rceil$ denote the upper integral part of $x \in R$). The LTS estimator is a special case of the LWS estimator for the choice $w(\alpha) = \mathrm{I}\{\alpha < \overline{\alpha}\}$ where $\mathrm{I}\{\dots\}$ is an indicator function. The main reason for developing the LWS estimator was to improve applicability. In the LTS estimator we can adjust only one constant but in the LWS estimator we can choose the whole weight function. This gives us a chance to increase efficiency or decrease gross error sensitivity.

The LWS estimator has nice properties. First of all the breakdown point can be computed immediately from the weight function. If $w(\alpha) > 0$ for $\alpha < \overline{\alpha}$ and $w(\alpha) = 0$ for $\alpha > \overline{\alpha}$ then the LWS estimator has the breakdown point equal to $\min\{1-\overline{\alpha}, \overline{\alpha}\}$. This means that the breakdown point is under control and we can choose it arbitrarily up to 0.5. Finally, the LWS estimator is regression and scale equivariant.

In the next section the conditions for consistency of the LWS estimator are provided. The condition for regressors is discussed in detail and is related to existing measures of linear dependency of regressors. Finally in section 3 the behaviour of the LWS and LTS estimators for linearly dependent subgroups of regressors is presented using a numerical example.

2 The weak consistency of the LWS regression estimator

The following assumptions will be needed throughout this section.

A1: The weight function w is nonincreasing, bounded and has first derivative almost everywhere. Moreover, there exists $0 < \overline{\alpha} < 1$ such that $w(\alpha) > 0$ for $\alpha \in (0, \overline{\alpha})$ and $w(\alpha) = 0$ for $\alpha \in (\overline{\alpha}, 1)$.

A2: The random errors Z_i have continuous distribution with $\mathrm{E}Z_i^2 < \infty$, distribution function F_Z and density f_Z. The density is bounded, symmetric, strictly decreasing on $(0, \infty)$, $f_Z(x) > 0$ for $x \in R$ and f_Z' exists everywhere.

A3: The random vectors of explanatory variables X_i have finite second moments and there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that

$$P\left(|X_i^T t| < \delta_1\right) \le \overline{\alpha} - \delta_2 \qquad \text{for } t \in H_1, \tag{7}$$

where $H_1 := \{t \in R^p : \|t\| = 1\}$.

Theorem 1 (Weak consistency of the LWS estimator) *Let conditions* **A1***,* **A2** *and* **A3** *be satisfied. Then the LWS estimator is weakly consistent estimator of* β_0*, i.e.* $\hat{\beta}_n^{LWS} \rightarrow_P \beta_0$.

Theorem 1 was presented and proved in [3]. The case of location parameter was discussed in [2] and $n^{1/4}$-consistency was proved. Let us now look deeply at the conditions in Theorem 1 namely at condition **A3**.

It can be shown that **A3** is equivalent to the following condition: the random variables X_i have finite second moments and there exists $\varepsilon > 0$ such that

$$P\left(|X_i^T t| = 0\right) \leq \overline{\alpha} - \varepsilon \qquad \text{for } t \in H_1. \tag{8}$$

The last inequality together with existence of second moments implies that the $p \times p$ matrix $\mathrm{E}X_i X_i^T$ is positive definite (positive definitenes of matrix $\mathrm{E}X_i X_i^T$ is necessary for consistency of the classical least squares). But condition **A3** is even stronger and is important for the consistency of the LWS estimator. Because $w(\alpha) = 0$ for $\alpha > \overline{\alpha}$ we try to fit only $\lceil n\overline{\alpha} \rceil$ observations. Suppose there exists $t \in R^p$ such that $P\left(|X_i^T t| = 0\right) > \overline{\alpha}$. Hence for large datasets (i.e. large n) there exists (with high probability) a subgroup of observations which contains at least $\lceil n\overline{\alpha} \rceil$ observations and their matrix of the explanatory variables is singular. Thus we expect that LWS estimator will not be stable in case of dependent regressors with higher probability than $\overline{\alpha}$.

Using condition (7) the measure of linear dependency of regressors can be defined as follows: let $\delta(\alpha)$ be the largest possible δ which satisfies inequality $P\left(|X_i^T t| < \delta\right) \leq \alpha$ for any $t \in H_1$, i.e.

$$\delta(\alpha) = \sup\left\{\delta \in R : P\left(|X_i^T t| < \delta\right) \leq \alpha \text{ for any } t \in H_1\right\}. \tag{9}$$

Notice $\delta(\alpha)$ is equal to the largest possible δ_1 in (7) for $\alpha := \overline{\alpha} - \delta_2$. Hence (7) is equivalent to: $\delta(\alpha) > 0$ for some $\alpha < \overline{\alpha}$.

An analogous measure for nonrandom regressors $x_1, \ldots, x_n$ was proposed by Davies (see [1]) where $\lambda_n(\alpha)$ is defined

$$\lambda_n(\alpha) = \min_{\substack{S \subset \{1,\ldots,n\} \\ \mathrm{card}\{S\} = \lfloor n\alpha \rfloor}} \left[\min_{\|\theta\|=1} \left(\max_{i \in S} \left|x_i^T \theta\right|\right)\right]. \tag{10}$$

The function $\lambda_n(\alpha)$ measures the worst possible conditioning of any $\lfloor n\alpha \rfloor$ subset of the explanatory variables for the linear regression model with fixed explanatory variables. In fact $\lambda_n(\alpha) = \delta(\alpha)$ if random variables X_i have uniform distribution on the set of points $\{x_1, \ldots, x_n\}$.

3 Numerical example

For $k \in \{15, 20, 25\}$ suppose the following linear regression model

$$Y_i = 1 + 2X_{i1} + 3X_{i2} + Z_i \qquad \text{for } i = 1, \ldots, 25 \tag{11}$$

where

$$\begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \right) \qquad \text{for } i = 1, \ldots, k \tag{12}$$

$$\begin{pmatrix} X_{i1} \\ X_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \qquad \text{for } i = (k+1), \ldots, 25 \tag{13}$$

and $Z_i \sim \mathrm{N}(0, 1)$. Notice that more than half of the observations have linearly dependent regressors because $X_{i1} \approx X_{i2}$ for $i = 1, \ldots, k$.

Theoretically all regressors should have the same distribution, i.e. we should choose a mixture of two normal distributions given in (12) and (13). But the behaviour of the LWS and LTS estimators depends on the number of linearly dependent regressors too much. Hence we use the model given by (11), (12) and (13) where the fraction of correlated regressors is under control (the number of them is equal to k).

Using model (11) we have made simulations as follows: for each $k \in \{15, 20, 25\}$ we have generated 100 datasets with regression model (11) and computed three estimators. First we computed the LWS estimator with weights

$$w_i := \frac{14 - i}{13} \qquad \text{for } i = 1, \ldots, 13 \tag{14}$$

$$:= 0 \qquad \text{for } i = 14, \ldots, 25, \tag{15}$$

i.e. the first half of weights is linearly decreasing and the second is equal to zero. Second we computed the LTS estimator with $\overline{\alpha} = 0.5$, i.e. $\lceil n\overline{\alpha} \rceil = 13$ in (6). Finally, for comparison we also computed the classical least squares estimator (LS).

For each estimated coefficient we calculate the error, i.e. the absolute difference between estimated and real regression coefficient. Tables 1, 2 and 3 give the number of cases with error in given intervals. For example the first number in line marked "$0 - 0.1$" is the number of cases with error of the intercept for LWS in interval $(0, 0.1)$, the second is the number of cases which have error of the first coefficient for LWS in interval $(0, 0.1)$, etc. Table 4 contains errors for the sum of the first and the second regression coefficients.

For computing the LTS we used iteration algorithm proposed by Víšek (see [6]). The idea of this algorithm is following. Having selected randomly h observations we apply LS on them and for the estimated regression coefficients we evaluate residuals for all n observations. Then we select h observations with the smallest squared residuals and again apply LS, again evaluate residuals for all n observations, etc. The repetitions are performed so long until an improvement in the sum of h smallest squared residuals is obtained (it can be easily proved that this sum is not increasing throughout this process). Then a new h observations are selected randomly and the whole process is repeated. The search is stopped when we arrive at the same model 20 times or until a prior given number of repetitions is exhausted. This algorithm can be also generalized for LWS applying WLS instead of LS where weights are assigned to observations by rule “larger squared residual – smaller weight” and vice versa.

Let us now look at the results more precisely. Table 1 contains data for the case of $k = 15$, i.e. approximately one half of the data has correlated regressor and one half has not. In this case all three estimators behave more or less the same. The LWS and LTS estimators have larger errors but the results are still reasonable. This is because the number of correlated observations (i.e. 15) is not much larger than the number of positive weights in the LWS and LTS estimators (i.e. 13).

The results for the second case of $k = 20$ are in Table 2. Here most of the observations have correlated regressor but there are still 5 observations which have not. Hence the whole matrix of regressors is of full rank but some submatrices containing 13 of the regressors are not well conditioned and these 13 regressors are “almost” dependent.

From Table 2 we see that there are many cases where the LWS and LTS estimators have very large errors. In fact the error of both coefficients for the LWS estimator was in three cases larger than five and for the LTS estimators in two cases. For both estimators more than 20 cases have the first or the second regression coefficient larger than 2. Hence both the LWS and LTS are not stable in this case.

Notice that the errors of the intercept for both estimators are not very large. This is because the first and the second regressor are dependent.

In Table 3 there are results for the case of $k = 25$. Here all observations have correlated regressors and hence all three estimators are unstable. The most unstable estimator here is the LWS estimator where many cases with errors larger than 5 arise, in some cases even larger than 10. The LTS estimator behave more or less the same but errors are rather smaller. This is because some weights in the LWS estimators are very small (compare the smaller weight $\frac{1}{13}$ with the larger 1) hence they do not have large influence on the value of the estimator. The errors of the intercept are again reasonable for all estimators.

Table 4 contains errors for the sum of both regression coefficients. For $k = 20$ and $k = 25$ there are many cases with large errors. But the errors of the sum of both coefficients are reasonable (all three errors in interval 2 – 5 are smaller than 2.2), i.e. if the first coefficient is large and positive the second has approximately the same absolute value but is negative and vice versa. This is because in observations $1, \ldots, k$ both regressors are positively corelated and hence they can substitute each other.

Finally we can say that the LWS and LTS estimators are very good in outlier detection (see [5], [7] or [8]). But they are very sensitive to linear dependency in subgroups of regressors in case of no contamination in data. The intercept is estimated well but there could be large errors in regression coefficients. Hence using these estimators we should be careful, namely in high dimensional datasets where this linear dependency can easily occur.

	LWS			**LTS**			**LS**		
Interval	**Int.**	**X1**	**X2**	**Int.**	**X1**	**X2**	**Int.**	**X1**	**X2**
0 – 0.1	13	20	20	13	23	25	27	45	44
0.1 – 0.5	51	56	49	50	47	43	68	54	54
0.5 – 1	26	19	22	27	23	24	5	1	2
1 – 2	10	5	9	10	6	8	0	0	0
2 – 5	0	0	0	0	1	0	0	0	0
5 – 10	0	0	0	0	0	0	0	0	0
$10 - \infty$	0	0	0	0	0	0	0	0	0

Table 1: $k = 15$.

	LWS			**LTS**			**LS**		
Interval	**Int.**	**X1**	**X2**	**Int.**	**X1**	**X2**	**Int.**	**X1**	**X2**
0 – 0.1	13	19	17	15	19	16	40	41	39
0.1 – 0.5	54	44	48	53	48	48	59	54	50
0.5 – 1	29	21	16	28	17	18	1	5	11
1 – 2	4	6	7	4	5	7	0	0	0
2 – 5	0	7	9	0	9	9	0	0	0
5 – 10	0	2	2	0	2	2	0	0	0
$10 - \infty$	0	1	1	0	0	0	0	0	0

Table 2: $k = 20$.

	LWS			**LTS**			**LS**		
Interval	**Int.**	**X1**	**X2**	**Int.**	**X1**	**X2**	**Int.**	**X1**	**X2**
0 – 0.1	23	1	1	14	2	1	31	6	4
0.1 – 0.5	50	8	8	59	7	9	68	14	14
0.5 – 1	24	9	11	24	12	10	1	19	18
1 – 2	3	21	14	3	19	18	0	31	27
2 – 5	0	34	34	0	39	36	0	28	34
5 – 10	0	22	25	0	17	21	0	2	3
$10 - \infty$	0	5	7	0	4	5	0	0	0

Table 3: $k = 25$.

	k=15		k=20		k=25	
Interval	**LWS**	**LTS**	**LWS**	**LTS**	**LWS**	**LTS**
0 – 0.1	16	15	17	21	12	11
0.1 – 0.5	47	45	43	39	40	43
0.5 – 1	21	25	33	31	35	32
1 – 2	15	14	6	9	13	14
2 – 5	1	1	1	0	0	0
5 – 10	0	0	0	0	0	0
10 – ∞	0	0	0	0	0	0

Table 4: The sum of errors of the first and the second regression coefficients.

References

[1] Davies L. (1990). *The Asymptotics of S-estimators in the linear regression model.* The Annals of Statistics, **18** (4), 1651 – 1675.

[2] Mašíček L. (2002). *Konzistence odhadu LWS pro parametr polohy (Consistency of LWS estimator for location parameter).* ROBUST'2002, J. Antoch, G. Dohnal and J. Klaschka (eds), JČMF, 240 – 246.

[3] Mašíček L. (2003). *Consistency of the least weighted squares regression estimator.* Proceedings of International Conference on Robust Statistics, Theory and Applications of Recent Methods, M. Hubert, G. Pison, A. Struyf, S. Van Aelst (eds), Birkhauser, Basel (in print).

[4] Rousseeuw P. J. (1984). *Least median of squares regression.* Journal of the American Statistical Associations **79**, 871 – 880.

[5] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* J.Wiley & Sons, New York.

[6] Víšek J. Á. (1996). *On high breakdown point estimation.* Comp. Stat. **11**, 137 – 146.

[7] Víšek J. Á. (2000). *Regression with high breakdown point.* ROBUST'2000, J. Antoch and G. Dohnal (eds), JČMF, 324 – 356.

[8] Víšek J. Á. (2000). *A new paradigm of point estimation.* Proceedings of "Data Analysis 2000 – Modern Statistical Methods – Modelling, Regression, Classification and Data Mining", K. Kupka (ed.), Trilobyte, Pardubice, 195 – 230.

Acknowledgement: Research was supported by grant of GA ČR 402/03/0084 and MSM113200008.

Address: L. Mašíček, Department of Probability and Statistics, Faculty of Mathmematics and Physics, Charles University, Sokolovská 83, 186 75 Prague, Czech Republic

E-mail: `masicek@karlin.mff.cuni.cz, libor@matfyz.cz`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

ON THE MAXIMAL SAMPLE COORDINATION

Alina Matei and Yves Tillé

Key words: Survey, coordination, IPF procedure.

COMPSTAT 2004 section: Official statistics.

Abstract: The sample coordination is a commonly faced problem in official statistics. We are interested in maximizing the overlap between two samples drawn in different time occasions. The form of the problem of sample coordination in the frame of a transportation problem enables us to compute the joint inclusion probability of two samples drawn on two different occasions, s_1 and s_2, and then the conditional probability $p(s_2|s_1)$. The problem is solved using linear programming. This solution is not computational fast due to the exponential number of possible samples. A new method to optimize coordination between two samples without using linear programming is proposed.

1 Introduction

The *problem of sample coordination* (PSC) is a commonly faced problem in official statistics. A population is sampled on two or more occasions, in order to obtain current estimates of a character (for example the number of unemployed). The major aim of the coordination is to control the overlap between samples. The population can change in time, due to births, deaths or changes in activity in the case of enterprises. Sample coordination is commonly applied in household, business and health surveys. Its applications in business studies is of increasing importance and interest. The coordination of business surveys is very important for statistical agencies that send out every year numerous questionnaires to the establishments.

Sample coordination can be either positive or negative. While in the former the number of common units between two or more samples is maximized, in the latter the number of common units is minimized. The positive and negative coordination can be seen as a dual problem. Thus, solving positive coordination problem can lead us to the solution of negative sample coordination and viceversa.

We focus on positive coordination and two occasions. Samples without replacement are selected on two distinct time periods. Let $U = \{1, .., k, .., N\}$ be the population studied, let π_k^1, π_k^2 for all $k \in U$ be the inclusion probabilities for the first and second time respectively, and let $\pi_k^{1,2}$ be the joint inclusion probability of unit k in the first and second occasion. A support $\mathcal{S}$ is a set of samples of U. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the sample supports in the first and second occasions, respectively.

The overlap between two samples is defined as the number of common units in both samples. Each unit $k \in U$ can be selected in both samples with probability at most $\min(\pi_k^1, \pi_k^2)$. An upper bound of the expected overlap is $\sum_{k\in U} \min_{k\in U}(\pi_k^1, \pi_k^2)$. We call this bound the *absolute upper bound*. It is reached when $\pi_k^{1,2} = \min_{k\in U}(\pi_k^1, \pi_k^2)$, for all $k \in U$. Only a few part of the already developed methods can reach the absolute upper bound in the case of two designs.

One way to solve PSC is to use mathematical programming to solve a transportation problem. PSC in the frame of a transportation problem enables us to compute the joint inclusion probability of two samples drawn in two different occasions (s_1 and s_2) and the conditional probability $p(s_2|s_1)$. The latter enables us to choose the sample s_2 drawn in a second occasion given that the sample s_1 was drawn in the first occasion. A bi-design denotes a couple of two marginal designs drawn in two occasions. We are interested to find the conditions when the *absolute upper bound* is reached in the case of a bi-design. We pose this problem because the value of the objective function in the case of an optimal solution given by the linear programming (the relative upper bound) is not necessarily equal to the absolute upper bound. In this article we develop a procedure to decide if the absolute upper bound can be reached or not. An algorithm based on the *Iterative Proportional Fitting* (IPF) procedure [3] is used to give an optimal solution in the case of PSC, without solving the linear program.

2 Transportation problem in sample coordination

Mathematical programming (more precisely linear programming) was used to solve the coordination problem of two samples as a transportation problem. The applications of the transportation problem in sample coordination are given in [11], [1], [2], [7], [5], [6], [8].
We use the following form of transportation problem presented in [2]:

$$\max \sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij} \tag{1}$$

subject to constraints

$$\left| \begin{array}{l} \sum_{j=1}^{q} p_{ij} = p_1(s_i^1), i = 1, \ldots, m, \\ \sum_{i=1}^{m} p_{ij} = p_2(s_j^2), j = 1, \ldots, q, \\ p_{ij} \geq 0, i = 1, \ldots, m, j = 1, \ldots, q, \end{array} \right.$$

with $c_{ij} = \#(s_i^1 \cap s_j^2)$, $p_1(s_i^1) = prob(s_i^1), p_2(s_j^2) = prob(s_j^2)$, $p_{ij} = prob(s_i^1, s_j^2)$. $s_i^1 \in \mathcal{S}_1$ and $s_j^2 \in \mathcal{S}_2$ denote the possible samples in the first and second occasion, respectively, with $\#(\mathcal{S}_1) = m$ and $\#(\mathcal{S}_2) = q$. We suppose that $p_1(s_i^1) > 0, p_2(s_j^2) > 0$ in order to compute the conditional probabilities. A modification of this problem has been done by Ernst [4]. In the case of two selected units per stratum, Ernst and Ikeda [7] simplify the computational aspect of the problem (1).

When only one unit is selected in each sample, we obtain a particular case of problem (1), that was presented by Raj [11] as follows:

$$\max \sum_{k=1}^{N} \pi_k^{1,2}, \tag{2}$$

subject to constraints

$$\left| \begin{array}{l} \sum_{\ell=1}^{N} \pi_{k\ell}^{1,2} = \pi_k^1, \\ \sum_{k=1}^{N} \pi_{k\ell}^{1,2} = \pi_\ell^2, \\ \pi_{k\ell}^{1,2} \geq 0, k, \ell = 1, \ldots, N, \end{array} \right.$$

where $\pi_{k\ell}^{1,2}$ is the probability to select the units k and ℓ in both samples. Arthanari and Dodge [1] showed that any feasible solution of problem (2), with $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$ for all $k \in U$, is an optimal solution. The Keyfitz method [9] gives an optimal solution to problem (2), without application of the simplex algorithm.

3 Maximal sample coordination

Our goal is to define a method which gives an optimal solution to problem (1) without using linear programming. We consider problem (1) as a two-dimensional distribution where only the two marginal distributions (the sums along the rows and columns) are given. Information about the inner distribution is available by using the propositions below. It is required to estimate the internal values. The technique is based on the IPF procedure [3].

A measure of positive coordination is the number of common sampled units in two occasions. Let n_{12} be this number. The goal is to maximize the expectation of n_{12} defined as

$$E(n_{12}) = \sum_{k \in U} \pi_k^{1,2} = \sum_{k \in U} \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p_{ij} = \sum_{s_i^1 \in \mathcal{S}_1} \sum_{s_j^2 \in \mathcal{S}_2} \#(s_i^1 \cap s_j^2) p_{ij},$$

which is the objective function of problem (1). To maximize $E(n_{12})$ it amounts to maximize the objective function of problem (1).

Similarly, the objective function of problem (2) is

$$\sum_{k=1}^{N} \#(\{k\} \cap \{k\}) Prob(\{k\}, \{k\}) = \sum_{k=1}^{N} \pi_k^{1,2}.$$

We define the *absolute upper bound* as $\sum_{k=1}^{N} \min(\pi_k^1, \pi_k^2)$. In general, $\sum_{s_i^1 \in \mathcal{S}_1} \sum_{s_j^2 \in \mathcal{S}_2} \#(s_i^1 \cap s_j^2) p_{ij} \leq \sum_{k=1}^{N} \min(\pi_k^1, \pi_k^2)$. The absolute upper bound is reached when $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$, for all $k \in U$.

Proposition 3.1. *We suppose to have $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$, for all $k \in U$. The following conditions are fulfilled:*

a1. *if* $(\min(\pi_k^1, \pi_k^2) = \pi_k^1$ *and* $k \in s_i^1, k \notin s_j^2)$ *then* $p_{ij} = 0$;

b1. *if* $(\min(\pi_k^1, \pi_k^2) = \pi_k^2$ *and* $k \notin s_i^1, k \in s_j^2)$ *then* $p_{ij} = 0$.

The reciprocal is also available.

Proof.

$$\begin{aligned}
\pi_k^1 &= \sum_{s_i^1 \ni k} p_1(s_i^1) \\
&= \sum_{s_i^1 \ni k} \sum_{s_j^2 \in \mathcal{S}_2} p(s_i^1, s_j^2) \\
&= \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p(s_i^1, s_j^2) + \sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p(s_i^1, s_j^2) \\
&= \pi_k^{1,2} + \sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p(s_i^1, s_j^2).
\end{aligned}$$

If $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$ and $\min(\pi_k^1, \pi_k^2) = \pi_k^1$, then $\sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p(s_i^1, s_j^2) = 0$, and $p(s_i^1, s_j^2) = p_{ij} = 0$, for $k \in s_i^1, k \notin s_j^2$. The proof is analogous for the second relation. □

The joint probabilities p_{ij} in problem (1) can be formulated as a matrix $\mathbf{P} = (p_{ij})_{m \times q}$. Proposition 3.1 enables us to set some p_{ij} to zero in order to find an optimal solution to problem (1).

Proposition 3.2. *A feasible solution of problem (1) with the properties:*

a2. $p_{ij} = 0$ *if there exists* $k \in s_i^1, k \notin s_j^2$ *and* $\min(\pi_k^1, \pi_k^2) = \pi_k^1$;

b2. $p_{ij} = 0$ *if there exists* $k \notin s_i^1, k \in s_j^2$ *and* $\min(\pi_k^1, \pi_k^2) = \pi_k^2$;

is an optimal solution.

Proof. Consider a feasible solution of problem (1), with the properties a2. and b2. Its objective function is equal to

$$\begin{aligned}
&\sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij} = \sum_{k \in U} \pi_k^{1,2} = \sum_{k \in U} \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p_{ij} \\
&\quad = \sum_{k \in U} \sum_{s_i^1 \ni k} \sum_{s_j^2 \ni k} p_{ij} + \sum_{\substack{k \in U, \\ \min(\pi_k^1, \pi_k^2) = \pi_k^1}} \sum_{s_i^1 \ni k} \sum_{s_j^2 \not\ni k} p_{ij} + \sum_{\substack{k \in U, \\ \min(\pi_k^1, \pi_k^2) = \pi_k^2}} \sum_{s_i^1 \not\ni k} \sum_{s_j^2 \ni k} p_{ij}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^1}} \Big(\sum_{s_i^1\ni k}\sum_{s_j^2\ni k} p_{ij} + \sum_{s_i^1\ni k}\sum_{s_j^2\not\ni k} p_{ij}\Big)\\
&\quad + \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^2}} \Big(\sum_{s_j^2\ni k}\sum_{s_i^1\ni k} p_{ij} + \sum_{s_j^2\ni k}\sum_{s_i^1\not\ni k} p_{ij}\Big)\\
&= \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^1}} \sum_{s_i^1\ni k}\sum_{s_j^2\in\mathcal{S}_2} p_{ij} + \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^2}} \sum_{s_j^2\ni k}\sum_{s_i^1\in\mathcal{S}_1} p_{ij}\\
&= \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^1}} \sum_{s_i^1\ni k} p_1(s_i^1) + \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^2}} \sum_{s_j^2\ni k} p_2(s_j^2)\\
&= \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^1}} \pi_k^1 + \sum_{\substack{k\in U,\\ \min(\pi_k^1,\pi_k^2)=\pi_k^2}} \pi_k^2\\
&= \sum_{k\in U} \min(\pi_k^1,\pi_k^2).
\end{aligned}$$

□

We note by $I = \{k \in U | \pi_k^1 \le \pi_k^2\}$ the set of "increasing" units and by $D = \{k \in U | \pi_k^1 > \pi_k^2\}$ the set of "decreasing" units.

Proposition 3.3. *Suppose that all samples have the corresponding probabilities* > 0, *and* $\pi_k^{1,2} = \min(\pi_k^1, \pi_k^2)$, *for all* $k \in U$. *Let* $s_1 \in \mathcal{S}_1$. *If at least one of the following relations is fulfilled for all* $s_2 \in \mathcal{S}_2$*:*

a3. $(s_1 \backslash s_2) \cap I \neq \emptyset$,
b3. $(s_2 \backslash s_1) \cap D \neq \emptyset$,

then the two designs cannot be maximally coordinated. This proposition holds in the symmetric sense, too (if s_2 *is fixed and at least one of the conditions a3. and b3. is fulfilled for all* $s_1 \in \mathcal{S}_1$*).*

Proof. Suppose that two designs can be maximally coordinated.

a3. $(s_1 \backslash s_2) \cap I \neq \emptyset \Longrightarrow$ it exists $\ell \in s_1 \backslash s_2$ and $\ell \in I \Longrightarrow$ from Proposition 3.1 $\Longrightarrow p(s_1, s_2) = 0$;
b3. $(s_2 \backslash s_1) \cap D \neq \emptyset, \Longrightarrow$ it exists $\ell \in s_2 \backslash s_1$ and $\ell \in D \Longrightarrow$ from Proposition 3.1 $\Longrightarrow p(s_1, s_2) = 0$;

We always have $p(s_1, s_2) = 0$, for all $s_2 \in \mathcal{S}_2 \Longrightarrow p_1(s_1) = 0 \Longrightarrow$ we obtain a contradiction with $p_1(s_1) > 0$. The proof is analogous for the symmetric sense of affirmation. □

Let $U = \{1, 2, 3, 4\}, I = \{3, 4\}$ and $D = \{1, 2\}$. We draw in two distinct occasions samples of size 2 and 3, respectively. Below, we note the zero values given by using the Proposition 3.1. By x we note the non-zero values. The sample $\{3, 4\}$ in the first occasion has on its row only the zero values. The maximal coordination is not possible. The same result is also available by using the Proposition 3.3.

	{1,2,3}	{1,2,4}	{1,3,4}	{2,3,4}
{1,2}	x	x	x	x
{1,3}	0	0	x	0
{1,4}	0	0	x	0
{2,3}	0	0	0	x
{2,4}	0	0	0	x
{3,4}	0	0	0	0

Impossible maximal coordination

4 The proposed algorithm

We propose the next algorithm based on the propositions 3.1 and 3.3:

Step 1. Let $\mathbf{P} = (p_{ij})_{m\times q}$ be the matrix given by the independence between both designs: $p_{ij} = p_1(s_i^1)p_2(s_j^2)$, for all $i = 1, \ldots, m, j = 1, \ldots, q$.

Step 2. Put the zeros in $\mathbf{P}$ using the Proposition 3.1.

Step 3. If the conditions of Proposition 3.3 are fulfilled, stop the algorithm and give the message "the absolute upper bound cannot be reached"; else apply the IPF procedure to restore the margins.

The correctitude of the algorithm is assured by the Proposition 3.2.

Concerning the IPF procedure, in a first iteration indicated by the exponent (1) calculate for all rows $i = 1, \ldots, m$

$$p_{ij}^{(1)} = p_{ij}^{(0)} \frac{p_1(s_i^1)}{p_1^{(0)}(s_i^1)}, \text{ for all } j = 1, \ldots, q, \tag{3}$$

where $p_{ij}^{(0)} = p_1(s_i^1)p_2(s_j^2)$ and $p_1^{(0)}(s_i^1) = \sum_{j=1}^{q} p_{ij}^{(0)}$. Now the total rows $p_1(s_i^1)$ are satisfied. Calculate in a second iteration for all columns $j = 1, \ldots, q$

$$p_{ij}^{(2)} = p_{ij}^{(1)} \frac{p_2(s_j^2)}{p_2^{(1)}(s_j^2)}, \text{ for all } i = 1, \ldots, m, \tag{4}$$

where $p_2^{(1)}(s_j^2) = \sum_{i=1}^{m} p_{ij}^{(1)}$. Now the total columns $p_2(s_j^2)$ are satisfied. In a third iteration, the resulting $p_{ij}^{(2)}$ are used in recursion (3) for obtaining $p_{ij}^{(3)}$, and so on until convergence is attained.

We take the following example from [2]. The two designs are one PSU per stratum. The population has size 5. The inclusion probabilities are

$0.5, 0.06, 0.04, 0.6, 0.1$ for the first design and $0.4, 0.15, 0.05, 0.3, 0.1$ for the second design. In the first design, the first three PSU's were in one initial stratum and the other two in a second initial stratum. There are $m = 12$ possible samples given in the table below with the corresponding probabilities:

$$0.15, 0.018, 0.012, 0.24, 0.04, 0.3, 0.05, 0.036, 0.006, 0.024, 0.004, 0.12.$$

The second design consists of five PSU's, $q = 5$. The authors of [2] solve the mathematical program associated to the problem and give the value 0.88 for the objective function. Yet, $\sum_{k \in U} \min(\pi_k^1, \pi_k^2) = 0.9$. We have $I = \{2, 3, 5\}, D = \{1, 4\}$. Using Proposition 3.3 we observe that the samples $\{2, 5\}$ and $\{3, 5\}$ have in theirs rows only values equal to zero, and then the two designs cannot be maximally coordinated. We modify the example by letting $\pi_5^1 = 0.2$. Now, $I = \{2, 3\}, D = \{1, 4, 5\}$ and the samples in the first design have the corresponding probabilities:

$$0.1, 0.012, 0.008, 0.24, 0.08, 0.3, 0.1, 0.036, 0.012, 0.024, 0.008, 0.08.$$

We apply the proposed algorithm on matrix $\mathbf{P}$. The absolute upper bound is now reached. The matrix $\mathbf{P}$ after the application of the Steps 1 and 2, the matrix $\mathbf{P}$ after the Step 3 and the values of c_{ij} are given below.

	{1}	{2}	{3}	{4}	{5}	Σ
{1}	0.04	0.015	0.005	0	0	0.0600
{2}	0	0.0018	0	0	0	0.0018
{3}	0	0	0.0004	0	0	0.0004
{4}	0	0.036	0.012	0.072	0	0.1200
{5}	0	0.012	0.004	0	0.008	0.0240
{1,4}	0.12	0.045	0.015	0.09	0	0.2700
{1,5}	0.04	0.015	0.005	0	0.01	0.0700
{2,4}	0	0.0054	0	0	0	0.0054
{2,5}	0	0.0018	0	0	0	0.0018
{3,4}	0	0	0.0012	0	0	0.0012
{3,5}	0	0	0.0004	0	0	0.0004
∅	0	0.012	0.004	0	0	0.0160
Σ	0.200	0.144	0.047	0.162	0.0180	1

Table 1: The matrix $\mathbf{P}$ after the application of the Steps 1 and 2.

5 Conclusions

It is possible to construct an algorithm to compute the conditional probability $p(s_2|s_1)$ for two samples s_1 and s_2 drawn on two different occasions, without solving a linear program. The proposed algorithm has the complexity $O(m \times q \times$ number of iterations in IPF procedure$)$, which is low compared to the linear program.

	{1}	{2}	{3}	{4}	{5}	Σ
{1}	0.098570	0.001287	0.000143	0	0	0.100
{2}	0	0.012	0	0	0	0.012
{3}	0	0	0.008	0	0	0.008
{4}	0	0.009583	0.001065	0.229352	0	0.240
{5}	0	0.003194	0.000355	0	0.076451	0.080
{1,4}	0.226073	0.002952	0.000328	0.070648	0	0.300
{1,5}	0.075358	0.000984	0.000109	0	0.023549	0.100
{2,4}	0	0.036	0	0	0	0.036
{2,5}	0	0.012	0	0	0	0.012
{3,4}	0	0	0.024	0	0	0.024
{3,5}	0	0	0.008	0	0	0.008
∅	0	0.072	0.008	0	0	0.080
Σ	0.400	0.150	0.050	0.300	0.100	1

Table 2: The matrix **P** after the application of the Step 3.

	{1}	{2}	{3}	{4}	{5}
{1}	1	0	0	0	0
{2}	0	1	0	0	0
{3}	0	0	1	0	0
{4}	0	0	0	1	0
{5}	0	0	0	0	1
{1,4}	1	0	0	1	0
{1,5}	1	0	0	0	1
{2,4}	0	1	0	1	0
{2,5}	0	1	0	0	1
{3,4}	0	0	1	1	0
{3,5}	0	0	1	0	1
∅	0	0	0	0	0

Table 3: Values of c_{ij}.

References

[1] Arthnari T., Dodge Y. (1981). *Mathematical programming in statistics.* John Wiley & Sons, Inc., New York.

[2] Causey B.D., Cox L.H., Ernst L.R. (1985). *Application of transportation theory to statistical problems.* Journal of the American Statistical Association **80**, 903−909.

[3] Deming W., Stephan F. (1940). *On a least squares adjustment of a sampled frequency table when the expected marginal totals are know.* Annual Mathematical Statistics **11**, 427−444.

[4] Ernst L.R. (1986). *Maximizing the overlap between surveys when information is incomplete.* European Journal of Operational Research **27**, 192 – 200.

[5] Ernst L.R. (1996). *Maximizing the overlap of sample units for two designs with simultaneous selection.* Journal of Official Statistics, **12** (1), 33 – 45.

[6] Ernst L.R. (1998). *Maximizing and minimizing overlap when selecting a large number of units per stratum simultaneously for two designs.* Journal of Official Statistics **14**, 297 – 314.

[7] Ernst L.R., Ikeda M.M. (1995). *A reduced-size transportation algorithm for maximizing the overlap between surveys.* Survey Methodology **21**, 147 – 157.

[8] Ernst L.R., Paben S.P. (2002). *Maximizing and minimizing overlap when selecting any number of units per stratum simultaneously for two designs with different stratifications.* Journal of Official Statistics **18**, 185 – 202.

[9] Keyfitz N. (1951). *Sampling with probabilities proportional to size: adjustment for changes in the probabilities.* Journal of American Statistics Association **46**, 105 – 109.

[10] Ohlsson E. (1996). *Methods for pps size: One sample coordination.* Research report **194**, Stockholm University, Sweden.

[11] Raj D. (1968). *Sampling theory.* McGraw-Hill, New York.

Address: A. Matei, Y. Tillé, Groupe de statistique, Université de Neuchâtel, Espace de l'Europe 4, CP 827, 2002, Neuchâtel, Switzerland

E-mail: `alina.matei@unine.ch, yves.tille@unine.ch`

COMPSTAT'2004 Symposium

DIAGNOSTIC DATA TRACES USING PENALTY METHODS

Lauren McCann and Roy E. Welsch

Key words: Influential data, penalty methods, outliers, regression.

COMPSTAT 2004 section: Robustness.

Abstract: This paper explores the idea of using the mean shift-outlier model on nearly all observations to assess their impact individually and collectively on regression parameter estimates and other statistics of interest. A small subset of "good" observations is found via high-breakdown robust methods and mean-shift dummy variables are added for all of the remaining observations. It is then possible to address the impact of additive outliers via variable selection methods. We focus on penalty methods and use the penalty parameter as a way to move through the space of mean-shift dummy variables to make various plots which aid in understanding the relationship between the data and regression statistics.

1 Introduction

Leave-one-out model and data diagnostics as discussed by Belsley, Kuh, and Welsch [2] and many other authors are well-known to be affected by masking and swamping. These difficulties can be ameliorated by using some form of bounded-influence and/or high-breakdown robust estimation to extract a "good" subset of the data and then use those observations to identify outliers and leverage observations and groups.

If an observation is exceptionally informative for a least-squares regression parameter, it will necessarily have a small residual and its outlyingness is directly linked to the presence or absence of the corresponding parameter (explanatory variable) in the model. An extreme case is the mean shift outlier model where one column of the explanatory variable matrix contains all zeros except for a one in a single row. In this case, a parameter is completely dedicated to that row and the fit is exact (residual is zero). Stated another way, the regression parameter estimates for all other variables are the same as would have been obtained if the row with a one in the dummy column had been completely removed from the computation.

This idea can be carried further and many dummy columns (selected columns from an identity matrix) could be added to begin to address masking, swamping, etc. Looked at this way, outlier detection is related to variable selection among those dummy variables. However, we do not know, a priori, what dummy columns to add to our original explanatory variables.

One approach is to first find a "good" subset of the data, add dummy variables for all remaining observations, and then do some form of variable selection on the dummy variables.

This is computationally efficient since there are good algorithms for least-squares variable selection (especially with sparse matrices). Our focus in this paper is on diagnostics (and detecting subsets of influential observations) and not necessarily on estimation of the original model parameters.

Atkinson and Riani [1] have recently proposed a method called forward search which also begins from a "good" subset of the data and works forward by adding data in a structured way. They, too, are interested in "detecting and investigating observations that differ from the bulk of the data."

In what follows, we propose low cost diagnostic procedures based on finding p (the number of model parameters) "good" observations and selecting from $n-p$ added dummy variables using penalty methods. We then compare these methods to the "forward search" of Atkinson and Riani [1].

2 Background and methodology

We consider the standard regression model with n observations on a response variable $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ with corresponding instances of explanatory variables $\boldsymbol{x}_1 \in \mathbb{R}^{p-1}, \ldots, \boldsymbol{x}_n \in \mathbb{R}^{p-1}$. We assume the observations are realizations of

$$Y_i = E(Y_i \mid \boldsymbol{x}_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and have common mean 0 and variance σ^2. Let $\mathbf{X}$ denote the $n \times p$ matrix of explanatory variables and

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Our linear predictors are of the form

$$\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i, \qquad i = 1, \ldots, n.$$

To find a "good" subset of size p we use the least median of squares (LMS) approach [10]. For each such subset (indexed by j) compute the least-squares coefficients $\hat{\boldsymbol{\theta}}_j$ and find

$$P_j = \underset{i=1,\ldots,n}{\text{median}} \; (y_i - \hat{\boldsymbol{\theta}}_j^{\mathrm{T}}\mathbf{x}_i)^2.$$

We use the subset J^* such that

$$P_{J^*} = \min_j P_j. \tag{1}$$

Of course there are $\binom{n}{p}$ subsets of size p and when this number is large, sampling is required. Such sampling is now an accepted practice in robust estimation and we will use it as necessary in spite of the loss of uniqueness and complete optimality.

There are a number of other methods that could be considered, including least trimmed squares (LTS). If the fraction of contaminated data is like $1/p$, then bounded-influence methods (which require much less computation than LMS or LTS) could also be considered [7].

Given the starting subset, we form a matrix $\mathbf{A}$ consisting of $n-p$ columns with all zeros except for a single 1 in rows not in the set J^*. Stated another way, $\mathbf{A}$ is an identity matrix with columns indexed by J^* omitted. Our new $n \times n$ matrix of "explanatory" variables is

$$\mathbf{Z} = (\mathbf{X} \mid \mathbf{A})$$

and $\hat{\boldsymbol{\beta}}$ now denotes the OLS coefficient estimates based on $\mathbf{Z}$. This allows us to implement the mean shift outlier model for the $n-p$ observations not in the good starting set.

With $\mathbf{Z}$ as our "model," we would get an exact fit for least-squares and the non-dummy coefficients would be the $\hat{\boldsymbol{\theta}}_{J^*}$ from the p "good" starting observations.

In many data-mining applications [5], penalty methods are used to address situations where model complexity (number of fitted parameters) is high relative to the number of observations. In order to vary the "number" of dummy variables and hence the complexity, we solve for $\hat{\boldsymbol{\beta}}(k)$ the penalized least-squares problem

$$\min_{\hat{\beta}(k)} \sum_{i=1}^{n} (y_i - \hat{\boldsymbol{\beta}}(k)^{\mathrm{T}} \mathbf{z}_i)^2 + k \sum_{j \notin J^*} \hat{\beta}_j^2(k). \tag{2}$$

The penalty is applied only to coefficients associated with the observations not in the "good" subset of size p. When k is zero, we obtain the least-squares solutions based only on the starting "good" subset. When $k = \infty$, we obtain the least-squares fit using all n observations with no dummy variables. We could also consider adding a penalty for the non-dummy variables to facilitate variable selection as well as row selection. One such approach is contained in Morgenthaler, et al. [9].

We will vary k to see how going from the starting p observations to all of the data affects coefficient estimates, t-statistics, residuals, etc. in a manner similar to the ridge trace used in ridge regression. This is just one (low-cost) path through all possible subsets of the dummy variables (observations).

Atkinson and Riani [1] choose another path. They start with a "good" subset of size p based on LMS and then go to subsets of size $p+1$, $p+2$ by successively adding back observations with the smallest squared residual based on OLS fits of size p, $p+1$, etc. A larger subset may not always contain all of the observations in a previous smaller subset, but this is a reasonable way to proceed to add observations and examine the effects of doing so. We will compare these methods on some examples later in this paper.

Since k is a continuous parameter and not discrete as is the case for adding observations (or removing them), it is useful to have some measure of how many parameters are in the fitted model. The estimated coefficients for the penalized model (2) are

$$\hat{\boldsymbol{\beta}}(k) = [\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + k\mathbf{A}\mathbf{A}^{\mathrm{T}}]^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y}.$$

A common measure of complexity or equivalent degrees of freedom [5] is

$$c(k) = \text{trace } \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + k\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{Z}^{\mathrm{T}}. \tag{3}$$

This reduces to the hat or projection matrix in least-squares regression when $k = 0$. When $k = \infty$, the complexity will be p, since we are not penalizing the p non-dummy explanatory variables.

3 Diagnostic traces

It is natural to display various components of a regression fitting process (coefficients, residuals, etc.) as a function of the penalty parameter, k, or of the equivalent degrees of freedom. These are useful but suffer from the fact that our penalty function (designed for low-cost computation) does not force “small” coefficients of the dummy variables to be hard zeros, thereby increasing the equivalent degrees of freedom count. Penalty functions with hard zeros are available at additional cost (see [11], and [3] and will be explored in a later paper.

Instead of making our diagnostic traces as functions of k directly, we make them as functions of the number of significant dummy variable t-statistics. The t-statistics we use are

$$t_j(k) = \frac{\hat{\beta}_j(k)}{\hat{\sigma}\sqrt{[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + k\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + k\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}]_{jj}}}, \quad \text{with}$$

$$\hat{\sigma}^2 = \frac{1}{n - c(k)} \sum_{i=1}^{n} (y_i - \hat{\boldsymbol{\beta}}(k)^{\mathrm{T}}\mathbf{z}_i)^2.$$

At the start, with p good data points and $n - p$ dummies allowed ($k = 0$) we get an exact fit and infinite t-statistics. As k increases, we expect large t-statistics associated with outlying data in the beginning. These will decrease as k gets larger and we move toward the least-squares solution ($k = \infty$) which tries to fit outliers. A large t-statistic on a dummy variable signals an additive outlier and we remove it. The number of observations minus the number of dummies that remain is the subset size for that point on the plot.

4 Where to look

Although diagnostic traces are designed to provide an informal look at how our data is influencing the model, it is useful to have some idea of a region of k-values that might provide a good trade-off between bias and the loss of efficiency due to omitting observations. There are, of course, many ways to do this based on the ideas in the vast literature related to ridge regression and

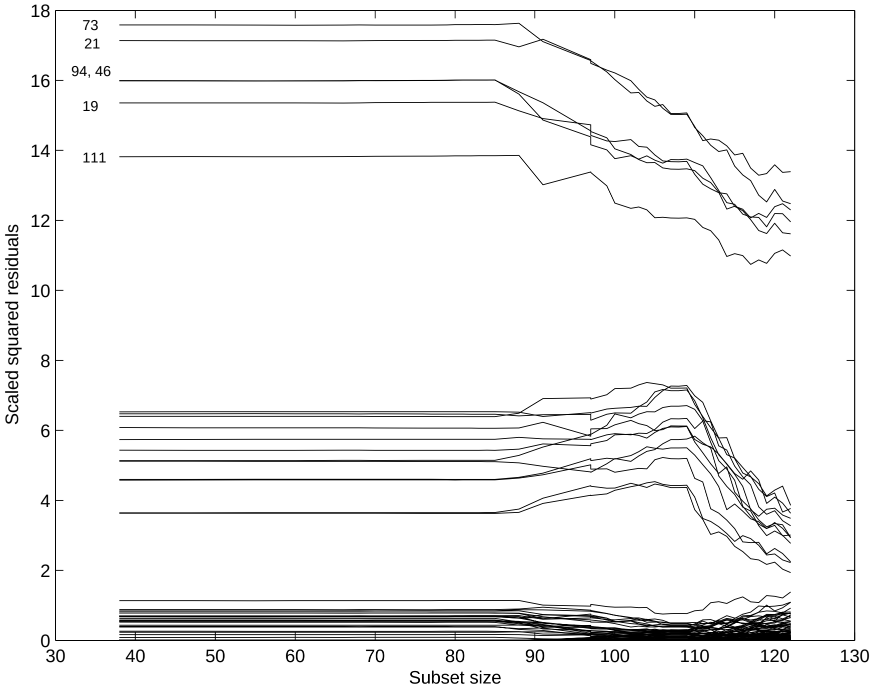


Figure 1: Hawkins' data scaled squared residuals as $k \to \infty$.

penalized regression. We find rules based on prediction and cross-validation to be appealing, and they have proved to be successful in a number of situations [4].

Since we have invested the effort to find p "good" points, we use them with one-at-a-time cross-validation to choose a working value for k. For each of the p observations in the good set and each k (on a grid of k-values), we predict that observation using the penalized regression coefficients estimated from all data except the one being predicted. We compare this prediction with the actual value and use a robust measure of prediction error, median absolute deviations from the median (MAD). We then choose as our working value of k the k that minimizes this measure of prediction error.

It is known that one-at-a-time cross-validation is not consistent in many situations, but is consistent when the number of predictors (real plus dummy in our case) increases as n increases, which is the situation we are in [8].

5 An example

Hawkins, et al. [6] simulated some data that was designed to give data analysts a difficult time. A convenient reference for the data is Appendix A.4 of Atkinson and Riani [1]. In this data set $n = 128$ and there are eight ex-

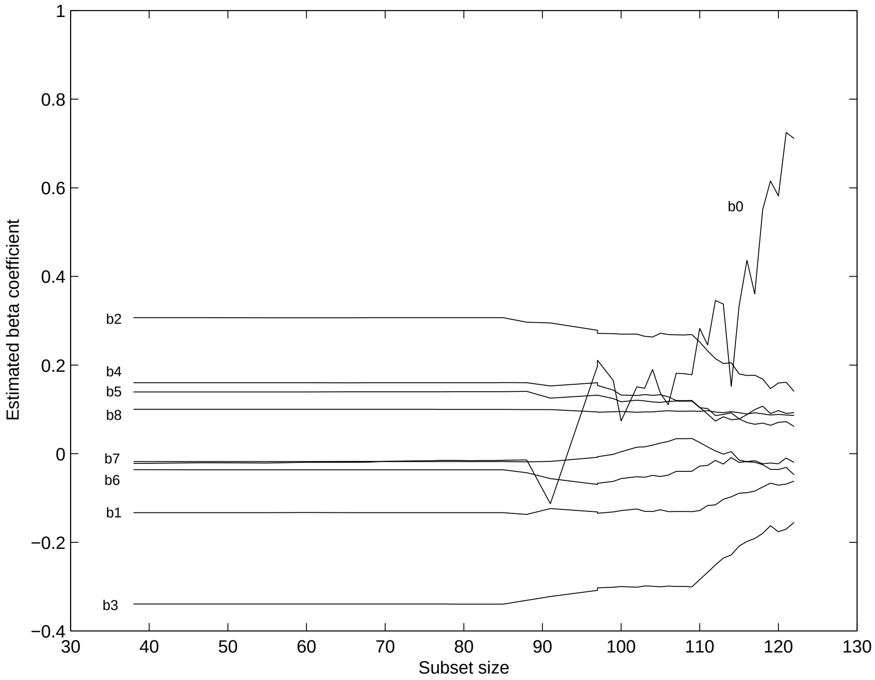


Figure 2: Hawkins' data estimated coefficients as $k \to \infty$.

planatory variables excluding the intercept. Therefore, $p = 9$ in our notation. Figure 1 is a plot of the squared residuals divided by $\hat{\sigma}_f^2$ which is obtained from the least-squares fit using all 9 variables (with no dummies) and all of the data. 2. The x-axis scale is the number of observations used to compute the estimated coefficients as described in Section 3. The plot begins around 40 since the residuals are essentially constant below 40. There are 86 "good" observations in the data set and the residuals show two major groups of outlying observations. These begin to enter the fitting process after 80 on the plot and have an increasing impact on the residuals. This plot is qualitatively similar to Figure 3.3 of Atkinson and Riani [1]. The story is the same for the plot of the coefficients in Figure Figure 3 is obtained by computing

$$(y_i - \hat{\boldsymbol{\beta}}(k)^{\mathrm{T}} \mathbf{z}_i)/\hat{\sigma}_f$$

for all $i \in J^*$, the original subset of good data. The leave-one-out "where to look" procedure from Section 4 gives a k-value of .02 which is about 86 on the x-axis observation scale. The right-hand end of these plots does not agree perfectly with the least-squares estimates for all explanatory and no dummy variables. This is because the t-statistic cut-off at a 95% significance level still finds a few dummy variables not significant and includes them in the fit.

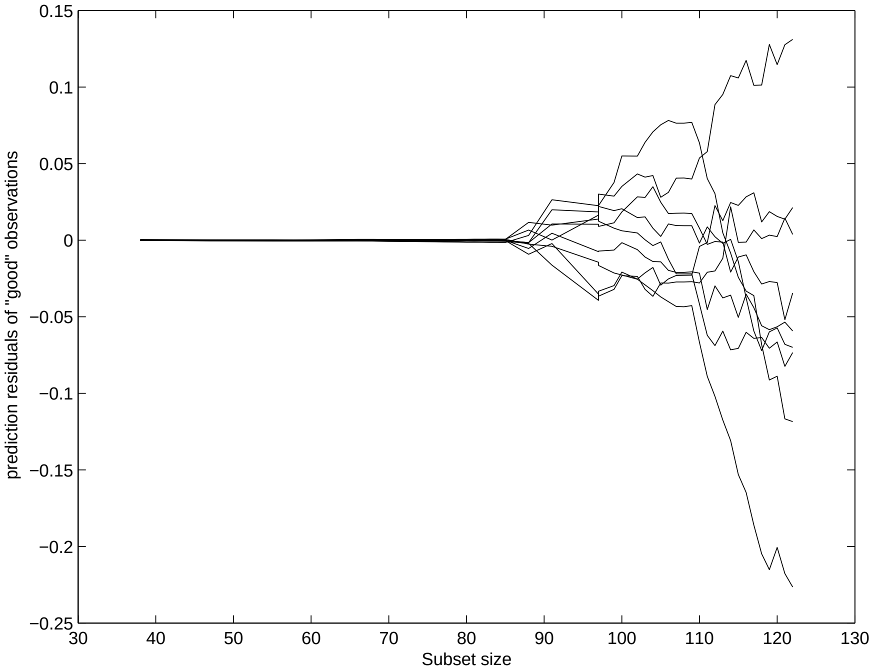


Figure 3: Hawkins' data prediction residuals for original good data as $k \to \infty$.

This is a form of prediction residual as k goes from 0 (only use "good" data) to $k = \infty$ which is OLS. We see significant changes as the clumps of "bad" data enter.

6 Conclusions

Penalty methods with the mean-shift outlier model provide a low-cost way to explore the impact of groups of unusual observations on regression models. Results obtained are comparable to those obtained from the forward search procedure of Atkinson and Riani [1]. In many applications today $p > n$ and a subset of p "good" observations cannot be found directly. This situation can be addressed by adjoining a full $n \times n$ identity matrix to the $n \times p$ matrix $\mathbf{X}$. This leads to an underdetermined system, but penalty methods can be quite effective there as well. Some examples are contained in Morgenthaler, et al. [9].

References

[1] Atkinson A.C., Riani M. (2000). *Robust diagnostic regression analysis.* Springer Verlag, New York.

[2] Belsley D.A., Kuh E., Welsch R.E. (1980). *Regression diagnostics.* New York, Wiley.

[3] Fan J., Li R. (2001). *Variable selection via nonconcave penalized likelihood and its oracle properties.* Journal of the American Statistical Association **96** (456), 1348 − 1360.

[4] Frank I., Friedman J. (1993). *A statistical view of some chemometrics regression tools (with discussion).* Technometrics **35** (2), 109 − 148.

[5] Hastie T., Tibshirani R., Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction.* Springer, New York.

[6] Hawkins D.M., Bradu D., Kass G.V. (1984). *Location of several outliers in multiple-regression data using elemental sets.* Technometrics, **26**, 197 − 208.

[7] Krasker W., Welsch R.E. (1982). *Efficient bounded-influence regression estimation.* Journal of the American Statistical Association **77**, 596 − 604.

[8] Li, Ker-Chau (1987). *Asymptotic optimality for C_p, C_L cross-validation and generalized cross-validation: discrete index set.* The Annals of Statistics **15**, (3), 958 − 975.

[9] Morgenthaler S., Welsch R.E., Zenide A. (2004). *Algorithms for robust model selection in linear regression.* Theory and Applications of Recent Robust Methods, M. Hubert, G. Pison, A. Struyf, S. Van Aelst (eds), Series: Statistics for Industry and Technology, Birkhauser, Basel (to appear).

[10] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* Wiley, New York.

[11] Tibshirani R. (1996). *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society, Series B **5**, 267 − 288.

Acknowledgement: This work was supported, in part, by the Center for Computational Research in Economics and Management Science at the Massachusetts Institute of Technology, Sloan School of Management. We would also like to acknowledge the technical assistance of Christine Liberty at MIT.

Address: L. McCann, R.E. Welsch, Massachusetts Institute of Technology, Sloan School of Management, 77 Massachusetts Avenue, E53-383, Cambridge, MA 02139, USA

E-mail: `rwelsch@mit.edu`

PREDICTION OF HIGH INCREASES IN STOCK PRICES USING NEURAL NETWORKS

Krzysztof Michalak and Piotr Lipinski

Key words: Neural networks, multi-layer perceptron, financial forecasting, stock market data prediction.

COMPSTAT 2004 section: Applications, Neural networks.

Abstract: This paper addresses the problem of stock market data prediction. It discusses the abilities of neural networks to learn and to forecast price quotations as well as proposes a neural approach to the future stock price prediction and detection of high increases or high decreases in stock prices. In order to validate the approach, a large number of experiments were performed on real-life data from the Warsaw Stock Exchange.

1 Introduction

In recent years, computational data analysis methods such as artificial intelligence or neural networks gain more and more popularity. Latest computer technologies enable an intensive development of these methods facilitating and speeding up computations. However, the large size of data as well as the long computing time remains the major constraint for computational methods.

Unlike many analytical approaches based on formal description of phenomena, that are the source of data, computational approaches focus mainly on data themselves. Since a large number of phenomena, not only natural phenomena, but also for instance stock market behavior, are complex and hard to describe by formal definitions, computational approaches turn out to be more efficient [4], [5], [6].

Financial data analysis still remains a grand challenge. Time series containing stock price quotations are chaotic to a large extent. Although these data are often suspected to be a random walk, there are a number of research that reject this hypothesis [3]. Nevertheless, stock price quotations are extremely hard to predict. The forecasting accuracy of fifty-five – sixty percent is often considered high.

This paper addresses the problem of stock market data prediction. It presents some investigations on price quotations forecasting using multi-layer perceptrons. The goal is to study abilities of neural networks to learn and to forecast price quotations as well as to propose and to validate a neural approach for stock market data prediction on real-life data from the Warsaw Stock Exchange. A large number of experiments were performed and described. Various sets of parameters concerning the network structure as well as the training process were considered.

This paper is structured in the following manner. Section 2 presents financial data from the Warsaw Stock Exchange and defines the problem. Section 3 describes the neural network structure and the training process. Some remarks on prediction of increases and decreases in stock prices are mentioned in Section 4. Section 5 discusses experiments performed on real-life data from the Warsaw Stock Exchange. Finally, Section 6 concludes the paper.

2 Data description and problem definition

This paper addresses the problem of the future stock price prediction on the basis of a number d of previous observations.

Let $p = (p_{-d+2}, p_{-d+3}, \ldots, p_{N+1})$ denote a vector containing price quotations of a given stock recorded at time $t = -d+2, -d+3, \ldots, N+1$ respectively (for the sake of simplicity of further definitions, the time instants and the coordinates of the vector p are numbered from $-d+2$ to $N+1$).

For a given time instant t, a number d of previous price quotations may be juxtaposed producing a data vector $x_t = (p_{t-d+1}, p_{t-d+2}, \ldots, p_t)$. In such a way, taking $t = 1, 2, \ldots, N$ in turn, a data matrix X of size $N \times d$ is obtained. The t-th row of the matrix X corresponds to a vector x_t containing d recent prices recorded until time t.

Let y_t denote the price recorded at the next moment in relation to time t, i.e. $y_t = p_{t+1}$. Let $y = (y_1, y_2, \ldots, y_N)'$.

This paper proposes a neural approach to prediction of the value y_t (the next-moment price quotation) on the basis of the observation x_t (the d previous price quotations). Beside the absolute value of price quotations, it may be also applied to the return rates. Moreover, the approach is especially efficient while detecting high increases (i.e. increases above a given threshold θ) or high decreases (i.e. decreases below a given threshold θ) in stock prices.

Evaluation is performed on four data sets. Each data set consists of a financial time series from the Warsaw Stock Exchange, which includes daily price quotations of a given stock over a period of about 2500 time instants from July 1994 until November 2003. Details on input data are presented in Table 1. The first column denotes the name of the stock, and the second column denotes the length of the time period, i.e. the number of stock price quotations recorded. Although the time period is the same for all experiments, the number of stock price quotations can differ, because for some stocks no quotations were recorded on some specific days.

3 Neural network structure and training process

The neural approach to the stock market data prediction proposed in this paper is based on a two-layer feed-forward perceptron.

Stock	Quotations
Tonsil	2660
Elektrim	2630
IndykPol	2259
DZ Bank Polska	2309

Table 1: Input data summary.

The size of the input layer is determined by the number d of previous price quotations which are used as the basis of each prediction. In experiments, d was usually equal to 64.

The hidden layer is composed of a number H of neuron units, where H varied between 10 and 5000 in experiments. All the neurons in the hidden layer use the *TANH* activation function [1].

The value returned by the neural network is required to be a single scalar. Therefore, the output layer contains a single neuron. This neuron uses logistic (sigmoid) activation function [1]. However, in some experiments, two other possibilities were considered, namely *softmax* and linear activation functions, but they are observed to cause the prediction to be heavily biased towards 0 or 1 and consequently impact the results adversely.

In the network considered, every two neurons in any two adjoining network layers are connected.

Before the training process starts, a training data set is chosen from input data. In experiments, the training data set consisted of 1000 first rows of the data matrix X and the corresponding part of the vector y, while remaining input data formed a test data set. The large number of observations guaranteed that the results obtained represented general nature of the stock market behavior and not some unusual phenomenon.

At the beginning of the training process, weights of connections between neurons are initialized randomly using the standard gaussian distribution. Next, training data vectors are fed one by one in a random order to the input of the neural network and signals of neurons in consecutive layers are propagated up to the output layer. The output generated by the perceptron is compared with the desired target data and weights of connections between neurons are optimized using scaled conjugate gradient algorithm [2], [7]. The entire process is repeated until the mean square error indicates that the training is completed. It usually happens after about 1000 iterations when MSE reaches about 10^{-15}.

In order to assess the quality of prediction, the mean square error evaluated on both the training and the test data sets is studied.

4 Prediction of increases or decreases

Although the neural network presented in the previous section is developed to predict the exact value of the future stock price, it may be also applied to detect high increases (i.e. increases above a specified threshold θ) or high decreases (i.e. decreases below a specified threshold θ) in stock prices after some modifications.

In this case, each target value, which is originally the next-moment price quotation, must be additionally processed to be equal to 1 if such an increase occurs, and to be equal to 0 otherwise. Moreover, the output of the neural network must be converted to 1 if a relative increase in stock prices exceeding the threshold θ is predicted, and to 0 otherwise.

The prediction is considered correct if and only if the converted output of the neural network is equal to the processed target value. In order to evaluate prediction reliability, a confusion matrix

$$M = \begin{pmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{pmatrix}$$

is used. Rows of the matrix M correspond to the target value desired, while columns correspond to the output obtained. Thus, m_{00} and m_{11} are the numbers of correct predictions, while m_{01} and m_{10} are the numbers of incorrect predictions. Confusion matrices are calculated for both the training data set (M_{train}) and the test data set (M_{test}) to investigate the ability of the neural network to adapt to training data and to extend the solution also to test data.

In order to assess the quality of the prediction further, three additional indices based on the confusion matrix M_{test} are derived:

$$P_0 = \frac{m_{00} + m_{11}}{N_{test}} \quad P_1 = \frac{m_{11}}{m_{01} + m_{11}} \quad P_2 = \frac{m_{11}}{N_{inc}}$$

where $N_{test} = m_{00}+m_{01}+m_{10}+m_{11}$ denote the total number of observations in the test data set, and $N_{inc} = m_{10}+m_{11}$ denote the number of observations where relative increases in stock prices exceed the threshold θ.

Thus, P_0 represents overall quality of predictions, P_1 reflects the percentage of the actual increases among all the increases predicted (correctly or incorrectly), and P_2 represents the percentage of the increases detected among all the increases occurred.

Since time series containing stock price quotations are chaotic to a large extent, a prediction is considered relatively efficient if the frequency of correct forecasts exceeds the N_{inc}/N ratio.

5 Experiments

In order to validate the approach proposed, a large number of experiments were performed. Evaluation was carried out on real-life data from the Warsaw Stock Exchange described in Table 2.

Stock	N	N_{test}	N_{inc}	N_{inc}/N
Tonsil	2660	1596	201	0.1259
Elektrim	2630	1566	159	0.1015
IndykPol	2259	1195	107	0.0895
DZ Bank Polska	2309	1245	90	0.0723

Table 2: Input data characterization.

Initial experiments concerned prediction of the exact value y_t of the future stock price on the basis of the observation x_t composed of $d = 64$ previous price quotations. In such an approach based on absolute price values, increases and decreases in stock prices were hard to detect, because their size was extremely small in relation to the absolute price value.

Next experiments were carried out using relative daily return rates instead of absolute price values, i.e. $r_t = \frac{p_t - p_{t-1}}{p_{t-1}}$ instead of p_t was used in construction of data vectors x_t and target values y_t.

Main experiments concerned prediction of increases above a specified threshold θ or decreases below a specified threshold θ in stock prices. Generally, the threshold value $\theta = 0.05$ was used, which ensures that reasonably many increases or decreases exceeding this limit exist in input data. However, other values of the θ parameter were also investigated. The results concerning high increases and high decreases were slightly different in some details. Nevertheless, due to size constrains, only experiments concerning high increases are presented further.

The aim of these experiments was to determine the optimal structure of the neural network that was apt to efficiently solve the problem considered. For each data set described in Table 2, the P_0, P_1 and P_2 indices were evaluated using a neural network with $H = 10, 50, 100, 500, 1000, 5000$ neuron units in the hidden layer.

Table 3 presents the summary of experiments for the *Tonsil* data set. Although the highest prediction accuracy was achieved for a very large neural network ($H = 5000$), the result is not really relevant, because this effect was caused by a small number of increases, which were usually predicted by such a large neural network. Excepting this experiment, significant results were obtained using neural networks with $H = 50$ and $H = 100$ hidden neurons.

Experiments for the *Elektrim* data set are summarized in Table 4. In this case, a small neural network ($H = 50$) produced the best results. It is worthy noticing that the prediction accuracy defined by P_1 decreases if the neural network becomes too large ($H = 5000$).

Summary of experiments for the *IndykPol* data set are shown in Table 5. One can see that a bit larger network ($H \geq 500$) perform more efficiently. Similarly as in the case of the *Elektrim* data set, the prediction accuracy decreases if the neural network becomes too large ($H = 5000$), but the decrease is smaller than previously. This effect may be related to a relatively good adaption of the neural network to the training data set.

H	P_0	P_1	P_2
10	0.7619	0.1255	0.1493
50	0.7832	0.1778	0.1990
100	0.7600	0.1679	0.2289
500	0.7575	0.1477	0.1940
1000	0.7588	0.1541	0.2562
5000	0.8340	0.3200	0.0400

Table 3: Results for the *Tonsil* data set.

H	P_0	P_1	P_2
10	0.7989	0.1389	0.1887
50	0.8487	0.1750	0.1321
100	0.8314	0.1329	0.1195
500	0.8359	0.1400	0.1195
1000	0.8697	0.1538	0.0629
5000	0.8851	0.0800	0.0126

Table 4: Results for the *Elektrim* data set.

H	P_0	P_1	P_2
10	0.7983	0.1278	0.2150
50	0.8042	0.1412	0.2336
100	0.8008	0.1497	0.2617
500	0.8343	0.1679	0.2150
1000	0.8109	0.1600	0.2617
5000	0.8569	0.1522	0.1308

Table 5: Results for the *IndykPol* data set.

Finally, Table 6 contains summary of experiments for the *DZ Bank Polska* data set. Although a small neural network with $H = 10$ hidden units produced the best results, it is worthy noticing that the overall prediction quality gradually increases with $H > 500$ for large neural networks ($H > 500$). Similarly as in the previous case, this effect may be related to the almost perfect adaption of the neural network to the training data set.

Average values of the P_0, P_1 and P_2 indices for various neural network structures are summarized in Table 7.

When considering significance of the prediction, the major emphasis must be put on P_1 and P_2 as they reflect the accuracy of the prediction and the extent to which stock price increases can be detected in advance. P_0 is less important as it concerns also detection of stock price decreases. Maximization of both P_1 and P_2 equally weighted can be simply achieved by maximizing the sum of the two indices. Experiments showed, that the maximum is reached

H	P_0	P_1	P_2
10	0.8498	0.1120	0.1556
50	0.8755	0.0886	0.0778
100	0.8987	0.1068	0.1222
500	0.8450	0.0360	0.0374
1000	0.8522	0.0566	0.0667
5000	0.8522	0.0566	0.0667

Table 6: Results for the *DZ Bank Polska* data set.

H	P_0	P_1	P_2
10	0.80223	0.12605	0.17715
50	0.82790	0.14565	0.16063
100	0.82273	0.13933	0.18308
500	0.81818	0.12290	0.14148
1000	0.82290	0.13113	0.16188
5000	0.86568	0.16498	0.06253

Table 7: Average values of P_0, P_1, P_2.

Stock	P_1	N_{inc}/N	$\frac{P_1}{N_{inc}/N}$
Tonsil	0.1679	0.1259	1.3336
Elektrim	0.1329	0.1015	1.3094
IndykPol	0.1497	0.0895	1.6726
DZ Bank Polska	0.1068	0.0723	1.4772

Table 8: Comparison of P_1 with the N_{inc}/N ratio.

when $H = 100$. This neural network structure was used to compare the value of prediction accuracy factor P_1 with the probability of high increase occurrences estimated by the ratio N_{inc}/N. Results of comparison are summarized in Table 8.

As the results contained in Table 8 show, proposed method of prediction allows for finding from 31% up to 67% more stock price increases than expected average.

6 Conclusion

This paper addresses the problem of stock market data prediction. It proposes a neural approach based on a two-layer feed-forward perceptron for the future stock price prediction on the basis of a number of previous observations as well as for detection of high increases or high decreases in stock prices.

Experiments performed on real-life data from the Warsaw Stock Exchange revealed some abilities of neural networks to learn and to forecast stock price quotations.

Results presented in this paper prove that prediction of high increases in stock prices may be carried out using neural networks of relatively small size, with about 100 neuron units in the hidden layer. Moreover, although the stock market financial time series are chaotic to a large extent, efficiency of such a prediction exceeds by far the expected average of price increases occurrences.

However, further research is needed, especially on different training methods as well as different data pre- and postprocessing methods.

References

[1] Bishop Ch.M. (1995). *Neural networks for pattern recognition.* Oxford University Press.

[2] Blue J.L., Grother P.J. (1992). *Training feed-forward neural networks using conjugate gradients.* SPIE **1661**, 179 – 190.

[3] Blasco N., Del Rio C., Santamaria R. (1997). *The random walk hypothesis in the Spanish stock market: 1980-1992.* Journal of Business Finance & Accounting **24** (5), 667 – 684.

[4] Kimoto T., Asakawa K., Yoda M., Takeoka M. (1990). *Stock market prediction system with modular neural networks.* International Joint Conference on Neural Networks (IJCNN 90), Piscataway, NJ, USA, 1 – 6.

[5] Korczak J., Lipinski P., Roger P. (2002). *Evolution strategy in portfolio optimization.* Artificial Evolution, Lecture Notes in Computer Science, Springer, **2310**, 156 – 167.

[6] Lipinski P. (2003). *Dependency mining in large sets of stock market trading rules.* Proceedings of 10th International Multi-Conference on Advanced Computer Systems, Technical University of Szczecin, Szczecin, Poland, (electronic edition, to appear also in a Kluwer edition).

[7] Fodslette Moller M. (1993). *A scaled conjugate gradient algorithm for fast supervised learning.* Neural Networks, **6** (4), 525 – 533.

Address: K. Michalak, Technical University of Wroclaw, Wroclaw, Poland
P. Lipinski, Institute of Computer Science, University of Wroclaw, Wroclaw, Poland
LSIIT, CNRS, Universite Louis Pasteur, Strasbourg, France
E-mail: `michalak@zacisze.wroc.pl`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

A NORMALISING TRANSFORMATION OF NONCENTRAL F VARIABLES WITH LARGE NONCENTRALITY PARAMETERS

Tetsuhisa Miwa

Key words: Industrial experiments, Taguchi's signal-to-noise ratio.
COMPSTAT 2004 section: Applications.

Abstract: In this article we consider a variance-stabilising transformation of noncentral F variables with large noncentrality parameters.

1 Introduction

Consider a noncentral F variable

$$F = (\chi_1^2/\nu_1)/(\chi_2^2/\nu_2) \ \sim \ F(\nu_1, \nu_2; \lambda), \tag{1}$$

where χ_1^2 is distributed as noncentral χ^2 with ν_1 degrees of freedom and noncentrality parameter λ, and χ_2^2 is independently distributed as central χ^2 with ν_2 degrees of freedom

Historically the noncentral F distribution has been studied in connection with the calculation of power in the analysis of variance. Then attention has primarily been directed towards small or moderate values of noncentrality parameters. In the latest development of signal-to-noise ratio approach proposed by Taguchi [5] in industrial experiments, we are forced to compare unknown noncentrality parameters whose values are usually very large [3].

In section 2, we propose a normalising transformation of noncentral F variables. Its variance is evaluated both by asymptotic expansion in $1/\lambda$ and by Monte Carlo simulation. In section 3, we formulate statistical tests for comparing two or more noncentrality parameters. Type I error rates are examined by Monte Carlo simulation.

2 Normalisation of noncentral F distributions

2.1 A variance-stabilising transformation

The mean and variance of F are given by

$$\mathrm{E}[F] = \frac{\nu_2}{\nu_2-2}\frac{\nu_1+\lambda}{\nu_1} = \mu, \quad \nu_2 > 2, \tag{2}$$

$$\begin{aligned}\mathrm{V}[F] &= \frac{2\nu_2^2[(\nu_1+\lambda)^2 + 2(\nu_2-2)(\nu_1+\lambda) - \nu_1(\nu_2-2)]}{\nu_1^2(\nu_2-2)^2(\nu_2-4)} \\ &= \frac{2}{\nu_2-4}\left[\mu^2 + 2\frac{\nu_2}{\nu_1}\mu - \frac{\nu_2^2}{\nu_1(\nu_2-2)}\right] = \sigma^2(\mu), \quad \nu_2 > 4. \end{aligned} \tag{3}$$

In this article we shall confine ourselves to cases with $\nu_2 > 4$. This condition, $\nu_2 > 4$, can be satisfied in many practical applications.

Laubscher [2] gives a transformation

$$\begin{aligned}
\tau(F) &= c \times \cosh^{-1}\left(\frac{F + \nu_2/\nu_1}{a}\right) \qquad (4)\\
&= c \times \left[-\ln a + \ln\left(F + \frac{\nu_2}{\nu_1} + \sqrt{F^2 + 2\frac{\nu_2}{\nu_1}F - \frac{\nu_2^2}{\nu_1(\nu_2-2)}}\right)\right],\\
& c = \sqrt{\frac{\nu_2-4}{2}}, \quad a = \frac{\nu_2}{\nu_1}\left(\frac{\nu_1+\nu_2-2}{\nu_2-2}\right)^{1/2},
\end{aligned}$$

from the relationship $(d/dt)\tau(\mu) \propto 1/\sigma(\mu)$. However, as Laubscher [2] points out, this transformation has a problem that $\tau(F)$ is only defined for

$$F \geq \frac{\nu_2}{\nu_1}\left[\left(\frac{\nu_1+\nu_2-2}{\nu_2-2}\right)^{1/2} - 1\right] > 0.$$

We shall consider the following modified transformation

$$g(F) = \ln\left(F + \frac{\nu_2}{\nu_1} + \sqrt{F^2 + 2\frac{\nu_2}{\nu_1}F}\right) \qquad (5)$$

which can be defined for any value of F $(0 \leq F < \infty)$. Being interested in large values of λ, we expand the mean and variance of $g(F)$ in terms of $1/\lambda$ and $1/\nu_2$:

$$\begin{aligned}
\mathrm{E}[g(F)] = g(\theta) + \frac{3}{3\nu_2-1} &+ O\left(\frac{1}{\nu_2^3}\right)\\
&- \frac{2}{\nu_1}\frac{1}{\theta} + O\left(\frac{1}{\theta^2}\right), \qquad (6)
\end{aligned}$$

$$\begin{aligned}
\mathrm{V}[g(F)] = \frac{2}{\nu_2-1} &+ O\left(\frac{1}{\nu_2^3}\right)\\
&+ \frac{6-2\nu_1}{\nu_1^2}\frac{1}{\theta^2} + O\left(\frac{1}{\theta^3}\right), \qquad (7)
\end{aligned}$$

where $\theta = \mathrm{E}[\chi_1^2/\nu_1] = (\nu_1+\lambda)/\nu_1$. The variance of $g(F)$ is of the order $1/\lambda^2$, so it quickly approaches the known limiting value $2/(\nu_2-1)$ as λ gets large. (The transformation (5) was first derived so that the terms of order $1/\lambda$ and $1/\lambda\nu_2$ are eliminated. However it turned out that all terms of order $1/\lambda\nu_2^i$ vanish in $\mathrm{V}[g(F)]$.)

2.2 The limiting distribution

We shall find the limiting distribution of $g(F)$ as λ approaches ∞. From (5) it follows that

$$g(F)-g(\theta)=\ln\left(\frac{\dfrac{\chi_1^2/(\nu_1\theta)}{\chi_2^2/\nu_2}+\dfrac{\nu_2}{\nu_1\theta}+\sqrt{\left(\dfrac{\chi_1^2/(\nu_1\theta)}{\chi_2^2/\nu_2}\right)^2+2\dfrac{\nu_2}{\nu_1\theta}\left(\dfrac{\chi_1^2/(\nu_1\theta)}{\chi_2^2/\nu_2}\right)}}{1+\dfrac{\nu_2}{\nu_1\theta}+\sqrt{1+2\dfrac{\nu_2}{\nu_1\theta}}}\right).$$

Since $\mathrm{E}[\chi_1^2/(\nu_1\theta)]=1$ and $\mathrm{V}[\chi_1^2/(\nu_1\theta)]=2(2\theta-1)/(\nu_1\theta^2)\to 0$ $(\theta\to\infty)$, $\chi_1^2/(\nu_1\theta)$ converges to unity in probability. Then $g(F)-g(\theta)$ converges to $Y=-\ln(\chi_2^2/\nu_2)$ in law. The skewness and kurtosis of Y are

$$\gamma_3 = \sqrt{\frac{2}{\nu_2}}\left\{1+\frac{1}{2\nu_2}+O\left(\frac{1}{\nu_2^2}\right)\right\}, \qquad (8)$$

$$\gamma_4 = \frac{4}{\nu_2}\left\{1+\frac{1}{\nu_2}+O\left(\frac{1}{\nu_2^2}\right)\right\}. \qquad (9)$$

2.3 Monte Carlo calculations of moments

From equation (7) we can expect that the variance of $g(F)$ is stable for large values of λ. We examined the mean and variance of $g(F)$ for moderate values of λ by Monte Carlo simulation (Tables 1 and 2). Each entry in Tables 1 and 2 gives the sample mean and variance calculated from 10,000 randomly generated $g(F)$ numbers. These tables show that the variance of $g(F)$ is stabilised around the limiting value $2/(\nu_2-1)$ for $\lambda/\nu_1\geq 10$.

	$\nu_2=10$		$\nu_2=20$		$\nu_2=40$	
λ/ν_1	$\bar{g}$	V_g	$\bar{g}$	V_g	$\bar{g}$	V_g
0	2.679	0.088	3.252	0.040	3.869	0.0192
10	3.687	0.230	3.983	0.108	4.392	0.0519
20	4.138	0.227	4.338	0.105	4.660	0.0522
30	4.443	0.225	4.597	0.104	4.858	0.0527
40	4.676	0.218	4.795	0.106	5.018	0.0505
50	4.869	0.223	4.955	0.104	5.155	0.0525
100	5.490	0.230	5.514	0.106	5.633	0.0515
$2/(\nu_2-1)$		0.222		0.105		0.0513

$\bar{g}$ and V_g are sample means and variances of 10,000 random $g(F)$ numbers.

Table 1: Monte-Carlo means and variances ($\nu_1=1$).

	$\nu_2 = 10$		$\nu_2 = 20$		$\nu_2 = 40$	
λ/ν_1	$\bar{g}$	V_g	$\bar{g}$	V_g	$\bar{g}$	V_g
0	1.673	0.135	2.063	0.058	2.559	0.0263
10	3.331	0.221	3.412	0.106	3.595	0.0508
20	3.920	0.218	3.938	0.105	4.052	0.0506
30	4.283	0.223	4.278	0.103	4.365	0.0519
40	4.548	0.222	4.539	0.104	4.597	0.0513
50	4.761	0.217	4.743	0.107	4.788	0.0516
100	5.429	0.222	5.388	0.103	5.408	0.0523
$2/(\nu_2 - 1)$		0.222		0.105		0.0513

$\bar{g}$ and V_g are sample means and variances of 10,000 random $g(F)$ numbers.

Table 2: Monte-Carlo means and variances ($\nu_1 = 5$).

3 Comparisons of noncentrality parameters

3.1 Comparison of two samples

We consider two independent noncentral F variables

$$F_i \sim F(\nu_1, \nu_2; \lambda_i), \quad i = 1, 2$$

and the null hypothesis

$$H_0 \colon \lambda_1 = \lambda_2$$

against the alternative

$$\begin{aligned} H_1 &: \quad \lambda_1 \neq \lambda_2 \quad \text{or} && (10) \\ H_2 &: \quad \lambda_1 > \lambda_2. && (11) \end{aligned}$$

From transformation (5), we obtain

$$g(F_i) \sim N\left(g(1 + \frac{\lambda_i}{\nu_1}) + \frac{3}{3\nu_2 - 1}, \frac{2}{\nu_2 - 1}\right), \quad i = 1, 2. \tag{12}$$

which gives the test statistic

$$Z = \frac{g(F_1) - g(F_2)}{\sqrt{4/(\nu_2 - 1)}}. \tag{13}$$

We can reject the null hypothesis $H_0 \colon \lambda_1 = \lambda_2$ when $|Z| > u_{\alpha/2}$ for the two-sided alternative (10), or when $Z > u_\alpha$ for the one-sided alternative (11), where u_α is the upper α point of the standard normal distribution.

When ν_2 is not large enough, the positive kurtosis (9) could affect the Type I error rate. We can refine the test statistic by the Cornish-Fisher expansion

$$Z' = Z - \frac{1}{24}\frac{\gamma_4}{2}(Z^3 - 3Z) = Z - \frac{1}{12\nu_2}(Z^3 - 3Z). \qquad (14)$$

(See e.g. Stuart and Ord [4] for the Cornish-Fisher expansion.)

3.2 Monte Carlo calculations of error rates

Type I error rates were calculated by Monte Carlo simulation for the same combinations of ν_1, ν_2 and λ as in Tables 1 and 2. Each entry in Tables 3 and 4 shows the percentage of false rejections out of 10,000 two-sided tests at the 5% level under the null hypothesis $H_0\colon \lambda_1 = \lambda_2 = \lambda$.

Tables 3 and 4 show that the Type I error rates are stable around 5% for $\lambda/\nu_1 \geq 10$. With small values of ν_2, the simple normalised statistic (13) is a little liberal, showing error rates slightly higher than 5%. This liberalness is corrected by the Cornish-Fisher expansion (Column II).

	$\nu_2 = 10$		$\nu_2 = 20$		$\nu_2 = 40$	
λ/ν_1	I	II	I	II	I	II
0	0.79	0.74	0.47	0.47	0.36	0.36
10	5.52	5.35	5.53	5.43	4.90	4.87
20	5.12	4.96	4.72	4.67	5.19	5.16
30	5.29	5.14	5.26	5.17	5.35	5.30
40	5.28	5.13	5.02	4.97	4.78	4.76
50	5.52	5.41	5.01	4.88	4.88	5.05
100	5.13	4.92	5.66	5.59	5.03	4.99

Column I: simple normalised statistic (13)
Column II: Cornish-Fisher expansion (14)

Table 3: Type I error rates by simulation ($\nu_1 = 1$).

3.3 Multiple comparisons of noncentrality parameters

Let F_i $(1 \leq i \leq k)$ be k independent noncentral F variables

$$F_i \sim F(\nu_1, \nu_2; \lambda_i), \quad 1 \leq i \leq k.$$

The uniformity null hypothesis

$$H_0\colon \lambda_1 = \cdots = \lambda_k \qquad (15)$$

can be rejected when

$$\sum_{i=1}^{k}(g(F_i) - \bar{g})^2 > \frac{2}{\nu_2 - 1}\chi^2_{k-1,\alpha}, \quad \bar{g} = \sum g(F_i)/k, \qquad (16)$$

	$\nu_2 = 10$		$\nu_2 = 20$		$\nu_2 = 40$	
λ/ν_1	I	II	I	II	I	II
0	1.44	1.39	0.98	0.96	0.71	0.71
10	5.08	4.90	4.68	4.62	4.87	4.84
20	5.11	4.82	5.34	5.22	5.16	5.14
30	4.69	4.58	4.97	4.89	4.76	4.73
40	5.10	5.08	4.74	4.65	5.28	5.20
50	5.18	4.96	5.06	4.96	4.91	4.90
100	5.18	5.08	5.36	5.30	4.82	4.78

Column I: simple normalised statistic (13)
Column II: Cornish-Fisher expansion (14)

Table 4: Type I error rates by simulation ($\nu_1 = 5$).

where $\chi^2_{k-1,\alpha}$ is the upper α point of χ^2 distribution with $k-1$ degrees of freedom. If we are interested in pairwise comparisons of noncentrality parameters, we can use the criterion

$$\max_{i,j} |g(F_i) - g(F_j)| > \sqrt{2/(\nu_2 - 1)}\; q_{k,\infty,\alpha}, \tag{17}$$

where $q_{k,\infty,\alpha}$ is the upper α point of range distribution.

Other multiple comparison procedures, e.g. Dunnett's procedure or Ryan-Einot-Gabriel-Welsh procedure, can be applied similarly. (See e.g. Hsu [1] for a general discussion of multiple comparison procedures.)

References

[1] Hsu J.C. (1996). *Multiple comparisons: theory and methods.* London: Chapman and Hall.

[2] Laubscher N.F. (1960). *Normalizing the noncentral t and F distributions.* Annals of Mathematical Statistics **31**, 1105 – 1112.

[3] Miwa T. (1994). *Statistical inference on non-centrality parameters and Taguchi's SN ratios.* Proceedings of International Conference – Statistics in Industry, Science and Technology, Tokyo, 66 – 71.

[4] Stuart A., Ord J.K. (1994). *Kendall's advanced theory of statistics.* London: Edward Arnold, 5th ed. **1**.

[5] Taguchi G. (1977). *Jikken Keikaku Hou.* Tokyo: Maruzen, 3rd ed. **2** ("The Desing of Experiments" in Japanese).

Acknowledgement: The author gratefully acknowledges the reviewers for their helpful comments and suggestions.

Address: T. Miwa, National Institute for Agro-Environmental Sciences, 3-1-3 Kannondai, Tsukuba 305-8604, Japan

E-mail: miwa@niaes.affrc.go.jp

COMPSTAT'2004 Symposium

CLUSTERING METHODS FOR FUNCTIONAL DATA: K-MEANS, SINGLE LINKAGE AND MOVING CLUSTERING

Masahiro Mizuta

Key words: Functional data analysis, MST, computer graphics.

COMPSTAT 2004 section: Functional data analysis.

Abstract: In this paper, we deal with methods for clustering for functional data and propose a new method for them. Clustering is a fundamental method for data analysis and data mining. The target data sets of clustering are very varieties from the viewpoint of data structure. Conventional methods of clustering assume that the data structures are multidimensional or dissimilarities between objects. Nowadays, functional data are paid attention for their abilities to represent data attributes.

We have already proposed several clustering methods for functional data. At first, the overview of functional clustering techniques, including functional k-means method and functional single linkage method, are shown. Then, we propose a new method: moving functional clustering.

1 Introduction

In most conventional data analysis methods, we assume that data are regarded as a set of numbers with some structures, i.e. a set of vectors or a set of matrices etc. Nowadays, we must often analyze more complex data. One type of the complex data is functional data structure; data themselves are represented as functions. Typical functional data are time series data [2], [4]. There are many other functional data, of course. Ramsay and Silverman [10] have studied function data analysis (FDA) as the analysis method to function data from the 1990's. They published another book on FDA [11] in 2002.

Data mining has been a field of active research and is defined as the process of extracting useful and previously unknown information out of large complex data sets. Cluster analysis is a powerful tool for data mining. The purpose of cluster analysis is to find relatively homogeneous clusters of objects based on measured characteristics. There are two main approaches of cluster analysis: hierarchical clustering methods and nonhierarchical clustering methods. Single Linkage is a kind of hierarchical clustering and is a fundamental method. k-means is a typical nonhierarchical clustering method.

We have studied clustering methods for functional data: functional single linkage [8] and functional k-means [9] etc. In this paper, we deal with functional clustering methods, including these two methods.

2 Functional data analysis and cluster analysis

In this section, functional data analysis (FDA) and cluster analysis are described briefly.

2.1 Functional data

Typical functional data are given by a set of functions: $\{\vec{x}_i(t); i = 1, 2, \ldots, n\}$. There are many techniques in FDA including functional regression analysis [3], [4], [12], functional principal components analysis, functional discriminant analysis, functional canonical correlation and functional clustering [1], [13], [14], [15]. We can get excellent lists of bibliography on FDA from [10] and `http://ego.psych.mcgill.ca/misc/fda/index.html`. We have proposed several methods for functional data: functional multidimensional scaling [6], extended functional regression analysis [12] etc. Functional data can be considered as infinity-dimensional data. Most methods in the book (Ramsay and Silverman [10]) for functional data are based on an approximation with finite expansions of the functions with basis functions. Once the functional data can be thought as finite linear combinations of the basis functions, functional data analysis methods for the functional data are almost the same as those of conventional data analysis methods. But, there is some possibility of using different approaches.

2.2 Cluster analysis

Sometimes, we divide methods of cluster analysis into two groups: hierarchical clustering methods and nonhierarchical clustering methods.

Hierarchical clustering refers to the formation of a recursive clustering of the objects data: a partition into two clusters, each of which is itself hierarchically clustered. We usually use data set of dissimilarities or distances; $S = \{s_{ij}; i, j = 1, 2, \ldots, n\}$, where s_{ij} are dissimilarity between object i and object j, and n is the size of the objects. Single linkage is a typical hierarchical clustering method. We start with each object as a single cluster and in each step merge two of them together. In each step, the two clusters that have minimum distance are merged. At the first step, the distances between clusters are defined by the distance between the objects. However, after this, we must determine the distances between new clusters. In single linkage, the distance between two clusters is determined by the distance of the two closest objects in the different clusters. The results of hierarchical clustering are usually represented by dendrogram. The height of dendrogram represents the distances between two clusters. It is well known that the result of Single Linkage and the minimal spanning tree (MST) [1] are equivalent from the computational point of view[5].

[1] Among all the spanning tree of weighted and connected graph, the one with the least total weight is called MST.

Nonhierarchical clustering partitions the data based on a specific criterion. Most nonhierarchical clustering methods do not deal with a set of dissimilarities directly. They use a set of p-dimensional : $X = \{\vec{x}_i; i = 1, 2, \ldots, n\}$. k-means method is a kind of nonhierarchical clustering and is to choose initial seeds points and assign the cases to them using a minimum the trace of within variances.

3 Functional clustering

We will describe clustering methods for functional data. Typically, there are two types of functional data to be applied to clustering methods. The first type of functional data is p-dimensional functions. The other type is a set of dissimilarity functions. We denote the functions for n objects depending on a variable t as $X(t) = \{\vec{x}_i(t)\}(i = 1, 2, \ldots, n)$ and denote dissimilarity data as $S(t) = \{s_{ij}(t); i, j = 1, 2, \ldots, n\}$, where $s_{ij}(t)$ are dissimilarity functions between object i and object j. We also use the notations $X = \{\vec{x}_i\}$ and $S = \{s_{ij}\}$ for ordinary data, as defined in section 2.2.

3.1 Using conventional clustering methods

Simple idea for clustering methods for functional data is to transform functional data to ordinary data and we apply conventional clustering methods to these data.

There are several methods to derive ordinary data $S = \{s_{ij}\}$ or $X = \{\vec{x}_i\}$ from functional dissimilarity data $S(t)$ or p-dimensional functional data $X(t)$. The most natural method may be to use integration in the domain or max operator:

$$\begin{aligned} s_{ij} &= \int \| \vec{x}_i(t) - \vec{x}_j(t) \|^2 \, dt, \\ s_{ij} &= \int s_{ij}(t) dt, \\ s_{ij} &= \max_t s_{ij}(t), \\ \vec{x}_i &= \int \vec{x}_i(t) dt. \end{aligned}$$

Then we can use ordinary clustering method to set of dissimilarity data $S = \{s_{ij}\}$ or p-dimensional data $X = \{\vec{x}_i\}$.

But the results of this simple method are not depended on the variable t. In order to derive clusters depended on t, i.e. *functional clustering.* We developed several methods: functional k-means and functional Single Linkage. We will describe these methods in the following subsections.

3.2 Functional k-means method

Functional k-means method is proposed in this section. We assume that we have p-dimensional functional data $X(t) = \{\vec{x}_i(t)\}(i = 1, 2, \ldots, n)$. It is realistic that values $X(t_j), j = 1, 2, \ldots, m$ are given. We restrict ourselves to two dimensional functional data and the one dimensional domain. The number of clusters k is prespecified as a user parameter.

The idea of functional k-means method is repetitive procedure of the conventional k-means method. At first, we apply conventional k-means method to $X(t_j)$ for each t_j. Then, we adjust labels of clusters. Even if we fix clustering method, there is freedom of labeling. We denote $C_i(t)$ as the cluster label of the i-th object at t for fixed K. We discuss about adjusting the cluster labeling of $C_i(t_2)$ with fixed $C_i(t_1)$. $C_i^*(t_2)$ are new cluster labels of the object that $\Sigma_i \sharp\{C_i(t_1) = C_i^*(t_2)\}$ takes the maximum value, where $\sharp$ indicates the size of the set. A simple method for adjusting the cluster labels is to use the cluster centers of the previous clustering for initial guesses for the cluster centers.

We must pay attention to the fact that even if two objects belong to the same cluster at t_j, it is possible that the two objects belong to different clusters at t'_j.

The results of the proposed functional k-means method can be represented graphically. We use the three dimensional space for the representation. Two dimensions are used for the given functional data, one is for t, and the clusters are shown with colors. Figure 1 is a conceptual display of the display for 7 functional data with 3 clusters. The figure is a snapshot of computer display. If we may use dynamic graphics, the results can be analyzed more effectively with interactive operations: rotations, slicing, zooming etc. We can analyze the result deeply.

3.3 Functional single linkage method

We introduce the algorithm of single linkage for functional dissimilarity data $S(t) = \{s_{ij}(t); i, j = 1, 2, \ldots, n\}$. The basic idea of this method is that we apply conventional single linkage to $S(t)$ and get functional MST, say MST(t). Then we calculate functional configuration and adjust labels of objects. We mention the detail of these two steps in the following.

We must construct a functional configuration of vertices of MST(t); functional multidimensional scaling [6] is useful to get them. The MST(t) is represented on the 2 dimensional space. It is possible to represent the results of functional single linkage for the given number of clusters K. Especially when we can use dynamic graphical display, the representation is much interactive.

But, even if we fix hierarchical clustering method, there is freedom of labeling of clusters. We adjust the cluster labels used the same method as functional k-means, described in section 3.2.

We show an example of the proposed method with 150 objects represented

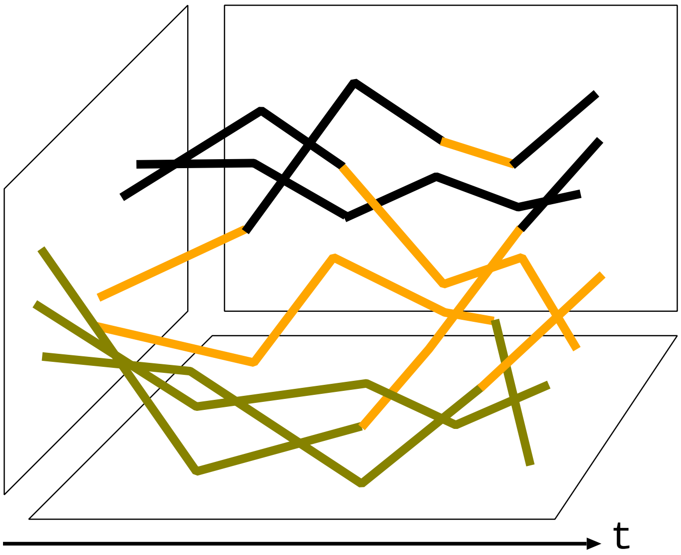


Figure 1: Functional k-means (concept).

by functional dissimilarities: $\{s_{ij}(t); i,j = 1,2,\ldots,150\}$. We use several values of t: $t_1, t_2, \ldots, t_7$. Figure 2 (a) shows the dendrograms for the results of conventional single linkage. It is very difficult to analyze the relations among them. Figure 2 (b) shows the MSTs of the result with the proposed method. Unfortunately we can not tell you the colors of the figures.

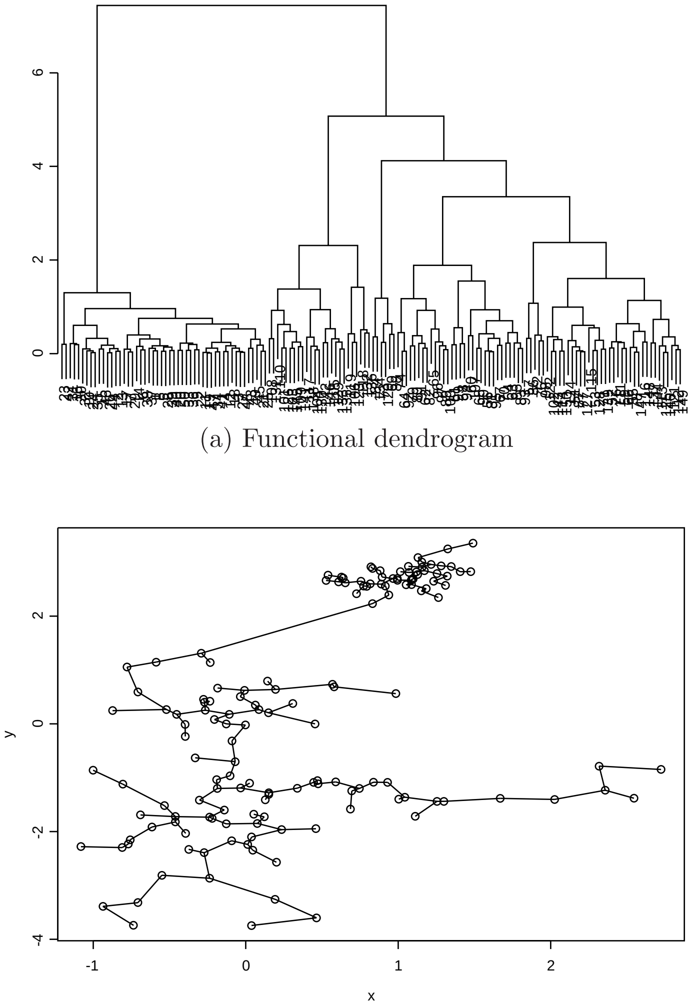
4 Moving functional clustering

In the previous sections, we deal with clustering method for functional data. When we must analyze one dimensional functional data $x_i(t)(i = 1, \cdots, n)$, functional k-means method or functional single linkage method can be applicable formally. But, these methods do not use multidimensional information of the data. We propose a method for clustering for one dimensional functional data.

In order to actively use the information from the functional data, we apply *windows* of the domain of the functions. We define the dissimilarities functions of the two functions:

$$s_{ij}(t)_d = \int_{t-d}^{t+d} (x_i(u) - x_j(u))^2 du.$$

These $s_{ij}(t)_d$ are called moving dissimilarities functions with windows range d. $s_{ij}(t)_d$ represent the degree of the closeness between functions $x_i(t)$ and $x_j(t)$ in the interval $[t-d, t+d]$.



(a) Functional dendrogram

(b) Functional minimal spanning tree (MST)

Figure 2: Functional Single Linkage (snap shot).

We can apply functional single linkage clustering, in the subsection 3.3, to these functional dissimilarities $s_{ij}(t)_d$.

The parameter d affects the $s_{ij}(t)_d$ and the clusters as results. But, in general, cluster analysis is an exploratory method. So the freedom of d is important for finding interesting results. For example, when we analysis stocks data, we can see interesting structures using $d = 3$ (days).

We assumed the dimension of the domain of the functions is one. It is easy to extend this method to more than one dimensional domain. The integration domain in the definition of $s_{ij}(t)_d$ is changed to multidimensional neighborhood; sphere or hyper sphere.

5 Concluding remarks

We discussed about clustering methods for functional data. We would summarize the method of functional clustering.

If you would like to use conventional clustering methods, the functional data are transformed to the ordinal data set with integration on t or maximum operation. When the functional dissimilarities data $S(t)$ are given, functional single linkage method is available. The results of functional clustering can be represented by functional minimum spanning tree (MST). When p-dimensional functions $X(t)$ are given, functional k-means is a candidate for the method. Specially when $p = 1$, moving functional clustering is helpful in analysis.

References

[1] Abraham C., Cornillon P.A., Matzner-Lober E., Molinari N. (2003). *Unsupervised curve clustering using B-splines.* Scand. J. Statist **30**, 581 − 595.

[2] Bosq D. (2000). *Linear processes in functional spaces: theory and applications.* Lecture Notes in Statistics **149**, New York: Springer-Verlag.

[3] Cardot H., Ferraty F., Sarda P. (2003). *Spline estimators for the functional linear model.* Statistica Sinica **13**, 571 − 591.

[4] Ferraty F., Vieu P. (2004). *Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination.* Nonparametric Statistics, **16** (1-2), 111 − 125.

[5] Gower J.C., Ross G.J.S. (1969). *Minimum spanning trees and single linkage cluster analysis.* Appl. Stat. **18**, 54 − 64.

[6] Mizuta M. (2000). *Functional multidimensional scaling.* Proceedings of the Tenth Japan and Korea Joint Conference of Statistics, 77 − 82.

[7] Mizuta M. (2002). *Cluster analysis for functional data.* Proceedings of the 4th Conference of the Asian Regional Section of the International Association for Statistical Computing, 219 − 221.

[8] Mizuta M. (2003). *Hierarchical clustering for functional dissimilarity data.* Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics, Volume V, 223 − 227.

[9] Mizuta M.(2003). *K-means method for functional data.* Bulletin of the International Statistical Institute, 54th Session, Book 2, 69 − 71.

[10] Ramsay J.O., Silverman B.W. (1997). *Functional data analysis.* New York: Springer-Verlag.

[11] Ramsay J.O., Silverman B.W. (2002). *Applied functional data analysis – methods and case studies.* New York: Springer-Verlag.

[12] Shimokawa M., Mizuta M., Sato Y. (2000). *An expansion of functional regression analysis (in Japanese)*. Japanese Journal of Applied Statistics **29-1**, 27 – 39.

[13] Tarpey T., Kinateder K.K.J. (2003). *Clustering functional data*. J. of Classification **20**, 93 – 114.

[14] Tokushige S., Inada K., Yadohisa H. (2001). *Dissimilarity and related methods for functional data* Proceeding of the International Conference on New Trends in Computational Statistics with Biomedical Applications, 295 – 302.

[15] Tokushige S., Yadohisa H., Inada K. (2002). *Fuzzy k-means Clustering for Functional Data. (In Japanese)*. Symposium of Statistical Models and their Applications, Hakodate.

Acknowledgement: This work was supported by Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (B) 14380122.

Address: M. Mizuta, Information Initiative Center, Hokkaido University, Sapporo 060-0811, Japan

E-mail: `mizuta@cims.hokudai.ac.jp`

BAYESIAN PREDICTION FOR A NOISY LOG-GAUSSIAN SPATIAL MODEL

M. Mohammadzadeh and Khaledi M. Jafari

Key words: Log-Gaussian, noisy spatial data, Bayesian predictive distribution.

COMPSTAT 2004 section: Spatial statistics.

Abstract: Sometimes a logarithm transformation can be successful in representing positive spatial data which may contain noise. In this case, we consider a noisy Log-Gaussian random field and use the Bayesian approach to extend the prediction problem of a noisy Gaussian model. Then the posterior distributions and the Bayesian predictive distribution are analytically determined by a discretisation method. Finally, a Bayesian spatial predictor based on the absolute error loss function is derived.

1 Introduction

Usually spatial data, collected from some applied disciplines such as petroleum engineering, civil engineering, geography, geology, meteorology and epidemiology, are thought as samples from realizations of a random field $S(\cdot) = \{S(t)\ ,\ t \in D\}$, where D is an index set in R^d, $d \geq 1$. A common scientific purpose in spatial data analysis is the prediction of the random field $S(\cdot)$ in an unmeasured site t_0, say $S(t_0)$, based on measured data in some sampled sites $t_1, \cdots, t_n$ within D. Usually, because of various reasons, in particular measurement error, $S(\cdot)$ is not directly observable. In this case, instead of $S = (S(t_1), \ldots, S(t_n))$, the noisy measurements $Z = (Z(t_1), \ldots, Z(t_n))$ are taken at the sample locations $t_1, \cdots, t_n$. Often, with Gaussian assumption for $S(\cdot)$, the spatial prediction of $S(t_0)$ is carried out. Cressie [4] and Chiles and Delfiner [2] considered this matter in a noisy Gaussian model by using the frequentist approach. Diggle and Ribeiro [7] proposed applying the Bayesian approach due to the fact that the model parameters uncertainty is fully accounted for when performing prediction. But, in some applications, data give evidence of non-Gaussian features. In this case, the Gaussian model is inappropriate. For modeling positive data with skewed sampling distribution, De Oliveira [5] proposed a Bayesian transformed Gaussian (BTG) model by applying a family of parameterizations transformations, such as the Box-Cox family. There are two complex problem related to this model. First, the selection of a specific transformations family may influence the prediction performance. Second, this model demand a specification of a joint prior distribution for the model parameters. Because the interpretation of the other parameters depends on the value of transformation parameter, determination of the joint prior distribution will be difficult. Nevertheless, he used an improper prior, but it is an unusual prior, in the sense that it depends on the observations. Therefore, it is undesirable to consider the transformation parameter as a random variable, consequently the use of BTG model may not be sensible. Sometimes, a logarithm transformation can be successful in representing this type of data. In this case, we consider a noisy Log-Gaussian random field and use the Bayesian approach to extend the prediction problem of the noisy Gaussian model. In this matter, the posterior distributions and the Bayesian predictive distribution are analytically determined by a discretisation method. Finally, a Bayesian spatial predictor based on the absolute error loss function is derived. Then a noisy Log-Gaussian model is described in section 2. Section 3 deals with the determination of a Bayesian spatial prediction for the considered model. A numerical example, in section 4 illustrates the calculation of the predictor. Finally, a discussion is presented in section 5.

2 Statistical model

Let $\{S(x), x \in D\}$ be a Log-Gaussian random field of interest such that $\{Y_s(t) = \ln S(t) \ , \ t \in D\}$ is a (nearly) Gaussian random field with the following mean and covariance functions

$$\begin{aligned} E[Y_s(t)] = f'(t)\beta &= \Sigma_{j=1}^{p}\beta_j f_j(t) \\ Cov[Y_s(u), Y_s(t)] &= \sigma^2\rho(u,t;\theta) \end{aligned}$$

where $\beta = (\beta_1, \ldots, \beta_p)' \in R^p$ are unknown regression parameters, $f(t) = (f_1(t), \ldots f_p(t))'$ are known location-dependent covariates, $\sigma^2 = Var[Y_s(t)]$ is the fixed variance of Y_s, $\rho(u,t;\theta)$ is a spatial correlation function and $\theta = (\theta_1, \ldots, \theta_q)' \in \Theta \subseteq R^q$ are parameters that control geometric aspects of the random field, such as the range and smoothness, as well as the other aspects of the spatial data association structure. Alternatively, it is assumed that the random field $S(\cdot)$ is not directly observable. Instead, the random vector $Z = (Z(t_1), \ldots, Z(t_n))$ represents the data measured at the sampling locations $t_1, \ldots, t_n \in D$ such that

$$\ln Z(t_i) = \ln S(t_i) + \varepsilon(t_i) \qquad i = 1, \ldots, n$$

where $\varepsilon(\cdot)$, representing the noise, is a zero mean Gaussian white noise random field with $Var[\varepsilon(t)] = \sigma^2\alpha^2$ and is independent of $S(.)$. The parameter $\tau^2 = \sigma^2\alpha^2$ is called nugget effect. By the stated assumptions, we have

$$Y_z = \ln(Z) = (\ln Z(t_1), \ldots, \ln Z(t_n)) \sim N_n(X\beta, \sigma^2 V_\phi)$$

where $X = (f_j(t_i))$ is a known full rank $n \times p$ matrix, $n > p$, $V_\phi = \Sigma_\theta + \alpha^2 I$ is a positive definite $n \times n$ matrix with $\phi = (\theta, \alpha^2)$, $\Sigma_\theta = (\rho(t_i, t_j; \theta))$ and I is the identity matrix. The likelihood function of the model parameters $\eta = (\beta, \sigma^2, \phi)$ based on the positive observed data $z = (z(t_1), \ldots, z(t_n))$ is given by

$$L(\eta; z) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} |V_\phi|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\ln z - X\beta)' V_\phi^{-1}(\ln z - X\beta)\right\} J$$

where $J = (\prod_{i=1}^n z_i)^{-1}$ is the Jacobian of the logarithm transformation. The joint distribution of $(Y_s(t_0), Y_z)$ is given by

$$N_{n+1}(\left(\begin{array}{c} f'(t_0)\beta \\ X\beta \end{array}\right) \ , \ \sigma^2 \left(\begin{array}{cc} 1 & r_\theta \\ r'_\theta & \Sigma_\theta \end{array}\right) + \tau^2 I)$$

where $r_\theta = (\rho_\theta(t_i, t_0))$ is a $n \times 1$ vector. Then, it can be shown that $Y_s(t_0)$ given z and η normally distributed as

$$(Y_s(t_0)|z, \eta) \sim N(\mu_1, \sigma^2\rho_1)$$

where

$$\begin{array}{rcl} \mu_1 & = & f'(t_0)\beta + r'_\theta V_\phi^{-1}(\ln z - X\beta), \\ \rho_1 & = & (1 + \alpha^2 - r'_\theta V_\phi^{-1} r_\theta). \end{array}$$

We are interested in predicting of noiseless random field $S(\cdot)$, not that of the noisy $Z(\cdot)$, using the noisy observed data. Given η is known, the predictor distribution is a Log-Normal distribution, $LN(\mu_1, \sigma^2\rho_1)$, with density function

$$f(s_o|z, \eta) = \left(\frac{1}{2\pi\sigma^2 s_0^2 \rho_1}\right)^{1/2} \exp\left\{-\frac{(\ln s_0 - \mu_1)^2}{2\sigma^2\rho_1}\right\} \ , \ s_0 > 0. \qquad (1)$$

Now, the optimal predictor corresponding to the absolute error loss function is given by

$$\hat{S}(t_0) = Median\ of\ (S(t_0)|z, \eta) = e^{\ \mu_1}. \qquad (2)$$

When η is unknown, Christensen et al. [3] proposed the plug-in approach, in which a maximum likelihood estimate of η is replaced in (2). Due to the fact that parameter uncertainty can be considered in Bayesian prediction, Handcock and Stein [9], Diggle et al. [8] and De Oliveira & Ecker [6] have studied different aspects of this matter. In this paper, we use the Bayesian approach such that the parameter uncertainty is fully accounted for when performing spatial prediction.

3 Bayesian satial prediction

In Bayesian analysis, the parameters are thought as random variables, therefore we need to specify prior distribution for the model unknown parameters. Berger et al. [1] showed that for the case of Gaussian random fields, some of the priors proposed in the literature, give rise to improper posterior distributions. Here, we consider proper priors for parameters which assure a proper posteriors. We also assume that the vector ϕ to be independent of the parameters β and σ^2, and θ to be independent of α^2, so that the prior densities satisfy

$$\pi(\eta) = \pi(\beta, \sigma^2, \theta, \alpha^2) = \pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\theta)\pi(\alpha^2)$$

Because of analytical convenience, we choose the following conjugate priors for β and σ^2

$$\begin{aligned} \pi(\beta|\sigma^2) &\sim N_p(\beta_0, \sigma^2 V_0) && (3)\\ \pi(\sigma^2) &\sim \chi^2_{Inv}(a,b) \qquad \left(i.e.\ \frac{ab}{\sigma^2} \sim \chi^2(a) \right) && (4) \end{aligned}$$

also $\pi(\theta)$ and $\pi(\alpha^2)$ are arbitrary proper priors. Now, the joint posterior distribution of the model parameters is given by

$$\pi(\eta|z) = \pi(\beta|z, \sigma^2, \phi)\pi(\sigma^2|z, \phi)\pi(\phi|z). \qquad (5)$$

It can be shown that

$$\begin{aligned} (\beta|z, \sigma^2, \phi) &\sim N_p(\beta_1, \sigma^2 V_1) && (6)\\ (\sigma^2|z, \phi) &\sim \chi^2_{Inv}(a+n, S_1^2) && (7) \end{aligned}$$

where

$$\begin{aligned} \beta_1 &= (V_0^{-1} + X' V_\phi^{-1} X)^{-1}(V_0^{-1}\beta_0 + X' V_\phi^{-1} \ln z)\\ V_1 &= (V_0^{-1} + X' V_\phi^{-1} X)^{-1}\\ S_1^2 &= \frac{ab + nS_2^2 + \beta_2' V_2 \beta_2 + \beta_0' V_0^{-1} \beta_0 - m' V_1 m}{a+n} \end{aligned}$$

with

$$\begin{aligned} S_2^2 &= \frac{1}{n}(\ln z - X\beta_2)' V_\phi^{-1}(\ln z - X\beta_2)\\ \beta_2 &= (X' V_\phi^{-1} X)^{-1} X' V_\phi^{-1} \ln z\\ V_2 &= (X' V_\phi^{-1} X)^{-1}\\ m &= (V_2^{-1}\beta_2 + V_0^{-1}\beta_0). \end{aligned}$$

To compute $\pi(\phi|z)$, we note that

$$\begin{aligned} \pi(\phi|z) &= \frac{\pi(\beta, \sigma^2, \phi|z)}{\pi(\beta|z, \sigma^2, \phi)\pi(\sigma^2|z, \phi)}\\ &\propto \frac{f(z|\beta, \sigma^2, \phi)\pi(\beta|\sigma^2)\pi(\sigma^2)}{\pi(\beta|z, \sigma^2, \phi)\pi(\sigma^2|z, \phi)}\pi(\phi). \end{aligned}$$

Applying (4) and (5) in the numerator of the above expression, we get

$$\pi(\phi|z) \propto g(\phi; z)\pi(\theta)\pi(\alpha^2)$$

where

$$g(\phi; z) = |V_1|^{\frac{1}{2}}|V_\phi|^{-\frac{1}{2}}(S_1^2)^{-\frac{a+n}{2}} \tag{8}$$

and the proportionality constant is independent of ϕ. However, this expression does not define a standard probability distribution. To have the feasible computation of $\pi(\phi|z)$, we use a discretisation method by choosing a set of values $A = \{\phi_i\}_{i=1}^m$, with $\phi_i \in \Theta \times [0, \infty)$, and assigning a discrete joint prior on A, say $\pi(\phi_i)$. Then we have

$$\pi(\phi_i|z) = \frac{g(\phi_i; z)\pi(\theta_i)\pi(\alpha_i^2)}{\sum_{j=1}^m g(\phi_j; z)\pi(\theta_j)\pi(\alpha_j^2)}.$$

Now, we consider the prediction of $S(t_0)$ based on the Bayesian predictive distribution, defined by

$$\begin{aligned} f(s_o|z) &= \int_\Omega f(s_o, \eta|z)d\eta \\ &= \int_\Omega f(s_o|z, \eta)\pi(\eta|z)d\eta \end{aligned} \tag{9}$$

where $\Omega = R^p \times (0, \infty) \times \Theta \times [0, \infty)$ and $f(s_o|z, \eta)$, $\pi(\eta|z)$ were given in (1) and (3). Since the analytical solution of the integration in (6) is very difficult, first the distributions of $(s_o|z, \sigma^2, \phi)$ and $(s_o|z, \phi)$ are identified, then they will be used to determine the distribution of $(s_o|z)$. If (1) and (4) are used, it can be shown that $f(s_o|z, \sigma^2, \phi)$ is a log-normal distribution

$$(s_o|z, \sigma^2, \phi) \sim LN(\mu_2, \sigma^2\rho_2) \tag{10}$$

where

$$\begin{aligned} \mu_2 &= (f'(t_0) - r'_\theta V_\phi^{-1}X)(V_0^{-1} + X'V_\phi^{-1}X)^{-1}V_0^{-1}\beta_0 \\ &+ (r'_\theta V_\phi^{-1} + (f'(t_0) - r'_\theta V_\phi^{-1}X)(V_0^{-1} + X'V_\phi^{-1}X)^{-1}X'V_\phi^{-1}) \ln z \\ \rho_2 &= (1 - r'_\theta V_\phi^{-1}r_\theta) \\ &+ (f'(t_0) - r'_\theta V_\phi^{-1}X)(V_0^{-1} + X'V_\phi^{-1}X)^{-1}(f'(t_0) - r'_\theta V_\phi^{-1}X)'. \end{aligned}$$

Also, from (5) and (7), we can show that $(s_o|z, \phi)$ has a log t distribution with $n + a$ degree of freedom denoted by $LT_{n+a}(\mu_2, S_1^2\rho_2)$ whose density function is given by

$$f(s_o|z, \phi) = \frac{1}{s_0}\frac{\Gamma(\frac{n+a+1}{2})}{\Gamma(\frac{n+a}{2})(\pi S_1^2\rho_2)^{1/2}}\left[1 + \frac{(\ln s_0 - \mu_2)^2}{S_1^2\rho_2}\right]^{-\frac{n+a+1}{2}}, \ s_0 > 0.$$

Next, based on the discretisation method, we obtain

$$\begin{aligned} f(s_o|z) &= \sum_{i=1}^{m} f(s_o|z,\phi_i)\pi(\phi_i|z) \\ &= \frac{\sum_{i=1}^{m} f(s_o|z,\phi_i)g(\phi_i;z)\pi(\theta_i)\pi(\alpha_i^2)}{\sum_{i=1}^{m} g(\phi_i;z)\pi(\theta_i)\pi(\alpha_i^2)}. \end{aligned} \tag{11}$$

If we choose a continuous distribution for ϕ, then analytical determination of the Bayesian predictive distribution becomes intractable, and importance sampling (Tanner, 1996) can be used to approximate it. For this instance, we generate m independent random values $\theta_1, \cdots, \theta_m$ from distribution $\pi(\theta)$ and $\alpha_1^2, \cdots, \alpha_m^2$ from $\pi(\alpha^2)$, then, for a sufficiently large m, an approximate Bayesian predictive distribution is given by

$$\begin{aligned} f(s_o|z) &= \int_{\Omega} f(s_o|z,\phi)\pi(\phi|z)d\phi \\ &\approx \frac{\sum_{i=1}^{m} f(s_o|z,\phi_i)g(\phi_i;z)}{\sum_{i=1}^{m} g(\phi_i;z)}. \end{aligned}$$

Since the exponential of a log t distribution does not have finite moment, the minimum square predictor does not exist. To circumvent this, we consider the absolute error loss function. Now, using (8) the Bayesian predictive distribution or its approximate (9), we can obtain the Bayesian spatial predictor or its approximate of $S(t_0)$ as

$$\hat{S}(t_0) = Median\ of\ (S(t_0)|z) = e^{\,\mu_1}.$$

4 Numerical example

To illustrate the proposed Bayesian spatial prediction method in section 3, let us apply it on a rainfall data set. Table 1 shows the rainfall at 20 sites in a region with 192×160 squared kilometers at north of Iran on the last month of winter 2002. For Bayesian spatial prediction of rainfall at any given site t_0, we assume that the data are realizations of a noisy Log-Gaussian model with fixed mean $E[Y_s(t)] = \beta^*$ and an exponential isotropic correlation function $\rho(u,t;\theta) = \theta^{||u-t||}$, where $\theta \in \Theta = (0,1)$ is the unknown parameter.

Since the hyper parameters β_0, V_0, a and b of the prior distributions (3) and (4) are unknown, we use the limit prior distributions (as V_0 and a tend to zero)

$$\begin{aligned} (\beta|z,\sigma^2,\phi) &\sim N_p(\beta_2, \sigma^2 V_2) \\ (\sigma^2|z,\phi) &\sim \chi^2_{Inv}(n-1, S_2^2) \end{aligned}$$

where $V_2 = \frac{1}{\mathbf{1}'V_\phi^{-1}\mathbf{1}}$ and $\beta_2 = V_2(\mathbf{1}'V_\phi^{-1}\ln z)$. For this case (8) becomes

$$g(\phi;z) = |V_2|^{\frac{1}{2}}|V_\phi|^{-\frac{1}{2}}(S_2^2)^{-\frac{n-1}{2}}$$

t_i	$z(t_i)$	t_i	$z(t_i)$	t_i	$z(t_i)$	t_i	$z(t_i)$
(1 , 8)	33.10	(5 , 8.5)	26.83	(7 , 9)	21.70	(9.5 , 9)	17.72
(2 , 5.5)	38.24	(5.5, 2.5)	26.76	(7 , 7.5)	18.24	(10 , 9)	31.14
(3 , 3)	24.35	(6 , 5)	29.55	(8 , 7)	33.98	(11 , 2)	23.69
(3.5 , 6)	46.06	(6 , 1)	24.47	(9 , 1.5)	41.85	(11 , 6)	22.82
(4 , 4)	23.60	(6.5 , 4)	30.57	(9 , 5)	19.31	(12 , 8)	31.27

Table 1: Rainfall at 20 stations in north of Iran.

and $f(s_o|z, \phi)$ is a log t distribution $LT_{n-1}(\mu_3, \frac{n}{n-1}S_2^2\rho_3)$, where

$$\mu_3 = \left(r'_\theta V_\phi^{-1} + \frac{(1 - r'_\theta V_\phi^{-1}\mathbf{1})\mathbf{1}' V_\phi^{-1}}{\mathbf{1}' V_\phi^{-1}\mathbf{1}} \right) \ln z$$

$$\rho_3 = (1 - r'_\theta V_\phi^{-1} r_\theta) + \frac{(1 - r'_\theta V_\phi^{-1}\mathbf{1})^2}{\mathbf{1}' V_\phi^{-1}\mathbf{1})}.$$

Now we generate $m = 30$ independent random values for $\phi = (\theta, \alpha^2)$ in $(0, 1) \times [0, \infty)$. For Uniform prior distributions $\pi(\theta_i) = \pi(\alpha_i^2) = \frac{1}{m}, \quad i = 1, \ldots, m$, the approximate Bayesian predictive distribution is simplified as

$$f(s_o|z) = \frac{\sum_{i=1}^m f(s_o|z, \phi_i) g(\phi_i; z)}{\sum_{i=1}^m g(\phi_i; z))}.$$

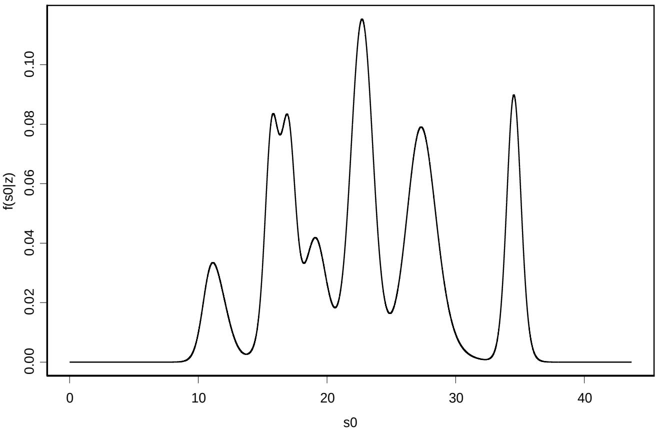


Figure 1: Bayesian Predictive Distribution

Figure 1 shows the curve of the predictive distribution of $S(t_0)$ at site $t_0 = (4, 8)$. The median of this distribution, $\hat{S}(t_0) = Median[S(t_0)|z]$ is computed using a function provided in S-Plus system. The Bayesian spatial prediction of rainfall at site $t_0 = (4, 8)$ is equal to 22.68.

5 Discussion

In some applications, the Gaussian assumption for spatial data is not appropriate. When a logarithm transformation is desirable in representing the positive spatial data, this work provides a Bayesian model for analyzing them. Actually, the Bayesian prediction of the noisy Gaussian model is extended to a noisy Log-Gaussian random field. Similarly, this model can be extended to a noisy Trans-Gaussian spatial model. Furthermore, to avoid the mentioned problem related to the BTG model, we would propose to use a transformation which is physically interpretable (e.g. a logarithm transformation) in a Bayesian frame work. In the numerical example our method is used to determine the Bayesian predictive distribution and deriving the Bayesian spatial prediction of rainfall at a given site based on the absolute error loss function.

References

[1] Berger J. O., De Oliveira V., Sanso B. (2001). *Objective Bayesian analysis of spatially correlated data.* Journal of the American Statistical Association **96**, 1361 – 1374.

[2] Chiles J.P., Delfiner P.(1999). Geostatistics; modelling spatial uncertainty. John Weily, New York.

[3] Christensen O.F., Diggle P.J., Ribeiro Jr. P.J. (2001). *Analysing positive-valued spatial data: the transformed Gaussian model.* Geostatistics for Environmental Applications. **11**, 287 – 298.

[4] Cressie N. (1993). *Statistics for spatial data.* John Weily, New York.

[5] De Oliveira V. Fokianos, K. Kedem B. (2003). *Bayesian transformed Gaussian random field: a review.* Japanese Journal of Applied Statistics. (to appear).

[6] De Oliveira V., Ecker M.D. (2002). *Bayesian hot spot detection in the presence of spatial trend: application to totorial nitrogen cocentration in the chesapeake bay.* Environmetrics. **13**, 85 – 101.

[7] Diggle P.J., Ribeiro Jr. P.J. (2003). *Bayesian inference in Gaussian model based geostatistics.* Geographical and Environmental Modelling. (to appear).

[8] Diggle P.J., Tawn J.A., Moyeed R.A. (1998). *Model based geostatistics.* Applied Statistics. **47**, 299 – 350.

[9] Handcock M.S., Stein M.L. (1993). *A Bayesian analysis of kriging.* Technometrics **35**, 403 – 410.

[10] Tanner M.A. (1996). *Tools for statistical inference.* Springer Verlag, Berlin: New York.

Acknowledgement: We are thankful to referees for their comments.

Address: M. Mohammadzadeh, K.M. Jafari, Department of Statistics, Tarbiat Modarres University, P.O.Box 14115-175, Tehran, Iran

E-mail: `mohsen_m@modares.ac.ir`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

FLEXIBLE DISCRETE EVENTS SIMULATION OF CLINICAL TRIALS USING LEANSIM(R)

T. Monleón, J. Ocaña, E. Vegas, P. Fonseca, A. Riera, J. Montero, I. Abbas, J. Casanovas, E. Cobo, J.A. Arnaiz, X. Carne and J.M. Gatell

Key words: Simulation, mixed models, discrete events, clinical trials, HIV, SAS.

COMPSTAT 2004 section: Simulations.

Abstract: We illustrate the process of use the realistic simulator LeanSim(r), for the concrete case of pilot clinical trial devoted to testing efficacy of the antiretroviral didanosine in a group of 50 patients with virologic failure presenting different mutations in the gen of HIV reverse transcriptase. It can origin resistance to transcriptase inhibitors and it can affect the efficacy of treatment designed to reduce the viral load in AIDS patients. LeamSim(r) is basically a discrete event simulation tool developed in C/C++ that can be applied in any ambit, but accepts personalised elements construction, in order to adapt the simulation model to the reality that want to be simulated (this is the lean simulation metaphor). Another important LeanSim aspect is the way in with the model is constructed, this is, via the process definition witch is closer to the clinical trials experiment definition. A simulation model based in Linear Mixed Models was estimated and during the simulation different scenarios were simulated to optimise sample size, effect of missing values, number of centres recruiting patients and the variability inter/intra patients. The results suggest that simulation may be a good exploratory tool in the design of future clinical trials and that LeanSim(r) is a promising simulation tool, flexible and powerful, thanks to the use of scripts and other techniques, to define models and simulation schema.

1 Introduction

Complex, long in time (usually more than 10 years) and large-scale medical clinical trials (MCT), with highly standardized procedures and phases (I,II,III, IV) are required to verify the efficacy and the safety on new drugs. These trials are based on the main statistical designs like parallel and crossover, see for example [11]. In the last twenty years, some researchers have suggested and tested the possible use of simulation as an exploratory tool to reduce the duration of MCT and to seek for possible problems during the true experiments with human patients. These experiences have propitiated the publication of normative guidelines, like [3], [4] in consonance with the tendency to international unification and regulation of methodologies in MCT, like ICH [5].

The work presented here has been performed under the Spanish research project (PTR1995-0518-OP-02-02) devoted to the simulation of clinical trials. The project main goal is to simulate a wide variety of true MCT and, at the same time, to develop a suitable tool to easily perform future simulations in this field. The long term objective is to asses the simulation adequacy in future clinical trials designs, as an exploratory tool to be employed before its effective realization. The study presented here, jointly with some similar ones [7], [8], [9], [10], will be used as a basis for the simulation of a pilot true MCT on AIDS, to be performed in the near future. It will be devoted to testing the efficacy of the antiretroviral didanosine in a group of patients with virologic failure, presenting different mutations in the HIV reverse transcriptase gene. These mutations of HIV may origin resistance to reverse transcriptase inhibitors, affecting the efficacy of treatments designed to reduce the viral load in AIDS patients. Here we present the design and the main results of the simulation of a previous MCT oriented to the evaluation of 2 antiretroviral pharmacological combinations, during 2 years [8]. The main objective is to illustrate the usage and the possible adequacy of the chosen simulator program, called "LeanSim(r)", which is based on a flexible concept of the discrete event simulation paradigm.

2 Methods

2.1 Simulation methodology

LeanSim(r) is basically a discrete simulation tool developed in C/C++ that can be applied in any ambit but that accepts personalised elements construction, in order to adapt the simulation model to the reality to be simulated -the "lean" simulation metaphor. The "world view" of LeanSim(r) is based on the process interaction approach [2], adequate to describe MCT experiments entities (patients) that are processed (visited, treated) by other entities like hospitals, etc. LeanSim(r) allows the simulation of population variability, both inter and intra patients, for example using specific functions to generate normal multivariate variability.

All the clinical trial simulations are based on a generic clinical trial template *(Figure 1)*. The simulated entities represent patients. These entities are managed by three main simulator objects: entity Generators (or Initialisators), entity Terminators (or Destructors) and a simulation Iterator which performs the entity processing main role. The concrete simulated MCT may be specified by mechanisms close to the object oriented paradigm, like specialization of pre-existing concepts or, alternatively, by calling scripts from the simulator objects. For the moment, only Visual Basic scripts are allowed. This last approach, less efficient but easier to implement, is illustrated here.

Six scripts parameterise the concrete simulation experiment and the clinical trial outcome (Viral load in this study) in form of an statistical model. These scripts may be automatically generated by two complementary tools, developed by the first author: “Hipocrates” [7] and a user interface to specify mixed models. Hipocrates is a program devoted to rationalise the specification of MCT design and to manage the data produced by them.

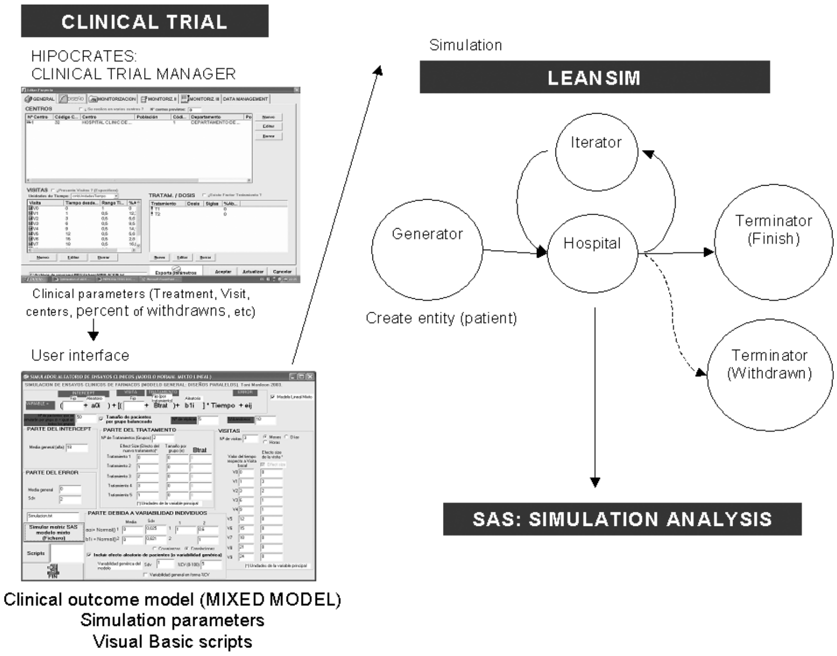


Figure 1: Event diagram of the clinical trials simulation model and interaction with the clinical trial specifications stored in Hipocrates. The conceptual model on LeanSim(r) simulator describes the life cycle of an entity (patient) during the clinical trial simulation.

Despite the apparent diagram simplicity, the main complexities are associated with the Iterator object. The model describes the life cycle of an entity (patient) involved in a clinical trial. In the next lines there is a brief description of this cycle.

- Entity generation (recruitment in the clinical trial) and assignation of its attributes (usually randomly generated using specified probability distributions).
- Transfer to the Iterator object that repeatedly computes the patient’s viral load in each visit. The visits are also scheduled by the Iterator. Also according to some probabilities there is the chance of entering a Terminator object representing withdraws.
- Withdrawn registration.
- Trial completion registration.

The total simulation time depends on the behaviour of an object of CleanIteratorMachine object and more specifically on the time to the last scheduled visit of the last patient to leave (withdrawn) the trial. In Multicentric trials (more than one hospital), each centre is represented by its own diagram and processing in all centres is performed (logically, not necessarily physically) in parallel. In the simulation model, the patient's recruitment time was simulated according to an exponential or a normal distribution. The time until the next scheduled visit was generated according also to an exponential distribution of mean 15 days. Other possible distributions were also evaluated.

The specification of the simulation parameters to the LeanSim(r) generator is based on the use of Visual Basic scripts, this novel concept provides flexibility and power to the simulation. It comes from the revision of a first version of the LeanSim(r) clinical trials conceptual model, described in [8], specially due to its inflexible initial nature, not adequate to specify the clinical outcome model (viral load). The communication between scripts and LeanSim(r) is performed by "call-backs" sent from the scripts. The simulator script engine receives the "call-backs", and allows the interaction with the simulation engine of LeanSim(r).

A compiled interface was used to help biomedical users to use mixed linear models. It simply describes a quite general form of mixed linear model for repeated measures [6], [12], [13] (adequate for most MCT) in a more friendly way for the clinical practitioner, and automatically creates a script describing it. It is intended as a tool for easy specification of model variants and parameter values, to be tested by simulation.

2.2 Clinical model estimation

A linear mixed model was fitted for the parallel antiretroviral AIDS clinical trial described above. The fitting process is described in more detail in [6], [7]. It is based on a linear mixed model. The parameters were estimated using the PROC MIXED procedure of SAS [6], [12], [13] . The finally resulting model was:

$$Y = (8.34 + b_{0i}) + (-4 + \alpha + b_{1i})Visit_{ij} + e_{ij}$$

where i stands for the i-th patient and j for the j-th visit. It is assumed that $b_{0i} \sim N(0, 0.618)$, $b_{1i} \sim N(0, 0.236)$, $e_{ij} \sim N(0, 2)$ and α is the slope of each genetic HIV group and has values 0, 1, 2, 3, 1 in the initial simulated scenario. According to the original protocol, and concerning the slope fixed parameter (that is the speed in the viral load decrease), there were significant differences between treatments, and we suppose no significant interactions between visit and treatment. The effect of the visit was clearly detectable. We assume small correlations between the random factors b_0 and b_1 not significantly different from 0. In the simulations, these random factors were assumed to be independent, and thus independently generated but, using LeanSim(r), it would be possible also to generate them as a multivariate normal vector [8].

2.3 Some simulation results

More than 200 simulations were performed in LeanSim(r) under diverse scenarios (*Figure 2* and *Table 1*), to study the effect of sample size, the effect of missing values, the number of hospitals recruiting patients, effect size (δ) and the inter/intra patients variability. The statistical analysis was performed in SAS 8.

	Percent withdrawn					
n per group	5%	0%	5%	10%	10%	0%
3	0.1558	-	-	-	-	-
5	0.9506	-	-	-	-	-
10	0.5742	0.1049	-	-	-	-
15	0.2512	-	-	-	-	-
20	0.9831	0.8193	-	-	-	-
25	0.9614	0.1064	-	-	-	-
30	-	0.0045	0.1326	-	-	-
35	-	-	0.1252	-	-	-
40	-	-	0.0191	0.2509	-	-
50	-	-	-	0.350	-	-
60	-	-	-	0.0342	0.4172	0.4388
100	0.3474	-	-	-	-	-
Genetic group maximum Effect size	d=0.5	d=1.7	d=1.7	d=1.7	d=1	d=1

Table 1: Statistical analysis of simulations of pilot AIDS clinical trial under diverse scenarios: percent of withdrawn, group effect size (δ), sample size (n). Treatment p-value is indicated, using a longitudinal model of 5 groups and 3 visits.

3 Discussion

The results suggest that simulation may be a good exploratory tool in the design of future clinical trials and that LeanSim(r) is a promising simulation tool, flexible and powerful, thanks to the use of scripts and other techniques, to define models and simulation schema. On the other hand, considerable work remains to make LeanSim(r) useful for routine work in MCT. Now we are trying to simulate some other, more complex, previously performed clinical trials, in order to acquire some additional experience on the adequacy and the limitations of LeanSim(r) and to consequently improve it. In any case, we believe that simulation will in the future be a valuable tool to design clinical trials. It would be helpful in the detection of possible future problems (for example, those associated to withdraws) and for optimisation purposes.

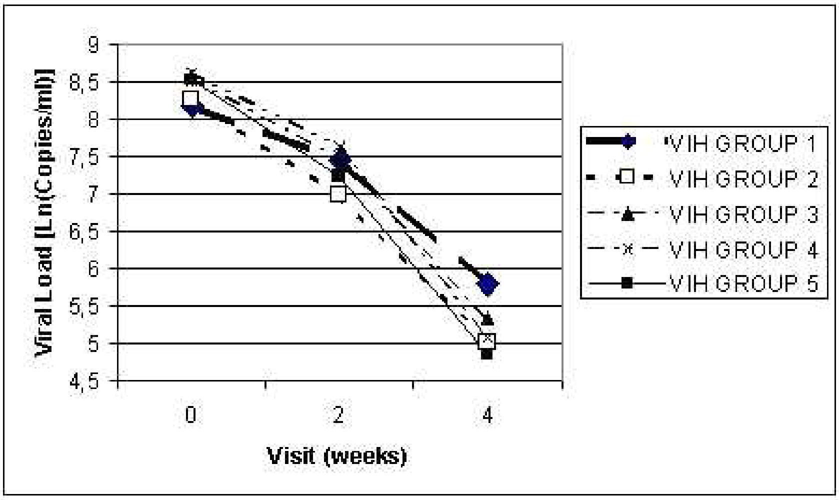


Figure 2: Simulation by LeanSim(r) of the pilot AIDS clinical trial during 4 weeks of treatment with didanosine and 5 HIV genetic groups which affects treatment efficacy.

References

[1] Abbas I., Cobo E., Casanovas J., Romeu J., Monleón T., Ocana J. (2003). *Optimising clinical trials design using simulation.* Proceedings of the Third Annual meeting of ENBIS and ISIS3. 21-23th August. Barcelona (Spain).

[2] Fonseca P., Casanovas J., Montero J. (2003). *LeanSim: Un sistema de simulación para el entrenamiento de personal especializado dentro de sistemas complejos.* Proceedings of the Second Iberoamerican Conference in Systems, Cybernetics and Computer Science CISCI. Vol I. Orlando FL (USA).

[3] Holford N.H.G., Hale M., Ko H.C., Steimer J.L., Sheiner L.B., Peck C.C. (1999). *Simulation in drug development: good practices.* Guidelines. `http://www.dml.georgetown.edu/cdds/SDDGP.html.`

[4] Holford N.H.G., Kimko H.C., Monteleone J.P.R., Peck C.C. (2000). *Simulation of clinical trials.* Annual review of pharmacology and toxicology **40**, 209−234.

[5] ICH (International Conference on Harmonization). (1997). *Guidance for Good Clinical Practice.* `http://www.ich.org.`

[6] Lindsey J.K. (1999). *Models for repeated measurements.* Oxford University press. USA: New York.

[7] Monleón T. (2003). *Hipócrates: software to perform management, monitoring and data management of clinical trials.* Users Manual.

[8] Monleón T., Ocaña J., Abbas I., Casanovas J., Cobo E., Arnaiz J.A., Carner X. (2003). *Simulación de ensayos clínicos de medicamentos. Ajuste del modelo.* Proceedings of the XXVII National Congress of Statistics. 8-11th April. Lleida (Spain).

[9] Monleón T., Ocaña J., Vegas E. (2003). *Realistic simulation about AIDS Clinical Trial.* Proceedings of the IX Spanish Conference of Biometry. Vol 1, 225 – 228. 28-30th May. La Coruna (Spain).

[10] Monleón T., Ocaña J., Vegas E., Fonseca P., Abbas I., Casanovas J., Cobo E., Arnaiz J.A., Carner X., Gatell. (2003). *Optimization of AIDS pilot clinical trial using LeanSim.* Value in Health Num 6, November/December, **6**, 794.

[11] Roset M., Bonfill X., Monleón T. (2003). *Curso de metodología de investigación y estadística en oncologíia y hematología I.* (Methodology of research and statistics in oncology and haematology I). Novartis Oncology. Spain: Barcelona.

[12] Verbeke G., Molenberhs G. (2001). *Mixed Models for longitudinal data with SAS.* SEA (Servei d'Estadistica Universitat Autónoma de Barcelona). Spain: Barcelona.

[13] Vonesh E.F., Chinchilli. V.M. (1997). *Linear and nonlinear models for the analysis of repeated measurements.* Marcel Dekker, Inc.

Address: T. Monleón, J. Ocaña, E. Vegas, Departament d'Estadística. Universitat de Barcelona. Avda Diagonal 645. 08028 Barcelona, Spain
P. Fonseca, A. Riera, J. Montero, I. Abbas, J. Casanovas, E. Cobo, Departament d'Estadística. Escola de Matemàtiques i Estadística. UPC. /Pau Gargallo s/n. 08028 Barcelona, Spain
J.A. Arnaiz, X. Carne and J.M. Gatell, Unitat d'Assaigs Clínics. Servei de Farmacologia Clínica. Hospital Clínic. C/Villarroel 170. 08036 Barcelona, Spain

E-mail: amonleong@wanadoo.es, jocana@ub.edu., pau@fib.upc.es, arriera@fib.upc.es, monty@fib.upc.es, esmail@fib.upc.es, josepk@fib.upc.es, erik.cobo@upc.es, paufonseca@msm.com, jaarnaiz@clinic.ub.es, xcarne@clinic.ub.es

COMPSTAT'2004 Symposium

ORTHOGONAL SCORE ESTIMATION WITH VARIABLE SELECTION IN MULTIVARIATE METHODS

Yuichi Mori, Kaoru Fueda and Masaya Iizuka

Key words: Variable selection, givens transformation, information criteria.

COMPSTAT 2004 section: Dimensional reduction.

Abstract: There are many criteria and procedures by which to select a reasonable subset of variables, most of which are based on the distance between the distributions of selected variables and original variables. On the other hand, when we have some authorized score, this score should be estimated by selected variables. In the present paper, we propose three procedures by which to select a subset of variables for the estimation of a specified score. In addition, we derive information criteria by which to compare these procedures.

1 Introduction

Consider a situation in which we wish to select items or variables so as to delete the redundant variables or to make a low-dimensional rating scale to measure latent traits. Validity requires all of the variables to be included. On the other hand, practical application requires that the number of variables be as small as possible.

These studies sought to obtain ordinary principal components(PCs) based on a subset of variables in such a way that these PCs retain as much information as possible compared to PCs based on all of the variables: Jolliffe's methods [5], [6] consider PC loadings, and the methods of McCabe [9] and Falguerolles and Jmel [2] use a partial covariance matrix to select a subset of variables, which maintains information on all variables to the greatest extent possible. Robert and Escoufier [12] and Bonifas et al. [1] used the RV-coefficient and Krzanowski [7], [8] used Procrustes analysis to evaluate the similarity between the configuration of PCs computed based on selected variables and that based on all variables. Tanaka and Mori [16] discuss a method called the "modified PCA" (M.PCA) to derive PCs that are computed using only a selected subset of variables but which represent all of the variables, including those not selected. Since M.PCA naturally includes variable selection procedures in the analysis, its criteria can be used directly to detect a reasonable subset of variables (e.g. [10]). Furthermore, other criteria can be considered, such as criteria based on influence analysis of variables using the concept reported in Tanaka and Mori [16] and criteria based on predictive residuals using the concept reported in Krzanowski [8]. Mori et al. [11] developed statistical software "VASPCA" to select a subset of variables, and Iizuka et al. [3] proposed computer intensive methods to determine the number of variables to observe.

Thus, the existence of several methods and criteria is one of the typical characteristics of variable selection in multivariate methods without external variables such as PCA, where the term "external variable" is used to indicate a variable to be predicted or to be explained using the information derived from other variables. Moreover, the existing methods and criteria often provide different results (selected subsets of variables), which is regarded as another typical characteristic. This occurs because each criterion or PC procedure has its own reasonable purpose for selecting variables. Therefore, we can not say that one procedure is better than another.

In practical applications, we have used some authorized scores, for example, principal components in principal component analysis and factors in factor analysis, and in many case, the scores are orthogonal (uncorrelated). Since the scores are obtained based on all variables, we would like to estimate them based on some subset of variables. In this paper, we propose a procedure by which to select a reasonable subset of variables we herein formulate methods for estimating the orthogonal scores. In addition, in Section 2, we propose information criteria by which to compare these methods and provide numerical examples in Section 3.

2 Formulation

2.1 Procedure 1

Suppose we have some p-variable data which is then reduced to r scores by a dimensional reduction method. Now, suppose that we would like to observe only q variables ($r \leq q \leq p$) in order to reduce the measurement cost. Let standardized observed p-value data be

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix},$$

i.e. $\sum_{i=1}^{n} y_{ij} = 0$ and $\sum_{i=1}^{n} y_{ij}^2 = 1$ for $j = 1, \ldots, p$. Suppose we have r scores such that

$$z_{ij} = a_{j1}y_{ij} + \cdots + a_{jp}y_{ij}, (i = 1, \ldots, n,\ j = 1, \ldots, r).$$

The coefficients $\{a_{ij}\}$ are derived from original data, for example PCA, or are given by prior research. In many cases, the scores $\{Z_1, \ldots, Z_p\}$ are uncorrelated.

Now, for selected variables $\{Y_{j_1}, \ldots, Y_{j_q}\}$, the least squares estimator of $\{Z_1, \ldots, Z_p\}$ is given by

$$\hat{Z} = Y_1 B,$$

where $B = (Y_1^T Y_1)^{-1} Y_1^T Z$ and

$$Y_1 = \begin{pmatrix} y_{1j_1} & \cdots & y_{1j_q} \\ \vdots & \ddots & \vdots \\ y_{nj_1} & \cdots & y_{nj_q} \end{pmatrix}.$$

This procedure is referred to herein as Procedure 1. However, the columns of $\hat{Z}$ are not correlated. We then consider the least squares problem subject to "$\hat{Z}^T\hat{Z} = B^TY_1^TY_1B$ as a diagonal matrix".

2.2 Sphering

Since $Y_1^TY_1$ is a non-negative definite symmetric matrix, there is an orthogonal matrix P such that

$$P^TY_1^TY_1P = diag\,[\lambda_1, \ldots, \lambda_q]$$

and $\lambda_i \geq 0$, for $i = 1, \ldots, q$. If some of the λs are zero, then at least one of the variables $\{Y_{j_1}, \ldots, Y_{j_q}\}$ should change the other variables. Then, we may assume that $\lambda_i > 0, i = 1, \ldots, q$. Let

$$S = diag\left[\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_q}\right]P^T,$$

then S is a nonsingular matrix and $S^TS = Y_1^TY_1$, and the condition "$B^TY_1^TY_1B = B^TS^TSB$ is a diagonal matrix". Letting $C = SB$, then the condition is "C^TC is a diagonal matrix" and the residual sum of squares is

$$tr((Z - Y_1B)^T(Z - Y_1B)) = tr((Z - Y_1S^{-1}C)^T(Z - Y_1S^{-1}C)).$$

Thus, setting $\tilde{Y_1} = Y_1S^{-1}$, we have the minimization problem

$$\begin{aligned}&\text{minimize } tr((Z - \tilde{Y_1}C)^T(Z - \tilde{Y_1}C))\\ &\text{subject to } C^TC \text{ is a diagonal matrix.}\end{aligned}$$

We then denote the solution of the minimization problem without an orthogonal condition as C_0, and we have

$$C_0 = (\tilde{Y_1}^T\tilde{Y_1})^{-1}\tilde{Y_1}^TZ = \tilde{Y_1}^TZ,$$

and the residual sum of squares is

$$\begin{aligned}&tr((Z - \tilde{Y_1}C)^T(Z - \tilde{Y_1}C))\\ &= tr\left((Z - \tilde{Y_1}C_0)^T(Z - \tilde{Y_1}C_0)\right) + 2tr\left((Z - \tilde{Y_1}C_0)^T(\tilde{Y_1}C_0 - \tilde{Y_1}C)\right)\\ &+tr\left((\tilde{Y_1}C_0 - \tilde{Y_1}C)^T(\tilde{Y_1}C_0 - \tilde{Y_1}C)\right).\end{aligned}$$

The first term does not depend on C, and the second term is zero. We should therefore minimize the third term

$$tr\left((\tilde{Y_1}C_0 - \tilde{Y_1}C)^T(\tilde{Y_1}C_0 - \tilde{Y_1}C)\right) = tr\left((C_0 - C)^T(C_0 - C)\right)$$

As such, we should minimize the sum of squares of the length of column vectors of $C_0 - C$ in q-dimensional space, rather than in the original p-dimensional space. Note that the solution of Procedure 1 is $C = C_0$.

2.3 Procedure 2

For the case in which the order of importance in r scores is $\{Z_1, \ldots, Z_p\}$, e.g. the first principal component and the second principal component, we propose the following procedure:

We first find the best estimator for the most important score, and then find the best estimator for the next important score subject to the second estimator being orthogonal to the first estimator, and so on. Setting $C_0 = \left(C_0^{(1)}, \ldots, C_0^{(r)}\right), C = \left(C^{(1)}, \ldots, C^{(r)}\right)$, we have C as

$$\begin{aligned}
C^{(1)} &= C_0^{(1)} \\
C^{(2)} &= C_0^{(2)} - \frac{\left\langle C_0^{(2)}, C^{(1)} \right\rangle}{\left|C^{(1)}\right|^2} C^{(1)} \\
\cdots & \quad \cdots\cdots\cdots \\
C^{(r)} &= C_0^{(r)} - \sum_{k=1}^{r-1} \frac{\left\langle C_0^{(3)}, C^{(k)} \right\rangle}{\left|C^{(k)}\right|^2} C^{(k)}
\end{aligned}$$

This procedure is referred to herein as Procedure 2.

2.4 Procedure 3

In many cases, there is no order of importance in r scores, and to obtain the solution for the minimization problem, we must perform the following procedure with iteration. This procedure is referred to herein as Procedure 3.

For a given $q \times r$ column-orthogonal matrix $P = (\mathbf{p}_1, \ldots, \mathbf{p}_r)$, i.e. $P^T P$ is an r-th identity matrix, we have the minimizer C of $tr\left((C_0 - C)^T (C_0 - C)\right)$ such that each column vector $C^{(i)}$ of the matrix C is parallel to the column vector $\mathbf{p}_i$ of the matrix P as

$$C = (\alpha_1 \mathbf{p}_1, \ldots, \alpha_r \mathbf{p}_r) = (\mathbf{p}_1, \ldots, \mathbf{p}_r) diag\,[\alpha_1, \ldots, \alpha_r]$$

where $\alpha_i = \left\langle C_0^{(i)}, \mathbf{p}_i \right\rangle$. Then we have

$$|C_0^{(i)} - C^{(i)}|^2 = |C_0^{(i)}|^2 - \left\langle C_0^{(i)}, \mathbf{p}_i \right\rangle^2 .$$

Since $\left|C_0^{(i)}\right|^2$ is constant, we should maximize $\sum_{i=1}^{r} \left\langle C_0^{(i)}, \mathbf{p}_i \right\rangle^2$.

2.5 Givens transformation

For

$$C = (\mathbf{p}_1, \dots, \mathbf{p}_r) diag\,[\alpha_1, \dots, \alpha_r]\,,$$

we consider rotations in each two-dimensional plane. Since we do not consider rotations in r-dimensional space spanned by the column vectors $(\mathbf{p}_1, \dots, \mathbf{p}_r)$ of P, but rather those in q-dimensional space, we add vectors $\mathbf{p}_{r+1}, \dots, \mathbf{p}_q$ such that $\{\mathbf{p}_1, \dots, \mathbf{p}_q\}$ is an ortho-normal base of R^q. Now we perform the transform

$$\begin{cases} \mathbf{p}'_i = \cos\theta \mathbf{p}_i + \sin\theta \mathbf{p}_j \\ \mathbf{p}'_j = -\sin\theta \mathbf{p}_i + \cos\theta \mathbf{p}_j. \end{cases}$$

For $1 \le i, j \le r$, we maximize

$$\begin{aligned} & \left\langle C_0^{(i)}, \mathbf{p}'_i \right\rangle^2 + \left\langle C_0^{(j)}, \mathbf{p}'_j \right\rangle^2 \\ & = \left(\left\langle C_0^{(i)}, \mathbf{p}_i \right\rangle^2 + \left\langle C_0^{(j)}, \mathbf{p}_j \right\rangle^2 \right) \frac{1 + \cos 2\theta}{2} \\ & + \left(\left\langle C_0^{(i)}, \mathbf{p}_j \right\rangle^2 + \left\langle C_0^{(j)}, \mathbf{p}_i \right\rangle^2 \right) \frac{1 - \cos 2\theta}{2} \\ & + \left(\left\langle C_0^{(i)}, \mathbf{p}_i \right\rangle \left\langle C_0^{(i)}, \mathbf{p}_j \right\rangle - \left\langle C_0^{(j)}, \mathbf{p}_i \right\rangle \left\langle C_0^{(j)}, \mathbf{p}_j \right\rangle \right) \sin 2\theta \\ & \text{for simplicity, we put} \\ & = A \sin 2\theta + B \cos 2\theta + C, \end{aligned}$$

7 and find θ which maximizes the transform. For $-\pi < 2\theta \le \pi$, we have $-\pi/2 < \theta \le \pi/2$. Thus, $\cos\theta \ge 0$ and the sign of $\sin\theta$ are identical to those one of $\sin 2\theta$ and A. Then, we have

$$\begin{aligned} \sin\theta &= \pm\sqrt{\frac{1 - \cos 2\theta}{2}} = \pm\sqrt{\frac{\sqrt{A^2 + B^2} - B}{2\sqrt{A^2 + B^2}}} \\ \cos\theta &= \sqrt{\frac{1 + \cos 2\theta}{2}} = \sqrt{\frac{\sqrt{A^2 + B^2} + B}{2\sqrt{A^2 + B^2}}}. \end{aligned}$$

For $1 \le i \le r < j \le q$, we maximize

$$\left\langle C_0^{(i)}, \mathbf{p}'_i \right\rangle^2 = \left\langle C_0^{(i)}, \cos\theta \mathbf{p}_i + \sin\theta \mathbf{p}_j \right\rangle^2 = \left(\cos\theta \left\langle C_0^{(i)}, \mathbf{p}_i \right\rangle + \sin\theta \left\langle C_0^{(i)}, \mathbf{p}_j \right\rangle \right)^2.$$

Based on the above, the maximum value is reached when $(\cos\theta, \sin\theta)$ is parallel to $\left(\left\langle C_0^{(i)}, \mathbf{p}_i \right\rangle, \left\langle C_0^{(i)}, \mathbf{p}_j \right\rangle \right)$, and for such θ we have

$$\left\langle C_0^{(i)}, \mathbf{p}'_j \right\rangle^2 = \left(-\sin\theta \left\langle C_0^{(i)}, \mathbf{p}_i \right\rangle + \cos\theta \left\langle C_0^{(i)}, \mathbf{p}_j \right\rangle \right)^2 = 0.$$

Thus, we have that $\mathbf{p}'_j$ is orthogonal to $C_0^{(i)}$ for all $j = r+1, \dots, q, i = 1, \dots, r$. We therefore take $\mathbf{p}_1, \dots, \mathbf{p}_r$ as an ortho-normal base of r-dimensional space spanned by $C_0^{(1)}, \dots, C_0^{(r)}$.

2.6 Information criteria

Since we consider the estimation problem of r uncorrelated scores, AIC is given by

$$AIC = n\sum_{k=1}^{r} \log \frac{|Z^{(k)} - Y_1 B^{(k)}|^2}{n} + 2 \times (\text{number of parameters}),$$

where superscript (k) represents the k-th column vector of the matrix. The number of parameters is qr for Procedure 1 and $r(q + (q - r + 1))/2$ for Procedures 2 and 3.

2.7 Algorithm

Based on the above considerations, we present the following algorithm by which to obtain the subset of variables. We first decide the number of selected variables q due to observation cost.

1. For each subset of q variables, using the sphering described in Section 2.2, we have the solution of Procedure 1.

2. For the solution of Procedure 1, we perform orthogonalization as described in Section 2.3 to obtain the solution of Procedure 2.

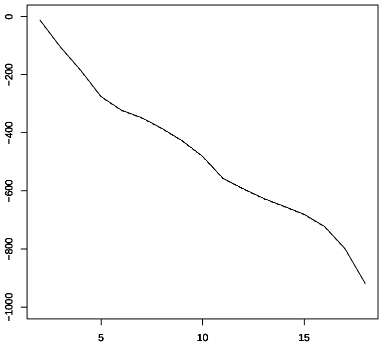
3. We iterate the Givens transform for $j = 2, \ldots, r, i = 1, \ldots, j - 1$ to obtain the solution of Procedure 3. We use the solution of Procedure 2 as an initial value.

4. Calculate AIC for each subset.

Finally, we select a subset of q variables which minimizes AIC.

3 Examples

Next, we show the result of applying the above procedure to “Alate” data and “MDOC” data. The Alate adelges (winged aphids) data set was analyzed originally by Jeffers [4] using ordinary PCA, and later by various researchers, including Jolliffe [6] and Krzanowski [7], [8], using PCA with variable selection functions. We applied our variable selection method to the data set given in Krzanowski [7]. The data set consists of 40 individuals and 19 variables. Eigenvalues and their cumulative proportions of the data are 13.8379 (72.83%), 2.3635 (85.27%), 0.7480 (89.21%), ... Therefore, as in previous studies we used two PCs. In addition, a mild disturbance of consciousness (MDOC) data set was originally analyzed by Sano et al. [13] using FA, and later by Tanaka and Kodake [15] and Tanaka [14] using principal factor analysis with variable selection functions. This data set consists of 25 variables and 87 individuals. For both data sets, we estimated the first and second principal components of full variables using q selected variables.



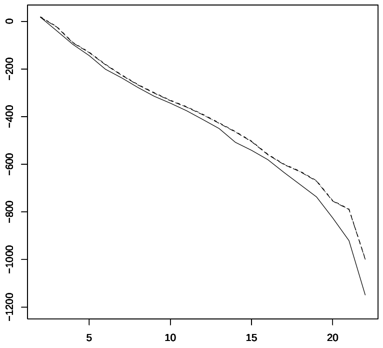


Figure 1: (left) AIC (vertical) for number (horizontal) of variable subsets. (Alate data, r=2)

Figure 2:(right) AIC (vertical) for number (horizontal) of variable subsets. solid line: present procedure, dotted line: Modified PCA. (MDOC data, r=2)

Figure 1 shows values of AIC for numbers of selected variables for alate data, and Figure 2 shows those for MDOC data. Although there are three lines corresponding to Procedures 1, 2 and 3 in the figures, the AICs of these procedures are so similar that the three lines appear as one.

Figure 2 also shows the values of the AIC of subset selected by our procedure and by Modified PCA. The values for these procedures are similar, especially when only a few variables are selected.

References

[1] Bonifas I., Escoufier Y., Gonzalez P.L., Sabatier R. (1984). *Choix de variables en analyse en composantes principales.* Rev. Statist. Appl. **23**, 5 − 15.

[2] Falguerolles A., De et Jmel S. (1993). *Un critere de choix de variables en analyse en composantes principales fonde sur des modeles graphiques gaussiens particuliers.* Rev. Canadienne Statist. **21** (3), 239 − 256.

[3] Iizuka M., Mori Y., Tarumi T., Tanaka Y. (2003). *Computer intensive trials to determine the number of variables in PCA.* Journal of the Japanese Society of Computational Statistics **15**, 337 − 345.

[4] Jeffers J.N.R. (1967). *Two case studies in the application of principal component analysis.* Appl. Statist. **16**, 225 − 236.

[5] Jolliffe I.T. (1972). *Discarding variables in a principal component analysis I - Artificial data -.* Appl. Statist. **21**, 160 − 173.

[6] Jolliffe I.T. (1973). *Discarding variables in a principal component analysis II - Real data -.* Appl. Statist. **22**, 21 − 31.

[7] Krzanowski W.J. (1987a). *Selection of variables to preserve multivariate data structure, using principal components.* Appl. Statist. **36**, 22 – 33.

[8] Krzanowski W.J. (1987b). *Cross-validation in principal component analysis.* Biometrics **43**, 575 – 584.

[9] McCabe G.P. (1984). *Principal variables.* Technometrics **26**, 137 – 44.

[10] Mori Y. (1997). *Statistical software VASPCA - Variable selection in PCA* -. Bulletin of Okayama University of Science **33**(A), 329 – 340.

[11] Mori Y., Iizuka M., Tarumi T., Tanaka Y. (2000). *Statistical software "VASPCA" for variable selection in principal component analysis.* Compstat2000 Proceedings in Computational Statistics (Short Communications), W. Jansen and J.G. Bethlehem (eds.), 73 – 74.

[12] Robert P., Escoufier Y. (1976). *A unifying tool for linear multivariate statistical methods: the RV-coefficient.* Appl. Statist. **25**, 257 – 265.

[13] Sano K., Manaka S., Kitamura K., Kagawa M., Takeuchi K., Ogashiwa M., Kameyama M., Tohgi H., Yamada H. (1977). *Statistical studies on evaluation of mind disturbance of consciousness – Abstraction of characteristic clinical pictures by cross-sectional investigation.* Sinkei Kenkyu no Shinpo. **21**, 1052 – 1065. (in Japanese)

[14] Tanaka Y. (1983). *Some criteria for variable selection in factor analysis.* Behaviormetrika **13**, 31 – 45.

[15] Tanaka Y., Kodake K. (1981). *A method of variable selection in factor analysis and its numerical investigation.* Behaviormetrika **10**, 49 – 61.

[16] Tanaka Y., Mori Y. (1997). *Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis.* American Journal of Mathematics and Management Science **17**, 61 – 89.

Acknowledgement: The authors would like to thank Professor Yutaka Tanaka and Professor Tomoyuki Tarumi for providing advice and numerous suggestions.

Address: Y. Mori, Department of Socio-Information, Okayama University of Science, Ridai-cho 1-1, Okayama, 700-0005, Japan
K. Fueda, Faculty of Environmental Science and Technology, Okayama University, Naka 3-1-1, Tsushima, Okayama, 700-8530, Japan
M. Iizuka, Faculty of law, Okayama University, Naka 3-1-1, Tsushima, Okayama, 700-8530, Japan

E-mail: `mori@soci.ous.ac.jp`, `fueda@ems.okayama-u.ac.jp`, `masa@law.okayama-u.ac.jp`

AUTOMATIC VALIDATION OF HIERARCHICAL CLUSTERING

Hans-Joachim Mucha

Key words: Hierarchical cluster analysis, validation, statistical computing.

COMPSTAT 2004 section: Clustering.

Abstract: Cluster analysis methods can be generalized in order to taking into account weights of observations. Using special weights leads to well-known resampling techniques. The prototype software ClusCorr98 offers some automatic validation techniques that can be considered as a so-called in-built validation of the number of clusters and of each cluster itself, respectively. Moreover this in-built validation can be used for finding the appropriate cluster analysis model. In that way clustering becomes a little bit intelligent. As an illustration of an application, the validation of results of hierarchical clustering based on the adjusted *Rand*'s measure is presented.

1 Introduction

Cluster analysis aims at finding interesting partitions or hierarchies without taking into account any background knowledge [5], [6]. Hierarchical clustering is in some sense more general than partitional clustering because a hierarchy (this is usually the result of a hierarchical cluster analysis) is a sequence of nested partitions. Here a partition is treated as an elementary component of a hierarchy. In the following partitions $\mathrm{P}(I,K)$ of the set of I objects (observations) into K non-empty clusters (subsets, groups) C_k are considered. The clusters are assumed pair-wise disjoint and the partition is an exhaustive subdivision. In this case a general way of validation of hierarchies will be recommended.

Simple model-based Gaussian clustering techniques can be expressed in terms of pairwise data clustering [2], [7]. Starting from pair-wise distances one can carry out both hierarchical and partitional clustering [9]. Moreover, weighted observations can be used in order to generalize the models. Otherwise it is well-known that the principle of weighting of observations is a key idea in data mining for handling cores (representatives of dense regions) and outliers. In the case of outliers one has to downweight them in order to reduce their influence [7]. Special weights are used in the proposed automatic validation technique that is applied to demographical data.

2 Simple model-based Gaussian clustering

Let $\mathbf{X}$ be the $(I \times J)$-data matrix under investigation consisting of I observations (objects) and J variables. Further, consider the well-known sum-of-

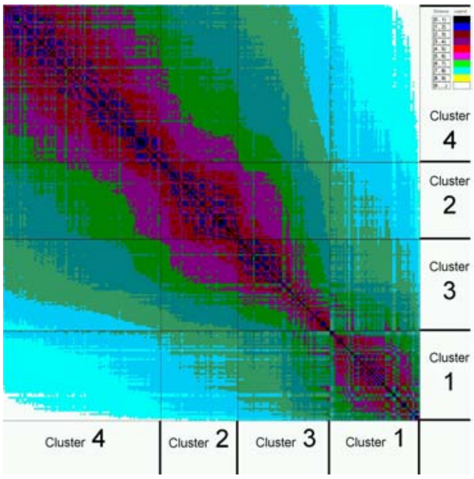


Figure 1: Fingerprint of a distance matrix with four clusters.

squares criterion

$$V_K = \sum_{k=1}^{K} tr(\mathbf{W}_k), \tag{1}$$

that has to be minimized concerning the partition $\mathrm{P}(I,K)$. Herein $\mathbf{W}_k = \sum_{i \in C_k} (\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T$ is the sample cross-product matrix for the k-th cluster C_k, and $\overline{\mathbf{x}}_k$ is the usual maximum likelihood estimate of expectation values in cluster C_k.

Criterion (1) can be written in the following equivalent form without explicit specification of cluster centres $\overline{\mathbf{x}}_k$

$$V_K = \sum_{k=1}^{K} 1/n_k \sum_{i \in C_k} \sum_{l \in C_k, l>i} d_{il}, \tag{2}$$

where n_k is the cardinality of cluster C_k, and

$$d_{il} = d(\mathbf{x}_i, \mathbf{x}_l) = (\mathbf{x}_i - \mathbf{x}_l)^T (\mathbf{x}_i - \mathbf{x}_l)$$

is the pair-wise squared Euclidean distance between two observations i and l. This criterion can be minimized for a single partition $\mathrm{P}(I,K)$ by exchanging observations between clusters [9]. This is equivalent to *k-means* clustering. Otherwise the hierarchical *Ward* method [10] minimizes (2) in a stepwise manner by agglomerative hierarchical clustering. Mucha et al. [7] presented other model-based cluster analysis in the pair-wise distances fashion. Figure 1 shows a so-called fingerprint of a distance matrix. It expresses one benefit of clustering based on pairwise distances, namely the visualization of arbitrary high dimensional data in only two dimensions. One can get such a presentation also for more complex cluster analysis models. Here the grey scale expresses the level of distance between a pair of observations: The higher the distance the brighter the grey becomes. The data under investigation accrues from population statistics of 227 observations (see Section 5). The four clusters are found by the *Ward* method.

Another benefit of clustering based on pairwise distances over clustering based directly on the $(I \times J)$-data matrix $\mathbf{X}$ is the more general meaning of distances. For instance, applying distances allows cluster analysis of mixed data (quantitative and qualitative data, see, for example, Gower [3]). Usually this procedure yields at least practical useful exploratory results.

The expression in (2) can be generalized to

$$V_K = \sum_{k=1}^{K} \frac{1}{M_k} \sum_{i \in C_k} m_i \sum_{l \in C_k, l > i} m_l d_{il} \tag{3}$$

by using positive weights of observations, where $M_k = \sum_{i \in C_k} m_i$ and m_i denote the weight of cluster C_k and the weight of observation i, respectively.

3 Resampling by special weights of observations

The weights m_i can be used for resampling purposes. Considering, for instance, the right hand side of equation (3) one recognizes the independence of weights m_i and M_k, respectively, from the pair-wise distances d_{il}. The latter ones exist in any case and are independent from the weighting the observations. That means once a distance matrix is figured out it will be fixed in the simulations. One has to change the weights only for simulation purposes. For example, the well-known bootstrap resampling technique can be formulated by choosing the following weights of observations:

$$m_i = \begin{cases} n & \text{if observation } i \text{ is drawn } n \text{ times} \\ 0 & \text{otherwise} \end{cases}$$

Here $I = \sum_i m_i$ holds in the bootstrap-resampling with replacement. Other resampling techniques can be described in a similar fashion by introducing weights. In view of the validation measures recommended in the next section, effective simulations for almost all hierarchical cluster analysis techniques can be performed by assigning the special weights of observations:

$$m_i = \begin{cases} c & \text{if observation } i \text{ is drawn randomly } (c > 0) \\ 0 & \text{otherwise} \end{cases}$$

This resampling technique is without replication. Usually it is $c = 1$. The observations with $m_i > 0$ are called active objects whereas the ones with $m_i = 0$ are called supplementary objects. The latter ones do not affect the cluster analysis in any way. However they can be allocated after clustering

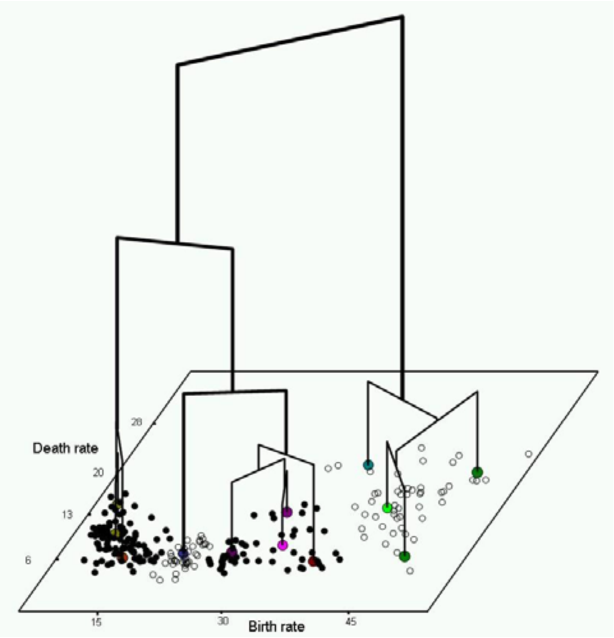


Figure 2: Graphical presentation of the result of hierarchical clustering.

into the partitions and hierarchies according to their distance values. This is an option of the software ClusCorr98 and can be done, for instance, by applying k nearest neighbour classification.

4 The adjusted *Rand*'s measure for comparing partitions

Partitions are basic results of cluster analysis that cover also hierarchies. Therefore comparing partitions becomes a basic and general tool for validation of cluster analysis results. The key approach for comparing partitions is based on the comparison of pairs of objects concerning their class membership [8]. For instance, to compare two partitions P(I,K) and Q(I,L) the *Rand* index $R^* = (a+d)/\binom{I}{2}$ (similarity index) can be applied. Here a and d count the pair-wise matches which are good in the sense of similarity (correspondence), see Table 1. Equivalently, the *Rand*'s index R^* can be expressed by using a contingency table obtained by crossing directly the two partitions P and Q:

$$R^* = [\binom{I}{2} + 2\sum_{k=1}^{K}\sum_{l=1}^{L}\binom{n_{kl}}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{l=1}^{L}\binom{n_{+l}}{2}]/\binom{I}{2}.$$

Partition Q

Partition P	Same cluster	Different clusters
Same cluster	a	b
Different clusters	c	d

Table 1: Contingency tables of pairs of observations concerning their cluster membership.

Partition Q

		1	*2*	...	l	...	L	Sum
	1	n_{11}	n_{12}	$\cdots$	n_{1l}	$\cdots$	n_{1L}	n_{1+}
	2	n_{21}	n_{22}	$\cdots$	n_{2l}	$\cdots$	n_{2L}	n_{2+}
	...	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
P	k	n_{k1}	n_{k2}	$\cdots$	n_{kl}	$\cdots$	n_{kL}	n_{k+}
	...							
	K	n_{K1}	n_{K2}	$\cdots$	n_{Kl}	$\cdots$	n_{KL}	n_{K+}
	Sum	n_{+1}	n_{+2}	$\cdots$	n_{+l}	$\cdots$	n_{+L}	$I = n_{++}$

Table 2: Contingency table by crossing two partitions P and Q.

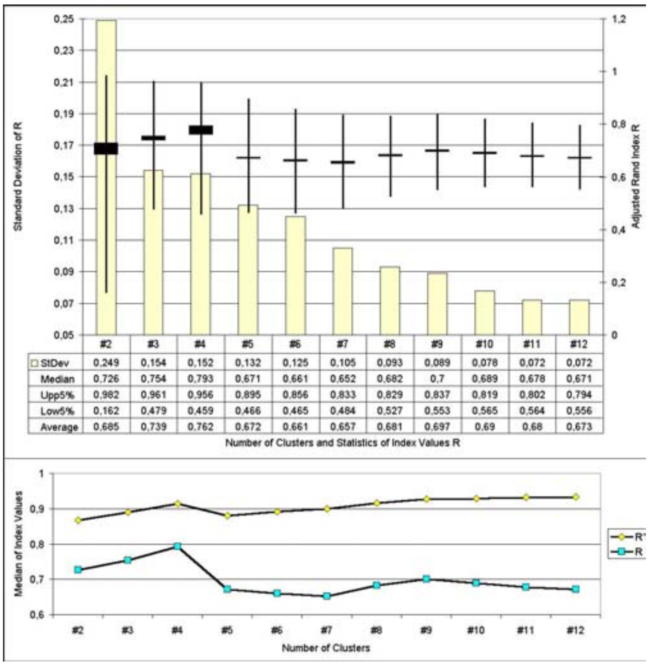
Table 2 shows such a contingency table with the elements n_{kl}. At the right hand side and at bottom there are the marginal sums n_{k+} and n_{+l}, respectively. The contingency table has the important advantage over Table 1 that the stability of every single cluster can be investigated additionally. But this is not the main topic of this paper. However this opportunity should be mentioned here. The measure R^* is dependent on the number of clusters K. The higher K the higher R^* becomes in average [6]. In order to avoid this disadvantage Hubert and Arabie [4] recommended the adjusted *Rand* index R based under the assumption of the generalized hypergeometric model:

$$R = \frac{\sum_{k=1}^{K} \sum_{l=1}^{L} \binom{n_{kl}}{2} - [\sum_{k=1}^{K} \binom{n_{k+}}{2} \sum_{l=1}^{L} \binom{n_{+l}}{2}]/\binom{I}{2}}{1/2[\sum_{k=1}^{K} \binom{n_{k+}}{2} + \sum_{l=1}^{L} \binom{n_{+l}}{2}] - [\sum_{k=1}^{K} \binom{n_{k+}}{2} \sum_{l=1}^{L} \binom{n_{+l}}{2}]/\binom{I}{2}}. \qquad (4)$$

This measure suits better for the decision about the number of clusters K because it takes the value 0 when the index R^* equals its expected value for each $k, k = 2, 3, \ldots, K$.

5 Simulation Studies

Hierarchical clustering gives a single unique solution (hierarchy). This is in opposition to some iterative method like *k-means* clustering that lead to locally optimal solutions depending on initial partitions. Figure 2 shows such



	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
□ StDev	0,249	0,154	0,152	0,132	0,125	0,105	0,093	0,089	0,078	0,072	0,072
Median	0,726	0,754	0,793	0,671	0,661	0,652	0,682	0,7	0,689	0,678	0,671
Upp5%	0,982	0,961	0,956	0,895	0,856	0,833	0,829	0,837	0,819	0,802	0,794
Low5%	0,162	0,479	0,459	0,466	0,465	0,484	0,527	0,553	0,565	0,564	0,556
Average	0,685	0,739	0,762	0,672	0,661	0,657	0,681	0,697	0,69	0,68	0,673

Figure 3: Simulation results of clustering the *Country* - data set.

a unique result of hierarchical clustering by the *Ward* method. The data matrix consists of 227 observations (countries) with the two variables birth rate and death rate in 1999. These variables are part of the population statistics published by CIA World Factbook [1]. Figure 3 consists of two parts. The upper part shows both the most important numerical results of simulations concerning the adjusted *Rand* index and a corresponding graphical representation of these univariate statistics.

The reading of this part of Figure 3 is as follows. The axis at the left hand side and the bars in the graphic are assigned to the standard deviation of R, whereas the axis at the right hand side and the box-plots are assigned to other statistics of R (Median, Average, upper and lower 5 percent quantile). The median of R for $K = 4$ takes the maximum value. That means, the four cluster solution is the most stable one. It can be confirmed in a high degree for almost all samples. For more than four clusters the median (or alternatively the average) of the adjusted *Rand* values becomes much lower. Therefore the number of cluster $K = 4$ is most likely. The axis at the left hand side and the bars in the graphic are assigned to the standard deviation of the adjusted *Rand*'s measure.

In the lower part of Figure 3 the *Rand* index R^* is compared with the adjusted *Rand* index R in the case of hierarchical cluster analysis by the

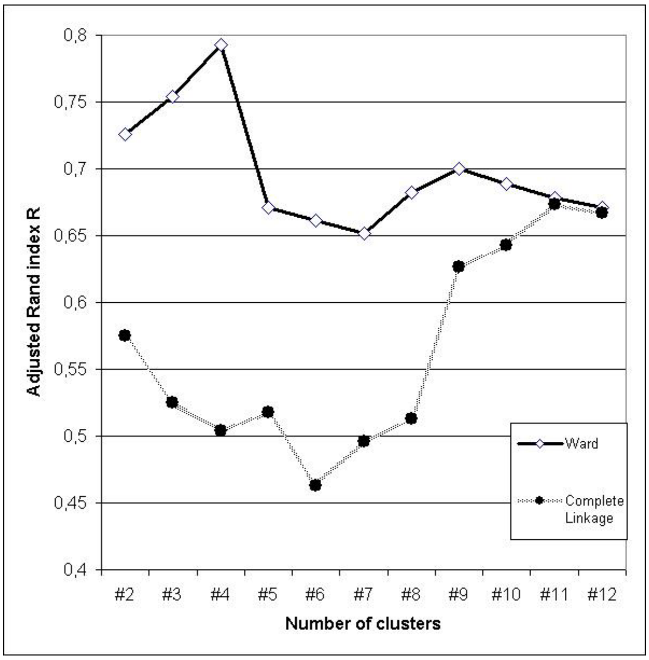


Figure 4: Comparison of the stability of two cluster analysis methods.

Ward method. Here the median out of 200 values represents an index value. One observes a rather continuous increase of the index R^* for increasing number of clusters, in contrast to the adjusted index R. The latter one is more appropriate to detect the right number of clusters.

Let us describe a little bit the favourite solution into 4 clusters. Cluster 1 contains 49 countries and is located at the upper right hand corner of Figure 2. Cluster 4 (85 countries) is located at the left hand side. Cluster 2 (41) is nearby cluster 4 whereas the cluster 3 (50) is located between cluster 1 and 2. In order to assess their stability the corresponding 200 (4×4)-contingency tables can be summarized, for example, by maximum correspondence to this unique partition into 4 clusters. As a result cluster 3 seems to be the most instable cluster because it counts a sum of 3443 discrepancies in comparison to to the other clusters where 798 (cluster 1), 647 (cluster 2) and 927 (cluster 4) discrepancies are counted. In relation to their cardinalities cluster 4 is the most stable one (10.91 relative discrepancies). Then cluster 2 follows with 15.78, cluster 1 (16.29), and at least cluster 3 (68.86). See also Figure 1 for a graphical impression about the clusters.

In Figure 4 two hierarchical clustering techniques are compared based on the adjusted *Rand* index R. The *Ward* method outperforms the Complete Linkage method for every number of clusters 2,3,... concerning to this measure of stability.

6 Conclusions

The principle of weighting of observations is a key idea for the in-built validation technique for hierarchical clustering. Using special weights leads to well-known resampling techniques. The proposed automatic validation techniques based on comparison of partitions is especially recommended for investigating the results of hierarchical clustering. Moreover this in-built validation is very easy to apply.

References

[1] CIA World Factbook (1999). *Population by country.* `http://www.geographic.org`

[2] Fraley C. (1996). *Algorithms for model-based Gaussian hierarchical clustering.* Technical Report, 311. Department of Statistics, University of Washington, Seattle.

[3] Gower J.C. (1971). *A general coefficient of similarity and some of its properties.* Biometrics **27**, 857 − 874.

[4] Hubert L.J., Arabie P. (1985). *Comparing partitions.* Journal of Classification **2**, 193 − 218.

[5] Kaufman L., Rousseeuw P.J. (1990). *Finding groups in data.* Wiley, New York.

[6] Mucha H.J. (1992). *Clusteranalyse mit Mikrocomputern.* Akademie Verlag, Berlin.

[7] Mucha H.J., Simon U., Brüggemann R. (2002). *Model-based cluster analysis applied to flow cytometry data of phytoplankton.* Weierstraß-Institute for Applied Analysis and Stochastic, Technical Report No. 5. `http://www.wias-berlin.de/.`

[8] Rand W.M. (1971). *Objective criteria for the evaluation of clustering methods.* Journal of the American Statistical Association **66**, 846 − 850.

[9] Späth H. (1985). *Cluster dissection and analysis.* Ellis Horwood, Chichester.

[10] Ward J.H. (1963). *Hierarchical grouping methods to optimise an objective function.* JASA **58**, 235 − 244.

Address: H.-J. Mucha, Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D-10117 Berlin, Germany

E-mail: `mucha@wias-berlin.de`

COMPSTAT'2004 Symposium

AN EXAMPLE OF D-OPTIMAL DESIGNS IN THE CASE OF CORRELATED ERRORS

Werner G. Müller and Milan Stehlík

Key words: Exact design, random field, numerical algorithms.

COMPSTAT 2004 section: Design of experiments.

Abstract: In this paper we consider an extension of a classic example in the design of experiments under the presence of correlated errors. This extension will allow to study the effect of the amount (range) of correlation on the design. Eventually we will gain valuable insights for the computational generation of optimal designs in more complex settings.

1 Introduction

Optimal design theory for regression experiments in the presence of correlated errors is a rarely visited topic, despite its apparent relevance in many application fields such as environmental science, geology, computer simulation, etc.

Generally one is confronted with selecting regressor values in a way such that the information about the parameters governing a posited model is maximized. In principle one can identify two sets of parameters of interests: one describing the trend (long term/distance fluctuations), the second describing the covariance function (short term/distance fluctuations).

And for the second characteristic there are again two ways on looking at it. On the one hand, covariance parameters can be considered as a nuisance quantity, which does not need to be estimated. On the other, which might be more realistic for most applications, covariance parameters need to be estimated as well and thus play a role in the objective function. In this paper we intend to demonstrate the difference between these two views on an illuminative simple case. Furthermore, the other main features that one can encounter with optimal designs in the correlated error setup (without replications) will be illustrated.

For this purpose we will consider a modification of the Example 6.4 studied in [5]. There we have the real random field

$$X(x) = \vartheta_1 + \vartheta_2 x + e(x) \tag{1}$$

(simple linear regression), with design space $V = [-1, 1]$ and a covariance function

$$\mathrm{cov}(x, z) = \begin{cases} 1 - \frac{d_{xz}}{\delta}, & \text{for } d_{xz} < \delta, \\ 0, & \text{for } d_{xz} \geq \delta. \end{cases} \tag{2}$$

with $\delta = 1$, where $d_{xz} = |z - x|$ denotes the distance between the design points. In this paper we will let the parameter δ vary, thereby allowing correlation structures with differing concentration, i.e. over which range in the design space observations will still be correlated.

Note that [5] has shown in Theorem 6.6 (a simplified proof of it is given in [4]) that since the covariance function can be expressed as a linear function of the responses, a uniformly optimal design is available in this case, which concentrates on the points $\{-1, 0, 1\}$. It will therefore be natural to eventually study 3-point D-optimal designs for our extended setting.

We further assume normal errors throughout the paper and maximize the log-likelihood function $L = \ln f(\vartheta, u)$, where $f(\vartheta, .)$ is the normal density of the vector $u = (X - \vartheta_1 - \vartheta_2 x, Y - \vartheta_1 - \vartheta_2 y, Z - \vartheta_1 - \vartheta_2 z)^T$ with the covariances given by (2); an explicite form is given in the Appendix. For brevity of the paper only the main ideas will be given in the proofs, more detailed derivations on particular aspects can be found in [6].

2 Unknown but fixed covariance parameters

In this section we suppose that δ is fixed. For $\delta \geq 0$ we introduce a function

$$J_\delta(x) = \begin{cases} 1, & \text{for } x \in [0, \delta), \\ 0, & \text{otherwise.} \end{cases}$$

We will in the following use $\delta = 0$ to represent the uncorrelated case.

2.1 Two-point designs

For simplicity let us first consider exact designs having only two design points $\{x, z\}$, $-1 \leq x < z \leq 1$. Here and further we use notation d for the distance when it is clear, which points are assumed. Then the elements $M_{i,j}$ of the Fisher information matrix M according to models (1) and (2) given above have the form:

$$M_{1,1} = E\left(-\frac{\partial^2 L}{\partial \vartheta_1^2}\right) = 2\frac{1 + (-1 + \frac{d}{\delta})J_\delta(d)}{1 - (1 - \frac{d}{\delta})^2 J_\delta(d)},$$

$$M_{1,2} = E\left(-\frac{\partial^2 L}{\partial \vartheta_1 \partial \vartheta_2}\right) = \frac{(x+z)(1 + (-1 + \frac{d}{\delta})J_\delta(d))}{1 - (1 - \frac{d}{\delta})^2 J_\delta(d)},$$

$$M_{2,2} = E\left(-\frac{\partial^2 L}{\partial \vartheta_2^2}\right) = \frac{x^2 + z^2 + 2xz(-1 + \frac{d}{\delta})J_\delta(d)}{1 - (1 - \frac{d}{\delta})^2 J_\delta(d)}.$$

The classic D-optimality criterion function has thus the form $\Phi(M) = -\ln \det M = -\ln(M_{1,1}M_{2,2} - M_{1,2}^2)$. In this setting we yield, not surprisingly:

Proposition 2.1. *The design $\{-1, 1\}$ is D-optimal for all $\delta \geq 0$.*

Proof *The criterion function can now after some algebra be written in the form*

$$\Phi(d) = \begin{cases} -\ln d^2, & \text{for } \delta \le d \le 2, \\ -\ln \frac{\delta^2 d}{2\delta - d}, & \text{for } 0 < d < \min\{\delta, 2\}. \end{cases}$$

Simple analysis shows that the minimum of $\Phi(d)$ *is attained for* $d = 2$*, which corresponds to the design* $\{-1, 1\}$. □

We should point out that this is one of the rare settings, in which the optimal design under correlation coincides with the uncorrelated case and does not depend upon the amount of

correlation reflected by the parameter δ. General conditions upon when this may happen are given in [1].

2.2 Three-point designs

Let us now consider the more interesting case of three design points, i.e. let $\{x, y, z\},\ -1 \le x < y < z \le 1$ be a proposed design, with fixed $\delta \ge 0$.

Theorem 2.1. *The following designs are D-optimal:*

- *For the uncorrelated case* ($\delta = 0$) *the D-optimal design exhibits repetitions, the criterion function attains its global minimum* $-\ln 8$ *for the designs* $\{-1, -1, 1\}$ *and* $\{-1, 1, 1\}$.
- *When both correlated and uncorrelated observations are possible* ($0 < \delta \le 2$)*, we obtain two D-optimal designs* $\{-1, -1 + \delta, 1\}$ *and* $\{-1, 1 - \delta, 1\}$ *with a common minimum value of the criterion function.*
- *When only correlated observations are possible* ($\delta > 2$)*, the minimum of the criterion function is attained for* $\{-1, y, 1\}$ *where* y *is arbitrary.*

Proof *The determinant of the information matrix can be written in the form*

$$\det M = \frac{2ad_{yz}d_{xy} - 2bd_{xz}d_{xy} - 2cd_{xz}d_{yz} + 3(x^2 + y^2 + z^2) - (x + y + z)^2}{1 - a^2 - b^2 - c^2 + 2abc},$$

where $a = (1 - \frac{d_{xz}}{\delta})J_\delta(d_{xz})$, $b = (1 - \frac{d_{yz}}{\delta})J_\delta(d_{yz})$ *and* $c = (1 - \frac{d_{xy}}{\delta})J_\delta(d_{xy})$. *When* $\delta = 0$*, the maximum of the* $\det M(x, y, z) = 3(x^2+y^2+z^2)-(x+y+z)^2$ *is attained for designs* $\{-1, -1, 1\}$ *and* $\{-1, 1, 1\}$. *When* $\delta > 2$*, the maximum of the* $\det M(d_{xz}) = \frac{\delta^2 d_{xz}}{2\delta - d_{xz}}$ *is attained for* $\{-1, y, 1\}$ *where* y *is arbitrary.*

Now let us suppose the most complex situation $0 < \delta \le 2$. *If* $d_{xz} < \delta$ *then we have the correlated case* ($d_{xy} + d_{yz} = d_{xz}$) *and for* $d_{xz} \to \delta$ *the supremum* δ^2 *of* $\det M$ *for the correlated case is reached. Now let* $d_{xz} = \delta + \Delta$ *for* $0 \le \Delta \le 2 - \delta$. *The maximum of the function* $d_{xy} \to \det M$ *under the constraint* $d_{xz} = \delta + \Delta$ *is attained for the two points* δ *and* Δ *with a common value* m_Δ. *We have* $\lim_{\Delta \to 0+} m_\Delta = \delta^2$ *and the function* $\Delta \to m_\Delta$ *is strictly increasing.*□

Figure 1, which plots the value of the design criterion for various values of y and δ ($x = -1, z = 1$), makes the behaviour of the optimal design very transparent. It is easy to see that minimal points lie along the main axes and that the design (at least the central point) gets irrelevant when the correlation is large

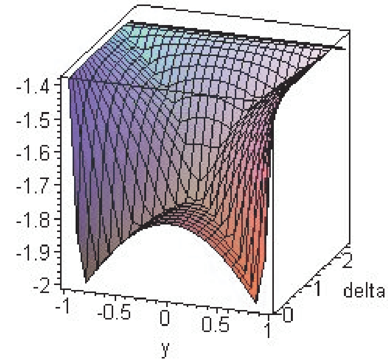


Figure 1: The criterion function with fixed $x = -1$ and $z = 1$.

Remark. Note that in the case $\delta = 1$ the optimal design is thus of the form $\{-1, 0, 1\}$, which coincides with the uniformly optimal design found by [5] in this setting.

However, it is evident that, unlike the two-point case, here we have a dependence of the optimal design on the value of the parameter δ, which for practical purposes must be resolved by e.g. employing local design criteria.

3 All parameters need to be estimated

The situation gets much more complicated (and interesting) when we also desire to estimate the parameter of the covariance function.

3.1 Two-point designs

Let us again start our considerations with the two points case, i.e. let $\{x, z\}$, $-1 \leq x < z \leq 1$ be a proposed design, with parameter δ, $0 < \delta$. Here the elements $M_{i,j}$ of the now 3×3 information matrix M have the form

$$M_{1,1} = E(-\frac{\partial^2 L}{\partial \vartheta_1^2}) = 2\frac{1 + (-1 + \frac{d}{\delta})J_\delta(d)}{1 - (1 - \frac{d}{\delta})^2 J_\delta(d)},$$

$$M_{1,2} = E(-\frac{\partial^2 L}{\partial \vartheta_1 \partial \vartheta_2}) = \frac{(x + z)(1 + (-1 + \frac{d}{\delta})J_\delta(d))}{1 - (1 - \frac{d}{\delta})^2 J_\delta(d)},$$

$$M_{2,2} = E(-\frac{\partial^2 L}{\partial \vartheta_2^2}) = \frac{x^2 + z^2 + 2xz(-1 + \frac{d}{\delta})J_\delta(d)}{1 - (1 - \frac{d}{\delta})^2 J_\delta(d)},$$

$$M_{3,3} = E(-\frac{\partial^2 L}{\partial \delta^2}) = J_\delta(d)\frac{d^2 - 2\delta d + 2\delta^2}{\delta^2(2\delta - d)^2},$$

$$M_{1,3} = M_{2,3} = 0.$$

Note that the information matrix is singular for $d \geq \delta$ and regular otherwise.

The D-optimality criterion function Φ has the form

$$\Phi(M) = -\ln\det M = -\ln M_{3,3}(M_{1,1}M_{2,2} - M_{1,2}^2).$$

and further on

$$\Phi(d) = \begin{cases} -\ln d\frac{d^2-2\delta d+2\delta^2}{(2\delta-d)^3}, & \text{for } 0 < d < \min\{\delta, 2\}, \\ +\infty, & \text{for } \delta \leq d \leq 2. \end{cases}$$

Proposition 3.1. *There exists an exact D-optimal design $\{-1, 1\}$ when $2 < \delta$ and there exists no exact D-optimal design for $\delta \leq 2$.*

Proof *The derivative of the function $d \to d\frac{d^2-2\delta d+2\delta^2}{(2\delta-d)^3}$ is positive for* $0 < d < \min\{\delta, 2\}$. □

Note, that the non-existence of an exact design for $\delta \leq 2$ is irrelevant for practical purposes, since for any finite approximation of the design space V we can always find a "best" design. The infinite value of the D-optimality criterion function can be interpreted as the impossibility to observe the variance parameter δ in the uncorrelated case. Thus we require a somewhat artificial composition of two criterion functions for the correlated and uncorrelated case, to maintain the comparison of both correlated and uncorrelated designs, which leads to a discontinuity and some interpretational issues. It is evident that this property is problematic for all criteria that rely on the regularity of the information matrix. It might be an advantage to resort to prediction rather than estimation based criteria, like the G-criterion, which minimizes the largest expected variance of prediction over the region of interest (see [3]), to compare such different designs.

3.2 Three-point designs

Also in the three point case, which relates to the Näther result, an exact global solution can only be found for $\delta > 2$. Let $\{x, y, z\}$, $-1 \leq x < y < z \leq 1$ be a proposed design, with δ as the parameter. Then the information matrix and D-optimality criterion function Φ has the following form

$$M_{1,1} = E(-\frac{\partial^2 L}{\partial \vartheta_1^2}) = \frac{2\delta}{2\delta - d_{xz}},$$

$$M_{1,2} = E(-\frac{\partial^2 L}{\partial \vartheta_1 \partial \vartheta_2}) = \frac{\delta(x+z)}{2\delta - d_{xz}},$$

$$M_{2,2} = E(-\frac{\partial^2 L}{\partial \vartheta_2^2}) = \delta \frac{2xz + \delta d_{xz}}{2\delta - d_{xz}},$$

$$M_{3,3} = E(-\frac{\partial^2 L}{\partial \delta^2}) = \frac{3d_{xz}^2 - 8\delta d_{xz} + 8\delta^2}{2\delta^2(2\delta - d_{xz})^2},$$

$$M_{1,3} = M_{2,3} = 0.$$

The D-optimality criterion function Φ does not depend on y and has form

$$\Phi(d_{xz}) = -\ln d_{xz} \frac{3d_{xz}^2 - 8\delta d_{xz} + 8\delta^2}{2(2\delta - d_{xz})^3}.$$

Theorem 3.1. *When both correlated and uncorrelated observations are possible ($0 < \delta \leq 2$), we have no exact optimal design (but still we can find the best one on a finite grid). When only correlated observations are possible ($2 < \delta$), as in the fixed case the minimum of the criterion function is attained for $\{-1, y, 1\}$ where y is arbitrary, which constitutes the set of D-optimum designs.*

Proof *As in the fixed case, the determinant of the information matrix can be written in the form* $\det M = g(d_{yz}, d_{yx}, d_{xz})$, *where g is a quite complex function (for further details see [5]). When $\delta > 2$, the maximum of the* $\det M = d_{xz} \frac{3d_{xz}^2 - 8\delta d_{xz} + 8\delta^2}{2(2\delta - d_{xz})^3}$ *is attained for $\{-1, y, 1\}$ where y is arbitrary.*

Now let us suppose the most complex situation $0 < \delta \leq 2$. First let us have $1 < \delta < 2$. If $d_{xz} < \delta$ then we have correlated case ($d_{yx} + d_{yz} = d_{xz}$) and for $d_{xz} \to \delta$ the supremum $\frac{3}{2}$ of $\det M$ *for the correlated case is reached. Now let $d_{xz} = \delta + \Delta$ for $0 \leq \Delta \leq 2 - \delta$. The element $M_{3,3} = E(-\frac{\partial^2 L}{\partial \delta^2})$ of the information matrix has the form $\frac{2\delta^2 - 2\delta d_{yx} + d_{yx}^2}{\delta^2(-2\delta + d_{yx})^2}$, for $0 < d_{yx} \leq \Delta$,*

$$\frac{2\Delta^3\delta + \Delta^4 + 12d_{yx}^2\Delta\delta - 2\Delta\delta^2 d_{yx} + 2\delta^3 d_{yx} + 2d_{yx}^2\delta^2}{\delta^2(-2\delta d_{yx} + 2d_{yx}^2 + \Delta^2 - 2\Delta d_{yx})^2} +$$

$$+ \frac{-8\Delta d_{yx}^3 + \Delta^2\delta^2 - 8\delta d_{yx}^3 + 8d_{yx}^2\Delta^2 - 4\Delta^3 d_{yx} + 4d_{yx}^4 - 8\delta d_{yx}\Delta^2}{\delta^2(-2\delta d_{yx} + 2d_{yx}^2 + \Delta^2 - 2\Delta d_{yx})^2},$$

for $\Delta < d_{yx} < \delta$ and $\frac{d_{yx}^2 - 2\Delta d_{yx} + \delta^2 + \Delta^2}{\delta^2(d_{yx} - \Delta + \delta)^2}$ for $\delta \leq d_{yx} < \delta + \Delta$. Function $d_{yx} \to M_{3,3}$ is strictly increasing for $0 < d_{yx} \leq \Delta$ and for $\frac{\delta+\Delta}{2} < d_{yx} < \delta$, and decreasing for $\Delta < d_{yx} < \frac{\delta+\Delta}{2}$ and for $\delta \leq d_{yx} < \delta + \Delta$. We have *a positive jump* $j(\delta) = \lim_{d_{yx}\to\Delta^+} M_{3,3} - \lim_{d_{yx}\to\Delta^-} M_{3,3} = \lim_{d_{yx}\to\delta^-} M_{3,3} - \lim_{d_{yx}\to\delta^+} M_{3,3} = \frac{1}{\Delta(2\delta-\Delta)}$. *There is no global maximum of* $\det M$ *for any* Δ.

Now let us have $\delta = 2$. *Then the function* $d_{yx} \to \det M$ *is strictly decreasing for* $0 < d_{yx} < 1$ *and strictly increasing for* $1 < d_{yx} < 2$ *and defined on the open set* $(0, 2)$ *(repetitions are not allowed). So there exist no global maximum of* $\det M$.

Now let $0 < \delta \le 1$. *If* $d_{xz} < \delta$ *then we have correlated case* ($d_{yx} + d_{yz} = d_{xz}$) *and for* $d_{xz} \to \delta$ *the supremum* $\frac{3}{2}$ *of* $\det M$ *for the correlated case is reached. Now let* $d_{xz} = \delta + \Delta$ *for* $0 \le \Delta < \delta$. *Then we obtain behaviour as in the previous case* ($1 < \delta \le 2$) *and have no optimal design. For* $\Delta \ge \delta$ *has the element* $M_{3,3} = E(-\frac{\partial^2 L}{\partial \delta^2})$ *form* $\frac{2\delta^2-2\delta d_{yx}+d_{yx}^2}{\delta^2(-2\delta+d_{yx})^2}$, *for* $0 < d_{yx} < \delta$, 0, *for* $\delta \le d_{yx} \le \Delta$ *and* $\frac{d_{yx}^2-2\Delta d_{yx}+\delta^2+\Delta^2}{\delta^2(d_{yx}-\Delta+\delta)^2}$ *for* $\Delta < d_{yx} < \delta + \Delta$. *We clearly have no global maximum of* $\det M$.$\square$

Remark. Although the expression for $M_{3,3} = E(-\frac{\partial^2 L}{\partial \delta^2})$ is quite complex, it can be written elegantly using special polynomials (see [6]) generated from the design points. This can be done in such a way, that they can be particularly useful for further computational purposes.

4 Outlook and computational aspects

The optimum designs that were calculated can be helpful in various ways. For more complex settings, where we have to employ numerical algorithms, such as the classic one suggested by [2], the designs will provide good starting values.

But more importantly, there are a number of lessons to be learned, when dealing with design problems in (potentially) correlated cases. These lessons continue to be of relevance for the computational solutions in practically more realistic situations. Firstly, we recommend not to use undifferentiable covariance functions, when it can be easily avoided. As what was seen, they cause discontinuities in the criterion function and may - in a technical sense - lead to non-existence of optimal designs. Secondly, at least in the case of estimable parameters in the covariance function, the use of design criteria and hence computational procedures on the basis of the information matrix will involve complex decisions on how to cope with singularities induced by "uncorrelated" situations. It will, however, be one topic of our future research to provide more general numerical algorithms for these purposes.

Appendix

For design $\{x, y, z\}$, $-1 \le x < y < z \le 1$ and fixed $0 \le \delta$ we have

$$-L(\vartheta_1, \vartheta_2) = \textit{constant on } \vartheta_i + \frac{1}{2\det\Sigma} \times \{(1-\operatorname{cov}^2(y,z))m(x,x)$$

$$+(1-\operatorname{cov}^2(x,z))m(y,y) + (1-\operatorname{cov}^2(x,y))m(z,z)+$$

$$2(\mathrm{cov}(z,y)\mathrm{cov}(x,z)-\mathrm{cov}(x,y))m(x,y)+2(\mathrm{cov}(x,y)\mathrm{cov}(y,z)-$$
$$-\mathrm{cov}(x,z))m(x,z)+2(\mathrm{cov}(x,y)\mathrm{cov}(x,z))m(x,y)-\mathrm{cov}(y,z)\},$$

where we use formal multiplication

$$m(x,y)=(X(x)-\vartheta_1-\vartheta_2 x)(Y(y)-\vartheta_1-\vartheta_2 y),$$

$\det\Sigma = 1-\mathrm{cov}^2(x,y)-\mathrm{cov}^2(y,z)-\mathrm{cov}^2(x,z)+2\mathrm{cov}(x,y)\mathrm{cov}(y,z)\mathrm{cov}(x,z)$. For $\delta > 0$ as a free parameter we have to add $\frac{1}{2}\log\det\Sigma$ to this expression.

References

[1] Bischoff W. (1993). *On D-optimal designs for linear models under correlated observations with an application to a linear model with multiple response.* Journal of Statistical Planning and Inference **37**, 69 – 80.

[2] Brimkulov U.N., Krug G.K., Savanov V.L. (1980). *Numerical construction of exact experimental designs when the measurements are correlated* (in Russian). Zavodskaya Laboratoria (Industrial Laboratory) **36**, 435 – 442.

[3] Fedorov V.V., Hackl P. (1997). *Model-oriented design of experiments.* Lecture Notes in Statistics, Volume **125**, Springer-Verlag, New York.

[4] Müller W.G., Pázman A. (2003). *Measures for designs in experiments with correlated errors.* Biometrika **90** (2), 423 – 434.

[5] Näther W. (1985). *Effective observation of random fields.* Teubner-Texte zur Mathematik, Band 72, BSB B. G. Teubner Verlagsgesellschaft, Leipzig.

[6] Stehlík M. (2004). *Further aspects on an example of D-optimal designs in the case of correlated errors.* Report #1 of the Research Report Series of the Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Austria.

Acknowledgement: Research of Milan Stehlík was partly supported by FWF-grant (Austrian grant agency) P16809-N2 and by the VEGA grant (Slovak grant agency) No 1/0264/03.

Address: W.G. Müller, Department of Statistics and Mathematics, University of Economics Vienna, Augasse 2-6, A-1090 Vienna, Austria
M. Stehlík, Department of Probability and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava, Mlynská Dolina, 842 48 Bratislava, Slovakia

E-mail: werner.mueller@wu-wien.ac.at, stehlik@fmph.uniba.sk

COMPSTAT'2004 Symposium

TAR-GARCH AND STOCHASTIC VOLATILITY MODEL: EVALUATION BASED ON SIMULATIONS AND FINANCIAL TIME SERIES

Pilar M. Muñoz, Maria Dolores Márquez, Manuel Martí-Recober, Cesar Villazón and Lesly Acosta

Key words: Stochastic volatility, TAR-GARCH, excess kurtosis, asymmetry.

COMPSTAT 2004 section: Times series analysis.

Abstract: The paper analyzes the empirical performance between the Stochastic Volatility (SV) and TAR-GARCH models. SV models are flexible enough to explain excess kurtosis, although the inclusion of TAR in the GARCH model improves the variance asymmetry in the time series. These models are used to analyze three daily time series (one simulated series and two financial time series) in order to illustrate the performance of both models. The analysis of residuals is used to evaluate the goodness of fit.We conclude that the SARV model is more useful for capturing the main features of the volatility time series than the TAR-GARCH model.

1 Introduction

Volatility is an important characteristic of financial markets. Understanding and modelling stock volatility is necessary to traders for hedging against risk, in options pricing theory or as a simple risk measure in many asset pricing models. A special feature of stock volatility is that it is not directly observable, but has some characteristics commonly seen in asset returns such as fat tails, volatility clustering, leverage effects, long memory and asymmetric patterns.

In the empirical financial literature, the most important class of nonlinear models is the family of conditional heteroscedastic models which can be classified into two general categories. The first category uses an exact function to govern the evolution of volatility, for example, generalized autoregressive heteroscedastic model (GARCH) proposed by Bollerslev [2]. The second category uses a stochastic function to describe the volatility; the Stochastic Volatility (SV) models of Taylor [13] belong to this category.

Finally, the Threshold models [14] are another, highly interesting alternative for modelling volatility. These models are based on the piecewise linearization of non-linear models over the state-space by the introduction of thresholds. The TAR model is a piecewise AR model, where the switching mechanism is controlled by the delayed process variable (threshold variable), not by the time index. The TAR model definition is very useful for handling non-linear features of the volatility such as asymmetric responses in volatility between positive and negative returns, time-irreversibility, limit cycle and jump phenomena. The above models can be combined to produce "second-generation" models[1], for instance, threshold GARCH (TGARCH) models [15]. In this field, the variety and complexity is clearly unlimited [12].

Many papers draw a comparison between two of the models mentioned above. For example, Carnero et al.[3] and Fleming and Kirby [5] compared the GARCH and SV models. Evaluated SETAR and GARCH are in [1].

Our approach in this paper is to analyze and compare the performance of the SV and TAR-GARCH models. The following section briefly introduces SV and TAR-GARCH(1,1) formulation models. In Section 3, we apply the models to analyze three daily time series. Finally, Section 4 draws conclusions from our results and provides suggestions for future work.

2 Models

In general, financial time series present excess kurtosis, asymmetric effects of positive or negative shocks, small first-order autocorrelation of squared observations and slow decay towards zero of the autocorrelation coefficients of squared observations. Many researchers argue that variance responds asymmetrically to past returns: an asymmetric effect is produced because variance tends to be higher under the influence of bad news than under the influence of good news. SV models are very flexible in capturing the excess-kurtosis observed [3] and an SV formulation with leverage effects captures the asymmetric behaviour in stock returns. Another possibility is to use a threshold model with conditional heteroscedasticity, for example, the TAR-GARCH models. The GARCH model can describe volatility clustering and excess kurtosis (although not entirely), whereas the TAR variance formulation captures the asymmetric patterns of volatility.

2.1 SARV(1) models

Let's consider the basic model of the stochastic volatility

$$y_t = \sigma_t \varepsilon_t \tag{1}$$

where y_t is the observed variable, in general the return on an asset, and σ_t is the unobserved volatility of y_t.

The evolution of volatility is governed by the equation

$$\log \sigma_t^2 = \mu + \phi(\log \sigma_{t-1}^2 - \mu) + \sigma_\nu \nu_t \tag{2}$$

where the errors ε_t and ν_t are Gaussian white sequences. This means that the logarithm of the conditional variance, $\log \sigma_t^2$, follows an autoregressive

[1] Tong [14] classified the non-linear models into "first-generation models" and "second-generation models". AR, ARMA, ARIMA,GARCH, SV and TAR models are included in the first class.

model. For this reason, this model is also known as the Stochastic Autoregressive Volatility model (SARV). Many authors recommend working with the logarithm of the variance in equation (1b) instead of the volatility to reduce the impact of the outliers.

In state space representation, equation 1 is called the observation equation, whereas equation 2 is the system equation. In this case, $x_t = \log \sigma_t^2$ is the state vector at time t. Obviously x_t is a latent variable.

The main feature of (1) is that they describe a discrete time, non-linear and Gaussian dynamic system, which evolves as a first-order Markov process. Many authors solve the estimation problem using the Bayesian approach [6]. Our main goal is to estimate the system state, i.e., the volatility. However, we need to implicitly estimate the unknown parameters $\theta = (\mu, \phi, \sigma_\nu)$ The usual approach is to include the parameters as part of the state vector $(x_t, \theta)\prime$ [10]. From a Bayesian point of view, we wish to obtain the *a posteriori* PDF $p(x_t, \theta|D_t)$ when the observation y_t arrives at time t. To accomplish this, a recursive formula based on Bayes' rule is used assuming that the *a priori* PDF[2] $p(x_{t-1}, \theta|D_{t-1})$ at time $t-1$ is known, where $D_t = \{y_1, \dots, y_t\}$ represents the available information at time $t-1$.

In this paper the *a posteriori* PDF is obtained using a modified version of the SIR (Sampling Importance Resampling) filter to find the parameter estimate when it is unknown. The standard SIR filter has the problem that the particles associated with the parameter do not regenerate, meaning that it has an impoverishment problem a few iterations later. To solve this situation, we propose introducing the jitter suggested in [8] at the end of each iteration, to ensure that the particles cover the *a posteriori* distribution.

The results presented in this paper corresponds to the priors proposed in [4] and [8]: $p(x_0) \sim N(0,1)$ where x_0 is the initial state

$$p(\mu) \sim N(-8, 5^2) \qquad p(\beta) \sim 2Beta(20, 1.5) - 1 \qquad p(\sigma_\nu^2) \sim 0.05/\chi_5^2$$

The priorfor the μ parameter is taken diffuse because it is not available prior information about this parameter. In this kind of series the persistence parameters β takes values about 0.9, is for this reason that would be appropriate an informative priori like the used in this case. A prior density for the variance cannot take on negative values,these authors proposed a prior from the inverted gamma family. The algorithm proposed is described in Annex 1.

2.2 TAR-GARCH(1,1) models

The most widely used models in financial time series are the GARCH models introduced Bollerslev [2]. The GARCH (1,1) model provides a simple parametric function to describe the volatility evolution and is capable of describing volatility clustering, and excess kurtosis (although not completely). Most non-linear extensions of the GARCH model are designed to allow asymmetric

[2] PDF: Probability Density Function

patterns, for example, the TAR model can be used to allow an asymmetric response in the conditional variance.

The TAR-GARCH(1,1) model was proposed in [11]. We can formulate the model as follows:

$$y_t = \sigma_t \varepsilon_t \tag{3}$$

$$\sigma_t^2 = \alpha_0^1 + \alpha_1^1 y_{t-1}^2 + \beta_1^1 \sigma_{t-1}^2 \qquad if \qquad y_{t-1} \leq 0 \tag{4}$$

$$\sigma_t^2 = \alpha_0^2 + \alpha_1^2 y_{t-1}^2 + \beta_1^2 \sigma_{t-1}^2 \qquad if \qquad y_{t-1} > 0 \tag{5}$$

$$\alpha_0^i > 0, \quad \alpha_1^i \geq 0, \quad \beta_1^i \geq 0 \quad i = 1,2 \tag{6}$$

where $\{\varepsilon_t\}$ is a sequence of iid random variables with mean 0 and variance 1.

Notice that in comparison with a threshold general model, in this case we set the delay in 4 and 5 to 1, and the threshold value to 0. These parameters may be estimated by the algorithm described in [9].

The non-negativity constraints on the parameters make these models linear and stationary. They allow for two-stage, least-squares estimation methods and provide simple test statistics for conditional homoscedasticity.

3 Empirical results

In order to illustrate the ability of the models to capture the characteristics of financial time series, we considered three series, one simulated and the other two, real. In the artificial time series, designated $Y(t)$ in the paper, we simulated a sample of 2000 observations using the SARV model described by equation 1 with the parameters proposed in [5] ($\mu = -0.632, \phi = 0.981$ and $\sigma_\nu = 0.194$). The two real time series were the daily log returns, in percentages and including dividends, of IBM stock (IBM), (3/7/1962 - 10/22/1998)[3] and the daily log returns of the S&P 500 index (S&P) (12/10/1990 - 12/28/2001)[4]. The observations analyzed in the IBM time series correspond to the residuals from an AR(2) model as proposed by Tsay [15]. Table 1 reports the time series statistics.

For the SARV model, we implemented the algorithm described in Appendix 1, which uses M =50,000 particles. We repeated the estimation process 200 times for each time series and reported the arithmetic average for each parameter estimated and the statistics values. The TAR-GARCH model estimation was done using maximum likelihood. In both cases, we used the R software package[5]. The parameters we estimated are listed in Table 2 for the TAR-GARCH(1,1) model and in Table 3 for the SARV model.

[3] http://gsb.uchicago.edu/fac/ruey.tsay/teachins/fts

[4] http://www.math.ku.dk /˜rolf /teaching /ctff03 /SP500.GARCH.R

[5] http://cran.r-project.org

	$Y(t)$	IBM	$S\&P$
N	2000	9142	2787
$Mean$	0.07	0.04	0.00
$St.D.$	0.75	1.49	0.01
$Skew$	-0.03	-0.34*	-0.26*
$Kurt.$	1.98*	14.59*	4.68*
$r(1)$	-0.05	0.00	0.01
$Q(20)$	37.8*	25.5	46.3*
$Autocorrelations\ y_t^2$			
$r_2(1)$	0.11*	0.16*	0.20*
$r_2(2)$	0.09*	0.08*	0.16*
$r_2(5)$	0.19*	0.08*	0.19*
$r_2(10)$	0.12*	0.02*	0.09*
$Q_2(20)$	515*	424*	679*

Table 1: Time series desc. statistic.

TARGARCH	*Y(t)*		*IBM*		*SP*	
(1, 1)	*1st. r.*	*2nd. r.*	*1st. r.*	*2nd. r.*	*1st. r.*	*2nd. r.*
N	1002	997	4691	4450	1321	1463
$\widehat{\alpha}_0$	0.024 (0.007)	0.028 (0.009)	2.022 (0.06)	0.012 (0.002)	1.3E-6 (3.3E-7)	1.3E-7 (7.4E-8)
$\widehat{\alpha}_1$	0.124 (0.019)	0.124 (0.022)	0.187 (0.006)	0.033 (0.002)	0.066 (0.008)	0.033 (0.004)
$\widehat{\beta}_1$	0.839 (0.022)	0.827 (0.035)	4.1E-15 (0.016)	0.963 (0.003)	0.925 (0.010)	0.966 (0.004)
$\widehat{\alpha}_1 + \widehat{\beta}_1$	0.968	0.951	0.187	0.995	0.991	0.999
log L	-271.563		-8008.95		11818.86	

Table 2: Estimated TAR-GARCH (1,1) model.

3.1 Model checking

The model diagnostics were based on the standardized observations, defined as $\widehat{\varepsilon}_t = \frac{y_t}{\widehat{\sigma}_t}$ where $\widehat{\sigma}_t$ is obtained by substituting the estimated parameters in equation 2 for the SARV model or 3 for the TAR-GARCH model. As suggested in [15], the Ljung-Box statistics of $\widehat{\varepsilon}_t$ can be used to check the adequacy of the mean equation and that of $\widehat{\varepsilon}_t^2$ can be used to test the validity of the volatility equation. The skewness, kurtosis, and quantile-to-quantile plot of $\{\widehat{\varepsilon}_t\}$ can be used to check the validity of the distribution assumption.

For the simulated series we can obtain the values for the mean absolute error (MAE) and the root mean square error (RMSE), in order to evaluate the forecasting performance of the models. Finally, the diagnostics for the series are contained in Table 4.

	Y(t)	*IBM*	*S&P*
N	2000	9142	2787
$\widehat{\mu}$	-0.763 (0.13)	0.324 (0.07)	-9.862 (0.07)
$\widehat{\phi}$	0.971 (0.01)	0.964 (0.00)	0.982 (0.00)
$\widehat{\sigma}_\nu$	0.191 (0.02)	0.188 (0.00)	0.197 (0.02)
L	-43.1	-6340	11545

Table 3: Estimated SARV model N: Sample Size ( ) St. dev. $r(k)$ and $r_2(k)$ Autoc.order k orig. and sq.observ.respec. $Q(20)$-$Q_2(20)$:Box-Ljung orig.and sq. obs. respectively.

	Y(t)		*IBM*		*SP*	
$\widehat{\varepsilon}_t = \frac{y_t}{\widehat{\sigma}_t}$	TG	SARV	TG	SARV	TG	SARV
$Mean$	0.0125	0.012	-0.0015	-0.011	0.0437	0.055
$St.Dev$	0.9703	0.945	0.950	0.983	0.9701	0.941
$Skew$	0.014	-0.01	-0.061*	0.066*	-0.475*	-0.169*
$Kurt.$	1.152*	-0.031	6.197*	2.94	2.513*	3.078
$r(1)$	-0.041	-0.042	0.013	0.024	0.040*	0.044*
$Q(20)$	27.058	22.75	28.425	25.61	37.95*	40.14*
$r_2(1)$	-0.023	-0.015	-0.001	0.018	0.009	-0.022
$r_2(10)$	0.008	0.001	0.028	-0.006	-0.002	0.016
$Q_2(20)$	28.813	20.334	248.159	31.29	11.114	16.07
$Obs > 3.5$	8(0.4%)	0(0%)	41(0.4%)	5(0%)	9(0.3%)	2(0%)
$RMSE$	0.555	0.482	* Significant at the 5% level			
MSA	0.451	0.381				

Table 4: Table 4: Descr.stat.stand. observ.

4 Conclusions

In conclusion, our results basically show that the SARV model is more useful than TAR-GARCH(1,1) for capturing the main features of volatility series.

The two real series exhibit skewness and kurtosis, while the simulated one $Y(t)$ only has kurtosis. Although the series are not correlated, the squared observations are correlated, which reflects the existence of volatility. The rejection of the hypothesis of linear independence is typical in time series with volatility.Our results show that the persistence of volatility estimated with the TAR-GARCH(1,1) model is higher than that estimated with the SARV model ,except for the IBM time series. This could be due to the fact the TAR-GARCH needs to have a persistence very close to one to explain high kurtosis and low $r_2(1)$. The persistence determines how fast the autocorrelation decreases towards zero; if it is close to one, the autocorrelations decrease only very gradually. This implies that the impact of shocks on conditional variance diminishes only very slowly.

In the simulated series and SARV models the results are like to the obtained in [5]. Carnero et al [3] showed a SARV model to the S&P time series in that the estimated parameter and residual analysis are similar to the present in this paper, although the sample took in [3]is from 11/1987 to 12/1998. Finally, to the IBM series, the persistence estimated for the second regime in the TAR-GARCH model is closed to the estimation in [15].

It would be interesting to evaluate how these models perform in the economic sense and to analyze the availability of these models for practitioners. We believe the TAR-GARCH model works better for a composite index time series than a stock returns time series. An interesting alternative would be to extend the SARV model to the THSV (threshold stochastic volatility) model proposed in [12], which would allow this model to capture the mean and variance asymmetries in time series simultaneously.

Annex 1: Pseudo-code Sequential Importance Resampling with Jitter (SIRJ) Algorithm.

```
1. Initialization, $t = 0$
        ○ FOR  $j = 1 : M$
              Samples $x_0^{(j)} \sim p(x_0)$, $\theta_0^{(j)} \sim p(\theta_0)$
        ○ END FOR
    • FOR $t = 1 : N$

2. Importance sampling
        ○ FOR  $j = 1 : M$
              Prediction:
              - Generate a random number $\nu_t^{(j)}$, according to the noise
                density associated to the state (eq. (2b)), and
              - Calculate $x_t^{(j)} = \mu_{t-1}^{(j)} + \phi_{t-1}^{(j)}(x_{t-1}^{(j)} - \mu_{t-1}^{(j)}) + \sigma_\nu^{(j)} \nu_t^{(j)}$
              Filtering: Assign to each particle $x_t^{(j)}$ the new weight
               $w_t^{(j)} \propto N(y_t \mid 0, \exp(\ x_t^{(j)})$
        ○ END FOR
        ○ FOR $j = 1 : M$
              Normalize the importance weights: $\widetilde{w}_t^{(j)} = w_t^{(j)} / \sum_{i=1}^{M} w_t^{(i)}$
        ○ END FOR
```

```
3. Resampling
        ○Resampling with replacement M times the particles
        $\left\{x_t^{(1)},\ldots,x_t^{(M)},\theta_t^{(1)},\ldots,\theta_t^{(M)}\right\}$ with the importance weights
        $\left\{\widetilde{w}_t^{(1)},\ldots,\widetilde{w}_t^{(M)}\right\}$
4. Jitter
        ○ FOR $j = 1 : M$
            Sample a new parameter vector: $\theta_t^{(j)} \sim N(\bullet \mid m_t^{(j)}, h^2 V_t)$
            where $m_t^{(j)} = a\theta_t^{(j)} + (1-a)\bar{\theta}_t$ and $\bar{\theta}_t$ and $V_t$ are mean and
            variance of the Monte Carlo approximation to $p(\theta|D_t)$
            $a$ around $0.95 - 0.995$, $h^2 = 1 - a^2$
        ○ END FOR
    • END FOR
```

References

[1] Boero G., Marrocu E. (2002). *The performance of non-linear exchange rate models: a forecasting comparison.* Journal of Forecasting **21**, 513–542.

[2] Bollerslev T. (1986). *Generalized autoregressive conditional heterocedasticity.* Journal of Econometrics **31**, 307–327.

[3] Carnero M.A., Peña D., Ruíz E. (2001). *Is stochastic volatility more flexible than GARCH?* W.P. **01-08**. Universidad Carlos III de Madrid.

[4] Chib S., Nardari F., Shepahard N. (2002). *Markov chain Monte Carlo methods for stochastic volatility models.* Journal of Econometrics **108**, 281–316.

[5] Fleming J., Kirby C. (2003). *A closer look at the relation between GARCH and stochastic autoregressive volatility.* Journal of Forecasting (forthcoming).

[6] Harvey A., Ruiz E., Shepard N. (1994). *Multivariate stochastic variance models.* Review of Economic Studies **61**, 247–264.

[7] Li W.K., Lam K. (1995). *Modelling the asymmetry in stock returns by a threshold ARCH model.* The Statistician **44**, 333–341.

[8] Liu J., West M. (2001). *Combined parameter and state estimation in simulation-based filtering.* In: Sequential Monte Carlo Methods in Practice (eds. A. Doucet, M. De Freitas and N. Gordon), Springer-Verlag; New York.

[9] Márquez M. D. (2002). *Volatility of return time series with SETAR models: The improvement of the algorithm of identification.* Doctoral dissertation, Universitat Politècnica de Catalunya.

[10] Muñoz M.P., Egozcue J.J., Martí Recober M. (1988). *Estimació del pol i de la variància del soroll d'un model AR(1) mitjançant filtratge no lineal.* Qüestió **12**, 21 – 42.

[11] Rabemananjara J.M., Zakoïan R.Y. (1993). *Threshold ARCH and asymmetries in volatility.* Journal of Applied Econometrics **8**.

[12] So M.K., Li W.K., Lam K. (2002). *A threshold stochastic volatility model.* Journal of Forecasting **21**, 473 – 500.

[13] Taylor S.J. (1994). *Modelling stochastic volatility: A review and comparative study.* Mathematical Finance **4**, 183 – 204.

[14] Tong H. (1990). *Non linear time series: A dynamical system approach.* Oxford University Press.

[15] Tsay R.S. (2002). *Analysis of financial time series.* Wiley-Interscience.

Address: M.P. Muñoz, M. Martí-Recober, L. Acosta, Dep.Estadística i Investigació Operativa, UPC, Barcelona, Spain
Márquez, Villazón, Dep.Economia Empresa, UAB, Bellaterra (Barcelona), Spain

E-mail: (pilar.munyoz,manuel.marti-recober,lesly.acosta)@upc.es
(mariadolores.marquez,cesar.villazon)@uab.es

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

QUANTIFYING ULTRAMETRICITY

Fionn Murtagh

Key words: Classification, clustering, hierarchy, ultrametric, p-adic.

COMPSTAT 2004 section: Clustering.

Abstract: The ultrametric properties of hierarchic clustering are well-known. In recent years, there has been interest in ultrametric properties found in statistical mechanics, optimization theory, and physics. It has been shown that sparse, high-dimensional spaces tend to be ultrametric. Given the pervasiveness of ultrametricity, it is important to be able to quantify how close given metric data are to being ultrametric. In this article we assess previously used coefficients of ultrametricity. We present a new coefficient of ultrametricity, and exemplify its properties experimentally. Our immediate objective in this work is to show that sparse, high-dimensional spaces, that are typical of many new data analysis problems in such areas as genomics and proteomics, and speech, tend to be inherently ultrametric.

1 Introduction

Ultrametricity is defined mathematically in section 2.1 below, but can be informally described as follows: there is a natural hierarchical or embedded structure among the data observations under investigation. Hierarchical cluster analysis involves inducing an ultrametric set of relationships on the objects or observations. In the 1980s, ultrametric spaces came under investigation in physics. Some recent work has also used the perspective of ultrametric topology as part of a model of human cognition. An important finding [15] has been that sparse and high-dimensional spaces tend to be ultrametric. This means that such spaces, containing points associated with a set of observations, are characterized by ultrametric (or hierarchical) relationships. The implications of this are far reaching for the analysis of massive, high-dimensional data sets in such fields as speech processing, or proteomics, to name but two.

In this article we will show how sparse, high-dimensional data are found to be ultrametric. Our main focus in this work is the quantifying of ultrametricity.

An initial response to this requirement would be to take a large data set, and construct a hierarchical clustering on it using some suitable clustering criterion. A constructive assessment of ultrametricity is then simply quantifying the discrepancy between input data and induced ultrametric data structure (e.g. through the Euclidean or some other distance between initial pairwise dissimilarities, and induced ultrametric distances).

Examples of such constructive approaches to assessing ultrametricity include: use of any hierarchical clustering algorithm, many of which can be implemented in a stepwise way based on the Lance-Williams update relationship; specifically for quantifying ultrametricity with a well-defined coefficient, Rammal et al. [15] use the single link method; and one can of course use a criterion such as the commonly used least squares fit criterion [3], [5], [11], [6] although it is known than such an approach will only approximate an optimal result [9], [8], [4] for this NP-complete problem.

Quantifying ultrametricity using a constructive approach is less than perfect due to the following:

- Potential complications arising from known problems, e.g. chaining in single link, non-uniqueness of a minimal superior ultrametric (Benzécri, 1976), or inversions (non-compliance with Bruynooghe's reducibility property: see Murtagh [12].
- In the case of the single link method, empirically observed scaling regimes described theoretically and empirically by Rammal et al. [15].
- Suboptimal solutions and dependency on starting configurations in the case of seeking direct optimization of NP-complete problems.

The conclusion here is that the "measurement tool" used for quantifying ultrametricity itself occupies an overly prominent role relative to that which we seek to measure.

In section 2, we will give the formal definition of ultrametricity. In section 3, we will look at various direct approaches to quantifying the extent of ultrametricity in a data set. Section 3.1 details an approach due to Lerman, which is based on ranks of dissimilarities. Section 3.2 uses one particular constructive approach, the single link agglomerative method, and the discrepancy between the resulting ultrametric and the initial distances. Section 3.3 describes two other approaches used in the literature. In Section 3.4 then we describe a new coefficient of ultrametricity, and we describe how the properties of this ultrametricity coefficient are advantageous, and outperform previous results. Section 3.5 presents experimental support for this new ultrametricity coefficient.

2 Relevant ultrametric axioms and triangle properties

2.1 Isosceles triangles with base side smallest

The ultrametric relation implies that triangles among all triplets are isosceles, with base side of smallest length [1], [10].

Consider three points x, y, and z. Without loss of generality let (y, z) be one of the less long sides. (Hence it is either the short base side, or one of the long sides.) Then $d(x, y) \leq \max\{d(y, z), d(x, z)\}$ implies that $d(x, y) \leq d(y, z) \leq d(x, z)$.

Now we permute x and y. But doing this implies that $d(y, x) \leq d(x, z) \leq d(y, z)$. And the only way that we can have simultaneously $d(y, z) \leq d(x, z)$ and $d(x, z) \leq d(y, z)$ is for these to be equal.

Hence we have $d(x, y) \leq d(x, z) = d(y, z)$, QED.

2.2 Ultrametrics, ultramines and their intersection

The ultrametric inequality is: $d(x, y) \leq \max\{d(y, z), d(x, z)\}$ When: $d(x, y) \geq \min\{d(y, z), d(x, z)\}$ then Rizzi [16] terms this an ultramine. Replacing y with x gives $d(x, x) \geq d(x, z)$, so an ultramine is a similarity measure. Lerman [10] uses the term ultrametric proximity. Other than the strong triangular inequality, and symmetry, Lerman [10]) gives the remaining property of an ultramine or ultrametric proximity as: $d(x, y) = +\infty$ whenever $x = y$.

As already noted, a metric space is ultrametric iff all triangles are isosceles with base lesser than or equal to the side lengths. A metric space is ultramine iff all triangles are isosceles with base greater than or equal to the side lengths. The intersection of ultrametrics and ultramines is defined by equilateral triangles.

3 Approaches to quantifying ultrametricity

3.1 Lerman's H measure based on ranks of pairwise dissimilarities

Lerman's measure of ultrametricity is based on ranks of dissimilarities between observations. Use of ranks is for two main reasons: (i) "robustness" is ensured, i.e. limitation of effects of unusually large or unusually small dissimilarities, and (ii) an effective normalization of the dissimilarities results from this, so that comparability between different data sets becomes feasible.

For $x, y, z \in E$: we consider $d(x, y) \leq d(y, z) \leq d(x, z)$. By the ultrametric inequality as seen in section 2.1 we have: $d(x, z) \leq d(y, z)$. Therefore (x, z) and (y, z) must be in the same class of the preorder on $E \times E$.

Hence [10] for all triples x, y, z, if M is median and S is maximum, consider the open interval $]M(x, y, z), S(x, y, z)[$. If this open interval is empty, then the associated preorder is ultrametric.

Given a triplet $\{x, y, z\} \in J$ for which $(x, y) \leq (y, z) \leq (x, z)$, for preorder ω, the interval $]M(x, y, z), S(x, y, z)[$ is empty if ω is ultrametric. Relative to such a triplet, the preorder ω is "less ultrametric" to the extent that the cardinal of $]M(x, y, z), S(x, y, z)[$, defined on ω, is large. We consider the mapping of all triplets J into all pairs F for the given preorder ω. We then define discrepancy between the structure of ω and the structure of an ultrametric preordonnance where $|.|$ denotes cardinality:

$$H(\omega) = \sum_J |]M(x, y, z), S(x, y, z)[|/(|F| - 3)|J|$$

The value 3 subtracted from $|F|$ takes account of the presence of the least, median and maximum distances. If ω is ultrametric then $H(\omega) = 0$. We are basically saying: the (open) interval between median and maximum of a triplet of distances is examined and the number of distances falling in this interval is counted. The "openness" of the median/maximum interval is important: in practice it means that we do not include the median nor maximum value, nor any values tied with them.

As shown in simple cases by Lerman [10, p. 218], data sets that are "more classifiable" in an intuitive way, i.e. they contain "sporadic islands" of more dense regions of points – a prime example is Fisher's iris data contrasted with 150 uniformly distributed values in $\mathbb{R}^4$ – such data sets have a smaller value of $H(\omega)$. For Fisher's data we find $H(\omega) = 0.0899$, whereas for 150 uniformly distributed points in a 4-dimensional hypercube, we find $H(\omega) = 0.1835$.

Generating all unique triplets is computationally intensive: for n points, $n(n-1)(n-2)/6$ triplets have to be considered. Hence, in practice, we must draw triangles randomly (uniformly) from the given point set.

Murtagh [13] gives empirical results based on Lerman's H-classifiability. There are two problems with Lerman's index, however. Firstly, ultrametricity is associated with $H = 0$ but non-ultrametricity is not bounded (nor defined). In experimentation, we have found maximum values for H in the region of 0.24. The second problem with Lerman's index is that for floating point, and high dimensional, points, the strict equality necessitated for an equilateral triangle is nearly impossible to achieve. However our belief is that approximate equilateral triangles are very likely to arise in high-dimensional spaces, due to increasing sparseness. We would prefer therefore that the quantifying of ultrametricity should "gracefully" take account of triplets which are "close to" equilateral. Note that for some authors, the equilateral case is considered to be "trivial" or a "trivial limit" [17]. For us, however, it is an important case, together with the other important case of ultrametricity (i.e., isosceles with small base).

3.2 Discrepancy between subdominant ultrametric and input data

In the Introduction we have indicated that creation of a hierarchical clustering, followed by comparison between the ultrametric distances found and the input set of dissimilarities, was an evident way to quantify ultrametricity, but suffered from some disadvantages. The single link hierarchical agglomerative clustering method has some attractive (and some unattractive!) properties. A constructive quantifying of ultrametricity was based on it. This we will now describe.

The quantifying of how ultrametric a data set is by Rammal et al. [14], [15] is given as an ultrametricity index: $\sum_{x,y}(d(x,y) - d_c(x,y))/\sum_{x,y} d(x,y)$ where d is the metric distance being assessed, and d_c is the subdominant ultrametric. The Rammal index is bounded by 0 (= ultrametric) and 1. As pointed out in Rammal [14], [15], this index suffers from "the chaining effect and from sensitivity to fluctuations".

The chaining effect implies that for $d(x,y) \leq r_0, d(y,z) \leq r_0$ then $d(x,z) = 2r_0 - \epsilon$ for arbitrarily small ϵ. Hence $d(x,z)$ can be anomalously large. Another manifestation is the following pathology postulate of Watson [18]. The subdominant ultrametric d_c of a given metric d can be arbitrarily close to zero, even when there is an ultrametric quite close to d in the supremum norm. This is formulated as follows. If d is a metric on a finite set, then there is an ultrametric d_c which minimizes $\sup\{|d(x,y) - d_c(x,y)| : x, y \in X\}$ among those d_c such that $\forall x, y \in X, d_c(x,y) \leq d(x,y)$.

Rammal et al. [14], [15] discuss a range of important cases: a set of n binary words, randomly defined among the 2^k possible words of k bits; and n words of k letters extracted from an alphabet of size K. For binary words, $K = 2$; for nucleic acids, four nucleotids give $K = 4$; for proteins, twenty amino acids give $K = 20$; and for spoken words, around 40 phonemes give $K = 40$. Using the Rammal ultrametricity index, experimental findings demonstrate that random data, in the sparse limit, are increasingly ultrametric.

3.3 Distance-based measures

Treves [17] considers triplets of points giving rise to minimal, median and maximal distances. In the plot of $d_{\text{min}}/d_{\text{max}}$ against $d_{\text{med}}/d_{\text{max}}$, the triangular inequality, the ultrametric inequality, and the "trivial limit" of equilateral triangles, occupy definable regions.

Hartmann [7] considers $d_{\text{max}} - d_{\text{med}}$. Now, Lerman [10] uses ranks in order to give (translation, scale, etc.) invariance to the sensitivity (i.e., instability, lack of robustness) of distances. Hartmann instead fixes the remaining distance d_{min}.

3.4 A new measure based on angles

We seek to avoid, as far as possible, lack of invariance due to use of distances. We seek to quantify both isosceles with small base configurations, as well as equilateral configurations. Finally, we seek a measure of ultrametricity bounded by 0 and 1. We will therefore use a coefficient of ultrametricity – we will term it α – which is specified algorithmically as follows.

1. All triplets of points are considered, with a distance defined (by default, Euclidean). Since for a large number of points, n, the number of triplets, $n(n-1)(n-2)/6$ would be computationally prohibitive, we instead randomly (uniformly) sample coordinates $(i \sim \{1..n\}, j \sim \{1..n\}, k \sim \{1..n\})$.
2. We check for possible alignments (implying degenerate triangles) and exclude such cases.
3. Next we select the smallest angle as less than or equal to 60 degrees. (We use the well-known definition of the cosine of the angle facing side of length x as: $(y^2 + z^2 - xy)/2yz$.) This is our first necessary property for being an isosceles ($<$ 60 degrees) or equilateral ($=$ 60 degrees) ultrametric triangle.

4. For the two other angles subtended at the triangle base, we seek an angular difference of strictly less than 2 degrees (0.03490656 radians). This condition is an approximation to the ultrametric configuration. This condition is targeting a configuration that is not exactly ultrametric but nonetheless very close to ultrametric.
5. Among all triplets (1) satisfying our exact properties (2, 3) and close approximation property (4), we define our ultrametricity coefficient as the relative proportion of these triplets. Approximately ultrametric data will yield a value of 1. On the other hand, data that is non-ultrametric in the sense of not respecting conditions 3 and 4 will yield a low value, potentially reaching 0.

The Fisher iris data (150×4) gives $\alpha = 0.0162$, indicating some, limited, ultrametricity. By recoding the four iris variables into discrete (zero or one) categories, we find the following. Firstly, with two discrete categories (data now: 150×8), we find $\alpha = 0.0949$. For four discrete categories (data now: 150×16), we find $\alpha = 0.477327$. For eight discrete categories (data now: 150×32), we find $\alpha = 0.741361$. This shows how increasing dimensionality, and sparseness, lead to greater ultrametricity.

3.5 Ultrametricity scaling with data size, dimensionality, and sparseness

We use uniformly distributed data and also uniformly distributed hypercube vertex positions. The latter is used to simulate the multivalued words considered by Rammal et al. (see above at end of section 3.2). Random values are converted to hypercube vertex locations by use of complete disjunctive data coding [2]. Say a variable has maximum and minimum values $x_{\max}$ and $x_{\min}$. Say, further, that $K = 4$. We set a series of thresholds at intervals given by $(x_{\max} - x_{\min})/(K - 1)$. A value of x falling in the first category receives a 4-valued set: $1, 0, 0, 0$; a value of x falling in the second category receives the 4-valued set: $0, 1, 0, 0$; and so on. Such complete disjunctive coding is widely used in correspondence analysis. It is easily verified that the row marginals are constant.

- We find surprising independence of α relative to n, the number of points. Consider the following: we generate uniformly distributed data points in $\mathbb{R}^{10}$. For $n = 1000, 5000, 10000, 15000, 20000, 25000$, we find $\alpha =$ 0.096386, 0.078000, 0.077077, 0.075075, 0.079000, 0.71000. There appears to be a small decrease in ultrametricity due to increasing density of points. (We found the same result, i.e. independence relative to n, with Lerman's index: see Murtagh [13].)
- Sparsity of coding helps greatly with ultrametricity. We will again take the number of points, $n = 1000, 5000, 10000, 15000, 20000, 25000$. We will also use a 10-dimensional space with, on this occassion, the points

at the vertices of a hypercube. (We do this by generating uniformly in $\mathbb{R}^5$ and then quantizing each of the 5 variables to two discrete categories. See discussion above, earlier in this section). We find, respectively: $\alpha = 0.271630, 0.247495, 0.260563, 0.264056, 0.269076, 0.275275$. With sparsity we again find very little dependence on n. For varying n, these α results are quite similar. However we see a very big change between points in $\mathbb{R}^{10}$ (discussed under the previous bullet point) and points at the vertices of a 10-dimensional hypercube (discussed under this bullet point).

- Dimensionality helps greatly with ultrametricity. Using $n = 5000$ real-valued points, uniformly distributed in space of dimensionality $m =$ 50, 100, 500, 1000, 5000, we find: $\alpha =$ 0.183183, 0.271000, 0.544000, 0.707708, 0.979000.
- Dimensionality and sparsity, combined, force the tendency towards ultrametricity, but the compounding of these two data properties is not as pronounced as we might have expected. Again we take the number of points, $n = 5000$. Using uniform data in real spaces of dimensions 25, 50, 250, 500 and 2500, and then quantizing to two discrete response categories, gives us dimensionalities $m = 50, 100, 500, 1000, 5000$. Our n points are now at the vertices of hypercubes in spaces of dimensionality m. We find $\alpha = 0.179179, 0.172172, 0.454910, 0.588000, 0.934000$.

4 Conclusion

We have clearly shown the dependence of our new ultrametricity coefficient, α, on numbers of points, space dimensionality, and sparsity of this space. Murtagh [13] describes some of the computational implications of this work, for the processing of masssive high-dimensional data sets.

References

[1] Benzécri J.P. (1979). *La Taxinomie.* 2nd ed., Dunod, Paris.

[2] Benzécri J.P. (1992). *Correspondence analysis handbook.* Marcel Dekker, Basel, (transl. Gopalan T.K.).

[3] Chandon J.L., De Soete G. (1984). *Fitting a least squares ultrametric to dissimilarity data: Approximation versus optimization.* In Diday E., Jambu M., Lebart L., Pagès J., Tomassone R. (eds.), Data Analysis and Informatics III, North-Holland, Amsterdam, 213 – 221.

[4] Day W.H.E. (1996). *Complexity theory: an introduction for practitioners of classification.* In Arabie P., Hubert L.J., De Soete G. (eds), Clustering and Classification, World Scientific, 199 – 233.

[5] De Soete G. (1987). *Least squares algorithms for constructing constrained ultrametric and additive tree representations of symmetric proximity data.* Journal of Classification **4**, 155 – 173.

[6] De Soete G., Carroll J.D. (1996). *Tree and other network models for representing proximity data.* In Arabie P., Hubert L.J., De Soete G. (eds), Clustering and Classification, World Scientific, 157 – 197.

[7] Hartmann A.K. (1998). *Are ground states of 3D ±J spin glasses ultrametric?.* Europhysics Letters **44**, 249 – 254.

[8] Křivánek M., Morávek J. (1986). *NP-hard problems in hierarchical-tree clustering.* Acta Informatica **23**, 311 – 323.

[9] Křivánek M., Morávek J. (1984). *On NP-hardness in hierarchical clustering.* In Havránek T., Sidák Z., Novák M. (eds), Compstat 1984: Proceedings in Computational Statistics, 189 – 194, Physica-Verlag, Vienna.

[10] Lerman I.C. (1981). *Classification et analyse ordinale des donneés.* Paris: Dunod, 1981.

[11] Makarenkov V., Leclerc B. (1999). *An algorithm for fitting a tree metric according to a weighted least-squares criterion.* Journal of Classification **16**, 3 – 26.

[12] Murtagh F. (1985). *Multidimensional clustering algorithms.* Physica-Verlag, Würzburg.

[13] Murtagh F. (2004). *On ultrametricity, sparse coding, and computation.* Journal of Classification, submitted, 2004.

[14] Rammal R., Angles d'Auriac J.C., Doucot B. (1985). *On the degree of ultrametricity.* Le Journal de Physique – Lettres **46**, L-945 – L-952.

[15] Rammal R., Toulouse G., Virasoro M.A. (1986). *Ultrametricity for physicists.* Reviews of Modern Physics **58**, 765 – 788.

[16] Rizzi A. (2000). *Ultrametrics and p-adic numbers.* In Gaul W., Opitz O., Schader M. (eds), Data Analysis: Scientific Modeling and Practical Application, Springer-Verlag, 325 – 324.

[17] Treves A. (1997). *On the perceptual structure of face space.* BioSystems **40**, 189 – 196.

[18] Watson S. (2003). *The classification of metrics and multivariate statistical analysis.* Preprint, York University 27 pp.

Address: F. Murtagh, School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK

E-mail: `f.murtagh@qub.ac.uk`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

NONPARAMETRIC REGRESSION WITH FUNCTIONAL DATA FOR POLYMER CLASSIFICATION

Salvador Naya, Ricardo Cao and Ramón Artiaga

Key words: Functional data, nonparametric regression, TGA curve, thermo-gravimetric experiments.

COMPSTAT 2004 section: Simulation.

Abstract: A nonparametric method of functional regression for the classification of different types of materials studied by thermo-gravimetric analysis is proposed in this work. The method is illustrated with two case studies for classifying different materials. Some simulation study shows the performance of the method when the correlation between groups parameters and the error variance change.

1 Introduction

Thermo-gravimetric analysis is one of the most used thermal analysis techniques for polymer characterization. It is also applied to the study of other materials like minerals and metals. It consists in measuring the mass of a sample while it is subjected to a thermal program in a controlled environment. The dynamic mode of operation consists in ramping the temperature at a constant heating rate. The response variable is the sample mass along the experiment, while the explanatory variable is temperature or time. The curves obtained from polymers typically show weigh loss steps at specific temperatures that give information about the polymer degradation such as temperature dependence and kinetics.

The information obtained by thermal gravimetric analysis can be used to this aim. In this work nonparametric regression with functional data was used for the classification of different polymers. The method can be extended to any kind of material that can be analyzed by TGA.

Since the general aim of this paper is concerned with classification of observations in a finite number of classes it can be included in the general setup of pattern recognition (Watanabe, [3]). Classical approaches to this problem includes parametric discriminant analysis. Nevertheless, in this case the data are curves and thus non parametric functional models are more suitable, since they take into account all the information from the sample (Ramsay and Silverman, [2]). The method is illustrated with two examples to classify several PVC and wood samples. Finally, many simulated experiments were used to evaluate the accuracy of the method.

2 Nonparametric classification method

In order to construct our nonparametric classification method, kernel regression was chosen. This method has been proved to work well in many cases (see Ferraty and Vieu, [1]).

The nonparametric Bayes rule was used to classify a future observation in one of the existing groups. This rule minimizes the probability of incorrect classification. It assigns a future observation to the highest probability class. The observed TGA curves, $X_i = X_i(t)$, are a sample of the explanatory variable, while the response sample consists in the observations Y_i of a discrete random variable taking values in the set $\{0, 1, \ldots, G\}$.

Considering a new TGA curve, $x = x(t)$, obtained from a material to classify, the estimator of the posterior probability is given by:

$$\widehat{r}_n(x) = \frac{\sum\limits_{i=1}^{n} Y_i K\left(\frac{\|x-X_i\|}{h_n}\right)}{\sum\limits_{i=1}^{n} K\left(\frac{\|x-X_i\|}{h_n}\right)} \tag{1}$$

Equation (1) is a version of the Nadaraya-Watson estimator reported by Ferraty and Vieu, [1], where K is the Epanechnikov kernel, i.e. $K(x) = \frac{3}{4}\left(1-x^2\right) I_{(|x|\leq 1)}$, $\|\bullet\|$ denotes the L_1 norm and h is the bandwidth or smoothing parameter.

Using the estimator obtained in equation (1), the classification rule assigns the observed curve x to the group of index:

$$d_h(x) = \arg\max_{0\leq j\leq G} \left\{\widehat{r}_h^{(j)}(x)\widehat{p}_j\right\}$$

where $\widehat{r}_h^{(j)}$ denotes the estimation of the probability of the sample belonging to the j class.

The smoothing parameter h will be chosen as some estimator of the value that minimizes the probability of missclassifying a future observation. This bandwidth will be defined as the h that minimizes the following cross-validation function:

$$CV(h) = n^{-1} \sum_{i=1}^{n} 1_{\left\{Y_i \neq d_h^{-i}(X_i)\right\}}, \tag{2}$$

where d_h^{-i} is the classification rule built up without the i-th observation. Finally, given a new sample and its TGA trace, denoted as x, the distances from this trace to the others in the sample will be calculated and $\widehat{r}_h^{(j)}$ will be estimated for each class of material for $j \in \{0, 1, ..., G\}$. The new sample will be assigned to the class k that maximizes $\widehat{r}_h^{(j)}(x)$.

2.1 Choice of the seminorm

An important issue related to this method is the choice of the distance to be used in order to obtain small differences among curves within a group and large distances for TGA curves of different groups. This is measured in the estimator $\widehat{r}_n(x)$ by using some seminorm $\|\cdot\|$.

Several seminorms have been considered in this setup: the L_1 norm, the L_2 norm and the seminorms defined as the L_2 norm between υ-th order derivatives (with $\upsilon \leq k$, for some $k \in \mathbb{N}$). The last one constitute the following family

$$N_\upsilon = \{\|\cdot\|_\upsilon ; \upsilon = 0, 1, 2, ..., k\} \text{ with } \|f\|_\upsilon = \left(\int \left(f^{(\upsilon)}(t)\right)^2 dt\right)^{1/2}.$$

2.2 Computational issues

In the following we present several topics that are important in order to implement the classification rule presented above.

2.2.1 Distances between curves A first concern that appears when computing the distance between any pair of curves in the sample is that they need not to be evaluated in the same grid. As a consequence,the first step consists in approximating the value of all the curves in the same grid of points. In the horizontal axis (corresponding to the variable time) a total number of $m = 3000$ points have been chosen. The vertical axis is rescaled by considering the percentage of weight instead of the absolute weight. On the other hand, since the final weight was not the same for all the experiments, the time period considered for all the curves was cuted at the point where 60% of the weight has been lost. This means that the final parts of the spectrum have been eliminated.

For every pair of observations, f and g, in the sample we consider some numerical approximation of the L_1 distance between them:

$$\|f - g\| = \int |(f(t) - g(t))| \, dt \simeq \sum_{i=2}^{m} |f(t_i) - g(t_i)| \, (t_i - t_{i-1}),$$

where $t_1 < t_2 < \ldots < t_m$ is the common grid for all the curves in the sample. As explained above, the differences $|f(t_i) - g(t_i)|$ have been approximated by interpolating the sample curves in every point of the common grid.

2.2.2 Cross-validation bandwidth Once computed the distance between curves, $\|X_i - X_j\|$,for every $i \neq j$, the values $\widehat{r}_h^{(j)}$ are computed in order to obtain the cross-validation function. This function has been evaluated in a finite grid of values of h. More specifically, the values for h have been chosen in the interval $\left[\min_{i\neq j}\|X_i - X_j\|, 3\max\|X_i - X_j\|\right]$. The h-grid has been chosen multiplicative in such a way that every value is 1.5 times the previous one. The final h to be considered is that one that minimizes $CV(h)$ along these points.

It is important to note that the distance between two the TGA curves, $\|X_i - X_j\|$ for $i \neq j$, have been computed only once and then stored in a matrix in order to avoid repetitive calculations when changing from one value of h to another.

2.2.3 Classification of a future observation Given an observed TGA curve, x, of some material to be classified, its distance to all the curves in the sample will be computed: $\|x - X_j\|$, for $j = 1, 2, ..., n$. Then the values $\widehat{r}_h^{(k)}(x)$ have to be computed for $k \in \{0, 1, ..., G\}$ and the index k that maximizes $\widehat{r}_h^{(k)}(x)$ corresponds to the class where this observation will be assigned.

3 Practical application

In a first case study a total number of 19 samples of 7 types of wood have been considered. For each of these, its TGA spectrum has been obtained. The cross-validation function has been computed. Its minimal value is 0.64, which gives an estimation of a rather low probability of correct classification. This is probably caused by the limited sample size (relative to the number of classes) and the already known difficulty in classifying different types of wood by thermal analysis. This difficulty is also present in many other classification techniques, as those based in spectrometrics.

The second study is concerned with PVC classification. A number of 16 samples of PVC of two different groups (rigid and flexible PVC) have been considered. In this case, the estimated probability of correct classification (using the minimal value of the cross-validation function) has been much larger, namely 0.94.

4 Simulation study

A simulation study was performed in order to check the performance of the method. Three kinds of wood were chosen, since these materials are very much alike in composition and thermal behaviour. It is not easy to classify this kind of materials only by TGA experiments. From actual experiments of the three samples, two sets of experiments were simulated by a logistic mixture model. The simulation was performed for each of the three groups, using the function $\varphi^{(r)}(x)$:

$$\varphi^{(r)}(x) = \sum_{j=1}^{k_{(r)}} w_j^{(r)} f(a_j^{(r)} + b_j^{(r)} x), \quad f(x) = \frac{e^x}{1+e^x}$$

The parameters for the model

$$\left(\left(w_1^{(r)}, a_1^{(r)}, b_1^{(r)} \right), \left(w_2^{(r)}, a_2^{(r)}, b_2^{(r)} \right), ..., \left(w_k^{(r)}, a_k^{(r)}, b_k^{(r)} \right) \right),$$

were simulated following a $3k_{(r)}$-dimensional normal distribution. Two different situations were considered: independent and dependent parameters. For every type of wood the distribution of every vector of parameters need not to be the same.

For the r-th group, we consider

$$(x_1, y_1, z_1, ..., x_k, y_k, z_k) \overset{d}{=} N_{3k}\left(\mu_{(r)}, \Sigma_{(r)}\right),$$

which will be simulated again whenever $z_j < 0$, for some j. Then, the parameters of the model are defined:

$$a_j^{(r)} = x_j, b_j^{(r)} = y_j, w_j^{(r)} = \frac{z_j}{\sum_{l=1}^{k} z_l} \mathit{100},\ j = \mathit{1}, \mathit{2}, ..., k$$

Two different situations were considered for the covariance matrix of the model:

$$\varSigma_{(r)} = \begin{pmatrix} \Sigma_{(r)_{(1)}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_{(r)_{(2)}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_{(r)_{(k)}} \end{pmatrix}.$$

1.- Independence between the parameters*:*

$$\Sigma_{(r)_{(j)}} = \begin{pmatrix} \sigma_{1,j}^2 & 0 & 0 \\ 0 & \sigma_{2,j}^2 & 0 \\ 0 & 0 & \sigma_{3,j}^2 \end{pmatrix}.$$

2.- Dependence between the parameters with correlation coefficient ρ:

$$\Sigma_{(r)_{(j)}} = \begin{pmatrix} \sigma_{1,j}^2 & \rho_j \sigma_{1,j}\sigma_{2,j} & \rho_j \sigma_{1,j}\sigma_{3,j} \\ \rho_j \sigma_{1,j}\sigma_{2,j} & \sigma_{2,j}^2 & \rho_j \sigma_{2,j}\sigma_{3,j} \\ \rho_j \sigma_{1,j}\sigma_{3,j} & \rho_j \sigma_{2,j}\sigma_{3,j} & \sigma_{3,j}^2 \end{pmatrix}.$$

The values used for the means, standard deviations and correlations in the simulation study were fitted according to some existing data. In the following tables, these parameters are presented for the three groups (cypress, eucalyptus and oak).

	Mean	Standard deviation $\left(\sigma_{0,ij}^{(1)}\right)$
$w_1^{(1)}$	11.9067	2.7464
$a_1^{(1)}$	5.1067	0.4002
$b_1^{(1)}$	−0.0147	0.0046
$w_2^{(1)}$	38.402	2.5176
$a_2^{(1)}$	101.51	3.4954
$b_2^{(1)}$	−0.0540	0.0053
$w_3^{(1)}$	49.6907	5.2152
$a_3^{(1)}$	17.573	5.2036
$b_3^{(1)}$	−0.0093	0.0051

Table 1: Means and standard deviations for the cypress parameters.

	Mean	Standard deviation $\left(\sigma_{0,ij}^{(2)}\right)$
$w_1^{(2)}$	13.578	0.7318
$a_1^{(2)}$	5.6633	1.5211
$b_1^{(2)}$	−0.0160	0.0041
$w_2^{(2)}$	17.9827	1.4784
$a_2^{(2)}$	13.3733	0.5601
$b_2^{(2)}$	−0.0083	0.0015
$w_3^{(2)}$	33.2240	0.9643
$a_3^{(2)}$	103.0033	1.8083
$b_3^{(2)}$	−0.0537	0.0078
$w_4^{(2)}$	35.2153	1.2145
$a_4^{(2)}$	14.8233	1.5309
$b_4^{(2)}$	−0.0080	0.0020

Table 2: Means and standard deviations for the eucalyptus parameters.

4.1 Simulation results

Some sample of 90 TGA curves have been simulated according to the probabilities 1/3, 1/3, 1/3 for every type of wood. The cross-validation bandwidth, h_{CV}, was found, whose minimal value is included in Table 7. Then, 1000 new samples were simulated with the same probabilities as before. For every of these samples the estimated Bayes rule is computed and the observation is assigned to one group. Comparing it with the true group, the probability of correct classification has been estimated either in the dependent case (PCCDep) or in the independent case (PCCIndep), that are also included in Table 7.

	Mean	Standard deviation $\left(\sigma_{0,ij}^{(3)}\right)$
$w_1^{(3)}$	6.0497	0.8251
$a_1^{(3)}$	7.4507	2.0019
$b_1^{(3)}$	−0.0163	0.0072
$w_2^{(3)}$	58.5593	2.4852
$a_2^{(3)}$	77.6833	10.2494
$b_2^{(3)}$	−0.0667	0.0057
$w_3^{(3)}$	35.3913	3.1471
$a_3^{(3)}$	11.3367	1.2536
$b_3^{(3)}$	−0.0043	0.0031

Table 3: Means and standard deviations for the oak parameters.

Component j	1	2	3
$\rho_{0,j}^{(1)}$	0.653	−0.627	−0.600

Table 4: Correlation coefficients for the cypress parameters.

Component j	1	2	3	4
$\rho_{0,j}^{(2)}$	−0.569	0.683	−0.517	−0.663

Table 5: Correlation coefficients for the eucalyptus parameters.

Component j	1	2	3
$\rho_{0,j}^{(3)}$	0.143	−0.470	−0.329

Table 6: Correlation coefficients for the oak parameters.

Table 7 contains the simulation results for a range of standard deviations obtained from Tables 1–3: $\sigma_{ij}^{(r)} = \lambda\sigma_{0,ij}^{(r)}$, for some values of λ. On the other hand, the sensitivity of the proposed classification method to the amount of dependence has also been explored. Table 8 exhibits the probability of correct classification for a range of possible correlations: $\rho_j^{(3)} = \delta\rho_{0,j}^{(3)}$, for some values of δ.

As expected, the percentage of correct classification increases as the variance decreases up to levels of 92.4%−95.4% with factors of 1/8 of the original standard deviations. When increasing the standard deviation up to 4 times, the probability of correct classification decreases to about 64.8% − 72.2%. Generally speaking, the results obtained under dependence are slightly better than those for independence (see Table 8).

λ	PCCIndep	PCCDep	$CV(h_{CV})$ Indep	$CV(h_{CV})$ Dep
4	64.8	72.2	0.19	0.18
2	73.6	75.9	0.16	0.16
1	77.8	82	0.15	0.14
1/2	75.2	87	0.07	0.07
1/4	89.5	88.5	0.05	0.03
1/8	92.4	95.4	0.03	0.01
1/10	92.4	97.8	0.01	0.01
1/20	94.3	99.8	0.01	0
1/40	98.9	100	0	0
1/50	99.8	100	0	0

Table 7: Percentages of correct classification (PCC) under dependence (PCCDep) and independence (PCCIndep).

δ	PCC	$CV(h_{CV})$
1.4	78.6	0.12
1	80.7	0.15
1/2	79.9	0.16
1/4	78.2	0.16

Table 8: Influence of the correlation coefficient in the probability of correct classification.

References

[1] Ferraty F., Vieu P. (2002). *The functional nonparametric model and application to spectrometric data.* Computational Statistics **17** (4), 545 – 564.

[2] Ramsay J.O., Silverman B.W. (1997). *Functional data analysis.* Springer.

[3] Watanabe S. (1985). *Pattern recognition.* Human and Mechanical. Wiley, New York.

Acknowledgement: This work was partially supported by Grants BFM2002-00265 (European FEDER support included), PGIDIT03PXIC10505PN and PGIDT01PXI10505PR.

Address: S. Naya, R. Cao and R. Artiaga, Departamento de Matemáticas and Departamento de Ingeniería Industrial II. Universidad de A Coruña, A Coruña 15071, Spain

E-mail: `salva@udc.es`

ESTIMATING THE RISK-ADJUSTED PREMIUM FOR THE LARGEST CLAIMS REINSURANCE COVERS

Abdelhakim Necir and Kamal Boukhetala

Key words: Heavy tails, extreme value theory, premium calculation, large claim, reinsurance treaty.

COMPSTAT 2004 section: Nonparametrical statistics.

Abstract: We consider Wang's premium principle to estimate the net premium for actuarie's largest claims with heavy-tailed distribution. Under the second order regular variation, asymptotic normality of such estimator is given. Our results provide an asymptotic confidence interval for an adequate net premium for tariffing risks that overshoot a high threshold.

1 Motivation

In insurance business a few large claims hitting a portfolio usually represent the greatest part of the indemnities paid by the company. These extreme events (or risks) are, therefore, the prime interest for actuaries. An example of such problem has been discussed by Cebrián, Denuit and Lambert [2] for the medical insurance large claims database. To determinate an adequate price or premium for these risks one use an appropriate pricing principle. Clearly, premiums cannot be too low because this would result in unacceptably large losses for the insures. On the other hand, premium cannot to be too high either because of competition between insures. There are several variants to define premium principles (see for instance, Bühlman [1], Kaas, van Heerwaarden and Goovaerts [10] and Wang [13], [14], [15]). Wang [13] proposes a premium principle based on a proportional transformation of the hazard function. This premium principle corresponds to the certainty equivalent of the dual theory of expected utility developed by Yaari [17]. These approaches to pricing insurance contracts treat insurance losses as positive random variables and produce premiums that are higher than the expected value of the insurance loss. The so-called *risk-adjusted principle* (see for instance, Christofides [3] or Rolski et al. [11, page 82]) belongs to this class of premiums that given by Wang [13].

In this paper we consider this approach of premium calculation principle to derive an estimator of the risk-adjusted premium for the largest claims. Moreover, the asymptotic normality of such estimator is given. This result provides an asymptotic confidence interval for the net premium of risks that overshoot a high threshold, and reopen the discussion on the usefulness of including the largest claims in the decision making procedure. Our arguments, in this paper, are based on the extreme value theory and the second order regular variation of heavy-tailed distributions.

2 The sample risk-adjusted premium of loss

Let $X,\ X_1,\ X_2,\ldots$ be a sequence of independent and identically distributed risks with common distribution function $F(x) = P(X \leq x),\ x \in \mathbb{R}$. Wang [13] proposes to compute the *risk-adjusted premium* of a risk X as a "distorted" expectation of X :

$$\Pi(X) := \int_0^\infty (1 - F(s))^{1/p}\, ds,$$

for some $p \geq 1$. The parameter $p \geq 1$ is called *the distortion coefficient.* The corresponding *risk-adjusted premium of loss* being defined, for a high threshold $u > 0$, as follows:

$$\Pi_u(X) := \int_u^\infty (1 - F(s))^{1/p}\, ds.$$

Notice that for a suitable economic interest, the threshold u must be so large and depends of sample size $n \geq 1$ of claims income. For this reason we will suppose that $u = u_n,\ n \geq 1$. This leads to rewrite $\Pi_u(X)$ into

$$\Pi_{u_n}(X) = \int_{u_n}^\infty (1 - F(s))^{1/p}\, ds.$$

Let $k = k_n$ be a sequence of integers satisfying $1 \leq k \leq n,\ k \to \infty$ and $k/n \to 0$ such that $u_n := Q(1 - k/n)$, where

$$Q(s) := \inf\{t \in \mathbb{R},\ F(t) \geq s\},\ 0 \leq s < 1,$$

is the quantile function or the generalized inverse of F. This follows that

$$\Pi_{u_n}(X) = \int_{Q(1-k/n)}^\infty (1 - F(s))^{1/p}\, ds.$$

For each $n \geq 1$, let $X_{1,n} \leq \cdots \leq X_{n,n}$ be the order statistics based on $X_1,\ldots,X_n$. By replacing $Q(1-k/n)$ and $F(\cdot)$ by their suitable estimators, we derive an estimator of *the risk-adjusted premium of loss* pertaining to the sample $X_1,\ldots,X_n$, that is

$$\widehat{\Pi}_{u_n} = \int_{X_{n-k,n}}^\infty (1 - F_n(s))^{1/p}\, ds,$$

where $F_n(x) = n^{-1}\#\{X_i \leq x : 1 \leq i \leq n\}$ for $x \in \mathbb{R}$, with $\#\Omega$ denoting cardinality of Ω, is the empirical distribution function pertaining to the sample $X_1,\ldots,X_n$. The corresponding empirical quantile function is defined as

$$\begin{aligned} Q_n(s) &= \inf\{t \in \mathbb{R},\ F_n(t) \geq s\},\ 0 < s \leq 1 \\ &= X_{i,n}, \quad (i-1)/n < s \leq i/n,\ i = 1,\ldots,n, \end{aligned}$$

with $Q_n(0) = X_{1,n}$. Observe now that

$$\widehat{\Pi}_{u_n} = -\int_0^{k/n} s^{1/p} dQ_n\left(1-s\right).$$

By integration by parts we get

$$\begin{aligned}
\widehat{\Pi}_{u_n} &= p^{-1}\int_0^{k/n} s^{1/p-1} Q_n\left(1-s\right) ds - \left(k/n\right)^{1/p} X_{n-k,n} \\
&= p^{-1}\sum_{i=1}^{k}\left(\int_{(i-1)/n}^{i/n} s^{1/p-1} ds\right) X_{n-i+1,n} - \left(k/n\right)^{1/p} X_{n-k,n} \\
&= \sum_{i=1}^{k}\left\{\left(\frac{i}{n}\right)^{1/p} - \left(\frac{i-1}{n}\right)^{1/p}\right\} X_{n-i+1,n} - \left(k/n\right)^{1/p} X_{n-k,n}. \\
&= \sum_{i=1}^{k}\left(\frac{i}{n}\right)^{1/p}\left\{X_{n-i+1,n} - X_{n-i,n}\right\}, \quad p \geq 1.
\end{aligned}$$

Note that for $p = 1$, the statistics $n\widehat{\Pi}_{u_n}$ corresponds to the large claims reinsurance treaty ECOMOR (excédent du coût moyen relatif) introduced by Thépaut [12] (see also, Rolski et al. [11, page 17]).

Since we interest to rare events, we must then suppose that the tail of distribution function F is of Pareto-type. On the other word, we must assume that F is *heavy-tailed* or *regularly varying at infinity*. Namely, there exists a positive constants $\gamma > 0$, such that

$$\lim_{t\to\infty}\frac{1-F\left(tx\right)}{1-F\left(t\right)} = x^{-1/\gamma}\ , \text{ for any } x > 0. \tag{1}$$

We also have a definition, that equivalent to (1), in term of the quantile function, that is

$$\lim_{t\to 0}\frac{Q\left(1-tx\right)}{Q\left(1-t\right)} = x^{-\gamma}, \text{ for any } x > 0. \tag{2}$$

(see, e.g., [7]). The real parameter γ is called *the tail index* of F. In the literature there are several estimators for $\gamma > 0$ among them is the well known Hill estimator (see [9]).

Further, we assume that the distribution tail $1 - F$ is second order regularly varying with first parameter $-1/\gamma$ and second parameter $\rho \leq 0$, that is, there exists a function $A\left(t\right) \to 0$ as $t \to 0$ which ultimately has constant sign such that the following refinement of (2) holds:

$$\lim_{t\to 0}\left(A\left(t\right)\right)^{-1}\left\{\frac{Q\left(1-tx\right)}{Q\left(1-t\right)} - x^{-\gamma}\right\} = x^{-\gamma}\frac{x^{\rho\gamma}-1}{\rho\gamma}, \text{ for any } x > 0. \tag{3}$$

If $\rho = 0$, interpret $\dfrac{x^{\rho\gamma}-1}{\rho\gamma}$ as $\log x$ (see, e.g., [6]).

In the sequel, $\overset{\mathcal{D}}{\rightarrow}$ and $\overset{\mathcal{D}}{=}$ stand, respectively, for convergence and equality in distribution. For further use, let $\mathcal{N}\left(0,\eta^2\right)$ denote the normal distribution with mean 0 and variance η^2.

3 Main results

Our main result is the following theorem which gives asymptotic normality of $\widehat{\Pi}_{u_n}$.

Theorem. *Assume that* (3) *holds and* $s^{1/p}Q(1-s) \rightarrow 0$ *as* $s \rightarrow 0$ *for some* $\gamma - 1/2 < 1/p \leq 1$. *Let* $k = k_n$ *is such that* $k \rightarrow \infty$, $k/n \rightarrow 0$ *and* $(k/n)^{1/2} A(k/n) \rightarrow 0$ *as* $n \rightarrow \infty$. *Then,*

$$\frac{(k/n)^{-1/p}\, k^{1/2}}{Q(1-k/n)} \left\{\widehat{\Pi}_{u_n} - \Pi_{u_n}\right\} \overset{\mathcal{D}}{\rightarrow} \mathcal{N}\left(0,\sigma^2(p,\gamma)\right), \ \textit{as } n \rightarrow \infty,$$

where

$$\sigma^2(p,\gamma) := \gamma^2 \begin{cases} p^{-2}\frac{1}{\lambda(\lambda+1)} - p^{-1}\frac{2}{\lambda+1} + 1, & \text{for } 1/p \neq \gamma \\ 2p^{-2} - 2p^{-1} + 1, & \text{for } 1/p = \gamma. \end{cases}$$

with $\lambda := 1/p - \gamma$.

3.1 Conclusion

In view of the above result we may readily derive an asymptotic confidence interval for the risk adjusted premium Π_{u_n}. The computation of bounds of this confidence interval depend upon the tail index γ of the distribution function F and the sample fraction k. To estimate γ we propose the Hill estimator [9] and for the optimal choice of k we suggest the method of the bootstrap given recently by Danielsson et al. [5].

4 Proof of the Theorem

Let $\xi_1, \xi_2, \ldots,$ be a sequence of independent uniform $(0,1)$ random variables. For each integer $n \geq 1$, the empirical quantile function is defined by

$$\mathbb{V}_n(t) = \xi_{i,n}\ , \quad \text{for } (i-1)/n < t \leq i/n,\ i = 1, \ldots, n,$$

with $\mathbb{V}_n(0) = \xi_{1,n}$, where $\xi_{1,n} \leq \cdots \leq \xi_{n,n}$ denote the order statistics based on $\xi_1, \ldots, \xi_n$. Assume without loss of generality, that the random variables $(X_n)_{n\geq 1}$ are defined on a probability space $(\Omega, \mathcal{A}, P)$ which carries the sequence $(\xi_n)_{n\geq 1}$ is such a way that $X_n = Q(\xi_n)$, for $n = 1, 2, \ldots$, and, consequently, $X_{i,n} = Q(\xi_{i,n})$ for all $1 \leq i \leq n$ and $n \geq 1$. Observe that for each integer $n \geq 1$, we have $\mathbb{V}_n(1 - i/n) = \xi_{n-i+1,n}$, and $Q(\mathbb{V}_n(1 - i/n)) = X_{n-i+1,n}$, $i = 1, 2, \ldots$ This follows that

$$\widehat{\Pi}_{u_n} = (k/n)^{1/p} \left\{ \int_0^1 s^{1/p-1} Q_n (1 - ks/n) \, ds - X_{n-k,n} \right\}.$$

On the other hand, since $s^{1/p} Q(1-s) \to 0$ as $s \to 0$, it is easily verified that, by integration by parts we get

$$\Pi_{u_n} = p^{-1} (k/n)^{1/p} \int_0^1 s^{1/p-1} Q (1 - ks/n) \, ds - (k/n)^{1/p} Q (1 - k/n) .$$

Observe now that

$$\frac{(k/n)^{-1/p} k^{1/2}}{Q(1-k/n)} \left\{ \widehat{\Pi}_{u_n} - \Pi_{u_n} \right\} = T_n^* - S_n^*,$$

where

$$T_n^* := p^{-1} k^{1/2} \int_0^1 s^{1/p-1} \frac{Q_n (1 - ks/n) - Q (1 - ks/n)}{Q (1 - k/n)} ds,$$

and

$$S_n^* := k^{1/2} \frac{X_{n-k,n} - Q (1 - k/n)}{Q (1 - k/n)}.$$

We have

$$\begin{aligned} T_n^* &= p^{-1} k^{1/2} \int_0^1 s^{1/p-1} \frac{Q (1 - ks/n)}{Q (1 - k/n)} \left[\frac{Q_n (1 - ks/n)}{Q (1 - ks/n)} - 1 \right] ds \\ &= \widetilde{T}_n + T_n, \end{aligned}$$

where

$$\widetilde{T}_n := p^{-1} k^{1/2} \int_0^{1/k} s^{1/p-1} \frac{Q (1 - ks/n)}{Q (1 - k/n)} \left[\frac{Q_n (1 - ks/n)}{Q (1 - ks/n)} - 1 \right] ds,$$

and

$$T_n := p^{-1} k^{1/2} \int_{1/k}^1 s^{1/p-1} \frac{Q (1 - ks/n)}{Q (1 - k/n)} \left[\frac{Q_n (1 - ks/n)}{Q (1 - ks/n)} - 1 \right] ds.$$

Further, for the term T_n, we may write

$$T_n := T_{n1} + T_{n2} + T_{n3},$$

where

$$\begin{aligned} T_{n1} &:= p^{-1} k^{1/2} \int_{1/k_n}^1 s^{1/p-1} \frac{Q (1 - ks/n)}{Q (1 - k/n)} \\ &\quad \times \left[\frac{Q_n (1 - ks/n)}{Q (1 - ks/n)} - \left(\frac{1 - V_n (1 - ks/n)}{ks/n} \right)^{-\gamma} \right] ds, \end{aligned}$$

$$\begin{aligned} T_{2n} &=: p^{-1} \int_{1/k_n}^{1} s^{1/p-1} \left[\frac{Q(1-ks/n)}{Q(1-k/n)} - s^{-\gamma} \right] \\ &\quad \times \left[\left(\frac{1-V_n(1-ks/n)}{ks/n} \right)^{-\gamma} - 1 \right] ds, \end{aligned}$$

and

$$T_{3n} := p^{-1} k^{1/2} \int_{1/k_n}^{1} s^{1/p-1} s^{-\gamma} \left[\left(\frac{1-V_n(1-ks/n)}{ks/n} \right)^{-\gamma} - 1 \right] ds.$$

For term T_{3n} we have

$$T_{3n} := p^{-1} k^{1/2} (k/n)^{\gamma} \int_{1/k_n}^{1} s^{1/p-1} \left\{ (1-V_n(1-ks/n))^{-\gamma} - (sk/n)^{-\gamma} \right\} ds.$$

Observe that T_{3n} may be rewritten into

$$T_{3n} = p^{-1} (k/n)^{\gamma-1/p} k^{1/2} \int_{1/n}^{k/n} s^{1/p-1} \left\{ (1-V_n(1-s))^{-\gamma} - s^{-\gamma} \right\} ds.$$

Note that $1-V_n(1-s) \overset{D}{=} V_n(s)$. Making use the mean value theorem, we get that

$$\begin{aligned} T_{3n} &\overset{D}{=} -p^{-1} \gamma (k/n)^{\gamma-1/p} k^{1/2} \int_{1/n}^{k/n} s^{1/p-1} (s+\zeta_{s,n})^{-\gamma-1} \\ &\quad \times \{V_n(s) - s\} ds, \end{aligned}$$

with $|\zeta_{s,n}| \le |V_n(s) - s|$, $1/n \le s \le k/n$. Let now $\theta_n := (k/n) k^{\frac{-\delta-1/2}{1/p-\gamma}}$, $0 < \delta < 1/2$, be a sequence of positive real number that satisfies $1/n < \theta_n < k/n$. Then T_{3n} may be decomposed into tow parts

$$\begin{aligned} T_{3n} &\overset{D}{=} -p^{-1} \gamma (k/n)^{\gamma-1/p} k^{1/2} \int_{1/n}^{\theta_n} s^{1/p-1} (s+\zeta_{s,n})^{-\gamma-1} \{V_n(s) - s\} ds \\ &-p^{-1} \gamma (k/n)^{\gamma-1/p} k^{1/2} \int_{\theta_n}^{k/n} s^{1/p-1} (s+\zeta_{s,n})^{-\gamma-1} \{V_n(s) - s\} ds. \end{aligned}$$

Following now the same arguments as used in the proof of Lemma 4.1 of Greoneboom et al. [8] and by using Theorem 2.1 of Csörgő, Csörgő, Horváth and Mason [4], we arrive to, for all large n

$$T_{3n} \overset{\mathcal{D}}{=} (1+o_p(1)) p^{-1} \gamma \int_0^1 s^{1/p-\gamma} W_n(s) ds.$$

Likewise we show that, for all large n

$$S_n^* \overset{\mathcal{D}}{=} (1+o_p(1)) \gamma W_n(1).$$

where $W_n(\cdot)$ is a sequence of standard Brownian motion. On the other hand, by using the second order regular variation (3) , Potter's inequalities (see for instance assertion (3.1) of Danielsson et al. [5] and limit theorems of Wellner [16] together, we omit details, we show that, for all large n, we have

$$\widetilde{T}_n = T_{n1} = T_{n2} = o_p(1).$$

Therefore, we get

$$\frac{(k/n)^{-1/p}k^{1/2}}{Q(1-k/n)}\left\{\widehat{\Pi}_{u_n} - \Pi_{u_n}\right\} \overset{\mathcal{D}}{=} p^{-1}\gamma\left\{\int_0^1 s^{1/p-\gamma-1}W_n(s)\,ds - W_n(1)\right\}$$
$$+o_p(1), \text{ as } n\to\infty.$$

This implies, since $W_n(\cdot)$ is a sequence of standard Brownian motion, that

$$\frac{(k/n)^{-1/p}k^{1/2}}{Q(1-k/n)}\left\{\widehat{\Pi}_{u_n} - \Pi_{u_n}\right\} \overset{\mathcal{D}}{\to} \mathcal{N}\left(0,\sigma^2(p,\gamma)\right), \text{ as } n\to\infty,$$

where

$$\sigma^2(p,\gamma) := \gamma^2 C^2(p,\gamma),$$

with

$$C^2(p,\gamma) := var\left\{p^{-1}\int_0^1 s^{1/p-\gamma-1}W_n(s)\,ds - W_n(1)\right\}.$$

Let $\lambda := 1/p-\gamma$. Since $EW_n(u)W_n(v) = \min(u,v)$, then

$$\begin{aligned}
C^2(p,\gamma) &= p^{-2}\int_0^1\int_0^1 t^{1/p-\gamma-1}s^{1/p-\gamma-1}E\{W_n(t)\,EW_n(s)\}\,dsdt + \\
&\quad +E\{W_n(1)\}^2 - 2p^{-1}\int_0^1 s^{1/p-\gamma-1}E\{W_n(s)\,W_n(1)\}\,ds \\
&= p^{-2}\int_0^1\int_0^1 t^{1/p-\gamma-1}s^{1/p-\gamma-1}\min(s,t)\,dsdt + 1 - \\
&\quad -2p^{-1}\int_0^1 s^{1/p-\gamma-1}sds \\
&= \begin{cases} p^{-2}\frac{1}{\lambda(\lambda+1)} - p^{-1}\frac{2}{\lambda+1} + 1, & \text{for } 1/p \neq \gamma \\ 2p^{-2} - 2p^{-1} + 1 & \text{, for } 1/p = \gamma. \end{cases}
\end{aligned}$$

This completes the proof of the Theorem. □

References

[1] Bühlmann H.(1980). *An economic premium principle.* ASTIN Bulletin **11**, 52 – 60.

[2] Cebrián, A. C., Denuit, M. and Lambert, P. (2003). *Generalized Pareto Fit to the Society of Actuaries' large Claims Database.* North American Actuarial Journal, **7**, 18 – 36.

[3] Christophides, S. (1998). *Principle for risk in finantial transactions.* Proceeding of the GISG/ASTIN Joint Meeting in Glasgow, **2**, 63 – 109.

[4] Csörgő, M., Csörgő, S., Horváth, L. and Mason, D. (1986). *Weighted empirical and quantile processes.* Ann. Probab., **14**, 31 – 85.

[5] Danielsson, J., de Hann, L., Peng, L. and de Vries, C. G. (2001). *Using a boostrap Method to choose the sample fraction in tail estimation. J. Multivariate Annal., **76**, 226 – 248.*

[6] de Haan, L., and Stadtmüler, U. (1996). *Generalized regular variation of second order.* J. Australian Math. Soc., (serie A), **61**, 381 – 395.

[7] Embrechts, P., Klüppelberg, C. and Mikosch,T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

[8] Groeneboom, P., Lopuhaä, H. P. and de Wolf, P.P. (2003). *Kernel estimators for the extreme value index.* Ann. Statist., **31**, 1956 – 1995.

[9] Hill, B. M. (1975). *A simple approach to inference about the tail of a distribution.* Ann. Statist., **3**, 1136 – 1174.

[10] Kaas, R., van Heerwaarden, A. E. and Goovaerts, M. J. (1994). *Ordering of Actuarial Risks.* Caire Education Serie 1, Brussels.

[11] Rolski, T., Schimidli, H., Schmidt, V. and Teugels, J.L. (1999). *Stochastic Processes for Insurance and Finance.* John Wiley & Sons, Chichester.

[12] Tépaut, A. (1950). *Une nouvelle forme de réassurance: le traité d'excédent du coût moyen relatif (ECOMOR).* Bull. Trimestriel Inst. Actuair. Français, **49**, 273 – 343.

[13] Wang, S.S. (1996). *Premium Calculation by Transforming the Layer Premium Density*, ASTIN Bulletin, **26**, 71 – 92.

[14] Wang, S.S. (1998). *An Actuarial index of right-tail risk.* North American Actuarial Journal, **2**, (2), 88 – 101.

[15] Wang, S.S. (2000). *A Note on Christofides conjecture regarding Wang's premium principle.* North ASTIN bulletin, **30**, 13 – 17.

[16] Wellner, J. A. (1978). *Limit theorems for the ratio of the empirical distribution function to the true distribution function.* Z. Wahrsch. verw. Gebiete, **45**, 73 – 88.

[17] Yaari, M. E., (1987). *The Dual Theory of Choice Under Risk.* Econometrica, **55**, 95 – 115.

Address: A. Necir, Laboratory of Applied Mathematics, University of Biskra, PO box 145 RP, 07000 Biksra, Algeria

K. Boukhetala, Department of Probability and Statistics, Faculty of Mathematics, PO box 32, El Alia, Bab-Ezzouar, USTHB, Algeria

E-mail: necirabdelhakim@yahoo.fr, kboukhetala@usthb.dz

MIXTURE OF GLMS AND THE TRIMMED LIKELIHOOD METHODOLOGY

N. Neykov, P. Filzmoser, R. Dimova and P. Neytchev

Key words: Trimmed likelihood estimator, robust estimator, mixture of GLMs, FlexMix, R.

COMPSTAT 2004 section: Robustness.

Abstract: The Maximum Likelihood Estimator (MLE) has been widely used to estimate the unknown parameters in the finite mixture of Generalized Linear Models (GLMs). However, the MLE can be very sensitive to outliers in the data. In this paper we consider an approach based on the Trimmed Likelihood Estimator (TLE) to estimate mixtures of GLMs in a robust way. The superiority of this approach in comparison with the MLE is illustrated through a simulation study.

1 Introduction

Finite mixture of distributions have been widely used to model a wide range of heterogeneous data, e.g., [15] or [29]. In most applications the mixture model parameters are estimated by the MLE. It is well known, however, that the MLE can be very sensitive to outliers in the data. In fact, even a single outlier can ruin completely the MLE. To overcome this, robust parametric alternatives of the MLE have been developed, e.g., [6], [7], [8], [11], [4], [14], [19], [21], [28]. Few of these alternatives have been used in fitting finite mixtures of GLMs. For instance mixtures of Poissons and normals based on the weighted MLE technique are discussed in [13]. To reduce the outliers influence on parameter estimates of a mixture of two normals, the Median of the negative Likelihood Estimator (MedLE) is recommended in [24], the Breakdown Point (BP) properties of which were studied in [25] and [26].

An indirect technique for the detection of interesting multiple structures in data by means of redescending M-estimators is suggested in [16] tracing all possible solutions to the M-estimating equations. This approach is extended further in [10]. Another way of doing robust estimation in mixtures of location-scale models has been the replacement of the classical multivariate location and scatter with their robust counterparts based on M and MCD estimates, as in [1], [5] and [20]. Mixtures of t-distributions are recommended in [15], but this approach is not resistant against leverage points. Thus robustness has been adapted to meet some problems with outliers in clustering and the clusterwise regression, a particular case of mixtures of GLMs. Generally speaking, robust fitting of mixtures has not been well developed yet.

Thus, after many years of parallel development of cluster analysis, outlier detection and robust techniques, the need for a synthesis between all of them has become apparent. It was demonstrated in [9] and [20] that such a synthesis can be a flexible and powerful tool for an effective analysis of heterogeneous data. So the aim of this paper is to make a step toward the achievement of this goal by offering a unified approach based on the trimmed likelihood methodology for fitting finite mixtures of distributions. The superiority of this approach in comparison with the MLE is illustrated through a simulation study in the mixtures framework of GLMs.

2 Trimmed likelihood methodology

The Weighted Trimmed Likelihood estimator is defined in [7] and [27] as

$$\mathrm{WTL}_k := \arg\min_{\theta\in\Theta^p} \sum_{i=1}^{k} w_{\nu(i)} f(y_{\nu(i)};\theta), \tag{1}$$

where $f(y_{\nu(i)};\theta) \le f(y_{\nu(i+1)};\theta)$, $f(y_i;\theta) = -\log\varphi(y_i;\theta)$, $y_i \in \mathcal{Y} \subset R^q$ for $i = 1,\dots,n$ are i.i.d. observations with probability density $\varphi(y,\theta)$, which depends on an unknown parameter $\theta \in \Theta^p \subset R^p$, $\nu = (\nu(1),\dots,\nu(n))$ is the corresponding permutation of the indices, which depends on θ, k is the trimming parameter and the weights $w_i \ge 0$ for $i = 1,\dots,n$ are associated with $f(y_i,\theta)$ and are such that $w_{\nu(k)} > 0$.

The basic idea behind the trimming in this estimator is in the removal of those $n-k$ observations whose values would be highly unlikely to occur if the fitted model was true. Due to the representation $\min\limits_{\theta\in\Theta^p} \sum\limits_{i=1}^{k} w_{\nu(i)} f(y_{\nu(i)};\theta) = \min\limits_{\theta\in\Theta^p} \min\limits_{I\in I_k} \sum\limits_{i\in I} w_i f(y_i;\theta) = \min\limits_{I\in I_k} \min\limits_{\theta\in\Theta^p} \sum\limits_{i\in I} w_i f(y_i;\theta)$, where I_k is the set of all k–subsets of the set $\{1,\dots,n\}$, it follows that all possible $\binom{n}{k}$ partitions of the data have to be fitted by the MLE. The WTL_k estimator is given by that partition with that MLE fit for which the negative log likelihood is minima.

The WTL_k estimator reduces to: (i) the MLE if $k = n$; (ii) the TLE if $w_{\nu(i)} = 1$ for $i = 1,\dots,k$ and $w_{\nu(i)} = 0$ otherwise, the MedLE if $w_{\nu(k)} = 1$ and $w_{\nu(i)} = 0$ otherwise, e.g., [19]. If $\varphi(y,\theta)$ is the multivariate normal density function then the MedLE and TLE coincide with MVE and MCD estimators of [21], if $\varphi(y,\theta)$ is the normal regression error density the MedLE and TLE coincide with the LMS and LTS estimators of [21]. Detailes can be found in [26] and [27].

General conditions for the existence of a solution of (1) can be found in [3] whereas the consistency is proved in [2]. In the GLMs framework, the BP properties of (1) are studied in [17]. For the particular cases of normal, logistic and log-linear regression models it is proved that the BP of the WTL_k estimator is $\frac{1}{n}\min\{n-k+1, k-\mathcal{N}(X)\}$, where $\mathcal{N}(X) := \max_{0\ne\beta\in R^p} \operatorname{card}\left\{i \in \{1,\dots,n\};\ x_i^\top\beta = 0\right\}$ is the maximum number of carriers $x_i \in R^p$ lying in a subspace, $X := (x_i^\top)$ is the carriers data matrix and $x_i^\top\beta$ is the so called linear predictor. If x_i are linearly independent then $\mathcal{N}(X) = p-1$. The BP can be exemplified by the range of the values of k.

For increasing k the estimator will possess a BP point less than the highest possible but it will be more efficient at the same time.

Computing the WTL$_k$ estimator is infeasible for large data sets because of its combinatorial and nonlinear optimization nature. To get approximate TLE an algorithm called FAST-TLE was developed in [18]. It reduces to the FAST-LTS/LMS/LTA or FAST-MCD/MVE algorithms considered in [9], [22] and [23] in the normal regression or multivariate Gaussian cases. The basic idea behind the FAST-TLE algorithm consists of carrying out finitely many times a two-step procedure: a trial step followed by a refinement step. In the trial step a subsample of size k^* is selected randomly from the data sample and then the model is fitted to that subsample to get a trial ML estimate. The refinement step is based on the so called concentration procedure. The cases with the k smallest negative log likelihoods from the trial fit are found. Fitting the model to these k cases gives an improved fit. Repeating the improvement step yields an iterative process. The convergence is always guaranteed after a finite number of steps since there are only finitely many k–subsets out of $\binom{n}{k}$ in all. At the end of this procedure the solution with the lowest value of (1) is stored. There is no guarantee that this value will be the global minimizer of (1) but one can hope that it would be a close approximation to it. The trial subsample size k^* should be greater than or equal to $\mathcal{N}(X)+1$ which is needed for the existence of the MLE but the chance to get at least one outlier free subsample is larger if $k^* = \mathcal{N}(X)+1$. Any k within the interval [N(X)+1, n] can be chosen in the refinement step. A recommendable choice of k is $\lfloor (n+\mathcal{N}(X)+1)/2 \rfloor$ because then the BP of the TLE is maximized (see, [17]). The algorithm could be accelerated by applying the partitioning and nesting techniques as in [22] or [23]. We note that if the data set is small all possible subsets with size k can be considered.

3 Finite mixture of GLMs

Now a short reminder to mixtures will be given. Details can be found in [15]. Let (y_i, x_i) for $i = 1, \dots, n$ be a sample of i.i.d. observations such that y_i is coming from a mixture of $\psi_1(y_i; x_i, \theta_1), \dots, \psi_g(y_i; x_i, \theta_g)$ distributions, conditional on the variables $x_i \in R^p$, in proportions $\pi_1, \dots, \pi_g$ defined by

$$\varphi(y_i; x_i, \Psi) = \sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j), \tag{2}$$

where $\Psi = (\pi_1, \dots, \pi_{g-1}, \theta_1, \dots, \theta_g)^T$ is the unknown parameter vector, the proportions satisfy the conditions $\pi_j > 0$ for $j = 1, \dots, g$, and $\sum_{j=1}^{g} \pi_j = 1$. The log likelihood is given by $\log L(\Psi) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{g} \pi_j \psi_j(y_i; x_i, \theta_j)\}$. The EM algorithm is a standard technique to obtain the MLE of Ψ. It consists in maximizing the complete data log likelihood given by

$$\log L_c(\Psi) = \sum_{j=1}^{g} \sum_{i=1}^{n} z_{ij} \{\log \pi_j + \log \psi_j(y_i; x_i, \theta_j)\}, \tag{3}$$

where z_{ij} denote the component-indicator variables, depending on whether y_i does or does not belong to the jth component. The algorithm proceeds iteratively, alternating the E and M steps. In the $(l+1)$th iteration of the E-step the posterior probabilities for each observation are computed as $\hat{z}_{ij}^{(l+1)}(y_i; x_i, \Psi^{(l)}) = \pi_j^{(l)} \psi_j(y_i; x_i, \theta_j^{(l)}) / \sum_{j=1}^{g} \pi_j^{(l)} \psi_j(y_i; x_i, \theta_j^{(l)})$. In the $(l+1)$th M-step iteration the prior probabilities are computed by $\pi_j^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ij}^{(l+1)}(y_i; x_i, \Psi^{(l)})$ and then the function is maximized

$$\max_{\theta_1,\ldots,\theta_g} \sum_{j=1}^{g} \sum_{i=1}^{n} \hat{z}_{ij}^{(l+1)}(y_i; x_i, \Psi^{(l)}) \log \psi_j(y_i; x_i, \theta_j). \tag{4}$$

For mixtures of GLMs, $\theta_j = h(x^T \beta_j),\ j = 1, \ldots, g$, the function h is appropriately chosen and under the assumption that the parameters $\beta_1, \ldots, \beta_g$ have no elements in common a priori (4) is maximized for each component separately using the posterior probabilities as weights (see, [15]).

4 Adjustments of the FAST-TLE to mixture of GLMs

The FAST-TLE algorithm can be easily implemented using the environment of software packages such as GLIM, S-PLUS, R, SAS, STATISTICA, etc., since the trial and refinement steps are based on a standard MLE procedure. In the following we illustrate this in the framework of mixtures of GLMs using the program FlexMix as a computational engine for fitting mixtures of GLMs models and model-based cluster analysis in R, described in [12]. The trial and refinement sample sizes k^* and k depend not only on the sample size but also on the number of mixture components and model parameters. As the linear predictor of a standard GLMs consists of an intercept and p carriers the number of the unknown parameters is $p+1$, hence in a mixture with g components this number is $g(p+1)$. Therefore the trial sample size k^* in a g components mixture of GLMs with random carriers has to be at least $g(p+1)$ to ensure the estimability in each component, otherwise $g(\mathcal{N}(X)+2)$. We recommend a larger trial subsample size in order to increase the chance for each component to contain at least $(p+1)$ cases. We also recommend a larger refinement sample size, say 80% or 90% of the data size, as in mixtures the majority of the data have to accommodate several heterogeneous components. Any prior information about the data structure could be useful at this stage. According to the FAST-TLE algorithm a trial MLE, $\widetilde{\Psi}$, is found by maximizing (3) over the trail subsample with size k^* instead of n. In the refinement step we are evaluating (2) at $\widetilde{\Psi}$ for $i = 1, \ldots, n$

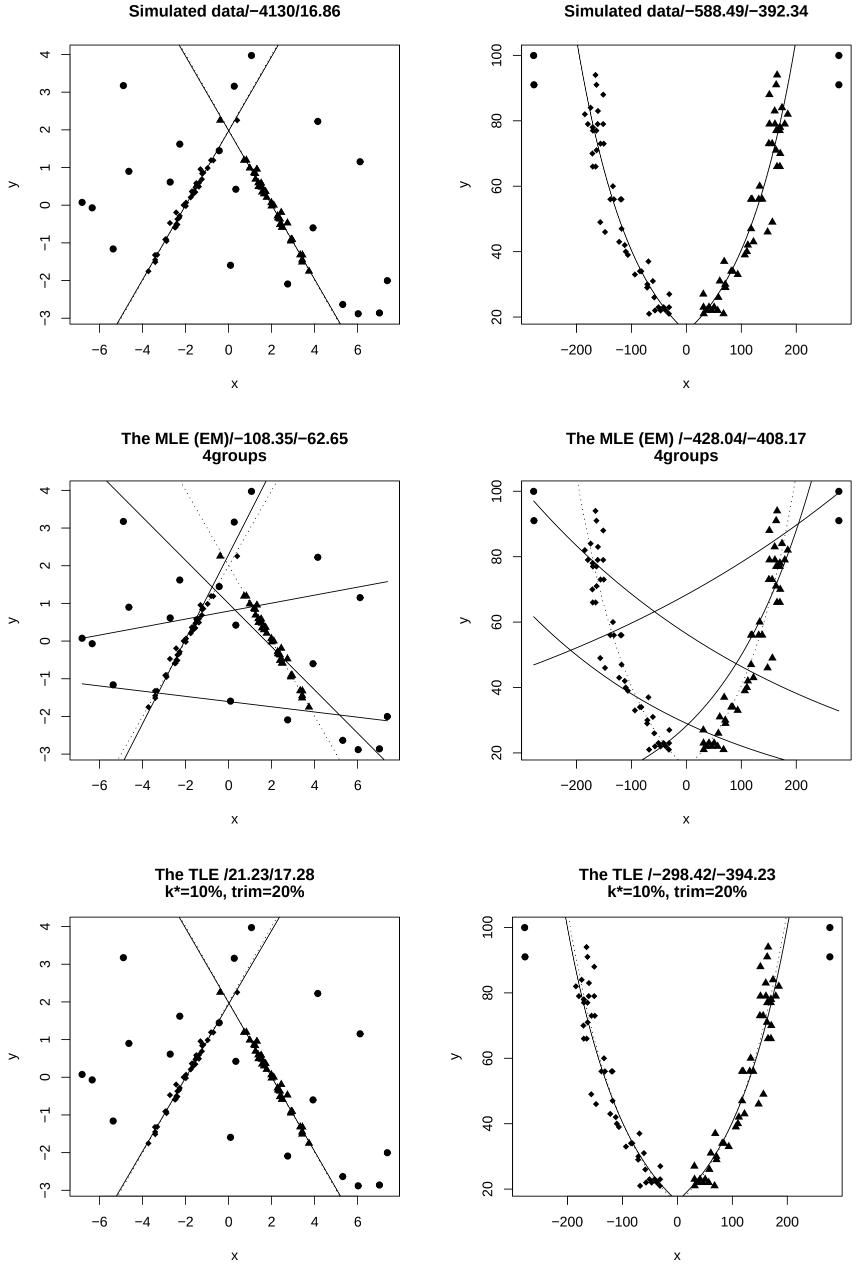
and then sorting $f(y_i; x_i, \widetilde{\Psi}) = -\log \varphi(y_i; x_i, \widetilde{\Psi})$ in ascending order to get the indices of the k smallest cases. The improved fit $\widehat{\Psi}$ is then obtained by maximizing (3) over these k cases.

5 Examples

Two artificially generated data sets, the mixture of normal and Poisson regression models, respectively, are shown on the upper two plots of Figure 1. On the left-hand upper plot, the points 1-40 are generated according to $x \sim N(2, 1)$, $y = 2 + x + \varepsilon$, $\varepsilon \sim N(0, 0.1)$, whereas the points 41-80 are their mirror pattern, and the points 81-100 are outliers uniformly distributed in the area $[x_{min}, x_{max}] \times [y_{min}, y_{max}]$. On the right-hand upper plot the points 1-48 are generated according to $x \sim U(20, 200)$, $\quad \mathrm{E}y = \log \lambda = 3 + 0.01x, \quad y \sim Poisson(\lambda)$, whereas the points 49-96 are their mirror pattern, and the points 97-100 are outliers. In both plots, the points that follow the models are marked by triangles and rhombs whereas the outliers are marked by bullets. The lines on the upper two plots, and their dotted analogs on the other 4 plots correspond to the true models. The continuous lines in the middle and bottom plots correspond to the fits. The upper plots heading values correspond to the negative log likelihood sums based on the whole samples and "good" subsamples evaluated at the true model parameters. The left heading values on the remaining plots correspond to the negative log likelihood and TLE minima based on the whole sample cases. The right heading values correspond to the negative log likelihood sums of the "good" cases evaluated at the MLE and TLE based on the whole samples.

The middle two plots give an impression about the MLE fits due to the program FlexMix starting and ending with a mixture of 4 components. The results of the mixture of Poisson models will be discussed only because of the similarity with the normal case. In fact we performed 4 experiments over the same data set in order to guarantee the reliability of the estimation procedure because of the internal random mechanism of the EM algorithm, respectively the FlexMix program, as regards the initial classification of the data. Each one of these experiments consists of 250 FlexMix runs starting respectively with 2, 3, 4 and 5 specified mixture components to assess the quality of the fits. As a result of these fits 4×250 plots were produced and examined. Only 12 times the mixture components were "correctly identified" which means: (i) on the background of 5 specified components the true components were 8 times nicely fitted in those 250 runs, however, 3 nonsense structures were also identified at the same time; (ii) on the background of 4 specified components the true components were 4 times nicely fitted in those 250 runs, however, 2 nonsense structures were also identified; (iii) in the remaining 500 trials neither a single nor 2 or 3 components of the mixture fit was satisfactory.

The bottom plots give an impression about the TLE fits due to the FAST-TLE algorithm using the FlexMix program with $k^* = 0.1n$ and $k = 0.8n$ in



Simulated data/−4130/16.86

Simulated data/−588.49/−392.34

The MLE (EM)/−108.35/−62.65
4groups

The MLE (EM) /−428.04/−408.17
4groups

The TLE /21.23/17.28
k*=10%, trim=20%

The TLE /−298.42/−394.23
k*=10%, trim=20%

Figure 1: The artificially generated data sets based on mixtures of two normal and two Poisson regression models with outliers are given on the left-hand and right-hand upper plots, respectively. The MLE and TLE fits are given on the middle and bottom plots, respectively.

both mixture types starting with mixtures of 2 components in 500 runs. The true structures were correctly identified in all runs within 30 repetitions of the procedure. Similar results were obtained with $k^* = 0.1n,\ k = 0.7n$. The MLE and TLE behavior was studied over many simulated mixtures of GLMs data with similar designs. The results were similar to the presented here.

References

[1] Campbell N.A. (1984). *Mixture models and atypical values.* Math. Geology **16**,465 – 477.

[2] Cizek P.(2002). *Robust estimation in nonlinear regression and limited dependent variable models.* http://econpapers.hhs.se/paper/wpawuwpem/0203003.htm.

[3] Dimova R., Neykov N.M. (2004). *Generalized d-fullness technique for breakdown point study of the trimmed likelihood estimator with applications.* In: Theory and Applications of Recent Robust Methods, M. Hubert, G. Pison, A. Struyf and S. Van Aelst, (eds.), Birkhäuser, Basel.

[4] Field C., Smith B. (1994). *Robust estimation - a weighted maximum likelihood approach.* Int. Statist. Rev. **62**, 405 – 424.

[5] Gallegos M.T. (2000). *A robust method for clustering analysis.* TR MIP-0013, Fakultät für Mathematik und Informatik, Universität Passau.

[6] Green P.J. (1984). *Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives.* J. Roy. Statist. Soc. Ser. B **46**, 149 – 192.

[7] Hadi A.S., Luceño A. (1997). *Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms.* Computational Statistics and Data Analysis **25**, 251 – 272.

[8] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust statistics. The approach based on influence functions.* Wiley, NY.

[9] Hawkins D.M., Olive D.J. (2002). *Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm* (with discussions). J. Amer. Statist. Assoc. **97**, 136 – 159.

[10] Hennig C. (2003). *Fixed point clusters.* J. of Classification **19**, 249 – 276.

[11] Huber P. (1981). *Robust statistics.* John Wiley & Sons, New York.

[12] Leisch F. (2003). *FlexMix.* Reference manual:
`http://cran.R-project.org/doc/packages/flexmix.pdf.`

[13] Markatou M. (2000). *Mixture models, robustness, and the weighted likelihood methodology.* Biometrics **56**, 483 – 486.

[14] Markatou M., Basu A., Lindsay B. (1997). *Weighted likelihood estimating equations: The discrete case with applications to logistic regression.* J. Statist. Plann. Inference **57**, 215 – 232.

[15] McLachlan G.J., Peel D. (2000). *Finite mixture models.* Wiley, NY.

[16] Morgenthaler S. (1990). *Fitting redescending M-estimators in regression.* In: Robust regression, H. D. Lawrence and S. Arthur (eds.), 105 – 128.

[17] Müller C.H., Neykov N.M. (2003). *Breakdown points of the trimmed likelihood and related estimators in generalized linear models.* J. Statist. Plann. Inference **116**, 503 – 519.

[18] Neykov N.M., Müller C.H. (2002). *Breakdown point and computation of trimmed likelihood estimators in generalized linear models.* In: Developments in robust statistics, R. Dutter, P. Filzmoser, U. Gather, P. Rousseeuw (eds.), Physica-Verlag, Heidelberg, 277 – 286.

[19] Neykov N.M., Neytchev P.N. (1990). *A robust alternative of the maximum likelihood estimator.* In: Short communications of COMPSTAT'90, Dubrovnik, 99 – 100.

[20] Rocke D.M., Woodruff D.L. (2002). *Computational connections between robust multivariate analysis and clustering.* In: Proc. of COMPSTAT, 255 – 260.

[21] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* Wiley, New York.

[22] Rousseeuw P.J., Van Driessen K. (1999). *Computing LTS regression for large data sets.* Technical report, University of Antwerp, (submitted).

[23] Rousseeuw P.J., Van Driessen K. (1999). *A fast algorithm for the MCD estimator.* Technometrics **41**, 212 – 223.

[24] Tibshirani R., Knight K. (1999). *Bootstrap bumping.* J. Comp. and Graph. Statist. **8**, 671 – 686.

[25] Vandev D.L. (1993). *A note on breakdown point of the least median squares and least trimmed squares.* Statistics and Probability Letters **16**, 117 – 119.

[26] Vandev D.L., Neykov N.M. (1993). *Robust maximum likelihood in the Gaussian case.* In: New directions in data analysis and robustness, S. Morgenthaler, E. Ronchetti, W.A. Stahel (eds.), (Birkhäuser Verlag, Basel, 259 – 264.

[27] Vandev D.L., Neykov N.M. (1998). *About regression estimators with high breakdown point.* Statistics **32**, 111 – 129.

[28] Windham M.P. (1995). *Robustifying model fitting.* J. Roy. Statist. Soc. Ser. B **57**, 599 – 609.

[29] Wedel M., Kamakura W.A. (1998). *Market segmentation: Conceptual and methodological foundations.* Dordrecht: Kluwer Academic Press.

Acknowledgement: The authors thank the referees for valuable comments. The authors would like to thank F. Leisch for the interesting discussions concerning mixtures and the program FlexMix.

Address: N. Neykov, R. Dimova, P. Neytchev, National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66 Tsarigradsko chaussee, Sofia 1784, Bulgaria

P. Filzmoser, Dept. of Statistics & Probability Theory, Vienna, University of Technology, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria

E-mail: Neyko.Neykov@meteo.bg, P.Filzmoser@tuwien.ac.at

COMPUTING THE DERIVATIVES OF THE AUTOCOVARIANCES OF A *VARMA* PROCESS

Jurek Niemczyk

Key words: Matrix differentiation, *VARMA* model, autocovariances, cross-covariances.

COMPSTAT 2004 section: Time series analysis.

Abstract: This paper describes the computation of the autocovariance function of a vector autoregressive moving average process (*VARMA*). In particular, it develops the derivatives of the autocovariance function with respect to all *VARMA* coefficients. Two effective applications are described and an example is given. The computation of the derivatives of the covariance between the variable and the lagged innovation is provided as well.

1 Introduction

There are several reasons for being interested in obtaining the exact derivatives of the autocovariance function of a *VARMA*(p,q) process and of the covariances between the variable and the lagged innovation. One of these reasons is to evaluate the quasi-maximum likelihood estimator for *VARMA* models. This is done using a numerical optimization algorithm where the objective function is the logarithm of the exact likelihood of the Gaussian *VARMA* process. This leads to better small-sample properties than least squares, especially when $q > 0$ (e.g Mélard, et al. [2], [14]). There are several algorithms for computing that exact likelihood, some of which include the use of the Kalman filter for the state space representation of the model. The necessary starting values (including the covariance of state vector at time 1) depend on the autocovariance function of the process and also on cross-covariances between the variable and the lagged innovation, for given values of the parameters. Numerical analysts recommend to use the exact derivatives of the objective function when it is available, since they are generally more accurate than numerical approximations of these derivatives. For computing the derivatives of the log-likelihood, we need also the derivatives of the starting values, hence those of the autocovariance function of the process and of the cross-covariances between the variable and the lagged innovation. Note that if this approach has been used for scalar *ARMA* processes (e.g. Mélard [9] and Kohn and Ansley [6], it has not been done for *VARMA* models, yet.

Another reason for being interested in the derivatives of the autocovariance function of a *VARMA* process is the evaluation of the exact Fisher information matrix, which is a new tool in describing the covariance structure of the *MLE*, see Klein et al [3]. In fact, Klein et al. [3] have observed that their program is sensitive to a correct routine for computing these derivatives. The method described in this paper will solve this shortcoming, and this was indeed the initial motivation of the present study.

Our computations rely on the algebra of Kronecker product and matrix vectorization, together with the matrix differential rules, Neudecker [13].

The paper is organized as follows. In section 2 the most important notations and definitions are given. In section 3 the autocovariances and cross-covariances of *VARMA* model and their derivatives are computed. In section 4 an example of the derivatives of the autocovariances of a 2-dimensional *VAR(*2*)* is provided. Section 5 is devoted to some conclusions.

2 Notations and definitions

Substantial notations and definitions which are used in this paper, in order to derive the autocovariances, cross-covariances and their derivatives, are presented in this section. They are based on Mittnick [12] and Magnus and Neudecker [7].

Let the matrix $\underset{(nm\times m)}{A}$ be of the form $\left[A_1' \dots A_n'\right]'$ and let $\underset{(m\times n)}{F}$ be the differentiable real matrix function of the matrix $\underset{(k\times l)}{X}$.

1. $\underset{(nm\times nm)}{\mathrm{T}_{\triangleright} A}$ is the lower triangular block Toeplitz matrix $\begin{bmatrix} A_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ A_n & \dots & A_1 \end{bmatrix}$.
2. $\underset{(nm\times nm)}{\mathrm{T}^{\triangleright} A}$ is the upper triangular block Toeplitz matrix $\begin{bmatrix} A_n & \dots & A_1 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A_n \end{bmatrix}$.
3. $\underset{(nm\times nm)}{\mathrm{H}^{\triangleright} A}$ is the upper counter-triangular block Hankel matrix $\begin{bmatrix} A_1 & \dots & A_n \\ \vdots & \ddots & \vdots \\ A_n & \dots & 0 \end{bmatrix}$.
4. $\underset{(mn\times kl)}{\frac{\partial \mathrm{vec} F}{\partial \mathrm{vec} X}}$ is the *Jacobian matrix* of F at X
5. $\mathbf{0}_{m\times n}$ denotes the $(m \times n)$ zero matrix.
6. P_{mn} is the commutation matrix defined as $\sum_{i=1}^{m}\sum_{j=1}^{n}(H_{ij} \otimes H_{ij}')$, where $H_{ij} = \mathrm{e}_i^m \mathrm{e}_j^{n\prime}$ and e_i^n denotes i-th column of the identity matrix I_n. Some of the properties of the commutation matrix, P_{mn}, are $P_{mn} = \sum_{j=1}^{n}(\mathrm{e}_j^n \otimes I_m \otimes \mathrm{e}_j^{n\prime}) = \sum_{j=1}^{m}(\mathrm{e}_j^m \otimes I_n \otimes \mathrm{e}_j^{m\prime})$ and $P_{mn} = P_{nm}'$.
7. $\mathrm{vec}(A)$ denotes the vectorization of a $(m \times n)$ matrix $A = [a_{ij}]$, $(\mathrm{a}_1', \dots, \mathrm{a}_n')'$, where a_i is the i-th column of the matrix A. In this paper we use the following properties of the vec operator: $\mathrm{vec}(ABC) = (C' \otimes A)\mathrm{vec}(B)$, for the conformable matrices A, B, C, and $\mathrm{vec}(A') = P_{mn}\mathrm{vec}(A)$.

3 The model

We deal with an $m-$dimensional vector *ARMA*(p,q) process satisfying

$$Y_t = A_1 Y_{t-1} + ... + A_p Y_{t-p} + Z_t - B_1 Z_{t-1} - ... - B_q Z_{t-q}, \tag{1}$$

where A_i, B_j $(i = 1, \dots, p,\ j = 1, \dots, q)$ are $(m \times m)$ matrices, $Z_t \sim \mathcal{D}(\mathbf{0}_{m\times 1}, \Sigma)$ is an m-dimensional *innovation process*, that is, $EZ_t = \mathbf{0}_{m\times 1}$ and $EZ_s Z_t' = \delta_{st}\Sigma$, with a symmetric, semi-positive definite matrix Σ. Using the *lag* operator L we can rewrite (1) as $A(L)Y_t = B(L)Z_t$, where $A(L) = I_m - A_1 L - ... - A_p L^p$ and $B(L) = I_m - B_1 L - ... - B_p L^q$ are *lag* polynomials. Process Y_t is assumed to be causal,[1] as a result, we have $Y_t = A(L)^{-1}B(L)Z_t \equiv C(L)Z_t = \sum_{i=0}^{\infty} C_i Z_{t-i}$, where the coefficients C_i can be obtained recursively

$$C_0 = I_m, \quad C_j = \sum_{i=1}^{j} A_i C_{j-i} - B_j, \quad j > 0. \tag{2}$$

3.1 Autocovariance function

In this subsection we derive the autocovariance function of the *VARMA* process. Postmultiplying (1) by Y_{t-k}' and taking the expectation we obtain

$$\Gamma(k) = \sum_{i=1}^{p} A_i \Gamma(k-i) + G(k), \tag{3}$$

where $G(k) = EZ_t Y_{t-k}' - B_1 EZ_{t-1} Y_{t-k}' - ... - B_q EZ_{t-q} Y_{t-k}'$. Denoting $g_i = \mathrm{vec}G(i)$, it can be shown that we can obtain g_i applying the following recurrence equation

$$g_k - \sum_{i=1}^{\min(p,q-k)} (A_i \otimes I_m) g_i = \sum_{j=1}^{q} (B_{j-k} \otimes B_j)\mathrm{vec}(\Sigma), \tag{4}$$

for $0 \le k \le q$ and $g_j = \mathbf{0}_{m^2\times 1}$ for $j > q$, where $B_0 = -I_m$. Now, let us denote matrix $\underset{(m^2\times m^2)}{A_i \otimes I_m}$ by $\mathbb{A}_i$, then we can rewrite (4) in the following matrix form, $\mathrm{Ng} = \mathbb{B}$, where

$$\mathrm{N} = \begin{cases} \mathrm{T}^{\triangleright}\big(\begin{bmatrix} -\mathbb{A}_q' & \dots & -\mathbb{A}_1' & I_{m^2} \end{bmatrix}'\big) & \text{if } q \le p \\ \mathrm{T}^{\triangleright}\big(\begin{bmatrix} \mathbf{0}_{m^2\times(q-p)m^2} & -\mathbb{A}_p' & \dots & -\mathbb{A}_1' & I_{m^2} \end{bmatrix}'\big) & \text{if } q > p \end{cases},$$

$$\underset{(q+1)m^2\times 1}{\mathrm{g}} = \begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_{q-1} \\ g_q \end{bmatrix} \quad \text{and} \quad \underset{(q+1)m^2\times 1}{\mathbb{B}} = \begin{bmatrix} \sum_{j=0}^{q}(B_j \otimes B_j)\mathrm{vec}\Sigma \\ \sum_{j=1}^{q}(B_{j-1} \otimes B_j)\mathrm{vec}\Sigma \\ \vdots \\ \sum_{j=q-1}^{q}(B_{j-q+1} \otimes B_j)\mathrm{vec}\Sigma \\ (B_0 \otimes B_q)\mathrm{vec}(\Sigma) \end{bmatrix}.$$

[1] We require that $det(A(z)) \neq 0$ for all $z \in \mathbb{R}$ such that $|z| \le 1$.

Since the matrix N is square and invertible, we have

$$\mathrm{g} = \mathrm{N}^{-1}\mathbb{B}. \tag{5}$$

Taking vec of (3), denoting $\gamma_i = \mathrm{vec}\Gamma(i)$ and $\mathbb{M}_i = I_m \otimes A_i$ we have

$$\gamma_k = \sum_{i=1}^{p} \mathbb{M}_i \gamma_{k-i} + g_k. \tag{6}$$

Within the matrix form $\mathbf{M}\gamma = \hat{\mathrm{g}}$, where

$$\underset{(p+1)m^2\times 1}{\hat{\mathrm{g}}} = \begin{cases} \begin{bmatrix} g_0' & \cdots & g_q' & \mathbf{0}_{1\times(p-q)m^2} \end{bmatrix}' & \text{if } q \leq p \\ \begin{bmatrix} g_0' & \ldots g_p' \end{bmatrix} & \text{if } q > p \end{cases}, \quad \underset{(p+1)m^2\times 1}{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_p \end{bmatrix},$$

$$\underset{(p+1)m^2\times(p+1)m^2}{\mathbf{M}} = \mathbf{M}^H + \mathbf{M}_T,$$

where $\mathbf{M}^H = \mathrm{H}^{\triangleright} \begin{bmatrix} I_{m^2} \\ -\mathbb{M}_1 P_{mm} \\ \vdots \\ -\mathbb{M}_p P_{mm} \end{bmatrix}$, $\quad \mathbf{M}_T = \begin{bmatrix} \mathbf{0}_{m^2\times m^2} & \mathbf{0}_{m^2\times pm^2} \\ \mathbf{0}_{pm^2\times m^2} & \mathrm{T}_{\triangleright} \begin{bmatrix} I_{m^2} \\ -\mathbb{M}_1 \\ \vdots \\ -\mathbb{M}_{p-1} \end{bmatrix} \end{bmatrix}.$

Since the matrix $\mathbf{M}$ is square and invertible, we have

$$\gamma = \mathbf{M}^{-1}\hat{\mathrm{g}}. \tag{7}$$

Result (7) is not new (see Ansley [1]) and was even slightly improved by Kohn and Ansley [5]. It is the basis of our contribution.

Remark 3.1. We used the fact that $\Gamma(-k) = \Gamma(k)'$ and $\mathrm{vec}\Gamma(k)' = P_{mm}\mathrm{vec}\Gamma(k)$. Having computed γ we can obtain higher order autocovariances γ_h, for $h > p$, using the recurrence relation (6) and remembering that $g_h \equiv \mathbf{0}_{m^2\times 1}$ if $h > q$.

3.2 Derivatives of the autocovariance function

In this subsection we deal with the derivatives of the autocovariance function. We use the facts that $\partial(XY) = \partial XY + X\partial Y$ and that $\partial(X\otimes Y) = \partial X \otimes Y + X \otimes \partial Y$. For all details see Magnus and Neudecker [8]. Denoting $F = \begin{bmatrix}\Gamma(0) & \ldots & \Gamma(p)\end{bmatrix}$ and $X = \begin{bmatrix} A_1 & \ldots & A_p & B_1 & \ldots & B_q \end{bmatrix}$, we need to obtain $\frac{\partial \mathrm{vec}F}{\partial \mathrm{vec}X} = \frac{\partial\gamma}{\partial\mathrm{vec}X}$. When we differentiate (7) we get

$$\begin{aligned} \partial\gamma &= \partial\mathbf{M}^{-1}\hat{\mathrm{g}} + \mathbf{M}^{-1}\partial\hat{\mathrm{g}} = \mathrm{vec}(\partial\mathbf{M}^{-1}\hat{\mathrm{g}}) + \mathbf{M}^{-1}\partial\hat{\mathrm{g}} \\ &= -((\mathbf{M}^{-1}\hat{\mathrm{g}})' \otimes \mathbf{M}^{-1})\partial\mathrm{vec}\mathbf{M} + \mathbf{M}^{-1}\partial\hat{\mathrm{g}} \end{aligned}$$

therefore, $\frac{\partial\mathrm{vec}F}{\partial\mathrm{vec}X} = -((\mathbf{M}^{-1}\hat{\mathrm{g}})' \otimes \mathbf{M}^{-1})\frac{\partial\mathrm{vec}\mathbf{M}}{\partial\mathrm{vec}X} + \mathbf{M}^{-1}\frac{\partial\hat{\mathrm{g}}}{\partial\mathrm{vec}X}$, where, depending on q,

$$\frac{\partial \hat{\mathrm{g}}}{\partial \mathrm{vec} X} = \begin{cases} \left[\frac{\partial g_0}{\partial \mathrm{vec}X}' \quad \dots \quad \frac{\partial g_q}{\partial \mathrm{vec}X}' \quad \mathbf{0}'_{(p-q)m^2 \times (p+q)m^2} \right]' & \text{if } q \leq p \\ \left[\frac{\partial g_0}{\partial \mathrm{vec}X}' \quad \dots \quad \frac{\partial g_p}{\partial \mathrm{vec}X}' \right]' & \text{if } q > p \end{cases},$$

where $\frac{\partial g_i}{\partial \mathrm{vec}X}$ are obtained by differentiating (5), $\frac{\partial \mathrm{g}}{\partial \mathrm{vec}X} = -((\mathbf{N}^{-1}\mathbb{B})' \otimes \mathbf{N}^{-1})\frac{\partial \mathrm{vec}\mathbf{N}}{\partial \mathrm{vec}X} + \mathbf{N}^{-1}\frac{\partial \mathbb{B}}{\partial \mathrm{vec}X}$.

Algorithm 3.1. Let $X_A \equiv [A_1 \dots A_p]$, $X_B \equiv [B_1 \dots B_q]$, as $\mathbf{M}([I_m, X_A])$ and $\mathbf{N}([I_m, X_A])$ are the linear matrix functions of the Kronecker products of I_m together with the submatrices of $[I_m, X_A]$, we have $\frac{\partial \mathrm{vec}\mathbf{M}}{\partial \mathrm{vec}X_A} = \left[M^1_{11} \dots M^1_{m1} \dots M^1_{1m} \dots M^1_{mm} \dots M^p_{11} \dots M^p_{m1} \dots M^p_{1m} \dots M^p_{mm}\right]$, where $M^k_{ij} = \mathrm{vec}\mathbf{M}\big([\mathbf{0}_{m^2 \times m^2}, [\mathbf{0}_{m^2 \times (k-1)m^2} \quad \mathrm{e}_i\mathrm{e}'_j \quad \mathbf{0}_{m^2 \times (p-k)m^2}]]\big)$, for $i, j = 1, \dots, m$, $k = 1, \dots p$. Since $\mathbf{M}$ is not a function of X_B, $\frac{\partial \mathrm{vec}\mathbf{M}}{\partial \mathrm{vec}X} = \left[\frac{\partial \mathrm{vec}\mathbf{M}}{\partial \mathrm{vec}X_A} \quad \mathbf{0}_{m^4(p+1)^2 \times m^2 q}\right]$. We proceed similarly with $\frac{\partial \mathrm{vec}\mathbf{N}}{\partial \mathrm{vec}X}$. Since $\mathbb{B}$ is the linear matrix function of the constant matrix $\mathrm{vec}\Sigma$ and the Kronecker products of I_m, together with the submatrices of $[I_m, X_B]$, defining

$$\mathbb{B}_1(\partial[B_0, X_B], [B_0, X_B], \Sigma) + \mathbb{B}_2([B_0, X_B], \partial[B_0, X_B], \Sigma) \equiv$$

$$\begin{bmatrix} \sum_{j=0}^{q}(\partial B_j \otimes B_j)\mathrm{vec}\Sigma \\ \sum_{j=1}^{q}(\partial B_{j-1} \otimes B_j)\mathrm{vec}\Sigma \\ \vdots \\ \sum_{j=q-1}^{q}(\partial B_{j-q+1} \otimes B_j)\mathrm{vec}\Sigma \\ (\partial B_0 \otimes B_q)\mathrm{vec}(\Sigma) \end{bmatrix} + \begin{bmatrix} \sum_{j=0}^{q}(B_j \otimes \partial B_j)\mathrm{vec}\Sigma \\ \sum_{j=1}^{q}(B_{j-1} \otimes \partial B_j)\mathrm{vec}\Sigma \\ \vdots \\ \sum_{j=q-1}^{q}(B_{j-q+1} \otimes \partial B_j)\mathrm{vec}\Sigma \\ (B_0 \otimes \partial B_q)\mathrm{vec}(\Sigma) \end{bmatrix} = \partial \mathbb{B},$$

we have, $\frac{\partial \mathrm{vec}\mathbb{B}}{\partial \mathrm{vec}X_B} = \left[B^1_{11} \dots B^1_{m1} \dots B^1_{1m} \dots B^1_{mm} \dots B^q_{11} \dots B^q_{m1} \dots B^q_{1m} \dots B^q_{mm}\right]$, where
$B^k_{ij} = \mathrm{vec}\mathbb{B}_1\big([\mathbf{0}_{m^2 \times m^2}, [\mathbf{0}_{m^2 \times (k-1)m^2} \quad \mathrm{e}_i\mathrm{e}'_j \quad \mathbf{0}_{m^2 \times (q-k)m^2}]], [B_0, X_B], \Sigma\big)$
$+\mathrm{vec}\mathbb{B}_2\big([B_0, X_B], [\mathbf{0}_{m^2 \times m^2}, [\mathbf{0}_{m^2 \times (k-1)m^2} \quad \mathrm{e}_i\mathrm{e}'_j \quad \mathbf{0}_{m^2 \times (q-k)m^2}]], \Sigma\big)$,
for $i, j = 1, \dots, m$ and $k = 1, \dots, q$. $\mathbb{B}$ is not a function of X_A, therefore $\frac{\partial \mathrm{vec}\mathbb{B}}{\partial \mathrm{vec}X} = \left[\mathbf{0}_{m^2(q+1) \times m^2 p} \quad \frac{\partial \mathrm{vec}\mathbb{B}}{\partial \mathrm{vec}X_A}\right]$. In order to compute the derivatives of higher order autocovariances we use the following recurrence $\partial\gamma_k = \sum_{i=1}^{p}(I_m \otimes \partial A_i)\gamma_{k-i} + \sum_{i=1}^{p}(I_m \otimes A_i)\partial\gamma_{k-i} + \partial g_k$, for $k > p$.

3.3 The derivatives of the covariance between the variable and the lagged innovation

To compute the derivatives of the covariance between the variable and the lagged innovation we use the following relation $EY_t Z'_{t-j} = C_j \Sigma$, for $j \geq 0$. Taking the vec of (2), denoting $c_i = \mathrm{vec}C_i$ and $b_i = \mathrm{vec}B_i$, for $1 \leq i \leq q$ and

$\mathbf{0}_{m^2\times 1}$ otherwise, we have

$$c_0 = \mathrm{vec} I_m, \qquad c_j = \sum_{i=1}^{j} (I_m \otimes A_i) c_{j-i} - b_j, \quad j > 0, \tag{8}$$

where $A_j = \mathbf{0}_{m\times m}$, for $j > p$. To obtain initial coefficients c_i, $i = 1, 2, ..., p$, we can solve the following system of equations $\mathbb{A}c = -b$, where
$\mathbb{A} = \mathrm{T}_{\triangleright}\big(\begin{bmatrix} I'_{m^2} & -\mathbb{M}'_1 & \dots & -\mathbb{M}'_p \end{bmatrix}'\big)$, $\underset{(p+1)m^2\times 1}{c} = \begin{bmatrix} c'_o & c'_1 & \dots & c'_p \end{bmatrix}'$ and

$$\underset{(p+1)m^2\times 1}{b} = \begin{cases} \begin{bmatrix} (\mathrm{vec} I_m)' & b'_1 & \dots & b'_q & \mathbf{0}_{1\times(p-q)m^2} \end{bmatrix}' & \text{if } p > q \\ \begin{bmatrix} (\mathrm{vec} I_m)' & b'_1 & \dots & b'_p \end{bmatrix}' & \text{if } p \leq q \end{cases}.$$

Given that $\mathbb{A}$ is invertible, $c = -\mathbb{A}^{-1}b$. To compute c_h, for $h > p$, we use the recurrence (8). Similarly, as in the previous subsection, we obtain the derivatives of c, $\frac{\partial c}{\partial \mathrm{vec} X} = \big((\mathbb{A}^{-1}b)' \otimes \mathbb{A}^{-1}\big) \frac{\partial \mathrm{vec}\mathbb{A}}{\partial \mathrm{vec} X} - \mathbb{A}^{-1} \frac{\partial b}{\partial \mathrm{vec} X}$. To compute $\frac{\partial \mathrm{vec}\mathbb{A}}{\partial \mathrm{vec} X}$ we proceed as in Algorithm 3.1. The computation of $\frac{\partial b}{\partial \mathrm{vec} X}$ is done in a similar fashion. In order to compute derivatives of c_h, for $h > p$, the following recurrence is used $\partial c_h = \sum_{i=1}^{p} (I_m \otimes \partial A_i) c_{h-i} + \sum_{i=1}^{p} (I_m \otimes A_i) \partial c_{h-i} - \partial b_h$, $h > p$. Let us denote $H \equiv \begin{bmatrix} EY_t Z'_t & \dots & EY_t Z'_{t-h} \end{bmatrix}$. Having computed c and $\frac{\partial c}{\partial \mathrm{vec} X}$ it suffices to premultiply the results by $\big(I_{h+1} \otimes (\Sigma' \otimes I_m)\big)$ to obtain $\mathrm{vec} H$ and $\frac{\partial \mathrm{vec} H}{\partial \mathrm{vec} X}$.

4 Example

In this section one example of the derivatives of the autocovariance function of two dimensional *VAR(2)* is presented. To check the accuracy of the program we can perform the following approximation of the derivatives. For some 'small' scalar m (e.g. 10^{-9}), $\partial\Gamma(X,h)/\partial x_{ijk} \approx (\Gamma(X^m_{ijk}, h) - \Gamma(X,h))/m$, where x_{ijk} is the ij-th element of the k-th sub-matrix of $X = [A_1 \dots B_q]$, with $1 \leq k \leq p+q$. X^m_{ijk} denotes the matrix which is obtained from X by adding the scalar m to its ijk-th element.
Let us consider the easiest example of 2-dimensional *VARMA(*$2,0$*)*, which is simply the process consisted of two *ARMA(*$2,0$*)* processes stacked together, $y^i_t = a^i_1 y^i_{t-1} + a^i_2 y^i_{t-2} + \epsilon^i_t, \quad i = 1, 2$. For example,

$$Y_t = A_1 Y_{t-1} + A_2 Y_{t-2} + Z_t, \qquad Z_t \sim \mathcal{N}(\mathbf{0}_{m\times 1}, \Sigma),$$

where $A_1 = \begin{bmatrix} a^1_1 & a^1_{12} \\ a^1_{21} & a^2_1 \end{bmatrix} = \left[\begin{smallmatrix} 0.7 & 0 \\ 0 & 0.1 \end{smallmatrix}\right]$, $A_2 = \begin{bmatrix} a^1_2 & a^2_{12} \\ a^2_{21} & a^2_2 \end{bmatrix} = \left[\begin{smallmatrix} -0.1 & 0 \\ 0 & 0.2 \end{smallmatrix}\right]$ and $\Sigma = \left[\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right]$.
Our program provides the following autocovariances and their derivatives

$$\begin{bmatrix} \Gamma(0) & \Gamma(1) & \Gamma(2) \end{bmatrix} = \begin{bmatrix} 3.3951 & 0 & 2.1605 & 0 & 1.1728 & 0 \\ 0 & 1.0582 & 0 & 0.1323 & 0 & 0.2249 \end{bmatrix},$$

$$\frac{\partial \mathrm{vec}\begin{bmatrix}\Gamma(0)' \\ \Gamma(1)' \\ \Gamma(2)'\end{bmatrix}'}{\partial \mathrm{vec}[B_1, B_2]} = \begin{bmatrix} 6.601 & 0 & 0 & 0 & 3.515 & 0 & 0 & 0 \\ 0 & 2.426 & 0.321 & 0 & 0 & 1.308 & 0.283 & 0 \\ 0 & 2.426 & 0.321 & 0 & 0 & 1.308 & 0.283 & 0 \\ 0 & 0 & 0 & 0.336 & 0 & 0 & 0 & 0.483 \\ 7.287 & 0 & 0 & 0 & 4.201 & 0 & 0 & 0 \\ 0 & 3.899 & 0.283 & 0 & 0 & 2.426 & 0.092 & 0 \\ 0 & 1.308 & 1.255 & 0 & 0 & 0.673 & 0.321 & 0 \\ 0 & 0 & 0 & 1.365 & 0 & 0 & 0 & 0.226 \\ 6.601 & 0 & 0 & 0 & 5.984 & 0 & 0 & 0 \\ 0 & 3.035 & 0.092 & 0 & 0 & 3.899 & 0.066 & 0 \\ 0 & 0.672 & 0.978 & 0 & 0 & 0.340 & 1.255 & 0 \\ 0 & 0 & 0 & 0.336 & 0 & 0 & 0 & 1.177 \end{bmatrix}.$$

The accuracy of the above *Jacobian matrix* connected to a_{12}^1, a_{21}^1, a_{12}^2 and a_{21}^2 has been confirmed using the proposed approximation,

$$\left.\frac{\partial \mathrm{vec}\begin{bmatrix}\Gamma(0)' \\ \Gamma(1)' \\ \Gamma(2)'\end{bmatrix}'}{\partial a_{ij}^k}\right|_{a_{ij}^k=0} \approx \frac{\mathrm{vec}\left\{\begin{bmatrix}\Gamma(X_{ijk}^m,0)' \\ \Gamma(X_{ijk}^m,1)' \\ \Gamma(X_{ijk}^m,2)'\end{bmatrix}' - \begin{bmatrix}\Gamma(X,0)' \\ \Gamma(X,1)' \\ \Gamma(X,2)'\end{bmatrix}'\right\}}{m}.$$

Moreover, a theoretical check has been carried out for those elements of the above *Jacobian matrix* that can be derived from the univariate *ARMA* processes. This proved their correctness. Finally, limitation of the available space allowed us to present the computation results only up to lag 2. The program was run with other causal *VARMA* specifications. The results were checked with the proposed approximation and theoretically, where possible.

5 Final remarks and conclusions

This paper has developed the algorithms which compute the exact derivatives of the autocovariance and the cross-covariance functions for a vector autoregressive moving average process. The procedure presented here is was aimed at facilitating the computation of the derivatives. More efficient approaches to computing the autocovariance function of a *VARMA* process can be found in Mittnik [11], [12], Tunnicliffe [15] and in forthcoming paper by Harti et al. [2]. However, these algorithms are probably more efficient but they are also much more complex. Therefore, obtaining the derivatives will also be more arduous. The *Matlab* program can be obtained from the author at address `http://homepages.ulb.ac.be/∼jniemczy`.

References

[1] Ansley C.F. (1980). *Computation of the theoretical autocovariance function for a vector ARMA process.* J. Statist. Comput. Simul. **12**, 15 – 24.

[2] Harti M., Mélard G., Pham D.T. (2004). *Computing covariances for scalar and vector ARMA processes.* Personal communication.

[3] Klein A., Mélard G., Zahaf T. (2000). *Construction of the exact Fisher information matrix of Gaussian time series models by means of matrix differential rules.* Linear Algebra and its Applications **321**, 209 – 232.

[4] Klein A., Mélard G., Zahaf T. (2003). *A program for computing the exact Fisher information matrix of a Gaussian VARMA model.* SIAM 2003 Conference on Applied Linear Algebra, July 15 - 19, Williamsburg.

[5] Kohn R., Ansley C. F. (1982). *A note on obtaining the theoretical autocovariance function for a vector ARMA process.* J. Statist. Comput. Simul. **15**, 237 – 283.

[6] Kohn R., Ansley C. F. (1985). *Computing the likelihood and its derivatives for a Gaussian ARMA model.* J. Statist. Comput. Simul. **22**, 229 – 263.

[7] Magnus J. R., Neudecker H. (1979). *The commutation matrix: some properties and applications.* The Annals of Statistics **2**, 381 – 394.

[8] Magnus J. R., Neudecker H. (1998). *Matrix differential calculus with applications in statistics and econometrics.* John Wiley & Sons, Chichester.

[9] Mélard G. (1985). *Exact derivatives of the likelihood of ARMA processes.* Proceedings of the Statistical Computing Section, American Statistical Association, Washington D.C., 187 – 192.

[10] Mélard G., Roy R., Saidi A. (2004). *The exact maximum likelihood estimation of structured or unit roots multivariate time series models.* Submited.

[11] Mittnik S. (1990). *Computation of theoretical autocovariance matrices of multivariate autoregressive moving average time series.* J. Roy. Statist. Soc. Ser. B **52**, 151 – 155.

[12] Mittnik S. (1993). *Computation of theoretical autocovariance matrices of multivariate autoregressive moving average by using a block Levinson method.* J. Roy. Statist. Soc. Ser. B **55**, 435 – 440.

[13] Neudecker H. (1969). *Some theorems on matrix differentiation with special reference to Kronecker matrix products.* The Journal of American Statistical Association **327**, 953 – 963.

[14] Shea B.L. (1989). *The exact likelihood of a vector autoregressive moving average model.* J. Roy. Statist. Soc. Ser. C **38**, 161 – 184.

[15] Tunnicliffe W. (1979). *Some efficient computational procedures for high order ARMA models.* J. Statist. Comput. Simul. **8**, 303 – 309.

Acknowledgement: This paper has benefited from an IAP-network in Statistics grant, contract P5/24, financially supported by the Belgian Federal Office for Scientific, Technical and Cultural Affairs (OSTC). The author is grateful to Guy Mélard for his support and useful comments whose incorporation improved the paper.

Address: J. Niemczyk, ISRO, Campus Plaine ULB CP 210 Boulevard du Triomphe, B-1050 Bruxelles, Belgium

E-mail: `jniemczy@ulb.ac.be`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

OPTIMALITY OF TWO-STAGE HYPOTHESIS TESTS

Andrei Novikov

Key words: Hypothesis testing, two-stage test, optimal decision, local asymptotic normality.

COMPSTAT 2004 section: Optimization.

Abstract: This paper deals with optimal two-stage tests for two simple hypotheses. The structure of both the optimal decision rule and the optimal continuation rule is given. The results are applied to the optimal two-stage tests for a Wiener process with a linear drift, and to obtain an asymptotically optimal test for two close hypotheses in the case of locally asymptotically normal statistical experiment. The numerical results of comparison between the optimal Neyman-Pearson test, Wald's SPRT and the proposed optimal two-stage test are given.

1 The structure of an optimal two-stage test

In this section, we give the structure of an optimal two-stage test for two simple hypotheses.

Let us assume that we can observe in a statistical experiment a random variable X (the first stage of the experiment), and, depending on it, either stop at the first stage or get to a second stage, obtaining an additional portion of observations Y. In both cases we have to take a final decision about the distribution from which X and Y come. This type of experiment can be thought of as an alternative to fixed-size sampling, as in the Neyman-Pearson test, and to completely sequential tests like the Wald's sequential probability ratio test (SPRT).

Let us assume that the vector (X, Y) follows a parametric distribution with a probability density function $f_\theta(x, y)$ with respect to a product-measure $\mu_1 \times \mu_2$ on the space of values of (X, Y), so $f_\theta(x) = \int f_\theta(x, y) d\mu_2(y)$ being the marginal density function of the first-stage component X with respect to μ_1.

For two simple hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ let us define a test as a triplet of measurable functions $(\phi_1(x), \phi_2(x, y), \chi(x))$, all of them taking values in $[0, 1]$, interpreting them as follows: $\phi_1(x)$ being the conditional probability, given a first-stage observation x, to reject H_0, $\phi_2(x, y)$ the conditional probability, given observations up to the second stage (x, y), to reject H_0, and $\chi(x)$ being the conditional probability, given the first-stage observation x, to get to the second stage (to continue sampling).

So the power function of the test will be defined as

$$P(\theta) = E_\theta \left[\phi_1(X)(1 - \chi(X)) + \phi_2(X, Y)\chi(X)\right]$$

(the total probability to reject H_0 given θ).

We are interested in minimizing $P(\theta_0)$ and $1 - P(\theta_1)$ which are, respectively, the error probabilities of the first and the second kind, and some quantities related to a cost of observations. As the first stage is always present, the only variable part is related to $C(\theta) = E_\theta\chi(x)$, which is the probability of continuing observations up to the second stage, given θ.

As usual in statistical hypotheses testing, we start from a sort of Bayesian set-up: we will be interested in finding tests which minimize the average total loss (ATL):

$$\pi_0 P(\theta_0) + \pi_1(1 - P(\theta_1)) + \pi_0 c_0 C(\theta_0) + \pi_1 c_1 C(\theta_1)$$

where π_0 and π_1 can be interpreted as prior probabilities of H_0 and H_1, respectively, and c_0 and c_1 some constants giving some weight to any of the two average observation costs measured by $C(\theta_0)$ and $C(\theta_1)$.

Let a^- be equal to a, if $a < 0$, and $a^- = 0$ otherwise, and let $I(A)$ be the indicator function of the event A.

The following theorem gives the structure of the test with the minimum ATL.

Theorem 1. *The minimum average total loss is equal to* π_1+

$$\int \left[l_1(x)^- + \left(\int l_2(x,y)^- d\mu_2(y) - l_1(x)^- + \pi_0 c_0 f_{\theta_0}(x) + \pi_1 c_1 f_{\theta_1}(x) \right)^- \right] d\mu_1(x)$$

where $l_1(x) = \pi_0 f_{\theta_0}(x) - \pi_1 f_{\theta_1}(x)$, $l_2(x,y) = \pi_0 f_{\theta_0}(x,y) - \pi_1 f_{\theta_1}(x,y)$, *and this minimum is achieved by a test with*

$$\begin{aligned} \phi_1(x) &= I(\{l_1(x) < 0\}) \\ \phi_2(x,y) &= I(\{l_2(x,y) < 0\}) \qquad (1) \\ \chi(x) &= I(\{\int l_2(x,y)^- d\mu_2(y) - l_1(x)^- + \pi_0 c_0 f_{\theta_0}(x) + \pi_1 c_1 f_{\theta_1}(x) < 0\}) \end{aligned}$$

Proof. For any test $(\phi_1(x), \phi_2(x,y), \chi(x))$ let us represent the ATL$-\pi_1$ as

$$\begin{aligned} &\int l_1(x)\phi_1(x)(1 - \chi(x)) d\mu_1(x) \\ &+ \int \left(\int l_2(x,y)\phi_2(x,y)\chi(x) d\mu_2(y) \right) d\mu_1(x) \qquad (2) \\ &+ \int (\pi_0 c_0 f_{\theta_0}(x) + \pi_1 c_1 f_{\theta_1}(x))\chi(x) d\mu_1(x) \end{aligned}$$

The first term in (2) is greater or equal than

$$\int l_1(x) I(\{l_1(x) < 0\})(1 - \chi(x)) d\mu_1(x) \qquad (3)$$

because $l_1(x)(\phi_1(x) - I(\{l_1(x) < 0\}))(1 - \chi(x)) \geq 0$ for any $0 \leq \phi_1(x) \leq 1$ (this is an almost literal repetition of the proof of the Neyman-Pearson's theorem).

The second term in (2) is greater or equal than

$$\int\left(\int l_2(x,y)I(\{l_2(x,y)<0\})\chi(x)d\mu_2(y)\right)d\mu_1(x), \tag{4}$$

because $l_2(x,y)(\phi_2(x,y)-I(\{l_2(x,y)<0\}))\chi(x)\geq 0$ for any $0\leq\phi_2(x,y)\leq 1$.

So from (2-4) we have that the (2) is greater or equal than

$$\begin{aligned}&\int l_1(x)I(\{l_1(x)<0\})(1-\chi(x))d\mu_1(x)\\&+\int\left(\int l_2(x,y)I(\{l_2(x,y)<0\})\chi(x)d\mu_2(y)\right)d\mu_1(x)\\&+\int(\pi_0c_0f_{\theta_0}(x)+\pi_1c_1f_{\theta_1}(x))\chi(x)d\mu_1(x)\end{aligned} \tag{5}$$

which is equal to

$$\begin{aligned}&\int l_1(x)^-d\mu_1(x)\\&+\int\left(\int l_2(x,y)^-d\mu_2(y)-l_1(x)^-+\pi_0c_0f_{\theta_0}(x)+\pi_1c_1f_{\theta_1}(x)\right)\\&\chi(x)d\mu_1(x)\end{aligned}$$

and in the same way as above this is greater or equal than

$$\begin{aligned}&\int l_1(x)^-d\mu_1(x)\\&+\int\left(\int l_2(x,y)^-d\mu_2(y)-l_1(x)^-+\pi_0c_0f_{\theta_0}(x)+\pi_1c_1f_{\theta_1}(x)\right)^-d\mu_1(x)\end{aligned}$$

which proves the first affirmation of Theorem 1.

The second one is immediate in view of the above proof (any step in it does not take to an inequality if the functions ϕ_1, ϕ_2 and χ are defined as in (1)), so test (1) is the optimal one.

2 Testing hypotheses about a drift of a Wiener process

2.1 The structure of the optimal two-stage test

Let us assume that we observe a Wiener process with a linear drift $W(t)+\theta t$. Without loss of generality we can assume that $W(t)$ is standard and that we are interested in testing the null hypotheses that $\theta=\theta_0=0$, taking as the alternative some $\theta\neq 0$.

At the first stage of the experiment, we observe the process up to a time t_1, keeping observing, if necessary, a time t_2 more at the second stage.

So, in terms of the above section, denoting by $\varphi(x)$ the standard normal probability density function we have:

$$f_\theta(x) = f_\theta^1(x) = \frac{1}{\sqrt{t_1}}\varphi(\frac{x-\theta t_1}{\sqrt{t_1}}), \quad f_\theta(x,y) = f_\theta^1(x)f_\theta^2(y)$$
$$f_\theta^2(y) = \frac{1}{\sqrt{t_2}}\varphi(\frac{x-\theta t_2}{\sqrt{t_2}})$$

the two components (X,Y) being independent.

By Theorem 1, for any given $\pi_0, \pi_1, c_0, c_1, t_1, t_2$ the optimal two-stage test is given by

$$\begin{aligned} \phi_1(x) &= I(\{Z_1(x) > \pi_0/\pi_1\}) \\ \phi_2(x,y) &= I(\{Z_1(x)Z_2(y) > \pi_0/\pi_1\}) \\ \chi(x) &= I(\{E\{(\pi_0 - \pi_1 Z_1(X)Z_2(Y))^-|X=x\} \\ &\quad -(\pi_0 - \pi_1 Z_1(x))^- + \pi_0 c_0 + \pi_1 c_1 Z_1(x) < 0\}) \end{aligned}$$

where $Z_1(x) = \exp(x\theta - \theta^2 t_1/2)$, $Z_2(y) = \exp(y\theta - \theta^2 t_2/2)$ are so-called likelihood ratios:

$$Z_1(x) = \frac{f_\theta^1(x)}{f_0^1(x)}, \quad Z_2(x) = \frac{f_\theta^2(y)}{f_0^2(y)}.$$

It is easy to see that the continuation rule is based on the function

$$g(z) = E_0(\pi_0 - \pi_1 z Z_2(Y))^-$$

and is equivalent to proceed to the second stage if and only if $z = Z_1(X)$ is such that

$$g(z) < (\pi_0 - \pi_1 z)^- - \pi_0 c_0 - \pi_1 c_1 z. \tag{6}$$

It is easy to observe that the function $g(z)$ is concave, and the right-hand side of (6) is piece-wise linear and concave, too. So if there are z satisfying (6), it is equivalent to $a < z < b$ with some a, b such that $a < \pi_0/\pi_1 < b$.

Due to this fact, it is obvious that the optimal two-stage test has the form:

$$\begin{aligned} \phi_1(x) &= I(\{Z_1(x) > \pi_0/\pi_1\}) \\ \phi_2(x,y) &= I(\{Z_1(x)Z_2(y) > \pi_0/\pi_1\}) \\ \chi(x) &= I(\{a < Z_1(x) < b\}), \end{aligned} \tag{7}$$

or, in other words, the optimal rule says:

1. Observe X. Stop observations at this stage if $Z_1(X) < a$ (accepting H_0), or if $Z_1(X) > b$ (rejecting H_0); continue observing otherwise.
2. At the second stage, obtain Y. Accept H_0 if

   $$Z_1(X)Z_2(Y) < \pi_0/\pi_1$$

   and reject it otherwise.

It is interesting to note that of c_0 and/or c_1 (the costs of additional observations) are sufficiently large, there are no solutions to (6), so the optimal test will stop at the first stage.

2.2 Comparison between optimal tests

Let us now pose a more realistic problem in relation with the two-stage tests.

Let us suppose that t_1 and t_2 are not fixed in advance, but are to be sought in order to minimize the average total loss of the form

$$\pi_0 P(\theta_0) + \pi_1(1 - P(\theta_1)) + \pi_0 c_0 N(\theta_0) + \pi_1 c_1 N(\theta_1), \tag{8}$$

say, where $N(\theta) = t_1 + t_2 C(\theta)$ is the average "sample number" in the two-stage experiment, given θ.

For any fixed t_1 and t_2 the solution is given by the above test, and the problem turns to be essentially numerical: to find (t_1, t_2) giving a minimum to (8) using the optimal test in the form of (7), say. It is obvious that test (7) has four parameters (t_1, t_2, a, b), and the minimum ATL in (8) can be calculated minimizing over all of them. Properly saying, parameters a and b are uniquely defined by (6) for any t_1, t_2, but there is no explicit way to calculate them, so we prefer to optimize over them as well, which is an equivalent procedure due to the results above.

We developed a program module (unit, in terms of Borland Pascal 6.0) for numerical optimization of (8), given any π_0,π_1,c_0,c_1. This module is available from the author.

Below we present some results of evaluation of optimal tests.

The most appropriate context of such an evaluation seems to be a comparison between different competing tests including the optimal two-stage test above.

So we compare the optimal two-stage test with the classic Neyman - Pearson and Wald's test, similar to [1]. Obviously, an optimal test (7) with $P(\theta_0) = \alpha$ and $1 - P(\theta_1) = \beta$, would minimize

$$\pi_0 c_0 N(\theta_0) + \pi_1 c_1 N(\theta_1)$$

among all the (two-stage) tests with error probabilities of the first and second kind not exceeding α and β, respectively. So it is interesting to compare its average sample number(ASN) $N(\theta_0)$ and $N(\theta_1)$ with the ASN of the Neyman-Pearson and Wald's test (see [1], see also [4]) .

The following table contains the respective characteristics of the three competing tests for a series of $\alpha = \beta$ evaluated for the null hypothesis $\theta = 0$ against the alternative $\theta = 1$. The numbers in the respective columns are the ASN of the three tests which correspond to the same level of $\alpha = \beta$ indicated in column "α".

The results above give a very clear evidence that two-stage tests have rather competitive properties concerning the average sample number.

The following table gives an idea about the parameters of the respective optimal two-stage test. Because $Z_1(x)$ and $Z_2(y)$ in (7) are monotone functions of x and y respectively, the parameters of the two-stage test are given in terms of x and y rather than in terms of Z_i. For example, to

α	**Wald**	**Neyman- Pearson**	**Two-Stage**
0.0072	9.71	23.94	15.17
0.0150	8.12	18.83	12.48
0.0231	7.15	15.91	10.84
0.0313	6.44	13.87	9.66
0.0396	5.87	12.32	8.73
0.0480	5.40	11.09	7.97
0.0563	5.00	10.07	7.33
0.0646	4.65	9.20	6.77
0.0729	4.34	8.46	6.28
0.0811	4.07	7.81	5.85
0.1563	2.32	4.08	3.22
0.2161	1.46	2.47	2.00

Table 1: Average sample number.

achieve $\alpha = \beta = 0.0072$, you have first to observe the process up to the time 11.54 (t_1), then if the value x of the process at that time is less than 2.58 (a), then accept H_0 and stop observing. If x is greater than 8.96 (b) then stop observing as well, and reject H_0. Otherwise keep observing for 16.53 time units more (t_2), obtaining the value y of the process at the end of this period. Based on this, accept H_0 if $y < 14.03$ (c) and reject H_0 otherwise.

α	a	b	c	t_1	t_2
0.0072	2.58	8.96	14.03	11.54	16.53
0.0150	2.01	7.29	11.08	9.30	12.87
0.0231	1.66	6.31	9.38	7.97	10.79
0.0313	1.41	5.62	8.20	7.03	9.36
0.0396	1.22	5.09	7.29	6.31	8.28
0.0480	1.06	4.66	6.57	5.72	7.42
0.0563	0.93	4.30	5.97	5.23	6.72
0.0646	0.82	3.99	5.47	4.81	6.12
0.0729	0.72	3.72	5.03	4.44	5.62
0.0811	0.64	3.49	4.65	4.12	5.18
0.1563	0.16	2.06	2.45	2.22	2.67
0.2161	-0.03	1.39	1.49	1.36	1.61

Table 2: The optimal two-stage test.

3 Asymptotically optimal two-stage tests for LAN experiments

In this section we will show how the results of the previous section can be applied to construct asymptotically optimal tests for a rather broad class of locally asymptotically normal experiments (LAN).

Let us say that a statistical experiment $\{X_1, X_2 \dots X_n\}$ with independent and identically distributed observations is locally asymptotically normal if for any $\epsilon > 0$ there exists $n = n(\epsilon)$ such that the likelihood ratio for two simple hypotheses θ and $\theta + \epsilon$

$$Z_\epsilon^n = \prod_{i=1}^n f_{\theta+\epsilon}(X_i)/f_\theta(X_i)$$

converges weakly, when $X_1, \dots X_n$ follow the distribution with the parameter θ, to that of two normal distributions:

$$Z = \exp\{\xi - 1/2\},$$

where ξ is a standard normal random variable (cf., e.g., [2]).

The aim of this section is to construct a test of $H_0 : \theta$ vs $H_1 : \theta + \epsilon$ with error probabilities α and β which asymptotically minimizes a weighted average sample number, as $\epsilon \to 0$.

Theorem 2. *Let π_0, π_1, c_0,c_1 be such numbers that there exists a two-stage test (7) minimizing (8) with $P(\theta_0) = \alpha$ and $1 - P(\theta_1) = \beta$. Then the two-stage test taking $n_1 = [t_1 n(\epsilon)]$ observations at the first stage, and additional $n_2 = [t_2 n(\epsilon)]$ observations at the second stage and defined as*

$$\begin{array}{rcl} \phi_1 & = & I(\{Z_\epsilon^{n_1} > \pi_0/\pi_1\}) \\ \phi_2 & = & I(\{Z_\epsilon^{n_1+n_2} > \pi_0/\pi_1\}) \\ \chi & = & I(\{a < Z_\epsilon^{n_1} < b\}) \end{array} \tag{9}$$

is asymptotically optimal in the sense that it minimizes

$$\lim_{\epsilon \to 0} (\pi_0 c_0 N_\epsilon(\theta) + \pi_1 c_1 N_\epsilon(\theta + \epsilon))/n(\epsilon)$$

in the class of all two-stage tests whose error probabilities of the first and the second kind asymptotically do not exceed α and β, respectively.

Proof is rather straightforward if we note that due to the LAN condition and independence of the observations the distributions of $Z_\epsilon^{n_1}$ and $Z_\epsilon^{n_2}$ defining test (9) converge weakly to the distribution of Z_1 and Z_2 in test (7), so its error probabilities converge to that of test (7), and so the continuation probability, and thus the average sample number $N_\epsilon(\theta)$ of the test in Theorem 2 normalized by $n(\epsilon)$ tends to $N(\theta_0)$ of test (7). The rest of the proof is due to the optimality of (7).

Another promising application of two-stage tests seems to be the construction of a test similar to that of Theorem 2 for statistical experiments with Markov dependent observations (see [3]), in which case the likelihood ratio behaves exactly the same way as in the case of independent observations. Unfortunately, a proof as in Theorem 2 does not proceed, because for dependent observations the structure of continuation rule is not as simple as in (7) any more. So the problem of finding an optimal sequential test for non-independent observations is still open even in the simplest case of two-stage tests.

References

[1] Aivazjan S.A. (1959). *A comparison of the optimal properties of the Neyman-Pearson and the Wald sequential probability ratio test.* Theory Prob. Appl. **4**, 86 – 93.

[2] Le Cam L. (1986). *Asymptotic methods in statistical decision theory.* Springer Series in Statistics. Springer-Verlag, New York-Berlin.

[3] Novikov A. (2001). *Uniform asymptotic expansion of likelihood ratio for Markov dependent observations.* Ann. Inst. Statist. Math. **53** (4), 799 – 809.

[4] Novikov A. (2002). *Efficiency of sequential hypotheses testing.* Aportaciones matemáticas. Serie Comunicaciones **30** 71 – 79.

Acknowledgement: My deep gratitude to Mexico and its wonderful people.

Address: A. Novikov, Universidad Autonoma Metropolitana - Unidad Iztapalapa, Departamento de Matematicas, San Rafael Atlixco #186, col. Vicentina, C.P. 09340, Mexico D.F., Mexico

E-mail: an@xanum.uam.mx

MODELLING RESIDUALS IN DYNAMIC REGRESSION: AN ALTERNATIVE USING PRINCIPAL COMPONENTS ANALYSIS

Francisco M. Ocaña-Peinado and Mariano J. Valderrama

Key words: Transfer function model, ARIMA model, principal component analysis, Standard & Poor's 500, IBEX 35.

COMPSTAT 2004 section: Time series analysis.

Abstract: In this paper IBEX 35 time series is represented as a transfer function model using the Standard & Poor's 500 index as input series by two different ways: first (as usually), by an ARIMA modelisation of the residuals, and secondly residuals are modeled using principal component analysis (PCA). Smoothing and predictions of IBEX 35 time series in the second form, will be more accurate than the ones using usual ARIMA models for the residuals.

1 Introduction

Classical regression models are often inadequate for representing relations between time series, fundamentally for two reasons:

- Only instantaneous relations are supposed, ignoring possible dynamic relations, that is, past relations between dependent and independent variables are not taken into account .
- The non-explained part is represented by white noise. Therefore, a possible time structure for the residuals is neglected.

Dynamic Regression Models, also known as transfer function models (TFM), solve these two problems; giving to the non explained part of the model, a time structure, and capturing the dynamic relations between time series. Since Box and Jenkins [1] introduced TFM, these models have been used mainly in the Economics field.
The Box-Jenkins modelling of the dynamic regression linear models is briefly described as follows. A complete analysis of these models is developed by Box and Jenkins [1], Brockwell and Davis [2] and Pankratz [4].

Let Y_t be the time series, called output, which we try to forecast in terms of other time series, called input, and that we denote as X_t. The linear model to be developed, taking into account possible dynamic relations between input and output, is:

$$Y_t = \upsilon_0 X_t + \upsilon_1 X_{t-1} + \upsilon_2 X_{t-2} + ... + N_t = \upsilon(B) X_t + N_t$$

where B is the backshift operator and the inertia process N_t, represents the non-explained part of Y_t, and will have an $ARIMA$ time structure as follows:

$$\nabla^d N_t = [\theta(B)/\phi(B)]a_t$$

where a_t is zero-mean white noise with variance σ_a^2.
Construction of TFM has three stages in the same way as ARIMA models described by Box and Jenkins.

In the application (section 3), we will study the correlation between the behavior of the European markets and the United States one. To study this phenomenon, it will be modeled the main Spanish index, IBEX 35, from the variations of the Standard & Poor's 500 index of New York (S&P 500).

2 Modelling residuals by PCA

In the previous section the classical modelling of the residuals is presented, using ARIMA models. In this section, we propose to apply the PCA to obtain an alternative model of them. The objective will be to explain the variability in the residual series by means of the components.
We start from a set of time observations of the input series as well as of the output series. From these, following the described model in last section, the transfer function $\widehat{\upsilon}(B)$ is estimated.
The estimated inertia process $\widehat{N}_t$ will not be more than the time series obtained as difference between the output series Y_t and the estimation of the transfer function $\upsilon(B)$:

$$\widehat{N}_t = Y_t - \widehat{\upsilon}(B)X_t$$

Supposing this process as stationary, a principal components analysis (PCA) is carried out, for which is considered the time series structured as r realizations of itself in h different moments of time as shows the Table 1.

	$t=1$	$t=2$	$\dots$	$t=h$
$\omega=1$	N_{11}	N_{12}	$\dots$	N_{1h}
$\omega=2$	N_{21}	N_{22}	$\dots$	N_{2h}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\omega=p$	N_{p1}	N_{p2}	$\dots$	N_{ph}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\omega=r$	N_{r1}	N_{r2}	$\dots$	N_{rh}

Table 1: Residual series tabulated for applying PCA.

Thus, we represent the time series in each one of its r sample-paths, as linear combination of the eigenvectors associated to its components:

$$\widehat{\mathbf{N}}(\omega) = \sum_{i=1}^{h} \mathbf{u}_i \xi_i(\omega) \tag{1}$$

where ξ_i are the principal components, and $\underline{u}_i$ denote the eigenvectors associated to them.
The proportion of the total variability of the inertia process explained by the i-*th* component ξ_i, is the quotient between its associated eigenvalue, and the total variance of the process. Besides, the proportion of variance accumulated by the k principal components with highest variance is the sum of the proportions of the total variance explained by each one of them, due to the uncorrelated character of the ξ_i.

So, we can approximate the process in an optimal way, in function of the k first principal components:

$$\widehat{\mathbf{N}} \cong \sum_{i=1}^{k} \mathbf{u}_i \xi_i$$

for which the time model to predict the output in terms of the input approximation with the k first principal components, will have the following expression:

$$y_t = \widehat{v}(B) x_t + \sum_{i=1}^{k} \mathbf{u}_i \widehat{\xi_i} \tag{2}$$

We refer to it as *TF-PCA* model.

3 Application

In this section we describe the derivation and application of the TF-PCA model in the stock market, and its comparison with the classical TFM. We have studied the behavior of IBEX 35 and S&P500 during almost a complete year. The period was since the 10th of May 2002 until the 7th of May 2003 (total number of observations is 250).

3.1 Obtaining the models

The existence of some linear correlation between both indexes during this period is showed in Figure 1. The first step in the modelling procedure, is to find the linear transfer function (TF) that relates the output series (IBEX 35), in terms of the input series (S&P500). This TF, obtained with the program ITSM for Windows (2000), was the following one:

$$y_t = 4.5742 x_t + 2.6226 x_{t-1} \tag{3}$$

where y_t is the IBEX 35 and x_t is the S&P500.

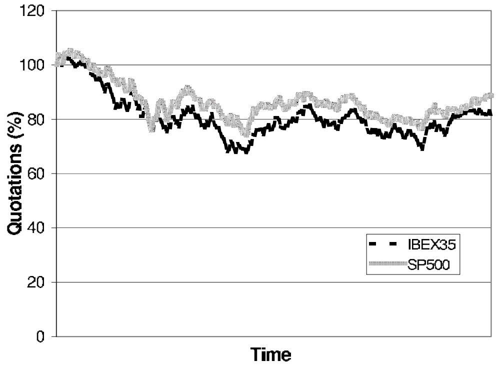


Figure 1: Quotations of both stock market indexes during the study.

The next step is to model the residuals in two ways: as usually (ARIMA model), and using PCA (in order to obtain the FT-PCA model). In the first way, the next ARIMA(0,1,1) model was found for the residuals:

$$\nabla \widehat{N}_t = n_t = a_t + 0.296 a_{t-1} \tag{4}$$

where a_t is zero-mean white noise with $\sigma_a^2 = 10073.4$. From equations (3) and (4) we have the following TFM in the classical sense:

$$y_t = 4.5742 x_t + 2.6226 x_{t-1} + n_t \tag{5}$$

For the FT-PCA model, a PCA was applied to the residual series once it was divided with $r = 50$ and $h = 5$ (see Table 1). The Table 2 reflects the percentage of total variance explained by the principal components:

ξ_i	Eigenvalue	Percentage of variance	Cumulative Percentage
1	4.12682	82.536	82.536
2	0.446848	8.937	91.473
3	0.189644	3.973	95.266
4	0.131522	2.630	97.897
5	0.105164	2.103	100

Table 2: Results of PCA for residual series.

The two first principal components are selected, that explain more than the 90% of the variability of the inertia process. So, the TF-PCA model given in (2) has the next form:

$$y_t = \widehat{v}(B) x_t + \mathbf{u}_1 \widehat{\xi}_1 + \mathbf{u}_2 \widehat{\xi}_2 \tag{6}$$

where the eigenvectors obtained in the PCA are respectively the following ones:

$$\mathbf{u}_1 = (0.423292, 0.456529, 0.462727, 0.450288, 0.442188)'$$

$$\mathbf{u}_2 = (-0.695175, -0.340398, 0.125934, 0.456328, 0.420436)'$$

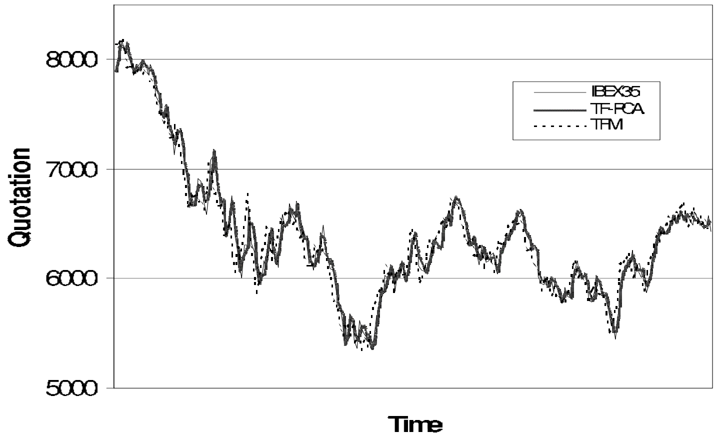


Figure 2: Smoothing of IBEX 35 with TFM and TF-PCA models.

3.2 Smoothing and predictions

In this section, we are going to compare the goodness of the classical TFM against the TF-PCA model in terms of smoothing and predictions.

3.2.1 Smoothing From the model given by (5) a smoothing for the output series was realized. The first step is to use (3) for smoothing the output, and the second step is to smooth residuals with the ARIMA model gives by (4), obtaining therefore the residual smoothing. Adding both smoothed structures (output with TF, and residuals), the smoothing is completed.

Using an analogous strategy for the TF-PCA model, a smoothing of the residuals has been made from $\widetilde{\xi_1}$ and $\widehat{\xi_2}$ and their respective associated eigenvectors. Note, that for each ω we have an estimation for each component, and we use them for smoothing the output series.

The results of the smoothing is showed in Figure 2. Table 3 shows both mean square errors (MSE).

Model	MSE
TFM	9974.61
TF-PCA	2973.22

Table 3: MSE obtained by smoothing with both models.

3.2.2 Predictions We forecast the output series IBEX 35 with the TFM and TF-PCA model for a week (five consecutive days in the stock market, starting in the first day after the last day included in order to obtain both models).

Predictions with TFM model are made in the usual form, that is, from the model given in (5) and with the values of the input series, S&P500 in this period. The MSE obtained for this model is 18792.45.

For the TF-PCA model, we have to predict the value of ξ_1 and ξ_2 for the period $\omega = 51$. As we already said previously in section 3.2.1, we have the sample values of ξ_1 and ξ_2 for $\omega = 1, 2, \ldots, 50$. The strategy in order to obtain an estimation of the principal components for $\omega = 51$, is to consider ξ_1 and ξ_2 as linear functions of the average residuals ($\overline{n}_t$), in the previous period. Formally:

$$\widehat{\xi}_i(r+1) = a_i + b_i \overline{n}_t(r) \qquad i = 1, 2.$$

So, the linear functions obtained for each component are:

$$\widehat{\xi}_1 = -144.89 - 1.5163\overline{n}_t \qquad \widehat{\xi}_2 = -55.5146 - 0.4079\overline{n}_t \tag{7}$$

Expressions given in (7) provide estimations of each component, that will be used in order to predict input series from the model given in equation (6). The MSE obtained using this strategy is 4999.98, that indicates that it has been reduced significantly with PCA for residuals with respect to the classical Box-Jenkins TFM. The Table 4 shows the real value of IBEX 35 and the predictions for it with the TF-PCA model:

Time	IBEX 35	Predictions TF-PCA
8th of May	6300.5	6334.7
9th of May	6387.8	6357.5
12th of May	6395.0	6441.1
13th of May	6376.0	6481.0
14th of May	6363.4	6462.1

Table 4: Predictions with TF-PCA for the input series.

References

[1] Box G.E.P., Jenkins G.M. (1976). *Time series analysis: forecasting and control.* Holden Day, 2nd edition, San Francisco.

[2] Brockwell P.J., Davis R.A. (1991). *Time series: theory and methods.* Springer, New York.

[3] Brockwell P.J., Davis R.A. (2000). *Time series: ITSM for Windows.* Springer, New York.

[4] Pankratz A. (1991). *Forecasting with Dynamic Regressions Models.* Wiley, New York.

Acknowledgement: This paper was supported by Project BFM2000-1466 of the Ministerio de Ciencia y Tecnología, Spain, and the Research Group FQM-307 of the Consejería de Educación y Ciencia de la Junta de Andalucía.

Address: F.M. Ocaña-Peinado, M.J. Valderrama, Department of Statistics and Operations Research, University of Granada, 18071-Granada, Spain

E-mail: `fmocan@ugr.es valderra@ugr.es`

STATE-SPACE MODEL FOR SYSTEM WITH NARROW-BAND EXCITATIONS

Monica Ortega-Moreno and Mariano J. Valderrama

Key words: Narrow-band process, state space model.

COMPSTAT 2004 section: Functional data analysis.

Abstract: The objective of this paper is to study dynamic systems with random narrow-band external excitations, by applying the properties of such a process, and derive a state space model in this situation.

1 Introduction

The problem of working with dynamic systems in continuous time has a great interest in several applied areas, mainly in communication and detection. These ones are defined in terms of stochastic differential equations with an input term.

In many times, the input of a dynamic system is the result of one or more narrow-band processes. These cases are of practical importance. Some recent studies concerning the response of a system to random narrow-band excitations can be found in Rong et al. [4] and Arena et al. [1], [2].

In this paper, narrow-band random external excitations are investigated and a state-space model (SSM) is proposed, on the basis of the properties of these processes.

The paper is structured as follows: Section 2 contains the formulation of the problem and shows different narrow-band processes, as well as linear combinations. In Section 3 a state-space model is proposed when the excitation of the system is a combination of narrow-band processes. Finally, in Section 4, we present a comparative study of forecasting with applications from a non stationary process, the Brownian motion, affected by different random narrow-band external excitations.

2 Formulation of the problem

Consider a dynamic system under random external excitations, derived by the following time dependent equation

$$\dot{X}(t) = F(t)X(t) + G(t)U(t) + W(t) \tag{1}$$

where dots indicate differentiation with respect to the time t; $X(t)$ is the n dimensional state vector, $U(t)$ represents the k dimensional input or system excitation in t, $F(t)$ and $G(t)$ are the system matrixes, with dimensions $n \times n$ and $n \times k$ respectively, and $W(t)$ is a n dimensional error vector distributed as a white noise with zero mean and whose variance-covariance is non-negative defined.

We suppose that the external random excitation,

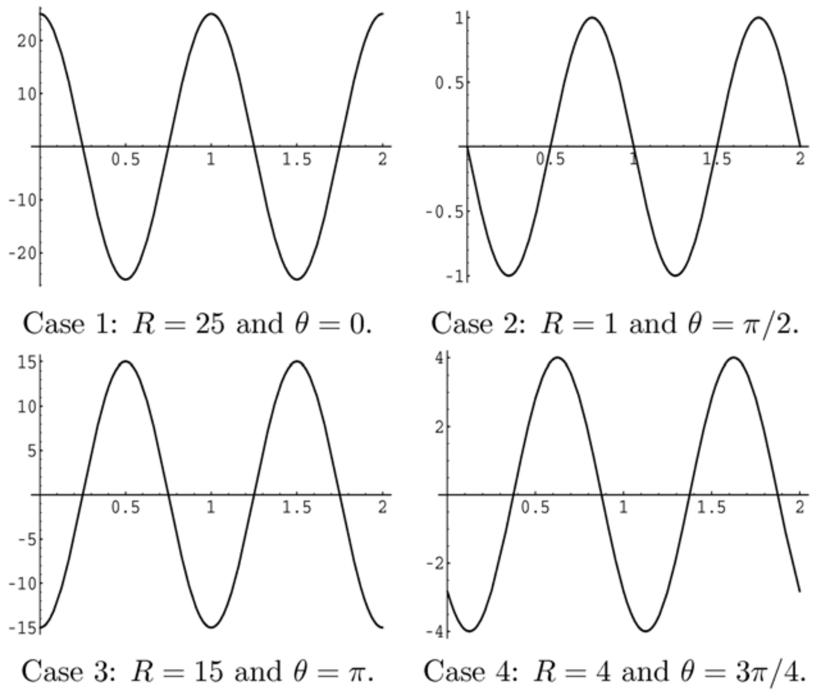
$$U(t) = \begin{bmatrix} u_1(t) & u_2(t) & \dots & u_k(t) \end{bmatrix}^T$$

is composed of narrow-band stochastic processes, so that:

$$u_i(t) = R_i \cos[2\pi t + \theta_i], \qquad \forall \quad 0 \le t \le T.$$

where θ and R are independent variables. The variable θ is called phase of the oscillator and it has uniform distribution on $[0, T]$. On the other hand, a necessary and sufficient condition for this process to be Gaussian is that the amplitude variable R has a Rayleigh distribution.

Figure 1 shows narrow-band processes for different values of amplitude R and phase θ.



Case 1: $R = 25$ and $\theta = 0$.

Case 2: $R = 1$ and $\theta = \pi/2$.

Case 3: $R = 15$ and $\theta = \pi$.

Case 4: $R = 4$ and $\theta = 3\pi/4$.

Figure 1: Narrow-band processes for different values of amplitude R and phase θ.

We can observe from equation (1) that the input term is affected by a time dependent matrix $G(t)$, so that it has different behaviours. Figure 2 shows some examples.

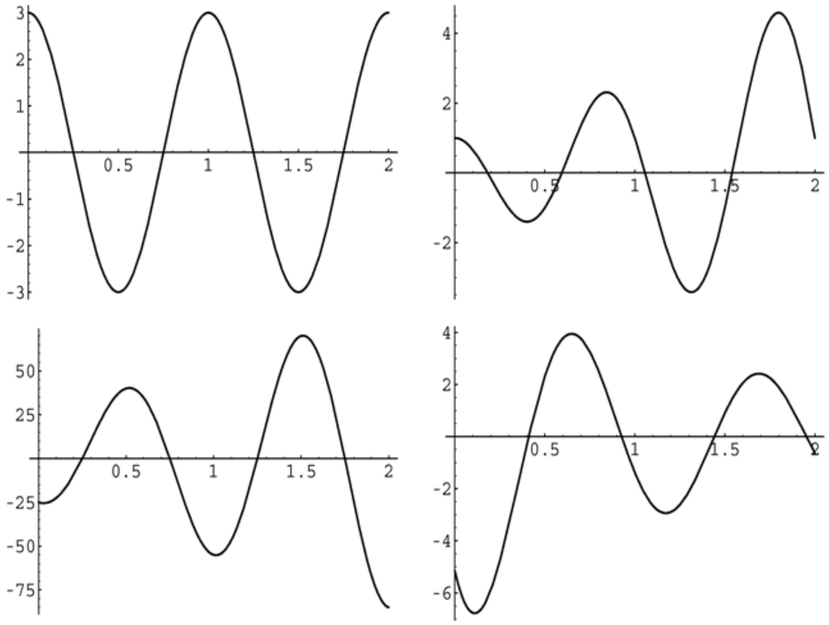


Figure 2: Results of combinations of narrow-band processes.

3 Derivation of a SSM

To obtain a SSM from equation (1) we take into account that:

$$\ddot{U}(t) = -4\pi^2 U(t).$$

By considering the new $(n+2k)$ dimensional state vector:

$$Z(t) = \left[\begin{array}{ccc} X(t) & U(t) & \dot{U}(t) \end{array} \right]^T$$

we can derive the state equation of the model:

$$\dot{Z}(t) = \left[\begin{array}{ccc} F(t) & G(t) & 0 \\ 0 & 0 & I \\ 0 & -4\pi^2 I & 0 \end{array} \right] Z(t) + \left[\begin{array}{c} W(t) \\ 0 \\ 0 \end{array} \right] \tag{2}$$

On the other hand, let us suppose that it is not possible to observe the signal process $Z(t)$ in a direct way, but it is corrupted by a vector of zero-mean Gaussian white noise $V(t)$, so that the measurement equation is:

$$Y(t) = \left[\begin{array}{ccc} I & 0 & 0 \end{array} \right] Z(t) + V(t) \tag{3}$$

where H is the measure matrix, with dimension $1 \times (n+2k)$, and the variance-covariance matrix of $V(t)$ is positive defined. Then a not invariant time SSM is given by equations (2) and (3).

Let us observe that an optimum estimation of the random input is obtained through the Kalman-Bucy filter at the same time as the optimum estimation of the process $X(t)$. It is possible because we have increased the dimensions of the system from n to $n + 2k$.

4 Applications

In order to evaluate from a practical point of view the behaviour of the proposed model we are going to simulate a sample path of standard Brownian motion on [0,2]. Then we have derived a minimal dimension SSM dependent on the time from the functional Principal Component Analysis (PCA) of the stochastic process, once their original sample-paths have been approximated by a suitable basis of B-splines functions, as is developed in Ortega-Moreno et al. [3] and Valderrama et al. [5].

After that we have excited the Brownian motion by different narrow-band random external excitations and have obtained the results of the optimal filter with the SSM given by equations (2) and (3).

Once the parameters of the system and initial conditions are specified it is possible to obtain the Kalman-Bucy estimate. Figure 3 shows the Brownian motion excited by a narrow-band process superposed with the forecasts performed by the filter with our model.

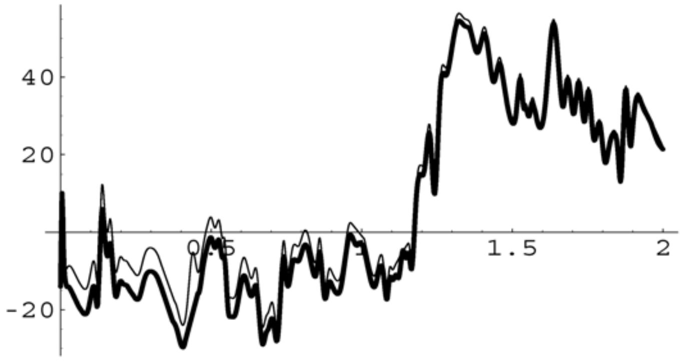


Figure 3: Sample path of a Brownian motion excited by a narrow-band process (thin line) and estimated (thick line).

Figure 4 shows the Brownian motion excited by a combination of narrow-band processes superposed with the forecasts performed by Kalman-Bucy filter with our model.

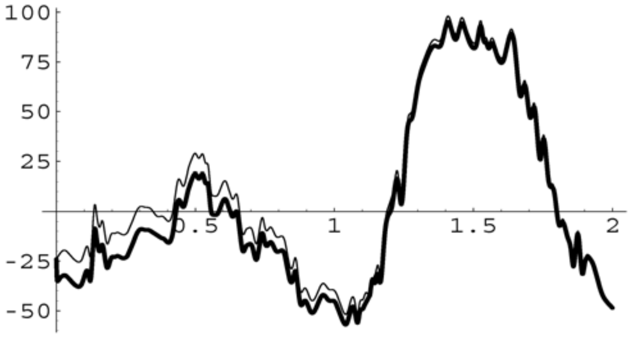


Figure 4: Sample path of a Brownian motion excited by a combination of narrow-band processes (thin line) and estimated (thick line).

We have evaluated the mean square errors (MSE) given by:

$$MSE(t) = \frac{1}{T}\int_0^T \left(Z_\omega(t) - \widehat{Z}_\omega(t)\right)^2 dt$$

associated to the predictions on the interval and the results have been shown in Table 1.

Dynamic system	**MSE(t)**
without excitation	0.0093
excited by $u_1(t) = 15\cos[2\pi t+\pi]$	31.676
excited by $u_2(t) = 25\cos[2\pi t]$	108.128
excited by $2\ t \cdot u_1(t) - u_2(t)$	107.624

Table 1: MSE associated to the predictions on the interval $[0, 2]$.

References

[1] Arena F., Fedele F. (2002). *A family of narrow-band non-linear stochastic processes for the mechanics of sea waves.* Eur.J.Mech.B.Fluids **21**, 125 – 137.

[2] Arena F., Fedele F. (2002). *Non-linear wind-generated waves forces on a vertical wall.* 15th ASCE Engineering Mechanics Conference. June 2-5, Columbia University, New York, NY.

[3] Ortega-Moreno M., Valderrama M.J., Ruiz-Molina J.C. (2002). *A staate space model for non-stationary functional data.* Compstat 2002, Physica-Verlag, Berlin, 135 – 140.

[4] Rong H.G., Meng G., Xu W., Fang T. (2003). *Response statistics of three-degree-of-freedom nonlinear system to narrow-band random excitation*. Nonlinear Dynamics **32**, 93 – 107.

[5] Valderrama M.J., Ortega-Moreno M., González P., Aguilera A.M. (2003). *Derivation of a state-space model by functional data analysis*. Computational Statistics **18**, 533 – 546.

Acknowledgement: This work was supported by the Research Group FQM307 financed by Conserjería de Educación y Ciencia de la Junta de Andalucía.

Address: M. Ortega-Moreno, Department of General Economy ans Statistics, University of Huelva, 21002-Huelva, Spain
M.J. Valderrama, Department of Statistics and Operations Research, University of Granada, 18071-Granada, Spain

E-mail: `ortegamo@uhu.es, valderra@ugr.es`

FINDING DIRECTED ACYCLIC GRAPHS FOR VECTOR AUTOREGRESSIONS

Les Oxley, Marco Reale and Granville Tunnicliffe-Wilson

Key words: Graphical models, structural VAR, causality.

COMPSTAT 2004 section: Time series analysis.

Abstract: In this paper graphical modelling is used to select a sparse structure for a multivariate time series model of New Zealand interest rates. In particular, we consider a recursive structural vector autoregressions that can subsequently be described parsimoniously by a directed acyclic graph, which could be given a causal interpretation. A comparison between competing models is then made by considering likelihood and possibly economic theory.

1 Introduction

Technology has impacted extensively on the operations of financial markets which are inhabited by a rich array of fixed-income securities, each bearing a particular rate of interest. The relationship between the yields on these various securities is the province of the term structure of interest rates literature which has a long history and can be traced-back formally to Keynes.

With the popularity of cointegration and VAR/SVAR approaches to estimation in econometrics, a separate literature using these approaches to estimate and test term structure models and implications has developed.

Here the papers are typically motivated by a concern to understand the term structure for the related monetary policy control issues and focus either upon technical estimation issues and often the validity of inferences derived including, importantly, causal inference, the effects of structural change or the testing of various hypotheses. Causality is a particularly important and popular issue given the role of monetary policy intervention.

In this paper we wish to add a significant extra dimension to the debate by using graphical modelling to identify causal mechanisms within multivariate time series models.

This paper considers an application to the term structure of interest rates where little consensus seems to exist on the causal nexus and direction between long and short rates of interest. In particular, there are three alternative views on causality; short rates *cause* long rates (broadly the traditional *Expectations Hypothesis* view); long rates *cause* short rates (here rational inflation expectations have a role); or the *market segmentation*, or *preferred habitat* approaches, where causality is discontinuous across maturity periods. The outcome in an empirical sense will be crucial for the efficacy of monetary policy design and implementation. In Section 2 of the paper the Graphical

Modelling (GM) approach will be outlined as it is a relatively new statistical approach. Section 3 will identify the relationship between the acyclic graph and more traditional multivariate structural VAR models. Section 4 will present the empirical example which relates to New Zealand interest rate data. The results from the research presented here can be used to assess (expost) empirical support for the choices made and as an example of how the GM technique can be used in practice.

2 Graphical modelling

Graphical modelling (GM) is a relatively new statistical approach, whose initial ideas were proposed by Dempster [5] and later devoleped by Darroch *et al.* [4]. The major attraction of the approach in empirical research is its ability to provide a convenient way to present pairwise relationships between random variables taken from a multivariate context.

The initial step in the approach is the computation of the partial correlations between the variables in the particular multivariate system under study. Once the numerical values are known we can test their significance by using an opportune statistic. Finally the results are presented as a graph, where the random variables are represented by nodes and a significant partial correlation between two random variables is denoted by a line that links them named *edge*. If the variables in the graph are jointly distributed as a multivariate Gaussian distribution, a significant partial correlation implies the presence of conditional dependence. For this reason the graph is called a conditional independence graph or (CIG).

A more informative object in GM is the directed acyclic graph (DAG). This is a directed graph where there are arrows linking the nodes and where the joint distribution of the variables can be expressed as a sequence of marginal conditional distributions.

Although the DAG and the CIG represent a different definition of the joint probability, there is a correspondence between the two which is embodied by the moralization rule (Lauritzen and Spiegelhalter [10]): because of this result we can obtain the CIG from the DAG by transforming the arrows into lines and linking unlinked parents with *moral* edges.

While the CIG represents the associations among the variables either in terms of conditional dependence or simply in terms of partial correlation, the DAG has a natural interpretation in terms of causality. As it is not the aim of this paper to enter into a philosophical, we just refer to some of the main contributions on the causality implied by directed acyclic graphs: Lauritzen [8], Pearl [12], Spirtes *et al.* [16], Lauritzen and Richardson [9].

The DAG is a very attractive because of its causal interpretation but in practice all we can observe is the CIG implied by the sample partial correlations. In order to obtain the DAG from the CIG we have to apply the inverse operation of the moralization, we name it *demoralization*. Unfortunately while the transformation of a DAG into a CIG is unique, the inverse operation of identification and removal of moral edges is not. To this end we need to use all the information we have about the relationships among the random variables in the system.

In this paper we apply this process within the context of multivariate structural VAR models considering first its *saturated* specification, where there are links between every pair of variables (including the contemporaneous variables), with the aim of finding a parsimonious form. As an illustration we consider the application to the transmission mechanism between interest rates in New Zealand.

3 The multivariate time series context

The relationship between several autoregressions can be modelled via the vector autoregression

$$x_t = c + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \ldots + \Phi_k x_{t-k} + e_t \tag{1}$$

of order k, VAR(k), where $x_t, x_{t-1}, \ldots, x_{t-k}$ are n-dimensional vectors with the corresponding coefficient vectors $\Phi_1, \Phi_2, \ldots, \Phi_k$, c is the constant and e_t is the error vector, which is assumed IID. If the covariance matrix, H, of e_t is not diagonal, the set of linear equations (1) corresponds to a system of seemingly unrelated regressions (Zellner [18]) where the relations among the components of x_t are hidden in H. To highlight such relations we can represent the canonical VAR(k) in (1) in its structural form (SVAR):

$$\Theta_0 x_t = d + \Theta_1 x_{t-1} + \Theta_2 x_{t-2} + \ldots + \Theta_k x_{t-k} + u_t \tag{2}$$

where $\Theta_i = \Theta_0 \Phi_i$ for $i = 0, \ldots, k$, $d = \Theta_0 c$ and $u_t = \Theta_0 e_t$ with covariance matrix $\Theta_0 H \Theta_0' = D$, which is diagonal.

If there are no zeros in the coefficient vectors, the SVAR is saturated, but in many cases some lagged variables on the RHS in (2) do not play any role in explaining the current variables, x_t. In this case the value of the corresponding coefficient is zero and hence the SVAR is sparse. An examination of the covariance matrix of the variables involved, both current and lagged, can assist in identifying the sparse structure by the computation of the partial correlations. Their significance can be tested using the appropriate sampling properties (Reale and Tunnicliffe Wilson [13] [14]). The model (2) may be represented by a directed acyclic graph (DAG) in which the components of x_t, x_{t-1}, ..., x_{t-p} form the nodes, and causal dependence is indicated by arrows linking nodes. The nature of the model is that all arrows end in nodes representing the contemporaneous variables on the left hand side of (2). Some arrows will start from past values, and some from other contemporaneous variables.

The coefficients can be estimated by single equation ordinary least squares (OLS) regression which is fully efficient under the assumption that the vector series is Gaussian but is also applicable and the properties of the estimates reliable, under wider conditions, such as e_t being I.I.D.

Next consider the exploratory tools used to identify the model. The first step is to identify the overall order p of a VAR model for the series. The second and central step is to construct a sample conditional independence graph (CIG) for the variables x_t, x_{t-1},..., x_{t-p} which form the nodes of the graph. At this stage the only causality we can assume is the one indicated by the arrow of time. Nevertheless, it may serve well to suggest the direction of dependence between contemporaneous variables. The corresponding structural VAR models are then fitted and refined by regression and a model selection criterion such as AIC, Akaike [1], used to select the best in terms of likelihood.

The statistical procedures are based on a data matrix X which in the general case consists of $m(P+1)$ vectors of length $n = N - P$, composed of elements $x_{i,t-u}$, $t = P+1-u, \dots N-u$, for each series $i = 1, 2, \dots, m$, and each lag $u = 0, 1, \dots, P$, for some chosen maximum lag P. In the first stage of overall order selection, for each order p we fit, by OLS, the saturated structural VAR regressions of the m contemporaneous (lag 0) vectors on all the vectors up to lag p. Using the sums of squares S_i from these regressions we form the AIC as $n \sum \log S_i + 2k$, where $k = pm^2 + m(m-1)/2$ is the total number of regression coefficients estimated in the regressions. For the saturated model the causal order of the contemporaneous variables does not affect the result, each one is included only as a regression variable for a subsequent variable in the chosen ordering. Then select the order p which minimizes the AIC.

The next step is to construct the sample CIG for the chosen model order p. In general a CIG is an undirected graph, defined by the *absence* of a link between two nodes if they are independent, conditional upon *all* the remaining variables. Otherwise the nodes are linked. In a Gaussian context this conditional independence is indicated by a zero partial autocorrelation:

$$\rho\left(x_{i,t-u}, x_{j,t-v} | \{x_{k,t-w}\}\right) = 0, \tag{3}$$

where the set of conditioning variables on the right is the whole set up to lag p, excluding the variables on the left.

The set of all such partial correlations required to construct the CIG is conveniently calculated from the inverse W, of the covariance matrix V of the whole set of variables, as

$$\rho\left(x_{i,t-u}, x_{j,t-v} | \{x_{k,t-w}\}\right) = -W_{rs}/\sqrt{(}W_{rr}W_{ss}) \tag{4}$$

where r and s respectively index the lagged variables $x_{i,t-u}$ and $x_{j,t-v}$ in the matrices V and W.

In the wider linear least squares context, defining linear partial autocorrelations as the same function of linear unconditional correlations as in the Gaussian context, the absence of a link still usefully indicates a lack of linear predictability of one variable by the other given the inclusion of all remaining variables.

To estimate the CIG we replace V with the sample covariance matrix $\hat{V}$ formed from the data matrix X, but including only lags up to p. From here we need a statistical test to decide which links are absent in the graph. We are only concerned with links between contemporaneous variables and between contemporaneous and lagged variables, because these are the only ones that appear in the structural model DAG. The test we use is to retain a link when $|\rho| > z/\sqrt{(}z^2 + \nu)) \approx z/\sqrt{n-p}$, where z is an appropriate critical value of the standard normal distribution. This derives from two results. The first is the standard, algebraic, relationship between a sample partial correlation $\hat{\rho}$ and a regression t value given by $\hat{\rho} = t/\sqrt{(}t^2 + \nu)$ (see Greene [6, p. 180]). The second is the asymptotic normal distribution of the t value for time series regression coefficients, given for example by Anderson [2, p. 211]. Generally, we might wish to apply multiple testing procedures when applying the test simultaneously to all sample partial autocorrelations, but that is not a practical option. Here we follow the arguments of Box and Jenkins [3] in the identification of autoregressive models using time series partial autocorrelations. The application of GM to VAR systems has been extended by demonstrating that the sampling properties of GM's for stationary VAR's are still valid for for I(1) VAR processes (Tunnicliffe Wilson and Reale [17]).

We then specify the DAG's as recursive equation systems which can be estimated by ordinary least squares.

The next stage in the process is to establish which DAG representations are consistent with the CIG or are nearly so, allowing for statistical uncertainty, considering *demoralization*.

As we mentioned above by this term we mean the inverse operation of moralization which allows to construct a CIG from a given DAG by inserting an undirected link between any two nodes a and b when there is another node c with incoming directed edges $a \to c$ and $b \to c$. In this case c is known as a common child of a and b, and the insertion of a new, moral, link will marry the parents. After this operation for the whole graph, the directions are removed from the original links.

Of course we attach the arrow of time to links from the past to the present, so the challenge is to clarify the directions of the recursive ordering of contemporaneous variables. Normally there are alternative competitive models and eventually we compare them by using likelihood based methods.

4 Identifying an interest rate transmission for New Zealand

We apply the methodology explained in the previous sections to the interest rate mechanism in New Zealand after the implementation of the Reserve Bank Act in February 1990. To this aim we consider the model proposed by Oxley [11] who identified a structural VAR using standard procedures. The paper by Oxley also provides a thorough discussion of the economic background for the interested reader.

The application is of interest to the economists as the issues involved for this New Zealand case are multi-faceted and involves the presence of inderect effects.

The data used are monthly, seasonally unadjusted interest rates taken from the Reserve Bank of New Zealand Financial Statistics database for the period February 1990 - April 2002. The individual series considered are the rates on money at call (denoted A); 90 day bank bills (B); the yield on 1, 3 and 5 year Government stock (C, D and E respectively); base lending rate (F) first mortgage housing rate (G) and the uncovered interest parity with the US (H).

We identified a VAR(2) and hence considered all the variables up to the second lag. Once the sample partial correlation matrix was computed we tested with the appropriate procedures explained above the significance of its elements and constructed the CIG in figure 1.

We then considered all the models consistent with the CIG and used subset regression to eliminate the moral links.

The final step was to used likelihood based measures to compare the different models. In particular we considered the Akaike information criterion, the Schwarz information criterion [12] (SIC) and the Hannan-Quinn information criterion [7] (HIC).

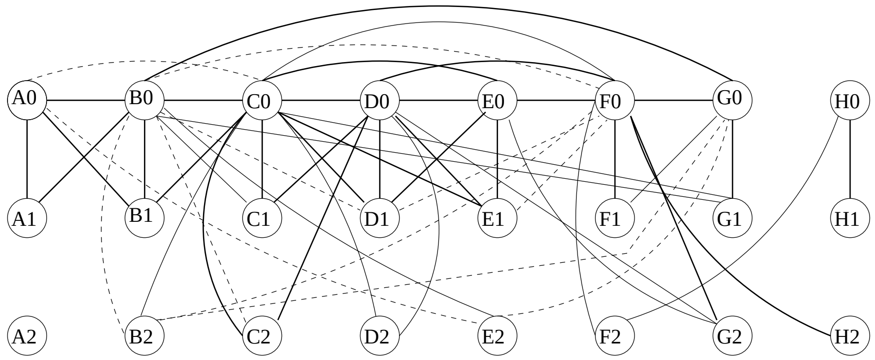


Figure 1: Conditional independence graph.

Here we present the two best models together with a table providing their values in terms of parameters, deviance and the different information criteria. They are represented in figure 2 and 3.

For both of them we can observe some common features as the lack of relevance of the uncovered interest parity and the central role of the 90 day bank bills interest rate. This application although interesting for the economist is meant as an example and so we will not make here more economic considerations.

We just conclude by saying that this methodology can be very useful for the applied scientist as it allows for any prior information when considering possible alternative DAG's.

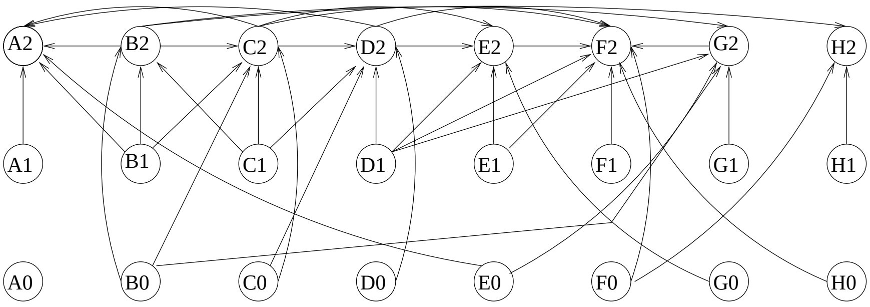


Figure 2: Best model.

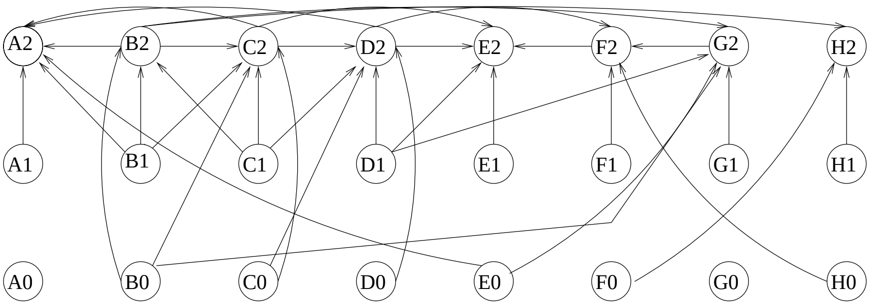


Figure 3: Alternative model.

Model	k	Dev	AIC	HIC	SIC
Best	42	130.15	-97.85	-230.68	-424.75
Alternative	37	141.13	-96.87	-235.53	-438.11

Table 1: Information criteria.

References

[1] Akaike H. (1973). *A new look at statistical model identification.* IEEE Transactions on Automatic Control. **AC-19**, 716 − 723.

[2] Anderson T.W. (1971). *The statistical analysis of time series.* Wiley, New York.

[3] Box G.E.P, Jenkins G.M. (1976). *Time series analysis, forecasting and control.* Holden Day, Oakland.

[4] Darroch J.N., Lauritzen S.L., Speed T.P. (1980). *Markov fields and log-linear interaction models for contingency tables.* Annals of Statistics **8**, 522 − 539.

[5] Dempster A.P. (1972). *Covariance selection.* Biometrics **28**, 157 − 175.

[6] Greene W.H. (1993). *Econometric analysis.* Prentice-Hall, Englewood Cliffs.

[7] Hannan E.J., Quinn B.G. (1979). *The determination of the order of an autoregression.* Journal of the Royal Statistical Society B **41**, 190 – 195.

[8] Lauritzen S.L. (2001). *Causal inference from graphical models.* In Complex Stochastic Systems, Cox D.R., Klüppleberg C. (eds). Chapman & Hall, London.

[9] Lauritzen S.L., Richardson T.S. (2002). *Chain graph models and their causal interpretations.* Journal of the Royal Statistical Society Series B, **64**, 321 – 361.

[10] Lauritzen S.L., Spiegelhalter D.J. (1988). *Local computations with probabilities on graphical structures and their applications to expert systems.* Journal of the Royal Statistical Society Series B **50**, 157 – 224.

[11] Oxley L. (2000). *Identifying an interest rate transmission mechanism for New Zealand.* Mimeo.

[12] Pearl J. (2000). *Causality.* Cambridge University Press, Cambridge.

[13] Reale M., Tunnicliffe W.G. (2002). *The sampling properties of conditional graphs for structural vector autoregressions.* Biometrika **89**, 457 – 461.

[14] Reale M., Tunnicliffe W.G. (2001). *Identification of vector AR models with recursive structural errors using conditional independence graphs.* Statistical Methods and Applications **10**, 49 – 65.

[15] Schwarz G. (1978). *Estimating the dimension of a model.* The Annals of Statistics **6**, 461 – 464.

[16] Spirtes P., Glymour C., Scheines R. (2000). *Causation, prediction and search.* MIT Press, Cambridge, MA.

[17] Tunnicliffe W. G., Reale M. (2002). *Causal diagrams for I(1) structural VAR models.* University of Canterbury Mathematics and Statistics Department Research Reports, **UCDMS2002/6**.

[18] Zellner A. (1962). *An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias.* Journal of the American Statistical Association **57**, 348 – 368.

Address: L. Oxley, Economics Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
M. Reale, Mathematics and Statistics Department, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
G. Tunnicliffe-Wilson, Mathematics and Statistics Department, Lancaster University, Lancaster LA1 4YF, UK.

E-mail: les.oxley@canterbury.ac.nz, marco.reale@canterbury.ac.nz, g.tunnicliffe-wilson@lancaster.ac.uk

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

CONFIDENCE INTERVALS AND TESTS FOR CONTRASTS BETWEEN COMBINED EFFECTS IN GENERALLY BALANCED DESIGNS

Roger W. Payne

Key words: Combination of information, effective degrees of freedom, experimental design, general balance.

COMPSTAT 2004 section: Design of experiments.

Abstract: General balance provides a unifying theory for experimental designs with several *block* (or *error*) terms. The total sum of squares can here be partitioned into components known as *strata*, one for each block term. Each stratum contains the sum of squares for the treatment terms estimated between the units of that stratum, and a residual representing the random variability of those units. Implicit in the definition is the idea that treatment effects may be estimated in more than one stratum. General balance provides efficient methods for calculating a single set of *combined effects* from all these estimates, but did not provide methods for testing them. Effective degrees of freedom are derived below, using Satterthwaite's method, and their properties are studied by simulations.

1 Introduction to general balance

General balance [3] is a powerful unifying concept for designed experiments, which encompasses most of the traditional designs. Generally balanced designs are efficient to analyse [4]; [5], and their results are straightforward to interpret [6]. Their properties are that (i) the block (i.e. random) terms are mutually orthogonal, (ii) the treatment terms are also mutually orthogonal, and (iii) the contrasts of each treatment term all have equal efficiency factors in each of the strata where they are estimated.

For a mathematical definition, suppose the model to be fitted is

$$y \quad = \quad \sum_{\alpha=1}^{b} Z_\alpha \, \epsilon_\alpha \; + \; \sum_{i=0}^{t} X_i \, \theta_i \ ,$$

where X_i and θ_i are the design matrix and effects of treatment term i, with term 0 as the grand mean, and Z_α and ϵ represent the design matrices and effects (i.e. residuals) of block term α. So

$$y \; - \; \mathrm{E}(y) \quad = \quad \sum_{\alpha=1}^{b} Z_\alpha \, \epsilon_\alpha \ ,$$

$\mathrm{E}(\epsilon_\alpha) = 0$, and $\mathrm{Var}(\epsilon_\alpha) = {\sigma_\alpha}^2 \ I$.

The first condition for general balance requires an *orthogonal block structure*: i.e. for the block (or dispersion) structure of the data vector y to be expressible as

$$\mathrm{Var}(y) \quad = \quad V \quad = \quad \sum_{\alpha=1}^{b} \xi_\alpha \, \tilde{S}_\alpha \ ,$$

where the *stratum variances* ξ_α are positive real constants, and $\tilde{S}_\alpha$ (the projectors onto the strata) are known symmetric matrices such that

$$\tilde{S}_\alpha \ \tilde{S}_\beta = \delta_{\alpha,\beta} \ \tilde{S}_\alpha$$

(thus indicating the block terms are orthogonal), and

$$\textstyle\sum_{\alpha=0}^{b} \tilde{S}_\alpha \ = \ I \ ,$$

where $\delta_{\alpha,\beta}$ takes the value 1 if α=β and the value 0 otherwise, and $\tilde{S}_0$ is the projector for the grand mean.

The relationship between the stratum projectors $\tilde{S}_\alpha$ and the random terms in the mixed model is derived by Payne & Tobias [6] as follows. The projector S_α for term α is

$$S_\alpha = Z_\alpha \ ({Z_\alpha}' \ Z_\alpha \ )^- \ {Z_\alpha}' \ .$$

This projects the data vector into the vector space spanned by the possible values of the vector ϵ_α, by forming an average of the data values for each element of ϵ_α and then putting that average back into the appropriate elements of the data vector.

For valid randomisation [3],[1], the effects in each random vector ϵ_α must have an equal replication, n_α say. We can then write

$$S_\alpha = (1 \ / \ n_\alpha) \ Z_\alpha \ {Z_\alpha}' \ .$$

Random term β is said to be marginal to term α if

$$S_\alpha \ S_\beta = S_\beta \ ,$$

that is, β is marginal to α if its vector space is a subspace of the vector space of α. The stratum projector for α is formed by removing from S_α the spaces for the terms marginal to α. These marginal terms should all occur before α in the model (i.e. β<α if β is marginal to α). So, taking values of α running from 1 up to b, we can define recursively

$$\tilde{S}_\alpha = S_\alpha \ (\text{ I - } \textstyle\sum_{\{\beta\text{: term } \beta \text{ is marginal to term } \alpha\}} \tilde{S}_\beta \ ) \ ,$$

or

$$\tilde{S}_\alpha = S_\alpha \ (\text{ I - } \textstyle\sum_{\{\beta\text{: } \beta<\alpha\}} \tilde{S}_\beta \ ) \ .$$

The stratum variances ξ_α can then be calculated as

$$\xi_\alpha = n_\alpha \ {\sigma_\alpha}^2 + \textstyle\sum_{\{\beta\text{: term } \beta \text{ is marginal to term } \alpha\}} n_\beta \ {\sigma_\beta}^2 \ .$$

The second condition for general balance requires an *orthogonal treatment structure*: i.e. one with the form

$$\text{E}(y) = \tilde{T}_0 \ \tau_0 + \tilde{T}_1 \ \tau_1 + \ldots \ \tilde{T}_t \ \tau_t \ ,$$

where $\tilde{T}_0 = \tilde{S}_0$ (the projector for the grand mean), and the projector $\tilde{T}_i$ onto the space for treatment term i is a known symmetric matrix such that

$$\tilde{T}_i \ \tilde{T}_j = \delta_{i,j} \ \tilde{T}_i \ .$$

This can similarly be related to the terms in the mixed model: define

$$T_i = X_i({X_i}'X_i)^-{X_i}' \ .$$

Then

$$\tilde{T}_i = T_i \ (\ I \text{ - } \textstyle\sum_{\{j\text{: term } j \text{ is marginal to term } i\}} \tilde{T}_j \ ) \ ,$$

or

$$\tilde{T}_i = T_i \ (\ I \text{ - } \textstyle\sum_{\{j\text{: } j<i\}} \tilde{T}_j \ ) \ .$$

Also, if we write the full design matrix as $X = [\ X_0 \mid X_1 \mid X_2 \mid \ldots \ ]$, then

$$T = X\ (X'\ X\ )^{-}\ X = \tilde{T}_0 + \ldots + \tilde{T}_t\ ,$$

and

$$\mathrm{E}(y) = T\ \tau = \tilde{T}_0\ \tau_0 + \tilde{T}_1\ \tau_1 + \ldots + \tilde{T}_t\ \tau_t\ .$$

The orthogonal block structure allows an independent least squares analysis to be performed within each stratum. The residual sum of squares in stratum α is

$$(\tilde{S}_\alpha\ y - \tilde{S}_\alpha\ \tilde{T}\ \tau\ )'\ (\tilde{S}_\alpha\ y - \tilde{S}_\alpha\ \tilde{T}\ \tau)\ ,$$

and the normal equations are

$$T\ \tilde{S}_\alpha\ T\ \hat{\tau} \quad = \quad T\ \tilde{S}_\alpha\ y\ ,$$

i.e.

$$\textstyle\sum_{i=0}^{t} \tilde{T}_i\ \tilde{S}_\alpha\ \tilde{T}_i\ \hat{\tau}_{\alpha i} \quad = \quad \sum_{i=0}^{t} \tilde{T}_i\ \tilde{S}_\alpha\ y\ .$$

The third, and final, condition for general balance defines the relationship between the block and treatment structures:

$$\tilde{T}_i\ \tilde{S}_\alpha\ \tilde{T}_j = \delta_{i,j}\ \lambda_{\alpha i}\ \tilde{T}_i$$

The normal equations then become

$$\textstyle\sum_{i=0}^{t} \lambda_{\alpha i}\ \tilde{T}_i\ \hat{\tau}_{\alpha i} \quad = \quad \sum_{i=0}^{t} \tilde{T}_i\ \tilde{S}_\alpha\ y\ ,$$

which are solved by

$$\hat{\tau}_{\alpha i} \quad = \quad (1\ /\ \lambda_{\alpha i})\ \tilde{T}_i\ \tilde{S}_\alpha\ y\ ,$$

with variance-covariance matrix

$$\mathrm{var}(\ \hat{\tau}_{\alpha i}\ ) \quad = \quad (\ \xi_\alpha\ /\ \lambda_{\alpha i}\ )\ \tilde{T}_i\ .$$

So $\lambda_{\alpha i}$ is the *efficiency factor* for treatment term i in stratum α as defined by Yates [9]. Note that, if term i is orthogonal, then $\lambda_{\alpha i}$=1.

2 Effective degrees of freedom

Payne & Tobias [6] devised an efficient method for estimating the stratum variances from analysis of covariance of any generally balanced design. Given these estimates, an efficient overall estimate of the effects of the treatments, combining the information from all the strata where they are estimated, can be derived by generalized least squares estimation with weight matrix V^{-1}. Payne & Tobias showed that this leads to estimates for treatment term i given by

$$\hat{\tau}_i \quad = \quad \textstyle\sum_{\alpha=0}^{b} w_{\alpha i}\ \hat{\tau}_{\alpha i}$$

where

$$w_{\alpha i} \quad = \quad (\lambda_{\alpha i}\,/\,\xi_\alpha\,)\ /\ \textstyle\sum_{\beta=0}^{b}(\lambda_{\beta i}\,/\,\xi_{\beta i}\,)$$

and

$$\mathrm{cov}(\tau_{\alpha i}\ ,\ \tau_{\beta j}\ ) \quad = \quad \delta_{ij}\ \tilde{T}_i\ /\ \textstyle\sum_{\alpha=0}^{b}(\lambda_{\alpha i}\,/\,\xi_{\alpha i}\,)$$

In their discussion, however, Payne & Tobias [6] left as unanswered the question of how to calculate confidence limits for these combined effects or to perform tests of contrasts amongst them.

Kenward & Roger [2], taking the more general context of analysis of unbalanced linear mixed models by residual maximum likelihood, derived methods that, for single degrees of freedom, amounted to modifications of Satterthwaite's [8] method. This method is also customarily used to obtain effective degrees of freedom for tables of means that contain effects from more than one stratum (see e.g. Payne et al. [7, page 369]. So it may provide a feasible solution here too.

Satterthwaite's method leads to the equation

$$\text{effective d.f.} \;=\; \{\; \textstyle\sum_{\alpha=0}^{b} \mathrm{var}(\alpha) \;\}^2 \;/\; \textstyle\sum_{\alpha=0}^{b} \{\; \mathrm{var}(\alpha)^2 \,/\, \mathrm{df}(\alpha) \;\}$$

where $\mathrm{var}(\alpha)$ is the variance of the contribution to the combined estimate from stratum α, and $\mathrm{df}(\alpha)$ is the number of effective degrees of freedom there.

Payne & Tobias [6] showed that effective degrees of freedom for the estimated stratum variance for stratum α can be calculated as

$$d_\alpha \;=\; \mathrm{trace}(\tilde{S}_\alpha\,) \;-\; \textstyle\sum_{i=0}^{t} \{\, w_{\alpha i}\; \mathrm{trace}(\tilde{T}_i\,)\,\}$$

For the combined effects of treatment term i

$$\mathrm{var}(\alpha) = {w_{\alpha i}}^2\;(\xi_\alpha \;/\; \lambda_{\alpha i})\;\tilde{T}_{\mathrm{i}}\;.$$

So, applying Satterthwaite's method (by substituting for $\mathrm{var}(\alpha)$ and $w_{\alpha i}$ as defined above) gives the following effective degrees of freedom for any contrast amongst the combined effects

$$f_i \;=\; \{\; \textstyle\sum_{\alpha=1}^{b} \lambda_{\alpha\, i} \,/\, \xi_\alpha \;\}^2 \;/\; \{\; \textstyle\sum_{\alpha=1}^{b} (\,\lambda_{\alpha\, i} \,/\, \xi_\alpha\,)^2 \,/\, d_\alpha \;\}$$

Effective degrees of freedom may also be required for use when assessing contrasts within a table of means. Any such contrast may involve contributions from several treatment terms. Suppose the contrast of interest involves a contrast vector c_i (possibly null) from each treatment term i. The variances of the contributions to the estimated contrast from each stratum, α, can then be calculated as

$$\mathrm{var}(\alpha) = \textstyle\sum_{i\in\{I_\alpha\}}\; {w_{\alpha i}}^2\;(\xi_\alpha/\lambda_{\alpha i}) {c_i}'\;\tilde{T}_i c_i\;.$$

where $\{I_\alpha\}$ is the set of indexes of the treatment terms estimated in stratum α. This allows Satterthwaite's equation to be applied, as before. However, the resulting equation is not presented explicitly here as the result does not simplify into as elegant a form as that for the combined effects. (In statistical software, it would in any case, be clearer to make the calculation in two stages as defined above.)

The fact that the individual contributions to the combined effects would be assumed to have normal distributions, under the usual assumptions of the analysis of variance, suggests that the combined effects could be assumed approximately to follow t-distributions with the calculated number of effective degrees of freedom. This assumption is assessed in the simulated examples below.

3 Examples

To study the properties of the effective degrees of freedom, data sets were simulated for a balanced-incomplete-block design with split plots. As shown in the plan below there were 10 blocks of three plots. Factor A, with five levels, was applied to the plots and factor B, with two levels was applied to the subplots.

a_3b_1 a_3b_2	a_2b_1 a_2b_2	a_1b_1 a_1b_2	a_1b_1 a_1b_2	a_1b_1 a_1b_2	a_2b_1 a_2b_2	a_1b_1 a_1b_2	a_1b_1 a_1b_2	a_2b_1 a_2b_2	a_1b_1 a_1b_2
a_4b_1 a_4b_2	a_3b_1 a_3b_2	a_2b_1 a_2b_2	a_2b_1 a_2b_2	a_3b_1 a_3b_2	a_3b_1 a_3b_2	a_2b_1 a_2b_2	a_3b_1 a_3b_2	a_4b_1 a_4b_2	a_4b_1 a_4b_2
a_5b_1 a_5b_2	a_4b_1 a_4b_2	a_3b_1 a_3b_2	a_4b_1 a_4b_2	a_5b_1 a_5b_2	a_5b_1 a_5b_2	a_5b_1 a_5b_2	a_4b_1 a_4b_2	a_5b_1 a_5b_2	a_5b_1 a_5b_2

The skeleton analysis-of-variance table, below, shows that 1/6 of the information for the main effect of factor A is estimated in the Blocks stratum (i.e. between blocks), and 5/6 in the Blocks.Plots stratum (i.e. between the plots within each block). The main effect of factor B and the A.B interaction are estimated in the Blocks.Plots.Subplots stratum (i.e. between the subplots within each plot).

Source	d.f.	efficiency factor
Blocks stratum		
A	4	0.167
Residual	5	
Blocks.Plots stratum		
A	4	0.833
Residual	16	
Blocks.Plots.Subplots stratum		
B	1	1.000
A.B	4	1.000
Residual	25	

So, in the A by B table of means, a comparison between two means with the same level of A (e.g. $\{a_1b_1\}$ versus $\{a_1b_2\}$) will have an effective number of degrees of freedom equal to the number of residual degrees of freedom of the Blocks.Plots.Subplots stratum, as this involves only B effects and these are only estimated there. So there will be no Satterthwaite approximation, and the assessment will be by an ordinary t-test with 25 degrees of freedom. Conversely, a comparison between two means with different levels of A (e.g. $\{a_4b_1\}$ versus $\{a_5b_2\}$) will have an effective number of degrees of freedom calculated as described in Section 2, to take account of the fact that the variance has contributions from all the strata.

Data sets were simulated by adding together vectors formed by generating sets of normal random numbers for the block effects, for the sets of plot effects within each block, and for the sets of subplot effects within each plot. The variance component for the Blocks.Plots.Subplots stratum was fixed at 10, while those for the Blocks and for Blocks.Plots took all pairs of values from the set {5, 10, 20 and 40}. One thousand simulations were done for each set of values of the variance components, and a tabulation was made of the percentage of times that the contrasts defined above, within the A by B table of means, were significant at 5%, 1% and 0.1%, assuming t-distributions with numbers of effective degrees of freedom calculated as described in Section 2. Also, for comparison, the tabulations were done for the contrasts at different levels of A calculated from the usual, non-combined A by B table of means (formed using the A effects only from the lowest stratum where they are estimated, namely the Blocks.Plots stratum).

The first column of each table (labelling the rows) gives the variance component used for the effects from the Blocks.Plots stratum, while the first row (labelling the groups of three columns) gives the variance component used for the effects from the Blocks stratum.

	5			10			20			40		
	5%	1%	.1%	5%	1%	.1%	5%	1%	.1%	5%	1%	.1%
5	5.1	0.8	0.2	4.7	1.0	0.2	4.6	0.4	0.1	5.2	1.3	0.0
10	4.1	0.8	0.2	4.8	0.7	0.0	4.6	0.7	0.1	5.0	1.0	0.2
20	5.5	0.9	0.1	5.7	1.2	0.0	5.0	1.0	0.2	5.3	0.8	0.2
40	5.6	1.1	0.1	4.7	1.1	0.0	4.6	0.9	0.3	3.6	0.5	0.0

Table 1: Contrasts between combined means with the same level of A.

	5			10			20			40		
	5%	1%	.1%	5%	1%	.1%	5%	1%	.1%	5%	1%	.1%
5	5.5	1.5	0.0	6.3	0.9	0.0	6.3	1.4	0.1	6.4	1.0	0.0
10	5.2	1.2	0.1	4.5	1.1	0.3	4.7	1.2	0.0	5.5	1.0	0.0
20	4.6	1.0	0.3	5.7	0.9	0.0	5.4	1.3	0.2	6.3	1.1	0.1
40	5.1	1.1	0.1	6.4	1.1	0.0	6.2	1.2	0.0	6.0	1.3	0.0

Table 2: Contrasts between combined means with different levels of A.

	5			10			20			40		
	5%	1%	.1%	5%	1%	.1%	5%	1%	.1%	5%	1%	.1%
5	4.5	1.0	0.0	6.1	0.8	0.0	5.4	1.4	0.1	5.1	1.0	0.0
10	5.3	1.3	0.1	4.1	0.9	0.2	3.6	1.1	0.1	5.1	0.9	0.0
20	4.8	0.8	0.2	5.6	0.7	0.1	5.3	1.1	0.2	4.5	0.9	0.0
40	5.0	1.0	0.1	5.9	1.2	0.0	6.2	1.1	0.0	6.0	0.9	0.0

Table 3: Contrasts between non-combined means with different levels of A.

From Table 2, it can be seen that the number of significances arising from the use of the effective degrees of freedom for the contrasts between combined means with different levels of A were comparable to those obtained for those from the ordinary (non-combined) table of means (Table 3). Too many significant contrasts were obtained for some of the combinations, but the differences were relatively small (for example, no more than 6.4% in the entries where 5% was expected). This compares reasonably well with the largest discrepancy (5.7% instead of 5%) in Table 1, which assessed the contrasts at the same level of A using ordinary t-tests (and thus provides a guide to the range of variation to be found in ordinary circumstances).

4 Conclusion

The use of weights, estimated from the data, in the calculation of the combined effects, the variances and the effective degrees of freedom might have been expected to cause the tests of contrasts between combined means to be less accurate than tests of contrasts between the usual (non-combined) means. However, for the design above, the simulations above show little difference between these two situations (Tables 2 and 3). The tests seem to have been a little over-optimistic for some of the simulations in both of Tables 2 and 3. However, the effects are slight, and would certainly not be enough to discourage use of the effective degrees of freedom for combined effects in practice. The moral seems to be that the significance levels should be used with caution, but this is a good rule under any circumstances.

This year, 2004, marks the 80th birthday of John Nelder who initially devised the theory of general balance. The importance of general balance is that it provides a general theory which facilitates general algorithms for the analysis of a very wide range of designs. This allows potential users to choose from a wide repertoire of useful designs all of which, as shown by Payne & Welham [5], are efficient to analyse. The calculation of effective degrees of freedom for combined effects provides a solution to a question that was left unanswered by Payne & Tobias [6], and this has now been made available in the statistical system GenStat *for Windows* [7].

References

[1] Bailey R.A., Praeger C.E., Rowley C.A., Speed T.P., (1983). *Generalized wreath products of permutation groups*. Proceedings of the London Mathematical Society, **47**, 69-82.

[2] Kenward M.G. & Roger J.H. (1997). *Small sample inference for fixed effects from restricted maximum likelihood*. Biometrics, **53**, 983 – 997.

[3] Nelder J.A. (1965). *The analysis of randomised experiments with orthogonal block structure. I. block structure & the null analysis of variance. II. treatment structure & the general analysis of variance*. Proceedings of the Royal Society of London, **A283**, 147 – 178.

[4] Payne R.W. & Wilkinson G.N. (1977). *A general algorithm for analysis of variance*. Applied Statistics, **26**, 251 – 260.

[5] Payne R.W. & Welham S.J. (1990). *A comparison of algorithms for combination of information in generally balanced designs*. COMPSTAT 1990, Physica-Verlag, Heidelberg, 297 – 302.

[6] Payne R.W. & Tobias R.D. (1992). *General balance, combination of information and the analysis of covariance*. Scandinavian Journal of Statistics, **19**, 3 – 23.

[7] Payne R.W., Baird D.B., Cherry M., Gilmour A.R., Harding S.A., Kane A.F., Lane P.W., Murray D.A., Soutar D.M., Thompson R., Todd A.D., Tunnicliffe Wilson G., Webster R. & Welham S.J. (2003). *The guide to GenStat release 7.1, part 2: statistics*. VSN International, Oxford.

[8] Satterthwaite F.F. (1941). *Synthesis of variance*. Psychometrika, **6**, 309 – 316.

[9] Yates F. (1936). *Incomplete randomized blocks*. Annals of Eugenics, **7**, 121 – 140.

Address: R.W. Payne, Rothamsted Research, Harpenden, Herts AL5 2JQ, UK

E-mail: `roger.payne@bbsrc.ac.uk`

COMPSTAT'2004 Symposium

STUDENTISED BLOCKWISE BOOTSTRAP FOR TESTING HYPOTHESES ON TIME SERIES

Martin Peifer and Jens Timmer

Key words: Blockwise bootstrap, studentising, smooth function model, tremor.

COMPSTAT 2004 section: Resampling methods.

Abstract: Without having profound model knowledge, hypotheses testing on time series can often be accomplished by modified resampling methods such as the blockwise bootstrap. For this specific method a selection of the free parameter blocklength, and the studentisation is proposed. The whole procedure is applied for testing hypotheses on medical times series.

1 Introduction

Since the introduction of the bootstrap [3], methods based on resampling have been applied on numerous statistical problems. The success of bootstrap may be explained by its easy implementation whenever the data are identical distributed and statistically independent. If the statistical independence does not hold, statistical inference using bootstrap methods becomes much more difficult and is often based on profound model assumptions, e.g. [2], [1].

Instead, for blockwise bootstrap only some structural properties of the times series such as α-mixing and strong stationarity are needed [10], [12]. The blockwise bootstrap yields appropriate results if the free parameter - blocklength l – is adjusted adequately. The choice of this parameter is discussed in Section 4.

The second issue deals with the studentisation of the blockwise bootstrap method for which we assume that the statistic can be represented by a smooth function of means. The effect of the studentisation is the enhancement of the asymptotic properties of blockwise bootstrap which is then second order correct [5], [11]. It is therefore highly recommended to studentise the method.

2 Blockwise bootstrap

Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be an observed sample from a strongly stationary, α-mixing, p-variate sequence $(\mathbf{X}_t)_{t\in\mathbb{Z}}$. The real-valued statistics $T_n = T_n(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ is assumed to be invariant under permutations of the observations.

For the blockwise bootstrap, b subsamples or blocks of length l are formed from the observations. We further assume, without loss of generality, that $n/l \in \mathbb{N}$ (otherwise, the data sample is truncated until $n/l \in \mathbb{N}$ holds). In the framework of blockwise bootstrap, two kinds of building subsamples are predominating, the overlapping blocks and the non-overlapping blocks. The overlapping blocks are defined by

$$\mathbf{Y}_i = (\mathbf{X}_i, \cdots, \mathbf{X}_{i+l}) \qquad i = 1, \cdots, b = n - l - 1 \,, \tag{1}$$

and non-overlapping are defined as follows:

$$\mathbf{Y}_i = (\mathbf{X}_{(i-1)l+1}, \cdots, \mathbf{X}_{il}) \qquad i = 1, \cdots, b = \frac{n}{l} \,. \tag{2}$$

We decided to use non-overlapping blocks and briefly mention different results which are due to the other sampling scheme. Blockwise bootstrap is then realised by:

1. Drawing blocks with replacement from the realisations $\{\mathbf{y}_1, \cdots, \mathbf{y}_b\}$ and forming $\mathbf{y}_1^*, \cdots, \mathbf{y}_b^*$ by gluing the drawn blocks together.

2. Repeat 1. B times to generate B bootstrap samples $\mathbf{x}_1^*, \cdots, \mathbf{x}_B^*$.

3. Calculate $T_{n,k}^* = T_n(\mathbf{x}_{1,k}^*, \cdots, \mathbf{x}_{n,k}^*), \quad k = 1, \cdots, B$.

4. Finally, determine e.g. the bootstrap distribution function $F^*(\xi) = B^{-1} \sum_{k=1}^{B} \theta\left(\xi - T_{n,k}^*\right)$, which approximates the distribution function of T_n.

Here, $\theta(\cdot)$ denotes the step function. A proof of the consistency of the described method is e.g. given in [13].

3 Studentising the blockwise bootstrap

3.1 Smooth function model

Suppose an i.i.d. sample of p-variate random vectors, $\mathbf{W}_1, \cdots, \mathbf{W}_n$. Let $\mu = E[\mathbf{W}_i]$ and $\bar{\mathbf{W}} = n^{-1} \sum_{i=1}^{n} \mathbf{W}_i$. A function A is called a smooth function model if

$$A \ : \mathbb{R}^p \rightarrow D \subset \mathbb{R} \,, \qquad A \in C^\infty(\mathbb{R}^p, D) \qquad \text{and} \quad A(\mu) = 0 \,. \tag{3}$$

This concept can be applied on statistical problems in which the statistic can be expressed by a smooth function g such as $T_n = g(\bar{\mathbf{W}})$. The associated smooth function model then yields: $A(\mathbf{w}) = g(\mathbf{w}) - g(\mu)$. To approximate the variance of T_n let

$$\mathbf{Z} = n^{\frac{1}{2}} \left(\bar{\mathbf{W}} - \mu\right) \quad , \quad a_i = \left.\frac{\partial A(\mathbf{w})}{\partial w^{(i)}}\right|_{\mathbf{w}=\mu}$$

$$b_{ij} = \left.\frac{\partial^2 A(\mathbf{w})}{\partial w^{(i)} \partial w^{(j)}}\right|_{\mathbf{w}=\mu} \quad \text{and} \quad C_{ij} = E\left[(\mathbf{W}_1 - \mu)^{(i)} (\mathbf{W}_1 - \mu)^{(j)}\right] \,.$$

Because of the independence of the $\mathbf{W}_i$'s, $E\left[Z^{(i)}\right] = 0$ and $E\left[Z^{(i)}Z^{(j)}\right] = C_{ij}$. Now, Taylor-expanding the function $S_n = n^{\frac{1}{2}}\ A(\bar{\mathbf{W}})$ at $\mathbf{w} = \mu$, leads to:

$$S_n \quad = \quad \sum_{i=1}^{p} a_i Z^{(i)} + n^{-\frac{1}{2}}\ \frac{1}{2} \sum_{i,j=1}^{p} b_{ij}\ Z^{(i)} Z^{(j)} + O_p(n^{-1})\ . \tag{4}$$

And therefore the expectations $E\left[S_n\right]$ and $E\left[S_n^2\right]$ are given by:

$$\begin{aligned} E\left[S_n\right] &= n^{-\frac{1}{2}}\ \frac{1}{2} \sum_{i,j=1}^{p} b_{ij}\ C_{ij} + O(n^{-1})\ , \\ E\left[S_n^2\right] &= \sum_{i,j=1}^{p} a_i a_j\ C_{ij} + O(n^{-1})\ . \end{aligned}$$

Due to the expansion (4), assumption (3) can be weaken to $A \in C^3\left(\mathbb{R}^p, D\right)$ for our purposes. Using $\mathrm{Var}\left(T_n\right) = n^{-1}\mathrm{Var}\left(S_n\right) = n^{-1}\sum_{i,j=1}^{p} a_i a_j\ C_{ij} + O(n^{-2})$ and replacing C_{ij} by its plug-in estimator $\hat{C}_{ij}$ yields:

$$\hat{\sigma}_n^2 = \widehat{\mathrm{Var}}(T_n) = n^{-1} \sum_{i,j=1}^{p} \hat{a}_i \hat{a}_j\ \hat{C}_{ij}\ , \tag{5}$$

where $\hat{a}_i = \left.\frac{\partial A(\mathbf{w})}{\partial w^{(i)}}\right|_{\mathbf{w}=\bar{\mathbf{W}}}$. Extending these results to blockwise defined statistics leads to variance estimators which are sufficient for studentising the described bootstrap method.

3.2 The procedure of studentising blockwise bootstrap

Since Equation (5) does not give consistent results if the observation $\mathbf{W}_i$ are not i.i.d., a modification of the smooth function A is needed. To approximate the variance of T_n under conditions outlined in Section 3.1, it is assumed that the statistic can be written into $T_n = g(\bar{\mathbf{W}})$, where $\mathbf{W}_i = f(\mathbf{X_i})$ is a suitable transformation of the observations $\mathbf{X}_i$, e.g. $f(x) = (x^2, x)$ and $g(x_1, x_2) = x_1 - x_2^2$ for the sample variance. Again, the series $\mathbf{W}_i$ is not i.i.d.. To incorporate the dependence structure into $\hat{\sigma}_n^2$, the blocking scheme is implemented into the statistic. For this purpose define

$$\tilde{W}_i \quad = \quad \left(W^{(1)}_{(i-1)l+1}, \cdots, W^{(1)}_{il}, \cdots, W^{(p)}_{(i-1)l+1}, \cdots, W^{(p)}_{il}\right)\ , \quad i = 1, \cdots, \frac{n}{l}$$

for non-overlapping blocks and for overlapping blocks

$$\tilde{W}_i \quad = \quad \left(W^{(1)}_{i}, \cdots, W^{(1)}_{i+l}, \cdots, W^{(p)}_{i}, \cdots, W^{(p)}_{i+l}\right)\ , \quad i = 1, \cdots, n-l-1\ .$$

The statistic is then rewritten into

$$T_n = \tilde{g}\left(b^{-1} \sum_{i=1}^{b} \tilde{\mathbf{W}}_i\right)\ ,$$

where $\tilde{g}(w) = g\left(l^{-1}\sum_{i=1}^{l}\left(w^{(i)},\cdots,w^{((p-1)l+i)}\right)\right)$. And hence again, $A(\mathbf{w}) = \tilde{g}(\mathbf{w}) - \tilde{g}(\mu)$. Suppose that the blocklength is adequately chosen and therefore the blocks are approximately independent, such that the approximation in Equation (5) can be used.

Studentised bootstrap is realised by studentising each bootstrap sample $T^*_{stud,n} = \frac{T^*_n - T_n}{\hat{\sigma}^*_n}$. The bootstrap approximation of the studentised statistic is now determined similar to Section 2.

4 Blocklength selection

The success of the method depends on the choice of the blocklength l. Taking the mean squared error as objective measure to be minimised, the blocklength selection shows a tradeoff between bias and variance. For the sample mean, the mean squared error can be calculated and is in case of the bootstrap variance approximation [7],

$$\mathrm{MSE}(l) = E\left[\left(\mathrm{Var}^*(\bar{X}) - \mathrm{Var}(\bar{X})\right)^2\right] \approx \frac{1}{n^2 l^2}\, C_1 + \frac{l}{n^3}\, C_2\ , \tag{6}$$

where
$$C_1 = \left(\sum_{k=-\infty}^{\infty} |k|\ \gamma(k)\right)^2\ ,\ C_2 = 2v\left(\sum_{k=-\infty}^{\infty} \gamma(k)\right)^2\ .$$

Here, $\gamma(k)$ is the auto-covariance function of the process and $v = 1$ for non-overlapping blocks, $v = 2/3$ for overlapping. The optimal blocklength is therefore given by $l_{opt} = (2C_1/C_2)^{1/3}\ n^{1/3}$. To generalise this concept, the statistic T_n is linearised before, using again the smooth function model. The constants C_1 and C_2 are then determined by the covariance structure of the linearised statistic.

Since the correlated errors of the empirical auto-covariance function would be amplified by the factor $|k|$ in C_1, the plug-in principle leads to a bad estimation of l. To avoid this problem, the mixing coefficient is assumed to decay exponentially and thus the auto-covariance function. Using this assumption, some points of the correlation function are estimated: $\hat{\gamma}(k)$, $k = 0,\cdots,m < n$, and the function $f(k) = \phi^k, 0 \le \phi < 1$ is fitted to the envelope of $\hat{\gamma}(k)$. Then the estimated parameter ϕ contains the characteristic time scale of the process and the blocklength is finally estimated by replacing $\gamma(k)$ in C_1, C_2 with ϕ^k. The procedure of selecting the blocklength is therefore the following:

- Linearise the statistic T_n by transforming the data points to V_i such that
$T_n = g(\bar{\mathbf{W}}) \approx n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\partial g(\mathbf{w})}{\partial w^{(j)}}\bigg|_{\mathbf{w}=\bar{\mathbf{W}}} W_i^{(j)} = n^{-1}\sum_{i=1}^{n} V_i$
$\Rightarrow V_i = \sum_{j=1}^{p}\frac{\partial g(\mathbf{w})}{\partial w^{(j)}}\bigg|_{\mathbf{w}=\bar{\mathbf{W}}} W_i^{(j)}$ for the smooth function model of Section 3.1.

- Estimate the auto-covariance function of the transformed series V_i.
- Determine the envelope of the estimated auto-covariance function. In case of oscillatory processes we propose using the Hilbert transform [14].
- Fit $f(k) = \phi^k$ to the envelope.
- The blocklength $\hat{l}$ is then:

$$\hat{l} = (4\ n/v)^{1/3} \left(\frac{\phi}{1-\phi} + \frac{\phi^2}{(1-\phi)^2} \right)^{2/3} \left(1 + 2\ \frac{\phi}{1-\phi} \right)^{-2/3} ,$$

again, $v = 1$ for non-overlapping and $v = 2/3$ for overlapping blocks.

Due to the calculation of the envelope of the auto-covariance function this approach differs from fitting a single autoregressive process of order 1 to the process and determining the blocklength from the process parameter. Such a procedure was e.g. studied in [15] within the class of autoregressive-moving average processes.

5 Simulations

To investigate the effect of studentisation and the choice of the blocklength, two different data generating processes are chosen:

- Autoregressive process of order 1 (AR[1]), $X_t = a_1\ X_{t-1} + \epsilon_t$, $a_1 = \exp(-1/\tau)$, $\tau > 0$ and $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. sequence of $\mathcal{N}(0,1)$ random variables.
- Autoregressive process of order 2 (AR[2]), $X_t = a_1\ X_{t-1} + a_2\ X_{t-2} + \epsilon_t$, where $a_1 = 2\exp(-1/\tau)\cos\left(\frac{2\pi}{T}\right)$, $a_2 = -\exp(-2/\tau)$, $\tau, T > 0$ and $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. sequence of $\mathcal{N}(0,1)$ random variables.

The chosen parameters are: $\tau = 5$ for the AR[1]-process and $\tau = 10$, $T = 5$ for the AR[2]-process. For all simulations autocorrelations up to time-lag 256 are considered to select the blocklength. Beside the data generating processes, a specific statistic has to be chosen. Motivated by the application, Section 6, the sample variance is used throughout this section. In order to give a measure of the accuracy for the following bootstrap approximations, two-sided equally tailed 95%-confidence intervals are calculated. The coverage-error of the confidence intervals is then estimated by the relative frequency, in which the variance of the process is falling into the interval over 1000 independent runs. The sign of the coverage error was chosen to be negative for conservative and positive for anti-conservative confidence intervals. The results of the simulations are shown in Fig. 1, where the coverage error is determined in dependence on the amount of data. This is done for either studentised or non-studentised blockwise bootstrap. For both, the AR[1]-process and the AR[2]-process the asymptotic coverage is approached for only 1500 data points, when the studentised method is used. In contrast to

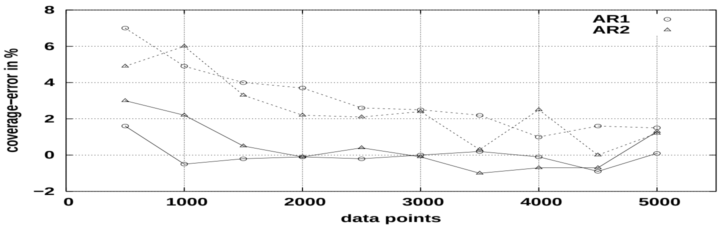


Figure 1: Coverage error of 95%-confidence intervals on dependence of the number of data points, where both, the non-studentised blockwise bootstrap (dotted lines) and the studentised blockwise bootstrap (solid lines) are considered. The data generating processes are AR[1]-process (○) and the AR[2]-process (△). Negative values of the coverage error denote conservative confidence intervals while positive values are corresponding to anti-conservative confidence intervals.

the non-studentised method, which is still showing a small coverage-error at $n = 5000$ data points. We therefore propose using the studentised bootstrap to enhance the rate of convergence, which is in accordance of many theoretical results concerning the bootstrap, see e.g. [8], [6], [9], [4], [16], [5].

Finally, in case of the AR[1]-process the theoretically calculated scaling constant C of the optimal blocklength $l_{opt} = C\ n^{1/3}$ is compared to the estimated constant. It turns out that both coefficients are matching for this situation.

6 Application

The proposed method is applied to test a time of the day (TOD) dependency in the variance of physiological (healthy) hand tremor. All time series are of length 30000 data points and are sampled with 1 kHz. At 4 different times, 9.00, 11.15, 13.30 and 15.30 the tremor of the outstrechted left hand was recorded. For each TOD, 3 data sets were recorded to test the consistency of the measurements. One recording at 13.30 was not used, because a drift is present, which is due to a slow hand movement. For a first inspection, 95%-confidence intervals of the sample variance for all datasets are estimated using the studentised blockwise bootstrap method. The results, Fig. 2, clearly show a TOD dependence. The confidence intervals further suggest, that the repeated measurements are consistent.

Testing the hypothesis of the TOD dependence and the consistency of the repeated measurements statistically, we choose to parameterise the bootstrap distributions of the variance estimators by χ^2-distributions. The degrees of freedom of the χ^2-distributions are estimated by minimising the Kolmogorov distance. This parameterisation is in good accordance to the bootstrap distribution. Finally, the parameterised distributions are log-transformed to yield a Gaussian error-model in good approximation. Now, a two factorial

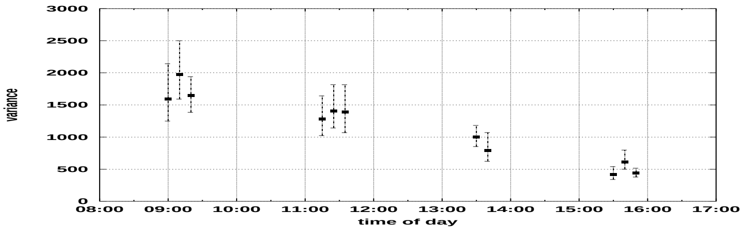


Figure 2: 95%-confidence intervals of the variance suggests a time of day dependency and the consistency of the repeated measurements of the tremor time series. The temporal distance of the repeated measurements are stretched for sake of clarity.

ANOVA has been carried out to test the hypotheses: 1. no over-all-effect is present, 2. there is no TOD dependence and 3. the repeated measurements are consistent. Choosing a level of significance of 1%, we can infer that an over-all-effect is present, which is the TOD dependence (p-values $< 10^{-5}$). The hypothesis of the consistency of the repeated measurements cannot be rejected. Hence, the test results are in perfect accordance of the intuition imparted by the visual inspection of the confidence intervals shown in Fig. 2.

7 Discussion and conclusion

The problem of hypotheses testing when the observed data are not statistically independent rises in many areas of applied statistics. It is often not possible to derive a suitable model of the data generating process. In these cases non-parametric methods like bootstrap are widely used.

The proposed blockwise bootstrap is a method, which can cope with the dependence structure of the observations. But some structural assumptions has to be fulfilled to achieve consistent results. Beside these assumptions, the free parameter of the blocklength has to be adjusted in order to minimise the bias and the variance of the approximation. The discussed method for selecting the blocklength is easy to apply and fast to compute. On two exemplary processes, the simulation study shows that the chosen blocklength gives suitable approximations.

The effect of studentising the statistic has also been studied. It turns out to use the studentised blockwise bootstrap is highly recommended to enhance the convergence rate. This result is in accordance with many past investigations of the bootstrap method.

To show the practical significance of the proposed method, we gave an application, in which a time-of-day dependence of the human hand tremor is tested. The test clearly confirms this dependence for the investigated subject. The discussed method seems to be appropriate for further studies in this area.

References

[1] Bühlmann P. (1997). *Sieve bootstrap for time series.* Bernoulli **3** (3), 123 – 148.

[2] Davison A.C., Hinkley D.V. (1997). *Bootstrap methods and their application.* Cambridge University Press.

[3] Efron B. (1979). *Bootstrap methods: Another look at the jackknife.* Ann. Statist. **7** 1 – 26.

[4] Fisher H.I., Hall P. (1990). *On bootstrap hypothesis testing.* Austral. J. Statist. **32** 177 – 190.

[5] Götze F., Künsch H.R. (1996). *Second-order correctness of the blockwise bootstrap for stationary observations.* Ann. Statist. **24** 1914 – 1933.

[6] Hall P. (1992). *The bootstrap and the edgeworth expansion.* Springer.

[7] Hall P., Horowitz J.L., Jing B. (1995). *On blocking rules for the bootstrap with dependent data.* Biometrika **82**, 561 – 574.

[8] Hall P., Martin M.A. (1988). *On bootstrap resampling and iteration.* Biometrika **75**, 661 – 671.

[9] Hall P., Wilson S.R. (1991). *Two guidelines for bootstrap hypothesis testing.* Biometrics **47**, 757 – 762.

[10] Künsch H.R. (1989). *The jackknife and the bootstrap for general stationary observations.* Ann. Statist. **17**, 1217 – 1241.

[11] Lahiri S.N. (1996). *On Edgeworth expansion and moving block bootstrap for studentized M-estimators in multiple linear regression models.* J. Multivariate Anal. **56**, 42 – 59.

[12] Liu R.Y., Singh K. (1992). *Moving blocks jackknife and boostrap capture weak dependence.* In R. Lepage and L. Billard (eds), Exploring the Limits of Bootstrap, 225 – 248. Wiley, New York.

[13] Naik-Nimbalkar U.V., Rajarshi M.B. (1994). *Validity of blockwise bootstrap for empirical processes with stationary observations.* Ann. Statist. **22**, 980 – 994.

[14] Oppenheim A.V., Schafer R.W. (1975). *Digital signal processing.* Prentice-Hall, Englewood Cliffs, NJ.

[15] Sherman M. (1998). *Efficiency and robustness in subsampling for dependent data.* J. Statist. Plan. & Infer. **75** 133 – 146.

[16] Timmer J., Lauk M., Vach W., Lücking C.H. (1999). *A test for a difference between spectral peak frequencies.* Comp. Stat. & Data Anal. **30**, 45 – 55.

Acknowledgement: Martin Peifer received financial support form the Deutsche Forschungsgemeinschaft (DFG).

Address: M. Peifer, J. Timmer, Freiburg Center for Data Analysis and Modelling, Eckerstr. 1, 79104 Freiburg, Germany

E-mail: `peifer@fdm.uni-freiburg.de`

SAMPLE SIZE DETERMINATION IN THE BAYESIAN ANALYSIS OF THE ODDS RATIO

Thu Pham-Gia and Noyan Turkkan

Key words: Bayesian approach, odds ratio, highest posterior density, sample size, 2 × 2 table.

COMPSTAT 2004 section: Bayesian statistics.

Abstract: The odds ratio is a measure of association between 2 proportions π_1 and π_2, related to two populations. Using two prior betas, and new results in distribution theory, we compute the exact minimum double-sample size (n_1, n_2) required in the Bernoulli sampling of two independent populations so that the expected length of the highest posterior density credible interval of $\Psi = \frac{\pi_1}{1-\pi_1}/\frac{\pi_2}{1-\pi_2}$ is less than a preset quantity. Other criteria commonly used in Bayesian Statistics, and in Bayesian Decision Theory, such as the Bayes risk and the Value of Sample Information will also be considered.

1 Introduction

The simplest form of the odds ratio is defined on a 2 × 2 contingency table, where π_1 and π_2 are, respectively, the independent proportions of elements in each population belonging to the first column. The odds coefficient in the first population, is $\pi_1/(1-\pi_1)$, and the one in the second is $\pi_2/(1-\pi_2)$, resulting in the odds ratio

$$\Psi = \frac{\pi_1}{(1-\pi_1)}/\frac{\pi_2}{(1-\pi_2)} \tag{1}$$

when both populations are considered. Immediate generalizations of (1) include the $(2\times n)$ table , and the $(2\times n\times k)$ table, where k can be considered as the number of strata [1]. Related results on the estimation of Ψ are obtained by Lui [6]. Alternately, we can also consider a vector $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ with $\sum_{i=1}^{2}\sum_{j=1}^{2}\pi_{ij} = 1$, and allow some dependence between π_{ij}, and define the odds ratio as

$$\Psi = \frac{\pi_{11}}{\pi_{21}}/\frac{\pi_{22}}{\pi_{12}} \tag{2}$$

When a Bayesian approach is adopted for the study of Ψ, according to (1), two beta independent distributions are considered for $\pi_1$and π_2, respectively [5], whereas for (2), a Dirichlet distribution is given to the vector $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$. Exact and approximate posterior distributions for this vector have been obtained by several authors [4], [3]. In this article we consider case 1 only, and will compute the sample size (n_1, n_2) required so that the $100(1-\xi)\%$ highest posterior density (hpd) interval of Ψ is less than a preset quantity. Similar results related to case 2 will be presented in a subsequent paper. In order to carry out the computations, we will use the exact expression of the ratio of two Beta-prime variables, established from Pham-Gia and Turkkan [8]. We encourage the readers to go through a companion paper of the present publication, Pham-Gia and Turkkan [9], where similar concepts are presented concerning the difference of two proportions. In Section 2, we first recall the exact expression of the distribution of Ψ. In Section 3, we determine the different regions in the (n_1, n_2)-plane so that the highest posterior density interval for Ψ has, respectively, its expected length, or its maximum length, less than a preset quantity. An efficient program developed by the authors is available upon request. In Section 4, the same approach is adopted in *Applied Bayesian Decision Theory*, with either the *Bayes risk* or the *Expected Value of Sample Information (EVSI),* as criterion. A numerical example, presented in Section 5, fully illustrates the results and methods presented earlier.

2 Sample size determination in Bayesian estimation theory

The sample size determination problem in *Bayesian Estimation Theory* has been the subject of several recent works, while the same problem related to *Bayesian Hypothesis Testing*, and based on Bayes Factor, attracted some attention only lately. In Bayesian Analysis, in the absence of a loss function, we consider the *highest posterior density* (denoted *hpd* in this paper from now on) credible interval and its length, which represents precision. Since data is not available prior to sampling, the sample size has to be based on either the *worst scenario* (max. or min. value of the criterion), or its average value, or another average concept, for example, the average probability coverage [2]. But only the *worst scenario* can guarantee that the precision condition will be satisfied, regardless of the sampling results. In *Bayesian Decision Theory*, when a loss function, or an utility function, is considered, attention shifts to the cost associated with the optimal decision, or more precisely, to the minimization of the posterior risk. Considering the average value, with respect to all sampling outcomes, of that posterior risk, also called the Bayes risk, we can impose a condition on that criterion [7]. On the other hand, under the quadratic loss the variance of the posterior mean represents the uncertainty in estimation, and we could be interested in knowing how large a sample size should be so that the average gain in information, or equivalently, the average loss in uncertainty, would be larger than a preset quantity. That potential gain is called the *Expected Value of Sample Information* (EVSI). Similar ideas, applied to the absolute value loss function, have lead to the use of the mean absolute deviation and related measures [10].

3 Exact prior and posterior distributions of the odds ratio

Let us consider the two independent beta random variables, denoted $\pi_i \sim beta(\alpha_i, \beta_i)$, $i = 1, 2$, with density $f(x_i) = x_i^{\alpha_i - 1}(1 - x_i)^{\beta_i - 1}/B(\alpha_i, \beta_i)$, $\alpha_i, \beta_i > 0, 0 \leq x_i \leq 1$, where $B(\alpha_i, \beta_i)$ is the beta function with parameters α_i and β_i. We know that each ratio $\frac{\pi_i}{1-\pi_i}$ then has a beta-prime distribution, with density:

$$f(\alpha_i, \beta_i; x) = \frac{x^{\alpha_i - 1}}{B(\alpha_i, \beta_i)(1 + x)^{\alpha_i + \beta_i}}, 0 \leq x, 0 < \alpha_i, \beta_i$$

Theorem 1: The density of the ratio of two independent beta-prime variables, $X_i \sim BP(\alpha_i, \beta_i), i = 1, 2$, denoted $RDF(\alpha_1, \beta_1; \alpha_2, \beta_2)$, has expression:

$$f_R(r) = KB(\alpha_1+\alpha_2, \beta_1+\beta_2)r^{\alpha_1 - 1}.{}_2F_1(\alpha_1+\alpha_2, \beta_1+\beta_2; \alpha_1+\alpha_2+\beta_1+\beta_2; 1-r), \tag{3}$$

where ${}_2F_1$ is Gauss Hypergeometric function in one variable, and

$$K = \frac{\prod_{i=1}^{2}\Gamma(\alpha_i + \beta_i)}{\prod_{i=1}^{2}\Gamma(\alpha_i)\Gamma(\beta_i)} \cdot \frac{\Gamma(\sum_{i=1}^{2}\alpha_i)\Gamma(\sum_{i=1}^{2}\beta_i)}{\Gamma(\sum_{i=1}^{2}(\alpha_i + \beta_i))}$$

Proof(summary): The proof follows the proof of the main theorem in Pham-Gia and Turkkan [8]. The beta-prime is a special case of the Generalized-F variables, and the expression of the density of the ratio of two such variables has been established there. Furthermore, it was established in Pham-Gia and Turkkan [9], that, if separate independent sampling results (n_1, x_1) and (n_2, x_2) are obtained for π_1 and π_2 , respectively, then the posterior distribution of Ψ is $RDF(\alpha_1^*, \beta_1^*; \alpha_2^*, \beta_2^*)$, as given by (1), where $\alpha_i^* = \alpha_i + x_i$ and $\beta_i^* = \beta_i + n_i - x_i$, $i = 1, 2$. Also, the predictive distribution of (X_1, X_2) with $0 \leq X_1 \leq n_1$, and $0 \leq X_2 \leq n_2$, can be obtained as the direct product of the separate beta-binomial distributions of X_1 and X_2. Hence, the hpd interval for Ψ can first be computed, using the expression of $RDF(\alpha_1^*, \beta_1^*; \alpha_2^*, \beta_2^*)$ and an algorithm developed by Turkkan and Pham-Gia [11] for the bivariate case. The expected value of its length can then be obtained.

4 The length of the highest posterior credible interval as a criterion

4.1 The general approach

Our approach in solving the sample size problem is computer-based, and can be thought as consisting of two steps: first, the computation of all numerical values of the chosen criterion in the quarter-plane ($n_1 \geq 0$ and $n_2 \geq 0$) (called the "*mapping out*" of that region) and second, the search for the numerical solution of an equation relating that criterion to the precision condition.

4.2 The expected length of the hpd interval as a criterion (ELC)

The hpd $(1-\xi)100\%$ credible interval can be obtained by solving a set of equations and, depending on that posterior, it could be a single interval, or a set of disjoint intervals. Let $beta(\alpha_1, \beta_1)$ and $beta(\alpha_2, \beta_2)$ be the two given prior distributions assigned to π_1 and π_2, respectively. We now proceed in two major steps.

A) "*Mapping out*" the quarter-plane:

a) We consider two cases:

1. *Small sample sizes $n_1$and n_2*: First, for ξ, n$_1$ and n$_2$ fixed, with $0 \leq n_1, n_2 \leq 30$, and for each pair (x_1, x_2) such that $0 \leq x_1 \leq n_2$ and $0 \leq x_2 \leq n_2$, we apply the above mentioned algorithm to compute the $(1-\xi)100$ % hpd interval for Ψ, using the expression of its posterior $RDF(\alpha_1^*, \beta_1^*, \alpha_2^*, \beta_2^*)$. Let $(\alpha_1, \beta_1, \alpha_2, \beta_2; n_1, x_1, n_2, x_2)$ be its total length.

2. *Large sample sizes n_1 and n_2*: For n_1 or n_2 larger than 30, we use normal approximation, since the posterior distribution of Ψ approaches quickly the normal distribution $N(\mu^{(n_1,n_2,x_1,x_2)}, \sigma^2_{(n_1,n_2,x_1,x_2)})$ where

$$\mu^{(n_1,n_2,x_1,x_2)} = \exp\left(\frac{1}{\alpha_2^*} + \frac{1}{\beta_1^*} - \log\frac{\beta_1^*\alpha_2^*}{\alpha_1^*\beta_2^*}\right),$$

and variance

$$\begin{aligned}\sigma^2_{(n_1,n_2,x_1,x_2)} &= \exp\left(3\left(\frac{1}{\alpha_2^*} + \frac{1}{\beta_1^*}\right) + \frac{1}{\alpha_1^*} + \frac{1}{\beta_2^*} - 2\log\frac{\beta_1^*\alpha_2^*}{\alpha_1^*\beta_2^*}\right) \\ &\quad - \exp\left(\frac{1}{\alpha_2^*} + \frac{1}{\beta_1^*} - 2\log\frac{\beta_1^*\alpha_2^*}{\alpha_1^*\beta_2^*}\right).\end{aligned}$$

At the $(1-\xi)100\%$ credible level, $\ell(\alpha_1, \beta_1, \alpha_2, \beta_2; n_1, x_1, n_2, x_2)$ is then simply $2z_{1-\xi/2}\sigma_{(n_1,n_2,x_1,x_2)}$, where $z_{1-\xi/2}$is the $(1-\xi/2)$-th percentile of the standard normal.

b) Secondly, we compute the expectation of this length,

$$\begin{aligned} K(n_1,n_2) &= E_{x_1,x_2}[\ell(\alpha_1,\beta_1,\alpha_2,\beta_2;n_1,x_1,n_2,x_2)] \\ &= \sum_{x_2=0}^{n_2}\sum_{x_1=0}^{n_1}\ell(\alpha_1,\beta_1,\alpha_2,\beta_2;n_1,x_1,n_2,x_2)g_1(x_1)g(x_2), \end{aligned} \tag{4}$$

using the marginal beta-binomial distributions of X_1 and X_2,

$$g(x_i)=\left(\begin{array}{c} n_i \\ x_i \end{array}\right)B(\alpha_i+x_i,\beta_i+n_i-x_i)/B(\alpha_i,\beta_i), 0\le x_i\le n_i, i=1,2$$

c) Thirdly, we let n_1 and n_2 vary from 1 to N, for N sufficiently large, with the choice of N depending naturally on the problem considered. We now have a surface S$_1$, with vertical coordinate $K(n_1,n_2)$, function of n_1 and n_2, with $0\le n_1,n_2\le N$.

B) *Sample sizes determination:* We can see that $K(n_1,n_2)$ is an decreasing function of n_1 and n_2 separately, considered here as continuous variables, and hence, if η_1 is the desired average precision, we only need to numerically solve: $K(n_1,n_2)\le\eta_1$. The sample size (n_1,n_2)required is then located in the region R$_1$ bounded below by the curve Ω_1. We can also fit a parametric curve Δ_1 to Ω_1.

4.3 The maximum length as a criterion (MLC)

In the above approach, if we take

$$M(n_1,n_2)=\max_{x_1,x_2}[l(\alpha_1,\beta_1,\alpha_2,\beta_2;n_1,x_1,n_2,x_2)],$$

$0\le x_1\le n_1, 0\le x_2\le n_2$, then we have the “worst of cases” criterion. Again, we use normal approximation for large sample sizes. The surface showing $M(n_1,n_2)$in function of (n_1,n_2) is S_2. Solving inequality: $M(n_1,n_2)\le\eta_2$ we have the sample region R$_2$ limited below by a curve Ω_2 in the (n_1,n_2) plane, similarly to the previous case.

5 The Bayesian decision theory criteria

In Bayesian Decision Theory, when the quadratic loss function is used , i.e. $L(\theta,a)=K|\theta-a|^2$, $K>0$, as presented in Pham-Gia and Turkkan [7], variances are of particular significance since they represent uncertainties in the decision process. Without any loss in generality, let $K=1$.

5.1 The Bayes risk for Ψ as a criterion(BRC)

With data not collected yet, the posterior risk itself is a random variable, with mean equal to the expected value w.r.t. data of the posterior variance (called the Bayes risk, $\rho(n)$) when the optimal decision is made. Hence, for the two-dimensional case, we have

$$\begin{aligned}\rho(n_1,n_2) &= E_{x_1,x_2}[\mathrm{VAR}_{\mathrm{post}}^{(x_1,x_2)}(\Psi)] \\ &= \sum_{x_2=0}^{n_2}\sum_{x_1=0}^{n_1}\left[\int_0^\infty r^2\mathrm{RDF}(\alpha_1^*,\beta_1^*,\alpha_2^*,\beta_2^*;r)dr\right. \\ &\quad \left.-\left(\int_0^\infty r\mathrm{RDF}(\alpha_1^*,\beta_1^*,\alpha_2^*,\beta_2^*;r)dr\right)^2\right]g_1(x_1)g_2(x_2)\end{aligned} \tag{5}$$

Numerically computing $\rho(n_1,n_2)$ in function of (n_1,n_2) gives surface S_3. Inequality: $\rho(n_1,n_2)\le\eta_3$ leads to the corresponding sample region R_3 bounded below by the level curve Ω_3, which can, again, be determined numerically.

5.2 The expected value of sample information as criterion (SIC)

For the single proportion case, we know that the posterior mean of the parameter Θ is itself a random variable, with mean equal to the prior mean , i.e. $E(\mu_{\mathrm{post}}^{\Theta})=\mu_{\mathrm{prior}}^{\Theta}$. On the other hand, its variance, which also represents the amount of information that still can be acquired by sampling, is an increasing function of n (see Pham-Gia and Turkkan [7]). A condition on EVSI(n) is in an order opposite to that of the previous criteria, i.e. we are now interested on the sample size required so that the related value of information is larger than a preset quantity. $\mathrm{EVSI}(n_1,n_2)=\mathrm{EVPI}-\rho(n_1,n_2)$ is given by surface S_4, where $\mathrm{EVPI}=\mathrm{Var}_{\mathrm{prior}}(\Psi)$ can be easily obtained by numerical integration of (3). Hence, the condition $\mathrm{EVSI}(n)\ge\eta_4$ gives the sample region R_4 for (n_1, n_2), bounded by the curve Ω_4, determined numerically.

6 A numerical example

We give a simple but representative example, where the numerical values of the criteria presented in the previous two sections are computed, for all values of n_1 and n_2 between 0 and 120. A significant amount of computer time is hence necessary to implement the first step of our approach because of the complex computations involved, and the use of precise analytic tools only. In general, depending on the prior distributions selected, posterior distributions can differ significantly. Results related to non-informative priors, and other priors of specific types, can be obtained from the authors.

Let's consider the quite general case where $\pi_1\sim beta(4.65,28.64)$ and $\pi_2\sim beta(36.26,5.67)$, with π_1 and π_2 independent. Hence, Ψ has a RDF(4.65, 28.64,36.26, 5.67) distribution, as given by (3). Let $\xi=0.05$ (or 95% hpd region). Fig.1 a) gives surface S_1 representing $\mathrm{K(n_1,\,n_2)}$, as given by (4), as a function of $(\mathrm{n_1,\,n_2})$, with $0\le\mathrm{n_1,\,n_2}\le120$, where $\mathrm{K(n_1,\,n_2)}$ is the average length at $(\mathrm{n_1,\,n_2})$ of the hdp interval for Ψ. $\mathrm{S_1}$is a convex surface, sloping downward as $\mathrm{n_1}$and $\mathrm{n_2}$ increase. Let's require, for example, $\mathrm{K(n_1,}$

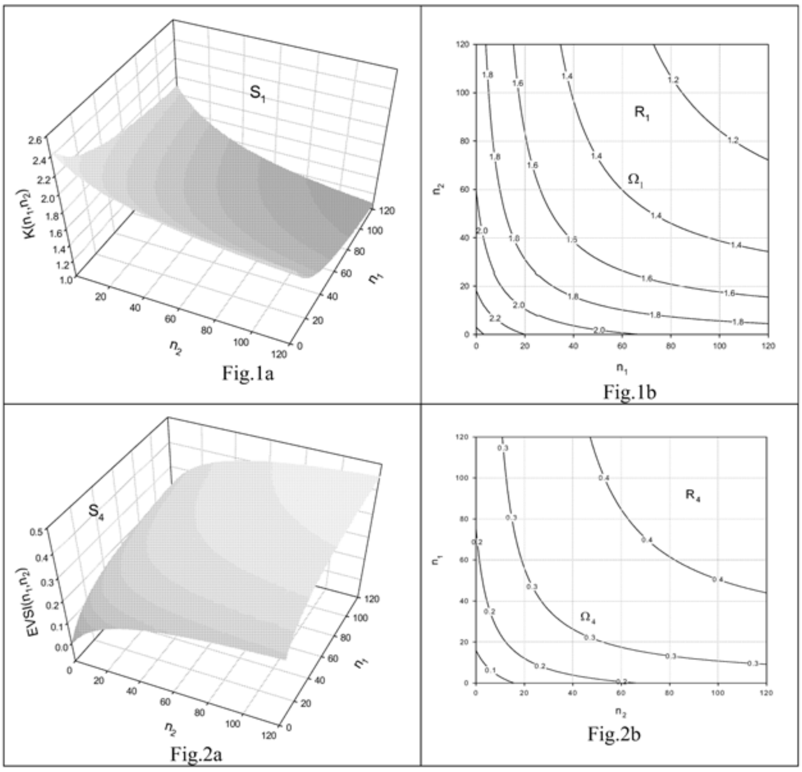


Fig.1a

Fig.1b

Fig.2a

Fig.2b

$n_2) \leq 1.40$. Solving numerically, we have the open region R_1, bounded below by the curve Ω_1 which is the intersection of S_1 and the horizontal plane with vertical coordinate 1.40. In Fig. 1 b), we can see that for $n_1 = 80$ and $n_2 = 60$ the mean length of the 95p.c. credible interval for the odds ratio is less than 1.4. Other level curves corresponding to different average lengths of the hpd interval , have been similarly determined and constitute a series of smooth convex curves intersecting the two axes at $(n_1,0\ )$ and $(0,n_2)$. These curves get more symmetric w.r.t. the first diagonal as n_1 and n_2 increase, reflecting the convergence of the posterior toward the normal. Similar conclusions and discussions hold for $M(n_1,\ n_2)$ and surface S_2 , and $\rho(n_1,\ n_2)$ and surface S_3, that have similar shapes, and are not shown here. EVSI(n_1, n_2), however, has its surface S_4 concave, sloping upward, reflecting the increase in average sample value, as n_1 and n_2 increase (Fig. 2a) and 2b)).

7 Conclusion

In view of numerous applications of the odds ratio in the life sciences, and even in a more general theoretical context of joint probability distributions [12], the above results could find real applications in other domains in the near future. On the other hand, generalizations of this study to the Generalized-F priors, as studied by Pham-Gia and Turkkan [8], and to other forms of odds-ratios, would follow essentially the same lines.

References

[1] Agresti A. (1980). *Generalized odds ratio for ordinal data.* Biometrics **36**, 59 − 67.

[2] Joseph L., Wolfson D.B., du Berger R. (1995). *Sample size calculations of binomial proportions via highest posterior density intervals.* The Statistician **44**, 143 − 154.

[3] Kateri M., Papaioannou T., Dellaportas P.(2001). *Bayesian analysis of correlated proportions.* Sankhya **63, series B**, 270 − 285.

[4] Latorre G. (1982). *The exact posterior distribution of the cross-ratio of a* 2×2 *contingency table.* J. Stat. Comput. Simul. **16**, 19 − 24.

[5] Lee P. (1998). *Bayesian statistics: an introduction.* 2^{nd} ed. London: Arnold.

[6] Lui K.J. (2002). *Interval estimation of generalized odds ratio in data with repeated measurements.* Stat. In Medicine, 3107 − 3117.

[7] Pham-Gia T., Turkkan N. (1992). *Sample size determination in Bayesian statistics.* The Statistician **41**, 389 − 397.

[8] Pham-Gia T., Turkkan N. (2002). *Operations on the generalized F-variables and applications.* Statistics **36**, 195 − 209.

[9] Pham-Gia T., Turkkan N.(2003). *Determination of the exact sample sizes in the Bayesian estimation of the difference of two proportions.* The Statistician **52**, 131 − 150.

[10] Pham-Gia T., Turkkan N. (2004). *Bayesian decision criteria in the presence of noises, under quadratic and absoute value loss functions.* Statistical Papers, to apear.

[11] Turkkan N., Pham-Gia T. (1997). *Highest posterior density credible region and minimum area confidence region: the bivariate case.* Journ. Royal Stat. Soc. **46, series C**, 131 − 140.

[12] Van Der Linde A., Osius G. *Discrimination based on an Odds Ratio Parametrization.* In Bayesian Statistics **7**, j.m. Bernardo, Ed.,Wiley, to appear.

Acknowledgement: Thanks to CRSNG grant A9249 (Canada) for its generous support, and to two referees for very helpful comments.

Address: T. Pham-Gia, N. Turkkan, Universite de Moncton, Department of Math/Stat, Moncton, NB, Canada E1A3E9

E-mail: phamgit@umoncton.ca

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

THE LEAST WEIGHTED SQUARES ESTIMATOR

Pavel Plát

Key words: Robust regression, the least weighted squares, algorithm, numerical examples.

COMPSTAT 2004 section: Robustness.

Abstract: The paper begins with a recapitulation of theoretical properties of the least weighted squares (LWS). The asymptotic normality and the asymptotic representation in the framework of nonrandom carriers is presented. An algorithm for evaluating the LWS estimator is proposed. LWS is compared with the least median of squares (LMS) and the least trimmed squares (LTS) by way of two numerical examples.

1 Theoretical results

First, let us introduce notations. Let N denotes the set of all positive integers, R denotes the set of all real numbers, and R^p denotes the p-dimensional Euclidean space. We will consider for any $n \in N$ the linear regression model

$$Y_i = x_i^T \beta^0 + e_i, \quad i = 1, 2, \ldots, n$$

which can be rewritten equivalently in usual matrix notation

$$Y = X\beta^0 + e \tag{1}$$

where $Y = (Y_1, Y_2, \ldots, Y_n)^T$ $(Y_i \in R)$ is the response variable, $X = (x_1, x_2, \ldots, x_n)^T$ $(x_i \in R^p)$ is the design matrix, β^0 is the "true" vector of regression coefficients, and finally $e = (e_1, e_2, \ldots, e_n)$ $(e_i \in R)$ is the vector of random fluctuations (disturbances). We will consider model with nonrandom explanatory variables, i.e. x_i's are nonrandom vectors from R^p, and let us notice that in the case when the intercept is included in the model, the first coordinates of all vectors x_i's are assumed to be equal to 1.

For any $\beta \in R^p$ let us put $r_i(\beta) = Y_i = x_i^T \beta$, i. e. $r_i(\beta)$ denotes the i-th residual when we assume β to be the vector of regression coefficients. Further, the order statistics of squared residuals will be denoted by $r^2_{(i)}(\beta)$, $i = 1, 2, \ldots, n$. To be more explicit, it means that for any $\beta \in R^p$

$$0 \le r^2_{(1)}(\beta) \le r^2_{(2)}(\beta) \le \cdots \le r^2_{(n)}(\beta).$$

Now, let us recall the definition of the least weighted squares estimator.

Definition 1.1. *A nonincreasing continuous function* $w : [0,1] \to [0,1]$ *such that* $w(0) = 1$ *and* $w(1) = 0$ *will be called the weight function.*

Definition 1.2. *Let $\mathcal{K} \subset R^p$ is a compact set such that $\beta^0 \in \mathcal{K}^o$, and w is a weight function. The estimator given as*

$$\hat{\beta}^{(LWS,n,w)} = \arg\min_{\beta \in \mathcal{K}} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r_{(i)}^2(\beta) \qquad (2)$$

will be called the least weighted squares (LWS).

The following theorem brings the asymptotic representation of the least weighted squares estimator. The proof of the theorem can be found in [5]. Some assumptions as follows are necessary for the proof.

Assumptions $\mathcal{A}$ *The sequence of disturbances $\{e_i\}_{i=1}^{\infty}$ is a sequence of independent identically distributed random variables. The distribution function F of the random fluctuation e_1 is absolutely continuous with a bounded density f which is positive on R, and has bounded first derivative. Moreover, $I\!\!E e_1^4 = \kappa_4 \in R^+$.*

Further, the sequence $\{x_i\}_{i=1}^{\infty}$ is a fix sequence of nonrandom vectors from R^p such that $\sum_{i=1}^{n} \|x_i\|^4 = \mathcal{O}(n)$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = Q$ where $Q \in R^{p,p}$ is a regular matrix (and convergence is of course assumed coordinatewise).

Further, let G denotes distribution function of e_1^2, and for any $\alpha \in [0,1]$, u_α^2 is the upper α-quantile of $G(z)$, i. e. $P(e_1^2 > u_\alpha^2) = 1 - G(u_\alpha^2) = \alpha$. Further, let us define $\sigma_z^2 = \int\limits_{-u_z}^{u_z} z^2 \mathrm{d}F(z)$.

Theorem 1.1. *Let Assumption $\mathcal{A}$ be fulfilled. Further, let $\mathcal{K} \subset R^p$ is a compact set such that $\beta^0 \in \mathcal{K}^o$, and w is a weight function. Then*

$$\sqrt{n}\left(\hat{\beta}^{(LWS,n,w)} - \beta^0\right) = \mathcal{O}_p(1),$$

and $\hat{\beta}^{(LWS,n,w)}$ is asymptotically normal with mean value equal to β^0 and covariance matrix

$$V\left((\hat{\beta}^{(LWS,n,w)}, F\right) = -\frac{2\int\limits_0^1 \sigma_{1-z}^2 w(z)\mathrm{d}w(z)}{\left(\int\limits_0^1 \left(z - 2u_{1-z} f\left(u_{1-z}\right)\right) \mathrm{d}w(z)\right)^2} \cdot Q^{-1},$$

i. e.

$$\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(LWS,n,w)} - \beta^0\right)\right) \to \mathcal{N}\left(\ 0\ ,\ V\left(\hat{\beta}^{(LWS,n,w)}, F\right)\ \right) \qquad \text{as } n \to \infty.$$

The least weighted square estimator will be compared with the least trimmed squares estimator and the least median of squares estimator in examples in the section 3. Therefore, let us recall also their definitions.

Definition 1.3. *Let $\mathcal{K} \subset R^p$ is a compact set such that $\beta^0 \in \mathcal{K}^o$. The estimators given as*

$$\hat{\beta}^{(LTS,n,h)} = \arg\min_{\beta \in \mathcal{K}} \sum_{i=1}^{h} r_{(i)}^2(\beta), \tag{3}$$

$$\hat{\beta}^{(LMS,n,h)} = \arg\min_{\beta \in R^p} r_{(h)}^2(\beta) \tag{4}$$

will be called the least trimmed squares (LTS) and the least median of squares (LMS), respectively.

2 The algorithm

Now, we turn our attention to the problem how to evaluate the least weighted square estimator. It is easily to see that (2) can be rewritten as

$$\hat{\beta}^{(LWS,n,w)} = \arg\min_{\beta \in \mathcal{K}, \pi \in \Pi_n} \sum_{i=1}^{n} w\left(\frac{\pi(i)-1}{n}\right) r_i^2(\beta) \tag{5}$$

where Π_n denotes the set of all permutations of the set $\{1, 2, \ldots, n\}$. It is clear from (5) that precise solution of extremal problem (2) can be found by means of a complete inspection over all permutations $\pi \in \Pi_n$. However, it is possible only for small values of n (say up to 20). Now, we are going to describe algorithm (see [3]) which is suitable for evaluation of an approximation to the precise solution of (2). This iterative algorithm is a generalization of the algorithm given in [7] or [8] for the least trimmed squares estimator.

1. Select randomly p points and find the regression plane going through them.
2. Evaluate residuals for all points with respect to this regression plane and define $\{k_i\}_{i=1}^n$ such that $r_{k_i}^2(\beta) = r_{(i)}^2(\beta)$ for all $i \in \{1, 2, \ldots, n\}$.
3. Evaluate the sum $S_n^{(j)} = \sum_{i=1}^n w(\frac{k_i-1}{n}) r_i^2(\beta)$.
4. In the case that the sum $S_n^{(j)}$ is smaller then the sum $S_n^{(j-1)}$ from the previous step, find a new hyperplane by using the (classical) weighted least square estimator with weights $w(\frac{k_1-1}{n})$, $w(\frac{k_2-1}{n})$, $\ldots$, $w(\frac{k_n-1}{n})$, then go to 2. Otherwise go to 5.
5. In the case that the same model has been found repeatedly (20 times) (with the smallest sum of weighted residuals) or an a priori given number of repetitions (say 10000) has been already accomplished, finish the algorithm. Otherwise go to 1.

Let us add some remarks. First, even if for deriving of theoretical results is important difference between regression models with random and nonrandom explanatory variables, this algorithm can be used for both these models.

Further, due to the fact that the continuity of weighted function is not necessary, it is easy to see that this algorithm is really generalization of the algorithm for the least trimmed squares estimator. It is sufficient to define $w(x)$ equal to 1 for $x \in [0, h/n)$, and equal 0 for $x \in [h/n, 0]$.

Finally, let us notice that the algorithm from [1] which uses dual formulation of the problem of linear programming and simplex method will be used in the case of the least median squares estimator. The algorithm appeared to give really tight approximation to the precise solution of the extremal problem (4) (see again [7] or [8]).

3 Numerical examples

Now, we are going to present two examples which enlighten some properties of the least weighted squares.

The *Educational data* have been chosen for the first example since they are well known their analysis by several authors are available, see [2], [6], [7] or [8]. We compare and comment on the results obtained by the least median of squares, the least trimmed squares and the least weighted squares estimators. Some comment will be also made.

The second example demonstrates by way of the set of simulated data the diference between *subsample sensitivity* of the least trimmed squares and the least weighted squares. In fact, the subsample stability of the least trimmed squares is quite low, see [8]. The second example shows that the behaviour of the least weighted squares (of course, with continuous weighted function w, see Definition 1.1) can be markedly better.

First of all, let us specify the weight function we will use

$$w(x) = \min\left\{1, \max\left\{\frac{(kb-k-1)\,x - kab + ka + b}{b-a}\ ,\ kx - k\right\}\right\} \qquad (6)$$

(see also Figure 1) where $a \in [\ 0\ ,\ 1\ )$, $b \in (\ a\ ,\ 1\ ]$ a $k \in (\ \frac{1}{a-1}\ ,\ 0\ ]$ are parameters by which we can modify properties of the least weighted square estimator.

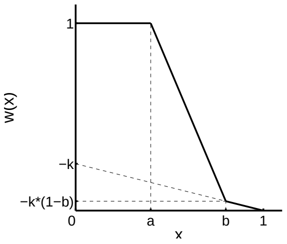


Figure 1: Weighted function.

Example 1. (Educational Data, 50 cases.) (see [2], [6], [7] or [8].)

The data deal with education expenditure variables for 50 U. S. states. The aim is to explain the per capita expenditure on public education in a state, projected for 1975 (Y), by number of residents per thousand residing in urban areas in 1970 (Residents), per capita personal income in 1973 (Income), and number of residents per thousand under 18 years of age in 1974 (Young). As already mentioned, we use the weight function (6). In this example we denote $w_a^{(1)}$ the function (6) where $b = \frac{49}{50}$, $k = -\frac{1}{10}$, and $w_a^{(2)}$ the function (6) where $b = \frac{40}{50}$, $k = -\frac{1}{10}$. The parameter a is chosen $\frac{44}{50}$, $\frac{34}{50}$ and $\frac{26}{50}$, respectively for both $w_a^{(1)}$ and $w_a^{(2)}$. The least median of squares and the least trimmed squares are evaluated for h equals to 45, 35 and 27, respectively. The estimates of the regression coefficients can be found in Tables 1, 2 and 3 and for completeness the ordinary least square estimate is given in Table 1. Notice that we compare results for such values of parameters h and a ($a = \frac{h-1}{n}$) because both LWS and LTS assign the weight equal to 1 just to h points for this choice.

Let us turn our attention to the last three rows of Table 1 and the last four rows of Table 2. Neither the least median of squares nor the least trimmed squares is better than the least weighted squares in the following sense. The hth order statistics among the squared residuals is smaller for $LWS_a^{(j)}$ than for LTS_h, $a = \frac{44}{50}$, $j = 1$, $h = 45$ and $\frac{34}{50}$, $j = 1, 2$, $h = 35$, and the sum of the h order statistics $r_{(i)}^2(\beta)$ is smaller for $LWS_a^{(j)}$ than for LMS_h, $a = \frac{44}{50}$, $j = 1$, $h = 45$ and $a = \frac{34}{50}$, $j = 2$, $h = 35$. Of course, the sum of weighted order statistics $r_{(i)}^2(\beta)$ is the smallest for LWS with corresponding weight function.

Method	LS	LMS_{45}	LTS_{45}	$LWS^{(1)}_{44/50}$
Intercept	-556.6	-213.2	-267.8	-266.9
Residents	-0.004	0.064	0.073	0.073
Income	0.072	0.039	0.044	0.046
Young	1.55	0.880	0.897	0.874
$\lvert r\rvert_{(45)}$		49.22	61.44	58.52
$\sum_{i=1}^{45} r^2_{(i)}$		45845	35787	35998
$\sum_{i=1}^{n} w^{(1)}(\frac{i-1}{n})r^2_{(i)}$		54687	46173	45948

Table 1: Results - LS, LMS_h, LTS_h ($h = 45$), $LWS_a^{(1)}$ ($a = \frac{44}{50}$, $b = \frac{49}{50}$).

Further, turn our attention to Table 3. In the table we can see that for the smallest considered h and a (27 and $\frac{26}{50}$) there is large difference between LMS and LTS on one hand and LWS on the other hand. Compare, please, the values in the last four rows of Table 3. The 27th order statistics among the squared residuals as well as the sum of the 27 order statistics $r_{(i)}^2(\beta)$ is now larger for both $LWS^{(1)}_{\frac{26}{50}}$ and $LWS^{(2)}_{\frac{26}{50}}$ than for LMS_{27} and LTS_{27} and

Method	LMS_{35}	LTS_{35}	$LWS^{(1)}_{34/50}$	$LWS^{(2)}_{34/50}$
Intercept	-283.3	-196.4	-281.4	-241.7
Residents	0.180	0.110	0.094	0.113
Income	0.033	0.033	0.042	0.036
Young	0.875	0.722	0.919	0.825
$\lvert r\rvert_{(35)}$	29.388	38.457	36.544	35.198
$\sum_{i=1}^{35} r^2_{(i)}$	13068	11367	13504	11908
$\sum_{i=1}^{n} w^{(1)}(\frac{i-1}{n})r^2_{(i)}$	33063	33436	30038	31445
$\sum_{i=1}^{n} w^{(2)}(\frac{i-1}{n})r^2_{(i)}$	18564	16974	17848	16501

Table 2: Results - LMS_h, LTS_h $(h = 35)$, $LWS_a^{(1)}$, $LWS_a^{(2)}$ $(a = \frac{35}{50}$, $b^{(1)} = \frac{49}{50}$, $b^{(2)} = \frac{40}{50})$.

Method	LMS_{27}	LTS_{27}	$LWS^{(1)}_{26/50}$	$LWS^{(2)}_{26/50}$
Intercept	-210.8	-143.5	-256.6	-210.7
Residents	0.040	0.043	0.094	0.116
Income	0.043	0.035	0.042	0.034
Young	0.742	0.639	0.832	0.744
$\lvert r\rvert_{(27)}$	16.65	19.04	23.02	23.83
$\sum_{i=1}^{27} r^2_{(i)}$	3728.6	3414.5	5061	4435.6
$\sum_{i=1}^{n} w^{(1)}(\frac{i-1}{n})r^2_{(i)}$	29754	30163	22463	23609
$\sum_{i=1}^{n} w^{(2)}(\frac{i-1}{n})r^2_{(i)}$	14479	14662	12347	11637

Table 3: Results - LMS_h, LTS_h $(h = 27)$, $LWS_a^{(1)}$, $LWS_a^{(2)}$ $(a = \frac{26}{50}$, $b^{(1)} = \frac{49}{50}$, $b^{(2)} = \frac{40}{50})$.

vice versa the sums of weighted order statistics $r^2_{(i)}(\beta)$ are larger for both LMS_{27} and LTS_{27} than for $LWS^{(1)}_{\frac{26}{50}}$ and $LWS^{(2)}_{\frac{26}{50}}$. The reason for this can be found in the fact that the least median of squares and the least trimmed squares choose only 27 observations and for these observations they find the best fitting model. On the other hand, the least weighted squares (although it assign to some observations weight less than 1) take always into account all observations.

Example 2. (Simulated Data, 13 cases.) In this example we will consider the data from the Table 4 which were generated by the model

$$Y_i = 4 - 0.5x_i + e_i$$

where $\mathcal{L}(e_i) = \mathcal{N}(0, 0.25)$, $i = 1, 2, \ldots 12$ and point 13 was contaminated.

Let us look at Tables 5 and 6. In the first one there are estimates of the regression coefficients when all observations are considered and LS, LTS_h $(h = 11$, 9 and 7) and LWS_a $(a = \frac{10}{13}$, $\frac{8}{13}$ and $\frac{6}{13})$.

Index	1	2	3	4	5	6	7	8	9	10	11	12	13
x	0.481	0.887	1.18	1.38	1.78	1.91	2.27	2.66	2.80	3.13	3.44	3.74	7.02
Y	3.86	3.36	3.55	3.44	2.94	2.93	2.55	2.35	2.80	2.27	2.33	2.16	1.11

Table 4: Simulated data.

Method	LS	LTS_{11}	LTS_9	LTS_7	$LWS_{10/13}$	$LWS_{8/13}$	$LWS_{6/13}$
Intercept	3.79	3.99	3.51	3.59	3.82	3.91	3.92
$\hat{\beta}_x$	-0.417	-0.533	-0.356	-0.365	-0.417	-0.470	-0.474

Table 5: Results - the full data.

Method	LS^{n-1}	LTS_{11}^{n-1}	LTS_9^{n-1}	LTS_7^{n-1}	$LWS_{10/12}^{n-1}$	$LWS_{8/12}^{n-1}$	$LWS_{6/12}^{n-1}$
Intercept	3.79	4.01	4.14	4.33	3.80	3.83	3.92
$\hat{\beta}_x$	-0.417	-0.534	-0.637	-0.746	-0.415	-0.414	-0.477

Table 6: Results - the data without point 11.

As the weight function is used again the function (6). This time with $b = 1$. Now, let us delete the observation number 11 and evaluate once again the estimates for the same values of parameters h, a and b. The results can be found in the second of the mentioned tables. In the case of LTS_h the differences between the results for the full data and for data without point 11 are rather large (especially for smaller values of h) whereas in the case of LWS the results are similar both with and without point 11. See also Figure 2 where "the worst results" of both method are shown. Let us try explain the behaviour of LTS_h. As already said, LTS_h select h observations and find the best fitting model for them. In the case of the full data set, LTS_h includes the 11th observation among those considered as "noncontaminated" for all values of $h = 7, 8, \ldots, 12$. After deleting point 11, LTS_h have to include other observations among the h points, and that is the reason of rather different results. On the other hand, LWS work always with all observations. Although the influence of some of them is reduced by small weights, they have still some influence on the results and hence the similar results both with and without point 11 are obtained.

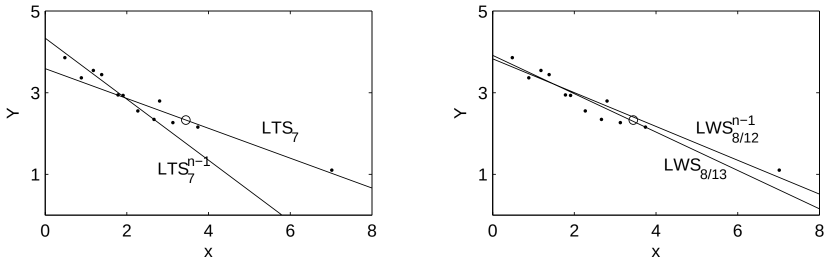


Figure 2:

(A) $LTS_7 \times LTS_7(n-1)$ (B) $LWS_{\frac{8}{13}} \times LWS_{\frac{8}{12}}$

4 Conclusions

The least weighted squares estimator presents reasonable generalization of the least trimmed square estimator. LWS as well as LTS is *affine*, *regression* and *scale equivariant* and it is estimator with controllable (and hence possibly hight) breakdown point. Further, as we presented LWS (as well as LTS) is $\sqrt{n}$*-consistent* and *asymptotically normal*. The *asymptotic representation* is also available, see [5]. Notice, that similar result which we recalled for nonrandom explanatory variables can be found in framework of random carriers in [4]. Moreover, in Example 2 it was demonstrated that the *subsample sensitivity* of LWS can be smaller than in the case of LTS. Precise evaluation of LWS is unfortunately possible only for small values of n and p. However simple algorithm for evaluating an approximation to the precise value of estimator was proposed. Finally, two numerical examples (one with real data set, the other with simulated data) gave a comparison between LWS, LTS and LMS methods. It revealed that LWS can be useful for the flexible data analysis.

References

[1] Boček P., Lachout P.(1995). *Linear programming approach to LMS-estimation.* Memorial volume of Comput. Statist. & Data Analysis **19**, 129 – 134.

[2] Chatterjee S., Price B. (1977). *Regression analysis by example.* J. Wiley *&* Sons, New York.

[3] Kalina J. (2003). *Autocorrelated disturbances of robust regression.* Personal communicate.

[4] Mašíček L. (2003). *Diagnostika a senzitivita robustních modelů (Diagnostics and sensitivityof robust models - in Czech).* PhD disertation, Fac. of Mathematics and Physics, Charles University.

[5] Plát P. (2003). *Odhad metodou nejmenších vážených čtverců (The least weighted squares estimator - in Czech).* Degree Project, Fac. of Nuclear Sc. and Physical Eng., Czech Technical University.

[6] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* J.Wiley *&* Sons, New York.

[7] Víšek J. Á. (1996). *On high breakdown point estimation.* Computational Statistics **11**, 137 – 146.

[8] Víšek J. Á. (2000). *On the diversity of estimates.* Computational Statistics and Data Analysis **34**, 67 – 89.

Acknowledgement: Research was supported by grant of GA ČR
no. 402/03/0084

Address: P. Plát, Faculty of Nuclear Sc. and Physical Eng., Czech Technical University, Trojanova 13, CZ - 120 00 Prague 2, Czech Republic

E-mail: plat@sin.cvut.cz

A PARAMETRIC FRAMEWORK FOR DATA DEPTH CONTROL CHARTS

Giovanni C. Porzio and Giancarlo Ragozini

Key words: Nonparametric statistical process control, multivariate Shewhart chart, average run length.

COMPSTAT 2004 section: Statistical process control.

Abstract: In the framework of Multivariate Statistical Process Control, nonparametric methods appear to be particularly useful. In this work data depth multivariate control charts are considered and a parametric setting for them is introduced and discussed. The proposed framework, that is based on the *Beta* distribution family, allows us to define an appropriate likelihood ratio test and to derive Average Run Length functions for nonparametric charts.

1 Introduction

In on-line production process monitoring, the item quality should be evaluated by measuring many of its features. As a consequence, multivariate statistical process control (SPC) techniques have been developed to jointly control many involved variables, exploiting multivariate data analysis methods.

The well known T^2 statistic is still used in the common practice of parametric multivariate SPC [7], and many other contributions relying on multivariate normality have been proposed in the literature (see e.g. [3, Chap. 8], [8, Chap. 10], and references therein). However, the multivariate normal model may not be adequate to describe real process data.

Unfortunately, when normality is not verified, in more than one dimension it is difficult to make and then manage other parametric assumptions. Alternatively, nonparametric multivariate techniques can be exploited to design SPC methods.

Among the possible techniques, data depth has found growing interest in multivariate quality control. Simplicial depth has been exploited to construct a quality index and related charts [6], [4], further investigated in [14]. Half-space depth has been considered to define a chart based on a bivariate boxplot [13]. A Shewhart-type chart has been designed exploiting a kind of convex hull peeling depth [10], [12].

In this paper, we aim to define an appropriate framework to investigate data depth control charts. Depth quantile properties allow us to approach nonparametrically SPC issues, ending up with a parametric setting. The key idea is to consider the *Beta* density as a model for a data depth control measure. It adequately describes in-control and out-of-control process status, and permits us to define a Shewhart sequential quality control procedure.

We focus to the case of change in location and/or increase in spread. For such a case, an hypothesis system is proposed, a likelihood ratio test is defined, appropriate control limits and Average Run Length (*ARL*) functions are derived. Different out-of-control cases could be studied in the same framework as well.

The paper is organized as follows. We discuss first how out-of-control indices can be derived from data depth measures (Section 2). Then we motivate the introduction of the *Beta* distribution family for data depth control chart in Section 3. Section 4 defines a Shewhart chart through an appropriate likelihood ratio test, control limit, and average run length. Finally, Section 5 offers some concluding remarks and further developments.

2 Nonparametric out-of-control measures

Data depth is a function $D_F(y)$ that measures the centrality of a point $y \in \Re^k$ with respect to a given multivariate distribution F. The deepest points lie at the core of the distribution, while points with lower depth values are located in the distribution tails. Although many notions of depth are available in the literature (see for discussions: [5], [16], [9], for our purposes it is not necessary to adopt any of them.

While first applications of data depth have been multivariate center-outward ordering of data scatters, and robust estimates of location and dispersion [15], [2], more recently data depth has been used within a multivariate statistical process control setting.

Specifically, let $Y \in \Re^k$ be the vector of the quality measures to be monitored, F_0 be a given in-control multivariate distribution for Y, and $D_{F_0}(.)$ be a depth function defined on F_0. The depth function contours of the in-control distribution are defined as:

$$C(d) = \{y \in \Re^k : D_{F_0}(y) = d\}.$$

If the region $R(d) = \{y \in \Re^k : D_{F_0}(y) \geq d\}$ enclosed by the contour $C(d)$ of depth d has a probability content equal to p under F_0, and F_0 is absolutely continuous and its density function is nonzero everywhere, depth contours are coincident with the p-th center-outward quantile Q_p of F_0:

$$Q_p = \{y \in \Re^k : D_{F_0}(y) = d_p\}$$

where d_p is such that $P(Y \in R(d_p)) = p$. Center-outward quantiles define a sequence of nested convex regions of increasing depth.

Assuming that the center of the in-control distribution is the quality target to be achieved, the deepest points will correspond to items of higher quality. Therefore, with respect to the process, the outer-inward sequence of these quantiles define a sequence of increasing quality levels.

Consider now the function $p_{F_0}(.) : \Re^k \rightarrow \Re^1$:

$$p_{F_0}(y) = P(Y \in R(d_p)) = p,$$

that maps a point y, having $D_{F_0}(y) = d_p$, with the probability content p of the center-outward quantile Q_p to which it belongs. For simplicity, in the following we will write $p(y)$ for $p_{F_0}(y)$, as in our setting such probability is always evaluated through the F_0 center-outward quantiles. However, note that, while the $p(y)$ values depend on F_0, it is not necessary that $Y \sim F_0$.

For each item, if $p(y)$ is close to zero (i.e. y belongs to a deepest center-outward quantiles of F_0), the process will be considered in-control. Vice versa, if $p(y)$ is close to one, the point y will be in the distribution tails, and the process is out-of-control.

Consequently, data depth control charts can be defined through the values of $p(y)$ that can be associated to each item. Liu [4] defined some Shewhart and CUSUM charts based on a quality index that is equivalent to $1 - p(y)$. By analogy with the T^2 control scheme, Porzio and Ragozini [12] proposed a nonparametric procedure based on the center-outward quantiles, and designed a chart where the $p(y)$ values are directly plotted.

3 Defining a parametric setting

Beyond the out-of-control measure one chooses, SPC procedures generally require an inferential setting that provides rules for users to take decisions over the running process.

In a sequential quality control scheme, let $Y_1, Y_2, \dots$ be the sequence of vectors representing the quality measurements of the items produced as time goes on. The process is in control as long as all the Y_i are i.i.d. from F_0, whereas the process will be declared out of control if at an unknown time point ν the multivariate distribution of $Y_\nu, Y_{\nu+1}, \dots$ changes to $F_1 \neq F_0$, with F_1 the out-of-control distribution.

In this framework, the aim of an SPC procedure is to determine if such a change has occurred. With this goal the following hypotheses can be tested:

$$\begin{array}{ll} H_0: & Y_1, ..., Y_{\nu-1}, Y_\nu, Y_{\nu+1}, ... \sim F_0 \\ H_1: & Y_1, ..., Y_{\nu-1}, \sim F_0 \\ & Y_\nu, Y_{\nu+1}, ... \sim F_1, \end{array} \tag{1}$$

with ν unknown. The time τ, $\tau \geq \nu$, at which the change is detected (i.e. when H_0 is rejected) is called stopping time.

Assuming that the corresponding density functions f_0 and f_1 exist and are known, a likelihood ratio test (LRT) can be defined for the hypotheses in (1). In the Shewhart control scheme only the last observation Y_n is used to decide upon the process, and hence the likelihood ratio is reduced to the comparison between the Y_n densities. The stopping time τ will be the first time the log-likelihood ratio is greater than a given threshold. In brief, the stopping rule will be [1]:

$$\tau = \inf \left\{ n | \log \frac{f_1(Y_n)}{f_0(Y_n)} \geq h \right\}. \tag{2}$$

The *LRT* setting allows the study of the control procedure properties in terms of *ARL* functions. In other words, it is possible to investigate the average run length under H_0 (ARL_0), i.e. the expected time to have a false alarm $E(\tau|H_0)$, and the average run length under H_1 (ARL_1), i.e. the expected time to decide correctly that the process is out-of-control $E(\tau|H_1)$. Knowledge of these quantities is important for comparing control procedures and to correctly implement them in practice.

However, if both F_0 and F_1 are completely unknown, the LRT approach is unfeasible and nonparametric procedures have to be adopted. As a consequence, ARL functions for nonparametric schemes can generally be studied only through simulations for specific cases.

In this work, we propose a possible mapping among the undefined hypotheses in (1) and a well-defined parametric hypothesis system. This approach, that is valid for the data depth control procedures described above in Section 2, allows us to study ARL functions for some alternative hypotheses.

In order to define our setting consider, first, the following lemma.

Lemma 3.1. Let $D_{F_0}(Y)$ have a continuous distribution, and let $Y \sim F_0$. Then $p(Y)$ is uniformly distributed on $[0, 1]$.

Proof. Note that

$$\begin{aligned} p(y) \quad &= P(Y \in R(d_p)) = P(D_{F_0}(Y) \geq d_p) = P(D_{F_0}(Y)) \geq D_{F_0}(y)) = \\ &= 1 - F_{D_{F_0}(Y)}(D_{F_0}(y)) \end{aligned}$$

with $F_{D_{F_0}(Y)}(D_{F_0}(y))$ the cdf of $D_{F_0}(Y)$. From the probability integral transformation we have

$$p(Y) = 1 - F_{D_{F_0}(Y)}(D_{F_0}(Y)) \sim U(0, 1).$$

□

In other words, it holds that

$$Y \sim F_0 \qquad \Rightarrow \qquad p(Y) \sim U(0, 1).$$

That is, as long as the process is in control the $p(y)$ values will come from a uniform distribution with support in $[0, 1]$. When the process goes out-of-control the $p(y)$'s will be generated from a different distribution, obviously still supported in $[0, 1]$.

Among the univariate distributions supported in $[0, 1]$, and including the Uniform distribution as a special case, we propose to consider the $Beta(a, b)$ distribution as a reasonable model to rewrite the hypotheses in (1) as:

$$\begin{array}{ll} H_0: & p(Y_1), \dots, p(Y_{\nu-1}), p(Y_\nu), p(Y_{\nu+1}), \dots \sim Beta(1, 1) \\ H_1: & p(Y_1), \dots, p(Y_{\nu-1}), \sim Beta(1, 1) \\ & p(Y_\nu), p(Y_{\nu+1}), \dots \sim Beta(a, b) \qquad a, b \neq 1 \end{array} \tag{3}$$

Rejecting the null hypothesis in (3) implies the rejection of the null in (1). On the other hand, parameters a and b ($a, b \neq 1$) describe out-of-control distributions. As the a and b values change, the unknown F_0 change in location, in scale or both.

To provide further arguments for this setting, let us consider for the sake of simplicity the case of shift in location: F_1 differs from F_0 just in its position in the multivariate space. Let $F_1^{\Delta L}$ be such a distribution, and let us evaluate similarity between F_0 and $F_1^{\Delta L}$ in terms of coverage probability, that is in the terms of the amount of probability for which Y occurs under $F_1^{\Delta L}$ in the inner region $R(d_p)$ of F_0. We have $P(Y \in R(d_p)|F_0) = p$, and then obviously $P(Y \in R(d_p)|F_1^{\Delta L}) < p$. In particular, as long as $F_1^{\Delta L}$ goes further from F_0, this probability decreases, and hence $P(Y \in R(d_p)|F_1^{\Delta L})$ is a measure of the difference between locations. Such probability can be parametrized in terms of the width of the shift $s, s \geq 0$, through any continuous function $g_p(s)$ decreasing in s. Specifically, among the possible $g_p(s)$, if it is assumed that $P(Y \in R(d_p)|F_1^{\Delta L}) = p^{(s+1)}$, then it can be proved that $p(Y|F_1^{\Delta L}) \sim Beta(s+1, 1)$. Hence, the *Beta* parameter $a = s+1$ (when $b = 1$), can be interpreted as a measure of the shift width as well. In particular, we note that for $a = 1$ ($s = 0$), it holds that $P(Y \in R(d_p)|F_1^{\Delta L}) = p$, and $F_1^{\Delta L} = F_0$. Following the same arguments, it can be shown that the $Beta(a, 1)$ density also describes increases in spread.

Both shift in location and increase in scale are to be considered a worsening of the process quality, and we focus our attention on these cases.

4 A Shewhart chart for changes in location and increases in scale

As discussed above, a strictly increasing *Beta* density ($\{a > 1, b = 1\}$) characterizes a shift in location and/or an increase in scale of F_1 with respect to F_0. In such case, the $Y_\nu, Y_{\nu+1}, ...$ will belong to outer F_0 center-outward quantiles with higher probability, and hence the $p(Y)$ will assume values close to 1 with higher probability.

To design the corresponding control chart consider then the likelihood ratio test in (2) for the hypothesis in (3) with $p(Y_\nu), p(Y_{\nu+1}), \dots \sim Beta(\theta_1)$, $\theta_1 \in \Theta_1 = \{a > 1, b = 1\}$:

$$\begin{aligned} \tau &= \inf\left\{n|\log\tfrac{f_1(p(Y_n))}{f_0(p(Y_n))} \geq h\right\} = \inf\left\{n|\log\, a[p(Y_n)]^{a-1}) \geq h\right\} = \\ &= \inf\left\{n|\, p(Y_n) \geq \exp\{[h - \log(a)]/(a-1)\} = \tilde{h}\right\}. \end{aligned}$$

The test statistic is then $p(Y_n)$, and therefore a Shewhart chart for detecting changes in location and scale could be designed plotting the $p(y_i)$ values against time. The Upper Control Limit (UCL) in the chart is defined by the cut-off value $\tilde{h}$, and the process is declared out-of-control as long as $p(y_n)$ lie above the UCL. The value $\tilde{h}$, that depends on the $p(Y)$ null distribution, is fixed by the user considering either the amount of false-positive or the ARL_0

desired. In the first case, for a significance level α, $\{\tilde{h} : P(p(Y_n) \geq \tilde{h}) = \alpha\}$. For Lemma 3.1., $\tilde{h} = 1 - \alpha$.

For this chart, the *ARL* $E_a(\tau)$ is a function of the *Beta* parameter a. As τ is the waiting time for the first alarm, it can be described by a Geometric distribution with parameter $\pi_a = P_a(p(Y) \geq \tilde{h})$, with $p(Y) \sim Beta(a, 1)$, $a \geq 1$. Hence:

$$\begin{aligned} E_a(\tau) &= (1/\pi_a) = \left[P_a(p(Y) \geq \tilde{h})\right]^{-1} = \\ &= \left[1 - F_{p(Y)}(\tilde{h})\right]^{-1} = \left[1 - \textstyle\int_0^{\tilde{h}} a x^{a-1} dx\right]^{-1} = \\ &= \left[1 - \tilde{h}^a\right]^{-1} = \left[1 - (1-\alpha)^a\right]^{-1}. \end{aligned}$$

The *ARL* is then a decreasing function of a ($a \geq 1$), and for $a = 1$ we have $ARL_0 = 1/\alpha$. *ARL* functions for different significance levels ($\alpha = 0.05, 0.027, 0.01$) are drawn in Figure 1.

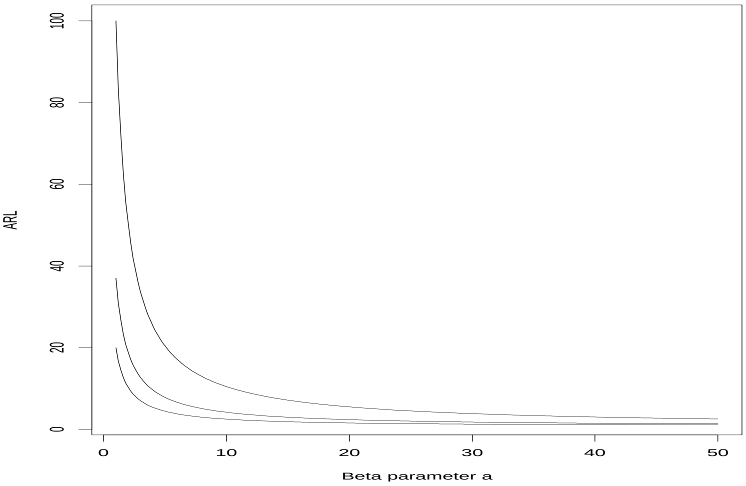


Figure 1: Average Run Length functions in terms of the Beta parameter a for $\alpha = 0.05, 0.027, 0.01$ (from the bottom to the top).

As discussed above, parameter a somehow measures differences in location and/or scale between the F distributions. For better interpretation and use of the chart, it can be expressed as a function of the coverage probability $P(Y \in R(d_p)|F_1)$. In particular, for a given significance level α, we have:

$$a = \frac{\log(P(Y \in R(d_{(1-\alpha)})|F_1))}{\log(P(Y \in R(d_{(1-\alpha)})|F_0))} = \frac{\log(P(Y \in R(d_{(1-\alpha)})|F_1))}{\log(1-\alpha)}.$$

As an example, for $\alpha = 0.01$, if $P(Y \in R(d_{0.99})|F_1) = 0.95$, i.e. F_1 is close to F_0, then $a = 5.1$ and the $ARL = 20$. On the other hand, if F_0 and

F_1 are quite different (say $P(Y \in R(d_{0.99})|F_1) = 0.2$), then $a = 160$ and $ARL = 1.25$.

5 Some final remarks

In this paper we have offered a parametric setting for multivariate nonparametric control charts based on data depth. The proposed approach relies on the *Beta* distribution family. As an additional point, we have found that a shift in location and/or increase in scale can be described by strictly increasing *Beta* densities.

Moreover, following similar arguments, it is possible to design a chart for the detection of a decrease in scale considering strictly decreasing *Beta* densities, where a decrease in scale has to be read as an improvment in the process quality. On the other hand, not strictly increasing or decreasing *Beta* densities correspond to out-of-control cases that do not occur in the SPC practice.

The proposed approach can also be exploited to design more complex stopping rules for Shewhart charts (e.g. m observations outside some warning limits), some CUSUM charts, and to derive corresponding ARL functions.

Finally, we recall that both parametric and nonparametric multivariate control charts generally require a (quite large) preliminary sample to be implemented in practice. Such a sample, previously taken from the process while it is in control, in our case has to be used to estimate depth quantiles and to evaluate the $p(y_i)$ out-of-control measures for the incoming observations. A discussion of which depth function could be chosen for the quantile estimation and related computational issues, although in a different setting, can be found in [11]. In further work, it would be of interest to investigate to what extent the introduced ARL functions depend on such a choice.

References

[1] Antoch J, Jaruskova D. (2002). *On-line statistical process control*. In: Multivariate Total Quality Control, Lauro C, Antoch J, Esposito Vinzi V, Saporta G (eds). Physica-Verlag, Heidelberg, 87− 124.

[2] Barnett V. (1976). *The ordering of multivariate data (with discussion)*. Journal of the Royal Statistical Society A **139**, 318− 354.

[3] Hawkins D.M., Olwell D.H. (1998). Cumulative Sum Charts and Charting for Quality Improvement, Springer-Verlag, New York.

[4] Liu R.Y. (1995). *Control Charts for Multivariate Process*. Journal of the American Statistical Association **90**, 1380−1387.

[5] Liu R.Y., Parelius J.M., Singh K. (1999). *Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference*. The Annals of Statistics **27**, 783−858.

[6] Liu R.Y., Singh K. (1993). *A Quality Index Based on Data Depth and Multivariate Rank Tests*. Journal of the American Statistical Association **88**, 252 – 260.

[7] Mason R.L., Young J.C. (2001). *Multivariate Statistical Process Control with Industrial Application.* ASA-SIAM Series on Statistics and Applied Probability **9**, Philadelphia.

[8] Montgomery, DC. (2001). Introduction to Statistical Quality Control, 4th ed., Wiley, New York.

[9] Mosler, K. (2002). Multivariate Dispersion, Central Regions and Depth, Springer-Verlag, New York.

[10] Porzio G.C., Ragozini G. (2001a). *A Nonparametric Shewhart Chart for Multivariate Process Control.* Cladag 2001 Book of Short Papers, 181 – 184.

[11] Porzio G.C., Ragozini G. (2001b). *Testing through Empirical Center-Outward Quantiles.* In: Modelli Complessi e Metodi Computazionali per la Stima e la Previsione, Provasi C. (ed.) Cleup, Padova, 409 – 414.

[12] Porzio G.C., Ragozini G. (2002). *A Nonparametric Approach to Monitor Multivariate Processes (in italian).* In: Analisi Multivariata per la Qualità totale, Lauro N.C., Scepi G. (eds). Franco Angeli, Milano, 211 – 223.

[13] Romanazzi M., Riani M. (2000). *Bivariate Boxplots and Quality Control.* SIS 2000 Proceedings, Sorrento.

[14] Stoumbos Z.G., Jones L.A., Woodall W.H., Reynolds M.R. Jr (2001). *On Nonparametric Multivariate Control Charts Based on Data Depth.* In: Frontiers in Statistical Quality Control, Lenz H.J., Wilrich P.T. (eds). Physica-Verlag, Heidelberg, 207 – 227.

[15] Tukey J.W. (1975). *Mathematics and the Picturing of Data.* In: Proceedings of International Congress of Mathematicians, James R.D. (ed). Vancouver, **2**, 523 – 531.

[16] Zuo Y., Serfling R. (2000). *General notions of statistical depth function.* Annals of Statistics **28**, 461 – 482.

Acknowledgement: This work was partially supported by M.I.U.R. grant # 2003130242: Multivariate statistical and visualization methods to analyze, to summarize, and to evaluate performance indicators.

Address: G.C. Porzio, G. Ragozini, Dept. of Economics, University of Cassino, Via Mazzaroppi, I-03043 Cassino (FR), Italy; Gino Germani Dept. of Social Sciences, Federico II University of Naples, Vico Monte della Pietà 1, I-80100 Napoli, Italy

E-mail: `porzio@eco.unicas.it, giragoz@unina.it`

COMPSTAT'2004 Symposium

SOME REMARKS TO TESTING OF HETEROSKEDASTICITY IN AR MODELS

Zuzana Prášková

Key words: Heteroskedasticity tests, random coefficient autoregression.

COMPSTAT 2004 section: Time series analysis.

Abstract: In the paper, test procedures for various sources of heteroskedasticity in autoregressive models are discussed. Asymptotic distribution of test statistics are studied and empirical powers are compared in a simulation experiment.

1 Constancy test in RCA(1) under heteroskedasticity

First, let us consider the random autoregression model

$$X_t = b_t X_{t-1} + Y_t, t = 1, \dots, n \tag{1}$$

where b_t is a random parameter such that $\mathrm{E}\, b_t = \beta, \mathrm{Var}\, b_t = \sigma_B^2$ and Y_t are zero mean errors with variances σ_t^2. The sequence $\{X_t\}$ is not stationary in general. Alternatively, we can write $b_t = \beta + B_t$ with $\mathrm{E}\, B_t = 0, \mathrm{Var}\, B_t = \mathrm{E}\, B_t^2 = \sigma_B^2$. Obviously, b_t is constant if and only if $\sigma_B^2 = 0$.

Problem of testing hypothesis $H_0 : \sigma_B^2 = 0$ was solved in literature and various test statistics were introduced (see e.g. Lee [3] or Ha and Lee [1] for references). Lee [3] developed a test which is locally best invariant (LBI) with respect to scale transformation under assumption that Y_t are normally distributed with constant variances. Prášková [5] considered the problem of testing constancy under assumption of heteroskedastic errors and obtained the asymptotic distribution of test statistic without assumption of normality. The test statistic which is a modification of the LBI test statistic by Lee [3] take in this case form (for $n = 2k$)

$$T_n = (\overline{R}_n)^{-2} \left[\sum_{t=1}^{n} (Z_t - \overline{Z}_n)(R_t - \overline{R}_n) + 2\frac{1}{n} \sum_{t=1}^{n} Z_t R_t \right] \tag{2}$$

where $Z_t = X_{t-1}^2/\sigma_t^2,\ R_t = r_t^2/\sigma_t^2$ and $\ r_t = X_t - \widehat{\beta} X_{t-1}$ or $r_t = X_t - \widetilde{\beta} X_{t-1}$, $\widehat{\beta}$ and $\widetilde{\beta}$ denote the least-squares estimator (LSE), respectively the weighted least-squares estimator of β with weights $\frac{1}{\sigma_t^2}$, based on $X_0, X_1, \dots, X_n$, and finally, $\overline{R_n}, \overline{Z_n}$ denote the arithmetic means of R_t, Z_t, respectively.

For $n = 2k + 1$ statistic T_n takes form

$$\begin{aligned} T_n &= (\overline{R}_n)^{-2} \left[\sum_{t=1}^{n} (Z_t - \overline{Z}_n)(R_t - \overline{R}_n) \right. \\ &+ \left. 2\frac{1}{n} \sum_{t=1}^{n} Z_t R_t - \frac{2n}{n-1} \overline{R}_n\ \overline{Z}_n \right]. \end{aligned}$$

Large values of T_n reject H_0.

The result can be formulated as follows.

Theorem 1.1. *Consider model (1) and suppose that the following assumptions hold:*

1. $\mathrm{E}\,X_0 = 0,\ \mathrm{Var}\,X_0 = \sigma_0^2 > 0, \mathrm{E}\,|X_0|^{4+\delta} \le K$ *for a constant* $K > 0$.

2. $\{b_t\}, \{Y_t\}$ *are mutually independent sequences of independent random variables, independent of* X_0.

3. $\mathrm{E}\,b_t = \beta, \mathrm{Var}\,b_t = \sigma_B^2, \mathrm{E}\,|b_t|^4 =$ *const.* $\forall\ t,\ \sup_t \mathrm{E}\,|b_t|^{4+\delta} < 1$.

4. $\mathrm{E}\,Y_t = 0, \mathrm{Var}\,Y_t = \sigma_t^2 \ge d > 0,\ \mathrm{E}\,|Y_t|^{4+\delta} \le K \quad \forall\ t$.

5. $\sigma_t^2 \to \sigma^2, \quad \frac{1}{n}\sum_{t=1}^n \mathrm{E}\,|Y_t|^4 \to m_4 > 0$.

Denote

$$S_n^2 = \frac{1}{n}\sum_{t=1}^n (Z_t - \overline{Z}_n)^2 (R_t - \overline{R}_n)^2. \tag{3}$$

Then under the hypothesis $H_0 : \sigma_B^2 = 0$, *the asymptotic distribution of* $T_n/(\sqrt{n}S_n)$ *is* $\mathcal{N}(0,1)$. *Under* $H_1 : \sigma_B^2 > 0$, *the distribution of* $T_n/(\sqrt{n}S_n)$ *converges to* $+\infty$ *in probability.*

Proof. See Prášková [5], Theorem 2. □

Assumption $\sigma_t^2 \to \sigma^2$ can be replaced by

$$\frac{1}{n}\sum_{t=1}^n \sigma_t^2 \to \sigma^2,\ \frac{1}{n}\sum_{t=1}^n (\mu_t - \bar{\mu}_n)^2 \to 0,$$

where $\mu_t = \mathrm{E}\,X_{t-1}^2$. Moreover, it can be proved, that under H_0 the distribution of T_n is asymptotically the same as that of modified statistic

$$\widetilde{T}_n = \sum_{t=1}^n (Z_t - \overline{Z}_n)(R_t - \overline{R}_n) + 2\frac{1}{n}\sum_{t=1}^n Z_t R_t. \tag{4}$$

Hence, we can obtain the following corollary.

Corollary 1.1. *Under assumptions of Theorem 1.1 and under* H_0, *the asymptotic distribution of statistic* $\widetilde{T}_n/(\sqrt{n}S_n)$ *is* $\mathcal{N}(0,1)$ *and the distribution of* $\widetilde{T}_n^2/(nS_n^2)$ *is asymptotically* $\chi^2(1)$ *as* $n \to \infty$.

2 Connection to general test of heteroskedasticity

Model (1) can be represented by

$$X_t = b_t[b_{t-1}X_{t-2} + Y_{t-1}] + Y_t = \sum_{j=0}^{t} \prod_{i=0}^{j} b_{t-i} Y_{t-j}$$

(for a convenience we denote $Y_0 := X_0$) and from here we get

$$\mathrm{E}\, X_t^2 = \sum_{j=0}^{t} \sigma_{t-j}^2 (\sigma_B^2 + \beta^2)^j \tag{5}$$

which is not constant neither for constant variances σ_t^2.

Alternatively, model (1) can be rewritten in the form

$$X_t = b_t X_{t-1} + Y_t = (\beta + B_t)X_{t-1} + Y_t = \beta X_{t-1} + u_t \tag{6}$$

which is $AR(1)$ model with constant parameter β and martingale differences errors $u_t = B_t X_{t-1} + Y_t$ such that $\mathrm{E}\,(u_t|\mathcal{F}_{t-1}) = 0, \mathrm{Var}\,(u_t|\mathcal{F}_{t-1}) = X_{t-1}^2\sigma_B^2 + \sigma_t^2$, where $\mathcal{F}_0 = \sigma(X_0), \mathcal{F}_t = \sigma(X_0, Y_1, B_1, \dots, Y_t, B_t), t = 1, 2, \dots,$ are σ-fields generated by corresponding random variables. Thus, $\mathrm{Cov}\,(u_t u_s) = 0$ for $t \neq s$ and

$$\mathrm{Var}\, u_t = \sigma_B^2 \mathrm{E}\, X_{t-1}^2 + \sigma_t^2 \tag{7}$$

and it is seen from (5) and (7) that model (1) can be considered heteroskedastic $AR(1)$ model with two sources of heteroskedasticity caused either of randomness of parameter ($\sigma_B^2 > 0$) or varying variances σ_t^2 of error terms Y_t.

To test presence of heteroskedasticity in a regression model

$$Y_i = \beta X_i + \epsilon_i, \quad i = 1, \dots, n,$$

where (X_i, ϵ_i) are independent random vectors such that $\mathrm{E}\,(X_i\epsilon_i) = 0$ for all $i = 1, \dots, n$, White [6] proposed a test of the joint null hypothesis

$$H_0' : X_i \textit{ and } \epsilon_i \textit{ are independent and } \mathrm{E}\,\epsilon_i^2 = \sigma^2, i = 1, \dots, n,$$

and established the asymptotic distribution of test statistic under H_0'.

Application of White's procedure to a general $AR(1)$ model

$$X_t = \beta X_{t-1} + \epsilon_t, \quad t = 1, \dots, n, \tag{8}$$

gives statistic

$$Q_n = \frac{n\left[\frac{1}{n}\sum_{t=1}^{n}(z_t - \bar{z}_n)(r_t - \bar{r}_n)\right]^2}{\frac{1}{n}\sum_{t=1}^{n}(r_t - \bar{r}_n)^2(z_t - \bar{z}_n)^2} \tag{9}$$

where $z_t = X_{t-1}^2, r_t = (X_t - \hat{\beta}X_{t-1})^2$, $\hat{\beta}$ is LSE of β and $\bar{z}_n$ and $\bar{r}_n$ are arithmetic means of z_t and r_t, respectively.

Though the condition of independence of (X_{t-1}, ϵ_t) for $t = 1, \dots, n$ is not satisfied in model (8), for model (1) considered as $AR(1)$ with errors variances (7) we can easily find a connection between $\widetilde{T}_n$ and Q_n.

Theorem 2.1. *Under assumptions of Theorem 1.1 and under joint hypothesis $H_0 : \sigma_B^2 = 0, \sigma_t^2 = \sigma^2$ for all t,*

$$\frac{\widetilde{T}_n^2}{nS_n^2} = Q_n + o_p(1).$$

Proof. A straightforward computation gives that under assumption $\sigma_t^2 = \sigma^2$

$$\frac{\widetilde{T}_n}{\sqrt{nS_n^2}} = T_n^{(1)} + T_n^{(2)}$$

where

$$T_n^{(1)} = \frac{\frac{1}{\sqrt{n}} \sum_{t=1}^n (z_t - \bar{z}_n)(r_t - \bar{r}_n)}{[\frac{1}{n} \sum_{t=1}^n (r_t - \bar{r}_n)^2 (z_t - \bar{z}_n)^2]^{1/2}}$$

$$T_n^{(2)} = \frac{2}{n} \frac{\frac{1}{\sqrt{n}} \sum_{t=1}^n r_t z_t}{[\frac{1}{n} \sum_{t=1}^n (r_t - \bar{r}_n)^2 (z_t - \bar{z}_n)^2]^{1/2}}$$

and under assumption $\sigma_B^2 = 0$ and under assumptions of Theorem 1.1 we get similarly as in the proof of Theorem 2 in Prášková [5], that $T_n^{(1)}$ is asymptotically $\mathcal{N}(0,1)$ while $T_n^{(2)}$ is $o_p(n^{-1/2})$. Since $\left(T_n^{(1)}\right)^2 = Q_n$, the proof is finished. □

We can find the asymptotic distribution of statistic Q_n under the hypothesis $H_0 : \sigma_t^2 = \sigma^2$ even in more general case.

Theorem 2.2. *Consider model (8), where ϵ_t are martingale differences with finite variances $\sigma_t^2 \geq \sigma^2 > 0$, such that*

1. $\mathrm{E}\,|\epsilon_t|^{4+\delta} < C$ *for all t and a constant C,*
2. $\frac{1}{n} \sum_{t=1}^n \mathrm{E}\,(X_{t-1}^2 \epsilon_t^2) \geq c > 0,$
3. $\frac{1}{n} \sum_{t=1}^n \mathrm{E}\,\epsilon_t^4 \to M$ *for a positive constant M, and $|\beta| < 1$.*

Then under hypothesis $H_0 : \sigma_t^2 = \sigma^2$, the asymptotic distribution of statistic Q_n is $\chi^2(1)$.

Proof. The proof follows the same ideas as the proof of Theorem 2 [6] in which asymptotic results for $AR(1)$ process with martingale differences errors by Nicholls and Pagan [4] are utilized. Details are omitted. □

3 Estimation of parameters

Assumption of known variances σ_t^2 can be hardly satisfied in practice. Assuming that $\sigma_t^2 = \sigma^2$ is constant, we can use the standard approach and estimate both σ^2 and σ_B^2 in model (1), respectively in its representation (6) from the regression

$$\mathrm{E}\,(u_t^2|\mathcal{F}_{t-1}) = \sigma^2 + \sigma_B^2 X_{t-1}^2$$

with u_t replaced by their LSE residuals.

The same can be done if we suppose that σ_t^2 take a few different values only, for instance, if they are periodically changing according to scheme

$$\sigma_t^2 = \sigma^{2[i]} \text{ for } t \in I_i = \{i, k+i, \dots, (m-1)k+i\}, i = 1, \dots, k \tag{10}$$

where k is such that $n = mk, m \in \mathbb{N}$. Such assumption can sometimes explain some seasonal pattern of data. For asymptotic considerations we suppose that k is fixed and m is sufficiently large.

Janečková [2] studied estimators of seasonal parameters (10) and obtained their strong consistency and asymptotic normality under assumptions 1-5 of Theorem 1.1, under assumption that Y_t take more then two values and some technical assumption.

Due to simple character of heteroskedasticity in this case, where the data set is split into finite number of groups, inserting these estimations into statistic T_n does not change the asymptotic behaviour of the test statistic. Nevertheless, the correct proof is more complicated and we will published it elsewhere.

4 Simulation experiment

We evaluated the performance of test statistics $\widetilde{T}_n, \widetilde{T}_n^2$, and Q_n, respectively, by simulation. We generated random process (1) with independent errors $Y_t \sim \mathcal{N}(0, \sigma_t^2)$ for various values of σ_t^2 and with iid random parameters b_t having either normal distribution $\mathcal{N}(\beta, \sigma_B^2)$ for various values of β and σ_B^2, or uniform distribution $\mathcal{R}(0,1)$.

In Table 1, number of rejections of statistic $\widetilde{T}_n$ is demonstrated both under $\sigma_B^2 = 0$ and $\sigma_B^2 > 0$. In Figures 1-2, the numbers of rejections of statistics $\widetilde{T}_n, \widetilde{T}_n^2$ and Q_n are compared for various values of σ_t^2 and σ_B^2, the value of β was fixed.

Both statistics $\widetilde{T}_n^2$ and Q_n behave quite similarly in case of homoskedastic errors or in case that the errors exhibit heteroskedasticity but the variances quickly converge to a constant (Figure 1). Statistic $\widetilde{T}_n$ rejects the null hypothesis more often then the remaining two in all cases. In case of too pronounced heteroskedasticity of errors (Figure 2), statistic Q_n rejects as expected the null hypothesis $\sigma_B^2 = 0$ even if it is true (heteroskedasticity is presented), but the empirical level of rejection is poor especially for small and mild sample sizes. Under H_1, the behaviour of Q_n is worse then that of $\widetilde{T}_n$ and $\widetilde{T}_n^2$. It is also seen that all statistics work well for larger sample sizes.

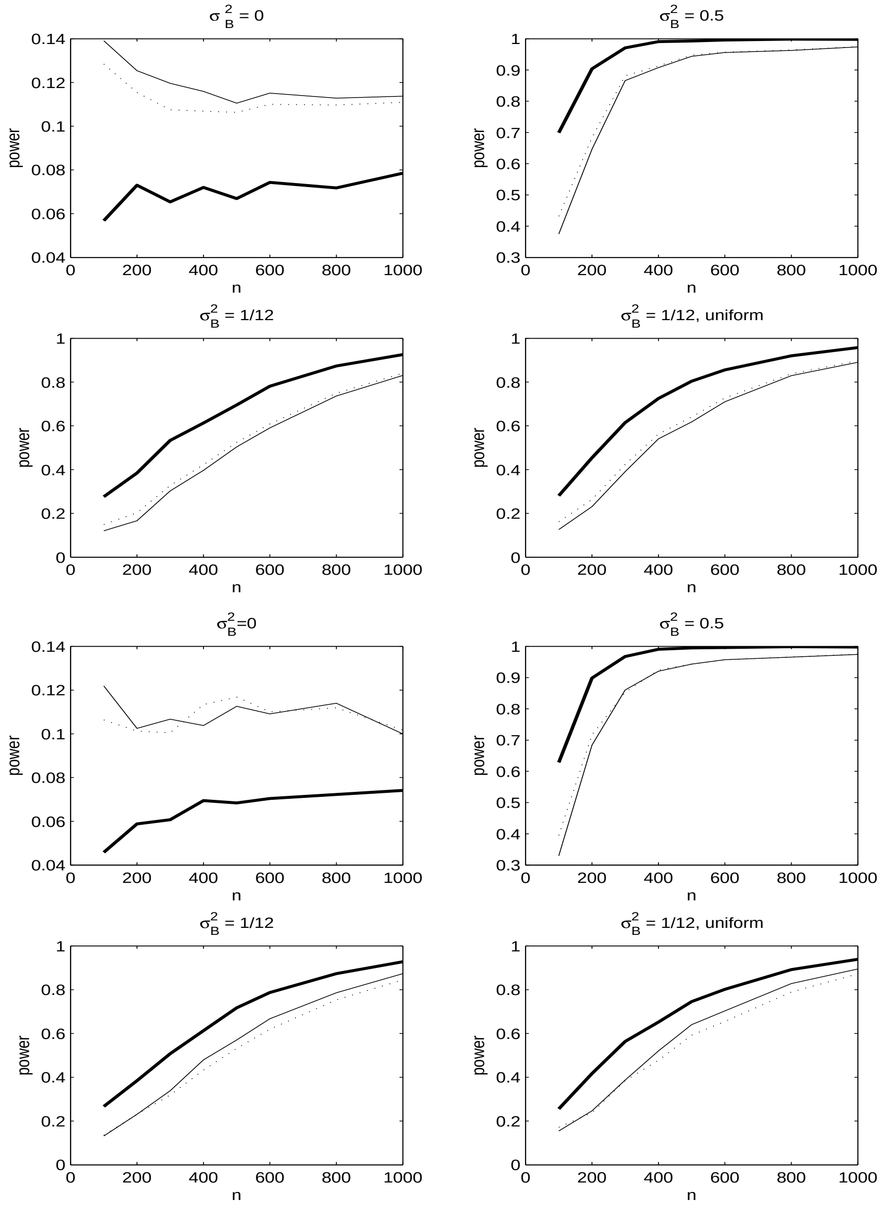


Figure 1: Relative number of rejections of $H_0 : \sigma_B^2 = 0$ against $H_1 : \sigma_B^2 > 0$ in 10 000 simulations for $\widetilde{T}_n$, (thick line), Q_n (thin line) and $\widetilde{T}_n^2$ (dotted line). Nominal level $\alpha = 0.1,\ \beta = 0.5$. Top four panels: $\sigma_t^2 \equiv 1$, bottom four panels: $\sigma_t^2 = 1 + (0.95)^{t/4}$.

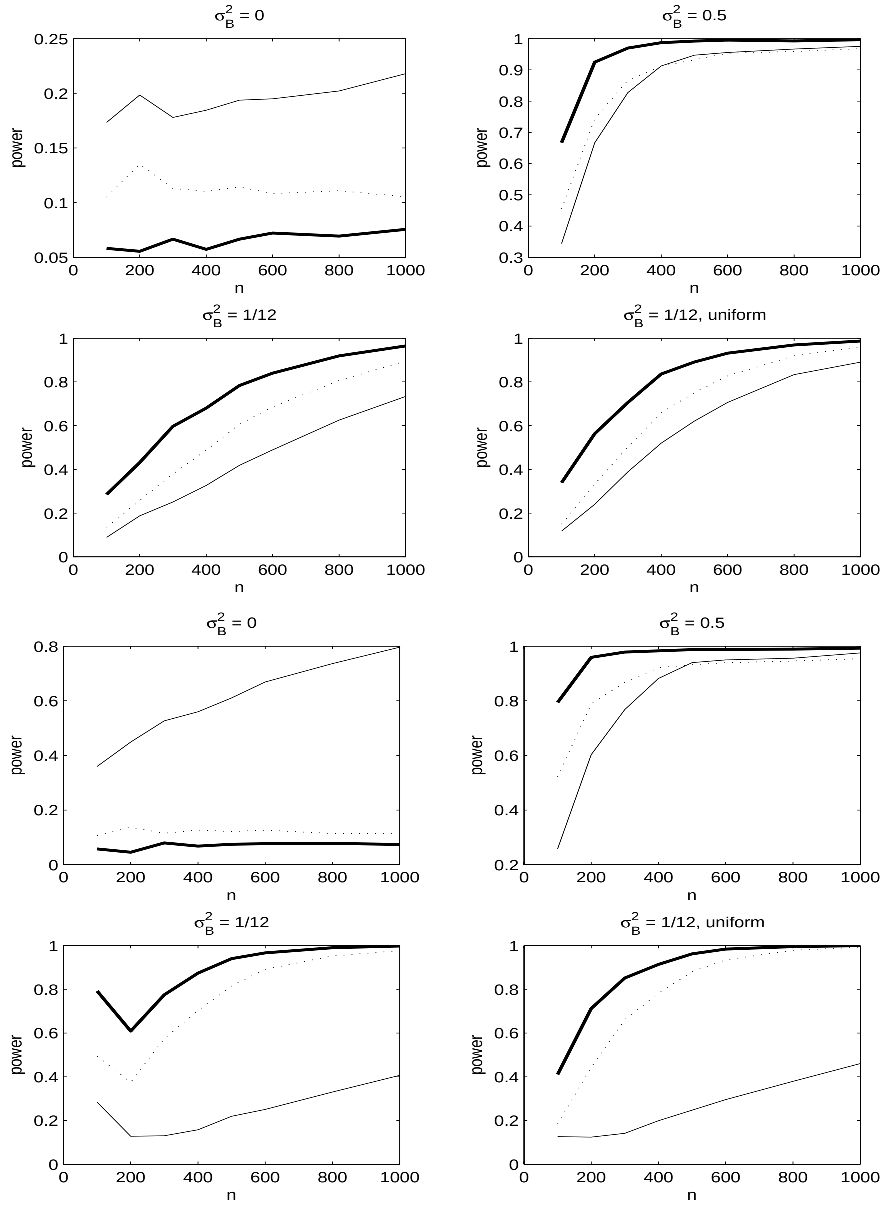


Figure 2: Relative number of rejections of $H_0 : \sigma_B^2 = 0$ against $H_1 : \sigma_B^2 > 0$ in 10 000 simulations for $\widetilde{T}_n$, (thick line), Q_n (thin line) and $\widetilde{T}_n^2$ (dotted line). Nominal level $\alpha = 0.1$, $\beta = 0.5$. Top four panels: $\sigma_t^2 = 1 + \frac{1}{4}(-1)^t$, bottom four panels: $\sigma_t^2 = 1 + \frac{1}{2}(-1)^t$.

$\sigma_B^2 = 0$ (H_0)				$\sigma_B^2 = 0.25$ (H_1)			
n	$\beta = 0.1$	$\beta = 0.5$	$\beta = 0.8$	n	$\beta = 0.0$	$\beta = 0.1$	$\beta = 0.5$
100	0.083	0.046	0.052	100	0.551	0.583	0.628
200	0.075	0.059	0.063	200	0.823	0.806	0.898
300	0.065	0.061	0.048	300	0.940	0.928	0.968
400	0.078	0.070	0.066	400	0.966	0.972	0.991
500	0.075	0.068	0.064	500	0.992	0.987	0.995
600	0.070	0.070	0.063	600	0.998	0.996	0.996
800	0.082	0.072	0.067	800	1.000	0.999	0.999
1000	0.074	0.074	0.068	1000	1.000	1.000	0.998

Table 1: Relative number of rejections of $H_0 : \sigma_B^2 = 0$ in 10 000 repetitions for statistic $\widetilde{T}_n$ and various values of β; $\sigma_t^2 = 1 + (0.95)^{t/4}$, nominal level $\alpha = 0.1$.

Testing heteroskedasticity in autoregression should therefore follow two steps:

1. Test general heteroskedasticity by using statistic Q_n. If the null hypothesis (homoskedasticity and constancy of autoregressive parameter) is rejected, then

2. Detect random character of autoregressive parameter by using statistic T_n, respectively $\widetilde{T}_n$.

References

[1] Ha J., Lee S. (2002). *Coefficient constancy test in AR-ARCH models.* Statist. Probab. Lett. **57**, 65 – 77.

[2] Janečková H. (2002). *Estimation of variances in a heteroscedastic RCA(1) model.* Kybernetika **38**, 405 – 424.

[3] Lee S. (1998). *Coefficient constancy test in a random coefficient autoregressive model.* J. Statist. Plann. Inference **74**, 98 – 101.

[4] Nicholls D. F., Pagan A. R. (1983). *Heteroscedasticity in models with lagged dependent variables.* Econometrica **51**, 1233 – 1242.

[5] Prášková Z. (2003). *Testing constancy in a heteroskedastic RCA model.* Bull. 54 Session ISI, Berlin 2003, CD-ROM.

[6] White H. (1980). *A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity.* Econometrica **48**, 817 – 838.

Acknowledgement: Partially supported by grant GAČR No. 201/03/0945 and by grant MSM 113200008

Address: Z. Prášková, Charles University, Faculty of Mathematics and Physics, Dept. of Probability and Mathematical Statistics, Sokolovská 83,
186 75 Prague, Czech Republic

E-mail: `praskova@karlin.mff.cuni.cz`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

STATISTICAL MODELLING OF LACTATION CURVE DATA

N. Quinn, L. Killen and F. Buckley

Key words: Lactation curves, analysis of residuals, regression.

COMPSTAT 2004 section: Applications.

Abstract: Empirical models of lactation curves using Irish data are examined in this study. 14,956 lactation records from commercial and experimental herds including both autumn and spring calving animals were used for the analysis. A number of models were evaluated on their "goodness-of-fit" and their adherence to the assumptions of regression analysis. Multicollinearity was found to be a severe problem in the application of the model of Ali and Shaeffer [1], but this was eliminated by omitting one of the variables from the estimation procedure. The modified Ali and Scheffer model (referred to as the Ali-B model) still provided a good fit, and met all the regression assumptions. This is a robust model, which is relatively simple to estimate and use, and is accurate in predicting milk yield.

1 Introduction

A lactation curve is a plot of yield of milk from a cow throughout her production period and typically has a shape as in Figure 1. A typical cow gives milk for on average 44 weeks until she is dried off two months before her next calving.

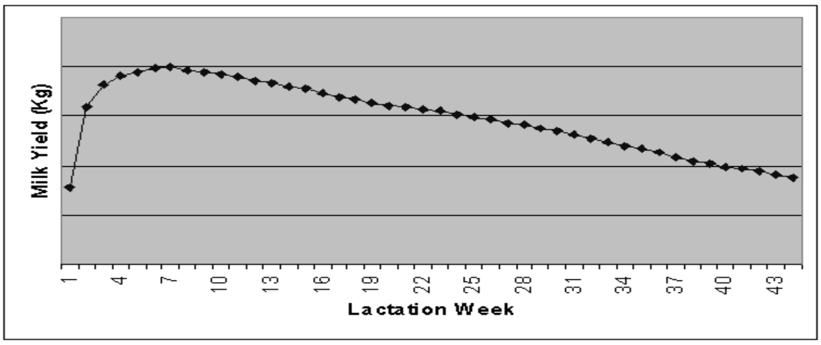


Figure 1: A Typical Lactation Curve.

Well-fitting models of lactation curves have many uses in farm management, such as enabling accurate prediction of total yield from incomplete records. To ensure accurate decisions pertinent to individual animals or herds it is essential that cumulative yield is predicted with minimum error and from relatively few test dates, the latter reducing the cost and inconvenience of milk recording. In addition, good models can improve the accuracy of genetic predictions of sires and dams [15] and is also important for the estimation of breeding values [17]. From the bio-economist's viewpoint, the lactation curve must accurately depict what is expected at farm level.

Many authors have contributed to the advancement in this area of research, but the model of Wood [29] is the basis for studies involving empirical equations [21]. Wood's model: $Y_n = an^b e^{-cn}$, where Y_n is the yield in week n, uses the method of least squares to obtain estimates for three parameters in this incomplete gamma function: a is a scaling factor associated with the average yield, b is related to pre-peak curvature and c is related to post-peak curvature. Many alternative models ([30], [6], [12], [1], [28]) have been proposed as a result of the lack of fit of Wood's model under certain circumstances. Yadav et al. [30] in India and Wilmink [28] in Canada developed models for this reason and in subtropical and tropical climates Kellogg et al. [13] and Shanks et al., [24] found that Wood's model was very poor in fitting their data. Killen and Keane [14] tested Wood's model and examined the shape of lactation curves for milk, fat and protein production using Irish data. Currently at industry level in Ireland, the SLAC (Standard Lactation Curve Method) of Olori and Galesloot [16] is the preferred methodology for predicting milk yield. This method incorporates 2,160 lactation curves, accounting for variation in effect of season of calving, calving age and level of production. While it is acknowledged that having a library of equations from which the most appropriate one is chosen will almost inevitably give accurate predictions, a single equation model has many advantages. It is simple to incorporate a single equation model into computer programs which can be easily updated and re-examined in the light of new data. It would also be considered more appropriate for use by bio-economists because they need to update and re-create the model parameters for many circumstances.

As management technology, cow production potential, and procedures to evaluate lactation curve models have advanced enormously in Ireland over the last 25 years, it has resulted in a renewed interest in re-examining lactation curve models under Irish conditions. The more recent alternatives to Wood's model, such as those proposed by Wilmink [28], Ali and Schaeffer [1], and Guo and Swalve [10] may be worthy of investigation because they have proven to have a better fit than Wood's model in their respective studies. The objective of this study is to compare the goodness-of-fit of a number of empirical models under present day Irish production circumstances and to analyse the residuals. The aim is to arrive at a well-fitting, robust, single equation model for lactation curve data. As well as fitting various models using regression analysis this study describes an analysis of the residuals, which does not appear to have been carried out in previous studies of modelling lactation curves.

2 Data

The dataset consists of 14,954 lactations, after editing, of which 13,227 were monthly records and 1,727 were daily records. The test day yields were collected, by the DairyMis system in Teagasc [7], over a period from 1995 to 2001. It consisted of both spring and autumn calving cows of various breeds, from 85 herds. The data which was collected included year of production, parity (lactation number), calving month, lactation week, milk yield, fat yield and protein yield.

3 Models and statistical analysis

Since early in the 20^{th} century models of the shape of the lactation curve have been proposed by authors such as Brody, Ragsdale and Turner [3], Brody, Turner and Ragsdale [4], Sikka [25], Dave [9], Wood [29], Wilmink [28], Ali and Schaeffer [1] and Guo and Swalve [10]. An inital examination showed that some of these models were inadequate in describing the Irish data (i.e. $R^2 < 0.70$). However, there is a danger in simply comparing the R^2 of the inverse polynomial and a logarithmic transformation of the incomplete gamma, as measures of the goodness of fit of both models causes uncertainty about the conclusions drawn [27]. The R^2 of a nonlinear equation is calculated as: $1-\frac{SSE}{CSS}$ where SSE is the error sum of squares and CSS is the corrected total sum of squares for the dependent variable. This formula is based on linear models and does not hold for non-linear models; it is not correct statistically to compare the R^2 of functions when using various PROC procedures in SAS as when finding the R^2 of a non-linear function the R^2 is “no longer bounded by zero and one” [23].

The Mean Square Prediction Error (MSPE) value was used as a measure of goodness-of-fit [11] for making comparisons between models. The models of Brody et al. [3], Brody et al. [4], Sikka [25], Dave [9], Wood [29], Wilmink [28], Ali and Schaeffer [1] and Guo and Swalve [10] were fitted to our data using the PROC REG procedure for estimating linear regression models and the PROC NLIN procedure for nonlinear regression models, in SAS version 8.2e. The effect of parity, calving month and herd were removed as they are known to have a significant effect on milk yield in Ireland [8].

Brody et al.'s [4] equation failed to converge,using the Irish data, and other models such as those of Brody et al. [3], Sikka [25] and Dave [9] were also eliminated because of high MSPE values (greater than 370). An initial examination of the results showed that the models of Wood (nonlinear form), Singh, Wilmink, Ali and Schaeffer and Guo and Swalve, had acceptable MSPE values. Wood's linear form was also included for comparison purposes as it is considered as the basic reference among the empirical equations [21]. The MSPE values found in this study for the linearised and non-linear versions of Wood's model reinforce the point made by Cobby and Le Du [6], that non-linear regression would prove to be a better method of fitting the model to the data. According to Cobby and Le Du [6] weighting improves the variance distribution of the residuals also, especially at the beginning of the lactation. Wood's linear model was examined initially without statistical weights and later by weighting the logarithm of the milk yields (i.e. the dependent variable) proportionally to the square of the milk yield. When examining Wilmink's model, the fourth parameter, k, is held constant, thus reducing the number of parameters to be estimated from four to three and greatly simplifying the fitting of the curve. Brotherstone, White and Meyer [5] showed k to be constant, at 0.1, over lactations and over age groups within parity for UK data. Olori et al. [17] estimated k, following a preliminary analysis, to be 0.61 for their data set, where k was estimated as being the best fitting value for the herd mean data. This study carried out a similar analysis on the fourth parameter by performing regression analysis on Wilmink's model, allowing the four parameters to be variable. An analysis of variance was then carried out on parameter k and it was found that it was constant over parity, calving month and herd effect for the Irish data, with a value of 0.10.

The appropriateness of the models is also tested by examining the assumptions of regression analysis which are as follows:

- Independent errors
- Equal error variance σ^2
- Indepenent explanatory variables
- Errors ϵ_i are distributed $N(0, \sigma^2)$

In any practical problem, the assumptions required for fitting models using regression analysis could be in doubt [19] and a second phase of analysis was carried out to test the validity of these assumptions by analysis of the residuals.

4 Results

4.1 Goodness-of-fit

The MSPE values found when fitting weekly milk yield using a number of existing models are given in Table 1. It can be seen that Ali and Schaeffer's model gave the best fit with a MSPE of 135.94, while Wood's model (linear form) was the least satisfactory in this regard (MSPE of 419.42).

4.2 Analysis of residuals

Wood's models (both linear and weighted linear forms) were therefore omitted from further consideration because of their high MSPE values. The Durbin-Watson statistic, d, was found to be between d_u and $4 - d_u$, for first order autocorrelation, for the models of Wood (nonlinear form), Wilmink, and Guo

Model	**MSPE**
Wood (Linear Form)	419.42
Wood (Weighted Linear Form)	384.06
Wood (Nonlinear Form)	315.23
Wilmink	324.430
Ali & Schaeffer	135.94
Guo & Swalve	252.98

Table 1: Goodness-of-fit statistics of expected curves for milk yield.

and Swalve indicating that autocorrelation was not present in these models. For Ali and Schaeffer's model the test for autocorrelation was inconclusive for first order but from an examination of second order autocorrelation it was concluded that autocorrelation was not significant in this model either.

The multicollinearity diagnostics were examined and they revealed that in Ali and Schaeffer's model multicollinearity was severe with a condition index of 3778 (Table 2). In Guo and Swalve's model and Wood's (nonlinear form) model it was moderate to strong (condition index values of 81.65 and 41.25 respectively), whereas in Wilmink's model, multicollinearity was weak (condition index value of 19.60). Ali and Schaeffer's model was then re-examined with each of the variables removed in turn thus omitting one of the parameters from the model. It was found that the condition index could be decreased from 3778 to 100.71, the greatest improvement occurring when parameter b was removed; this modified form of Ali and Schaeffer's model will be referred to as the Ali-B model. The MSPE value for the Ali-B model was 185.23, which is higher than for Ali and Schaeffer's model but the problem of multicollinearity among the independent variables inflating the standard errors has been reduced.

Heteroskedasticity was not a problem in any of the models as all models had a p-value for White's test of greater than 0.05. The Kolmogorov-Smirnov test statistic *(D)*, varied from 0.15 to 0.17 across the models indicating that the normality assumption was not violated, though it is accepted that a test's ability to reject the null hypothesis increases with the sample size [22]. As kurtosis varied between 1.35 and 0.85, and skewness varied between -0.28 and -0.08, it was concluded that there was no significant deviation from normality in the distribution of the residuals. When all of the assumptions are taken into account it appears that the Ali-B model is in fact the most robust model as it adheres to assumptions and it also provides a reasonable fit to the data. When estimating the parameter values of Ali and Schaeffer's model, certain records were removed because the upward slope was not being recognised. The records removed were those in Data Set 1 which had their first recording taken after week seven of lactation. These did not appear to affect the other models to any great extent which suggests that it is necessary to have a reading within the early weeks of a lactation to accurately estimate the upward slope for Ali and Schaeffer's model.

Test	**Guo & Swalve**	**Ali & Schaeffer**[1]	**Ali-B**[1]
Functional Form	$Y_n = a + b\sqrt{n} + c\ log(n)$	$Y_n = a + b\gamma + c\gamma^2 + d\omega + e\omega^2$	$Y_n = a + c\gamma^2 + d\omega + e\omega^2$
MSPE	252.98	135.94	185.23
Durbin-Watson	No 1^{st} Order	No 2^{nd} Order	No 1^{st} Order
Condition Index	81.65	3778	100.70
White's Test	0.21	0.33	0.31
Kolmogorov-Smirnov	0.16	0.16	0.15

Table 2: Diagnostics for three models.

5 Conclusions

The goal of this study was to examine a number of single equation models to explain the production of milk yield throughout a lactation. The results show that the modified Ali and Schaeffer model (Ali-B) is the most reliable model at predicting individual weekly cow milk yield.

The model's ability to the adhere the assumptions, which are made when fitting the models using regression analysis, were examined. The only assumption that was not satisfied was that of the explanatory variables being independent in every case (multicollinearity). Multicollinearity was extremely strong for Ali and Schaeffer's model but when parameter b was removed it was found that the resulting model (Ali-B) was the most satisfactory in that it satisfied all assumptions tested and multicollinearity was no longer a major issue. While there was still some correlation between the explanatory variables this is inevitable in non-linear models [19]. The Ali-B model also had a relatively good MSPE value, and it was concluded necessary to sacrifice some goodness-of-fit for a model's ability to satisfy the assumptions.

It has therefore been possible to arrive at a single, well-fitting, and robust model to represent the shape of lactation curves for Irish dairy cows. Before using this model to predict the milk yield for a specific cow the seasonal effects have to be added to the Ali-B model. These effects vary from region to region accounting for variation in climate, soil quality, environment etc. In conclusion the Ali-B model is the best fitting model to Irish data and it can be easily updated for different regional effects.

References

[1] Ali T.E., Schaeffer L.R. (1987). *Accounting for covariance among test day milk yields in dairy cows.* Canadian Journal of Animal Science **67**, 637 – 644.

[2] Belsley D., Kuh E., Welsch E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity.* John Wiley & Sons.

[3] Brody S., Ragsdale A.C., Turner C.W. (1923). *The rate of decline of milk secretion with the advance of the period of lactation.* Journal of General Physiology, **5**, 441 – 444.

[4] Brody S., Turner C.W., Ragsdale A.C. (1924). *The relation between the initial rise and the subsequent decline of milk secretion following parturition.* Journal of General Physiology**6**, 541 – 545.

[5] Brotherstone S., White I.M.S., Meyer K. (2000). *Genetic modelling of daily milk yield using orthogonal polynomials and parametric curves.* Animal Science **70**, 407 – 415.

[6] Cobby J.M., Le Du Y.L.P. (1978). *On fitting curves to lactation data.* Animal Production **70**, 407 – 415.

[7] Crosse S. (1986). *"The Development and implementation of a computerized management system for Irish Dairy farmers"*. PhD Thesis. National University of Ireland, Cork.

[8] Cunningham E.P. (1972). *Components of variation on dairy cow production.* Irish Journal of Agricultural Research **11**, 1 – 10.

[9] Dave B.K. (1971). *First lactation curve of the Indian water buffalo.* JNKVV Research Journal **5**, 93 – 98.

[10] Guo Z., Swalve H.H. (1995). *Modelling of lactation curve as a sub-model in the evaluation of test day records.* Paper presented at the Interbull Meeting, 7-8th September, 1995, Prague, Czech Republic.

[11] Kvanli AH., Guynes C.S., Pavur R.J. (1986) *Introduction to business statistics.* Fourth Edition. West Publishing Company.

[12] Keown J.F., Van Vleck L.D. (1972). *Extending lactation records in progress to 305-day equivalent.* Journal of Dairy Science **56**, 1070 – 1079.

[13] Kellogg D.W., Urquhart N.S., Ortega A.J. (1972). *Estimating Holstein lactation curves with a gamma curve.* Journal of Dairy Science **60**, 1308 – 1315.

[14] Killen L., Keane M. (1978). *The shape of lactation curves in Irish dairy herds.* Irish Journal of Agricultural Research **17**, 267 – 282.

[15] Koonawootrittriron S., Elzo M.A., Tumwasorn S., Sintala W. (2001). *Lactation curves and prediction of daily and accumulated milk yields in a multibreed dairy herd in Thailand using all daily records.* Thai Journal of Agricultural Science **34**, 123 – 139.

[16] Olori V.E., Galesloot P.J.B. (1999). *Projection of partial lactation records and calculation of 305-day yields in the Republic of Ireland.* Proceedings of the 1999 Interbull Meeting in Zurich, Switzerland, August 26-27, Pg 149 – 154.

[17] Olori V.E., Brotherstone S., Hill W.G., McGuirk B.J. (1999). *Fit of standard models of the lactation curve to weekly records of milk production of cows in a single herd.* Livestock Production Science **58**, 55 – 63.

[18] Madalena F.E., Martinez M.L., Freitas A.F. (1979) *Lactation curves of Holstein-Friesian and Holstein-Friesian x Gir cows.* Animal Production **29**, 101 – 107.

[19] Maddala G.S. (1992). *Introduction to econometrics.* 2nd Edition. Macmillan Publishing Company, New York.

[20] Molina J.R., Boschini C. (1979). *Adjustment of dairy curve of Holstein herd with a linear model.* Agronomia Costarricense **3**, 167 – 174.

[21] Pérochon L., Coulon J.B., Lescourret F. (1996) *Modelling lactation curves of dairy cows with emphasis on individual variability.* Animal Science **63**, 189 – 200.

[22] SAS®Online Doc, Version 8, 1999. SAS Inst., Inc. Cary NC.

[23] SAS®FAQ#983, http://support.sas.com/faq/009/FAQ00983.html site visited on 20th July 2003.

[24] Shanks R.D., Berger P.J., Freeman A.E., Dickinson F.N. (1981). *Genetoc aspects of lactation curves.* Journal of Dairy Science **64**, 1852 – 1860.

[25] Sikka L.C. (1950). *A study of lactation as affected by heredity and environment.* Journal of Dairy Research **17**, 231 – 252.

[26] Singh R.P., Gopal R. (1982). *Lactation curve analysis of buffaloes maintained under village conditions.* Indian Journal of Animal Sciences **52**, 1157 – 1160.

[27] Tozer P.R., Huffaker R.G. (1999). *Mathematical equations to describe lactation curves for Holstein-Friesian cows in New South Wales.* Australian Journal of Agricultural Research **50**, 431 – 440.

[28] Wilmink J.N.M. (1987). *Adjustment of test day milk, fat and protein yield for age, season and stage of lactation.* Livestock Production Science **16**, 335 – 348.

[29] Wood P.D.P. (1967). *Factors affecting the shape of the lactation curve in cattle.* Nature, London **216**, 164 – 165.

[30] Yadav M.P., Katpatal B.G., Kaushik S.N. (1977). *Components of inverse polynomial function of lactation curve, and factors affecting them in Hariana and its Friesian crosses.* Indian Journal of Agricultural Science **47**, 777 – 781.

Acknowledgement: The authors would like to acknowledge the technical assistance of Mr. David Cliffe, Teagasc, Moorepark and Mr. Gary Keogh, from DCU, who gave advice on the analysis of residuals.

Address: N. Quinn, L. Killen F. Buckley, School of Computing, Dublin City University, Dublin 9, Ireland

E-mail: `nquinn@computing.dcu.ie`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

THE ITALIAN JUDICIAL STATISTICAL INFORMATION SYSTEM

M. Renzetti, G. Sindoni, L. Tininini and A. Urbano

Key words: Statistical information systems, data aggregation, data integration.

COMPSTAT 2004 section: Statistical information systems.

Abstract: This paper presents an integrated, Web-based system of statistical indicators measuring quantity and quality of judicial demand and supply by different Italian territorial units. The user can browse tables without predefined access paths, independently choosing survey, item, time period and level of territorial detail (administrative or judicial) and performing drill-down and roll-up operations. The navigation system allows complete separation of data and applications and complete control of the confidentiality and significance of published data. The system also provides two software tools which allow researchers to set up and organize statistical information for publication on the Internet and manage it throughout its life cycle.

1 Introduction and project description

One of the tasks of national statistical offices is to disseminate data in an effective and timely way. As Internet use spreads, Italy's National Statistical Institute (ISTAT) disseminates ever more statistical information on line. The Italian judicial system has been heavily reformed by a set of laws modifying its organization and procedures (supply side) to improve its efficiency in spite of an increasing number of cases and appeals (demand side). As a consequence, it was necessary to create an integrated system of statistical indicators measuring quantity and quality of justice demand and supply by different Italian territorial units.

In accordance with these principles, the *"Justice Territorial Information System (SITG)"* was created[1] to enhance the spread of statistical data on the justice system and improve the level of territorial and thematic detail, as well as flexibility of access to this kind of information, while respecting the law regulating protection of privacy (Law n. 675/96).

Upkeep and updating of the system are key issues, as they are crucial to its survival and informative efficacy. An autonomous system was thus also created (*SITG-Manager*) to manage general administration and updating of the judicial system's various statistical components.

SITG is based on the careful conceptual planning and statistical checking of all elements. The statistical topics were chosen to satisfy the needs and

[1] The information system was designed and developed by a team of ISTAT researchers, supported by external IT staff

requests made by the various users of judicial statistics. The system manages 25 different surveys and generates absolute and derived indicators from surveys conducted by ISTAT or Ministries and other Public Institutions.

According to defined conceptual design, the system will disseminate indicators and metadata on the following areas and issues: **Civil Justice** (separations, divorces, minors, labour suits, protests, bankruptcies, Judicial office activities), **Administrative Justice** (Regional Administrative Courts, Council of State, etc.), **Notary Activity**, **Penal Justice** (Convicts, suicides and attempted suicides; crimes reported to the police, criminality, juvenile criminality, Judicial office activities), **Prison world**.

The indicators give information on different aspects of Italy's social-judicial situation and can be used for evaluation and management by judicial system administrators. The definition and choice of indicators were based on the researchers' experience, external requests and a survey carried out by statistical information centres located in ISTAT's regional offices throughout the country. The following indicators are produced for every subject down to the minimum territorial detail (administrative and judicial) allowed by data availability: *social-demographic indicators* (describing the characteristics of judicial events and involved persons); *management indicators* (describing the operating of judicial offices). The planned indicators therefore cross all the contemplated territorial levels and are updated annually. Below are examples of the main conceptual units and related indicators in the judicial Web site:

- Married couple instability: number of separations and divorces (absolute and rates per 100,000 married people), custody of minors awarded, details on marriage and husband/wife, maintenance allowances
- Business insolvency: number of declared and closed bankruptcies, crisis indices, enterprise structural features and financial loss
- Efficiency of judicial system: duration figures, rates of extinction and turn over of proceedings
- Criminality risk: number and quotients of crimes, kinds of reported crimes, principal characteristics (sex, age, citizenship) of perpetrators
- Prison situation: new arrivals and inmates present at year end by some social-demographic attributes, overcrowding indices.

Indicators are put together in thematic tables describing homogeneous topics and combined to optimise their information efficacy and spreading. Table flank is normally (but not necessarily) territorial level, while the table header changes according to subject and argument.

2 The Web based navigation system

SITG is made up of the following sections, each with an inner division: **Statistical Data Browsing.** It is divided into "Predefined Table Consultation" and "Dynamic Table Consultation" (available from next year). The first includes: 1) *Data from 2000* - The user can browse tables without predefined access paths, independently choosing survey, item, time period and level of territorial detail (administrative or judicial); 2) *Data before 2000* - Contains a set of summary tables that can be visualized and saved in spreadsheet format on a personal computer. **Download Area.** Here statistical tables can be saved for each subject in spreadsheet format on a personal computer. **International Comparisons.** This section is intended to be a landmark in comparison of judicial statistics, as justice must also be seen in an European dimension. There are three main areas: 1) *Countries*: outlines the organization and functioning of judicial systems in some European countries (currently France, Germany, Spain, England and Wales), whilst the item dedicated to "Other Countries" offers only links to national Web sites with judicial statistics and information; 2) *Statistical Comparisons*: presents comparable tables for some judicial statistics referring to the five European countries mentioned above. A methodological note explains utilized methods and problems connected to statistical comparisons in this area; 3) *International Institutions*: shows in detail activities and projects carried out by the Council of Europe, European Union and United Nations, as well as a set of links to other Web sites dealing with judicial issues. **Documentation.** This contains a glossary of the most important statistical terms used in the Web site tables, informative cards about surveys, main publications dealing with justice statistics, laws regulating the different phenomena surveyed, instructions.

A user guide (explaining the browsing mechanism) and a user comments form are also available. Finally, other Web site options are textual research and analytical index of indicators (with link to the corresponding table).

Dynamic tables may be consulted through a *generalised* spatio-temporal Web warehouse[4] system known as *DaWinci*, based on a previous prototype[5]. It was originally conceived for publication of Population and Housing Census data[2] and, due to the requirement for incremental evolutions to suit new application contexts, it was designed according to the following principles:

- **Complete separation between data and applications.** Not only aggregate data, but also information describing their meaning and guiding user navigation, are stored in the system database. This allows code rewriting to be minimised when changes to the application behaviour are needed.
- **Usability of the navigation interface.** The graphic user interface has been designed to be as self-explanatory as possible and to minimize the number of mouse clicks needed to access a new table[1].

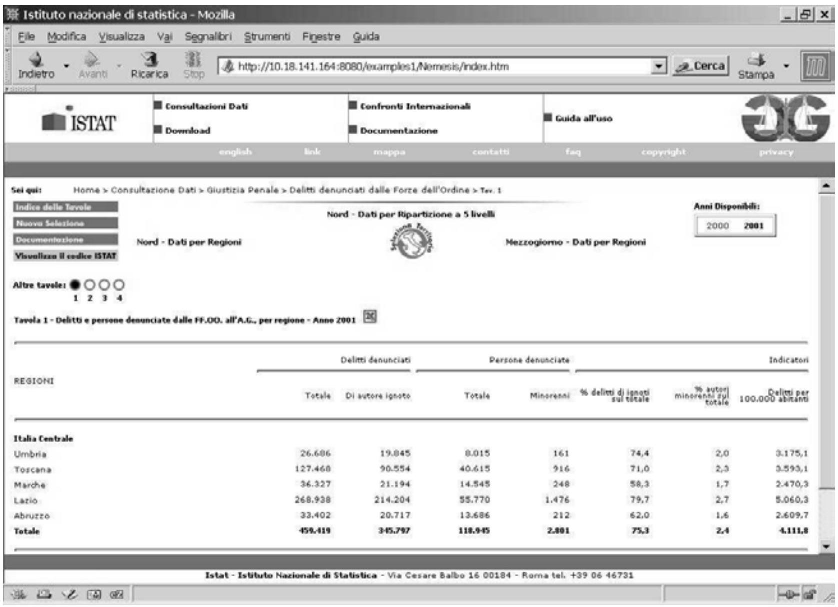
[2] Available on the Web at the address `http://dawinci.istat.it/pl/index_eng.html`.

- **Complete control on the confidentiality and significance of published data.** The dynamic table building process does not perform any calculation: all displayed values are pre-aggregated by the data aggregation tools described below. In addition to optimising system performance, this allows complete control of the quality features of published data in terms of secondary disclosure control and significance of aggregates from sample data.

The system's basic navigation and display units are its statistical tables, which are dynamically built by extracting data from a relational database and are displayed as HTML files. The system has a three layer architecture which will be described in section 3.

2.1 Navigation functions and system architecture

The published statistical tables are classified into *Categories* (for example "Civil Justice", "Administrative Justice", etc.), hierarchically organised so that users can progressively restrict the set of available tables by choosing to display only the list of tables for the category of interest. The same table can be available for more than one year and users can choose the year of interest. Figure 1 presents the table display page.



http://10.18.141.164:8080/examples1/Nemesis/index.htm

Sei qui: Home > Consultazione Dati > Giustizia Penale > Delitti denunciati dalle Forze dell'Ordine > Tav. 1

Tavola 1 - Delitti e persone denunciate dalle FF.OO. all'A.G., per regione - Anno 2001

REGIONI	Delitti denunciati		Persone denunciate		Indicatori		
	Totale	Di autore ignoto	Totale	Minorenni	% delitti di ignoti sul totale	% autori minorenni sul totale	Delitti per 100.000 abitanti
Italia Centrale							
Umbria	26.686	19.845	8.015	161	74,4	2,0	3.175,1
Toscana	127.468	90.554	40.615	916	71,0	2,3	3.593,1
Marche	36.327	21.194	14.545	248	58,3	1,7	2.470,3
Lazio	268.938	214.204	55.770	1.476	79,7	2,7	5.060,3
Abruzzo	33.402	20.717	13.686	212	62.0	1.6	2.609.7
Totale	**459.419**	**345.797**	**118.945**	**2.801**	**75,3**	**2,4**	**4.111,8**

Istat - Istituto Nazionale di Statistica - Via Cesare Balbo 16 00184 - Roma tel. +39 06 46731

Figure 1: The table display and navigation page.

Each table is available at different territorial levels and areas and the following navigation functions are available as Web links from the table display page: *drill down*, i.e. visualising the same table at a more detailed territorial level; *roll up*, i.e. visualising the same table at a less detailed territorial level; *horizontal navigation*, i.e. visualising the same table for a neighbouring area; *territory change*, i.e. visualising the same table at a completely different territorial level and/or area. The system is able to manage more than one territorial hierarchy. For example, the Italian Judicial information system publishes tables referring to both the administrative and Justice hierarchy (which are structurally different to each other).

Each table, or all tables of the same category, can be downloaded in spreadsheet format. Each spreadsheet is dynamically built by means of a specific module, based on Open Source technology[3].

2.2 The data model

The navigation system is based on a generalised data model, specifically tailored for statistical table storage, management and extraction. The model is outlined in figure 2, where mono-directional arrows represent one-to-many relationships between entities while bidirectional ones represent many-to-many relationships.

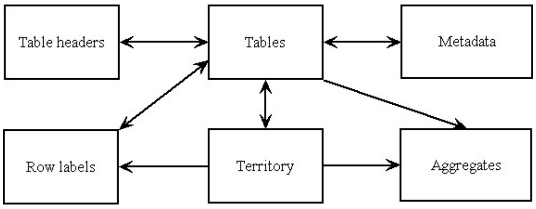


Figure 2: Outline of the data model.

A statistical table is identified in the database by a code and can span across more than one *year* and *territorial level*. Each table contains *aggregates*, whose values are explicitly stored with information on their position in the table display area. Table *headers* are stored as the set of their component labels (statistical objects, classifications and modalities). The way in which the labels must be displayed is also stored using a suitable structure. Table *flanks* are sequences of *row labels* which are also stored using a suitable data structure. The flank of a table is normally (but not necessarily) composed of names of *territorial entities* (municipalities, provinces, etc.), as the system is strongly focused on the data's territorial aspects. The territorial hierarchies are stored as temporal objects, that is, each territorial object is stored together with its reference year and information on its parent in the hierarchy. It is thus possible to keep track of hierarchy updates, such as creation and deletion of objects and changes in the hierarchical relationships[2].

Documentation metadata are also stored in the database and are structured to allow for the definition, modification and deletion of new *types* and *instances.* Explicit links to files of many formats (spreadsheet, PDF, etc.) are also stored and these can be dynamically displayed in metadata pages.

3 Software tools for data loading and aggregation and statistical table management

Although the main features *SITG* provides are related to Web navigation, the system also provides two software tools, *SITG-Load* and *SITG-Manager*, which allow researchers to set up and organize statistical information for publication on the Internet and to manage it throughout its lifecycle.

This complex goal can be achieved following the stream of a main process that starts with steps through which survey microdata are loaded into a database server and processed and aggregated to obtain macrodata for use in statistical table production. At the same time, a process of statistical table definition allows researchers to specify information to be included for each table, rules to be applied for calculation and layouts to be used for presentation. Based on macrodata availability and statistical table definitions, it is possible to produce and publish the whole set of predefined data and indicators.

Described below are the tasks making up the main process and the way in which *SITG-Load* and *SITG-Manager* can help to accomplish them:

SITG-Load (microdata and macrodata management)

1. *Microdata loading.* For each survey, validated microdata, typically available as flat files, are loaded into the database server. Researchers can load microdata for each survey and a specified year by means of a simple user interface which allows them to select a flat file on the hard disk of their own desktop and run a microdata load procedure.

2. *Macrodata calculation.* After some automatic checks which verify that the loading process has been completed correctly, users can initiate a set of procedures, for a group of surveys and a selected year, which processes and aggregates microdata to produce macrodata for storage in the database server and use in statistical tables production.

SITG-Manager (statistical content management)

1. *Statistical variables definition.* Personnel with in depth knowledge of the database schema define a number of statistical variables and map them with database tables columns. As a result, statisticians can handle statistical data simply and straightforwardly, without any concerns about the database's structure.

2. *Statistical indicators definition.* Starting from statistical variables, researchers can enter a definition for a number of indicators. Some information must be provided for each indicator: identifier, type (absolute

or derived), numeric format, calculation formula, territorial hierarchy (administrative, judicial or both), first and last year of validity.

3. *Statistical tables definition.* From a set of indicators, researchers can define any new statistical table. Some information must be given for each table: identifier, title, argument, ordered list of indicators to be used and related heading titles, notes.

4. *Statistical tables production.* This step allows researchers to produce one or more statistical tables, simply choosing a year to be used for the calculation. At the end of the process, statistical tables are automatically loaded into an Intranet database, to allow researchers to use the same navigation system available to the Internet community and verify that the new tables are correct.

5. *Statistical tables publication.* All validated statistical tables can be easily published on the Web with a simple click. Table data are copied into an exposed database server for browsing by Internet users.

In addition to the main process of statistical table management, *SITG-Manager* implements a metadata loading and publishing process. This can be considered as a simplified version of the former process, through which a researcher can load a set of documents or information related to one or more specific surveys and make them available to Internet users. Of course, all the above steps run in parallel with a statistical tables conceptual design task,[3] which is active throughout the lifecycle of the main process.

These software tools have been designed to allow researchers to handle all tasks belonging to the main process effectively and with a negligible impact on their day-by-day activities. Through a number of functions for statistical content management, they also allow easy modification as required of the structure and organisation of tables and documents published on the Web.

From a technological point of view, *SITG-Load* and *SITG-Manager* use a traditional client/server software architecture. The first was developed in a MS Visual Basic environment, whilst the second is an Oracle Forms 6i application. Both tools run a set of PL/SQL stored procedures on the server side to perform some back-end operations. However, the navigation system uses a 3-tier, Web based software architecture and provides a predefined set of statistical tables which Internet users can access through simple thin-client PCs with a common Web browser installed.

There are two distinct database servers: an Intranet server used by researchers to safely test newly produced statistical tables and an Internet host where tables and metadata available to the Internet community are loaded.

[3] This is the real core of statistical analysis and deals with conceiving, defining, calculating and validating statistical indicators and tables.

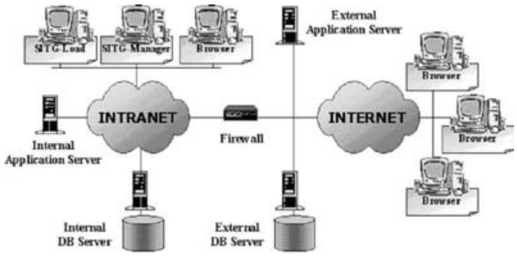


Figure 3: Outline of the system architecture.

Each database server is an IBM AIX host with an Oracle 8i instance, for safe, reliable and high available data management, and a set of PL/SQL stored procedures, mainly used to produce macrodata from microdata and statistical tables from macrodata. The application server is an IBM AIX host with a Apache/TomCat Web server installation and a set of JSP's (Java Server Pages), to implement the navigation system.

4 Future developments

A new release will be realized next year. It will offer dynamic data access (so Internet users can create personalized tables by choosing their desired variables), interactive cartography and time series of statistical data.

References

[1] Atzeni P., Merialdo P., Sindoni G. (2001). *Web site evaluation: methodology and case study.* Proceedings of the International Workshop on Data Semantics in Web Information Systems (DASWIS-2001).

[2] Hornsby K., Egenhofer M.J. (2000). *Identity-based change: a foundation for spatio-temporal knowledge representation.* International Journal of Geographical Information Science **3**, 207 – 224.

[3] The Apache Jakarta Project. *POI-HSSF - Java API To Access Microsoft Excel Format Files.* http://jakarta.apache.org/poi/hssf/

[4] Kimball R. (1996). *The data warehouse toolkit.* John Wiley & Sons.

[5] Sindoni G., De Francisci S., Paolucci M., Tininini L. (2002). *Experiences in developing a spatio-temporal information system.* Research in Official Statistics **5**, 45 – 57.

Address: M. Renzetti, G. Sindoni, L. Tininini and A. Urbano, National Institute of Statistics (Istat), via Cesare Balbo 16, 00184, Roma, Italy

E-mail: `<marenzet, sindoni, tininini, urbano>@istat.it`

THE MULTIVARIATE LEAST WEIGHTED SQUARED DISTANCES ESTIMATOR

Ella Roelant, S. Van Aelst and Gert Willems

Key words: Robust estimation, statistical computing.

COMPSTAT 2004 section: Robustness.

Abstract: In this paper the multivariate least weighted squared distance estimator is studied. A fast algorithm to compute the MLWSD estimator is developed and compared with the MCD algorithm of Rousseeuw.

1 The estimator

Let $X_n = \{\mathbf{x_1}, \dots, \mathbf{x_n}\}$ be a data set of p-variate observations. In this paper we will consider an estimator for the location μ of these data. Our estimator minimizes a weighted sum of the squared Mahalanobis distances where the weights are applied on the ranks of these distances. We are mainly interested in functions $a_n(i) = h^+(i/(n+1))$, $i = 1, \dots, n$ with $h^+ : (0,1) \to [0,\infty)$ such that

$$\sup\{u; h^+(u) > 0\} = 1 - \alpha,$$

with $0 \leq \alpha \leq \frac{1}{2}$. Hence a proportion α of the observations $\mathbf{x_i}$ are given weight 0, which makes sure that we obtain a robust estimator (see also [4]). More precisely the breakdown point ε^*, i.e. the smallest fraction of observations from X_n that need to be replaced by arbitrary values to carry the estimate beyond all bounds [3], will be at least α because a proportion α of the observations with largest robust distances does not affect the estimator.

We define the estimator as follows:

Definition 1.1. *The multivariate least weighted squared distances (MLWSD) estimator is any solution of*

$$\hat{\mu}_{MLWSD}(X_n) = \underset{\mu,\Sigma;|\Sigma|=1}{argmin}\, D_n(\mu, \Sigma)$$

where the objective function D_n is defined as

$$D_n(\mu, \Sigma) = \sum_{i=1}^{n} a_n(R_i) d_i^2(\mu, \Sigma)$$

with $d_i^2(\mu, \Sigma) = (\mathbf{x_i} - \mu)'\Sigma^{-1}(\mathbf{x_i} - \mu)$ and R_i is the rank of $d_i^2(\mu, \Sigma)$ among $d_1^2(\mu, \Sigma), \dots, d_n^2(\mu, \Sigma)$.

An equivalent formulation of the MLWSD estimator is obtained as follows. Let $\mathcal{R}$ be the set of all possible rankings of the observations. In other words,

$\mathcal{R}$ consists of all permutations of $\{1, \ldots, n\}$. Then, for any rank vector $R \in \mathcal{R}$, consider the following weighted mean and covariance matrix:

$$\begin{aligned}\hat{\mu}_{a_n}(R) &= \frac{\sum_{i=1}^n a_n(R_i)\mathbf{x_i}}{\sum_{i=1}^n a_n(R_i)} \\ \hat{\Sigma}_{a_n}(R) &= c_{h^+}\frac{\sum_{i=1}^n a_n(R_i)(\mathbf{x_i} - \hat{\mu}_{a_n})(\mathbf{x_i} - \hat{\mu}_{a_n})'}{\sum_{i=1}^n a_n(R_i)}\end{aligned}$$

where c_{h^+} is a consistency factor. The following proposition shows that the MLWSD estimator can be characterized in terms of minimizing the determinant of $\hat{\Sigma}_{a_n}(R)$ over all possible rankings of the observations.

Proposition 1.1. *For data sets in general position it holds that*

$$\begin{aligned}&\left\{\tilde{\mu}|(\tilde{\mu}, \tilde{\Sigma}) \in \underset{\mu,\Sigma;|\Sigma|=1}{\mathit{argmin}} \sum_{j=1}^n a_n(R_j)d_j^2(\mu, \Sigma)\right\} \\ &= \left\{\hat{\mu}_{a_n}(R)|R \in \underset{R\in\mathcal{R}}{\mathit{argmin}}\det(\hat{\Sigma}_{a_n}(R))\right\}\end{aligned}$$

This second characterization clearly shows that the MLWSD estimator actually generalizes the minimum covariance determinant estimator (MCD) [6]. Indeed, the MCD estimator is obtained as a special case by choosing as a weight function $a_n(i) = I(i \leq k)$ with $n/2 \leq k \leq n$. In some sense, while the MCD searches for an optimal subset of k observations, the more general MLWSD estimator searches for an 'optimally ranked' set of k observations.

Remark: Assume that the observations come from a unimodal elliptical distribution $F_{\mu,\Sigma}$ with density

$$f_{\mu,\Sigma}(x) = |\Sigma|^{-1/2}g((x - \mu)'\Sigma^{-1}(x - \mu))$$

where $\mu \in \mathbb{R}^p$, $\Sigma \in \text{PDS}(p)$, the class of positive definite symmetric matrices of order p and $g : [0, \infty) \to [0, \infty)$ has a strictly negative derivative. Then following Lopuhaä [5] the consistency factor is given by $c_{h^+} = c_1/c_3$ with

$$\begin{aligned}c_1 &= \frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty h^+(G(r^2))g(r^2)r^{p-1}dr \\ c_3 &= \frac{2\pi^{p/2}}{\Gamma(p/2)}\int_0^\infty \frac{1}{p}h^+(G(r^2))g(r^2)r^{p+1}dr\end{aligned}$$

where $G(t) = \mathcal{P}_{F_{0,I}}(Z'Z \leq t)$.

2 Algorithm

In this section we develop a fast algorithm to compute the MLWSD estimator which is similar to the MCD algorithm of Rousseeuw and Van Driessen [7] and the MLTS algorithm of Agulló et al. [1]. We make a distinction between two types of weight functions in our algorithm: a non-increasing weight function and a function that is non-decreasing on the part where its function value is not equal to zero.

2.1 Non-increasing weight function

The basis of our algorithm with this type of weight function is the following proposition:

Proposition 2.1. *Consider a data set $X_n = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ of p-variate observations and a non-increasing weight function a_n. Denote $Q_1 := \sum_{i=1}^{n} a_n(R_{1i}) d_1^2(i)$. R_1 is the rank vector of $d_1^2(i) = (\mathbf{x_i} - \hat{\mu}_1)' \tilde{\Sigma}_1^{-1} (\mathbf{x_i} - \hat{\mu}_1)$, $i = 1, \ldots, n$ where $\hat{\mu}_1 \in \mathbb{R}^{p \times 1}$ and $\tilde{\Sigma}_1 \in \mathbb{R}^{p \times p}$ with $|\tilde{\Sigma}_1| = 1$. Now compute $\hat{\mu}_2 := \hat{\mu}_{a_n}(R_1)$ and $\hat{\Sigma}_2 := \hat{\Sigma}_{a_n}(R_1)$. Denote $\tilde{\Sigma}_2 = (\det \hat{\Sigma}_2)^{-1/p} \hat{\Sigma}_2$ and $d_2^2(i) = (\mathbf{x_i} - \hat{\mu}_2)' \tilde{\Sigma}_2^{-1} (\mathbf{x_i} - \hat{\mu}_2)$, $i = 1, \ldots, n$ with corresponding rank vector R_2. With $Q_2 := \sum_{i=1}^{n} a_n(R_{2i}) d_2^2(i)$ we then have*

$$Q_2 \leq Q_1.$$

We call this construction, where a new weight vector is based on the distances $d^2(\hat{\mu}_{a_n}, \hat{\Sigma}_{a_n})$ using the weights of the old weight vector, a C-step. Following Rousseeuw and Van Driessen [7], C stands for “concentration”. Due to the use of the non-increasing weight function we make sure that the smallest distances get the highest weights, so in that sense the most concentrated points become highly weighted.

We can now use this C-step to develop the algorithm. We therefore start by drawing 1000 random $(p+1)$ subsets J_m of $\{1, \ldots, n\}$ and we compute the corresponding classical sample mean $\hat{\mu}_m$ and sample covariance matrix $\hat{\Sigma}_m$. If $\det(\hat{\Sigma}_m) = 0$ for some subset J_m then we add points to J_m until $\det(\hat{\Sigma}_m) > 0$ or $\#J_m = n$. For each subset we compute the corresponding distances $d_i(\hat{\mu}_m, \tilde{\Sigma}_m)$ $(i = 1, \ldots, n)$, where $\tilde{\Sigma}_m$ is a rescaling of $\hat{\Sigma}_m$ in order to have determinant 1, and compute the weights of these distances. Then we apply some C-steps (e.g. two) lowering each time the value of the objective function. In the following step we select the 10 subsets J_m which yielded the lowest values of the objective function and for them we carry out further C-steps until convergence. The final solution reported by the algorithm is the $\hat{\mu}$ and $\tilde{\Sigma}$ that correspond to the lowest objective value of these 10.

2.2 Non-decreasing weight function

In this case we can no longer completely rely on Proposition 2.1 and we have altered our algorithm as follows. We start with the same setting as for the non-increasing weight function using the algorithm given above. If we now apply C-steps it is not assured that the C-step each time lowers the value of the objective function. Hence we follow the general principle that if the

C-step does not lower the value of the objective function then we keep the earlier result as the final solution for that subset (and so we stop applying C-steps). We apply a maximum of 4 C-steps on each subset.

We then select the 10 subsets J_m which yielded the lowest values of the objective function and for them we carry out further C-steps until the objective function value does not decrease anymore (We set a maximum of 30 C-steps in this second phase to make sure that the algorithm stops by a given time).

3 Simulations

3.1 Finite-sample performance

In this section we investigate the finite-sample performance of the MLWSD estimator. We consider both a non-increasing and a non-decreasing weight function and compare these results with the finite-sample efficiencies of the MCD estimator. We performed several simulations as follows. For three different sample sizes ($n = 100$, 300 and 500), and for $p = 3$ we generated 1000 data sets of size n from a multivariate standard normal distribution $N(0, I_p)$. We used two different types of weight function, namely a non-increasing function:

$$\begin{aligned} a_n(i) &= \Phi^{-1}\left(\tfrac{n-i+n+1}{2(n+1)}\right), & i &= 1,\dots,k \\ &= 0, & i &= k+1,\dots,n \end{aligned}$$

and a non-decreasing function:

$$\begin{aligned} a_n(i) &= \Phi^{-1}\left(\tfrac{i+n+1}{2(n+1)}\right), & i &= 1,\dots,k \\ &= 0, & i &= k+1,\dots,n \end{aligned}$$

¿From now on we will refer to the MLWSD estimator with non-increasing weight functions as the MLWSD↓ estimator and for a non-decreasing weight function as the MLWSD↑ estimator. We considered two typical choices for k namely, $k = [(n + p + 2)/2]$ (corresponding to $\alpha = 0.50$) and $k \approx 0.75n$ (corresponding to $\alpha = 0.25$).

For each simulated data set $Z^{(l)}, l = 1,\dots,1000$ we computed the location estimator $\hat{\mu}^{(l)}_{MLWSD}$. We computed the Monte Carlo variance of the estimated vector $\hat{\mu}_{MLWSD}$ as $n \underset{1\le j\le p}{\text{ave}}\ \underset{l}{\text{var}}((\hat{\mu}_{MLWSD})^{(l)}_j)$ and we used its inverse to measure the finite-sample efficiency of the location estimator. Table 1 lists the finite-sample efficiencies for the MLWSD↓, the MLWSD↑ and the MCD estimator with $\alpha = 0.50$ and $\alpha = 0.25$. We use the notation MLWSD↓50 for $\alpha = 0.50$ and MLWSD↓25 for $\alpha = 0.25$ and similarly for the MLWSD↑ and MCD estimators. We first note that similarly to the MCD estimator [2] we find that the efficiencies for $\alpha = 0.25$ are always higher than the corresponding efficiencies for $\alpha = 0.50$. Furthermore we see that for normal data neither the MLWSD↓ nor the MLWSD↑ yields an improvement over the MCD. The MLWSD↓ estimator gives the lowest efficiencies in each case. For $\alpha = 0.50$ the efficiency loss is small but for $\alpha = 0.25$ the differences become larger.

n	100	300	500
MLWSD↓50	0.2309	0.1948	0.1908
MLWSD↑50	0.2557	0.2151	0.2092
MCD50	0.2654	0.2198	0.2127
MLWSD↓25	0.3955	0.3933	0.3861
MLWSD↑25	0.4514	0.4288	0.4029
MCD25	0.4939	0.4870	0.4877

Table 1: Finite-sample efficiencies for the MLWSD↓, the MLWSD↑ and the MCD estimator with $\alpha = 0.50$ and $\alpha = 0.25$.

3.2 Finite-sample robustness

We also performed simulations with contaminated data sets to study the finite-sample robustness of the MLWSD↓ and the MLWSD↑ estimator and compared these results with those obtained for the MCD estimator. To generate contaminated data sets with outliers we started with the uncontaminated data sets as before and we then replaced 20% or 40% of the data points $\mathbf{x_i}$ by p variables distributed according to $N(s\sqrt{\chi^2_{p,0.99}}, 1.5)$ with $s = 5, 3, 1$. For each data set $Z^{(l)}, l = 1, \ldots, 1000$ we computed the location estimator $\hat{\mu}^{(l)}_{MLWSD}$. The mean squared error of the vector $\hat{\mu}_{MLWSD}$ is given by

$$\text{MSE}(\hat{\mu}_{MLWSD}) = n \underset{1\leq j\leq p}{\text{ave}} \underset{l}{\text{ave}}[\{(\hat{\mu}_{MLWSD})^{(l)}_j\}^2]$$

and the bias is

$$\text{bias}(\hat{\mu}_{MLWSD}) = \sqrt{\underset{1\leq j\leq p}{\text{ave}}[\{\underset{l}{\text{ave}}(\hat{\mu}_{MLWSD})^{(l)}_j\}^2]}.$$

Table 2 shows the MSE and the bias of the different estimators for $p = 3$ and sample sizes $n = 100$, 300 and 500 and for 40% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 1.5)$. We notice here that for the larger data sets the bias of both the MLWSD↑ and the MLWSD↓ estimator is smaller than the bias of the MCD estimator. (The results for $s = 5$ and $s = 3$ for 20% and 40% outliers and $\alpha = 0.50$, which are not reported here, yield the same conclusions in case of the non-decreasing weight function. The non-increasing weight function gives no improvement to the MCD estimator for these situations.) If we compare the two different weight functions, we see that for small data sets the MLWSD↓ estimator yields the best results. On the other hand for large data sets the MLWSD↑ has a smaller MSE and bias.

We repeated these simulations for the different sample sizes 100, 300 and 500, but this time with more concentrated outliers, distributed according to $N(s\sqrt{\chi^2_{3,0.99}}, 0.1)$ with $s = 5, 3, 1$. Table 3 shows the MSE and the bias for the MLWSD↓ and MCD estimator for 20% outliers for the case $s = 1$. For

$n = 100$ we see that the bias of the MLWSD↓ estimator is smaller than the bias of the MCD estimator. For $n = 500$ this is no longer true.

n	100		300		500	
	MSE	bias	MSE	bias	MSE	bias
MLWSD↓50	2.7391	0.00470	2.9580	0.00261	3.0831	0.00087
MLWSD↑50	5.6873	0.02642	2.9603	0.00219	3.0180	0.00020
MCD50	2.8442	0.00729	2.6207	0.00305	2.7519	0.00144

Table 2: 40% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 1.5)$.

n	100		300		500	
	MSE	bias	MSE	bias	MSE	bias
MLWSD↓50	5.1106	0.00946	4.6254	0.00400	4.0047	0.00212
MCD50	5.0533	0.01614	3.5652	0.00586	3.8183	0.00163
MLWSD↓25	1.9474	0.00536	1.7598	0.00290	1.9921	0.00197
MCD25	3.2058	0.01826	1.4193	0.00230	1.5839	0.00123

Table 3: 20% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 0.1)$.

n	100		300		500	
	MSE	bias	MSE	bias	MSE	bias
MLWSD↑50	444.59	2.09991	1399.12	2.15572	2365.40	2.17250
MCD50	645.16	2.53583	2072.52	2.62674	3514.28	2.65013

Table 4: 40% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 0.1)$.

Table 4 shows the MSE and the bias for the MLWSD↑ and MCD estimators for 40% outliers from $N(\sqrt{\chi^2_{3,0.99}}, 0.1)$. We see here that the bias and the MSE of the MLWSD↑ estimator are smaller than those of the MCD estimator.

Although the MLWSD estimators do not improve the efficiency of the MCD for normal data, depending on the type of contamination they can give lower bias. In general we see that the MLWSD↑ usually has a lower bias than the MCD while not losing much efficiency.

4 Example

To illustrate our estimator we use the Philips data set which was also used in Rousseeuw and Van Driessen [7]. The data set consists of nine characteristics measured on $n = 677$ diaphragm parts for TV sets. Because recently a new production line was started, the goal is to gain insight in the production process and the interrelations between the nine measurements and to find out whether deformations or abnormalities have occurred and why.

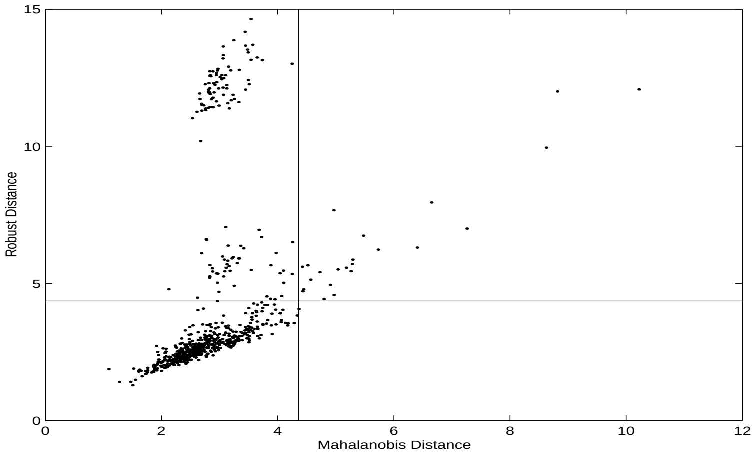


Figure 1: Philips data (using MLWSD↑25): Distance-distance plot.

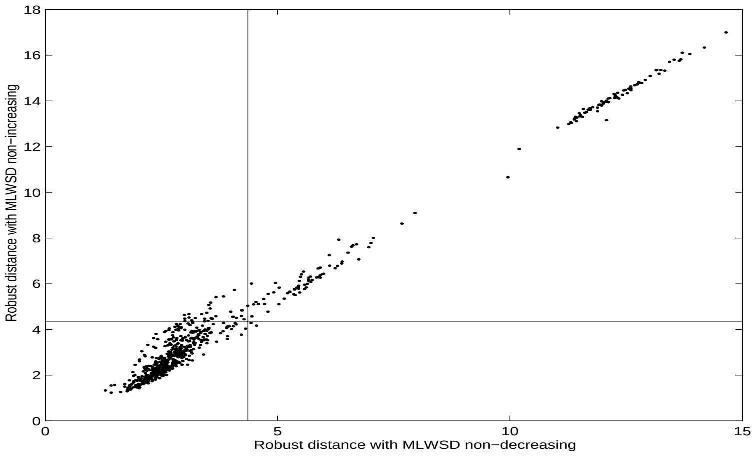


Figure 2: Philips data: Plot of the robust distances of the MLWSD↓ estimator versus the robust distances of MLWSD↑ estimator for $\alpha = 0.25$.

We will use the MLWSD estimator to compute the robust distances of the data points. Figure 1 shows a distance-distance plot which plots the robust distances (based on the MLWSD↑ estimator for $\alpha = 0.25$) versus the classical Mahalanobis distances. We have indicated the usual cutoff value $\sqrt{\chi^2_{9,0.975}} = 4.361$. This plot shows that a group of strongly deviating observations is detected by the robust MLWSD↑ estimator but not by the classical Mahalanobis distances derived from the empirical mean and covariance. To compare the MLWSD estimators with non-increasing and non-decreasing weight functions, Figure 2 shows a plot of the robust distances of the MLWSD↓ estimator versus the robust distances of MLWSD↑ estimator for $\alpha = 0.25$. The lines again correspond to the cutoff value 4.361. Both estimators clearly detect the strong outliers lying far away from the majority of the data. However we notice that there are several points that lie above the cutoff for the MLWSD↓ estimator and under the cutoff for the MLWSD↑ estimator. A look at the data teaches us that these points mainly belong to the first 100 measurements. From Rousseeuw and Van Driessen [7] we know that after these measurements the production line was adjusted and that we can consider these data points as intermediate outliers. The treatment of the intermediate outliers differs between both estimators. The MLWSD↓ reveals this effect because it gives low weight to the points further from the center. On the other hand the MLWSD↑ gives a high weight to these intermediate outliers so that they are now masked.

References

[1] Agulló J., Croux C., Van Aelst S. (2001). *The multivariate least trimmed squares estimator*. Submitted.

[2] Croux C., Haesbroeck G. (1999). *Influence function and efficiency of the minimum covariance determinant scatter matrix estimator*. Journal of Multivariate Analysis **71**, 161 − 190.

[3] Donoho D.L., Huber P.J. (1983). *The notion of breakdown point*. A Festschrift for Erich Lehmann, P.J. Bickel, K.A. Doksum and J.L. Hodges, (eds.), Belmont, Wadsworth, 157 − 184.

[4] Hössjer O. (1994). *Rank-based estimates in the linear model with high breakdown point*. Journal of the American Statistical Association **89**, 149 − 158.

[5] Lopuhaä H.P. (1999). *Asymptotics of reweighted estimators of multivariate location and scatter*. The Annals of Statistics **27**, 1638 − 1665.

[6] Rousseeuw P.J. (1984). *Least median of squares regression*. Journal of the American Statistical Association **79**, 871 − 880.

[7] Rousseeuw P.J. and Van Driessen K. (1999). *A fast algorithm for the minimum covariance determinant estimator*. Technometrics **41**, 212 − 223.

Address: E. Roelant, S. Van Aelst and G. Willems, Dept. of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium

E-mail: `Ella.Roelant@UGent.be`

COMPSTAT'2004 Symposium

QUANTILE ESTIMATION WITH CALIBRATION ESTIMATORS

M. Rueda García, S. Martínez Puertas, H. Martínez Puertas, Y. Román Montoya, S. González Aguilera

Key words: Calibration, auxiliary information, finite population quantile.

COMPSTAT 2004 section: Others.

Abstract: In this work we are going to construct new estimators for the finite population quantile of a study variable defined in a finite population by means of estimators of the distribution function obtained with the calibration method. Comparisons are made with existing estimators in a simulation study using a natural population.

1 Introduction

In sample surveys, auxiliary population information is often used at the estimation stage to increase the precision of estimators of a population total, mean or distribution function. To incoporate the auxiliary information in the estimation of the distribution function we will use the calibration method of Deville and Särndal [2] and we will obtain new estimators of the distribution function. We study the principal properties of this new estimators and under some conditions, this calibrated estimators are distribution function. This property is not usual in most of the estimators that use the auxiliary information to estimate the distribution function. Finally we use the calibrated estimators of the distribution function to obtain new estimators of the quantiles of the study variable. A simulation study is included to compare the precision of the new estimators of quantiles with the usual estimators.

2 Calibration estimators of the distribution function

Consider a finite population $U = \{1, \dots, k, \dots, N\}$, consists of N different elements. Let $s = \{1, \dots, n\}$ be the set of n units included in a sample, selected according to a specified sampling design which inclusion probabilities π_k and π_{kl} are assumed to be strictly positive. Let y_k be the value of the variable of interest y, for the kth population element, with which also is associated and auxiliary vector value $\mathbf{x}_k = (x_{k1}, x_{k2}, \dots, x_{kJ})$. The values $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ are known for the entire population but y_k is known only if the kth unit is selected in the sample s. The finite population distribution function of the study variable y, is defined by

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k)$$

with $\Delta(t-y_k)=0$ if $t<y_k$ and $\Delta(t-y_k)=1$ if $t\geq y_k$. Is known that the distribution function $F_y(t)$ can be estimated by the Horvitz-Thompson estimator, given by

$$\widehat{F}_{YH}(t)=\frac{1}{N}\sum_{k\in s}d_k\Delta(t-y_k)$$

with $d_k=1/\pi_k$, the basic design weights. The estimator $\widehat{F}_{YH}(t)$ is unbiased, but it present the objection that, in general, is not a distribution function and does not use the auxiliary information provided by the vector $\mathbf{x}$. We shall modify the estimator $\widehat{F}_{YH}(t)$ to obtain new estimators of $F_y(t)$, replacing the basic design weights d_k by a new weights ω_k which are constructed with the calibration techniques. First, we shall consider the calibration estimator

$$\widehat{F}_{yc}(t)=\frac{1}{N}\sum_{k\in s}\omega_k\Delta(t-y_k)$$

where the new weights ω_k are modified from $d_k=1/\pi_k$ minimizing the chi-square distance measure

$$\Phi_s=\sum_{k\in s}\frac{(\omega_k-d_k)^2}{d_kq_k} \tag{1}$$

with q_k known positive constants unrelated to d_k, subject to the calibration equation

$$\frac{1}{N}\sum_{k\in s}\omega_k\Delta(t_0-g_k)=F_g(t_0) \tag{2}$$

The values g_k that appear in the constranint (2) are given by $g_k=\widehat{\beta}'\mathbf{x}_k$ with

$$\widehat{\beta}=\left(\sum_{k\in s}d_kq_k\mathbf{x}_k^{'}\mathbf{x}_k\right)^{-1}\cdot\sum_{k\in s}\mathbf{x}_ky_k \tag{3}$$

a weighted estimator of the multiple regression coefficient between y and $\mathbf{x}$; $F_g(t_0)$ denotes the finite distribution function of the variable g evaluated in an arbitrary chosen point t_0. Now, to obtain the weights ω_k we have to minimize (1) subject to (2). Minimization leads to the calibrated weights

$$\omega_k=d_k+\frac{d_kq_kN\Delta(t_0-g_k)\Big(F_g(t_0)-\widehat{F}_{GH}(t_0)\Big)}{\displaystyle\sum_{k\in s}d_kq_k\Delta(t_0-g_k)} \tag{4}$$

with $\widehat{F}_{GH}(t_0)$ denotes the Horvitz-Thompson estimator for the distribution function of g in the point t_0 and assuming that

$$\sum_{k\in s}d_kq_k\Delta(t_0-g_k)\neq 0$$

for which is only sufficient choose t_0 adequately. The resulting estimator of $F_y(t)$ is

$$\widehat{F}_{yc}(t) = \widehat{F}_{YH}(t) + \Big(F_g(t_0) - \widehat{F}_{GH}(t_0)\Big)\widehat{B} \tag{5}$$

with

$$\widehat{B} = \frac{\displaystyle\sum_{k\in s} d_k q_k \Delta(t_0 - g_k)\Delta(t - y_k)}{\displaystyle\sum_{k\in s} d_k q_k \Delta(t_0 - g_k)} \tag{6}$$

The next question is if the estimator $\widehat{F}_{yc}(t)$ is a distribution function. For it, we have to verify if the properties of the distribution function are satisfied. Its easy to verify that $\widehat{F}_{yc}(t)$ is continue on the right. On the other hand, $\widehat{F}_{yc}(t)$ is not monotone nondecreasing, in general and we have:

$$\lim_{t\to-\infty} \widehat{F}_{yc}(t) = \lim_{t\to-\infty} \frac{1}{N}\sum_{k\in s} \omega_k \Delta(t - y_k) = 0$$

and

$$\lim_{t\to+\infty} \widehat{F}_{yc}(t) = \frac{1}{N}\sum_{k\in s} d_k - \frac{1}{N}\sum_{k\in s} d_k\Delta(t_0 - g_k) + F_g(t_0)$$

This value is not equal to the unit, in general. We are going to solve these problems and we will begin with the condition ii). Since $\widehat{F}_{yc}(t)$ is a calibration estimator, it is easy to see that $\widehat{F}_{yc}$ is monotone nondecreasing if and only if ω_k are positives for all sample units. For this reason we shall seek the conditions that guarantee that $\omega_k \geq 0$. It's clear that ω_k are positive if $q_k = c$ for all sample units. The uniform constans $q_k = 1$ is likely to dominate in applications (see *Deville* and *Särndal* [2]), therefore in the most of cases the calibrated weights ω_k are positive and then $\widehat{F}_{yc}(t)$ is monotonically increasing. To solve the condition

$$\lim_{t\to+\infty} \widehat{F}_{yc}(t) = \frac{1}{N}\sum_{k\in s} \omega_k = 1$$

we can choose t_0 sufficiently big, but this choice joined with $q_k = c$ produces the Háyek estimator. Another way is to add the restriction

$$\frac{1}{N}\sum_{k\in s} \omega_k = 1 \tag{7}$$

to the calibration process. This condition is equivalent to the condition

$$\frac{1}{N}\sum_{k\in s} \omega_k \Delta(t_i - g_k) = F_g(t_i)$$

when t_i is sufficiently big. Thus we shall consider the calibrated weights that minimizing the chi-square distance (1) subject to the calibration equations

$$\frac{1}{N}\sum_{k\in s}\omega_k\Delta(t_j-g_k)=F_g(t_j)\quad j=1,2,\ldots,P \tag{8}$$

where t_j; $j=1,2,\ldots,P$ are points that we choose arbitraryly and assumed that $t_1\leq t_2\leq\ldots\leq t_P$; with t_P sufficiently big to guarantee the condition (7). Now, we denote by

$$\mathbf{t}^{'}=(t_1,\ldots,t_P)\ ;\ \Delta(\mathbf{t}-g_k)^{'}=\Big(\Delta(t_1-g_k),\ldots,\Delta(t_P-g_k)\Big)$$

$$F_g(\mathbf{t})=\Big(F_g(t_1),\ldots,F_g(t_P)\Big)\ ;\ \widehat{F}_{GH}(\mathbf{t})=\Big(\widehat{F}_{GH}(t_1),\ldots,\widehat{F}_{GH}(t_P)\Big)$$

In accordance with this new calibration process, we have to minimize (1) subject to (8) and the new weights obtained are

$$\omega_k=d_k+d_kq_kN\Big(F_g(\mathbf{t})-\widehat{F}_{GH}(\mathbf{t})\Big)^{'}\cdot T^{-1}\cdot\Delta(\mathbf{t}-g_k) \tag{9}$$

assuming that the inverse of

$$T=\sum_{k\in s}d_kq_k\Delta(\mathbf{t}-g_k)\Delta(\mathbf{t}-g_k)^{'}$$

exists. The resulting estimator can be written as

$$\widehat{F}^{*}_{yc}(t)=\widehat{F}_{YH}(t)+\Big(F_g(\mathbf{t})-\widehat{F}_{GH}(\mathbf{t})\Big)\cdot\widehat{D} \tag{10}$$

where

$$\widehat{D}=T^{-1}\cdot\sum_{k\in s}d_kq_k\Delta(\mathbf{t}-g_k)\Delta(t-y_k)$$

If T is singular, the calibration process doe not have solution and we take $\widehat{F}^{*}_{yc}(t)=\widehat{F}_{YH}(t)$. It's clear that $\widehat{F}_{yc}(t)$ is a particular case of $\widehat{F}^{*}_{yc}(t)$.

Next, we shall obtain a expression of T^{-1} and then we can give another expression of the weights (9) with the aim to study better the properties of the estimator $\widehat{F}^{*}_{yc}(t)$. For it, we remember that the points t_j are ordered, that is $t_1<t_2<\ldots<t_P$ and we consider the g-values of sample units in ascending order $g_{(1)}\leq g_{(2)}\leq\ldots g_{(n-1)}\leq g_{(n)}$ and we suppose that the value t_i is bigger than the first k_i sample values of the variable g, with $k_i>k_{i-1}$ for $i=2,\ldots,P$ and $k_p=n$. Under these considerations T^{-1} is a $P\times P$ symmetric matrix $(a_{ij})_{i,j=1}^{P}$ with

$$a_{11} = \frac{1}{\sum_{k=1}^{k_1} d_{(k)} q_{(k)}} + \frac{1}{\sum_{k=k_1+1}^{k_2} d_{(k)} q_{(k)}} \quad ; \quad a_{12} = \frac{-1}{\sum_{k=k_1+1}^{k_2} d_{(k)} q_{(k)}}$$

and $a_{1j} = 0$; $j = 3, 4, \dots P$. For $i = 2, 3, \dots, J-1$

$$a_{i,i} = \frac{1}{\sum_{k=k_{i-1}+1}^{k_i} d_{(k)} q_{(k)}} + \frac{1}{\sum_{k=k_i+1}^{k_{i+1}} d_{(k)} q_{(k)}} \quad ; \quad a_{i,i+1} = \frac{-1}{\sum_{k=k_i+1}^{k_{i+1}} d_{(k)} q_{(k)}}$$

with $a_{ij} = 0$ if $j \neq i-1$; $j \neq i$ and $j \neq i+1$ and

$$a_{JJ} = \frac{1}{\sum_{k=k_{J-1}+1}^{n} (d_k - 1) q_k}$$

Thus, if $k_1 < k_2 < \cdots < k_{J-1} < k_J$ the matrix T^{-1} is not singular. Once determined T^{-1} we can get a new expression for the weights substituting the expression of T^{-1} in (9) and

$$\omega_{(j)} = d_{(j)} + \frac{d_{(j)} q_{(j)} N \Big(F_g(t_1) - \widehat{F}_{GH}(t_1) \Big)}{\sum_{k=1}^{k_1} d_{(k)} q_{(k)}} \qquad j = 1, 2, \dots, k_1$$

In general, the weights for the sample units $j = k_{i-1} + 1, \dots, k_i$ with $i = 2, \dots, J$ are

$$\omega_{(j)} = d_{(j)} - \frac{d_{(j)} q_{(j)} N \Big(F_g(t_{i-1}) - \widehat{F}_{GH}(t_{i-1}) \Big)}{\sum_{k=k_{i-1}+1}^{k_i} d_{(k)} q_{(k)}} + \frac{d_{(j)} q_{(j)} N \Big(F_g(t_i) - \widehat{F}_{GH}(t_i) \Big)}{\sum_{k=k_{i-1}+1}^{k_i} d_{(k)} q_{(k)}}$$

Now, we can study if $\widehat{F}^*_{yc}(t)$ is a distribution function. Clearly, $\widehat{F}^*_{yc}(t)$ is continue on the right and $\lim_{t \to -\infty} \widehat{F}^*_{yc}(t) = 0$. Also $\lim_{t \to +\infty} \widehat{F}^*_{yc}(t) = 1$; because the new system of weights (9) give perfect estimates in t_P, and then

$$1 = F_g(t_P) = \frac{1}{N} \sum_{k \in s} \omega_k \Delta(t_P - g_k) = \frac{1}{N} \sum_{k \in s} \omega_k$$

In general, $\widehat{F}^*_{yc}(t)$ is not monotone nondecreasing but $\widehat{F}^*_{yc}(t)$ is a calibrated estimator, consequently $\widehat{F}^*_{yc}(t)$ is monotone nondecreasing if and only if ω_k are positive. It is possible to demostrate that if $q_k = c$ with c a positive constan, then the weights (9) are positive and $\widehat{F}^*_{yc}(t)$ is monotone nondecreasing.

3 Quantile estimation with calibration estimators of of the distribuction function $\mathbf{F_y(t)}$

In this section we use the calibrated estimators for the estimation of β-quantile of the study variable y

$$Q_y(\beta) = F_y^{-1}(\beta) = inf\{t : F_y(t) \geq \beta\}$$

with $F_y(t)$ the distribution function of y evaluated at t.

A general procedure to estimate $Q_y(\beta)$ consists of constructing of a estimator $\widehat{F}_y(t)$ of $F_y(t)$, that must be monotone nondecreasing and then the quantile $Q_y(\beta)$ is estimated by the inverse of the estimator $\widehat{F}_y^{-1}(t)$, that is

$$\widehat{Q}_y(\beta) = \widehat{F}_y^{-1}(\beta) = \inf\{t : \widehat{F}_y(t) \geq \beta\}$$

Now, we use the estimator $\widehat{F}_{yc}(t)$ and $\widehat{F}^*_{yc}(t)$ to obtain a new estimators of $Q_y(\beta)$ following the general procedure previously described. Both estimators are monotone nondecreasing if $q_k = c$ for all units sample, therefore we use this choice to estimate $Q_y(\beta)$. Since the estimator $\widehat{F}_{yc}(t)$ is used to estimate $Q_y(\beta)$ it is logical to take $t_0 = Q_g(\beta)$ and thus $\widehat{F}_{yc}(t)$ give perfect estimates for $F_g(Q_g(\beta)$. For the same reason, we take the points $t_j = Q_g(\alpha_j)$ for $j = 1, 2, \ldots, P-1$ and $t_P = \max_{k \in U} g_k$, to construct $\widehat{F}^*_{yc}(t)$ with $0 < \alpha_j < 1$ and $\alpha_i = \beta$ for some $i \in \{1, 2, \ldots P-1\}$. The new estimators of $Q_y(\beta)$ are

$$\widehat{Q}_{yc}(\beta) = inf\{t : \widehat{F}_{yc}(t) \geq \beta\} \tag{11}$$

$$\widehat{Q}^*_{yc}(\beta) = inf\{t : \widehat{F}^*_{yc}(t) \geq \beta\} \tag{12}$$

If we want to estimate the median $M_y = Q_y(0.5)$ of the study variable y, then the election of t_0, to construct $\widehat{F}_{yc}(t)$ is $t_0 = M_g = Q_g(0.5)$ and to obtain $\widehat{F}^*_{yc}(t)$ we can take

$$t_1 = Q_g(0.25) = Q_{1g} \;;\; t_2 = M_g \;;\; t_3 = Q_g(0.75) = Q_{3g} \text{ and } t_4 = \max_{k \in u} g_k$$

The estimator of M_y are given by

$$\widehat{M}_{yc} = inf\{t : \widehat{F}_{yc}(t) \geq 0.5\} \tag{13}$$

$$\widehat{M}^*_{yc} = inf\{t : \widehat{F}^*_{yc}(t) \geq 0.5\} \tag{14}$$

4 Simulation study

In this section we compare the precision of the estimators $\widehat{M}_{yc}$ and $\widehat{M}^*_{yc}$ with the following estimators, $\widehat{M}_{CD}$ the estimator for the median obtained with the Chambers-Dunstan estimator $\widehat{F}_{CD}(t)$ [1], $\widehat{M}_{RKM}$ the estimator of the median derived from the Rao-Kovar-Mantel estimator $\widehat{F}_{RKM}(t)$ [4], $\widehat{M}_{ps}$ the median estimator derived from $\widehat{F}_{ps}(t)$, [5] and the estimators $\widehat{M}_d$ and $\widehat{M}_R$ derived from the difference estimator $\widehat{F}_d(t)$ and the ratio estimator $\widehat{F}_R(t)$, respectively.

The empirical study has been carried out with a natural population called Sugar Cane, formed by 338 farms dedicated to the production of the sugar cane. This population was used originally by Chambers and Dunstan [1]. The study variable is y=*Income of sugar cane* and only we will take an auxiliary variable $\mathbf{x} = x_1$ =*Area of sugar cane.*

We selected 1000 samples for three different sample sizes, $n = 50$; $n = 75$ and $n = 100$ under simple random sampling without replacement (SRSWOR) and for every estimators of median of the study variable y we calculate the relative bias (RB) and de coefficient of variation (CV).

Estimator	RB	CV	RB	CV	RB	CV
	$n = 50$		$n = 75$		$n = 100$	
$\widehat{M}_{CD}$	0,080	0,042	0,062	0,035	0,056	0,031
$\tilde{M}_d$	0,001	0,078	0,001	0,063	0,003	0,055
$\tilde{M}_R$	-0,061	0,092*	-0,04	0,080**	-0,035	0,054***
$\widehat{M}_{ps}$	-0,001	0,071	0,001	0,059	-0,001	0,0481
$\tilde{M}_{RKM}$	0,002	0,0674	0,002	0,056	0,001	0,0467
$\widehat{M}_{yc}$	0,004	0,085	0,002	0,058	-0,005	0,036
$\widehat{M}^*_{yc}$	-0,001	0,070	0,001	0,058	0.000	0,047

* only with 233 data, ** only with 85 data and *** only with 32 data

Firstly, the table reveals that the Chambers-Dunstan estimator have the less variance, but the bias of this estimator is very large. The behaviour of the ratio estimator is very bad: the estimator present high bias and variance; moreover in a lot of samples the estimator presents values out of range. The performance of calibration estimators is acceptable: their bias and variances are small for all sample sizes. The calibration estimators are easy to implement, and make a valid alternative to other estimators of distribution function.

References

[1] Chambers, R.L, Dunstan, R. (1986). *Estimating distribution function from survey data.* Biometrika **73**, 597 – 604.

[2] Deville, J.C., Särndal. (1992). *Calibration estimators in survey sampling.* Journal of the American Statistical Association **79**, 376 – 604.

[3] Rao J.N.K. (1994). *Estimating totals and distribution functions using auxiliary information at the estimation stage.* Journal of Official Statistics **10**, 153 – 165.

[4] Rao J.N.K, Kovar J.G, Mantel H.J. (1990). *On estimating distribution functions and quantiles from survey data using auxiliary information.* Biometrika **77**, 365 – 375.

[5] Silva P.L.D, Nascimento, Skinner C.J. (1995). *Estimating distribution functions with auxiliary information using poststratification.* J. Official Statist **11**, 277 – 294.

Address: M. Rueda García, Department of Statistics & OR, University of Granada, Spain

S. Martínez Puertas, Department of Statistics & Matematic Applied, University of Almería, Spain

H. Martínez Puertas, Department of Statistics & OR, University of Granada, Spain

E-mail: `mrueda@ugr.es, spuertas@ual.es`

A BAYESIAN MODEL FOR BINOMIAL IMPERFECT SAMPLING

M. Ruiz, F.J. Girón, C. Rojano, C. Pérez and J. Martín

Key words: Bayesian inference, imperfect data, noisy Bernoulli process, Gibbs sampling.

COMPSTAT 2004 section: Bayesian methods.

Abstract: We develop a Bayesian model to estimate the proportion θ in a noisy Bernoulli process as well as the noise parameters λ_1 and λ_0. By introducing auxiliary or latent variables we are able to make not only inferences on θ, λ_1 and λ_0, but also inferences for the real status of a concrete individual. Furthermore, Gibbs sampling-based algorithms are easily implemented.

1 Introduction

In this paper we study the problem of making inferences with imperfect data from a noisy Bernoulli process. In real contexts, the obtained data sometimes do not reflect the true state of the elements in a sample. This fact is very common in election surveys (voters are reluctant to provide their true opinion), in medical diagnostics (test failures), in criminology studies (many crimes remain unreported) or in consumer surveys for marketing research (consumers may not remember their behavior or they misunderstand survey questions). The main consequence is that such distortions or noises can have an important effect on inferences because the effective amount of information obtained from the sample is reduced considerably.

Inferences about the proportion θ of individuals presenting a certain characteristic are usually based on data obtained from a dichotomous process. This process is usually modeled by using Bernoulli distributions with parameter θ. The usual presence of noises in this kind of processes makes inferences about θ difficult.

Based on likelihood criteria several techniques have been developed to solve this problem. However, all of them provide joint information about the parameter of interest and the parameters characterizing the distortion or noise that affect the data. By performing a likelihood analysis, Gaba and Winkler [3] found an identification problem in a dichotomous data model. However, this problem can be avoided by using a Bayesian approach. In this way, separate information for the two kinds of parameters (proportion and noise) can be obtained through their appropriate posterior distributions.

Winkler [6] studied a Bayesian model for imperfect sampling in a dichotomous process under the assumption of known noise parameters. For the most interesting case of unknown noise level, Winkler and Gaba [7] presented a Bayesian model with a single noise parameter. This work was generalized by Gaba and Winkler [3] by describing a Bayesian model with two noise parameters that are independent of the proportion θ. Later, Gaba [2] presented a model formalizing a prior dependence between the proportion and the noise parameter.

In this paper we develop a Bayesian model related to the proportion θ in a Bernoulli process and to the noise parameters denoted by λ_1 and λ_0. This Bayesian model is an augmented version of the one proposed by Gaba and Winkler [3]. This extension introduces explicitly auxiliary variables of great interest in this context. The introduction of auxiliary variables has the advantage of allowing a double perspective for the problem:

(1). Inferences about the proportion and the noise parameters can be carried out by using the appropriate posterior and predictive distributions.

(2). Inferences about the real status for one or more individuals can be obtained.

Furthermore, these auxiliary variables allow an easy implementation of Gibbs sampling-based algorithms.

The outline of the paper is as follows. Section 2 presents the model including basic concepts, applicability conditions, main results and a Gibbs sampling-based algorithm to solve the problem of estimating the proportion and the noise parameters. Section 3 presents the way to make inferences about the status of any individual. A Gibbs sampling-based algorithm is developed to estimate the real status of a particular individual. An illustrative example is provided in Section 4. Finally, conclusions are presented in Section 5.

2 The Bayesian model

Firstly, we define some basic concepts. Let O_n be an observed sample and h an individual from O_n.

Definition 2.1. *The position of h is a function $u(h) = u_h$ taking value 1 if $h \in A$ (positive position) and 0 if $h \in A^c$ (negative position).*

Definition 2.2. *The classification of h is a function $v(h) = v_h$ taking value 1 if h is classified in A (positive classification) and 0 if h is classified in A^c (negative classification).*

That is, u_h and v_h are the real and the reported values for h.

Definition 2.3. *The position (classification) vector of O_n is defined by $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)'$ ($\boldsymbol{v} = (v_1, v_2, \ldots, v_n)'$).*

Definition 2.4. *The effective (observed) frequency of O_n is defined by $x = \sum_{h=1}^n u_h$ ($y = \sum_{h=1}^n v_h$).*

Definition 2.5. *The partition $\mathcal{A} = \{A_{11}, A_{01}, A_{10}, A_{00}\}$ is defined in the set of individuals of O_n, where $A_{ij} = \{h : u_h = i, v_h = j\}$.*

Definition 2.6. *The status of the individual h of O_n is a vector function $\boldsymbol{s}(h) = \boldsymbol{s}_h$ taking values $(1,0,0,0)'$, $(0,1,0,0)'$, $(0,0,1,0)'$, $(0,0,0,1)'$ if h is in A_{11}, A_{01}, A_{10} or A_{00}, respectively.*

Definition 2.7. *The latent vector of O_n is $\boldsymbol{k} = (k, k', k'', k''')'$ where the components are the unknown cardinals of the sets in $\mathcal{A}$.*

Proposition 2.1. *The relations $k + k' + k'' + k''' = n$, $k + k'' = x$ and $k + k' = y$ hold for O_n.*

Proposition 2.2. *Each pair $(\boldsymbol{u}, \boldsymbol{v})$ produces a single vector called latent vector induced by $\boldsymbol{u}$ and $\boldsymbol{v}$. This vector is denoted by $\boldsymbol{k}_{\boldsymbol{u},\boldsymbol{v}} = (k_{\boldsymbol{u},\boldsymbol{v}}, k'_{\boldsymbol{u},\boldsymbol{v}}, k''_{\boldsymbol{u},\boldsymbol{v}}, k'''_{\boldsymbol{u},\boldsymbol{v}})'$.*

Firstly, we are interested in the following unknown parameters: the proportion θ of positive positions, the probability λ_1 of correct positive classification and the probability λ_0 of incorrect positive classification. The random vectors and other random variables we use to set the model are related to the concepts introduced in the previous definitions, i.e. $\boldsymbol{U} = (U_1, U_2, \ldots, U_n)'$, $X = \sum_{h=1}^{n} U_h$, $\boldsymbol{V} = (V_1, V_2, \ldots, V_n)'$, $Y = \sum_{h=1}^{n} V_h$, $\boldsymbol{S}_h$, and $\boldsymbol{K} = (K, K', K'', K''')'$.

The following conditions are assumed[1]:

(1). A joint prior density for $(\theta, \lambda_1, \lambda_0)$ depending on a known hyperparameter $\boldsymbol{\pi} = (\boldsymbol{\rho}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_0)$ is set. This density verifies $f(\theta, \lambda_1, \lambda_0|\boldsymbol{\pi}) = = f(\theta|\boldsymbol{\rho}) f(\lambda_1|\boldsymbol{\pi}_1) f(\lambda_0|\boldsymbol{\pi}_0)$.

(2). $U_h \perp U_k \,|\, \theta, \lambda_1, \lambda_0$ if $h \neq k$; $U_h \perp \lambda_1, \lambda_0 \,|\, \theta$ for $h = 1, 2, \ldots, n$.

(3). $V_h \perp U_k, V_k \,|\, U_h, \lambda_1, \lambda_0$ if $h \neq k$; $V_h \perp \theta \,|\, U_h, \lambda_1, \lambda_0$ for $h = 1, 2, \ldots, n$.

(4). $U_h \perp \theta, \lambda_1, \lambda_0 \sim \mathrm{Ber}(\theta)$; $V_h \perp U_h, \lambda_1, \lambda_0 \sim \mathrm{Ber}(\lambda_i)$ if $U_h = i$.

For reasons of space, we only present the main results. Neither the intermediate procedures nor the proofs are shown here. We focus on the joint posterior density $f(\theta, \lambda_1, \lambda_0|y, \boldsymbol{\pi})$ and on the posterior predictive distribution $f(x|y, \boldsymbol{\rho})$. The following results allow to simplify the generating process from these distributions by using Gibbs sampling.

Theorem 2.1. *The variables K and K'' are conditionally independent given θ, λ_1, λ_0, y, and their conditional distributions are respectively $f(k|\theta, \lambda_1, \lambda_0, y) \sim Bi\left(y, \frac{\theta\lambda_1}{\omega}\right)$ and $f(k''|\theta, \lambda_1, \lambda_0, y) \sim Bi\left(n - y, \frac{\theta(1-\lambda_1)}{1-\omega}\right)$ where $\omega = \theta\lambda_1 + (1-\theta)\lambda_0$.*

[1] $\cdot \perp \cdot | \cdot$ denotes conditional independence of the first two quantities given the third. Results about independence in imprecise data models can be found in Girón *et al [4]*.

Corollary 2.1. *The distribution of the vector $\boldsymbol{K}$ conditioned by θ, λ_1, λ_0, y, is* $f(\boldsymbol{k}|\theta,\lambda_1,\lambda_0,y) =$

$$= \begin{cases} f(k|\theta,\lambda_1,\lambda_0,y)f(k''|\theta,\lambda_1,\lambda_0,y) & \text{if } 0 \le k \le y,\ 0 \le k'' \le n-y \\ 0 & \text{otherwise} \end{cases}$$

Corollary 2.2. *The distribution of X conditioned by θ, λ_1, λ_0, y, is the convolution of the distributions of K and $K'' = X - K$, i.e.* $f(x|\theta,\lambda_1,\lambda_0,y) = = \sum_{k=0}^{x} f(k|\theta,\lambda_1,\lambda_0,y) f(x-k|\theta,\lambda_1,\lambda_0,y)$.

Lemma 2.1. *If the latent vector $\boldsymbol{k} = (k,k',k'',k''')'$ is known for the sample O_n, then* $f(\theta,\lambda_1,\lambda_0|\boldsymbol{k},\boldsymbol{\pi}) = f(\theta|\boldsymbol{k},\boldsymbol{\rho})f(\lambda_1|\boldsymbol{k},\boldsymbol{\pi_1})f(\lambda_0|\boldsymbol{k},\boldsymbol{\pi_0})$.

Theorem 2.2. *The joint posterior distribution $f(\theta,\lambda_1,\lambda_0|y,\boldsymbol{\pi})$ can be represented as* $\sum_{x=0}^{n}\sum_{k=0}^{y} f(k,x|y,\boldsymbol{\pi})f(\theta|x,\boldsymbol{\pi})f(\lambda_1|k,x,\boldsymbol{\pi})f(\lambda_0|y-k,x,\boldsymbol{\pi})$.

The following Gibbs sampling-based algorithm allows to generate from the joint posterior density $f(\theta,\lambda_1,\lambda_0|y,\boldsymbol{\pi})$ and from the posterior predictive distribution $f(x|y,\boldsymbol{\rho})$. Note that all necessary distributions are easy to generate from.

Algorithm 2.1. *Initial values $\theta^{(0)}$, $\lambda_1^{(0)}$, and $\lambda_0^{(0)}$ are fixed. For each iteration t, steps 1-3 must be followed.*

(1). Generate $\boldsymbol{k}^{(t)} \sim f(\boldsymbol{k}|\theta^{(t-1)},\lambda_1^{(t-1)},\lambda_0^{(t-1)},y)$. Theorem 2.1 allows to generate $k^{(t)} \sim Bi\left(y, \frac{\theta^{(t-1)}\lambda_1^{(t-1)}}{\omega^{(t-1)}}\right)$; $k''^{(t)} \sim Bi\left(n-y, \frac{\theta^{(t-1)}(1-\lambda_1^{(t-1)})}{1-\omega^{(t-1)}}\right)$. Then $\boldsymbol{k}^{(t)} = (k^{(t)}, y-k^{(t)}, k''^{(t)}, n-y-k''^{(t)})'$.

(2). Generate $x^{(t)} \sim f(x|\theta^{(t-1)},\lambda_1^{(t-1)},\lambda_0^{(t-1)},y)$. Applying Corollary 2.2, take $x^{(t)} = k^{(t)} + k''^{(t)}$.

(3). Generate $\theta^{(t)}$, $\lambda_1^{(t)}$, $\lambda_0^{(t)}$ from their densities conditioned by $\boldsymbol{\pi}$, $\boldsymbol{k}^{(t)}$ and y. By using Lemma 2.1 and Theorem 2.2:

$$\theta^{(t)} \sim f(\theta|\lambda_1^{(t-1)},\lambda_0^{(t-1)},\boldsymbol{k}^{(t)},y,\boldsymbol{\pi}) = f(\theta|x^{(t)},\boldsymbol{\rho})$$

$$\lambda_1^{(t)} \sim f(\lambda_1|\theta^{(t-1)},\lambda_0^{(t-1)},\boldsymbol{k}^{(t)},y,\boldsymbol{\pi}) = f(\lambda_1|\boldsymbol{k}^{(t)},\boldsymbol{\pi}_1)$$

$$\lambda_0^{(t)} \sim f(\lambda_0|\theta^{(t-1)},\lambda_1^{(t-1)},\boldsymbol{k}^{(t)},y,\boldsymbol{\pi}) = f(\lambda_0|\boldsymbol{k}^{(t)},\boldsymbol{\pi}_0)$$

For the particular case that the prior distributions are $f(\theta|\boldsymbol{\rho}) = Be(p,q)$, $f(\lambda_1|\boldsymbol{\pi}_1) = Be(\alpha_1,\beta_1)$ and $f(\lambda_0|\boldsymbol{\pi}_0) = Be(\alpha_0,\beta_0)$, we obtain $f(\theta|x,\boldsymbol{\rho}) = Be(p+x, q+n-x)$, $f(\lambda_1|\boldsymbol{k},\boldsymbol{\pi}_1) = Be(\alpha_1+k, \beta_1+k'')$ and $f(\lambda_0|\boldsymbol{k},\boldsymbol{\pi}_0) = Be(\alpha_0+y-k, \beta_0+n-y-k'')$. When non-standard distributions are used, several generation techniques are available.

3 Individual analysis

As defined in Section 2, vector $\boldsymbol{v}$ represents the particular classification of the individuals in the sample O_n. When this vector is known, inferences about the true position for one or more individuals can be obtained. In this case, we are concerned with the posterior predictive distribution $f(u_h|\boldsymbol{v},\boldsymbol{\pi})$. This is particularly interesting for diagnostic problems.

New definitions must be established to differentiate between the individual information and the information of the remainder of the sample. Let h be a particular individual.

Definition 3.1. *Let $\boldsymbol{z}$ $(\boldsymbol{r})$ define the position (classification) vector for the reduced sample $O_n - \{h\}$ and let z (r) denote the effective (observed) frequency.*

Note that the components of the position and classification vectors of O_n are respectively $(\boldsymbol{z}, u_h)$ and $(\boldsymbol{r}, v_h)$.

Definition 3.2. *Let $\boldsymbol{c}_{\boldsymbol{z},\boldsymbol{r}} = \sum_{h'\neq h} \boldsymbol{s}_{h'} = (c_{\boldsymbol{z},\boldsymbol{r}}, c'_{\boldsymbol{z},\boldsymbol{r}}, c''_{\boldsymbol{z},\boldsymbol{r}}, c'''_{\boldsymbol{z},\boldsymbol{r}})'$ denote the latent vector induced by $(\boldsymbol{z},\boldsymbol{r})$.*

Proposition 3.1. *The following statements hold:*
$\boldsymbol{k}_{\boldsymbol{u},\boldsymbol{v}} = \boldsymbol{c}_{\boldsymbol{z},\boldsymbol{r}} + \boldsymbol{s}_h$; $c_{\boldsymbol{z},\boldsymbol{r}} + c''_{\boldsymbol{z},\boldsymbol{r}} = z$; $c_{\boldsymbol{z},\boldsymbol{r}} + c'_{\boldsymbol{z},\boldsymbol{r}} = r$; $x = z + u_h$; $y = r + v_h$.

The random variables and vectors related to the previous definitions are denoted by R, Z, and $\boldsymbol{C} = (C, C', C'', C''')$. The following distributions of the random variables U_h, V_h, Z, and R are deduced from the conditional independence conditions exposed in Section 2, i.e.:

$$f(u_h, z|\theta, \lambda_1, \lambda_0) = f(u_h|\theta) f(z|\theta)$$
$$f(v_h|u_h, \theta, \lambda_1, \lambda_0) = f(v_h|u_h, \lambda_1, \lambda_0)$$
$$f(r|z, \lambda_1, \lambda_0) = \textstyle\sum_{c=0}^{r} f(c|z, \lambda_1) f(r-c|z, \lambda_0)$$
$$f(v_h, r|u_h, z, \theta, \lambda_1, \lambda_0) = f(v_h|u_h, \lambda_1, \lambda_0) f(r|z, \lambda_1, \lambda_0)$$
$$f(v_h, r, u_h, z|\theta, \lambda_1, \lambda_0) = f(v_h, r|u_h, z, \theta, \lambda_1, \lambda_0) f(u_h, z|\theta)$$

Theorem 3.1. *The posterior predictive distribution $f(u_h|\boldsymbol{v},\boldsymbol{\pi})$ only depends on v_h and r, i.e.: $f(u_h|\boldsymbol{v},\boldsymbol{\pi}) = f(u_h|v_h, r, \boldsymbol{\pi})$.*

The importance of Theorem 3.1 lies in the fact that we only need to know the classification of the individual h and the number of positive classifications. So, we do not need to know the classification for the remainder of the individuals in the sample.

Theorem 3.2 provides a Gibbs sampling-based algorithm to generate from the distribution of interest.

Theorem 3.2. *The random vectors $\boldsymbol{C}$ and $\boldsymbol{S}_h$ are conditionally independent given θ, λ_1, λ_0, and their conditional distributions are respectively:*

$$f(\boldsymbol{c}|\theta, \lambda_1, \lambda_0, r) =$$
$$= \begin{cases} f(c|\theta, \lambda_1, \lambda_0, r) f(c''|\theta, \lambda_1, \lambda_0, r) & \text{if } c \leq r,\ c'' \leq n-1-r \\ 0 & \text{otherwise} \end{cases}$$

$$f(\boldsymbol{s}_h|\theta,\lambda_1,\lambda_0,v_h) =$$
$$= \begin{cases} f(s_h|\theta,\lambda_1,\lambda_0,v_h)f(s''_h|\theta,\lambda_1,\lambda_0,v_h) & \textit{if } s_h \le v_h,\ s''_h \le 1-v_h \\ 0 & \textit{otherwise} \end{cases}$$

where $f(c|\theta,\lambda_1,\lambda_0,r) \sim Bi\left(r, \frac{\theta\lambda_1}{\omega}\right)$ $(\omega = \theta\lambda_1 + (1-\theta)\lambda_0)$

$$f(c''|\theta,\lambda_1,\lambda_0,r) \sim Bi\left(n-1-r, \frac{\theta(1-\lambda_1)}{1-\omega}\right)$$

$$f(s_h|\theta,\lambda_1,\lambda_0,v_h) = \begin{cases} Ber\left(\frac{\theta\lambda_1}{\omega}\right) & \textit{if } v_h = 1 \\ 0 & \textit{if } v_h = 0 \end{cases}$$

$$f(s''_h|\theta,\lambda_1,\lambda_0,v_h) = \begin{cases} Ber\left(\frac{\theta(1-\lambda_1)}{1-\omega}\right) & \textit{if } v_h = 0 \\ 0 & \textit{if } v_h = 1 \end{cases}$$

Algorithm 3.1. *Initial values $\theta^{(0)}$, $\lambda_1^{(0)}$, and $\lambda_0^{(0)}$ are fixed. For each iteration t, steps 1-4 must be followed.*

- *(1). Generate $\boldsymbol{c}^{(t)} \sim f(\boldsymbol{c}|\theta^{(t-1)},\lambda_1^{(t-1)},\lambda_0^{(t-1)},r)$ by using:*
  $c^{(t)} \sim f(c|\theta^{(t-1)},\lambda_1^{(t-1)},\lambda_0^{(t-1)},r)$,
  $c''^{(t)} \sim f(c''|\theta^{(t-1)},\lambda_1^{(t-1)},\lambda_0^{(t-1)},r)$.
  Then $\boldsymbol{c}^{(t)} = (c^{(t)}, r-c^{(t)}, c''^{(t)}, n-1-r-c''^{(t)})'$.
- *(2). Generate $\boldsymbol{s}_h^{(t)} \sim f(\boldsymbol{s}_h|\theta^{(t-1)},\lambda_1^{(t-1)},\lambda_0^{(t-1)},v_h)$ by using:*
  $s_h^{(t)} \sim f(s_h|\theta^{(t-1)},\lambda_1^{(t-1)},\lambda_0^{(t-1)},v_h)$,
  $s_h''^{(t)} \sim f(s''_h|\theta^{(t-1)},\lambda_1^{(t-1)},\lambda_0^{(t-1)},v_h)$.
  Then $\boldsymbol{s}_h^{(t)} = (s_h^{(t)}, v_h - s_h^{(t)}, s_h''^{(t)}, 1-v_h-s_h''^{(t)})'$.
- *(3). Generate from U_h, $\boldsymbol{K}$ and X by using the values obtained at step 2, i.e.: $u_h^{(t)} = s_h^{(t)} + s_h''^{(t)}$; $\boldsymbol{k}^{(t)} = \boldsymbol{c}^{(t)} + \boldsymbol{s}_h^{(t)}$; $x^{(t)} = k^{(t)} + k''^{(t)}$*
- *(4). Generate $\theta^{(t)}$, $\lambda_1^{(t)}$ and $\lambda_0^{(t)}$ in the same way as at step 3 in Algorithm 3.1 by using $\boldsymbol{k}^{(t)}$ and $y = r + v_h$.*

Theorem 3.2 must be used for steps 1 and 2.

4 Illustrative example

The range of applications for binomial sampling with imperfect data is very wide. A particular case is the diagnostic problem. The confirmation of many chronic diseases is a complex and expensive task. It is often necessary to propose screening tests that can be rapidly and economically applied to many people. However, these tests are not error-free. By using the notation introduced in Section 2, we have:

		Disease (Position)		
		+	−	Total
Test	+	True + (k)	False + (k')	$k + k' = y$
(Classification)	−	False − (k'')	True − (k''')	$k'' + k'''$
	Total	$k + k'' = x$	$k' + k'''$	n

Table 1: General representation of a diagnostic test.

For every diagnostic test there are two critical values that determine its accuracy: sensitivity and specificity. Sensitivity is the probability that a test turns out to be positive, given that the person has the disease, i.e, λ_1, while specificity is the probability that a test turns out to be negative, given that the person does not have the disease, i.e, $1-\lambda_0$. A valid test should have both a high sensitivity and specificity. Positive and negative predictive values are also very important in this context. For more information about diagnostic tests, see Altman [1].

We present an illustrative example about AIDS diagnostic. A random sample of 192,415 individuals is selected. Twenty individuals are classified as disease-affected by screening their blood samples. If we know that a particular individual has been classified as disease-affected or non disease-affected, we are concerned with the true state of the individual, i.e. if he is affected by AIDS or not. Also, we are interested in studying the posterior distributions for θ, λ_1 and λ_0. The prior densities proposed by Johnson and Gastwirth [5] are used. These are $\theta \sim Be(15, 94092)$, $\lambda_1 \sim Be(142, 1)$ and $\lambda_0 \sim Be(3, 1363)$.

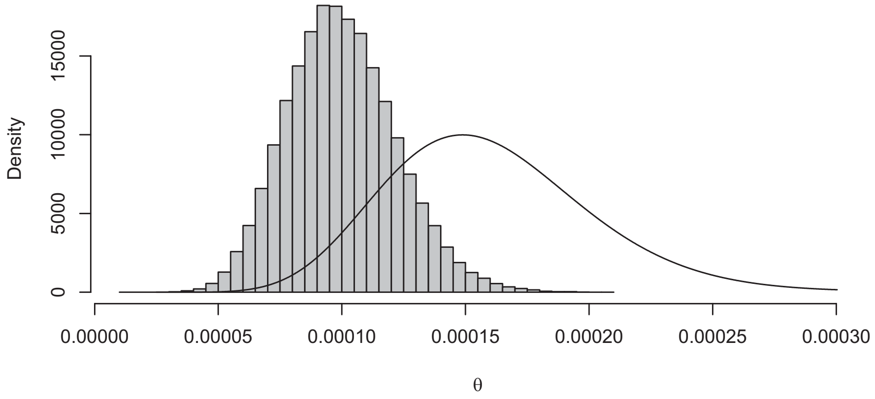


Figure 1: Prior (solid line) and posterior (histogram) densities for θ.

The individual I has been classified as non disease-affected. We now apply Algorithm 3.1. After we consider that the chain has converged, a sample of size 100,000 is generated from the predictive distribution $f(u_I|v_I = 0, r = 20, \boldsymbol{\pi})$, then the estimation of $P[U_I = 1|V_I = 0, r = 20, \boldsymbol{\pi}]$ is 0. We proceed similarly for the individual J that has been classified as disease-affected. In this case, an estimation of $P[U_J = 1|V_J = 1, r = 19, \boldsymbol{\pi}]$ is 0.6760 (0.0015). Monte Carlo standard error estimates are represented in brackets. The estimated posterior means for θ, λ_1, and λ_0 are respectively $9.9950 \cdot 10^{-5}$ $(6.9444 \cdot 10^{-8})$, 0.9928 $(2.2816 \cdot 10^{-5})$ and $4.9045 \cdot 10^{-5}(7.3045 \cdot 10^{-8})$. Finally, the prior density and the histogram for the posterior density of θ are represented in Figure 1.

5 Conclusion

We have developed a Bayesian model that involves the proportion of a Bernoulli process as well as the noise parameters. The introduction of auxiliary or latent variables allows us to make not only inferences for these parameters but also inferences for the real status of a particular individual. The introduction of auxiliary variables make the generation process easier thus allowing to implement Gibbs sampling-based algorithms.

References

[1] Altman, D.G. (1991). *Practical statistics for medical research.* Chapman and Hall.

[2] Gaba A. (1993). *Inferences with an unknown noise level in a Bernoulli process.* Management Science **39**, 1227 – 1237.

[3] Gaba A., Winkler R.L. (1992). *Implications of errors in survey data: a Bayesian approach.* Management Science **38**, 913 – 925.

[4] Girón F.J., Kadane J.B., Moreno E. (1997). *Independence issues in imprecise data models: a Bayesian approach.* C. R. Acad. Sci. Paris **324**, 1149 – 1153.

[5] Johnson W.O. (1991). *Bayesian inference for medical screening tests: approximations useful for the analysis of acquired immune deficiency syndrome.* J. R. Statist. Soc. **53**, 427 – 439.

[6] Winkler R.L. (1985). *Information loss in noisy and dependent processes.* Management Science In J.M. Bernardo et al. (Eds.) Bayesian Statistics 2, North Holland, Amsterdam, 559 – 570.

[7] Winkler R.L., Gaba A. (1990). *Inference with imperfect sampling from a Bernoulli process.* In S. Geisser et al. (Eds.) Bayesian and Likelihood Methods in Statistics and Econometrics: Essay in Honor of George A. Barnad, North Holland, Amsterdam, 303 – 317.

Address: M. Ruiz, F.J. Girón, C. Rojano, Departamento de Estadística e I.O.
Facultad de Ciencias, Universidad de Málaga, 29071 Málaga, Spain
C. Pérez, J. Martín, Departamento de Matemáticas, Facultad de Veterinaria,
Universidad de Extremadura, 10071 Cáceres, Spain

E-mail: {mruiz,fj_giron,rojano}@uma.es,{carper,jrmartin}@unex.es

A TWO-SYSTEM GOVERNED BY PH-DISTRIBUTIONS WITH MEMORY OF THE FAILURE PHASE

Juan Eloy Ruiz-Castro, Rafael Pérez-Ocón, Delia Montoro-Cazorla and Gemma Fernández-Villodres

Key words: Phase-type distributions, Markov process, two-system, algorithms, Matlab.

COMPSTAT 2004 section: Reliability.

Abstract: A two-unit system in a dynamic environment is considered, with the following assumptions: a) the operational and repair times follow phase-type distributions, b) operational times are affected by a factor that modified the following period, c) the system undergoes operational accidental failures in addition to wear-out ones, and d) the system remembers the phase of failure, and when it is repaired returns to the operational phase in which the failure occurred. A general Markov process with vectorial states is an appropiate structure for modelling this system. This system extends a previous work for a unit-system [2]. The transition probabilities and the stationary distribution are calculated, obtaining mathematical expressions in a well structured form using the Kronecker algorithm. Some performance measures of general interest in the study of systems are determined using an algorithmic approach, such us, the availability, and the rate of occurrence of failures. The results are used for studying a parallel system and a numerical example is analyzed, implementing computationally the formulae obtained throughout the paper in Matlab. The mathematical expressions are given by algorithmic methods, which highlight the utility of phase-type distributions in the analysis of lifetime data.

1 Introduction

In the study of dynamic systems modeled by Markov processes, the steady-state is frequently considered because of calculation simplicity. The transient behaviour is not so broadly studied, possibly due to the untractable expressions appearing in the calculations. When only exponential times are involved, some authors have considered the transient regime for studying particular problems. In general, the study of these topics is difficult and it is necessary to rely on computational calculations, and sometimes the expressions of the quantities of interest are not possible to determine. In this line, Neuts and Meier [1] considered a general Markov model and constructed an infinitesimal generator highly structured and formulated efficient algorithms to evaluate quantities of interest. More recently, Schouten et al. [5] studied a two-system with Markovian degrading units and two types of repairing, deriving explicit expressions for the Laplace transforms of the up and down periods of the system. Neuts et al. [2] considered a general Markov process for modelling a reliability system in stationary regime; some performance measures were calculated. In this work the unit of the system remembers the failure phase when the unit completed its repair, and it begins operating in that phase. Following the last line of work, we extend this model by considering a two-unit independent system and the transient and stationary analysis is carried out. In a certain way, our study also expands that of Schouten et al. [5], because the Markovian degrading in the units can be included when the lifetime of these are phase-type distributed. In the present paper, external and internal failures are considered. The former can be repairable or non-repairable, and the latter non-repairable. The successive operational times form a geometric process. For this system, the transition probabilities are explicitly calculated by using an algorithmic approach that involves the Kronecker product, which is very useful in the context in which we have used it and allows us to extend the results from a unit-system to a two-system in an natural way. Moreover, the performance measures calculated in the previous paper of Neuts et al. [2] in the steady-state and some other are now calculated in a transient state. We applied the model to a parallel system, the rate of occurrence of failures and other reliability measures are calculated for the general Markov model that governs the system. For the computational implementation of the expressions the MATLAB programme was used, and these are available from the authors.

2 The model

We consider a system having two independent units. Each unit is submitted to accidental and wear-out failures. A Poisson process with rate λ governs accidental failures, and they are repairable with probability p and non-repairable with probability $q = 1 - p$. Wear-out failures are non-repairable. All failures are independent and if a non-repairable failure occurs, the unit is replaced by a new and identical one. Each unit has its repairman. Also, each unit has an operational time governed by a phase-type distribution, and these successive random times form a geometric process. Each unit is replaced after a prefixed number of failures, N. All times are independent random variables. It will be assumed that initially the units are new. For the sake of simplicity, we will assume that each unit can be repaired only once, $N = 1$, so they are replaced when the second repairable failure or a non-repairable failure arrives. The probability distribution of the times to failure after repair $n(n = 0, 1)$ for the unit k is a phase-type distribution with representation $[\alpha(k), a_k^n \mathbf{T}(k)]_{m_k}$ with $a_k \geq 1$ and $k = 1, 2$. Analogously, the probability distribution of the repair times for the unit k has representation $[\beta(k), \mathbf{S}(k)]_{n_k}$ with $k = 1, 2$. The states in each unit are classified in two groups: operational and in repair. These states are macro-states. So, there are m_k operating phases and

n_k repair ones for the unit k. The macro-states of the system are defined as a couple of the macro-states defined in the unit system [2]. E will denote the state space of the two-unit system. We introduce the following notation $(k = 1, 2)$: wk: unit k is new and it is operational, wrk: unit k is operational and it have been repaired, and rk: unit k is in repair. The macro-states are classified depending on the operational or in repair stage of the units. Four classes are distinguished: 1) The two units are operational, $E_0 = \{E_{01} = (w1, w2), E_{02} = (w1, wr2), E_{03} = (wr1, w2), E_{04} = (wr1, wr2)\}$.2) Only unit 1 is operational, $E1 = \{E_{11} = (w1, r2), E_{12} = (wr1, r2)\}$. 3) Only unit 2 is operational, $E_2 = \{E_{21} = (r1, w2), E_{22} = (r1, wr2)\}$. 4) Both units are in repair, $E_3 = \{(r1, r2)\}$ The macro-states of E_0 are of order $1 \times m_1m_2$, the ones of E_1 are of order $1 \times m_1n_2m_2$, the ones of E_2 are of order $1 \times m_1n_1m_2$, and the ones of E_3 are of order $1 \times m_1n_1m_2n_2$. For instance, the state (i, j, k, l) of the macro-state E_3 indicates that the first unit is in repair phase j and it failed in the operating phase i, and the unit 2 is in repair phase l and it failed in the operating phase k. The generator matrix will be constructed using the macro-states. For illustrating the building of the generator we show the following cases. When both units are operating without repair and a repairable failure occurs in the second unit, there is a transition between the macro-states $E_{01} = (w1, w2)$ to $E_{11} = (w1, r2)$. In this transition the second unit from any functioning phase goes to the repair state according to the initial vector β_2. The possible transitions between phases in this case are $(i, j) \to (i, j, r)$ with $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$ and $r = 1, \ldots, n_2$, being i the operating phase of the unit one, j the operating phase of the unit two when this one failured and r the phase of repair. So, the transition rate matrix is given by $\mathbf{I}_{m_1} \otimes (\lambda p \mathbf{I}_{m_2} \otimes \beta_2)$, being $\mathbf{I}_{m_1}$ and $\mathbf{I}_{m_2}$ the identity matrix of m_1 and m_2 order, respectively. This is the block $(1, 1)$ of the matrix $\mathbf{Q}_{01}$. On the other hand, by considering that the second unit is in repair, if the first unit is operating after repair and a wear-out or an accidental (repairable or non-repairable) failure presents itself, there is a transition from the macro-state $E_{12} = (wr1, r2)$ to $E_{11} = (w1, r2)$. The possible transitions between phases are $(i, j, r) \to (s, j, r)$ for $i, s = 1, \ldots, m_1$, $j = 1, \ldots, m_2$ and $r = 1, \ldots, n_2$, where i and s are the operating phases of the unit one before and after the failure respectively, and j the operating state of the unit two when this one failed and r the repair phase. The transition rate matrix is given by $(a_1\mathbf{T}_1^0\alpha_1 + \lambda\mathbf{e}\alpha_1) \otimes \mathbf{I}_{m_2n_2}$, where $\mathbf{e}$ is a column vector of 1's with appropriate dimension. It is the block $(2, 1)$ of the matrix $\mathbf{Q}_{11}$. Following this reasoning the generator $\mathbf{Q}$ can be expressed in terms of the blocks that determine the macro-states E_i. It can be seen that the final form is:

$$\mathbf{Q} = \begin{array}{c} \\ E_0 \\ E_1 \\ E_2 \\ E_3 \end{array} \begin{array}{c} \begin{array}{cccc} E_0 & E_1 & E_2 & E_3 \end{array} \\ \begin{pmatrix} \mathbf{Q}_{00} & \mathbf{Q}_{01} & \mathbf{Q}_{02} & \mathbf{0} \\ \mathbf{Q}_{10} & \mathbf{Q}_{11} & \mathbf{0} & \mathbf{Q}_{13} \\ \mathbf{Q}_{20} & \mathbf{0} & \mathbf{Q}_{22} & \mathbf{Q}_{23} \\ \mathbf{0} & \mathbf{Q}_{31} & \mathbf{Q}_{32} & \mathbf{Q}_{33} \end{pmatrix} \end{array}.$$

These blocks are,

$$\mathbf{Q}_{01} = \begin{pmatrix} \mathbf{I}_{m_1} \otimes (\lambda p \mathbf{I}_{m_2} \otimes \beta_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m_1} \otimes (\lambda p \mathbf{I}_{m_2} \otimes \beta_2) \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

$$\mathbf{Q}_{02} = \begin{pmatrix} (\lambda p \mathbf{I}_{m_1} \otimes \beta_1) \otimes \mathbf{I}_{m_2} & \mathbf{0} \\ \mathbf{0} & (\lambda p \mathbf{I}_{m_1} \otimes \beta_1) \otimes \mathbf{I}_{m_2} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

$$\mathbf{Q}_{10} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_{m_1} \otimes (\mathbf{I}_{m_2} \otimes \mathbf{S}_2^0) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{m_1} \otimes (\mathbf{I}_{m_2} \otimes \mathbf{S}_2^0) \end{pmatrix},$$

$$\mathbf{Q}_{11} = \begin{pmatrix} \mathbf{A}_1 \oplus (\mathbf{I}_{m_2} \otimes \mathbf{S}_2) - \lambda \mathbf{I}_{m_1 m_2 n_2} & \mathbf{0} \\ (a_1 \mathbf{T}_1^0 \alpha_1 + \lambda \mathbf{e} \alpha_1) \otimes \mathbf{I}_{m_2 n_2} & a_1 \mathbf{T}_1 \oplus (\mathbf{I}_{m_2} \otimes \mathbf{S}_2) - \lambda \mathbf{I}_{m_1 m_2 n_2} \end{pmatrix},$$

$$\mathbf{Q}_{13} = \begin{pmatrix} (\lambda p \mathbf{I}_{m_1} \otimes \beta_1) \otimes \mathbf{I}_{m_2 n_2} \\ \mathbf{0} \end{pmatrix},$$

$$\mathbf{Q}_{20} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & (\mathbf{I}_{m_1} \otimes \mathbf{S}_1^0) \otimes \mathbf{I}_{m_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (\mathbf{I}_{m_1} \otimes \mathbf{S}_1^0) \otimes \mathbf{I}_{m_2} \end{pmatrix},$$

$$\mathbf{Q}_{22} = \begin{pmatrix} (\mathbf{I}_{m_1} \otimes \mathbf{S}_1) \oplus \mathbf{A}_2 - \lambda \mathbf{I}_{m_1 m_2 n_1} & \mathbf{0} \\ \mathbf{I}_{m_1 n_1} \otimes (a_2 \mathbf{T}_2^0 \alpha_2 + \lambda \mathbf{e} \alpha_2) & (\mathbf{I}_{m_1} \otimes \mathbf{S}_1) \oplus a_2 \mathbf{T}_2 - \lambda \mathbf{I}_{m_1 m_2 n_1} \end{pmatrix},$$

$$\mathbf{Q}_{23} = \begin{pmatrix} \mathbf{I}_{m_1 n_1} \otimes (\lambda p \mathbf{I}_{m_2} \otimes \beta_2) \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{Q}_{31} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_{m_1} \otimes \mathbf{S}_1^0 \otimes \mathbf{I}_{m_2 n_2} \end{pmatrix},$$

$$\mathbf{Q}_{32} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_{m_1 n_1} \otimes (\mathbf{I}_{m_2} \otimes \mathbf{S}_2^0) \end{pmatrix}, \quad \mathbf{Q}_{33} = (\mathbf{I}_{m_1} \otimes \mathbf{S}_1) \oplus (\mathbf{I}_{m_2} \otimes \mathbf{S}_2),$$

and the block $\mathbf{Q}_{00}$ is given by

$$\left(\begin{matrix} \mathbf{A}_1 \oplus \mathbf{A}_2 - 2\lambda \mathbf{I}_{m_1 m_2} & \mathbf{0} \\ \mathbf{I}_{m_1} \otimes (a_2 \mathbf{T}_2^0 \alpha_2 + \lambda \mathbf{e} \alpha 2) & \mathbf{A}_2 \oplus a_2 \mathbf{T}_2 - 2\lambda \mathbf{I}_{m_1 m_2} \\ (a_1 \mathbf{T}_1^0 \alpha_1 + \lambda \mathbf{e} \alpha 1) \otimes \mathbf{I}_{m_2} & \mathbf{0} \\ \mathbf{0} & (a_1 \mathbf{T}_1^0 \alpha_1 + \lambda \mathbf{e} \alpha 1) \otimes \mathbf{I}_{m_2} \end{matrix} \right.$$

$$\left. \begin{matrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ a_1 \mathbf{T}_1 \oplus \mathbf{A}_2 - 2\lambda \mathbf{I}_{m_1 m_2} & \mathbf{0} \\ \mathbf{I}_{m_1} \otimes (a_2 \mathbf{T}_2^0 \alpha_2 + \lambda \mathbf{e} \alpha 2) & (a_1 \mathbf{T}_1 \oplus a_2 \mathbf{T}_2) - 2\lambda \mathbf{I}_{m_1 m_2} \end{matrix} \right),$$

with $\mathbf{A}_1 = \mathbf{T}_1 + \mathbf{T}_1^0 \alpha_1 + \lambda q \mathbf{e} \alpha_1$ and being $\mathbf{I}_k$ the identity matrix of order k and $\mathbf{e}$ a column vector of 1's with appropiate dimension. Its initial probability vector depends on the initial conditions. If both units are new, this vector is $[\alpha(1) \otimes \alpha(2), \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}]$.

3 The units

3.1 Stationary distribution for each unit

The stationary distribution for each unit has been obtained in Neuts et al. [2]. Here we introduce the notation that will be used in advance. The stationary distribution for each unit k (k = 1, 2) is denoted by $\pi_k = [\pi_k(\mathbf{0}), \pi_k(\mathbf{1_R}), \pi_k(\mathbf{1})]$, where $\mathbf{0}$ denotes the macro-state when the unit k is operating initially, $\mathbf{1_k}$ when it is in repair and $\mathbf{1}$ when it is operating after repair.

3.2 Transition probabilities

Now let us consider a unit and eliminate the subindex k for simplicity. When the unit is in the macro-state $\mathbf{i}$, it can be in some of the m operational phases, and when it is in the macro-state $\mathbf{i_R}$ it can occupy some of the $m \times n$ phases (the n repair phases, and the m operational ones where the unit failed). If the unit is in macro-state $\mathbf{i}$, the state of the unit at time t is given by a couple $(\mathbf{i}, y)$ where y indicates the operational phase occupied by the unit at time t. On the other hand, if the unit is in macro-state $\mathbf{i_R}$, the state of the unit at time t is given by a couple $(\mathbf{i_R}, x, y)$ where y indicates the repair phase at time t and x the operational phase occupied by the unit when it failed. Let us assume that the unit begins in state 0. We use $P_v(\mathbf{i}, j, t)$ to denote the probability that the unit will be in the operational state $(\mathbf{i}, j)$ at time t, given that it initially occupies the state $(\mathbf{0}, v)$, for $j, v = 1, \ldots, m$; $i = 0, \ldots, N$ and $t \geq 0$. Similarly, we define $P_v(\mathbf{i_R}, j', k, t)$ as the probability that the unit will be in the repair state $(\mathbf{i_R}, j', k)$, given that it initially occupies the state $(\mathbf{0}, v)$, for $j' = 1, \ldots, m$; $k = 1, \ldots, n$; $\mathbf{i_R} = \mathbf{1_R}, \ldots, \mathbf{N_R}; v = 1, \ldots, m$ and $t \geq 0$. The transition probabilities among the macro-state $\mathbf{0}$ and the operational macro-state $\mathbf{i}$ or the repairing one $\mathbf{i_R}$ are given, respectively, by the matrices

$$\mathbf{P_i}(t) = (P_v(\mathbf{i}, j, t))_{v,j=1,\ldots,m}$$

$$\mathbf{P_{i_R}}(t) = (P_v(\mathbf{i_R}, j', k, t))_{v=1,\ldots,m;(j',k)=\{(r,s);r=1,\ldots,m,s=1,\ldots,n\}}.$$

Eliminating the subindices for simplicity, we write the matricial Kolmogorov forward equations as $\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}$, being the generator for an unit, and $\mathbf{P}(t)$ the transition probability matrix. These equations can be solved by taking Laplace transfom and solving the corresponding algebraic system. The calculations are available from the authors.

4 Two-unit system

4.1 Transition probabilities

When the two-system is considered, the corresponding forward Kolmogorov equations supply a great number of differential equations. We indicate the transition probability functions, given that initially the system occupies the macro-state E_{01}. For simplicity and because the two units are new, let us study the transition probabilities between the macro-state E_{01} and the others. Let $\mathbf{P}^k_{\mathbf{j}}(t)$ denote the probability vector that the unit k will occupy the macro-state $\mathbf{j}$ at time t, given that initially the unit is new; that is, it is in the macro-state $\mathbf{0}$, with $\mathbf{j} = \mathbf{0}, \mathbf{1_R}, \mathbf{1}$; $k = 1, 2$, and $t \geq 0$. Let $\mathbf{P}^S_i(t)$ denote the probability that the system occupies the macro-state i at time t given that initially it is in the macro-state E_{01}, with i in E, $t \geq 0$. These probabilities have been calculated by applying the Kronecker algorithm, considering the independence of the units. They can be expressed in the following way:

$$\begin{array}{ll}
\mathbf{P}^S_{E_{01}}(t) = \mathbf{P}^1_{\mathbf{0}}(t) \otimes \mathbf{P}^2_{\mathbf{0}}(t) & \mathbf{P}^S_{E_{02}}(t) = \mathbf{P}^1_{\mathbf{0}}(t) \otimes \mathbf{P}^2_{\mathbf{1}}(t) \\
\mathbf{P}^S_{E_{04}}(t) = \mathbf{P}^1_{\mathbf{1}}(t) \otimes \mathbf{P}^2_{\mathbf{1}}(t) & \mathbf{P}^S_{E_{11}}(t) = \mathbf{P}^1_{\mathbf{0}}(t) \otimes \mathbf{P}^2_{\mathbf{1_R}}(t) \\
\mathbf{P}^S_{E_{21}}(t) = \mathbf{P}^1_{\mathbf{1_R}}(t) \otimes \mathbf{P}^2_{\mathbf{0}}(t) & \mathbf{P}^S_{E_{22}}(t) = \mathbf{P}^1_{\mathbf{1_R}}(t) \otimes \mathbf{P}^2_{\mathbf{1}}(t)
\end{array}$$

$$\begin{array}{l}
\mathbf{P}^S_{E_{03}}(t) \ = \mathbf{P}^1_{\mathbf{1}}(t) \otimes \mathbf{P}^2_{\mathbf{0}}(t) \\
\mathbf{P}^S_{E_{12}}(t) = \mathbf{P}^1_{\mathbf{1}}(t) \otimes \mathbf{P}^2_{\mathbf{1_R}}(t) \\
\mathbf{P}^S_{E_3}(t) \ = \mathbf{P}^1_{\mathbf{1_R}}(t) \otimes \mathbf{P}^2_{\mathbf{1_R}}(t).
\end{array}$$

4.2 Stationary distribution

We denoted by π the stationary vector. It is useful to represent it according to the blocks of the matrix $\mathbf{Q}$. Considering the order described in 2 the vector π is $\pi = [\pi_{01}, \pi_{02}, \pi_{03}, \pi_{04}; \pi_{11}, \pi_{12}; \pi_{21}, \pi_{22}; \pi_{31}]$, where π_{01} is the vector of order $1 \times m1m2$ corresponding to $(w1, w2)$, and the same for $\pi_{02}, \pi_{03}, \pi_{04}$ and $(w1, wr2), (wr1, w2), (wr1, wr2)$, ordering properly the phases. The rest of subvectors of π built in the same way. The product $\pi\mathbf{Q}$ is done using the blocks in the vector and in the matrix. Making such product equal to zero, a matricial system with 9 equations is obtained. Instead of solving the system numerically, we use the independence of the units. Analogously to the transient case the Kronecker algorithm is applied. We illustrate this for the block π_{01}. In this case $\pi_{01} = \pi_1(\mathbf{0}) \otimes \pi_2(\mathbf{0})$.

5 Parallel system

A particular and usesul system is the parallel one. In this section is considered the transient and stationary behaviour. Some performance measures of interest are calculated. We show some of them, others are available from the authors.

Operational time

The random variable that denotes the lifetime of the system (time up to failure), T_S, follows a phase-type distribution being the absorbent state E_3. The representation for this time is denoted by $(\gamma_S, \mathbf{L}_S)$ being

$$\gamma_S = [\alpha(1) \otimes \alpha(2), \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}] \text{ and } \mathbf{L}_S = \begin{pmatrix} \mathbf{Q}_{00} & \mathbf{Q}_{01} & \mathbf{Q}_{02} \\ \mathbf{Q}_{10} & \mathbf{Q}_{11} & \mathbf{0} \\ \mathbf{Q}_{20} & \mathbf{0} & \mathbf{Q}_{22} \end{pmatrix}.$$

Availability
The availability is the probability that the system does not occupy the down macro-state at time t, thus,

$$\begin{aligned} A_S(t) &= 1-[\alpha(1)\otimes\alpha(2)]\mathbf{P}^S_{E3}(t)\mathbf{e} = 1-[1-[\alpha(1)\mathbf{P}^1_{\mathbf{1_R}}(t)\mathbf{e}\otimes\alpha(2)\mathbf{P}^2_{\mathbf{1_R}}(t)\mathbf{e}] \\ &= 1-\overline{A}_1(t)\overline{A}_2(t), \end{aligned}$$

$\overline{A}_k(t)$ being the probability that at time t the unit k is not operational, $k = 1, 2$. The stationary availability is $A = 1 - \pi_{31}\mathbf{e}$.
Reliability
The reliability is the probability that the system will be continuosly operational up to time t, given that initially the system is operational. This function can be expressed as,

$$R_S(t) = \gamma_S exp(\mathbf{L}_S t)\mathbf{e} = R_1(t)R_2(t),$$

being $R_k(t)$ the reliability of the unit k, with $k = 1, 2$.
Rate of occurrence of failures (ROCOF)
The mean number of repairable failures per unit time for the units of the system at time t is given by

$$\begin{aligned} v_1(t) &= \lambda p(\alpha(1)\otimes\alpha(2))(2\mathbf{P}^S_{E_{01}}(t)+\mathbf{P}^S_{E_{02}}(t)+\mathbf{P}^S_{E_{03}}(t)+\mathbf{P}^S_{E_{11}}(t) \\ &+ \mathbf{P}^S_{E_{21}}(t))\mathbf{e}_{m_1m_2}. \end{aligned}$$

The mean number per unit time of repairable failures in the system is

$$v^S_1(t) = \lambda p(\alpha(1)\otimes\alpha(2))(\mathbf{P}^S_{E_{11}}(t)+\mathbf{P}^S_{E_{21}}(t))\mathbf{e}_{m1m2}.$$

The ROCOF has also been calculated for non-repairable and wear-out failures, the expressions obtained are available from the authors. A stationary study has also been carried out.

6 Numerical application

We consider a parallel system with the following numerical values for the parameters and matrices (the unit time will be the hour): $\lambda = 0.00187$ accidental failures per hour; $p = 0.87$ proportion of repairable accidental failures; $a_1 = 1.25$; $a_2 = 1.35$ rates of the geometric processes of operational and repair times; $\beta_1 = \beta_2 = \alpha1 = \alpha_2 = (1, 0, 0)$ initial vectors. The **S** matrices are relative to the repair times and **T** matrices to the operational times for every unit:

$$\mathbf{T}_1 = \begin{pmatrix} -0.0027 & 0.0027 & 0 \\ 0 & -0.008 & 0.008 \\ 0 & 0 & -0.02878 \end{pmatrix} \quad \mathbf{S}_1 = \begin{pmatrix} -0.02 & 0.02 & 0 \\ 0.01 & -0.08 & 0.07 \\ 0.005 & 0 & -0.1 \end{pmatrix}$$

$$\mathbf{T}_2 = \begin{pmatrix} -0.012 & 0.012 & 0 \\ 0 & -0.075 & 0.075 \\ 0 & 0 & -0.05 \end{pmatrix} \quad \mathbf{S}_2 = \begin{pmatrix} -0.1 & 0.1 & 0 \\ 0.03 & -0.05 & 0.02 \\ 0.02 & 0 & -0.04 \end{pmatrix}.$$

Considering the stationary case: $A = 0.9786$ and $v_1 = 3.5632e - 004$. Other measures are available from the authors.

References

[1] Neuts M., Meier K.S. (1981). *On the use of Phase Type distributions in reliability modelling od systems with two components.* OR Spektrum **2**, 227 – 234.

[2] Neuts M.F., Pérez-Ocón, R., Torres-Castro I. (2000). *Repairable models with operating and repair times governed by Phase Type distributions.* Advances in Applied Probability **32**, 468 – 479.

[3] Pérez-Ocón, R. and Montoro-Cazorla, D. (2004). *A multiple system governed by a quasi-birth-and-death-process.* Reliability Engineering and System Safety, (accepted).

[4] Pérez-Ocón R., Ruiz-Castro J. Eloy (2004). *Two models for a repairable two-system with phase-type sojourn time distributions.* Reliability Engineering and System Safety, (accepted).

[5] Van der Schouten and Wartenhorst, F.A. (1994) *Transient analysis of a two-unit standby system with Makovian degrading units.* Management Science **43(2)**, 335 – 343.

Acknowledgement: The authors would like to thank the referee for their constructive comments. The first and the second author gratefully acknowledge the financial support of the Ministerio de Ciencia y Tecnología, Proyecto BFM2001-3802, Spain.

Address: J. E. Ruiz-Castro, Dpto. Estadística e Investigación Operativa. Universidad de Granada. Facultad de Ciencias. Campus Fuentenueva, 18071 Granada, Spain

E-mail: `jeloy@ugr.es`

SOME APPROACHES TO OVERLAPPING CLUSTERING OF BINARY VARIABLES

Hana Řezanková, Dušan Húsek and Alexander A. Frolov

Key words: Boolean factorization, factor analysis, fuzzy cluster analysis.
COMPSTAT 2004 section: Clustering.

Abstract: This paper describes some possible solutions on how to use standard statistical procedures for finding overlapping binary factors. The application of factor analysis as a base for further investigation and interpretation of the factor loading matrix by means of fuzzy cluster analysis is suggested. Some experiments were performed with using statistical packages STATISTICA and S-PLUS.

1 Introduction

As the size of analyzed databases increases, methods for reduction of dimensionality have become more important. Factor analysis is a well known method for this purpose. It is based on investigation of correlations between variables. The input data matrix can be expressed by means of multiplication of two matrices. One contains factor scores and the second one contains factor loadings. In this paper we will focus on the modification of factor analysis in which both input variables and new (latent) variables are dichotomous (binary).

We can mention the analysis of the input matrix $\mathbf{X}$ which contains information about N textual documents and M terms. We want to express these terms by means of L factors. Let us consider the case in which the input matrix contains only values "one" (the term is contained in the document) and "zero" (the term is not contained in the document). This matrix can be decomposed into two matrices. One expresses the occurrence of L groups of terms in individual documents and the second one expresses which terms are contained in individual groups. Both new matrices also contain only the values "zero" and "one".

Regarding [4], we suppose that each ith case (row) $\mathbf{X}_i$ ($i = 1, 2, \ldots, N$) of the analyzed data matrix can be expressed as a logical sum of weighted vectors of factor loadings: $\mathbf{X}_i = \bigvee_{l=1}^{L} f_{il}\boldsymbol{\alpha}_l$, where f_{il} are factor scores, $\boldsymbol{\alpha}_l$ are vectors of factor loadings ($l = 1, 2, \ldots, L$) and L is a number of factors. Such a case is considered in which $\boldsymbol{\alpha}_l \in B_p^M$ [1] (pM is a maintained constant) and $\mathbf{f}_l \in B_{p_f}^L$. Here p and p_f are sparseness (a ratio of the number of active elements to the total number of elements) of vectors of factor loadings with respect to variables and cases with respect to vectors of factor scores.

[1] $B_p^M = \{\mathbf{X} \mid X_j \in \{0,1\}, P\{X_j = 1\} = p \ \ \forall j = 1, 2, \ldots, M\}$

We are interested in different techniques for finding vectors of factor loadings α_l from the data. Formerly, the special procedures for Boolean factor analysis exited in BMDP (its name was 8M). We tried to use it once but we did not obtain satisfactory results. The problem is, that this process leads to finding a local minimum of the error function but it does not have to be a global minimum.

For solving of this task, the Hopfield-like neural network has been developed (see [4]). However, no commercial product exists for this purpose. In this paper, we will focus on the possibilities of finding an approximate solution by means of standard procedures implemented in the statistical packages STATISTICA (6.0) and S-PLUS (4.5). There is the task of finding overlapping clusters of asymmetric binary variables. The task of determination of the optimal number of clusters has been described, and that is why we are interested only in assignment of variables to clusters with the known number of them.

Some multivariate methods are based on the proximity matrix in which similarities (or dissimilarities) for all existing pairs of variables are evaluated. For factor analysis, the correlation matrix is the base. It means that similarity is expressed by a correlation coefficient. For asymmetric binary variables special coefficients exist. They investigate a similarity on the base of equality of active elements.

2 Suggested solution

In analyzed data files some variables can be assigned to only one factor, and the others to two or more factors. Statistical techniques for overlapping clustering are described theoretically but procedures for solving tasks of this type are rarely implemented in commercial statistical packages. Fuzzy cluster analysis (in the S-PLUS system) could be used for this purpose but it makes only clustering cases possible and thus the input matrix needs to be transposed. However, if we apply this technique for simulated data files, the solution does not correspond to input setting on the basis of which the data were generated.

One possibility of how to find overlapping clusters of variables by methods in common use is the application of different techniques for clustering. This consists mainly of different agglomerative techniques of hierarchical cluster analysis or techniques of multidimensional scaling. The suitable similarity measures for asymmetric binary data have to be used, such as Jaccard or Czekanowski (Dice) coefficients. A general formula for this class of coefficients by which a similarity between variables X and Y can be expresses is

$$s(X,Y) = \frac{\Theta \sum\limits_{i=1}^{N} x_i y_i}{\Theta \sum\limits_{i=1}^{N} x_i y_i + \sum\limits_{i=1}^{N} |x_i - y_i|} \; .$$

If $\Theta = 1$, it is Jaccard coefficient, and if $\Theta = 2$, we obtain Czekanowski similarity measure (these measures are included for example in the SPSS system).

Variables which belong to only one cluster are mostly identified identically by different techniques. Variables which belong to two or more clusters can be assigned differently by different methods. From obtained results we can come to a conclusion to which clusters variables can belong (to all clusters in which they appeared). The solution is approximate as variables can be assigned to more clusters than it was done in setting. If no special software exists, it is a way how to obtain an image about data structure. The disadvantage of this investigation is its complexity. Applications of the process mentioned above are described in [8] and [9].

The other way how to obtain an approximate solution of the task defined above is the use of factor analysis as the base for the further analysis. We came to a conclusion that the suitable technique for revelation of factors is the interpretation of the factor loading matrix by fuzzy cluster analysis. In this manner we obtain the solution which gets near to setting for generating data for the analysis.

We applied this process to the simulated data (see below). At first the "variables" in the input matrix containing only zero values were omitted. So, for an example, the input data matrix has 100 variables but only 65 variables were further analyzed. We used the STATISTICA system for factor analysis because it is user-friendly. We generated correlation matrices in the special form required by STATISTICA.

STATISTICA requires specifying a maximum number of factors as an input parameter. As mentioned before, for simplification we suppose that the number of factors (L) has been determined. A problem can occur when we obtain less than L factors. It makes factor interpretation difficult. That is why we decided to interpret results by fuzzy cluster analysis in S-PLUS. For this purpose, it is important to determine a number of factors (K) for factor analysis. We suggest setting this number so that the following relations were satisfied:

$$K_{\min} = 2, \quad 2^{K_{\max}-1} < L \quad \text{and} \quad 2^{K_{\max}} \quad \geq L \ .$$

We also applied this process to the input data matrix characterizing a collection of textual documents. Moreover, we compared it with the use of multiple correspondence analysis (by using the STATISTICA system) and interpretation of coordinates by fuzzy cluster analysis in the same manner as in the previous case.

3 Description of experiments and results

We performed the following experiments. We simulated data files for different parameters N, M, L and p. Additionally, we reflected complexity C which is a number of vectors of factor loadings by which each row of the data matrix is created. For example, if $C = 1$ then each row is one of the vectors, if $C = 2$ then each row is created by a combination of two vectors and so on.

First, we will describe the analysis of one simulated data file for the illustration of solving the task specified above. The data file was simulated with the following parameters: $N = 2000, M = 100, L = 11, p = 0.1$. It means that 10 variables are assigned to one factor. After omitting variables containing only zero values, we got the reduced matrix with 65 variables ($M = 65$).

The easiest case is that each row of the input data matrix is a certain factor directly ($C = 1$). It could seem that we would be able to find factors by factor analysis. We specified $K = 11$ as an input parameter in STATISTICA but we obtained only 10 factors for the factor loading matrix. We used rotation Varimax and tried to interpret factor loadings.

As it has been noted, some variables are assigned to only one factor. Most of them can be determined according to factor loadings which are greater than 0.7 (the default setting in STATISTICA) but not all. Similarly, variables assigned to 2 or more factor could be determined. The question is as follows: which values could be considered “high” for assignment of a certain variable to a certain factor and which ones could not. Using fuzzy cluster analysis makes our decision easy.

Our suggestion on how to reveal factors easily is to apply factor analysis to reduced correlation matrix (see above) with a suitable input parameter K (in our example $K = 4$). We obtain the factor loading matrix M x K (in our example 65 x 4) in which each row characterizes one variable. This matrix does not need to be rotated. Further analysis does not depend on whether the matrix is rotated or not. By application of fuzzy cluster analysis to this matrix (for L clusters) we obtain the M x L (in our example 65 x 11) matrix of membership coefficients. In this case, one cluster represents one factor. If the membership coefficient is greater than or equal to $1/L$ then the variable belongs to the particular factor; in the opposite case the variable is not assigned to this factor. If such a high value is in the row only once (usually greater than 0.5) then the variable is assigned to only one factor. As the result of our analysis we obtained values of coefficients for the assignments to only one factor greater than 0.9.

For binary factorization, it means that if a value of a membership coefficient is greater than or equal to $1/L$, then it can be replaced by the value 1. In the opposite case, it would be replaced by zero.

In the example described in this section, all variables assigned to only one factor were determined correctly and some factors were also determined exactly. However, some variables were assigned to more clusters than was correct. Since we did experiments with only small numbers of simulated data files, any conclusion about mistakes cannot be made.

We also tried to interpret factor loadings for $K = 10$ but we did not obtain the expected results. In some columns representing clusters all values of membership coefficients were lower than 0.5. In addition, we applied multidimensional scaling in STATISTICA to the proximity matrix of Jaccard coefficients; a proximity matrix was generated in the special form required by STATISTICA. For result interpretation of 4 dimensions, we used fuzzy cluster analysis. However, only a small amount of variables was assigned to only one cluster.

Taking into consideration $C = 2$ and repeating the analysis for the second data file, we obtained results almost identical to the previous case. The membership coefficients of variables assigned to only one cluster were greater than 0.89. For $C = 3$, the membership coefficients of variables assigned to only one cluster were greater than 0.88. When we changed a number of cases to $N = 1000$, the mentioned figure was 0.89. Differences in membership coefficients for the mentioned examples were insignificant for finding factors.

4 Application

We also tried to analyze the data matrix containing values about the occurrence terms in textual documents. We analyzed the data published in Berry et al. (15 documents and 18 terms). We applied both factor analysis and multiple correspondence analysis and then we used fuzzy cluster analysis for result interpretation. We interpreted only the part of output about coordinates from multiple correspondence analysis, i.e. terms in connection with the value “one”.

Output 1. Interpretation of the factor loading matrix

```
Membership coefficients:
              [1]   [2]   [3]   [4]
    ABNORMAL 0.84  0.05  0.05  0.06
         AGE 0.37  0.19  0.13  0.32
    BEHAVIOR 0.13  0.50  0.15  0.23
       BLOOD 0.70  0.09  0.10  0.10
       CLOSE 0.30  0.17  0.37  0.15
     CULTURE 0.21  0.17  0.10  0.52
    DEPRESSE 0.10  0.54  0.11  0.24
    DISCHARG 0.06  0.12  0.05  0.77
     DISEASE 0.35  0.19  0.29  0.17
        FAST 0.05  0.06  0.85  0.04
    GENERATI 0.18  0.26  0.40  0.16
    OESTROGE 0.07  0.69  0.09  0.15
    PATIENTS 0.08  0.12  0.05  0.76
    PRESSURE 0.07  0.07  0.82  0.05
        RATS 0.11  0.16  0.63  0.10
     RESPECT 0.78  0.07  0.07  0.09
        RISE 0.14  0.37  0.32  0.17
       STUDY 0.10  0.36  0.09  0.45
```

There were two factors in factor analysis and two coordinates in multiple correspondence analysis. The factor loadings and coordinates were interpreted by four clusters in fuzzy cluster analysis. Outputs from the S-PLUS system are shown in Output 1 and Output 2 (the values greater than $1/L$ and single in the row are written in bold).

On the basis of Output 1 we can identify four following groups of terms (the terms which are assigned to only one group are written in bold).

Group 1: **abnormalities**, age, **blood**, close, disease, **respect**
Group 2: **behavior, depressed**, generation, **oestrogen**, rise, study
Group 3: close, disease, **fast**, generation, **pressure, rats**, rise
Group 4: age, **culture, discharged, patients**, study

On the basis of Output 2 we obtain a very similar result with the following difference: behavior is assigned to two groups (groups 2 and 4 mentioned above) and generation is assigned to only one group (group 3 mentioned above). The application of multiple correspondence analysis is a little more complicated because we have to omit the rows for the zero values from the output.

Output 2. Interpretation of coordinates obtained by multiple correspondence analysis

Membership coefficients:

	[1]	[2]	[3]	[4]
ABNORMAL:1	**0.89**	0.04	0.03	0.04
AGE:1	0.31	0.38	0.17	0.14
BEHAVIOR:1	0.11	0.25	0.49	0.15
BLOOD:1	**0.76**	0.09	0.07	0.09
CLOSE:1	0.27	0.15	0.15	0.43
CULTURE:1	0.18	**0.57**	0.16	0.10
DEPRESSE:1	0.06	0.17	**0.68**	0.08
DISCHARG:1	0.06	**0.75**	0.13	0.05
DISEASE:1	0.33	0.17	0.16	0.34
FAST:1	0.08	0.07	0.08	**0.77**
GENERATI:1	0.14	0.14	0.19	**0.53**
OESTROGE:1	0.08	0.16	**0.63**	0.12
PATIENTS:1	0.06	**0.81**	0.09	0.04
PRESSURE:1	0.12	0.10	0.11	**0.67**
RATS:1	0.13	0.13	0.18	**0.55**
RESPECT:1	**0.83**	0.07	0.05	0.05
RISE:1	0.14	0.18	0.34	0.35
STUDY:1	0.09	0.44	0.38	0.09

5 Conclusion

By reason of a special procedure for Boolean factor analysis is missing in statistical packages, we have tried to find a way to determine L overlapping binary factors. We suggest applying factor analysis for K factors, where K is less than L and then interpreting the factor loading matrix by using fuzzy cluster analysis for L clusters. If a value of the membership coefficient is greater than or equal to $1/L$, then it can be replaced by the value 1. In the opposite case, it would be replaced by zero. In such a way, we can obtain the searched for factors in columns of the membership coefficient matrix.

We applied this technique to data files generated with different parameters and we obtained similar results. By using this technique we got a higher ratio of the number of active elements to the total number of elements for some factors with respect to variables. The results are not exact but for the users of statistical packages, it can be a suitable advice on how to solve tasks of this type.

We also analyzed the input data matrix characterizing a collection of textual documents. We found that the results obtained by the method described above are almost the same as results obtained by the application of multiple correspondence analysis and interpretation of coordinates by fuzzy cluster analysis.

Obviously, a special procedure for solving the described task would be a better solution. At the present time, when the use of artificial neural networks is popular for data mining, it seems that the development of neural networks for this purpose is helpful.

References

[1] Berry M.W., Dumais S.T., Letsche T.A. (1995). *Computational methods for intelligent information access.*
`http://www.cs.utk.edu/ berry/sc95/sc95.html`.

[2] Fichet B. (1986). *Distances and Euclidean distances for presence-absence characters and their application to factor analysis.* Multidimensional Data Analysis, Proceedings of a workshop, DSWO Press, Leiden, 23 – 46.

[3] Frolov A.A., Húsek D., Muraviev I.P. (1997). *Informational capacity and recall quality in sparsely encoded Hopfield-like neural network: Analytical approaches and computer simulation.* Neural Networks **10**, 845 – 855.

[4] Frolov A.A., Húsek D., Muraviev I.P., Řezanková H., Snášel V., Polyakov P.A. (2003). *Features extraction by Hopfield-like neural network.* Neural Network Engineering Experiences. University of Malaga, Malaga, 383 – 390.

[5] Hopfield J.J. (1982). *Neural network and physical systems with emergent collective computational abilities.* Proc. Natl. Acad. Sci. USA **79**, 2544–2548.

[6] Perez-Vicente C.J., Amit D. (1989). *Optimized network for sparsely encoded patterns.* J. of Physics A: Math. Gen. **22**, 559 – 569.

[7] Řezanková H., Húsek D. (2002). *Comparison of the SAS, SPSS and STATISTICA systems in the area of clustering variables.* Computational Statistics & Data Analysis (Statistical Software Newsletter) **41**, 331 – 339.

[8] Řezanková H., Húsek D. (2003). *Nonlinear factorization of binary variables in statistical packages.* Mundus Symbolicus **11** (1) 35 – 44.

[9] Řezanková H., Húsek D., Smid J., Snášel V. (2003). *Clustering of documents via similarity measures.* CIC'03, CSREA Press, Las Vegas, 292 – 299.

Acknowledgement: The presented paper is financially supported by grant GA MSK(CZ)LN00B096 awarded by Ministry of education of Czech Republic.

Address: H. Řezanková, University of Economics, W. Churchill Sq. 4,
130 67 Prague 3, Czech Republic
D. Húsek, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic
A.A. Frolov, Institute of Higher Nervous Activity and Neurophysiology of the Russian Academy of Sciences, Butlerova 5a, Moscow, Russia

E-mail: `rezanka@vse.cz, dusan@cs.cas.cz, aafrolov@mail.ru`

COMPSTAT'2004 Symposium

HOMOGENEITY ANALYSIS FOR SETS OF TIME SERIES

Pedro Saavedra, C.N. Hernández, I. Luengo, J. Artiles and A. Santana

Key words: Stationary processes, population spectrum.

COMPSTAT 2004 section: Time series analysis.

Abstract: A set of time series generated by stationary processes with absolutely continuous spectral distribution is considered and a statistic test for testing the spectral homogeneity is proposed. We approach its probability distribution under the null hypothesis taking into account asymptotic results for the periodograms of linear processes. The procedure is illustrated by means of a small simulation study.

1 Introduction

Let $\left(A, \Lambda, P^{A}\right)$ be a probability space related to a set of objects such that, on each one, a stationary stochastic process with absolutely continuous spectral distribution can be observed. The aim of this paper is to survey the homogeneity of such processes. Therefore, a random sample of objects $a_1, \ldots, a_r$ is drawn from A, and for all a_i, the processes $X\left(a_i, t\right)$ are observed at the same time points $t = 1, \ldots, N$. Thus, the available data for the analysis is the set of time series $\left\{X\left(a_i, t\right) : i = 1, \ldots, r; t = 1, \ldots, N\right\}$. We denote by $Q_i\left(\omega\right)$, for $|\omega| \leq \pi$, the spectral density function corresponding to the ith process $X\left(a_i, t\right)$. In these conditions, $Q_1\left(\omega\right), \ldots, Q_r\left(\omega\right)$ can be considered independent realizations of a stochastic process defined on the probability space $\left(A, \Lambda, P^{A}\right)$. The parameters of interest are $f\left(\omega\right) = E_A\left[Q_i\left(\omega\right)\right]$ and $\mathrm{var}_A\left(Q_i\left(\omega\right)\right)$ (E_A and var_A denote respectively the expectation and variance on the set A). The function $f\left(\omega\right)$ is the so called population spectrum [1]. The time series are generated by the same pattern when $\mathrm{var}_A\left(Q_i\left(\omega\right)\right) = 0$, for all $|\omega| \leq \pi$. We propose a procedure for testing the null hypothesis $H_0 : \mathrm{var}_A\left(Q_i\left(\omega\right)\right) = 0$, for $|\omega| \leq \pi$. In section 2 a brief revision of the spectral theory for the doubly stochastic process $\left\{X\left(a, t\right) : a \in A, t \in \mathbb{Z}\right\}$ is shown. For the aforementioned testing, a statistic test is proposed in section 3. We also give a method for approaching its probability distribution under the null hypothesis. This approach is based on the distribution of periodogram ordinates [7]. Finally, some simulations illustrate the proposed procedure.

2 The general model of random effects

We analyse the considered data set $\{X(a_i,t) : i = 1,\ldots,r; t = 1,\ldots,N\}$ in the frequency domain. Let

$$I_{a,N}(\omega) = \frac{1}{2\pi N}\left|\sum X(a,t)\exp(-i\omega t)\right|^2 \tag{1}$$

denote the periodogram of the time series observed on the object . According to Saavedra et al. [6], we suppose that the doubly stochastic process $\{X(a,t) : a \in A, t \in \mathbb{Z}\}$, with population spectrum $f(\omega)$ obeys to the general model of random effects in the sense that the ith periodogram verifies at the jth Fourier frequency $\omega_j = \frac{2\pi j}{N}$, $j = 1,\ldots,\left[\frac{N}{2}\right]$:

$$I_{i,N}(\omega_j) = f(\omega_j)\cdot Z_i(\omega_j)\cdot U_{ij}^{(N)} + R_{i,N}(\omega_j) \tag{2}$$

where $W_{ij}^{(N)} = Z_i(\omega_j)\cdot U_{ij}^{(N)}$ satisfies:

i. $E_A\left[W_{ij}^{(N)}\right] = 1$

ii. $E_A\left[W_{ij}^{(N)}\cdot W_{il}^{(N)}\right] = 1 + \operatorname{cov}_A(Z_i(\omega_j)\cdot Z_i(\omega_l))$

iii. $E_A\left[|R_{i,N}(\omega_j)|^2\right] = O\left(\frac{1}{N^2}\right)$, uniformly in j.

Saavedra et al give conditions for a doubly stationary process to obey the model (2). An interesting class are the moving average processes with random coefficients given by:

$$X(a,t) = \sum_{u=0}^{p-1} g_u(a)\cdot\xi(a,t-u) \tag{3}$$

where for each object $a \in A$, $\{\xi(a,t) : t \in \mathbb{Z}\}$ is a gaussian white noise with probability distribution independent from a, and $g\prime = (g_0, g_1, \ldots, g_{p-1})$ being a random vector defined on $(\mathbb{R}^p, \mathfrak{B})$, with expectation vector γ and covariance matrix C. The idea behind the model (2) is based on the asymptotic representation of the periodogram of linear processes. According to this representation, for each object $a \in A$, the periodogram of the process $X(a,t)$ satisfies $I_{a,N}(\omega_j) = Q_a(\omega_j)\cdot U_{a,j}^{(N)}$, where for $j = 1,\ldots,[N/2]$, the $U_{a,j}^{(N)}$ are asymptotically uncorrelated and exponentially distributed with parameter one random variables. If $f(\omega) = E_A[Q_a(\omega)]$ exists, the $Z_a\ (\omega) = \frac{Q_a(\omega)}{f(\omega)}$ can be defined and thus (2) follows. From the given data, each trajectory $Q_a(\omega)$ can be estimated as:

$$\hat{Q}_a(\omega,\lambda) = \frac{1}{N\lambda}\sum_{j=-\upsilon}^{\nu} K\left(\frac{\omega-\omega_j}{\lambda}\right)\cdot I_{a,N}(\omega_j) \tag{4}$$

being λ the bandwidth an $K(u)$ the kernel function. We suppose that $K(u)$ is a symmetric, nonnegative function, with compact support $[-\kappa, \kappa]$ and uniformly Lipschitz, being $\int K(u)\, du = \int u^2 K(u)\, du = 2\pi$. For each $a \in A$, the properties of the estimate $\hat{Q}_a(\omega; \lambda)$ can be obtained under the assumptions given by Franke y Härdle [3]. Under such conditions, and letting $\lambda \longrightarrow 0$, $N \longrightarrow \infty$ and $\left(N\lambda^4\right)^{-1} = O(1)$, the mean square error is:

$$\begin{aligned} \text{mse}\left(\hat{Q}_a(\omega, \lambda)\right) &= E_a\left[\left\{\hat{Q}_a(\omega, \lambda) - Q_a(\omega)\right\}^2\right] = \\ &O\left((N\lambda)^{-1}\right) + O\left(\lambda^4\right) + O\left(\left(\frac{\log N}{N}\right)^2\right) \end{aligned} \tag{5}$$

uniformly in $|\omega| < \pi - \kappa h$. Taking the bandwidth as $\lambda \sim N^{-\frac{1}{5}}$, it occurs that mse$\left(\hat{Q}_a(\omega, \lambda)\right) \sim N^{-\frac{4}{5}}$. Likewise, the population spectrum $f(\omega)$ is estimated as:

$$\hat{f}(\omega, h) = \frac{1}{Nh} \sum_{j=-\upsilon}^{\nu} K\left(\frac{\omega - \omega_j}{h}\right) \cdot \bar{I}_{\bullet, N}(\omega_j) \tag{6}$$

being $\bar{I}_{\bullet, N}(\omega_j) = \frac{1}{r}\sum_{i=1}^{r} I_{i,N}(\omega)$ the average periodogram. For the general model of random effects and being $f(\omega)$ twice continuously differentiable over $[-\pi, \pi]$, Saavedra et a.l prove that, for $N \longrightarrow \infty$, $h \longrightarrow 0$ and $\left(Nh^4\right)^{-1} = O(1)$

$$E_A\left[\hat{f}(\omega, h)\right] - f(\omega) = \frac{h^2}{2} f\text{"}(\omega) + o\left(h^2\right) + O\left(\frac{1}{N}\right) \tag{7}$$

uniformly in $|\omega| < \pi - \kappa h$. Moreover, if $\psi(u, v) = \text{cov}_A(Z_i(u), Z_i(v))$ is differentiable for $u, v \in [-\pi, \pi]$, then:

$$\begin{aligned} \text{var}_A\left(\hat{f}(\omega; h)\right) = &\frac{f^2(\omega)}{Nhr}\left\{E_A\left[Z_a^2(\omega)\right] + \text{var}_A(Z_a(\omega))\right\} \cdot \|K\|^2 + \\ &o\left((Nhr)^{-1}\right) + \frac{f^2(\omega)\, \text{var}_A(Z_a(\omega))}{r} + \\ &O\left(\frac{h^2}{r}\right)\left\{\text{var}_A(Z_a(\omega)) + \left(\frac{\partial}{\partial \omega}\right)\text{var}_A(Z_a(\omega))\right\} \end{aligned} \tag{8}$$

uniformly in $|\omega| < \pi - \kappa h$ and being $\|K\|^2 = \frac{1}{2\pi}\int K^2(u)\, du$.

3 The homogeneity test

The homogeneity hypothesis of the trajectories $Q_i(\omega)$ can be formalized as $H_0 : \text{var}_A(Q_a(\omega)) = 0$, for $|\omega| < \pi$. It seems clear that a test for this hypothesis must be based on a statistic as:

$$\frac{1}{r}\sum_{i=1}^{r}\int_{0}^{\pi}\left\{Q_i\left(\omega\right)-f\left(\omega\right)\right\}^2 d\tau\left(\omega\right) \tag{9}$$

for some measure defined on $[-\pi,\pi]$. Obviously, the population spectrum $f\left(\omega\right)$ is unknown and the trajectories $Q_i\left(\omega\right)$ can be not directly observed. Therefore, we consider a statistic of the form:

$$\frac{1}{r}\sum_{i=1}^{r}\int_{0}^{\pi}\left\{\hat{Q}_i\left(\omega;\lambda\right)-\hat{f}\left(\omega;h\right)\right\}^2 d\tau\left(\omega\right) \tag{10}$$

being the $\hat{Q}_i\left(\omega\right)$ estimations of the trajectories $Q_i\left(\omega\right)$ for $i=1,\ldots,r$ and $\hat{f}\left(\omega\right)$ an estimation of the population spectrum $f\left(\omega\right)$. We denote by $\delta^2_{r,N}$ the statistics resulting of taking as measure $\tau\left(\omega\right)$, which assigns mass $\frac{2\pi}{N}$ to each Fourier frequency. Thus, it is easy to obtain:

$$E_A\left[\delta^2_{r,N}\right]\leq$$
$$\frac{6\pi}{N}\sum_{j=0}^{\nu}\left\{\mathrm{var}_A\left(Q_i\left(\omega_j\right)\right)+\mathrm{mse}\left(\hat{f}\left(\omega_j;h\right)\right)+E_A\left[\mathrm{mse}\left(\hat{Q}_i\left(\omega_j\right);\lambda\right)\right]\right\} \tag{11}$$

where $\nu=\left[\frac{N}{2}\right]$. For the general model of random effects given by (2), $\mathrm{mse}\left(\hat{f}\left(\omega\right)\right)$ is of order r^{-1} if the null hypothesis fails and of order $\left(Nhr\right)^{-1}$ otherwise. Moreover, according to Franke and Härdle, it can be derived that under adequate assumptions $E_A\left[\mathrm{mse}\left(\hat{Q}_i\left(\omega_j;\lambda\right)\right)\right]\sim N^{-\frac{4}{5}}$, for $\lambda\sim N^{-\frac{1}{5}}$, $N\longrightarrow\infty$, We now consider the statistic test:

$$D^2_{r,N}=r\delta^2_{r,N}=\frac{2\pi}{N}\sum_{i=1}^{r}\sum_{j=1}^{\upsilon}\left\{\hat{Q}_i\left(\omega;\lambda\right)-\hat{f}\left(\omega;h\right)\right\}^2 \tag{12}$$

and give the following procedure for getting an approach of the probability distribution of $D^2_{r,N}$ under H_0 taking in account the asymptotic results for the periodograms of linear processes.

Step 1. We choose a estimate $\hat{f}\left(\omega;h\right)$ of the population spectrum $f\left(\omega\right)$, for some bandwidth h.

Step 2. For each $i=1,\ldots,r$, we simulate independent and exponentially distributed with parameter one random variables $V_{i,1},\ldots,V_{i,\nu}$. We define the periodograms $\tilde{I}_{i,N}\left(\omega_j\right)=\hat{f}\left(\omega_j;h\right)\cdot V_{i,j}$.

Step 3. From the periodograms $\tilde{I}_{i,N}\left(\omega_j\right)$, we obtain estimations $\tilde{Q}_i\left(\omega_j;\lambda\right)$ of the trajectory $Q_i\left(\omega_j\right)$ for $i=1,\ldots,r$. By means of (6), we also obtain a estimation $\tilde{f}\left(\omega;h_2\right)$ of $\hat{f}\left(\omega;h\right)$.

Step 4. We finally compute:

$$\tilde{D}^2_{r,N} = \frac{2\pi}{N}\sum_{i=1}^{r}\sum_{j=1}^{\nu}\left\{\tilde{Q}_i\left(\omega_j;\lambda\right) - \tilde{f}\left(\omega_j;h_2\right)\right\}^2$$

The B values $\tilde{D}^2_{r,N}$ obtained by iterating of the steps 2, 3 y 4, give an approach to the probability distribution under H_0 of the statistic test $D^2_{r,N}$.

4 Simulations

We illustrate the proposed procedure by simulations of a moving average process with random coefficients. Let $X\left(a,t\right) = \xi\left(t\right) + a_1 \cdot \xi\left(t-1\right) + a_2 \cdot \xi\left(t-2\right)$ be a stochastic process such that:

i.- $a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \approxeq N_2\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} \frac{\tau}{2} & \frac{\tau\sqrt{2}}{4} \\ \frac{\tau\sqrt{2}}{4} & \frac{\tau}{2} \end{pmatrix}\right)$

ii.- $\{\xi\left(t\right) : t \in \mathbb{Z}\}$ are independent and with standard normal distribution random variables.

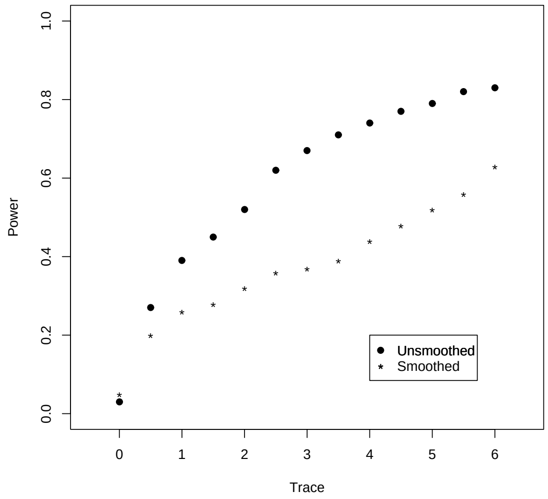


Figure 1: The power functions according to the statistic test is based or not on estimates smoothing of $Q_i\left(\omega\right)$ and $f\left(\omega\right)$.

We have considered a sample size of $r = 50$ objects and $N = 200$ observations per object. Note that if the trace of covariance matrix is $\tau = 0$, then, the distribution of $a' = (a_1, a_2)$ is degenerate, and thus, all time series are generated by the same pattern. Obviously, the homogeneity hypothesis can be expressed as $H_0 : \tau = 0$. statistic test given in (12) depends on the bandwidths h and λ.We could take the unsmoothed periodogram $I_{i,N}(\omega)$ as an estimate of the trajectory $Q_i(\omega)$ and the average periodogram $\bar{I}_{\bullet,N}(\omega)$ as an estimate of the population spectrum $f(\omega)$. However, it seems clear that the optimum power is obtained when the bandwidths $\lambda$and h are optimum. For a set of alternatives values τ,figure 1 shows the power functions obtained for both tests, according to the statistic test is based on the optimum estimations of $Q_i(\omega)$ and $f(\omega)$ or on the unsmoothed ones.

We have also approached the probability distribution under the null hypothesis $H_0 : \tau = 0$ of the statistic test $D^2_{r=50,N=200}$ according to the algorithm given in section 3, taking as number of iterations the value $B = 500$. Figure 2 shows the true probability density function of statistic test and the approach obtained.

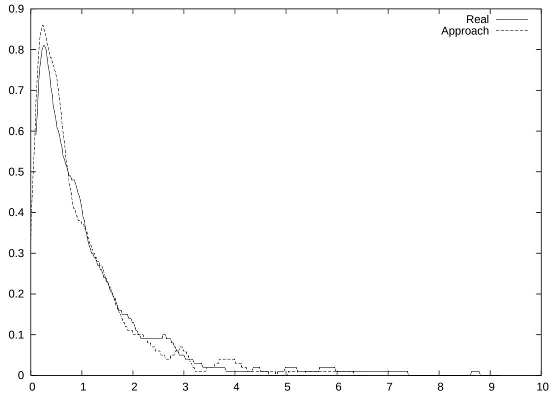


Figure 2: True probability density function of statistic test and the approach with 500 iterations.

5 Discussion

In the field of biomedical science, the survey of a marker prognosis value can require a previous testing of the homogeneity of that marker over the population considered. The procedure proposed can be applied to signals evaluated over patients that can be modelled as stationary linear processes. The decomposition of $E_A\left[\delta^2_{r,N}\right]$ given at (11), shows that the power of the test can be improved, minimizing the mean square errors of the estimates $Q_i(\omega)$ and $f(\omega)$ (see figure 2). Finally, we point out that the procedure for approaching the probability distribution under H_0 of the statistic test given in (12) is of easy implementation in a language for statistical computing such as S or R.

References

[1] Diggle P.J., I. Al-Wasel (1993). *On periodogram-based spectral estimation for replicated time series. In: Subba Rao (Ed)*, Developments in Time Series Analysis, (Chapman and Hall, Great Britain), 341 – 354.

[2] Diggle P.J., Al-Wasel I. (1997). *Spectral analysis of replicated biomedical time series.* Appl.Statist **46**, 31 – 71.

[3] Franke J., Härdle W. (1992). *On bootstraping kernel spectral estimates.* Ann. Stat. **30**, 121 – 145.

[4] Hernández-Flores C.N., Artiles-Romero J., Saavedra-Santana P. (1999). *Estimation of the population spectrum with replicated time series.* Comp. Stat and Data Anal. **30**, 271 – 280.

[5] Priestley M.B. (1981). *Spectral analysis and time series.* (Wiley, New York).

[6] Saavedra P., Hernández C.N., Artiles J. (2000). *Spectral analysis with replicated time series.* Communications. In Statistics Theory and Methods **29**, 2343 – 2362.

[7] Chen Z.-G. and Hannan E. J. (1980). *The distributions of periodogram ordinates.* Journal of Time Series Analysis. **1**, 73 – 82.

[8] Development Core Team (2003). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, http://www.R-project.org.

Acknowledgement: This paper has been part of the research program of I+D (reference BFM 2000-0517).

Address: P. Saavedra, C.N. Hernández, I. Luengo, J. Artiles A. Santana, Departament of Mathematics, University of Las Palmas de Gran Canaria, Spain

E-mail: saavedra@dma.ulpgc.es

PROPERTIES OF THE SLIDE VECTOR MODEL FOR ANALYSIS OF ASYMMETRY

Takayuki Saito

Key words: Algebraic solution, dissimilarity, line structure, multidimensional scaling, necessary condition, skew-symmetric matrix, sufficient condition.

COMPSTAT 2004 section: Dimensional reduction.

Abstract: The slide vector model, that has been proposed to account for asymmetric proximity data in multidimensional scaling (MDS), has a wide applicability for dimensional reduction because the model is represented by a general form in terms of distance and scalar product. Although an iterative procedure for fitting the model to data has been suggested, the mathematical structure of the model itself has never been clarified to date. In this paper, we clarify algebraic properties of the model, presenting necessary and/or sufficient conditions of the model to hold for a given set of data. On the basis of those conditions, we show an algebraic solution of the model in a closed form, of which role is discussed in comparison with the iterative procedure.

1 Introduction

Many models and methods have been developed to analyze asymmetric proximity (similarity or dissimilarity) data in the field of MDS. A comprehensive review is given by Zielman & Heiser [9]. Let us suppose an $n \times n$ asymmetric data matrix $\boldsymbol{\Delta} = (\delta_{jk})$ where δ_{jk} indicates the *dissimilarity* for (j, k), an ordered pair of objects j and k. All the *diagonal* elements are observed and not zero. For dissimilarity data of this type, analysis of asymmetry was proposed in a variety of ways (e.g. [1], [5], [3], [6].

Among them, we are concerned with the slide vector model(SVM) suggested by Zielman & Heiser [8]. Unlike most of the models in asymmetric MDS, which are viewed as extensions of distance models or scalar product models in symmetric MDS, SVM represents dissimilarity in the form of combination of distance and scalar product terms as in (4). This feature allows a possibility that SVM derives structural information from dissimilarity data that the other asymmetric models do not to good extent.

For data analysis by SVM, Zielman & Heiser [8] suggested an iterative procedure to estimate the model parameters. It maximizes the degree of fit of the model to data for specified dimensions. Then one would adopt the solution with the best fit, changing dimensions. After the treatment, however, it would remain unknown whether it is justified or not to assume the model structure to the data.

To cope with this problem, we clarify the properties of the model by an algebraic treatment, and provide necessary and/or sufficient conditions for the model to hold. The results serve for testing whether or not a set of data involves the structure of SVM. Furthermore, we suggest an algebraic solution under some assumptions, and then discuss its implication and utilization in data analysis.

2 Algebraic consideration about the slide vector model

The slide vector model

The model represents asymmetry by a uniform shift or translation of the difference vector between the two object points in a multidimensional space. The error-free model in r dimensions is stated as

$$\delta_{jk} = \left(\sum_{t=1}^{r} (x_{jt} - x_{kt} + z_t)^2 \right)^{\frac{1}{2}} \quad (j, k = 1, 2, \ldots, n)\,. \tag{1}$$

The dissimilarity is a function of $\boldsymbol{X}$ and $\boldsymbol{z}$, where $\boldsymbol{X} = (x_{jt})$ is an $n \times r$ matrix of coordinates, and $\boldsymbol{z} = (z_t)$ is an $r \times 1$ slide-vector. Denote the j-th row of $\boldsymbol{X}$ by $\boldsymbol{x}_{(j)} = (x_{j1}, \ldots, x_{jr})'$ and the t-th column by $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{nt})'$, then

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_r) = (\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \cdots, \boldsymbol{x}_{(n)})'\,. \tag{2}$$

The self-dissimilarity is equal to a constant γ,

$$\delta_{jj} = \left(\sum_{t=1}^{r} z_t^2 \right)^{\frac{1}{2}} = \gamma \quad (j = 1, 2, \ldots, n). \tag{3}$$

To find implications of the model, we expand (1) as

$$\delta_{jk}^2 \quad = \quad d_{jk}^2 + \gamma^2 + 2 \sum_{t=1}^{r} z_t (x_{jt} - x_{kt})\,. \tag{4}$$

On the right-hand side, the first term shows the squared Euclidean distance and the second the squared norm of $\boldsymbol{z}$, both of which contribute to the symmetric component of the dissimilarity. In contrast, the third term represents the skew-symmetric component as

$$\boldsymbol{z}'\boldsymbol{x}_{(j)} - \boldsymbol{z}'\boldsymbol{x}_{(k)} = \sum_{t=1}^{r} z_t (x_{jt} - x_{kt})\,. \tag{5}$$

Define a symmetric matrix $\boldsymbol{S} = (s_{jk})$ and a skew-symmetric matrix $\boldsymbol{A} = (a_{jk})$ as follows:

$$s_{jk} \quad = \quad (\delta_{jk}^2 + \delta_{kj}^2)/2 \tag{6}$$

$$a_{jk} \quad = \quad (\delta_{jk}^2 - \delta_{kj}^2)/2\,. \tag{7}$$

Then we have a decomposition of squared data as $\delta_{jk}^2 = s_{jk} + a_{jk}$, each component is represented as follows:

$$s_{jk} = d_{jk}^2 + \gamma^2\,, \tag{8}$$

$$a_{jk} = 2\boldsymbol{z}'(\boldsymbol{x}_{(j)} - \boldsymbol{x}_{(k)})\,. \tag{9}$$

Indeterminacy of the parameters

Let us consider identification of the model parameters. A translation of the origin of $\boldsymbol{X}$ by $\boldsymbol{\xi} = (\xi_t)$, which is denoted as $\tilde{x}_{jt} = x_{jt} + \xi_t$, does not affect the model representation. Consider an orthogonal matrix $\boldsymbol{\Gamma}$ and transformations of $\boldsymbol{X}$ and $\boldsymbol{z}$ such as $\boldsymbol{x}_{(j)}^* = \boldsymbol{\Gamma}'\boldsymbol{x}_{(j)}$ and $\boldsymbol{z}^* = \boldsymbol{\Gamma}'\boldsymbol{z}$. Apparently $d_{jk}(\boldsymbol{X}) = d_{jk}(\boldsymbol{X}^*)$. Substituting $\boldsymbol{X}^*$ and $\boldsymbol{z}^*$ in (9), we find invariance of the skew-symmetric part in such a way as

$$2\boldsymbol{z}^{*\prime}(\boldsymbol{x}_{(j)}^* - \boldsymbol{x}_{(k)}^*) = 2\boldsymbol{z}'(\boldsymbol{x}_{(j)} - \boldsymbol{x}_{(k)})\,. \tag{10}$$

Hence the model is invariant with respect to translations and rotations.

Properties of the skew-symmetric part

From (4) we have

$$\delta_{jk}^2 - \delta_{kj}^2 = 4\sum_{t=1}^{r} z_t(x_{jt} - x_{kt})\,. \tag{11}$$

Summing equations of this type over pairs (i,j), (j,k) and (k,i) gives

$$\delta_{ij}^2 + \delta_{jk}^2 + \delta_{ki}^2 = \delta_{ik}^2 + \delta_{kj}^2 + \delta_{ji}^2 \quad \text{for all triples} \quad (i,j,k). \tag{12}$$

Thus a relation of additivity, which may be called transitivity, holds with $\{\delta_{jk}^2\}$. Using (7), we have another expression of (12) as

$$a_{jk} = a_{ji} + a_{ik} \quad \text{for all triples} \quad (i,j,k). \tag{13}$$

Let $\delta_{j.}^2$ be the j-th row average and $\delta_{.j}^2$ the j-th column average of $\{\delta_{jk}^2\}$. Then (13) is rewritten as

$$a_{jk} = a_{j.} - a_{k.} \quad \text{where} \quad a_{j.} = \frac{1}{2}(\delta_{j.}^2 - \delta_{.j}^2)\,. \tag{14}$$

Let us define $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{z}$, of which element is

$$y_j = \boldsymbol{z}'\boldsymbol{x}_{(j)} = \sum_{t=1}^{r} z_t x_{jt} \quad (j = 1, 2, \ldots, n). \tag{15}$$

Thus y_j means a kind of closeness of object j to the reference point specified by the slide vector $\boldsymbol{z}$, since y_j is the projection of vector $\boldsymbol{x}_{(j)}$ on vector $\boldsymbol{z}$. Let $\boldsymbol{1}$ be an n-dimensional vector of ones. From (9) and (15), we write $\boldsymbol{A}$ as

$$\boldsymbol{A} = 2(\boldsymbol{y}\boldsymbol{1}^{'} - \boldsymbol{1}\boldsymbol{y}^{'})\,, \tag{16}$$

indicating that $\boldsymbol{A}$ is of rank 2 and $a_{jk} = 2(y_j - y_k)$. For convenience of exposition, we set $y_j = a_{j.}/2$, which leads to

$$\sum_{j=1}^{n} y_j = 0\,. \tag{17}$$

There could be a special case in which all $\boldsymbol{x}_{(j)}$ are orthogonal to $\boldsymbol{z}$. However, the possibility is extremely small in practice and can be safely ignored. Accordingly, in view of (15), we set a constraint for $\boldsymbol{X}$ to meet (17), such as

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_{(j)} = \boldsymbol{0}\,. \tag{18}$$

Through a simple manipulation, the singular value decomposition (SVD) of $\boldsymbol{A}$ is stated as

$$\boldsymbol{A} \;\; = \;\; \mu(\boldsymbol{w}_1\boldsymbol{w}_2^{'} - \boldsymbol{w}_2\boldsymbol{w}_1^{'})\,.$$

Here $\boldsymbol{w}_1 = (w_{j1})$ and $\boldsymbol{w}_2 = (w_{j2})$ are singular vectors associated with singular value μ, which are given by

$$\boldsymbol{w}_1 = \frac{1}{\|\boldsymbol{y}\|}\boldsymbol{y}\,, \quad \boldsymbol{w}_2 = \frac{1}{\sqrt{n}}\boldsymbol{1} \quad \text{and} \quad \mu = 2\sqrt{n}\|\boldsymbol{y}\|\,. \tag{19}$$

Plotting objects in terms of (w_{j1}, w_{j2}) on the plane spanned by those vectors [2], it will reveal a line pattern perpendicular to axis w_2.

However, we will not always obtain singular vectors of $\boldsymbol{A}$ in the form of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ from SVD of data matrix $\boldsymbol{A}$, because the decomposition is provided with rotational indeterminacy. To show this possibility, we take up $\boldsymbol{A}$ given by (16). By some manipulation on $\boldsymbol{A}$, it is found that the following equation holds,

$$\boldsymbol{A}\boldsymbol{A}\boldsymbol{1} \;\; = \;\; \lambda\boldsymbol{1}\,. \tag{20}$$

Here $\lambda = 4(\theta_1^2 - n\theta_2)$, $\theta_1 = \boldsymbol{1}^{'}\boldsymbol{y}$ and $\theta_2 = \boldsymbol{y}^{'}\boldsymbol{y}$. To compute SVD of $\boldsymbol{A}$ in terms of data, we deal with equations such as

$$\boldsymbol{A}\boldsymbol{u}_1 = -\mu\boldsymbol{u}_2 \quad \text{and} \quad \boldsymbol{A}\boldsymbol{u}_2 = \mu\boldsymbol{u}_1\,.$$

According to a theorem [4], condition (20) is necessary and sufficient in order that a line structure exists on the plane spanned by $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, such that $\boldsymbol{u}_2 = a + b\boldsymbol{u}_1$ where $b \neq 0$. Let $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2)$ and $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2)$. By the rotational indeterminacy in the SVD of $\boldsymbol{A}$, $\boldsymbol{U}$ is related to $\boldsymbol{W}$ with a 2×2 orthogonal matrix $\boldsymbol{T}$ in such a way as $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{T}$.

Thus the line pattern in terms of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ is a rotated one of the line pattern given by $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. Hence, when the model holds, SVD of the skew-symmetric part of $\{\delta^2_{jk}\}$ reveals a line pattern. The spacing of points on the line corresponds to the projection of object points on the slide vector.

Properties of the symmetric part

Summing equations of type (4) over pairs (j,k) and (k,j) gives a representation that

$$\delta^2_{jk} + \delta^2_{kj} - \delta^2_{jj} - \delta^2_{kk} = 2d^2_{jk}\,. \tag{21}$$

Define a matrix of transformed data $\boldsymbol{Q} = (q_{jk})$ where

$$q_{jk} = \frac{1}{2}(\delta^2_{jk} + \delta^2_{kj} - \delta^2_{jj} - \delta^2_{kk})\,. \tag{22}$$

Rewrite (21) as $q_{jk} = d^2_{jk}$. If the model holds with $\boldsymbol{\Delta}$, q_{jk} should be nonnegative, from which follows that

$$\delta^2_{jk} + \delta^2_{kj} \quad \geq \quad \delta^2_{jj} + \delta^2_{kk} \quad \text{for all pairs } (j,k)\,. \tag{23}$$

When this condition is satisfied, we define $\boldsymbol{P} = (p_{jk})$ where

$$p_{jk} \quad = \quad q^{\frac{1}{2}}_{jk} = (\delta^2_{jk} + \delta^2_{kj} - \delta^2_{jj} - \delta^2_{kk})^{\frac{1}{2}}\,. \tag{24}$$

By definition, $p_{jk} = p_{kj}$ and $p_{jj} = 0$. It is necessary for p_{jk} to be a metric, so that the triangular inequality should be satisfied,

$$p_{ij} + p_{jk} \geq p_{ik} \quad \text{for all triples} \quad (i,j,k). \tag{25}$$

When the model holds, p_{jk} should satisfy not only the metric axioms but also the requirement of Euclidean distance. Define a centering matrix $\boldsymbol{H}$ and also $\boldsymbol{B}$ in terms of $\{\delta_{jk}\}$ as

$$\boldsymbol{B}(\boldsymbol{\Delta}) = \boldsymbol{HQH} \quad \text{where} \quad \boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{11}'\,. \tag{26}$$

Using a theorem due to Young & Householder [7], we find that $\boldsymbol{B}$ is positive semidefinite with rank r if p_{jk} is Euclidean distance in r dimensions.

3 Necessary conditions

Given a dissimilarity data matrix $\boldsymbol{\Delta} = (\delta_{jk})$, we present necessary conditions for SVM to hold in r dimensions on the basis of the arguments above. We can examine whether the data satisfy the necessary conditions. Condition 1 is checked with $\boldsymbol{\Delta}$ or with $\boldsymbol{A}$ defined by (7), and condition 2 is checked with $\boldsymbol{A}$.

Condition 1: additivity

The data matrices satisfy the additivity, $\boldsymbol{\Delta}$ satisfies (12) or $\boldsymbol{A}$ does (13).

Condition 2: line pattern

Matrix $\boldsymbol{A}$ is of rank 2. The two-dimensional plot of objects in terms of its singular vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ reveals a line pattern on the plane.

We can examine whether the data involve Euclidean distance in r-dimensions, using matrices of transformed data $\boldsymbol{Q}$ given by (22), $\boldsymbol{P}$ given by (24), and $\boldsymbol{B}$ given by (26) as follows.

Condition 3: distance properties

The $\{q_{jk}\}$ *are nonnegative, equivalently,* $\boldsymbol{\Delta}$ *satisfy (23) or* $\boldsymbol{P}$ *does (25). Matrix* $\boldsymbol{B}$ *is positive semidefinite and of rank* r.

4 Sufficient condition

Let us consider sufficiency for the model to hold with $\boldsymbol{\Delta} = (\delta_{jk})$. First we assume that the data satisfy condition 3. By Young & Householder's theorem, we solve the eigenvalue problem of $\boldsymbol{B}(\boldsymbol{\Delta})$, and obtain an $n \times r$ matrix $\boldsymbol{X} = (x_{jt})$, a configuration of objects in r-dimensional Euclidean space. It is determined with the origin at the centroid, which results in (18). We retain notation (2). The r vectors $\{\boldsymbol{x}_t\}$ are linearly independent, and

$$\sum_{j=1}^{n} \boldsymbol{x}_{(j)} = \boldsymbol{X}^{'}\mathbf{1} = \mathbf{0}\,. \tag{27}$$

Denote the Euclidean distance by d_{jk} simply. From (21), we have

$$\delta_{jk}^2 + \delta_{kj}^2 = 2d_{jk}^2 + 2\gamma^2\,. \tag{28}$$

Next we assume that the data satisfy condition 1. According to a theorem [4], matrix $\boldsymbol{A}$ defined by (7) is of rank 2 and expressed as

$$\boldsymbol{A} = \boldsymbol{\eta}\mathbf{1}' - \mathbf{1}\boldsymbol{\eta}' \quad i.e. \quad a_{jk} = \eta_j - \eta_k\,, \tag{29}$$

where $\boldsymbol{\eta} = (\eta_j)$. Referring to (14), we have (30), which satisfies (31):

$$\eta_j = \frac{1}{2}(\delta_{j.}^2 - \delta_{.j}^2)\,, \tag{30}$$

$$\sum_{j=1}^{n} \eta_j = 0\,. \tag{31}$$

Let us represent $\boldsymbol{\eta}$ by a linear combination of the independent $\boldsymbol{x}_t$'s. For a purpose of exposition, we like to set

$$\boldsymbol{\eta} = 2X\boldsymbol{z} = 2(z_1\boldsymbol{x}_1 + z_2\boldsymbol{x}_2 + \cdots + z_r\boldsymbol{x}_r)\,. \tag{32}$$

It means that $\boldsymbol{\eta}$ indicates the projection of object points on a vector $\boldsymbol{z}$. Condition (31) is conformable to (27), because $0 = \mathbf{1}'\boldsymbol{\eta} = 2\,\mathbf{1}'\boldsymbol{X}\boldsymbol{z}$. To determine $\boldsymbol{z}$, we set a solution such as

$$\tilde{\boldsymbol{z}} = \frac{1}{2}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{\eta}\,, \tag{33}$$

among others. Considering that the scale unit of $\boldsymbol{z}$ is arbitrary in (32), we scale it as $\boldsymbol{z} = \beta\tilde{\boldsymbol{z}}$. Let us normalize $\boldsymbol{z}$ to meet $\boldsymbol{z}'\boldsymbol{z} = \gamma^2$, where γ is the identical value of δ_{ii} (see (3)). Then we set

$$\boldsymbol{z} = \gamma\,\tilde{\boldsymbol{z}}/\|\tilde{\boldsymbol{z}}\|\,. \tag{34}$$

From (7),(29) and (32), we obtain

$$\delta_{jk}^2 - \delta_{kj}^2 = 2(\eta_j - \eta_k) = 4\sum_{t=1}^{r} z_t(x_{jt} - x_{kt})\,. \tag{35}$$

Combining (28) with (35), we derive a representation such as

$$\begin{aligned}\delta_{jk}^2 &= d_{jk}^2 + \gamma^2 + 2\sum_{t=1}^{r} z_t(x_{jt} - x_{kt}) \qquad (36)\\ &= \sum_{t=1}^{r}(x_{jt} - x_{kt} + z_t)^2\,. \qquad (37)\end{aligned}$$

This is the slide vector model. Matrix $\boldsymbol{X}$ is determined with the translational and rotational indeterminacy. It is noted that (37) is expressed by such $\boldsymbol{X}$. Hence the representation of $\boldsymbol{z}$ in the form of 33) and (34), is not unique.

Condition 4: sufficiency

Given $\boldsymbol{\Delta}$, *it is sufficient for the model to hold that the data satisfy both condition 1 and condition 3. The representation of* $\boldsymbol{X}$ *and* $\boldsymbol{z}$ *is not unique.*

Remember that the two conditions are also necessary for the model. Therefore, *a necessary and sufficient condition* for the model to hold with a given dissimilarity matrix is that both condition 1 and condition 3 hold. If these conditions are not satisfied with data, it is not meaningful to fit SVM to the data.

5 Algebraic solution of the model

Given observations o_{jk} of model (1), one may set an error model $o_{jk} = \delta_{jk} + e_{jk}$. Fitting the model to the data, one would like to estimate $\boldsymbol{X}$ and $\boldsymbol{z}$ by optimizing a criterion. For such a purpose, Zielman & Heiser [8] suggested the iterative procedure mentioned above.

For error-free data which satisfy the sufficient condition exactly, we can derive a solution of the model parameters. Let us assume that the slide vector emanates from the origin, which is the centroid of the object configuration. In view of the rotational indeterminacy, it is not required to perform SVD of $\boldsymbol{A}$ but only to use (29). The algebraic procedure is outlined as follows.

step 1 Construct matrices $\boldsymbol{A}$ by (7) and $\boldsymbol{B}$ by (26).
step 2 Solve the eigenvalue problem of $\boldsymbol{B}$, and derive $\boldsymbol{X}$ of rank r.
step 3 Find $\boldsymbol{\eta}$ from $\boldsymbol{A}$ by (30).

step 4 Compute $\tilde{\boldsymbol{z}}$ by (33).
step 5 Determine $\boldsymbol{z}$ by 34).

For step 4, we have alternative treatments, using either (4) or (11) in stead of using (30). Using either of them yields the identical $\boldsymbol{z}$.

6 Discussion

Let us examine the assumptions put for sufficient condition, from which the algebraic solution has followed. It may be allowed to assume that the origin of the space for the judgment of dissimilarity coincides with the centroid of the object configuration. The slide vector is considered as the reference vector in the dissimilarity judgment. Then it can be assumed that the vector emanates from the centroid. This is postulated to cope with the indeterminacy of the model parameters and also to formulate the algebraic solution.

Thinking of real data, it seems to be seldom that the model holds with the data exactly. When observations include the error, even to a very small extent, the necessary conditions would not hold perfectly. In view of the possibility, the sufficient condition would not do, likewise. However, by checking the extent to which these conditions are satisfied with data, one may examine some justification for applying SVM. The algebraic solution for $\boldsymbol{X}$ and $\boldsymbol{z}$ is based on the sufficient condition. Then it tends to become unstable for error-perturbed data. Owing to the centralization (26), however, $\boldsymbol{X}$ given by step 2 is far stabler than $\boldsymbol{z}$. The algebraic solution in the closed form can be used as an initial solution for the iterative procedure.

In conclusion, we have clarified algebraic properties of SVM in new light. The results serve for theoretical examination about the model in applications, which has not been possible by model fitting work.

References

[1] Escoufier Y., Grorud A. (1980). *Analyse factorielle des matrices carrées non-symétriques.* In Diday, E. (Eds.) Data Analysis and Informatics, 263 – 276, New York: North Holland.

[2] Gower J.C. (1977). *The analysis of asymmetry and orthogonality.* In J.R. Barra et.al. (ed.), Recent Developments in Statistics, 109 – 123, North-Holland.

[3] Saito T.(1991). *Analysis of asymmetric proximity matrix by a model of distance and additive terms.* Behaviormetrika **29**, 45 – 60.

[4] Saito T. (1997). *Line structure derived from decomposition of asymmetric matrix.* Journal of the Japanese Society of Computer Statistics, **10**, 47 – 57.

[5] Saito T., Takeda S.(1990). *Multidimensional scaling for asymmetric proximities: Model and method.* Behaviormetrika **28**, 49 – 80.

[6] Weeks D.G., Bentler P.M. (1982). *Restricted multidimensional scaling model for nonsymmetric proximities.* Psychometrika **47**, 201 – 208.

[7] Young G., Householder A.S.(1938). *Discussion of a set of points in terms of their mutual distances*. Psychometrika **3**, 19 – 22.

[8] Zielman B., Heiser W.J. (1993). *Analysis of asymmetry by a slide vector model*. Psychometrika **58**, 101 – 114.

[9] Zielman B., Heiser W.J. (1996). *Models for asymmetric proximities.* British Journal of Mathematical and Statistical Psychology **49**, 127 – 146.

Address: T. Saito, Graduate School of Decision Science and Technology, Tokyo Institute of Technology, Ookayama 2-12-1, Meguroku, Tokyo 152-8552, Japan

E-mail: gsaito@valdes.titech.ac.jp

COMPSTAT'2004 Symposium

A STATISTICAL METHOD FOR MARKET SEGMENTATION USING A RESTRICTED LATENT CLASS MODEL

Naoko Sakurai, Michiko Watanabe and Kazunori Yamaguchi

Key words: Latent class model, market segmentation, customer identification.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: Along with the rapid growth in the distribution business computer infrastructure network, the ability to collect frequent shoppers program (FSP) typed data based on customer identification number is also showing a sharp upward trend. As these types of large databases increase, the current 'mass marketing' techniques will be replaced by "target marketing" based on finely defined market segmentation obtained from customer purchasing patterns. In this paper a new method for customer segmentation is proposed. The method employed is an extended version of a restricted latent class (LC) model. Data on purchases of milk products is analyzed, and customers were divided into two segments (LS and SS) based on degree of brand loyalty. The lower degree of loyalty, or SS segment, was then further divided into classes. Two calculations, one based on two SS analysis, one with two classes and the other with five, were implemented and the results compared. The response frequency of each brand was assumed to come from a mixture distribution of multinomial distributions and the distribution of the random variable for each frequency of purchasing a brand product was assumed as the Poisson distribution. The parameter estimation for the latent classes employed the EM algorithm. The analysis with five SS classes was found to be superior in terms of precision in interpretation of the data. Not only were clear patterns of preference identified, but the interpretation afforded an understanding of the motivation behind these purchasing patterns. These results show that the model tested here is a highly effective tool for segmenting consumers and analyzing purchasing patterns in precision target marketing.

1 Introduction

Along with the expansion of computer networks in general retail distribution, frequent shoppers program (FSP) data collection in the industry has gradually increased in scale. As FSP data includes a customer identification number, detailed information about each customer's purchasing tendencies can be analyzed among various products in a large database. To date, market segmentation analysis in these kinds of database, have centered on such methods as “RFM analysis” or “market basket analysis” (also called mass-marketing). These analytical methods, however, provide limited information on the motivation behind purchasing decisions. A new segmentation method, adapted to today’s diversified patterns of personal consumption, is thus urgently required. This paper proposes a new segmentation method for target marketing that analyzes purchasing patterns using a restricted latent class (LC) model.

Prior research on customer segmentation by buying behavior can be seen [2]. Grover and Srinivasan [4] has reported that customers can be classified into several latent segment classes with two purchase data for each customer. Wedel and Kamakura [9] has also reported that an LC model is effective for customer segmentation. In this paper, a restricted LC segmentation method utilizing large-scale purchase tendency data is proposed. This method allows accurate understanding of personal buying activities even for customers for which there is no such information in advance. Analogous customer groups, based on characteristic buying behavior, are analyzed as one segment of the market. All customers are placed into segment groups and classes, and it is possible to calculate each customer’s probability of belonging to each segment and class. This is extremely useful in various kinds of marketing promotion that require precise interpretation of purchasing behavior. This research has been executed based on the principles of a competitive market.

2 Market segmentation by latent class model

The following three steps are vital to LC target marketing as discussed in this paper:

- Dividing the market into segments and classes
- Analyzing the structure and the size of each segment and class
- Identifying effective target promotion methods for each segment and class

The first step is most important. Using conventional segmentation methods, based on sex, age, job and income distinctions, it is difficult to determine the correct motivation for each purchase. For example, the reasons why one customer bought a product of Brand L may include:

(1) Customer always buys only products of Brand L
(2) Customer bought products of Brand L because of low price
(3) Customer bought products of Brand L because of new product
(4) Customer buys various Brand products, including Brand L

By analyzing FSP data with customer identification, it is possible to closely follow, understand, and finely interpret personal purchase patterns. The results from this sort of analysis allow for extremely accurate and precise target marketing. This research employs the LC analysis method for segmentation of customers by purchase patterns and brand. This method illuminates the cause and effect relationships among the data and its structure. The chief characteristics of this method are as follows.

(1) Latent segments and classes composing the market are determined.
(2) Each customer is classified into a particular latent class.
(3) Characteristics of purchase patterns and heterogeneousness of each class is analyzed.

In this paper, the number of brands is defined as n and two kinds of segments; the Brand Loyal Segment (LS) and Brand Switching Segment (SS), are constructed. Customers in the LS class are loyal to one particular band, and never purchase other brands. SS customers show a probability of selecting among various brands. This type of segmentation allows each segment to be characterized by the brands which the customers have a high probability of selecting. In addition, both analysis of competition among brands, and relational analysis of customers' brand tastes and attributes can be implemented.

3 Explanation of data in analysis

3.1 Description of data

The data used in this research was supplied by the Japanese Data Analysis Competition 2001. The data was composed of 11,641 individual customer's purchases at an intermediate scaled supermarket located in southern of Japan, from April to October in 2000. The data included customer identification number, product code, number of items purchased, date, time (for most data) and total amount of purchase. Dairy products were selected for analysis because of their daily consumption patterns.

3.2 Brand setting and objective variables for analysis

The milk products were divided into five brand categories, A, B, C, D, E, based on maker and type of product, as shown in Table 1. The brand categories were selected from the FSP data, but categories that showed less than four purchases per month were excluded. The final total objective customer identification data for analysis was 3246, which is arranged into the format shown in Table 2. The number in each cell represents each customer's total number of purchases of that milk product during the research period. In this paper this total number of purchases has been taken as a random variable.

4 Model

Latent class modelling was introduced by Lazarsfeld and Henry [6] in dichotomous survey data. In LC model analysis it is assumed that latent variables

brand	explanation & products group
A	Major brand (general)
B	Major brand (functional; low fat or enriched)
C	Local brand (general) (low price)
D	Second major brand (functional; low fat or enriched)
E	Local brand (high quality) (high price)

Table 1: Milk brands.

customer ID	A	B	C	D	E
00000001	57	0	6	5	31
00000002	2	0	97	0	0
00000003	1	0	98	0	0
00000004	0	0	99	0	0
00000005	3	0	95	0	0
00000006	3	0	90	0	5
00000007	1	0	87	10	0
00000008	0	0	98	0	0

Table 2: Objective data for analysis.

are categorical [8]. The initial methodology of analysis was established by Goodman [3]. Since then, various research related to LC models has been implemented [7], [1]. As mentioned above, Grover & Srinivasan [4] have analysed two times purchase data of coffee with a LC model. In this paper, an extended version of the restricted model developed by Grover & Srinivasan, is employed. The formula, which denotes a mixture distribution model, is shown in (1) below

$$P(X_1, X_2, \ldots, X_n|S) = \sum_t V_t P_t(X_1, X_2, \ldots, X_n|S) \tag{1}$$

where S denotes the total number of purchases by each customer and t denotes the total number of segment classes.

$$X_1, X_2, \ldots, X_n$$

denote the number of purchases of each brand product by one customer, where n is the number of brands. V denotes the ratio of each segment class, as shown in Table 3, which introduces a hypothetical market segmentation analysis using this LC model, assuming P_t as the probability model and m as the number of switching segment classes. When X_j is a random variable denoting buying frequency of brand j, each X_j follows the Poisson distribution. This model assumes that random variables inside each segment class are mutually independent. The joint probability distribution can thus be represented as

$$P_t = \prod_{j=1}^{n} \exp(-S_j) S_j^{x_j} / x_j!$$

where $S_j = S \cdot P_{j,t}$.

For the LS case, the following are assumed;

$$P_{j,j} = 1, \quad P_{j,t} = 0 \quad \text{where} \quad j \neq t \quad \text{(See Table 3).}$$

	LS: Loyal Seqment				SS: Switching Seqment			
	1	2	...	n	$n+1$	$n+2$	...	$n+m$
ratio	v_1	v_2		v_n	v_{n+1}	v_{n+2}	...	v_{n+m}
brand 1	1	0		0	$P_{1,1}$	$P_{1,2}$		$P_{1,m}$
brand 2	0	1		0	$P_{2,1}$	$P_{2,2}$		$P_{2,m}$
...			...				...	
brand n	0	0		1	$P_{n,1}$	$P_{n,2}$		$P_{n,m}$

Table 3: Hypothetical segmentation class analysis.

5 Estimation

The Maximum likelihood method with the EM algorithm is used in parameter estimation for this model. The calculation uses a log-linear model [5]. The log likelihood of the hypothetic complete data has been assumed as in formula (2),

$$\log L = \sum_{k=1}^{r} \{-s_k P_j + x_{kj} \log s_k P_j\} + C = \sum_{k=1}^{r} s_k P_j + \sum_{k=1}^{r} x_{kj} \log s_k P_j + C' \tag{2}$$

where k denotes customer ID number and j denotes the number of brands. Following (2), the maximum likelihood estimator can be calculated from formula (3).

$$\hat{P}_j = \sum_{k=1}^{r} x_{kj} / \sum_{k=1}^{r} s_k \tag{3}$$

(3) represent an M step of the EM algorithm. On the other hand, in an E step, where the expectation value is calculated, the probability of each customer belonging to each segment would not be available. When observed variables are given, each probability can be calculated by the formula (4),

$$\begin{aligned} P(Z = t | X_1, \ldots, X_n) &= P(Z = t) P(X_1, \ldots, X_n | Z = t) / \\ &\quad / \sum_{l} P(Z = l) P(X_1, \ldots, X_n | Z = l) \end{aligned} \tag{4}$$

where $P(Z = t)$ denotes the size of each class $P(X_1, X_2, \ldots, X_n|t)$ is easily calculated based on the principle of local independence, as a product of the probability of a positive reply for each class.

6 Results

Table 4 and 5 represent analytical results for two different numbers (two in Table 4 and five in Table 5) of SS segment classes. A, B, C, D, E represent the product brands defined in Table 1.

	LS					SS	
	1	2	3	4	5	1	2
ration	0.0267	0.0012	0.3355	0.0080	0.0034	0.3631	0.2620
A	1	0	0	0	0	0.0665	0.4436
B	0	1	0	0	0	0.0291	0.0106
C	0	0	1	0	0	0.8679	0.2978
D	0	0	0	1	0	0.0315	0.1766
E	0	0	0	0	1	0.0051	0.0713

Table 4: Analytical results with two class model.

	LS				
	1	2	3	4	5
ration	0.0238	0.0012	0.2660	0.0079	0.0034
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1
	SS				
	1	2	3	4	5
ration	0.2647	0.1131	0.0433	0.1773	0.0994
A	0.0174	0.7379	0.1684	0.2193	0.1242
B	0.0146	0.0059	0.0072	0.0443	0.0191
C	0.9473	0.2161	0.3153	0.7185	0.3830
D	0.0179	0.0291	0.1019	0.0156	0.4668
E	0.0028	0.0109	0.4071	0.0024	0.0069

Table 5: Analytical results with five class model.

Interpreting the two-class model in Table 4, it can be seen that in class SS 1, Brand C is by far and away the most popular product. In SS 2, Brand A scored highest, followed by Brands C and D. As the number of classes in SS is increased , the data for each class gains precision. Table 6 shows a reasonable interpretation of the results of Table 5. As can be seen, the model allows for precise interpretation of the purchasing patterns shown by the customers in each class. Although the number of SS 3 customers is low, the data clearly indicates their extremely strong preference for expensive, high quality products. SS5 customers, on the other hand, have a clear preference for specialized products such as low fat or enriched calcium or protein, but also frequently purchase the less expensive local brand. This data clearly shows that the larger the number of SS classes, the more precise the interpretation.

	Explanation	*Interpretation*
SS1	about 95% buy Brand C	Prefers local lower price brand
SS2	about 74% buy Brand A and 22% buy Brand C	Prefers major brand & sometimes local
SS3	small group but 41% buy Brand E	Strongly prefers high quality product
SS4	opposite pattern of SS2	Prefers local brand & sometimes major
SS5	about 47% buy Brand D	Prefers major brand with function

Table 6: Interpretation of five class model.

7 Conclusion

This analysis assumes that a particular set of consumers is composed of some individuals with strong brand loyalty, and others with weaker loyalty. It is also assumed that those with weaker loyalty will show characteristic purchasing patterns with regard to price, quality and functionality of competing products. The results indicate that increasing the number of classes in the lower loyalty, or SS segment, produces greater precision in interpretation of the analysis. Each class was shown to be characterized by homogeneous and clearly identifiable purchasing patterns, and the interpretation, based on the attributes of the competing brands, provided an understanding of the motivation behind the purchasing patterns. In some cases, further increasing the number of classes may produce even more precise interpretation. The AIC or BIC can be employed in determining the ideal number of SS classes for a particular set of data. Caution, however, must be exercised in choosing the number of classes. In this research case, for example, use of the larger number of class would have confused the interpretation. In conclusion, the results shown here clearly indicate that an extended version of the restricted latent class model is a highly effective tool for segmenting and analyzing a static customer database such as FSP.

References

[1] Agresti A. (2002). *Categorical data analysis.* Second Edition. New York: Wiley.

[2] Goodhardt G.J., Ehrenberg A.S.C., Chatfield C. (1984). *The Dirichlet: A comprehensive model of buying behaviour.* J. of Royal Statistical Association A **147**, Part 5, 621 – 655.

[3] Goodman L.A. (1974). *Exploratory latent structure analysis using both identifiable and unidentifiable models.* Biometrika **61**, 215 – 231.

[4] Grover R., Srinivasan V. (1987). *A simultaneous approach to market segmentation and market structuring.* J. of Marketing Research **24**, 139 – 153.

[5] Hagenaars J.A. (1993). *Loglinear models with latent variables.* Series of Quantitative Applications in the Social Sciences **94**, Thousand Oakes, CA: Sage Publications.

[6] Lazarsfeld P.F., Henry N.W. (1968). *Latent structure analysi.* Boston: Houghton Mufflin.

[7] McCutcheon A.L. (1987). *Latent class analysis.* Series of Quantitative Applications in the Social Sciences **64**, Thousand Oakes, CA: Sage Publications.

[8] Vermunt J.K., Magidson J. (2000). *Latent class cluster analysis.* Hagenaars, J. A., McCutcheon, A.L. (eds.), Applied Latent Class Analysis, 89 – 106. Cambridge, UK: Cambridge University Press.

[9] Wedel M., Kamakura W.A. (1998). *Market segmentation: Concepts and methodological foundations.* Boston: Kluwer Academic Publishers.

Acknowledgement: We sincerely thank the Japanese Data Analysis Competition 2001 for supplying the FSP data used in this research.

Address: N. Sakurai, Tokyo University of Information Sciences, Wakaba-ku, Chiba, 265-8501, Japan, M. Watanabe, Toyo University, Bunkyo-ku, Tokyo, 112-8606, Japan, K. Yamaguchi, Rikkyo University, Toshima-ku, Tokyo, 171-8501, Japan

E-mail: `sakurai@rsch.tuis.ac.jp`, `michiko_watanabe@nifty.com`, `kyamagu@rikkyo.ac.jp`

A MIXTURE MODEL APPROACH FOR ON-LINE CLUSTERING

A. Samé, C. Ambroise and G. Govaert

Key words: EM algorithm, stochastic gradient.

COMPSTAT 2004 section: Clustering.

Abstract: This article presents an original on-line algorithm dedicated to mixture model based clustering. The proposed algorithm is a stochastic gradient ascent which maximizes the expectation of the classification likelihood. This approach requires few calculations and exhibits a quick convergence. A strategy for choosing the optimal number of classes using the Integrated Classification Likelihood (ICL) is studied using simulated data. The results of the simulations show that the proposed method provides a fast and accurate estimation of the parameters (including the number of classes) when the mixture components are relatively well separated.

1 Introduction

Generally, stochastic gradient algorithms are used for on-line parameter estimation in signal processing and pattern recognition for their algorithmic simplicity. They have been shown to be faster than standard algorithms. In clustering, MacQueen on-line kmeans algorithm [6] is the one commonly used.

In the context of a flaw detection problem using acoustic emission, we have been brought to classify under real time constraints a set of points located in a plane. The solution provided by the so-called CEM algorithm [4] applied using a gaussian mixture model provides a satisfactory solution for this problem and is faster than the EM algorithm [5] one's. However, in spite of its speed, CEM algorithm is not able to react in real time when the number of acoustic emissions becomes too large (more than 10000 points). In this work, we aim to develop an on-line mixture model based clustering algorithm which also allows us to choose the appropriate number of classes.

Let us suppose that data are independent observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \dots$ which are sequentially received and distributed following a mixture density of K components, defined on $\mathbb{R}^p$ by

$$f(\boldsymbol{x}; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}; \boldsymbol{\theta}_k),$$

with $\boldsymbol{\Phi} = (\pi_1, \dots, \pi_K, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ where $\pi_1, \dots, \pi_K$ denote the proportions of the mixture and $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ the parameters of each density component. We denote by $z_1, \dots, z_n, \dots$ the classes associated to the observations, where $z_n \in \{1, \dots, K\}$ corresponds to the class of $\boldsymbol{x}_n$.

To estimate the parameter $\mathbf{\Phi}$, we choose to use a stochastic gradient algorithm. These algorithms generally allow to optimize the expectation of a criterion [2], [3]

$$C(\mathbf{\Phi}) = E\left[J(\boldsymbol{x}, \mathbf{\Phi})\right],$$

where the expectation is computed using the unknown true parameter of the distribution function f. The criterion $J(\boldsymbol{x}, \mathbf{\Phi})$ measures the quality of the parameter $\mathbf{\Phi}$ given the observation $\boldsymbol{x}$. The stochastic gradient algorithm aiming to maximize the criterion C is then written

$$\mathbf{\Phi}^{(n+1)} = \mathbf{\Phi}^{(n)} + \alpha_n \nabla_{\mathbf{\Phi}} J(\boldsymbol{x}_{n+1}, \mathbf{\Phi}^{(n)}) \qquad (1)$$

where the learning rate α_n is a positive scalar or a positive definite matrix such that $\sum |\alpha_n| = \infty$ and $\sum |\alpha_n|^2 < \infty$.

In the second section, we present the Titterington on-line clustering approach [7]; the third section is devoted to the new stochastic gradient algorithm that we propose for on-line clustering; an experimental study is summarized in the fourth section.

2 Stochastic gradient algorithm derived from EM algorithm

Given the observed data $\mathbf{x}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and some initial parameter $\mathbf{\Phi}^{(0)}$, the standard EM algorithm [5] maximizes the log-likelihood $\log p(\mathbf{x}_n; \mathbf{\Phi})$ by maximizing iteratively the expectation of the complete data conditionally to the available data:

$$Q(\mathbf{\Phi}, \mathbf{\Phi}^{(q)}) = E[\log p(\mathbf{x}_n, \mathbf{z}_n; \mathbf{\Phi}) | \mathbf{x}_n, \mathbf{\Phi}^{(q)}]$$

where $\mathbf{z}_n = (z_1, \ldots, z_n)$.

Titterington on-line clustering approach [7] consists in using a special stochastic gradient algorithm which can be derived from the standard EM algorithm. For this purpose, we define in the same way as for the EM algorithm the quantity

$$Q_{\mathbf{x}_{n+1}}(\mathbf{\Phi}, \mathbf{\Phi}^{(n)}) = E[\log p(\mathbf{x}_{n+1}, \mathbf{z}_{n+1}; \Phi) | \mathbf{x}_{n+1}, \mathbf{\Phi}^{(n)}],$$

where, this time, the parameter $\mathbf{\Phi}^{(n)}$ has been computed from the observations $\mathbf{x}_n$. The maximization of $1/(n+1) Q_{\mathbf{x}_{n+1}}(\,\cdot\,, \mathbf{\Phi}^{(n)})$ using Newton method after replacing the hessian matrix term by its expectation which is the Fisher information matrix $I_c(\mathbf{\Phi}^{(n)})$ associated to one complete observation $(\boldsymbol{x}, z)$ results in the algorithm proposed by Titterington:

$$\mathbf{\Phi}^{(n+1)} = \mathbf{\Phi}^{(n)} + \frac{1}{n+1} \left(I_c(\mathbf{\Phi}^{(n)})\right)^{-1} \frac{\partial \log f(\boldsymbol{x}_{n+1}; \mathbf{\Phi}^{(n)})}{\partial \mathbf{\Phi}}. \qquad (2)$$

Fisher information matrix $I_c(\mathbf{\Phi}^{(n)})$ is positive definite for some density families like the regular exponential family and thus Titterington algorithm has the general form (1) of the stochastic gradient algorithms; which guarantees under some conditions [2], [3] that the criterion maximized by (2) is $E[\log p(\boldsymbol{x}; \mathbf{\Phi})]$.

3 Stochastic gradient algorithm derived from CEM algorithm

The criterion to be maximized in this section, by analogy with the classification likelihood maximized in the CEM algorithm [4] which can also be written $L_C(\mathbf{\Phi}) = \max_{z_1,\ldots,z_n} \log p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n, z_1,\ldots,z_n; \mathbf{\Phi})$, is the expected criterion

$$C(\mathbf{\Phi}) = E[\max_{1\leq z\leq K} \log p(\boldsymbol{x}, z; \mathbf{\Phi})],$$

where $\log p(\boldsymbol{x}, z; \mathbf{\Phi})$ is the complete log-likelihood of the parameter $\mathbf{\Phi}$ given the complete observation $(\boldsymbol{x}, z)$.

The application of algorithm (1) needs the gradient of the function $J(\boldsymbol{x}, \mathbf{\Phi}) = \max_{1\leq z\leq K} \log p(\boldsymbol{x}, z; \mathbf{\Phi})$ with respect to $\mathbf{\Phi}$ to be computed. However, this gradient does not exist for some values of $\boldsymbol{x}$ due to the well known non differentiability of the max function. In this situation which is very common, Bottou [2, 3] shows that it is sufficient to replace this gradient by a function $H(\boldsymbol{x}, \mathbf{\Phi})$ verifying $E[H(\boldsymbol{x}, \mathbf{\Phi})] = \nabla_{\mathbf{\Phi}} C(\mathbf{\Phi})$ on the one hand, and on the other hand that the functions $H(\boldsymbol{x}, \mathbf{\Phi})$ and $C(\mathbf{\Phi})$ verify certains conditions.

In the gaussian mixture case, we may consider the function $H(\boldsymbol{x}, \mathbf{\Phi})$ such that

$$H(\boldsymbol{x}, \mathbf{\Phi}) = \begin{cases} \nabla_{\mathbf{\Phi}} J(\boldsymbol{x}, \mathbf{\Phi}) & \text{if } \nabla_{\mathbf{\Phi}} J(\boldsymbol{x}, \mathbf{\Phi}) \text{ exists} \\ 0 & \text{otherwise.} \end{cases}$$

The parameters to be updated in the gaussian mixture case are both the proportions $\pi_1,\ldots,\pi_k$ and the parameters $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ of each gaussian distribution of the mixture. The direct updating rule (1) applied to the proportions does not guarantee in practice that $0 < \pi_k < 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Therefore to overcome this numerical instability, we use a logit parametrization [8] $w_k = \log(\pi_k/\pi_K)$. The resulting new variables $w_1,\ldots,w_{K-1}$ belong to $I\!R$. Finally the CEM stochastic gradient algorithm in the gaussian mixture model case can be defined as follows:

Step 0 initialization of the parameters $\pi_k^{(0)}$, $\boldsymbol{\mu}_k^{(0)}$ and $\mathbf{\Sigma}_k^{(0)}$

Step 1 (iteration $n+1$) assignation of the new observation $\boldsymbol{x}_{n+1}$ to the class k^* which maximizes the log-likelihood of the current parameter knowing this observation:

$$\begin{aligned} k^* &= \arg\max_{1\le k\le K}\Big(\log \pi_k^{(n)} - \\ &\frac{1}{2}\log\det(\boldsymbol{\Sigma}_k^{(n)}) - \frac{1}{2}(\boldsymbol{x}_{n+1}-\boldsymbol{\mu}_k^{(n)})^T{\boldsymbol{\Sigma}_k^{(n)}}^{-1}(\boldsymbol{x}_{n+1}-\boldsymbol{\mu}_k^{(n)})\Big)\end{aligned}$$

Step 2 (iteration $n+1$) updating of the parameters:

$$\begin{aligned} w_k^{(n+1)} &= w_k^{(n)} + \alpha_n(z_{n+1,k}-\pi_k^{(n)}) \quad for\ k=1,\ldots,K-1 \\ \pi_k^{(n+1)} &= \frac{\exp(w_k^{(n+1)})}{1+\sum_{\ell=1}^{K-1}\exp(w_\ell^{(n+1)})} \quad for\ k=1,\ldots,K-1 \\ \pi_K^{(n+1)} &= \frac{1}{1+\sum_{\ell=1}^{K-1}\exp(w_\ell^{(n+1)})} \\ \boldsymbol{\mu}_k^{(n+1)} &= \boldsymbol{\mu}_k^{(n)} + z_{n+1,k}\ \alpha_n{\boldsymbol{\Sigma}_k^{(n)}}^{-1}(\boldsymbol{x}_{n+1}-\boldsymbol{\mu}_k^{(n)}) \\ \boldsymbol{\Sigma}_k^{(n+1)} &= \boldsymbol{\Sigma}_k^{(n)} + z_{n+1,k}\ \alpha_n\ \cdot \\ &\Big(\frac{1}{2}{\boldsymbol{\Sigma}_k^{(n)}}^{-1}\big((\boldsymbol{x}_{n+1}-\boldsymbol{\mu}_k^{(n)})(\boldsymbol{x}_{n+1}-\boldsymbol{\mu}_k^{(n)})^T{\boldsymbol{\Sigma}_k^{(n)}}^{-1} - I\big)\Big)\end{aligned}$$

where $z_{n+1,k}$ equals 0 if $k=k^*$ and 1 otherwise.

Particularly, an algorithm equivalent to MacQueen on-line kmeans algorithm [6] can be recovered if we consider a gaussian mixture with identical proportions and spherical covariance matrices (equal to the identity matrix) with a learning rate $\alpha_n = 1/(n+1)$.

The proposed method for the choice of the number of classes consists in running the CEM stochastic gradient algorithm concurrently for models from 2 to K_{max} number of clusters and selecting the solution which maximizes the integrated classification likelihood criterion (ICL) proposed by Biernacki, Celeux and Govaert [1]. In our situation, the ICL criterion can be written as

$$ICL(m,K) = \log p(\boldsymbol{\Phi}^{(n)};\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n,z_1,\ldots,z_n) - \frac{\nu_{m,K}}{2}\log(n),$$

where $\boldsymbol{\Phi}^{(n)}$ is the parameter vector obtained at iteration n with the stochastic gradient algorithm, $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ the data available at time n (or at iteration n), $z_1,\ldots,z_n$ the corresponding classes computed by applying the *maximum a posteriori* rule with the parameter $\boldsymbol{\Phi}^{(n)}$ and $\nu_{m,K}$ the number of free parameters of the model. This approach is possible because few calculations are required by the CEM stochastic gradient algorithm and this allows us to compare different runs.

4 Simulations

The adopted strategy for simulations consists in initially drawing n observations according to a mixture of two bi-dimensional gaussian distribution, to apply the standard CEM algorithm on a few points (n_0 points) and finally to apply the CEM stochastic gradient algorithm on the rest of the points. The main parameters which control the simulations are:

- the samples sizes: $n = 100, n = 300, n = 500, n = 1000, n = 3000,$ $n = 5000$;
- the number n_0 of points initially processed with the CEM algorithm: $n_0 = 80$;
- the number of components of the mixture: $K_0 = 4$;
- the overlapping degree between the components of the mixture measured by the theoretical percentage of misclassified points which varies as a function of the distance between the class centers $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$, $\boldsymbol{\mu}_4$; the retained overlapping degrees are:
  - 5% of theoretical error: $\boldsymbol{\mu}_1 = (0; 0), \boldsymbol{\mu}_2 = (4; 0), \boldsymbol{\mu}_3 = (0; 4),$ $\boldsymbol{\mu}_4 = (4; 4)$
  - 14% of theoretical error: $\boldsymbol{\mu}_1 = (0; 0), \boldsymbol{\mu}_2 = (2.5; 0), \boldsymbol{\mu}_3 = (0; 2.5),$ $\boldsymbol{\mu}_4 = (2.5; 2.5)$
  - 20% of theoretical error: $\boldsymbol{\mu}_1 = (0; 0), \boldsymbol{\mu}_2 = (2.2; 0), \boldsymbol{\mu}_3 = (0; 2.2),$ $\boldsymbol{\mu}_4 = (2.2; 2.2)$
- the mixture proportions chosen equal: $\pi_k = 1/4$ for $k = 1, \ldots, 4$;
- the variance matrices fixed to the identity matrix.

The usual learning rate in stochastic approximation, $\alpha_n = 1/(an)$ $(a > 0)$, has been chosen in these simulations. By varying values of a, we observe that the algorithm perform better for values of a between 0,1 and 0,6. So we choose the leaning rate $\alpha_n = 1/(0.3n)$ in the current simulations.

Table 1 presents ICL criterion as a function of the number K of classes taken into account by the CEM stochastic gradient algorithm and the sample size n, for an overlap leading to 14% of theoretical error. We observe in this situation that the number of classes found by our method, that is the one for which the ICL criterion is greater, corresponds to the true simulated number of clusters which is 4 clusters. This observed behavior is the same, even for small values of $n(100, 300)$. The situation corresponding to 5% of theoretical error gives also good results. However, the true number of classes is not recovered in the situation leading to 20% of theoretical error (see table 2), even for the relatively large sample sizes $n(3000, 5000)$. This behavior is not surprising because CEM algorithm is known to provide biased estimations when the classes are not well separated.

	$n = 100$	$n = 300$	$n = 500$	$n = 1000$	$n = 3000$	$n = 5000$
$K = 2$	-0.0447	-0.1303	-0.2135	-0.4230	-1.2705	-2.1175
$K = 3$	-0.0440	-0.1275	-0.2115	-0.4190	-1.2570	-2.0925
$K = 4$	**-0.0437**	**-0.1250**	**-0.2052**	**-0.4065**	**-1.2135**	**-2.0175**
$K = 5$	-0.0459	-0.1298	-0.2115	-0.4185	-1.2465	-2.0600
$K = 6$	-0.0464	-0.1366	-0.2153	-0.4255	-1.2870	-2.1575
$K = 7$	-0.0487	-0.1335	-0.2180	-0.4325	-1.3200	-2.1175

Table 1: ICL criteria (divided by 10^4) as a function of the number of classes K and the sample sizes n, for an overlapping leading to 14% of theoretical error.

	$n = 100$	$n = 300$	$n = 500$	$n = 1000$	$n = 3000$	$n = 5000$
$K = 2$	**-0.0415**	**-0.1205**	**-0.2023**	-0.4050	-1.2075	-2.0200
$K = 3$	-0.0418	-0.1208	-0.2033	**-0.4045**	**-1.2015**	-2.0025
$K = 4$	-0.0430	-0.1263	-0.2122	-0.4165	-1.2480	-2.0950
$K = 5$	-0.0422	-0.1209	-0.2035	-0.4015	-1.2045	**-1.9975**
$K = 6$	-0.0457	-0.1266	-0.2120	-0.4155	-1.2180	-2.0900
$K = 7$	-0.0449	-0.1286	-0.2077	-0.4195	-1.2900	-2.0300

Table 2: ICL criteria (divided by 10^4) as a function of the number of classes K and the sample sizes n, for an overlapping leading to 20% of theoretical error.

5 Conclusion

This paper proposes an on-line estimation of a mixture model parameters. The proposed stochastic gradient algorithm is a generalization of the on-line kmeans algorithm introduced by MacQueen [6]. The few required computations allows several models to be estimated concurrently, defining an inexpensive strategy of model choice.

Although the proposed method provides reasonably good results, the convergence analysis of the CEM stochastic gradient algorithm toward a local maxima of the expected classification likelihood $E[\max_{1\leq z\leq K} \log p(\boldsymbol{x}, z; \boldsymbol{\Phi})]$ is met only under some conditions [2], [3] which are often difficult to prove. The verification of these conditions at least for some particular models remains the main prospect of this work.

References

[1] Biernacki C., Celeux G., Govaert G. (2000). *Assessing a mixture model for clustering with the integrated completed likelihood.* IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (7), 719−725.

[2] Bottou L. (1991). *Une approche théorique de l'apprentissage connexioniste; applications à la reconnaissance de la parole.* Thèse de doctorat, université d'Orsay.

[3] Bottou L. (1998). *Online learning and stochastic approximations.* In online learning in neural networks, D. Saad (ed.), Cambridge: Cambridge University Press.

[4] Celeux G., Govaert G. (1992). *A classification EM algorithm for clustering and two stochastic versions.* Computation Statistics and Data Analysis **14**, 315 – 332.

[5] Dempster A.P., Laird N.M., Rubin D.B. (1977). *Maximum likelihood from incomplete data via the EM algorithm (with discussion).* J. Royal Stat. Soc. B **39** (1), 1 – 38.

[6] MacQueen J. (1967). *Some methods for classification and analysis of multivariate observations.* In Proceedings of 5th Berkeley Symposium on Mathematics, Statistics and Probability **1**, 281 – 298.

[7] Titterington D.M. (1984). *Recursive parameter estimation using incomplete data.* J. Royal Statist. Soc. Series B **46**, 257 – 267.

[8] Yao J. F. (2000). *On recursive estimation of incomplete data models.* Statistics **34** (1), 27 – 51.

Address: A. Samé, C. Ambroise, G. Govaert, Université de Technologie de Compiègne, HEUDIASYC, UMR CNRS 6599, BP 20529, 60205 Compiègne Cedex, France

E-mail: {same,ambroise,govaert}@utc.fr

COMPSTAT'2004 Symposium

EXPERIMENTAL STUDY OF LEAF CONFIDENCES FOR RANDOM FORESTS

Petr Savický and Emil Kotrč

Key words: Decision trees, random forests, weights, leaf confidences.

COMPSTAT 2004 section: Neural networks and machine learning.

Abstract: Decision forests (ensembles of trees) achieve usually smaller generalization error compared to single trees. In the classical methods for growing forests, bagging and boosting, the individual trees are constructed by methods originally developed for growing a single tree as the final predictor. In particular, the trees are usually pruned. For such trees, using weights (confidences) for individual trees improves the accuracy of the prediction of the ensemble.

Random forests technique [4] uses a specific tree growing process, which does not produce good individual trees, but the whole ensemble frequently achieves better results than ensembles of trees obtained by classical bagging and boosting. One of the default features of Random Forests technique is that it does not use any weights. The current paper presents experiments demonstrating that in specific situations, appropriately chosen weights may improve the prediction for Random Forests of limited size.

1 Introduction

This paper is concerned with *two-class* classification problems and with the analysis of *Random Forests* [4] technique for constructing classifiers using a training data set. We describe a statistical model for simulation of the classification of Random Forests (RF) and we use it to analyze modifications of RF using weights based on *leaf levels of confidence.* In the situation with an overlap between positive and negative region in the predictor space, these modifications improve the prediction if the probability that a negative case is accepted is required to be very small.

Assume a domain space $X = \mathrm{R}^n$, where n is the number of numerical predictors, and the class label set $C = \{0, 1\}$. Our aim is to build a classifier which assigns a class (i.e. label 0 or 1) to cases from the space X. Formally, a classifier is a function $h : X \to C$. Such classifier h is constructed using a training set $L = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in X$ is a case and $y_i \in C$ indicates its class. Learning with a training set is also called *supervised learning.* Accuracy of the classifier is then estimated using a testing set $K = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ containing cases with the known classification.

Random forest is a collection of *decision trees.* Decision tree is a rooted tree whose leaves are assigned class labels, possibly with a level of confidence in the label. Each internal node is labeled by a test. Since we consider only problems with numerical predictors, the tests have the form $X_k \leq a$, where

X_k is one of the predictors and a is a threshold. When classifying a new case, the computation starts at the root. On each internal node, the result of the test determines whether the computation continues to the left or right subtree. The label finally reached determines the prediction of the tree.

2 Ensembles of trees

Ensemble of predictors (also a committee) is a collection of predictors for the same problem, which vote for classification of test (unseen) cases. Decision forest is an ensemble of trees and the rule for combining predictions of the trees. In order to construct a forest it is necessary to build several sufficiently different trees using the given learning data. There is a number of different methods for this, which include subsampling of the data (bagging [3]) and iterative reweighting of learning cases (adaptive boosting [5]).

Random Forests technique [4] uses bagging with replacement to produce different random samples of the same size as the original data. On each of these samples, a tree is constructed. At each node, a small group of input variables (predictors) to split on is selected at random and among these variables, the best variable according to the impurity of the node criterion (see [2]) is used for the split. Random Forests grow trees to maximum size, to get pure leaves on the subsample (leaves contains cases from one class only), and it does not prune them.

When used for classification, the predictions of the trees are summed up with equal weights and compared to a threshold. The default threshold is one half of the number of the trees, which produces the usual majority vote.

Let us consider a general scheme of combining the votes of individual trees based on assigning a level of confidence to each leaf. The *leaf level of confidence* is a real number, which combines the information contained in a class label together with a measure of confidence in the label. Positive values mean a preference for class 1, negative values mean a preference for class 0. The higher the absolute value, the stricter the preference. Pure class assignment may be represented by levels of confidence equal to ± 1.

Assume a decision forest for a two class problem consisting of N trees $T_1, \ldots, T_N$. For each leaf v of any of the trees, let $c(v)$ be its leaf level of confidence. For a case x, let $T_j(x)$ be the leaf of T_j, which is reached by x. The classification of new cases is obtained using the function

$$F_N(x) = \sum_{j=1}^{N} c(T_j(x))$$

and a threshold t. For a given value of t, the classification of a new case x is

$$G_{N,t}(x) = \begin{cases} 1 & \text{if } F_N(x) \geq t \\ 0 & \text{otherwise} \end{cases}$$

Leaf levels of confidence are used in boosting by default [7], [8]. They also help e.g. for bagging using trees constructed by CART methodology [1].

3 Suggestions for leaf confidences in Random Forests

Let us consider the leaf levels of confidence derived from the frequencies of classes in the leaf obtained by resubstituting the training data into the tree. Let us denote the number of training cases of class 1, 0, resp. in a leaf v by $pos(v)$, $neg(v)$, resp. The leaf levels of confidence used in the current paper have the form $c(v) = w(pos(v), neg(v))$, where $w : \mathcal{N} \times \mathcal{N} \to \mathcal{R}$ is an appropriate function.

RF technique constructs trees with pure leaves on random subsamples, but the leaves may not be pure on the whole learning set. The original voting scheme of RF cannot be exactly represented by leaf levels of confidence as described above, since RF assigns classes to each leaf only on the basis of the subsample used for tree growing and the result on this subsample may not be determined from $pos(v), neg(v)$. However, the default voting in RF may be closely approximated by using the level of confidence $rf(pos, neg) \overset{\text{def}}{=} \text{sign}(pos - neg)$. In our experiments, the results of the original RF are close to the result obtained using this level of confidence.

Let us start with confidences for weak predictors derived for boosting using minimization of exponential loss function, see [8]. This may be used to derive levels of confidence for the whole trees as well as for individual leaves. For leaf confidences, this yields several possible methods. The simplest of them depends only on $pos(v), neg(v)$. Augmented by a smoothing parameter ε, it is given as follows.

$$q_{(\varepsilon)}(pos, neg) \overset{\text{def}}{=} \frac{1}{2} \ln \frac{pos + \varepsilon}{neg + \varepsilon}$$

Since the leaf level of confidence $q_{(\varepsilon)}$ did not lead to satisfactory results, we tried several other possibilities, the most successful of which are now described.

Let us call the first of them *normalized difference* with parameters α and h, which is defined as

$$nd_{(\alpha,h)}(pos, neg) \overset{\text{def}}{=} \frac{pos - neg - h}{(pos + neg)^{\alpha}}$$

The simple rf level of confidence above may be approximated by a continuous function with a similar behavior. The level of confidence $\sigma_{(k)}(pos, neg)$ uses the sigmoidal function $s(x) = (1 - e^{-x})/(1 + e^{-x})$ with values in $(-1, 1)$ as follows

$$\sigma_{(k)}(pos, neg) \overset{\text{def}}{=} s\left(\frac{pos - neg}{k \cdot (pos + neg)}\right)$$

4 Statistical model of classification in RF

Constructing a forest is a random process. The aim of our experiment is to obtain estimate of the average behaviour of the forests of a given size expressed as an ROC (Receiver Operator Characteristic) curve. Instead of constructing several forests of the considered size and averaging their ROC curves, we adopted a statistical model of a forest of N trees based on the fact that the distribution of individual trees in the forest does not depend on the size of the forest.

Let $\tilde{T}$ be a random variable representing a single random tree in the forest. The properties of the distribution of $\tilde{T}$ were estimated on the basis of a sample of 500 independent random trees constructed as a single forest on the training set L.

Let K_0, and K_1, resp. be the negative, and positive cases resp. from the test set K. The ROC curve for a given forest of size N is the curve containing for each threshold t the point whose coordinates are the sample estimates of the conditional expected values $\mathrm{E}(G_{N,t}(x)|y)$ obtained on the test set K given by

$$\left[\frac{1}{|K_0|}\sum_{x_i\in K_0} G_{N,t}(x_i), \frac{1}{|K_1|}\sum_{x_i\in K_1} G_{N,t}(x_i)\right]. \tag{1}$$

For a fixed t, we estimate the expected value of the point (1) taken over the distribution $\mathcal{F}$ of the random forests of size N as follows. The expected value of the average of $G_{N,t}(x_i)$ over x_i is computed as an average over x_i of expected values of $G_{N,t}(x_i)$. For each point $x_i \in K$, each given leaf level of confidence function $c()$, and a fixed size of forest N, an estimate of $\mathrm{E}_{\mathcal{F}}[G_{N,t}(x_i)]$ is calculated using the normal approximation $\mathrm{N}(N\cdot\mu_i, N\cdot\sigma_i^2)$ of the distribution of $F_N(x_i)$, where μ_i and σ_i^2 are obtained as the sample estimates based on the sample $c(T_1(x_i)),\ldots,c(T_{500}(x_i))$ for the variable $F_1(x_i) = c(\tilde{T}(x_i))$.

The estimated points (1) for different t are used as an estimate of ROC curve for comparison of the classification accuracy for different leaf level of confidence functions $c()$.

In order to verify the accuracy of the model, we also calculated the ROC curves for several random forests for each function $c()$. The results imply very similar conclusion concerning the comparison of classification accuracy of different leaf level of confidence functions and will be discussed elsewhere.

5 Data sets

5.1 MAGIC data

As the first data set we used simulated test data generated by a Monte Carlo program, Corsika, described in [6]. The program simulates the events of detecting a gamma particle (interesting signal) and hadron particles (background) by ground-based atmospheric Cherenkov telescope MAGIC [1]. A comparison study of various techniques for classification of new events as signal or background may be found in [1]. The random forest technique appeared to be the best in a wide range of the ROC curve expressing the quality of the prediction. Each of the events is characterized by 10 numerical predictors used also in the real detector. Since the data are simulated, the class variable is present for all generated cases. The data were split into the training set (12679 cases) and test set (6341 cases), i.e. approximately in ratio 2:1.

5.2 Gaussian data

As the second data set, we used simple pair of distributions in 5 dimensions. The signal was generated as a vector of 5 independent variables, each with $N(0, 1)$ distribution restricted to the interval $[-5, 5]$. The background was generated as a vector uniformly distributed in the cube $[-5, 5]^5$. The training set contained 10000 cases and the test set 5000 independent cases.

6 Experimental results

In our experiments, we compared the prediction accuracy of several modifications of RF on the two data sets described above. Accuracy of the prediction of forests with various leaf confidences is measured in terms of ROC curves, which expresses the dependence between background acceptance (x-axis) and signal acceptance (y-axis). Backgound acceptance is the probability that the case from the class 0 is classified as 1. Signal acceptance is the probability that the case from the class 1 is classified correctly.

In order to obtain ROC curves for forests of different sizes, we used the statistical model developed in Section 4. Using the model, we tested several types of leaf confidences and selected a few of them, which performed best. In a table form, we present signal acceptancies for all selected types of leaf confidencies (conf.t.) and the three forest sizes (N). For each combination, the table contains the signal acceptancies for several fixed low values of background acceptance (1 to 4 percent). The last column of the table contains the average of the four previous columns and the confidence types are sorted according to this average.

For forests of size 20 and 80 for the MAGIC data, we present the best results in Figure 1. Namely, for each of these two forest sizes, we selected the leaf confidences with the best result for the given forest size and present the low part of the corresponding ROC curve (for low values of background acceptance) together with ROC curve for pure RF for comparison.

In our experiments with MAGIC data, we selected the leaf confidences $q_{(0.5)}$, $\sigma_{(1)}$, $nd_{(1,0)}$, $nd_{(0.9,0)}$ and $nd_{(0.9,2)}$. Classification using these leaf confidences is compared with the simulated pure RF. For the gaussian data, we selected the same leaf confidences with the same parameters as for MAGIC

[1] http://hegra1.mppmu.mpg.de/MAGICWeb/

data except that we used $q_{(0.01)}$ instead of $q_{(0.5)}$, since $q_{(0.01)}$ is mostly better for gaussian data.

6.1 MAGIC data

For MAGIC data, all of the presented leaf confidences are better than pure RF for forest size 20. For larger forests, the order of the confidence types changes. Confidence type $q_{(0.5)}$ becomes worse than pure RF for forest size 40 and $nd_{(0.9,0)}$ and $nd_{(0.9,2)}$ become worse than RF for forest size 80. On the other hand, confidences $nd_{(1,0)}$ and $\sigma_{(1)}$ improve their relative position in the table. The confidence type $\sigma_{(1)}$ appears to be the best for forest size 80, although the difference between $nd_{(1,0)}$ and $\sigma_{(1)}$ is negligible.

N	conf.t.	1%	2%	3%	4%	aver.
20	rf	0.2427	0.3828	0.4660	0.5341	0.4064
20	$q_{(0.5)}$	0.2947	0.4009	0.4783	0.5407	0.4287
20	$\sigma_{(1)}$	0.2939	0.4024	0.4800	0.5422	0.4296
20	$nd_{(1,0)}$	0.2961	0.4015	0.4805	0.5412	0.4298
20	$nd_{(0.9,2)}$	0.3069	0.4066	0.4799	0.5402	0.4334
20	$nd_{(0.9,0)}$	0.3090	0.4065	0.4811	0.5431	0.4349
40	$q_{(0.5)}$	0.2961	0.4046	0.4820	0.5452	0.4320
40	rf	0.2951	0.4092	0.4876	0.5509	0.4357
40	$nd_{(0.9,0)}$	0.3138	0.4098	0.4875	0.5512	0.4406
40	$nd_{(0.9,2)}$	0.3166	0.4108	0.4861	0.5493	0.4407
40	$\sigma_{(1)}$	0.3142	0.4182	0.4919	0.5535	0.4445
40	$nd_{(1,0)}$	0.3151	0.4183	0.4915	0.5536	0.4446
80	$q_{(0.5)}$	0.2977	0.4088	0.4849	0.5461	0.4344
80	$nd_{(0.9,0)}$	0.3165	0.4125	0.4923	0.5560	0.4443
80	$nd_{(0.9,2)}$	0.3223	0.4106	0.4924	0.5548	0.4450
80	rf	0.3149	0.4221	0.4965	0.5581	0.4479
80	$nd_{(1)}$	0.3284	0.4258	0.4980	0.5596	0.4529
80	$\sigma_{(1)}$	0.3293	0.4260	0.4989	0.5586	0.4532

Table 1: Results for MAGIC data. Description of the table is at the beginning of Section 6.

6.2 Gaussian data

The situation for the gaussian data has some common properties to the situation for MAGIC data. Namely, for small forest, the confidence type $nd_{(0.9,2)}$ is quite successful, but for larger forests, both confidences $nd_{(0.9,0)}$ and $nd_{(0.9,2)}$ are worse that pure RF and $\sigma_{(1)}$ and $nd_{(1,0)}$ become the best confidence types.

N	conf.t.	1%	2%	3%	4%	aver.
20	$q_{(0.01)}$	0.8246	0.9260	0.9644	0.9797	0.9237
20	rf	0.8315	0.9300	0.9649	0.9800	0.9266
20	$nd_{(0.9,0)}$	0.8347	0.9290	0.9651	0.9799	0.9272
20	$\sigma_{(1)}$	0.8351	0.9292	0.9649	0.9799	0.9273
20	$nd_{(1,0)}$	0.8358	0.9292	0.9650	0.9799	0.9275
20	$nd_{(0.9,2)}$	0.8366	0.9299	0.9653	0.9804	0.9280
40	$q_{(0.01)}$	0.8345	0.9272	0.9652	0.9803	0.9268
40	$nd_{(0.9,0)}$	0.8391	0.9294	0.9653	0.9804	0.9285
40	$nd_{(0.9,2)}$	0.8411	0.9305	0.9658	0.9809	0.9296
40	rf	0.8426	0.9310	0.9656	0.9808	0.9300
40	$nd_{(1,0)}$	0.8449	0.9296	0.9655	0.9806	0.9302
40	$\sigma_{(1)}$	0.8455	0.9297	0.9656	0.9806	0.9304
80	$q_{(0.01)}$	0.8430	0.9266	0.9652	0.9805	0.9289
80	$nd_{(0.9,0)}$	0.8433	0.9295	0.9655	0.9805	0.9297
80	$nd_{(0.9,2)}$	0.8455	0.9304	0.9660	0.9813	0.9308
80	rf	0.8498	0.9317	0.9660	0.9810	0.9321
80	$nd_{(1)}$	0.8526	0.9301	0.9656	0.9808	0.9323
80	$\sigma_{(1)}$	0.8527	0.9303	0.9657	0.9809	0.9324

Table 2: Results for gaussian data. Description of the table is at the beginning of Section 6.

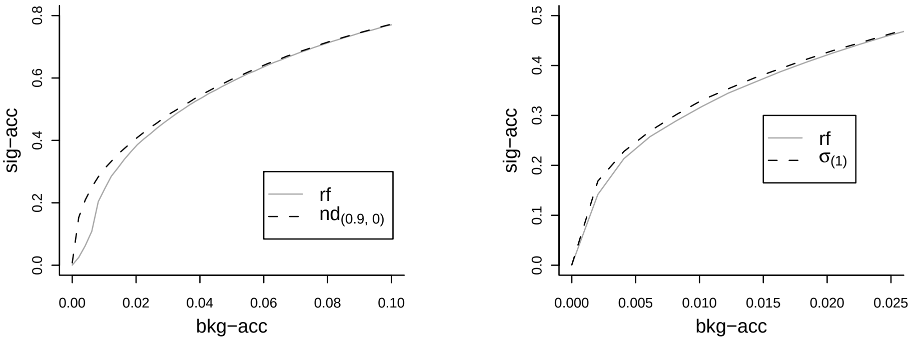


Figure 1: The graph on the left hand side compares rf and $nd_{(0.9,2)}$ for 20 trees, on the right hand side compares rf and $\sigma_{(1)}$ for 80 trees. Both graphs are for MAGIC data.

7 Discussion and conclusion

The current paper investigates Random Forests classifiers parametrized so that the probability of accepting a negative case is very small. Under these conditions, appropriately chosen leaf levels of confidence may improve the prediction. Two distributions are presented, for which the improvement may be observed for forests of size between 20 and 80. The advantage obtained using leaf confidences decreases with increasing forest size, however, in practical situations, improving predictions of forests of limited size is important.

References

[1] Bock R.K., Chilingarian A., Gaug M., Hakl F., Hengstebeck T., Jiřina M., Klaschka J., Kotrč E., Savický P., Towers S., Vaicilius A., Wittek W. (2004). *Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope.* Nucl. Instr. Meth. A **516**, 511 – 528.

[2] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984). *Classification and regression trees*. Belmont CA: Wadsworth.

[3] Breiman L. (1996). *Bagging predictors.* Machine Learning **24** (2), 123 – 140.

[4] Breiman L. (2001). *Random forests.* Machine Learning **45** (1), 5 – 32.

[5] Freund Y., Schapire R.E. (1997). *A decision-theoretic generalization of on-line learning and an application to boosting.* J. of Computer and System Sciences **55** (1), 119 – 139.

[6] Heck D. et al. (1998). *CORSIKA, A Monte Carlo code to simulate extensive air showers.* Forschungszentrum Karlsruhe FZKA 6019.

[7] Quinlan J.R. (1996). *Bagging, boosting and C45.* Proceedings of the Thirteenth National Conference on Artificial Intelligence, 725 – 730.

[8] Schapire R.E., Singer Y. (1999). *Improved boosting algorithms using confidence-rated predictions*. Machine Learning **37**, 297 – 336.

Acknowledgement: The authors were supported by the Ministry of Education of the Czech Republic under Research and Development project LN00A056, Institute of Theoretical Computer Science - ITI. Experiments were performed using $\mathcal{META}$ Center[2].

Address: P. Savický, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
E. Kotrč, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Department of Mathematics, Trojanova 13, 120 00 Prague 2, Czech Republic
E-mail: `savicky@cs.cas.cz`

[2] http://meta.ten.cz

COMPSTAT'2004 Symposium

STANDARD METHODS AND INNOVATIONS FOR DATA EDITING

Emanuela Scavalli

Key words: Error localisation, missing data, nonlinear relations, neural networks.

COMPSTAT 2004 section: Neural networks, data imputation.

Abstract: This paper informs about the results of the EUREDIT (The development and evaluation of new methods for editing and imputation) project aimed at improvement of the data editing.

1 Introduction

The necessity of having a reliable dataset presumes the treatment of the collected data (both administrative and survey ones), generally indicated as *data editing*, which consists of an integrated and complementary set of methodologies aimed at producing coherent and complete data. In this process, if from one side it is essential to verify the presence of possible errors, on the other side it is also important to find or impute the real value when it is missing, paying particular attention not to introduce errors.

In this sense the evaluation of the quality of the editing and imputation process assumes a fundamental role [6]. In general the error localisation procedures present high quality levels when they allow to identify the maximum number of errors, especially those which have more influence on the analysis results, while the imputation procedures are evaluated basically on their ability to preserve the structure (at micro and macro level) of the *real* data.

In general the application of these evaluation criteria requires the knowledge of the real values for a set of rough observed data.

The results from the EUREDIT project[1] represent a great contribution for the considerations that in this paper are reported. In fact, the project made a comparative evaluation of traditional and innovative data editing methods for different datasets in order to define a general strategy which associates the best methods to the different applicative situations. In this sense, a set of innovative methods[2] has been considered in order to integrate and improve the *Standard* techniques, i.e. those currently in use in the Na-

[1] The EUREDIT project ("*The development and evaluation of new methods for editing and imputation*") has been carried out under the 5th Framework Program for Research and Development of the European Commission from March 2000 to February 2003 and the consortium was composed by National Statistical Institutes, universities and private firms. www.cs.york.ac.uk/euredit

[2] Neural Networks Models, robust methods for the outlier detection, non-parametric regression methods for imputing in panel surveys or time series.

tional Statistical Institutes (NSIs).[3] *Standard* methods have been considered so to offer a benchmark for the evaluation of new methods. The comparison of the different techniques to different types of datasets, with continuous and discrete data, allowed to make some considerations and suggestions on the best practice use for each type of data.

2 Standard methods and innovations for data editing

The methods that the NSI's currently use for data *editing* (defined in the project *standard methods*), can be distinguished in two classes, according to their belonging to the *rules-based* methods or the *model-based* methods.

In particular, among the rule-based approaches the *Fellegi-Holt* method [5] and the *data-driven* approach are generally considered for the error localisation, while the single and multiple imputation techniques based on the *minimum distance donor* or on regression methods are adopted for the missing data imputation.

In the *Fellegi-Holt* methods a set of edit rules is used both for error localisation and for imputation.

For the error localisation, the subset of edits activated by a given record is processed in order to individuate the subset of variables which most likely contains the errors causing the activation of those edits. The F-H error localisation algorithm is based on the *minimum change principle*, i.e. the number of variables judged to be erroneous must be the minimum under the constraint to explain all edit failures. For continuous data only linear edits on non-negative variables are admissible. A variant to this approach is given by the Nearest-neighbour Imputation Methodology (NIM), accordingly to which the error localisation is no more based on the minimum change principle, but on the consideration of the differences between the current record (with edit failures) and a potential donor (a neighbour with no edit failures).

For the imputation process, a range of possible values to impute is determined in order to avoid values that might cause additional failures of edit rules; then, actual values can be assigned by using a number of different methods (nearest neighbour, regression imputation, etc.)

Different systems incorporating F-H methods have been developed by Statistics Canada, Statistics Netherlands, ISTAT and ONS, and they have been applied in the EUREDIT project. The NIM approach presents many advantages that make it preferable to the *Fellegi-Holt* methodology in some situations. One of them is the editing and imputation of complex hierarchical structures, such as households. NIM allows to consider an entire household as the record to be edited, and experiences carried out made it clear that its performance is higher than that of pure F-H systems or other systems.

In particular, the main softwares in use in this class of F-H methods are:

[3] The Fellegi-Holt method and the data-driven approach for the error localisation, the single and multiple imputation techniques based on the minimum distance donor or on regression methods for the missing data imputation.

CANCEIS and SCIA for categorical variables, GEIS for continuous variables, Cherry-PIE and E-C system for continuous variables, DIS only for imputation of continuous and categorical variables.

The *Model-based* methods, on the other hand, are founded on the definition of as many models (parametric and linear) as variables involved in the process of edit and imputation are. These models are used both for error localisation and imputation.

Error localisation is carried out by calculating, for each variable, an expected value conditioned on a set of covariates; then the actual value is compared to the expected value, and if the two values diverge too much, the actual value is considered erroneous.

Problems related to this approach can be represented by the presence of errors (or missing values) into the covariates; moreover, it is crucial the choice of the metric used for evaluating the proximity between actual and expected values.

The imputation is made on the basis of the defined model assigning the expected value to missing and erroneous data. Also in this case covariates may contain errors which can influence the predicted value. The imputation can be *deterministic* (the predicted value is directly imputed), or *stochastic* (the imputed value is drawn by from a conditional distribution), with the preservation of means and totals. Imputations carried out in this way generally do not take into account the coherence of imputed values with values of other variables contained into the record. However, the application of edit failures to the predicted values is therefore possible subsequent to the imputation.

In this class of methods the EM (*Expectation -Maximisation*) algorithm and the IMAI (*Integrated Modelling Approach to Imputation*) are currently used. The EM algorithm estimates distribution parameters in presence of missing data, under a specified super-population model and an ignorable non response mechanism [4] with the assumption of multi-normality.

The *standard* methods which are based on deterministic edit rules even if exactly determine the presence of errors or impute the right values but they require the possibility to define coherence rules. On the other side, *standard* methods based on models do not take into account of edits or coherences but they usually suppose linear relationships between variables.

In this logic, new methods which take into account of non linear relations between variables are arising. In particular neural networks represent a high potential instrument since they presume not linear relations between variables and, that is very innovative, do not a priori require the knowledge of the model, since they learn and build the model directly from data.

Therefore, the problem of presence of errors in the phase of neural networks model construction is reduced in respect of the *standard* methods, because they are less sensitive to anomalous values. This can means that the treatment of outlier has to be considered in a separate way.

3 New methods and the neural networks

The *new methods*, which were investigated in the EUREDIT project, have been subdivided in those belonging to the class of neural network methods, and those classifiable as robust methods which deal mainly with problems of *outlier detection* and *robust imputation*.

Different methods have been considered in this class, each of them with relevant peculiarities: *Multi-layer Perceptrons* (MLPs), *Tree-Structured Self-Organising Maps* (SOMs), *Correlation Matrix Memories* (CMMs), *Support Vector Machines* (SVMs). Among these methods very interesting results derive on the application of the MLPs.

In general a neural network is composed by a set of elementary units (neurones) linked by weighted connections (**weights**). The processing units are arranged in layers: an **input layer** with units representing the input fields, one or more **hidden layers**, and an **output layer** with a unit or units representing the output field(s). *A Multi-Layer Perceptron* is a network with almost one hidden layer.

In practice, to each input (X_i) a weight (W_i) is associated, then a combination of inputs and weights is obtained throughout a net potential function ($P = f(X_i, W_i)$); finally a transfer function $F = g(P)$ combines the output, which could be the final output or used as an input for a subsequent layer.

The network learns through training by examining individual records, generating a prediction for each record, and making adjustments to the weights whenever it makes an incorrect prediction. Initially, all weights are defined with random values. The network learns through training. Examples for which the output is known are repeatedly presented to the network, and the answers (predicted values) it gives are compared to the known outcomes. Information from this comparison is passed back through the network, gradually changing the weights. As training progresses, the network usually becomes increasingly accurate in replicating the known outcomes. This process is repeated many times and the network continues to improve its predictions until one or more stopping criteria (as accuracy, cycles, time running) have been met. Once trained, the network can be applied to future cases where the outcome is unknown.

In order to prevent *over-training*, i.e. the risk that the created network represents perfectly the relations between data without having the feature of generalisation, the initial dataset is split into training and validation sets. The network is trained on the training set, and accuracy is estimated by using the validation set.

The MLP can be used both for localising errors and for imputing missing values. The error localisation can be obtained by considering as target variable in the network the presence or absence or errors. For this approach the presence of a clean dataset is required for training networks: by comparing clean and perturbed data, an indicator of presence/absence of errors for each variable is calculated: MLPs are trained on this subset and then generated networks are applied on the perturbed datasets. Another approach consists in considering as target variable the variable itself (not the indicator): if the predicted value differs form the actual value then it can be considered erroneous. In both error localisation approaches, a threshold value above which corresponding values can be classified as erroneous has to be defined generally minimising the total amount of misclassifications.

As far as the imputation process is concerned, MLPs (with target variable equal to the variable itself) are trained on those records for which the target value is not missing and the so generated networks are applied for imputing missing values. What is very interesting is that results fundamentally does not change if data for the network training contains errors (except, obviously, for out of range values).

4 Some considerations

The definition of goodness of a method depends on the evaluation criteria adopted and on the contest in which the method is applied. In general, a method can present good performances related to some criteria, but not for others. For example, for the error localisation, in general the criteria adopted for measuring the performances are the capability of a method to correctly classify errors, or, conversely, its capability to minimise misclassifications; or the ability to detect the most influential errors, those with the highest impact on final estimates.

An imputation procedure, on the other side, should be evaluated with respect to the *predictive accuracy*, i.e. it should preserve single values, the *distributional accuracy*, i.e. it should preserve the distribution of true data, and the *estimation accuracy*, i.e. it should reproduce as much as possible the lower order moments of the distribution of true data (at least first and second moments). It should also be important that the imputed values are "plausible", i.e. coherent with other data and not failing any edit rule.

On the basis of the different evaluation criteria the best methods should be chosen. It is important to underline the fact that the concept of *best* is sometimes very relative, as performance for a given method may vary accordingly to the considered indicators and subsets of variables.

Among *standard methods*, from the results of the EUREDIT project it emerges that CANCEIS-SCIA revealed the best performance for categorical data, both for error localisation and imputation, CHERRY PIE was the best for error localisation in continuous data. For imputation, both multivariate regression plus hot deck method showed the best results, followed by IMAI predictive mean matching method.

Among *neural network based methods*, MLP applications always presents best performances, both for error localisation and imputation, followed by SOM.

By comparing all methods, *standard* and innovative ones, in some cases the traditional results better, especially for those datasets and variables for which it is possible to define edit rules. This was the case of the UK Census data where variables concerning *individuals* (relation to head, marital status, sex and age), and *households* (number of rooms and presence of bathroom) have been considered. CANCEIS-SCIA results the best for categorical variables because of the possibility of define many edit rules; for the continuous variable (age), MLP is the best for the influential error detection, while CANCEIS-SCIA shows the best performance for estimation accuracy. Instead, in terms of preservation of true values SVM presents better results followed by MLP and CANCEIS-SCIA.

In a dataset of the *UK Annual Business Inquiry* for the error localisation of the six most important economic variables the MLPs results the best method in terms of pure error localisation performance and differences between moments, while SOM is the better in terms of influential error detection. For the imputation, MLPs present the best values in term of predictive and distributional accuracy.

Results on the *Danish Labour Forces Survey (DLFS)*, where the only variable "income" contains missing values, MLP shows the best values in terms of *predictive accuracy* followed by the Linear Regression. In the *distributional accuracy* group of indicators, MLP is still among the best with SOM.

It is evident that in many cases new methods, especially the MLPs present better results than *standard* methods; this underline the potentiality of these methods, that can be further improved by using them in a combining way with *standard* methods: this could be done, for example, by applying first the edit-rule methods and then the neural networks, where a control on the coherency of the imputed values can be done a posteriori by using the traditional methods.

For all variables for which distributional hypothesis cannot be specified or characterised by non linear relationships, the MLPs represent an important tool for the *data editing* [7]. From the project results, the main feature that arise is that, even if the model does not recognise errors, but it almost never introduces new ones (false positive); this could represent a good quality for a localisation technique.

The main limits of the method consist in the fact that a clean dataset is required, it is also difficult to give an interpretation of the model parameters and no edit rules can be specified in the process of error localization and imputation.

References

[1] AA.VV. (2003). *Methods and experimental results from the Euredit project.* Deliverable 6.1, EUROSTAT.

[2] AA.VV. (2003). *Best methods for statistical edit and imputation- the Euredit Project.* Deliverable 6.2, EUROSTAT.

[3] Bishop M.C. (1995). *Neural network for pattern recognition.* Oxford Clarendon Press.

[4] Dempster A.P., Laird N.M., Rubin D.B. (1977). *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society B **39**, 1 − 38.

[5] Fellegi I.P., Holt D. (1976). *A systematic approach to automatic edit and imputation.* Journal of the American Statistical Association, **71**, 17 − 35.

[6] Luzi O., Scavalli E. (2001). *Evaluating the quality of data editing and imputation methods in the European EUREDIT project.* Proceedings of Intermediate Meeting of the Italian Statistical Society, Rome, 4-6 June.

[7] Scavalli E. (2003). *Neural networks for error localisation and data imputation in statistical surveys.* Proceedings of Intermediate Meeting of the Italian Statistical Society, Naples, 9-11 June.

[8] Verboon P., Schulte Nordholt E. (1997). *Simulation experiments for hotdeck imputation.* Statistical Data Editing, Methods and Techniques, Vol. II, **48**, UN/ECE, 22 − 29.

Address: E. Scavalli, ISTAT, National Statistical Institute, v. C. Balbo n.16
00184 Rome Italy

E-mail: `scavalli@istat.it`

INTERDEPENDENCE BETWEEN EMERGING AND MAJOR MARKETS

Abdel Sharkasi, H. Ruskin and M. Crane

Key words: Simple regression, volatility and wavelet analysis.

COMPSTAT 2004 section: Applications.

Abstract: In this paper, we investigate the price spillover effects among two developed markets, (the US and the UK), and two developing markets, (Irish and Portuguese), using a new testing method suggested by Lee (2002). We find that there are interrelationships between any two of the Irish, the UK and Portuguese markets and that the co-movements between the emerging markets and the US are statistically significant but weak. We also found that the US market is slightly influenced by the UK but not *vice versa*.

1 Introduction

The relationships between international stock markets have been investigated in several articles, especially after “Black Monday”, (October 1987). These studies indicated that co-movements among stock markets have increased the possibilities for national markets to be influenced by the changes in international ones ([12], [9], [6], [7] and [13]).

The advantage of global portfolio diversification has been noted in the finance literature for some time. Several studies ([11], [14] and [2]) showed that it is useful to spread content internationally, rather than locally, as stocks in different markets are less correlated than those within the same market. Tang [16] investigated, for instance, Asian emerging and mature markets and reported that an increase in the correlation between worldwide stock markets may cause the reduction of some or all of the diversification benefits and this means that diversification benefits depend upon the degree of the relationships among different stock markets. Tang [17] found that the intertemporal stability of the correlation matrix is important in examining the ex-ante diversification benefits and stock market co-movements. The potential diversification effects have decreased and become less important due to increase in the international co-movement among stock markets, especially since the mid 1990's ([15] and [16]).

More recently, Lee [10] developed a new testing technique based on the wavelet transform, in order to study the international transmission effects between three developed markets (the US, Germany and Japan) and two emerging markets in the MENA region, namely Egypt and Turkey. He documented that innovation from the major markets affected the emerging markets but the that opposite was not true.

In addition, Bessler and Yang [3] employed an Error Correction Model and Directed Acyclic Graphs (DAG) to study the co-integration among nine major markets namely Japan, the US, the UK, France, Switzerland, Hong Kong, Germany, Canada and Australia. Their results showed that changes in the UK, Switzerland, Hong Kong, France and Germany influenced the US market, while the US market is affected by its own innovation as well. Moreover, Brooks and Negro [4] studied the relationship between market co-integration and the degree to which companies operate internationally. They considered three factors, (global, country-specific and industry-specific), and found that the importance of the international factor has increased since the 1980s while that of the country-specific factor has decreased.

Furthermore, Wongswan [18] found strong evidence of international transmission from the US and Japanese markets to Korean and Thai markets during the late 1990's. Most recently, Antoniou et al. [1] applied a VAR-EGARCH model to study the relationships among three EU markets namely Germany, France and the UK and their results showed evidence of co-integration among those countries.

Our goal in this article is to study whether or not there is evidence of co-integration between four stock markets (Irish, Portuguese-as developing and the UK and the US-as mature). To examine this, we applied a testing method, (based on the wavelet transform), suggested by Lee [10].

The remainder of this paper is organized as follows: In Section 2, a brief description of the testing method is given. The data and empirical results are described in Section 3 and our conclusion is presented in the final section.

2 Brief description of the testing method

With the increase in media coverage of world events and a corresponding increase in access by the wider public to this coverge, global transmissions of information can be expected to be completed within a short period of time. The wavelet analysis and, in particular, the discrete wavelet transform (DWT), is very useful (for more detail see [5]) in splitting data series into different frequency wavelet crystals and high-frequency components which explain the short-term movements in the series . A new testing method based on wavelet analysis was developed by Lee [10] and it can be described as follows:

- Reconstruct the returns series using the first and the second high-frequency wavelet crystals (d_1 & d_2) separately.
- Estimate the simple regression and reverse regression models between each two using three different scales:
  - The row daily returns.
  - The returns series rebuilt form d_1.
  - The returns series rebuilt form d_1 plus that rebuilt from d_2.
- Test the significant of regression coefficient (slope) and R^2.

3 Data and empirical results

The data used in the following analysis consists of the daily prices of stock market indices for two emerging markets, namely Portuguese and Irish and two major markets, (the US and the UK), during the period from January 1^{st}, 1993 to September 30^{th}, 2003. We considered the indices ISEQ Overall, PSI20, FTSE All Share and S&P500 to be representative of the Irish, Portuguese, UK and US markets respectively.

As these markets use their local currencies for presenting the values of their indices, so we use the daily returns instead of using the daily prices, where the former equal the natural logarithm of the ratio between the closing price of index at time t and that at time $t - 1$. Some daily observations have been deleted because the markets we studied have different holidays and closing trading days, (as has been done by e.g. [10]).

Index→ **Measure↓**	**ISEQ**	**PSI20**	**FTSE**	**S&P500**
No. Observations	2556	2556	2556	2556
Mean	0.00052	0.00029	0.00012	0.00033
Std.Dev	0.0104	0.0109	0.0099	0.0111
Minimum	-0.0757	-0.0959	-0.0515	-0.0704
Maximum	0.0584	0.0694	0.0509	0.0557
Skewness	-0.3580**	-0.5760**	-0.1820	-0.021
Kurtosis	4.503**	6.849**	2.794**	3.077**
Jarque-Bera	2203.63**	5109.643**	840.70**	1002.87**

Note:** denotes statistically significant at 1% level.

Table 1: Descriptive statistics of the daily returns of the stock markets indices series.

Table 1 represents the descriptive statistics of the stock market indices and shows that the sample means of all indices are positive. We test whether or not the skewness and kurtosis of all these series are different from zero. The results show that the returns series of ISEQ and PSI20 indices have significant negative skewness, but those of FTSE and S&P500 are not significantly different from zero. The returns of all indices are leptokurtic and the results of a normal test (Jarque-Bera) also show that all returns series can not be regarded as normally distributed.

From Table 2, It can be seen that high-frequency components have more energy than low-frequency ones and this implies that the movements in all index returns are caused by the short-term fluctuations. It also implies that the first "d_1" and the the second "d_2" components of the wavelet transform account for more than 60% of the energy. This indicates that there are no long memory effects in the returns series of these indices.

In order to study the co-movements among those markets, firstly, we built simple regression models between each of the two European markets on the

Index → **Wavelet Crystals↓**	**ISEQ**	**PSI20**	**FTSE**	**S&P 500**
s_6	0.028	0.039	0.014	0.012
d_6	0.023	0.025	0.017	0.012
d_5	0.036	0.042	0.027	0.031
d_4	0.070	0.058	0.047	0.048
d_3	0.155	0.163	0.157	0.145
d_2	0.274	0.267	0.301	0.234
d_1	0.431	0.406	0.436	0.518

Table 2: Percentages of energy by wavelet crystals for the daily returns of indices series.

same trading day and similarly for each European market on the US market of the *previous* trading day. Secondly, we built a simple regression model of the US market on each European market on the same trading day and these models are estimated using the three different scales mentioned in Section 2. The results are given in Tables 3(A) to 3(F) for each case and clearly show that there are significant levels of inter-correlation between the Irish and UK markets and also between the Irish and Portuguese. However, the relationship between the Irish and US markets is weak. From Table 3 (D), (E) and (F), we can see that there is significant co-movement between Portuguese and UK markets and there are spillover effects from both Portuguese and UK markets on the US market but not *vice versa*.

Regression→	$M_t{}^{IRL}$ **on** $M_t{}^{UK}$			$M_t{}^{UK}$ **on** $M_t{}^{IRL}$		
Scales↓	**Constant**	**Slope**	R^2	**Constant**	**Slope**	R^2
Return	4.46E-04 (0.034)	0.592 (0.000)	0.322	-1.58E-04 (0.328)	0.544 (0.000)	0.322
Return.D1	-5.85E-07 (0.996)	0.509 (0.000)	0.251	-1.06E-06 (0.992)	0.492 (0.000)	0.251
Return.D1.2	6.18E-08 (1.000)	0.552 (0.000)	0.300	-3.31E-06 (0.981)	0.544 (0.000)	0.300

A: ISEQ Overall and FTSE

Regression→	M_t^{IRL} **on** $M_{t-1}{}^{US}$			$M_t{}^{US}$ **on** $M_t{}^{IRL}$		
Scales↓	**Constant**	**Slope**	R^2	**Constant**	**Slope**	R^2
Return	4.46E-04 (0.034)	0.356 (0.000)	0.145	1.93E-04 (0.365)	0.258 (0.000)	0.057
Return.D1	-1.94E-06 (0.988)	0.172 (0.000)	0.039	-2.65E-06 (0.987)	0.065 (0.007)	0.002
Return.D1.2	-3.41E-06 (0.983)	0.273 (0.000)	0.092	1.26E-06 (0.995)	0.155 (0.000)	0.019

B: ISEQ Overall and S&P500

Regression→	$M_t{}^{IRL}$ **on** $M_t{}^{P}$			$M_t{}^{P}$ **on** $M_t{}^{IRL}$		
Scales↓	**Constant**	**Slope**	R^2	**Constant**	**Slope**	R^2
Return	4.19E-04 (0.029)	0.340 (0.000)	0.128	9.67E-05 (0.632)	0.378 (0.000)	0.128
Return.D1	-1.28E-06 (0.992)	0.352 (0.000)	0.135	-6.94E-08 (1.000)	0.384 (0.000)	0.135
Return.D1.2	-3.64E-06 (0.995)	0.341 (0.000)	0.135	4.22E-06 (0.370)	0.370 (0.000)	0.126

C: ISEQ Overall and PSI20

Regression→	$M_t{}^{P}$ **on** $M_t{}^{UK}$			$M_t{}^{UK}$ **on** $M_t{}^{P}$		
Scales↓	**Constant**	**Slope**	R^2	**Constant**	**Slope**	R^2
Return	2.29E-04 (0.230)	0.517 (0.000)	0.221	-1.45E-06 (0.993)	0.428 (0.000)	0.221
Return.D1	2.84E-07 (0.998)	0.516 (0.516)	0.236	-1.51E-06 (0.989)	0.459 (0.000)	0.237
Return.D1.2	5.65E-06 (0.971)	0.505 (0.000)	0.231	-6.18E-06 (0.976)	0.458 (0.000)	0.231

D: PSI20 and FTSE

Regression→	$M_t{}^{P}$ **on** $M_{t-1}{}^{US}$			$M_t{}^{US}$ **on** $M_t{}^{P}$		
Scales↓	**Constant**	**Slope**	R^2	**Constant**	Slope	R^2
Return	2.29E-04 (0.280)	0.196 (0.000)	0.040	2.48E-04 (0.241)	0.266 (0.000)	0.066
Return.D1	-7.32E-07 (0.996)	3.41E-02 (0.058)	0.001	-2.62E-06 (0.987)	0.194 (0.000)	0.028
Return.D1.2	2.88E-06 (0.987)	0.122 (0.000)	0.017	1.22E-07 (0.999)	0.228 (0.000)	0.044

E: PSI20 and S&P500

Regression→	$M_t{}^{UK}$ **on** $M_{t-1}{}^{US}$			$M_t{}^{US}$ **on** $M_t{}^{UK}$		
Scales↓	**Constant**	**Slope**	R^2	**Constant**	**Slope**	R^2
Return	3.49E-05 (0.852)	0.272 (0.000)	0.092	2.69E-04 (0.179)	0.471 (0.000)	0.177
Return.D1	-1.81E-06 (0.989)	4.46E-03 (0.793)	0.000	-2.20E-06 (0.989)	0.300 (0.000)	0.060
Return.D1.2	-5.17E-06 (0.975)	0.151 (0.000)	0.029	2.59E-06 (0.989)	0.368 (0.000)	0.106

F: FTSE and S&P 500

- P-values of t-tests are given in parentheses.
- Where subscript refers to the day in question and the superscript indicates the market (e.g. IRL, P are the Irish and Portuguese markets respectively).
- Return.D1 is an indicator of the returns series, reconstructed using the first wavelet crystal (d_1).
- Return.D1.2 is an indicator of the returns series, reconstructed using the first and the second wavelet crystals (d_1 & d_2).

Table 3: Regression Analysis between each pair of four stock markets using three different scales.

Market **Explained↓**	Days **Ahead**	**Ireland**	**Portugal**	**The UK**	**The US**	**OM**
Ireland	5	60.77	1.20	26.15	11.88	39.29
	10	60.24	1.31	26.29	12.16	39.76
	15	60.21	1.32	26.30	12.17	39.79
Portugal	5	0.51	77.54	18.40	3.54	22.45
	10	0.83	76.61	18.83	3.73	23.39
	15	0.83	76.52	18.83	3.82	23.48
The UK	5	0.38	0.30	88.77	10.56	11.24
	10	0.59	0.54	87.99	10.87	12.00
	15	0.59	0.55	87.99	10.88	12.02
The US	5	0.37	0.78	19.87	78.98	21.02
	10	0.45	1.16	20.44	77.95	22.05
	15	0.45	1.17	20.45	77.93	22.07

Note: OM denotes the percentage of forecast error variance explained collectively by the other markets.

Table 4: The Percentages of error variance of the market in the first column explained by innovation in the market in the first row.

To compare our results with one of the common methods, we estimated the vector autoregressive (VAR) model of order 10 of the daily returns of these markets. The percentages of the decomposition of 5-day, 10-day and 15-day ahead forecasts of the returns series have been measured[1]. At 15 days ahead, for example, the results, given in Table 4, show that the most of the variance in these markets is explained by their own innovations and that the UK is the most influential market while the Irish is the most influenced market. The UK explains 26.30, 18.83 and 20.45 percent for Irish, Portuguese and the US respectively and the US explains 12.17, 3.82 and 10.88 percent of the variance of Irish, Portuguese and the UK respectively. We also found that the forecast error variance is very sensitive to the order of variables for orthogonalization and to the stability of these series and this suggests that the new technique, based on wavelet analysis, is more reliable than the VAR method.

4 Conclusion

Our objective in this paper has been to study the international transmission between four markets namely the Irish, Portuguese, UK and US. A new testing method suggested by Lee [10] has been applied to do so. Our results show that there are significant inter-correlations between each pair of Irish, Portuguese and UK markets separately. In addition, the indications are that the US has insignificant spillover effects from or on to the other markets. We can say that the emerging markets have significant spillover effects on each other but there is no co-integration between the major markets.

[1] The orthogonalization is ordered as the UK, Portuguese, the US and Irish.

Wavelet analysis

The Wavelet Transform (**WT**) has been explained in some detail, (particularly in [5] and [10]) and the following offers a brief explanation only. The **WT** has two types of wavelets called father and mother wavelets, ϕ and ψ respectively, where $\int \phi(t)dt = 1$ and $\int \psi(t)dt = 0$. These can be computed using the following equations

$$\phi(t) = \sqrt{2}\sum_k \ell_k \phi(2t-k) \tag{1}$$

$$\psi(t) = \sqrt{2}\sum_k \hbar_k \phi(2t-k) \tag{2}$$

The orthogonal wavelet series approximation to a given signal $f(t)$ is defined by

$$f(t) = \sum_k s_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) + \ldots + \sum_k d_{1,k}\psi_{1,k}(t) \tag{3}$$

where J is the number of multiresolution levels, (or crystals), and k ranges from 1 to the number of coefficients in the specified components (or levels). The coefficient $s_{J,k}$, $d_{J,k}$, …, $d_{1,k}$ are the wavelet transform coefficients given by

$$s_{J,k} = \int \phi_{J,k}(t)f(t)dt \tag{4}$$

$$d_{j,k} = \int \psi_{j,k}(t)f(t)dt \qquad (j = 1, 2, \ldots, J) \tag{5}$$

The discrete wavelet transform (**DWT**) computes the coefficient of the wavelet series approximation in Equation(3) for a discrete signal $f_1, \ldots, f_n$ of finite extent. The DWT maps the vector $f = (f_1, f_2, \ldots, f_n)'$ to a vector of n wavelet coefficients $w = (w_1, w_2, \ldots, w_n)'$ which contains the "smooth" coefficient $s_{J,k}$ and "detail" coefficients $d_{j,k}$ $[j = 1, 2, \ldots, J]$. The $s_{J,k}$ describes the underlying smooth behaviour of the signal at coarse-scale 2^J while $d_{J,k}$ describes the coarse-scale deviations from the smooth behaviour and the $d_{J-1,k}, \ldots, d_{1,k}$ provide progressively finer-scale deviations from the smooth behaviour.

References

[1] Antoniou A., Pescetto G., Violaris A. (2003).*Modelling international price relationships and interdependencies between the stock index and stock index futures markets of three EU countries: A Multivariate analysis.* Journal of Business Finance **30** (5)& (6), 645 – 67.

[2] Baily W., Stulz R. (1990). *Benefits of international diversification: the case of Pacific Basin stock markets.* Journal of Portfolio Managment **16** (4), 57 – 61.

[3] Bessler D.A., Yang J. (2003). *The structure of interdependence in international stock markets.* Journal of International Money and Finance **22**, 261 – 87.

[4] Brook R., Negro M.D. (2003). *Firm-level evidence on international stock market co-movement.* International Monetary Fund, IMF Working Papers **No: 03/55**, Washington, DC, USA.

[5] Bruce A., Gao H. Y. (1996). *Applied wavelet analysis with S-Plus.* New York: Springer-Verlag.

[6] Booth G. G., Martikainen T., Tse Y. (1997). *Price and volatility spillovers in Scandinavian stock market.* Journal of Banking and Finance **21**, 811 – 823.

[7] CVM (1998). *International transmission of stock market volatility spillover effect on Latin American markets.* the IOSCO's Emerging Markets Annual Meeting, (Conference on Management of Volatility in Turbulent Markets), Kuala Lumpur, Malaysia, May 1998.

[8] Francis B.B., Leachman L.L. (1998). *Superexogeneity and the dynamic linkages among international equity markets.* Journal of International Money and Finance **17**, 475 – 92.

[9] Kim S.W., Rogers J.H. (1995). *International stock price spillovers and market liberalization: evidence from Korea, Japan, and the United States.* Journal of Empirical Finance **2**, 117 – 33.

[10] Lee H. S.(2002). *International transmission of stock market movements: A wavelet analysis on MENA stock market.* Economic Research Forum, ERF Eighth Annual Conference, Cairo, Egypt, January 2002.

[11] Levy H., Sarnat M. (1970). *International diversification of investment potfolios.* American Economic Review **60**, 668 – 75.

[12] Lin W., Engle R.F., Ito T. (1994). *Do bulls and bears move across borders? international transmission of stock returns and volatility.* The Review of Financial Studies **7** (3), 507 – 38.

[13] Ng A. (2000). *Volatility spillover effect from Japan and the US to the Pacific-Basin.* Journal of International Money and Finance **19**, 207 – 33.

[14] Solnik B.H. (1974). *Why not diversify internationally rather than domestically?.* Financial Analysis Journal **30** (4), 48 – 54.

[15] Solnik B.H. (1990). *Pacifc Basin and international diversification.* In Capital Markets Research **2**, Ghe S. G and Chang R. P (ed), Elsevier Sicence Publishers B. V. (North-Holland).

[16] Tang G. Y. (1996). *Intervalling effect on intertemporal stability among Asian emerging markets and developed markets.* Journal of Business Research **36**, 257 – 65.

[17] Tang G. Y. (1998). *The intertemporal stability of the covariance and correlation matrices of Hong Kong stock market.* Applied Financial Economics **8**, 359 – 65.

[18] Wongswan J. (2003). *Transmission of information across international equity markets.* International Finance Discussion Papers **759**, Board of Governors of the Federal Reserve System, USA.

Acknowledgement: A.S would like to gratefully acknowledge the receipt of a grant from his government (Libya) in support of this research. Professers Gama and Duarte and J. A. Matos are thanked for facilitating access to the Portuguese data.

Address: A. Sharkasi, H. Ruskin, M. Crane, School of Computing, Dublin City University, Dublin-Ireland

E-mail: asharkasi, hruskin and mcrane@computing.dcu.ie

COMPSTAT'2004 Symposium

FLEXIBLE REGRESSION MODELING VIA RADIAL BASIS FUNCTION NETWORKS AND LASSO-TYPE ESTIMATOR

Teppei Shimamura and Masahiro Mizuta

Key words: Nonlinear regression, neural networks.

COMPSTAT 2004 section: Neural networks.

Abstract: Radial basis function networks provides a more flexible model and gives a very good performance over a wide range of applications. However, in the modeling process, care is taken not to choose the number of the basis functions and the positions of the centres, the regularization parameter and the smoothing parameter as appropriate according to the model complexity, they often gives poor generalization performance.

In this paper, we develop a new model building procedure based on radial basis function networks; positioning the centres with k-means clustering for the conditional distribution $\Pr(\boldsymbol{x}|y)$ and estimating the weights by maximum penalized likelihood with Lasso penalty. We present an information criterion for choosing the regularization and smoothing parameters in the models. The proposed procedure determines the proper number and location of the centres automatically. The simulation result shows that the proposed method performs very well.

1 Introduction

Neural networks have showed promise in a number of practical applications. Examples can be found in statistics, engineering, artificial intelligence, and other fields. In recent years, there have been many developments of the radial basis function network model in the areas of the radial basis function design, the choice of the centres and the widths, and learning algorithms.

In this paper, we consider nonlinear regression models based on radial basis function networks. Radial basis function networks provides a flexible model and gives a good performance over a wide range of regression problems. One of the often-cited advantages of the Radial basis function networks is its simplicity; once the basis functions are determined, there is no full nonlinear optimization scheme required, as compared with the multi-layer perceptoron. However, many of the heuristical or sophisticated approaches for centre placement and width determination not always gives good generalization performance. And then the use of unsupervised techniques to determine the basis function parameters is not in general and optimal procedure so far as the subsequent supervised training is concerned.

We propose a new model building procedure based on radial basis function networks. At first, we use k-means clustering to find the centres which are generated from the conditional distribution $\Pr(\boldsymbol{x}|y)$ instead of the marginal distribution $\Pr(\boldsymbol{x})$. The estimated centres are expected to have some information on the joint distribution $\Pr(\boldsymbol{x}, y)$. Next, we estimate weight parameters of the radial basis function network regression models by maximum penalized likelihood approach with Lasso penalty. One of the most outstanding features of penalized likelihood with Lasso penalty is the sparsity of the solution. The sparse estimates (i.e., in which irrelevant weight parameters are exactly set to be zero) are attractive because they are equivalent to a structural simplification of the estimated network model.

In penalized likelihood approach based on Lasso penalty, there is a regularization parameter which controls the balance of model fitting and model complexity, and controls the degree of sparseness of the estimates. In model building process, there are critical problems concerned with the choice of the regularization parameter and a smoothing parameter which controls the scale of the basis functions. These are commonly adjusted using cross-validation methods, and are time consuming. We derive an information criterion for choosing the additional parameters.

We investigate the performance of the proposed nonlinear regression modeling procedure and evaluation criterion with robot arm data. Finally, some concluding remarks are given.

2 Radial basis function network regression model

Suppose that we have n observations $\{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, n\}$ consisting of p-dimensional explanatory variable $\boldsymbol{x}$ and the corresponding value y, from a Gaussian nonlinear regression model as followed by

$$y_i = u(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Here, $u(\cdot)$ is an unknown function and ε_i are independent and identically normal distributed with mean 0 and variance σ^2.

The purpose of regression problem is to infer a functional relation $y = u(\boldsymbol{x})$ with a set of observations. We represent $u(\boldsymbol{x})$ as an expansion in radially symmetric nonlinear basis functions and take the following form

$$u(\boldsymbol{x}) = \sum_{k=1}^{M} w_k \phi_k(\boldsymbol{x}) + w_0 = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi} \tag{2}$$

where $\boldsymbol{\phi} = (\mathbf{1}, \phi_1(\boldsymbol{x}), \ldots, \phi_M(\boldsymbol{x}))^{\mathrm{T}}$ are a set of the constant $\mathbf{1}$ and M basis functions, and $\boldsymbol{w} = (w_0, w_1, \ldots, w_M)^{\mathrm{T}}$ are output weights and Radial basis function networks achieve flexibility by fitting simple models in a receptive region local to the data point $\boldsymbol{x}$. A localization is achieved via a radial basis function $\boldsymbol{\phi}(\cdot)$. In this paper, we use the following Gaussian basis function

$$\phi_k(\boldsymbol{x}) = \phi_k(\|\boldsymbol{x} - \boldsymbol{\mu}_k\|) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2}{2\nu}\right), \quad k = 1, \ldots, M \tag{3}$$

where $\boldsymbol{\mu}_k$ is a centre of the k-th basis function, ν is a smoothing parameter which controls the amount of overlapping among the basis functions, and $\|\cdot\|^2$ is taken to be l_2 norm. It follows from equation (1), (2) and (3) that the response variable y are drawn from a family of the distributions with densities

$$f(y_i|\boldsymbol{x}_i; \boldsymbol{w}, \sigma^2, \boldsymbol{\mu}_k, \nu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left\{y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \nu)\right\}^2\right]. \quad (4)$$

3 Estimation

To construct the nonlinear regression models based on Gaussian radial basis functions in (4), it is necessary to estimate the number of the basis function, M, the positions of the centres, $\boldsymbol{\mu}_k$, the weights, $\boldsymbol{w}$, the variance, σ^2, and the smoothing parameter, ν. There are several approaches to learning these parameters. The simplest approach is to maximize the log-likelihood of the nonlinear regression model with respect to all the parameters. However, this approach is a full nonlinear optimization problem which will typically be computationally intensive and may be prone to having nonconvex solution sets with multiple local minima. In addition, if the number of observed data points is less than the number of all the parameters, the problem is ill-posed and generally an infinite set of maximum likelihood solutions exist.

Another approach is to estimate the positions $\boldsymbol{\mu}_k$ of the centres separately from the weights $\boldsymbol{w}$. Having obtained the former, the latter is a simple optimization problem. The $\boldsymbol{\mu}_k$ are commonly chosen in an unsupervised way using the observation data $\boldsymbol{x}$ alone. The estimated $\hat{\boldsymbol{\mu}}_k$ are then kept fixed while the weights $\boldsymbol{w}$ are optimized. Next we discuss in detail how to choose the parameters in the two-stage procedure and propose a new procedure for model building. The stage in a simple implementation are as follows.

(1). Determine the positions of the centers by k-means clustering for conditional distribution.

(2). Determine a candidate value from each of the regularization parameter and the smoothing parameter.

(3). Solve for the weights and the variance using maximum penalized likelihood approach with Lasso penalty.

(4). Calculate the information criterion for the estimated model.

(5). Iterate the procedure from step 2 to 4 for all of the candidate values.

(6). Choose the additional parameters which minimize the information criterion, and then calculate the final output.

Note that the number of the basis functions need not to be estimated since the regularization parameter determine the degree of sparseness of the weights which is equivalent to the number of the basis functions.

3.1 k-means clustering for conditional distribution

A popular choice for determining the centres $\boldsymbol{\mu}_k$ is the use of clustering techniques to find a set of centres which more accurately reflects the marginal distribution $\Pr(\boldsymbol{x})$ of $\boldsymbol{x}$. The obvious drawback of these approaches is that the conditional distribution $\Pr(y|\boldsymbol{x})$ is having no say in positioning of the centres.

So we propose a clustering technique to find a set of centres which reflects the conditional distribution $\Pr(\boldsymbol{x}|y)$ instead of $\Pr(\boldsymbol{x})$. The key idea is Sliced Inverse Regression (SIR) proposed by Li [7]. SIR is a dimension reduction method based on partitioning the range of the one dimensional response variable y into a fixed number H of slices denoted by $\mathcal{S}_1, \ldots, \mathcal{S}_H$. Then, p dimensional variable $\boldsymbol{x}$ is regressed on $\tilde{y}$ which is the discrete version on y resulting from slicing its range; we can decompose the joint density of $\boldsymbol{x}$ and y to the conditional density $\Pr(\boldsymbol{x}|y)$ and marginal density $\Pr(y)$.

In this case, we apply k-means clustering for each slice and choose all of the estimated centres. But we have a serious problem; what number k of clusters is most appropriated for each slice? To solve the problem, we use the resampling approach proposed by Roth et al. [11] and choose k which minimize this criterion. The algorithm consists of the following steps.

(1). Sort the data $\{\boldsymbol{x}_i, y_i\}$ by y_i.

(2). Divide the data into the non-overlapping slices $\mathcal{S}_j,\ j = 1, \ldots, H$.

(3). Within each slice, split the object set of size $2n$ into two sets of equal size, $\mathcal{O}_1^n$ and $\mathcal{O}_2^n$.

(4). Apply k-means clustering.to the two datasets separately. The result are the mapping α_1 of each of the objects in $\mathcal{O}_1^n$ to one of k clusters and the mapping α_2 of each of the objects in $\mathcal{O}_2^n$ to one of k clusters.

(5). Compare two clustering solutions. See Roth et al. [11] for detail. Use α_2 to predict the cluster membership of all objects contained in the first set.

(6). Find the correct permutation of labels by using the Hungarian method since the label indices might be permuted.

(7). Calculate the costs for identifying labels i and j which are the number of miss-classifications with respect to the labels $Y^n = \alpha_i(\mathcal{O}_1^n)$.

(8). Iterate the procedure from step 3 to 7, average over assignment costs and compute the expected stability value.

(9). Iterate the procedure from step 3 to 8 for each k to be tested.

(10). Iterate the procedure from step 3 to 9 for each slice and set the centres of the cluster in each slice to be the centres of the basis functions.

3.2 Penalized likelihood and lasso-type estimators

Having obtained centres, we need to estimate the weight parameters $\boldsymbol{w}$. The maximum likelihood estimators of the weights $\boldsymbol{w}$ and the error variance σ^2

are given by

$$\hat{\boldsymbol{w}} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}\boldsymbol{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i;\nu)\right\} \tag{5}$$

where $\boldsymbol{\Phi} = (\boldsymbol{\phi}(\boldsymbol{x};\nu),\ldots,\boldsymbol{\phi}(\boldsymbol{x};\nu))^{\mathrm{T}}$ and $\boldsymbol{y} = (y_1,\ldots,y_n)^{\mathrm{T}}$. However, in the settings of nonparametric regression models, the maximum likelihood criterion often leads to be overfitting, so that it yields unstable parameter estimators and often tends to be ill-posed problem. To avoid overfitting and ill-posedness, we maximize the penalized log-likelihood

$$l_{\lambda}(\hat{\boldsymbol{w}}) = \sum_{i=1}^{n}\log f(y_i|\boldsymbol{x}_i;\boldsymbol{w}) - n\lambda\eta(\boldsymbol{w}) \tag{6}$$

where λ is a regularization parameter which controls model fitting and model complexity and $\eta(\boldsymbol{w})$ is a penalty term. Typical forms for the penalty term can be represented as a quadratic form $\eta(\boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{w}$. In this paper, we use an absolute penalty instead of the quadratic penalty as followed by

$$\eta(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 = |w_1| + |w_2| + \ldots + |w_M|. \tag{7}$$

This penalty is called Lasso penalty. Least absolute shrinkage and Selection Operator, called Lasso, is proposed by Tibshirani [13] within the framework of the least square problem. Because of the property of Lasso penalty, making λ sufficiently large cause some of the weights to be exactly zero.

3.3 Model evaluation criteria

In order to choose the adjusted parameters λ and ν, we present an information criterion for evaluating nonlinear regression models by radial basis function networks and Lasso-type estimators.

A information criterion proposed by Akaike [1], known as AIC, is a criterion for evaluating a statistical model estimated by maximum likelihood and is given by

$$\mathrm{AIC} = -2\sum_{i=1}^{n}\log f(\boldsymbol{x}|\hat{\theta}_{ML}) + 2p \tag{8}$$

where $\hat{\boldsymbol{\theta}}_{ML}$ is the maximum log-likelihood estimate and p is the number of estimated parameters within the model. For nonlinear models by maximum penalized likelihood method, we need to replace p by some measure of model complexity. Hastie and Tibshirani [3] proposed to use the trace of the "hat matrix" given by $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ where $\hat{\boldsymbol{y}}$ is the fitted value using the estimated model.

However, in this case, $\boldsymbol{H}$ is not well-defined since the penalized likelihood function in (6) is not differentiated. So we approximate the penalized likelihood function by a quadratic in the neighborhood of a maximum penalized likelihood estimator $\hat{\boldsymbol{w}}$. The approximation is then

$$\begin{aligned} l_\lambda(\boldsymbol{w}) &\approx l_\lambda(\hat{\boldsymbol{w}}) + \nabla l_\lambda(\hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{w}-\hat{\boldsymbol{w}}) \\ &\quad + \frac{1}{2}(\boldsymbol{w}-\hat{\boldsymbol{w}})^{\mathrm{T}}\nabla^2 l(\hat{\boldsymbol{w}})(\boldsymbol{w}-\hat{\boldsymbol{w}}) - \frac{n\lambda}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{w} \end{aligned}$$

where

$$\nabla l(\boldsymbol{w}) = \frac{\partial}{\partial \boldsymbol{w}} l(\hat{\boldsymbol{w}}), \quad \nabla^2 l(\hat{\boldsymbol{w}}) = \frac{\partial}{\partial \boldsymbol{w}\partial \boldsymbol{w}^{\mathrm{T}}} l(\hat{\boldsymbol{w}}) \tag{9}$$

and

$$\boldsymbol{Z} = \mathrm{diag}\ (I_1, \ldots, I_M)\,, \quad I_k = \begin{cases} 1, & \hat{w}_k \neq 0 \\ 0, & \hat{w}_k = 0 \end{cases} . \tag{10}$$

Let d to be the number of the nonzero weights. Then the "hat matrix" is given by

$$\boldsymbol{H} = \boldsymbol{\Phi}_* \left(\boldsymbol{\Phi}_*^{\mathrm{T}} \boldsymbol{\Phi}_* + n\lambda\hat{\sigma}^2 \boldsymbol{I}_* \right)^{-1} \boldsymbol{\Phi}_*^{\mathrm{T}} \tag{11}$$

where $I_* = \mathrm{diag}(0, 1, 1, \ldots, 1)$ is a $(d+1) \times (d+1)$ diagonal matrix, $\boldsymbol{\Phi}_* = (\boldsymbol{\phi}_*(\boldsymbol{x}_1; \nu), \ldots, \boldsymbol{\phi}_*(\boldsymbol{x}_n; \nu))^{\mathrm{T}}$ and $\boldsymbol{\phi}_*(\boldsymbol{x}_i; \nu) = (\mathbf{1}, \phi_1(\|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|), \ldots, \phi_d(\|\boldsymbol{x}_i - \boldsymbol{\mu}_d\|))^{\mathrm{T}}$. By replacing the number of parameters p in AIC in (11) by the trace of $\boldsymbol{H}$, we obtain an information criteria for the nonlinear Gaussian regression model based on radial basis function networks and Lasso-type estimators in the form

$$\mathrm{AIC}_M = n \log(2\pi\hat{\sigma}^2) + n + 2\mathrm{tr}\ \boldsymbol{H}. \tag{12}$$

where $\hat{\sigma}^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2/n$.

Hurvich, Simonoff & Tsai [5] gave an improved version of AIC [12] for choosing a smoothing parameter in various types of nonparametric regression models. This type of a criterion of the nonlinear Gaussian regression model with radial basis function networks and Lasso-type estimators is followed by

$$\mathrm{AIC}_c = n \log(2\pi\hat{\sigma}^2) + n + 2n\frac{\mathrm{tr}\ \boldsymbol{H} + 2}{n + \mathrm{tr}\ \boldsymbol{H} - 1}. \tag{13}$$

We choose the regularization parameter λ and the hyperparameter ν which minimized the AIC_M in (15) or the AIC_c in (16).

4 Simulation

We use robot arm data[1] as a benchmark to compare the proposed modelling procedure with other neural network approaches. This data set involves implementing a model to map a two-dimensional joint angle (x_1, x_2) to the end arm position (y_1, y_2). The true function is given by

$$y_1 = 2.0\cos(x_1) + 1.3\cos(x_1 + x_2) + \varepsilon_1 \tag{14}$$

$$y_2 = 2.0\sin(x_1) + 1.3\sin(x_1 + x_2) + \varepsilon_2 \tag{15}$$

[1] http://wol.ra.phy.cam.ac.uk/mackay/

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 0.05$. We use the first 200 observations of the data to estimate our models and the last 200 observations to evaluate them. In this example, we set the number of the slices to be 4.

Konishi et al. [6] proposed generalized Bayesian information criteria for model evaluation by maximum penalized likelihood method with quadratic penalty and apply this criteria for the selection of the smoothing parameter, the regularization parameter, and the number of basis functions (using k-means clustering for marginal distribution $\Pr(x)$) in radial basis function network models. They compares their modelling strategy with other neural network approaches [2], [4], [8], [9], [10]. The results are presented in Table 1. The proposed modelling procedure appears more accurate than other neural network approaches on this data set.

Method	Mean Square Error
Mackay's (1992) Gaussian Approximation with HE	0.00573
Mackay's (1992) Gaussian Approximation with LTE	0.00557
Neal's (1996) hybrid MC	0.00554
Neal's (1996) hybrid MC with ARD	0.00549
Rios Insua and Müller's (1998) MLP with R-J MCMC	0.00620
Holmes and Mallick's (1998) RBF with R-J MCMC	0.00535
Andrieu *et al.*'s (2001) R-J MCMC with Bayesian model	0.00502
Andrieu *et al.*'s (2001) R-J MCMC with MDL	0.00512
Andrieu *et al.*'s (2001) R-J MCMC with AIC	0.00520
Konishi *et al.*'s (2004) RBF with BIC	0.00509
Proposed modelling procedure with AICm	0.00302
Proposed modelling procedure with AICc	0.00301

Table 1: Mean squared errors for the robot arm data.

5 Conclusion

We have developed a new model building procedure based on radial basis function networks; positioning the centres with k-means clustering for the conditional distribution $\Pr(\boldsymbol{x}|y)$, estimating the weights by maximum penalized likelihood with Lasso penalty, and deriving a model evaluation criteria for choosing the regularization and smoothing parameters. Through the simulation, we show that our procedure catch out the true structure generating the data and perform better than some other methods.

References

[1] Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd Inter. Symp. on Information Teory* (Petrov, B. N. and Csaki, F. eds.), Akademiai Kiado, Budapest, 267-281.

[2] Andrieu C., De Freitas N. & Doucet A. (2001). Robust full Bayesian learning for radial basis networks. *Neural Comp.* **13**, 2359-407.

[3] Hastie T. J. & Tibshirani R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

[4] Holmes C. C. & Mallick B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural Comp.* **10**, 1217-233.

[5] Hurvich C. M., Simonoff J. S. & Tsai C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information Criterion. *J. R. Statist. Soc.* **B 60**, 271-293.

[6] Konishi S., Ando T. & Imoto S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91**, 1, 27-43.

[7] Li K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**(414), 316-327.

[8] MacKay D. J. C. (1992). A practical Bayesian framework for backpropagation networks. *Neural Comp.* **4**, 448-72.

[9] Neal R. M. (1996). *Bayesian Learning for Neural Networks.* Lecture Notes in Statistics 118. New York: Springer-Verlag.

[10] Rios Insua D. & Müller P. (1998). Feedforward neural networks for nonparametric regression. In *Practical Nonparametric and Semeparametric Bayesian Statistics*, ED. D. K. Dey, P. Müller and D. Sinha, 181-91. New York: Springer Verlag.

[11] Roth V., Lange T., Braun M. and Buhmann J. M. (2002). A Resampling Approach to Cluster Validation. *Computational Statistics - COMPSTAT 2002*, Physica Verlag.

[12] Sugiura N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. Theor. Meth.* **A 7**, 13-26.

[13] Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

Address: T. Shimamura, M. Mizuta, Graduate School of Engineering, Hokkaido University and Infomation Initiative Center, Hokkaido University, N. 11 W. 5, Kita-ku, Sapporo 060-0811, Japan

E-mail: shima@cims.hokudai.ac.jp

EWMA COMBINATION OF BOTH GARCH AND NEURAL NETWORKS FOR THE PREDICTION OF EXCHANGE RATE

So Young Sohn and Won H. Shin

Key words: Forecasting, GARCH, neural networks, EWMA, combining.

COMPSTAT 2004 section: Time series analysis.

Abstract: Exchange rate forecasting is an important and challenging task for both academic researchers and business practitioners. Several statistical or artificial intelligence approaches have been applied to forecasting exchange rate. The recent trend to improve the prediction accuracy is to combine individual forecasts in the form of the simple average or weighted average where the weight reflects the inverse of the prediction error. This kind of combination, however, does not reflect the current prediction error more than the relatively old performance. In this paper, we propose a new approach where the forecasting results of GARCH and neural networks are combined based on the weight reflecting the inverse of EWMA of the mean absolute percentage error (MAPE) of each individual prediction model. Empirical study results indicate that the proposed combining method has better accuracy than GARCH, neural networks, and traditional combining methods that utilize the MAPE for the weight.

1 Introduction

Exchange rate forecasting is an important and challenging task for both academic researchers and business practitioners. Recent trend of related research is use of either GARCH (generalized autoregressive conditional heteroskedastic) models or neural networks. GARCH was first proposed by Bollerslev [2] and was applied to exchange rate forecasting by Hsieh [3]. A typical finding is that GARCH models provide superior forecasts of volatility than most conventional econometric models which assume homoscedasticity.

In recent years, there has been a growing interest to adopt the artificial intelligence technologies to predicting exchange rate volatility. One stream of these advanced techniques has focused on the use of neural networks [7], [8].

The weakness of the individual forecasting based on either GARCH or neural network is its dependence on a single forecasting model that is expected to capture all aspects of the volatility formation process. To overcome the limitation of individual forecasting approaches, combining methods of individual forecasts have been applied. Bates and Granger [1] advocated the use of weighted average in the combining forecasts. The weights are calculated from the variance and covariance of different forecast errors. Similarly Russell and Adam [5], Schwaerzel and Rosen [6], and Zhang and Joung [8] combined individually predicted results using the weights which are obtained from MSE (Mean Squared Error), MAE(Mean Absolute Error), or MAPE (Mean Absolute Percentage Error) of individual models.

In the existing combining methods, the errors of between actual and predicted values are equally reflected to the weights regardless of time order in a forecasting horizon. This kind of weight is slow to adopt the dynamic changes in a timely manner. In this paper, we propose a new approach where the forecasting results of GARCH and neural networks are combined based on the weight reflecting the inverse of EWMA (exponentially weighted moving average) of the absolute percentage error (APE) of each individual prediction model. Our approach is then compared to other combining methods suggested by Russell and Adam [5], Schwaerzel and Rosen [6], and Zhang and Joung [8] along with GARCH and neural networks. The main advantage of EWMA lies in its dynamic property which guarantees the fast adaptability to the change of individual model's performance.

The remainder of this paper is organized as follows. Section 2 presents the individual prediction models along with combining algorithms. In section 3, comparison study based on exchange rate data is performed and the results are summarized. Conclusion is given in section 4.

2 Exchange rate forecasting model

In this section we describe four different models applicable to the prediction of exchange rate: GARCH, Neural networks, EWMA combination, and MAPE combination.

The GARCH model
For the parameterization of time varying conditional variance on past information, the GARCH(p, q) model for exchange rate (y_t) at time t specifies the conditional variance as the following linear function of q lagged ε_t^2 and p lagged h_t:

$$\begin{aligned} y_t &= y_{t-j}\gamma_{t-j} + \varepsilon_t, \quad \text{where} \quad \varepsilon_t|\Omega_{t-1} \sim N(0, h_t) && (1) \\ h_t &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} && (2) \end{aligned}$$

where y_{t-j} is a vector of explanatory variables including possibly lagged dependent variables and γ_{t-j} is a vector of coefficients, p and q are the orders of the process, and Ω_{t-1} is information set available at time t.

Neural networks model
A backpropagation neural network consists of different sets of nodes and the connections between one set of nodes to the other. Each connection between two nodes in different sets assigned a weight that shows the strength of connection.

For effective training of such neural networks, optimal network structure needs to be tuned. Like other application domain, problems associated with deciding proper number of hidden layers and nodes are still not resolved for time series forecasting as well.

EWMA combination

We introduce EWMA approach in an effort to combine individually predicted results from both GARCH and neural networks. In EWMA combining method, the weight associated with each individual prediction model is determined in a manner to reflect the current prediction error more than the relatively old performance. In other words, our proposed weights can reflect the change of performance of models dynamically on changing time t. In general, EWMA of time series data st is defined as follows [4]:

$$Z_t = \lambda s_t + (1-\lambda)s_{t-1}, \tag{3}$$

where $0 < \lambda \leq 1$ is a constant and $Z_0 = \overline{s}$.

It is well known that EWMA Z_t is reduced to the form of a weighted average of all the previous samples available:

$$Z_t = \lambda \sum_{j=0}^{t-1} (1-\lambda)^j s_{t-j} + (1-\lambda)' Z_0 \tag{4}$$

In the light of combining algorithm, we define s_t to be the absolute percentage error (APE) of individual prediction model such as GARCH or neural networks at time t. That is $s_t = |y_t - \hat{y}_t^i|/y_t$ where y_t is actual exchange rate at time t and $\hat{y}_t^i$ is a predicted value for time t by individual model i, respectively. Subsequently, Z_0 can be obtained as the mean absolute percentage error so that EWMA of APE of individual prediction model i at time t can be expressed as follows:

$$\mathrm{EWMA}_{i,t} = \lambda \sum_{j=0}^{t-1} (1-\lambda)^j \frac{|y_{t-j} - \hat{y}_{t-j}^i|}{y_{t-j}} + (1-\lambda)' \cdot \mathrm{MAPE}_i\,, \tag{5}$$

where MAPE_i is the mean of APE of individual model i.

Next, we propose the EWMA combining method by applying the inverse of EWMA as the weight for the individual forecasting model where the weight w_i can be written as follows at a fixed training time period T:

$$w_i = \frac{\mathrm{EWMA}_i^{-1}}{\sum_{i=1}^{N} \mathrm{EWMA}_i^{-1}} \tag{6}$$

where N indicates the total number of forecasts to be combined.

Based on equation (6), we combined individually predicted results $\hat{y}_{t-j}^i$ directly using weight w_i as follows:

$$\hat{Y}_{\mathrm{combine}} = \sum_{i=1}^{N} w_i \hat{y}_t^i\,. \tag{7}$$

In this way, the predicted value at time t from the individual model associated with less EWMA gets a larger weight, where the weight reflects more the current performance than the relatively old one.

On the other hand, in the existing MAPE combining, the weight wi can be re-written as follows:

$$w_i = \frac{\mathrm{MAPE}_i^{-1}}{\sum_{i=1}^{N} \mathrm{MAPE}_i^{-1}} \tag{8}$$

In the next section, using Korean won/US dollar exchange rate data, we compare EWMA combining method with GARCH, neural networks and the MAPE combining method. We use two popular criteria (RMSE and MAPE) to evaluate the predictive performance of forecasting models.

3 Case study

Research data used in this study is the exchange rate between Korean won and US dollar. The data contain the daily observations from January 1999 to February, 2002, giving 918 data points. The time series plotted in Fig. 1 show numerous turning points. Also, seasonality is observed along with slightly increasing trend.

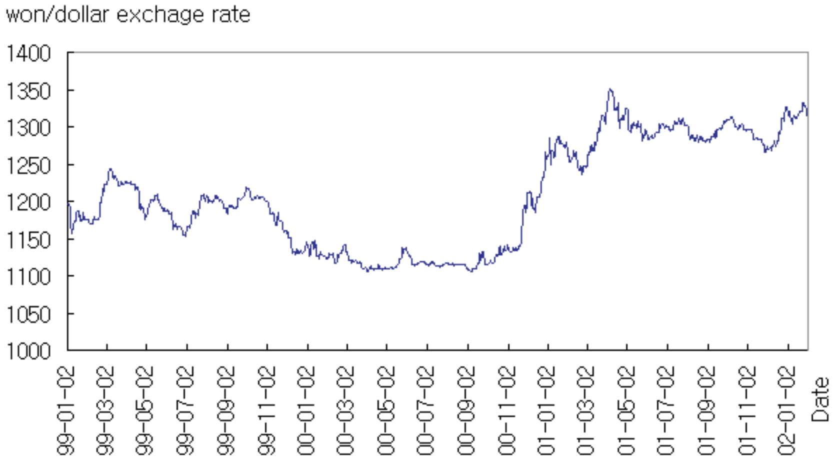


Figure 1: Daily won/dollar exchange rate series (1999 2002).

In order to fit individual forecasting models, we use three lagged variables (y_{t-1}, y_{t-2}, and y_{t-7}), which are selected based on AIC (Akaike 's Information Criterion).

We focus on one-step ahead forecasts. In one step ahead forecasting, the exchange rate values are forecasted one step ahead at a time and the actual values are then used for the next prediction in a forecasting horizon. We split the data into two: training data involves observations from January 1 1999 to January 31 2002 while the validation data involves observation from February 1 2002 to February 28 2002. For validation, parameters of individual models (GARCH, neural networks) were determined as follows. Parameters of GARCH are selected as GARCH(1,1) based on AIC. GARCH(1,1) process has been widely used for modeling of financial data since Bollerslev [1].

For construction of neural networks, fifteen and ten hidden nodes are assigned for two hidden layers, respectively and they are connected by logistic activation function. This model selection is determined by trial and error for optimal setting.

In EWMA combining method, we apply different values of 0.1, 0.5, and 0.9 for λ to see the performance change due to the value of λ. Since typically actual value for the forecasted one is not available, we apply the same weight for the forecasting period. Table 1 contains the MAPE and MSE of each model obtained based on the test data.

	MAPE (%)	MSE
GARCH (1,1)	2.5874	0.0068
Neural Network	2.8465	0.0087
MAPE combining method	2.5155	0.0045
EWMA Combining method ($\lambda = 0.1$)	2.5229	0.0043
EWMA Combining method ($\lambda = 0.5$)	2.3347	0.0041
EWMA Combining method ($\lambda = 0.9$)	2.0613	0.0023

Table 1: MAPE of won/dollar exchange rate forecasting model.

This empirical result represents that GARCH has better accuracy than neural networks. Also, both EWMA and MAPE combining methods are better than the individual forecasting model. This would be due to the fact that the most of the actual values are in between the individually predicted values. Thus we could obtain the weighted average effect. On the other hand, EWMA combining method has better accuracy than that of MAPE combining method. Table 2 shows the weights used for individual models for each combining method. Apparently, as the value of λ becomes large, the large weight is assigned on GARCH. It reflects that GARCH performs especially better than neural networks on relatively newer won/dollar exchange rate data. It means that GARCH has getting a smaller error gradually than neural networks as time elapsed on average. Therefore we assign a larger weight to GARCH than neural network, and the results show that EWMA combining method can do the proper work.

	GARCH (1,1)	Neural networks
MAPE combining	0.525	0.475
EWMA combining method ($\lambda = 0.1$)	0.612	0.388
EWMA combining method ($\lambda = 0.5$)	0.647	0.353
EWMA combining method ($\lambda = 0.9$)	0.792	0.218

Table 2: Weights of individual models for combining model.

4 Discussion

We propose a combining method by applying the inverse of EWMA of APE as the weight for the individual forecasting model at time t. In this way, the predicted value for time t from the individual forecasting model associated with less EWMA gets the larger weight. The weight reflects more the current performance than the relatively old one. Next, using won/dollar exchange rate data, we compare EWMA combining method with GARCH, neural networks and the combining method based on MAPE, that is time irrelevant. Results of empirical study indicates the following. First, GARCH has better accuracy than neural networks. Secondly, combining methods outperform individual models (GARCH, neural networks). Lastly, the proposed EWMA combining method has better accuracy than the others in won/dollar exchange rate data in terms of MAPE. This appears to be due to the fact that EWMA combining method has more weight on GARCH than neural networks.

We expect that EWMA combining method would have a better performance when the accuracy of individual model changes over time. In order to get more rigorous results, superiority of EWMA combining method needs verification based on Monte Carlo simulation. This is left for one of future study areas.

References

[1] Bates J.M., Granger C. (1969). *The combination of forecasts.* Operational Research Quarterly **20**, 451 – 468.

[2] Bollerslev T. (1986). *Generalized autoregressive conditional heteroskedasticity.* Journal of Econometrics **31**, 307 – 400.

[3] Hsieh D.A. (1988). *The statistical properties of daily foreign exchange rates 1974-1983.* Journal of International Economics **24**, 129 – 145.

[4] Montgomery D.C. (1996). *Introduction to statistical quality control.* John Wiley & Sons, New York.

[5] Russell T.D., Adam E.E. (1987). *An empirical evaluation of alternative forecasting combinations.* Management Science **33** (10), 1267 – 1376.

[6] Schwaerzel R., Rosen B. (1997). *Improving the accuracy of financial time series prediction using ensemble networks and high order statistics.* Proc. of the International Conference on Neural Networks **4**, 2045, – 2050.

[7] Yao J., Tan C.L. (2000). *A case study on using neural networks to perform technical forecasting of forex.* Neurocomputing **34** (1), 79 – 98.

[8] Zhang B., Joung, J. (1999). *Time series prediction using committee machines of evolutionary neural trees.* Proc. of the 1999 Congress on Evolutionary Computation **1**, 281-286.

Acknowledgement: This work was supported by grant No. (R04-2002-000-20003-0) from the Basic Research Program of the Korea Science & Engineering Foundation.

Address: H.W. Shin, ISamsung Economy Research Institute, Seoul, Korea
S.Y. Sohn, Dept. of Industrial Systems Engineering, Yonsei University, Seoul, Korea

E-mail: `won@seri.org, sohns@yonsei.ac.kr`

COMPSTAT'2004 Symposium

TREE HARVEST: METHODS, SOFTWARE AND SOME APPLICATIONS

Roberta Siciliano, Massimo Aria and Claudio Conversano

Key words: Fast algorithm, ensembles, multi-class response, discriminant trees, data editing.

COMPSTAT 2004 section: Tree-based methods.

Abstract: Harvesting Trees gathers past and recent results on tree-based methods focalizing the attention on both the software implementation of suitable procedures for special types of data sets (i.e., complex data structure, multi-class response, set of within-groups correlated predictors, missing data) and the perception of tree results in real world case studies applications. Main issue is to make feasible the idea of trees as a powerful tool to provide information which is *statistically reliable* and *with an added value* in terms of problem solving and knowledge discovery.

1 Trees and software

Segmentation [8] aims to build up an oriented tree graph playing the role of a model describing the supervised classification or regression for a response variable on the basis of a set of predictors (of categorical or numerical type), all variables measured on a sample of objects. A recursive split of the predictor space determines a split of the objects into two subgroups where the response variable presents an always decreasing degree of impurity, as measured by either heterogeneity for classification trees or variation for regression trees. Terminal nodes define a partition of the objects into exhaustive and disjoint groups, each labeled by either the modal class or the average value. Paths of exploratory trees, as described by the split conditions from the root node to terminal nodes, can be interpreted to explain the dependence relationships of the predictors on the response variable. Trees can be also used to assign a value/class to a new object for that only the measurements of the predictors are known, namely the object goes into the tree structure reaching a terminal node which label determines the final prediction. Accuracy can be evaluated considering an error rate estimator based on mean square error for regression trees and misclassification rate for classification trees. The trade-off between the tree size and the prediction accuracy is considered for identifying the decision rule for new objects prediction.[1] The great success of tree-based methods can be attributed to some key factors, the possibility to

[1] CART methodology considers a cost-complexity measure in a pruning algorithm to define a set of subtrees and to select the final decision tree optimizing the cross-validation or test sample estimates. Recent proposals are bagging and boosting estimators that improve the prediction accuracy.

analyse huge data sets, the simple structure of the partitioning algorithm to grow trees, the simplicity of interpretation of the tree results. Computational enhancements to accelerate the tree growing (i.e., fast algorithms) as well as to improve their prediction performance (i.e., bagging, boosting) increased their popularity as a fundamental tool for exploratory and decision purposes within the Knowledge Discovery Process.

Harvesting trees is the project of few researchers to develop methodological results and a specialized software for both statisticians and non-statisticians to handle data and to apply various tree-based methods in productive and fancy way. Main issue is to gather past and recent results on segmentation methods for standard as well as non standard data structures within a specialized software supported by an handy guide user interface. Special care is paid on the visualization aids associated to a tree graph and on alternative partitioning criteria to handle special structures of data as well as to add fruitful information within each node of the tree.

Tree Harvest (TH) is an interactive software developed in MATLAB 6.5 characterized by an intensive use of graphical tools and computational procedures in order to apply various segmentation routines in a new fashion[2]. Specific procedures have been implemented respectively for accelerating the ensemble algorithms, for defining a suitable multi-class classifier, for dealing with a set of within-group correlated predictors, for handling data editing tasks such as missing data and data validation. TH supports the analyst with an interactive guide user interface (fig. 1) and various features for data exploration and prediction evaluation (fig. 1), such as the use of colors to discover patterns in data exploration, the interactive description of local results within the node, the complementary use of trees and models.

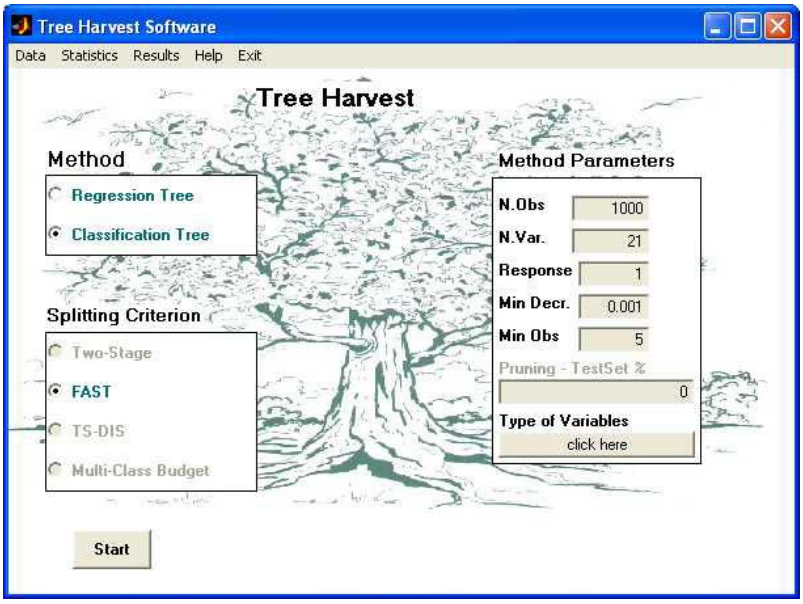
In the next section, we highliht some of the methodologies implemented in the TH software and provide main results deriving from applications on real and simulated data.

2 Tree-growing via FAST

The basic tree-growing routine in TH software is Fast Algorithm for Splitting Trees (FAST) [11]. It grows the tree structure maximizing the *Impurity Proportional Reduction* (IPR) at each node of the tree but using less computing time (in terms of number of splits to be tried out before finding the best split) than CART segmentation procedure. At each node, FAST iterates a two-stage splitting criterion maximizing first a global IPR[3] over all predictors to find the best predictor and then a local IPR over all splits generated

[2] TH is an extension of the software ET (*Exploratory Trees*) recently introduced by [2] and also described by [3]. It includes procedures not only for exploratory analysis but also for decision trees and model prediction [10]

[3] Global IPR is a measure of the relative decrease of impurity of Y given each modality of the predictor X. It is equivalent to the predictability index τ of Goodman and Kruskal when the impurity is Gini's index of heterogeneity for classification trees and Pearson's correlation coefficient when the impurity is variation for regression trees.



Method

○ Regression Tree
● Classification Tree

Splitting Criterion

○ Two-Stage
● FAST
○ TS-DIS
○ Multi-Class Budget

Method Parameters

N.Obs 1000
N.Var. 21
Response 1
Min Decr. 0.001
Min Obs 5
Pruning - TestSet % 0
Type of Variables click here

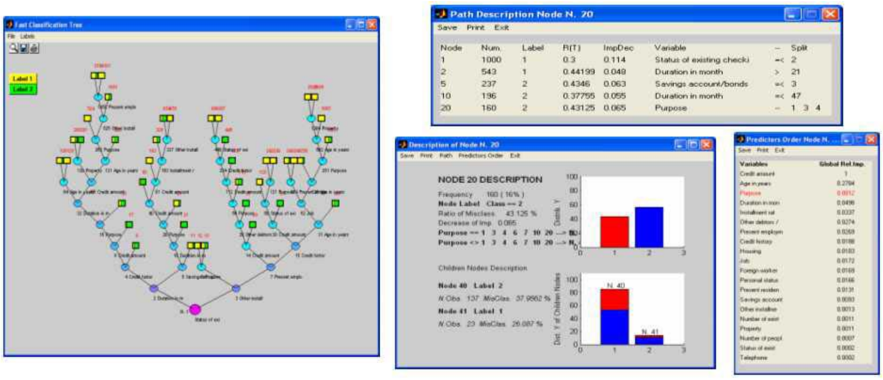
Figure 1: TH Graphical User Interface. User can select the type of analysis (classification or regression), the splitting criterion and specify the method parameters.

by the best predictor; it stops the iterations if the global IPR of the subsequent best predictor does not improve the local IPR of the current best split. FAST performance is preferred over standard segmentation especially in two case studies, namely in presence of variables with high predictability power on the response variable as well as when dealing with categorical predictors with many categories[4]. In both cases, FAST identifies the best split in very few iterations due to the selection of the best predictor in the first stage. A FAST tree-growing accelerates also the prediction estimate based on resampling techniques such as bagging and boosting as shown in table 1 and in figure 2).

3 Multi-class BUDGET TREE

Standard segmentation procedures fail when dealing with multi-attribute response variables, especially in presence of a non-uniform distribution of objects among the J classes. TH offers a solution through Multi-Class Budget

[4] Numerical predictors can be categorized according to a suitable algorithm based on a clustering optimization criterion [2] so that they can play the same role of categorical predictors in the analysis. In presence of mixed predictors, it is well known that splits at the upper part of the tree are likely generated by numerical predictors.



Path Description Node N. 20

Node	Num.	Label	R(T)	ImpDec	Variable	--	Split
1	1000	1	0.3	0.114	Status of existing checki	=<	2
2	543	1	0.44199	0.048	Duration in month	>	21
5	237	2	0.4346	0.063	Savings account/bonds	=<	3
10	196	2	0.37755	0.055	Duration in month	=<	47
20	160	2	0.43125	0.065	Purpose	--	1 3 4

Description of Node N. 20

NODE 20 DESCRIPTION

Frequency 160 (16%)

Node Label Class == 2

Ratio of Misclass 43.125 %

Decrease of Imp. 0.065

Purpose == 1 3 4 6 7 10 20 --> N. 40

Purpose <> 1 3 4 6 7 10 20 --> N. 41

Children Nodes Description

Node 40 Label 2

N.Obs. 137 *MisClas.* 37.9562 %

Node 41 Label 1

N.Obs. 23 *MisClas.* 26.087 %

Predictors Order Node N. ...

Variables	Global Rel.Imp.
Credit amount	1
Age in years	0.2794
Purpose	0.0912
Duration in mon	0.0498
Installment rat	0.0337
Other debtors /	0.0274
Present employm	0.0269
Credit history	0.0198
Housing	0.0183
Job	0.0172
Foreign worker	0.0169
Personal status	0.0166
Present residen	0.0131
Savings account	0.0093
Other installme	0.0013
Number of exist	0.0011
Property	0.0011
Number of peopl	0.0007
Status of exist	0.0002
Telephone	0.0002

Figure 2: A bottom-up representation of a Classification Tree using the FAST splitting algorithm (left panel) for the German Credit dataset (Machine Learning Repository). Different colors are assigned to different types of nodes and a visual inspection of the distribution of the response in each terminal node is allowed. A simple click on a terminal node allows to visualise interactively its main features. In the right panel the nodes description window is shown. The user can visually interact with the tree through queries. For each node, the software provides useful information (node label, misclassification rate, impurity reduction, splitting information, node paths). The same information is provided for the child nodes. For each split, a ranking of the predictors is provided on the basis of the IPR.

Trees (MC-BudgetTree) [1]. At each node of the tree, a given number of predictors is selected according to a global IPR measure and for each of them the modalities are cross-classified with the response classes. In this way, it is possible to fit a model for categorical data, such as the latent budget model[5], choosing the best among the selected predictors on the basis of a suitable goodness of fit measure. Main issue is the criterion used to partition the objects at each node into K groups on the basis of the mixing parameter estimates of the latent budget model. For binary trees, i.e., $K = 2$ latent budgets, objects in node t fall into the left child node t_{left} if the estimate[6] $\hat{\pi}_{i|k=1} \geq 0.5$ and into the right child node t_{right} if the estimate $\hat{\pi}_{i|k=2} > 0.5$[7].

[5] Latent Budget Model is a reduced-rank model to decompose the conditional distributions $\pi_{j|i}$ of a cross-classification of two categorical variables into a mixture of K latent budgets, where both mixing parameters $\pi_{k|i}$ and latent budgets $\pi_{j|k}$ are conditional probabilities summing up to one over the index k and the index j respectively.

[6] The parameter estimates are provided by a least-sqaures criterion and identified according to a suitable algorithm [1].

[7] The split of the I modalities into two subgroups generates the split of the objects, in particular the i-th modality is assigned to the k-th budget which is linked to the highest mixing parameter.

				N.Iter.	Boosting		Bagging	
Dataset	Obs.	Variables	Splits	Cart	Fast Aver.	Gain	Fast Aver.	Gain
German Credit	1000	20 (13)	542952	542952	903	600.27	893	607.00
Ionosfere	351	34 (0)	8114	8114	4662	0.74	4804	0.69
Breast Cancer	699	9 (9)	80	80	19	3.21	20	3.00
Pima Indians	768	8 (0)	1246	1246	898	0.38	932	0.34
SouthAfr. Heart.	463	9 (1)	675	675	631	0.07	587	0.15
Credit SPAD	468	11(2)	30	30	12	1.50	15	1.00
WDBC	569	30(0)	3306	3306	2011	0.64	2034	0.62
SpamBase Stump	4601	56 (0)	7389	7389	1874	2.94	1923	2.84
Simulated Stump	12000	10 (0)	99990	99990	13845	6.22	13973	6.15

Table 1: Performance of FAST algorithm for accelerating bagging and boosting procedures using real data sets (Machine Learning Repository and SPAD library) and the simulated data set stump (Hastie et al., 2002). There are reported the number of observations, the number of variables (in bracket the number of categorical predictors), the number of splits and the number of iterations at the top node, the gain measures (in terms of reduced number of iterations) when applying the FAST algorithm and the CART algorithm with bagging and boosting procedures.

The latent budget parameter estimates in the children nodes express the fuzzy assignment of the objects to the J classes of the response variable. In addition, a measure of discrepancy between the latent budget and the mean budget, i.e., $d_{j|k} = \frac{\pi_{j|k} - p_{.j}}{p_{.j}}$ (where $p_{.j}$ is the prior class proportion assignment), provides, by means of its sign, to assign each response class to one of the two child nodes. Thus, the MC-Budget Tree defines a strenght measure of class assignment enriching in this way the within node interpretation (fig. 4).

4 DIScriminant TREE

Exploratory trees can be also grown via discriminant functions (DIS-TREE). This will prove particularly convenient when dealing with a data set organized as a set of within-group correlated predictors and when the number of predictors is high with respect to the number of objects. The idea is to summarise the information through a set of discriminant variables, each is a combination of the original predictors with weights derived by the canonical variates. The procedure is implemented for both classification trees [4] and regression trees [12], the latter making use of the new concepts of prospective and retrospective splits[8]. DIS-TREE provides multiple splits at each node of the tree, which interpretation can be facilitated by looking at the discriminant analysis coefficients (fig. 5) in order to rank the discriminant variables importance.

[8] The procedure for regression trees determines, at each node, a dummy response variable maximizing the IPR with respect to the numerical response (i.e., the so-called best retrospective split) without caring of the predictors. A fast algorithm can be applied to find the best multiple split of the objects maximizing the IPR with respect to all candidate splits due to the discriminant variables (i.e., the so-called prospective split).

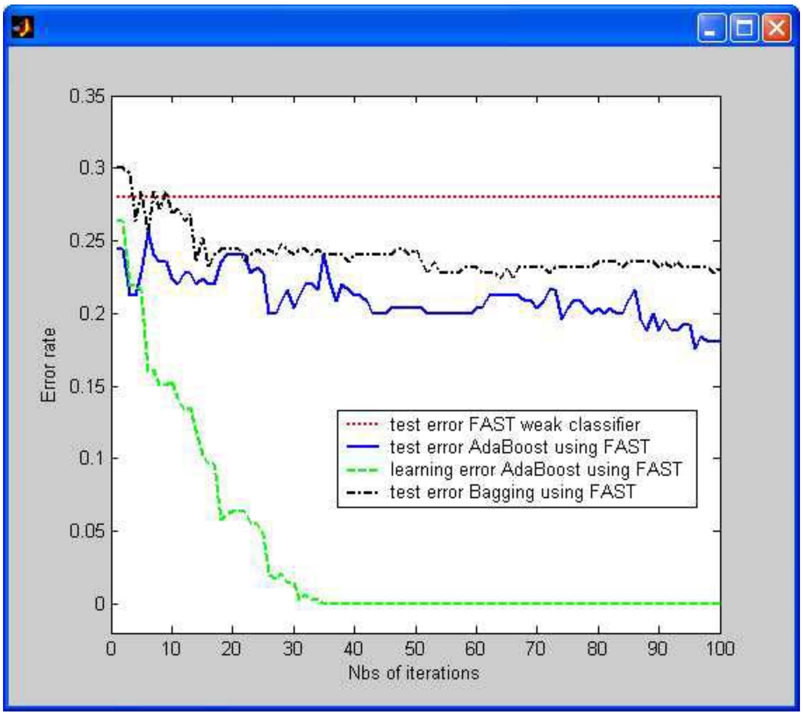
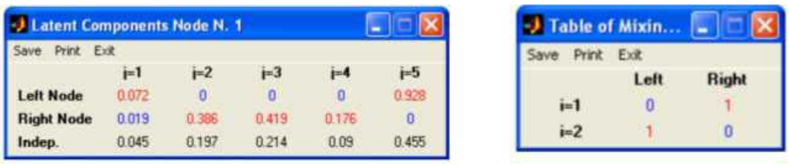


Figure 3: Performance of tree classifiers using the FAST splitting algorithm on the German Credit dataset(Machine Learning Repository).

5 Trees for data editing and data modeling

Tree-based methods have proved to be very useful for data editing as well as for data modelling. In particular, TH is going to include a missing data imputation procedure which is particularly useful in presence of multiple missing data. The algorithm first performs a lexicographic ordering of the objects with respect to the number of missing values of each row/column; the imputation of missing values is then performed incrementally using trees (Incremental Non-Parametric Imputation) [6]. A further implementation will concern a novel strategy based on TREEs and VALidation (TREEVAL) for the automated derivation of edits and data validation [9].

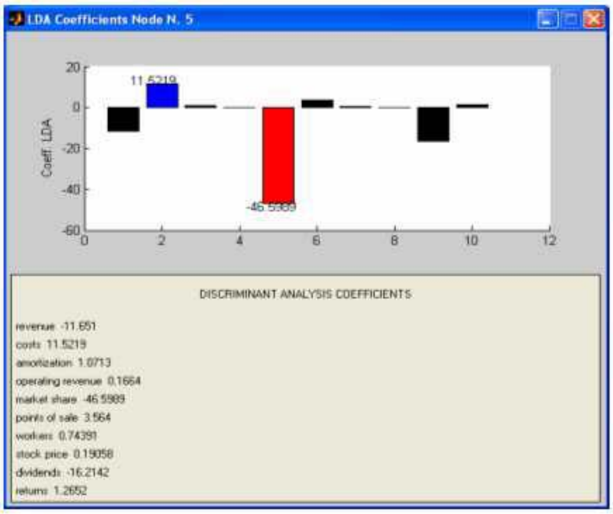
A final task will be to implement trees in semiparametric regression modelling framework to be used for classification and prediction. In fact, trees can be considered as an alternative to scatterplot smoother estimators in generalized additive regression modellng. This approach has been used in the so-called Generalized Additive Multiple-Models (GAM-M) and Generalized Additive Multi-Mixture Models (GAM-MM) frameworks [5]. Here, trees are used with classical scatterplot smoothers for defining a semiparametric regression model using different types of smoothers/classifiers or even mixtures



	i=1	i=2	i=3	i=4	i=5
Left Node	0.072	0	0	0	0.928
Right Node	0.019	0.386	0.419	0.176	0
Indep.	0.045	0.197	0.214	0.09	0.455

	Left	Right
i=1	0	1
i=2	1	0

Figure 4: Multi-class BUDGET TREE node description. For each terminal node, a distribution of both the latent components and the mixing parameters is provided.



DISCRIMINANT ANALYSIS COEFFICIENTS

revenue -11.651
costs 11.5219
amortization 1.0713
operating revenue 0.1664
market share -46.5989
points of sale 3.564
workers 0.74391
stock price 0.19058
dividends -16.2142
returns 1.2652

Figure 5: DIS-TREE Segmentation: node information about the splitting discriminant variable coefficients (simulated data).

of them. This approach, based on combination of classifiers is somewhat similar to Bagging and Boosting [8]. TH is going to implement these approaches with the aim to provide powerful and flexible tools for the anlysis of complex data structures.

6 Concluding remarks

Tree Harvest software collects various tree-based methods for dealing with standard as well as non standard data structures. TH satisfies properties to be *light* (with respect to the technological implementation), *easy* (concerning the interpretation capability in getting proper information), *direct* (with respect to the data which are employed), *accessible* (in terms of cost of implementation), *quick* (referring to the timeless of the strategies to be proposed to the final decision-maker). Such complementary tools allow for harvesting trees in various types of exploratory analysis and decision-making process.

References

[1] Aria M. (2004). *Multi-class budget exploratory trees.* In Studies in Classification, Data Analysis, and Knowledge Organization, Springer-Verlag, to appear.

[2] Aria M. (2003). *An interactive and accessible software for statistical learning through exploratory trees.* (In italian). Phd thesis in Computational Statistics, Dept. of Mathematics and Statistics, University of Naples Federico II, Italy.

[3] Aria M., Siciliano R. (2003). *Learning from trees: two-stage enhancements.* In Proceedings of CLADAG 2003, 22-24 September, Bologna.

[4] Cappelli C., Conversano C. (2002). *Canonical variates for recursive partitioning in data mining.* In Hardle W. and Ronz B. (eds.), COMPSTAT 2002 Proceedings, Physica-Verlag, Heidelberg, 213 – 218.

[5] Conversano C. (2003). *Bagged mixtures of classifiers using model scoring criteria.* Journal of Pattern Analysis and Applications **5**, (4), 351 – 362.

[6] Conversano C., Siciliano R. (2004). *Incremental tree-based imputation with lexicographic ordering.* In Minotte M., Swzychak A. (eds.), Interface 2003 Proceedings, Washington, to appear.

[7] Hand D.J., Mannila H., Smyth P. (2001). *Principles of data mining.* The MIT Press.

[8] Hastie T.J., Tibshirani R.J., Friedman J. (2001). *The elements of statistical learning.* Springer Verlag.

[9] Petrakos G., Conversano C., Farmakis G., Mola F., Siciliano R., Stavropulos P. (2004). *New ways to specify data edits.* Journal of the Royal Statistical Society **167**, Part 2, 249 – 274.

[10] Siciliano R. (1998). *Exploratory versus Decision Trees.* Invited lecture to COMPSTAT '98 (Bristol, August 24-28), in R. Payne, P. Green (Eds.): Proceedings in Computational statistics: 13th Symposium of COMPSTAT, Physica Verlag, Heidelberg (D).

[11] Siciliano R., Mola F. (2000). *Multivariate data analysis through classification and regression thees.* Computational Statistics and Data Analysis **32**, Elsevier Science, 285 – 301.

[12] Siciliano R., Mola F. (2002). *Discriminant analysis and factorial multiple splits in recursive partitioning for data mining.* In Roli, F., Kittler, J. (eds.), Proceedings of International Conference on Multiple Classifier Systems (Chia, June 24-26, 2002), 118 – 126, Lecture Notes in Computer Science, Springer, Heidelberg.

Acknowledgement: Financial support from MIUR of Italy.

Address: R. Siciliano, M. Aria, Dipartimento di Matematica e Statistica, Università di Napoli Federico II, Monte S. Angelo, via Cintia, 80126 Napoli, Italy

C. Conversano, Dipartimento di Economia e Territorio, Università di Cassino, via Mazzaroppi, 03043, Cassino Italy

E-mail: `roberta@unina.it; aria@unina.it; c.conversano@unicas.it`

APPROPRIATE CROSS-VALIDATION FOR REGULARIZED ERRORS-IN-VARIABLES LINEAR MODELS

Diana M. Sima and Sabine Van Huffel

Key words: Cross-validation, errors-in-variables model, regularization.

COMPSTAT 2004 section: Model selection.

Abstract: In the context of errors-in-variables ill-conditioned linear models, Tikhonov regularization is often employed. In this paper it is shown that the error function used by a cross-validation criterion for choosing a good regularization parameter must be based on the 'generalization error' instead of the 'prediction error', which is used in ordinary linear regression. This observation leads to a new cross-validation criterion that is based on orthogonal distances. A consistency theorem is also proved. Numerical experiments sustain the superiority of the new approach in comparison with classical methods for selecting the regularization parameter.

1 Introduction

Consider a slightly incompatible ill-conditioned linear system, $Ax \approx b$, $A \in \Re^{m\times n}$, $b \in \Re^m$, $x \in \Re^n$. In the Least Squares (LS) setting, it is assumed that the inconsistency of the system is due only to noise in b. In many applications, it is reasonable to take into account that the matrix A might also be contaminated by noise. This is an errors-in-variables model and, equivalently, it is the approach of the Total Least Squares (TLS) problem [4], [9]. Some linear problems have an intrinsic ill-conditioning; therefore, classical estimation methods (least squares, total least squares) provide physically meaningless solutions, due to numerical sensitivity. This motivates the introduction of regularization. The most commonly used method is the Tikhonov formulation of the Regularized Least Squares (RLS) problem:

$$\min_x \left(\|Ax-b\|_2^2 + \lambda\|Qx\|_2^2\right), \tag{1}$$

where $\lambda > 0$ is the *regularization parameter*, and Q is a given weight matrix through which the size of x is measured. Methods for choosing a good λ are based on the goal that the regularized solution is an appropriate solution for $Ax \approx b$, but the size of $\|Qx\|_2$ is also kept under control. Among the most popular methods are: the discrepancy principle [7], (generalized) cross-validation [3], L-curve criterion [5], etc.

In this paper, a new technique for regularization parameter selection in the context of the errors-in-variables model is proposed. It is based on the classical cross-validation criterion. Section 2 shortly introduces, in a general setting, the classical cross-validation criterion and its application in regularization problems. Section 3 highlights that a classical cross-validation criterion based on a *prediction error* loss function is not appropriate for errors-in-variables models. Instead, a cross-validation criterion based on a *generalization error* loss function is proposed. It is then proved that the latter criterion can be defined based on orthogonal distances; its most important property, similar to the consistency property of the Total Least Squares solution, is that for a growing degree of overdetermination (*i.e.*, m increasingly large) this criterion provides a consistent estimator of the *best regularized solution*. Finally, Section 4 shows convincing numerical results in favor of the new criterion.

2 Cross-validation criterion

Cross-validation is a widely-used technique that allows selecting amongst statistical models. Its various forms range from parameter selection to probability density estimation, from classification to stopping criteria in training neural networks.

Its simple philosophy relies on repeatedly splitting the available data into estimation and evaluation parts, and picking the model that minimizes a certain criterion applied to the evaluation parts.

In a simplified framework, let $\{Z_1, \dots, Z_m\}$ denote given data that comes from an unknown model $\mathcal{M}$, which is parameterized by an unknown parameter λ. Let $\{I_1, I_2, \dots, I_c\}$ be a partition of the set of indices $\{1, 2, \dots, m\}$, each of size p. (This implies the condition $m = pc$, which not a necessary restriction, but it is used to simplify notation.) For a fixed parameter λ and each set I_j, a 'partial' model $\mathcal{M}_{-I_j}(\lambda)$ can be estimated using only the data from $\{Z_1, \dots, Z_m\}$ with indices in $\{1, 2, \dots, m\} \backslash I_j$. Then the 'performance' of the partial model is estimated under a certain error function L, and the cross-validation function is defined as $V(\lambda) := \frac{1}{c} \sum_{j=1}^{c} L(\mathcal{M}_{-I_j}(\lambda), Z_{I_j})$. The non-negative-valued function L must measure the error of assuming that the partial model $\mathcal{M}_{-I_j}(\lambda)$ describes also the samples Z_{I_j} (which are not used in the process of constructing $\mathcal{M}_{-I_j}(\lambda)$).

The value of λ that minimizes V is selected as the cross-validation parameter and it is used to construct the cross-validated model $\mathcal{M}(\lambda)$.

2.1 Cross-validation for regularized least squares

In regularized least squares using Tikhonov regularization (see (1)), the regularized solution with regularization parameter λ is $x(\lambda) = (A^T A + \lambda Q^T Q)^{-1} A^T b$. Using the formalism introduced before, the Z_i variables are the rows $(A_i \ b_i)$ of the data matrix $[A \ b]$ and the model $\mathcal{M}(\lambda)$ is parameterized by the solution $x(\lambda)$. Similarly, the partial models $\mathcal{M}_{-I_j}(\lambda)$ are replaced by the partial solutions $x_{-I_j}(\lambda) = (A_{-I_j}^T A_{-I_j} + \lambda Q^T Q)^{-1} A_{-I_j}^T b_{-I_j}$. (Here, A_{I_j}, b_{I_j} denote the rows of A, elements of b, respectively, with indices in I_j, and A_{-I_j}, b_{-I_j} denote the rows of A, elements of b, respectively, with indices in $\{1, 2, \ldots, m\}\backslash I_j$.) The error function is the quadratic loss function and the classical cross-validation function for regularized least squares is:

$$L(x, [A_{I_j}\ b_{I_j}]) = \|A_{I_j} x - b_{I_j}\|_2^2, \quad V_{\text{RLS}}(\lambda) = \frac{1}{c}\sum_{j=1}^{c} \|A_{I_j} x_{-I_j}(\lambda) - b_{I_j}\|_2^2. \quad (2)$$

3 Cross-validation and errors-in-variables linear models

In this section, it will be assumed that the linear model generating the incompatible system $Ax \approx b$ is an *errors-in-variables model.* Specifically, it is assumed that there exists an exact linear relation $A_0 x_0 = b_0$ and that A and b are perturbed measurements of A_0 and b_0, $A = A_0 + \delta A$, $b = b_0 + \delta b$; moreover, the elements of $[\delta A\ \delta b]$ are independent, identically distributed, with zero mean and variance σ^2. The case of interest is when A_0 is ill-conditioned, and the noisy data matrix A (although it is *better conditioned* than A_0) is also ill-conditioned, with no significant gap in the singular values.

3.1 Prediction error vs. generalization error

Let $\hat{x}$ be an arbitrary estimator of x_0. In this context, it is important to make a distinction between *prediction* and *generalization errors.* Let $[C\ d]$ denote a row (or several rows) of measured data from the same errors-in-variables model. In the context of prediction, the estimator $\hat{x}$ is used to compute $\hat{d} := C\hat{x}$, which is a predicted value for d.

Definition 3.1. $\|d - \hat{d}\|_2^2$ *is called the prediction error.*

Let $\Delta d := \hat{d} - d$. Due to noise in $[C\ d]$ it is probable that $C\hat{x} = d$ is not readily satisfied. Using Δd, a compatible system is constructed: $C\hat{x} = d + \Delta d$. Note that this system corrects only the right-hand-side, whereas the noisy matrix C remains unaltered. A redundant way of defining Δd is via the trivial (feasible set is a singleton) optimization problem: $\min_{\Delta d} \|\Delta d\|_2^2$ subject to $C\hat{x} = d + \Delta d$. This optimization problem is introduced for comparison with the following similar problem, which is related to the *generalization* error:

$$\min_{\Delta C, \Delta d} \|[\Delta C\ \Delta d]\|_F^2 \quad \text{subject to } (C + \Delta C)\hat{x} = d + \Delta d. \quad (3)$$

In words, the optimal solution of problem (3) consists of the smallest corrections that must be added to $[C\ d]$ in order to make the equation $C\hat{x} \approx d$ compatible. Let $[\widehat{\Delta C}\ \widehat{\Delta d}]$ be the optimal solution of (3) and define $\hat{C} := C + \widehat{\Delta C}$ and $\hat{d} := d + \widehat{\Delta d}$. Then $\hat{C}\hat{x} = \hat{d}$ is satisfied.

Definition 3.2. $\|[C\ d] - [\hat{C}\ \hat{d}]\|_F^2$ *is called the generalization error.*

3.2 Cross-validation for prediction or generalization

The cross-validation function uses a certain error function L in order to assess the performance of the partial models (constructed from the estimation parts) onto the evaluation parts. Applied to the errors-in-variables context, two error functions are compared in the following. They are based on the prediction error and the generalization error, respectively.

Definition 3.3. *The prediction error function is defined as* $L^{pred}(\hat{x},[C\ d]) := \|C\hat{x}-d\|_2^2$; *the generalization error function is* $L^{gen}(\hat{x},[C\ d]) := \frac{\|C\hat{x}-d\|_2^2}{\|\hat{x}\|_2^2+1}$.

The justification of the previous definition is straightforward in the case of L^{pred}; for L^{gen}, it is clarified by the following lemma (see also [9, Thm.6.5]):

Lemma 3.1. *The optimal solution of the optimization problem (3) is given by*

$$\widehat{\Delta C} = -\frac{(C\hat{x}-d)\hat{x}^T}{\|\hat{x}\|_2^2+1}, \qquad \widehat{\Delta d} = \frac{C\hat{x}-d}{\|\hat{x}\|_2^2+1}, \tag{4}$$

and the optimal value of the generalization error is equal to $\frac{\|C\hat{x}-d\|_2^2}{\|\hat{x}\|_2^2+1}$.

Proof. Defining the Lagrangean of (3) as $\mathcal{L}(\Delta C, \Delta d, v) = \|[\Delta C\ \Delta d]\|_F^2 + 2v^T((C+\Delta C)\hat{x} - d - \Delta d)$ (with v - the vector of Lagrange multipliers), the formulas (4) are easily derived from the first order optimality conditions: $\Delta C = -v\hat{x}^T, \qquad \Delta d = v, \qquad (C+\Delta C)\hat{x} = d + \Delta d.$ □

The two error functions, L^{pred} and L^{gen}, lead to the definition of two different cross-validation functions,

$$V^{pred}(\lambda) = \frac{1}{c}\sum_{j=1}^{c}\|A_{I_j}x_{-I_j}(\lambda) - b_{I_j}\|_2^2, \quad V^{gen}(\lambda) = \frac{1}{c}\sum_{j=1}^{c}\frac{\|A_{I_j}x_{-I_j}(\lambda) - b_{I_j}\|_2^2}{\|x_{-I_j}(\lambda)\|^2+1}. \tag{5}$$

Note that V^{pred} is identical to the function V_{RLS} in (2).

3.3 Optimal regularization parameter

It should be noted at this point that, depending on the shape of the model $\mathcal{M}(\lambda)$ (which means $x(\lambda) = (A^TA + \lambda Q^TQ)^{-1}A^Tb$ in the Tikhonov regularization case), there may be several definitions of the 'optimal' regularization parameter λ. For the numerical experiments in Section 4, the optimal λ was defined as the minimizer of $\|x(\lambda) - x_0\|_2$. Another choice might be, for instance, the minimizer of the angle between $x(\lambda)$ and x_0.

Assuming that minimizing $\|x(\lambda) - x_0\|_2$ is a good criterion to obtain a meaningful solution, let λ^{opt} be the optimal regularization parameter. (Note that λ^{opt} can be computed effectively *only* in simulation examples, when x_0 is known.) Clearly, any method for choosing λ cannot give a better regularized solution. Therefore the aim is to find a λ that gives an $x(\lambda)$ as close to $x(\lambda^{opt})$ as possible.

3.4 Consistency of the regularized solution obtained with the generalization error cross-validation

The following theorem is closely related to the consistency discussion for the Least Squares and Total Least Squares solutions [9, Chapter 8]. The notation $\sigma_{\min}(M)$ or $\sigma_{\max}(M)$ will denote the minimum, respectively, maximum singular value of a matrix M.

Theorem 3.1. *Let* λ^{pred} *and* λ^{gen} *be the minimizers of the cross-validation functions* V^{pred} *and* V^{gen}, *respectively, and let* λ^{opt} *be the optimal regularization parameter,* i.e., *the minimizer of* $\|x(\lambda) - x_0\|_2$. *If*

$$\lim_{m\to\infty} \sigma_{\min}(A) = \infty, \text{ and } \lim_{m\to\infty} \frac{\sigma_{\min}(A^T A)}{\sigma_{\max}(A_{I_j})} = \infty, \quad \forall j \in \{1,\dots,c\}, \tag{6}$$

and if $\exists \lim_{m\to\infty} \frac{1}{m} A_0^T A_0 =: F \in \Re^{n\times n}$*, such that*

$$\lambda^{opt} \text{ is } \textbf{not} \text{ the minimizer of } \|x_0 + (F + \sigma^2 I_n)^{-1} x_0 - x(\lambda)\|_2^2, \tag{7}$$

then, as $m \to \infty$*,*

a. $\|x(\lambda^{pred}) - x(\lambda^{opt})\|_2$ *is asymptotically biased away from zero;*

b. $\|x(\lambda^{gen}) - x(\lambda^{opt})\|_2 \to 0.$

Proof. The guiding ideas for the proof are given below, with the help of several lemmas. For the first auxiliary result, the definition of the cross-validation function for an arbitrary error function L, $V(\lambda) = \frac{1}{c}\sum_{j=1}^{c} L(x_{-I_j}(\lambda), [A_{I_j}\ b_{I_j}])$, must be contrasted with the definition of the *conditional risk* [1]:

Definition 3.4. *The conditional risk function is*

$$\tilde{V}(\lambda) = \frac{1}{c}\sum_{j=1}^{c} E_{[\delta C\ \delta d]}\left[L(x_{-I_j}(\lambda), [A_{I_j}^0 + \delta C\ \ b_{I_j}^0 + \delta d])\right], \tag{8}$$

where $E_{[\delta C\ \delta d]}$ *denotes the expectation taken with respect to the common density function of the elements of* $[\delta C\ \delta d] \in \Re^{p\times n}$*, which have the same characteristics as the noise* $[\delta A\ \delta b]$*.*

Denote the optimal cross-validation parameter by $\hat{\lambda} = \arg\min V(\lambda)$ and the optimal conditional risk parameter by $\tilde{\lambda} = \arg\min \tilde{V}(\lambda)$. Note that what differs between the two formulations is that the cross-validation function uses a *particular* noise realization $[\delta A_{I_j}\ \delta b_{I_j}]$ that is added to the true data $[\delta A_{I_j}^0\ \delta b_{I_j}^0]$, whereas $\tilde{V}$ takes the expectation of any possible added noise. $\tilde{V}$ is uncomputable (since the exact $[A_0\ b_0]$ are unknown), but it is used in the proof of Theorem 3.1, because of the following property established in [1] (in a different context and for a different purpose, however):

Lemma 3.2. *Under certain assumptions (see [1, Thm.1]),*[1] $\lim_{m\to\infty} |\tilde{V}(\tilde{\lambda}) - \tilde{V}(\hat{\lambda})| = 0$.

In other words, the cross-validation parameter $\hat{\lambda}$ is asymptotically optimal for $\tilde{V}$. This will allow to replace (at the limit $m \to \infty$) the minimization of V with the minimization of $\tilde{V}$, in order to prove the properties of the cross-validation parameter. Another replacement that may be done in the limit is: $x_{-I_j}(\lambda)$ by $x(\lambda)$, where $x(\lambda)$ is computed from all the m rows of the given data $[A\ b]$.

Lemma 3.3. $\lim_{m\to\infty} \|x(\lambda) - x_{-I_j}(\lambda)\|_2 = 0, \quad \forall j \in \{1,\dots,c\}$.

The proof follows from the expansion $x_{-I_j}(\lambda) = (A^T A - A_{I_j}^T A_{I_j} + \lambda Q^T Q)^{-1} (A^T b - A_{I_j}^T b_{I_j}) = x(\lambda) + (A^T A + \lambda Q^T Q)^{-1} A_{I_j}^T (A_{I_j} x_{-I_j}(\lambda) - b_{I_j})$, by bounding from above the norm $\|x(\lambda) - x_{-I_j}(\lambda)\|_2$ with a term proportional to $\sigma_{\max}(A_{I_j})/\sigma_{\min}^2(A)$. From (6), this ratio goes to zero as m goes to infinity.

The last auxiliary result is:

[1] The assumptions are not listed here, due to lack of space. Among these assumptions, the most troublesome is that [1] allows only a finite number of models to select from. In the present context, this implies that λ is constrained to belong to a discrete set.

Lemma 3.4. *If $C = C_0 + \delta C \in \Re^{p\times n}$ and $d = d_0 + \delta d \in \Re^p$, and all elements of $[\delta C\ \delta d]$ are i.i.d., have zero mean and variance σ^2, then, for any $x \in \Re^n$,*

$$E_{[\delta C\ \ \delta d]}\left[\|Cx - d\|_2^2\right] = \|C_0 x - d_0\|_2^2 + p\sigma^2(\|x(\lambda)\|_2^2 + 1). \tag{9}$$

This follows clearly from $E\left[\|\delta C x(\lambda) - \delta d\|_2^2\right] = p\sigma^2(\|x(\lambda)\|_2^2 + 1)$.

Proof of point a: ¿From Lemma 3.2 the optimal λ^{pred} can be obtained (when $m \to \infty$) by minimizing $\tilde{V}^{pred}$. This is written as

$$\begin{aligned}
\min_\lambda \tilde{V}^{pred}(\lambda) &= \min_\lambda \frac{1}{c}\sum_{j=1}^{c} E_{[\delta C\ \ \delta d]}\left[\|(A^0_{I_j} + \delta C)x_{-I_j}(\lambda) - (b^0_{I_j} + \delta d)\|_2^2\right] \\
&\approx \min_\lambda \frac{1}{c}\sum_{j=1}^{c}\left[\|A^0_{I_j}x(\lambda) - b^0_{I_j}\|_2^2 + p\sigma^2(\|x(\lambda)\|_2^2 + 1)\right] \qquad (10) \\
&= \min_\lambda \frac{1}{c}\,\|A_0 x(\lambda) - b_0\|_2^2 + p\sigma^2(\|x(\lambda)\|_2^2 + 1) \\
&= \frac{1}{c}\min_\lambda \left\| \begin{bmatrix} A_0 \\ \sqrt{m\sigma^2} I_n \\ 0 \end{bmatrix} x(\lambda) - \begin{bmatrix} b_0 \\ 0 \\ \sqrt{m\sigma^2} \end{bmatrix} \right\|_2^2 . \qquad (11)
\end{aligned}$$

Line (10) follows from Lemma 3.4, but the approximation sign is used because x_{-I_j} is replaced by $x(\lambda)$ (Lemma 3.3). The minimization (11) is a constrained least squares problem (the constraint being represented by the parameterization $x(\lambda)$). In the unconstrained situation (*i.e.*, $x(\lambda)$ replaced by a free x) the least squares solution is given by $x_{\text{LS}} = (A_0^T A_0 + m\sigma^2 I_n)^{-1} A_0^T b_0$, which is a biased estimator of x_0. It is easy to show that the solution $x(\lambda^{pred})$ of (11) equals also $\arg\min_{x(\lambda)} \left\| \begin{bmatrix} A_0 \\ \sqrt{m\sigma^2} I_n \end{bmatrix} (x_{\text{LS}} - x(\lambda)) \right\|_2^2 = \arg\min_{x(\lambda)} \left\{ \|A_0\,(x_{\text{LS}} - x(\lambda))\|_2^2 + m\sigma^2\,\|x_{\text{LS}} - x(\lambda)\|_2^2 \right\}$. As $m \to \infty$, $x(\lambda^{pred})$ will tend to be as close as possible to x_{LS}. Note that $x_{\text{LS}} = x_0 + \left(\frac{A_0^T A_0}{m} + \sigma^2 I_n\right)^{-1} x_0$; therefore, as $m \to \infty$, λ^{pred} is the minimizer of $\|x(\lambda) - x_{\text{LS}}\|_2 = \|x(\lambda) - x_0 - (F + \sigma^2 I_n)^{-1} x_0\|_2$. ¿From assumption (7), it follows that λ^{opt} cannot be, at the limit, equal to λ^{pred}, and the bias between the prediction error model and the optimal regularized model is therefore concluded.

Proof of point b: As in the previous case, Lemmas 3.2- 3.4 help writing the minimization of V_E^{gen} as $\min_\lambda \frac{1}{c}\frac{\|A_0 x(\lambda) - b_0\|_2^2}{\|x(\lambda)\|_2^2 + 1} + p\sigma^2 \iff \min_\lambda \frac{\|A_0 x(\lambda) - b_0\|_2^2}{\|x(\lambda)\|_2^2 + 1}$. In the unconstrained case ($x(\lambda)$ replaced by a free x), this problem is a trivial (noiseless) TLS problem, which yields the exact solution x_0. In the singular value decomposition of $[A_0\ b_0]$, the largest n singular values go to infinity in the limit (see (6)); the smallest one is 0 and corresponds to the right singular vector $[x_0^T\ -1]^T/\|[x_0^T\ -1]^T\|_2$. This implies that the optimal solution $x(\lambda^{gen})$ should be as close as possible to x_0. From the definition of λ^{opt}, it follows $\lim_{m\to\infty} \|x(\lambda^{gen}) - x(\lambda^{opt})\|_2 = 0$. □

4 Numerical experiments

The following experiment illustrates the consistency property of the new cross-validation criterion. Random problems of growing size m are used. The matrix A_0 is generated with the function `regutm` from the Regularization Toolbox [6]; thus, A_0 is ill-conditioned, with exponentially decreasing singular values, and random (left and right) singular vectors. The exact solution is set to $x_0 = \left(\left(\frac{1}{n}\right)^2, \left(\frac{2}{n}\right)^2, \ldots, \left(\frac{n}{n}\right)^2\right)^T$, and b_0 is computed as $b_0 = A_0 x_0$. White noise of $\sigma\%$ is added to $[A_0\ b_0]$ in order to obtain $[A\ b]$; several different noise realizations are then used to compute average relative errors. Figure 1 shows the behavior of the average relative errors obtained with three methods for computing the regularization parameter. For reference, the best possible error (obtained for the λ that minimizes $\|x(\lambda) - x_0\|_2$) is shown. The experiment demonstrates that with increasing m the new cross-validation estimator performs better and better, while the other estimators don't have this 'consistency' property.

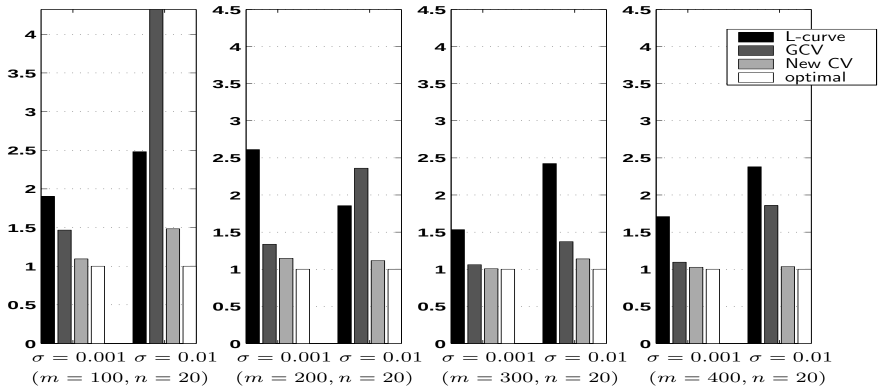


Figure 1: Average relative errors between Tikhonov-regularized solutions and exact solution x_0 when the regularization parameter is computed using the L-curve, the Generalized Cross-Validation criterion (from [6]), the new cross-validation for errors-in-variables and the 'optimal' regularization parameter. The latter is computed by minimizing the Euclidean distance between the regularized solution and the exact solution x_0. All values are scaled by dividing to the corresponding minimal average relative error (the fourth bar).

5 Conclusion

A new procedure for estimating regularization parameters in the context of linear errors-in-variables model was proposed. The advantages and applicability range of the new method were explained, and compared with the cross-validation criterium based on prediction errors. Numerical results confirmed these explanations.

On-going study is related to applying this type of cross-validation to a more specific formulation, namely when the estimated partial solutions are computed as *Regularized Total Least Squares* solutions [2], [8], instead of *Regularized Least Squares* (Tikhonov) solutions.

References

[1] Dudoit S., van der Laan M.J. (2003). *Asymptotics of cross-validated risk estimation in model selection and performance assessment.* U.C. Berkeley Division of Biostatistics Working Paper 126.

[2] Golub G.H., Hansen P.C., O'Leary D.P. (1999). *Tikhonov regularization and total least squares.* SIAM Journal on Matrix Analysis and Applications **21**, 185 − 194.

[3] Golub G.H., Heath M., Wahba G. (1979). *Generalized cross-validation as a method for choosing a good ridge parameter.* Technometrics **21**, 215 − 223.

[4] Golub G.H., Van Loan C.F. (1980). *An analysis of the total least squares problem.* SIAM Journal on Numerical Analysis **17**, 883 − 893.

[5] Hansen P.C. (1992). *Analysis of discrete ill-posed problems by means of the L-curve.* SIAM Review **32**, 561 − 580.

[6] Hansen P.C. (1994). *Regularization tools, a Matlab package for analysis of discrete regularization problems.* Numerical Algorithms **6**, 1 − 35.

[7] Morozov V.A.(1966). *On the solution of functional equations by the method of regularization.* Soviet. Math. Dokl. **7**, 414 − 417.

[8] Sima D.M., Van Huffel S., Golub G.H. (2003). *Regularized total least squares based on quadratic eigenvalue problem solvers.* Technical Report 03-07, ESAT-SISTA, K.U. Leuven. To appear in BIT.

[9] Van Huffel S., Vandewalle J. (1991). *The total least squares problem: computational aspects and analysis.* Volume 9 of *Frontiers in Applied Mathematics.* SIAM, Philadelphia.

Acknowledgement: Dr. Sabine Van Huffel is a full professor and Diana M. Sima is a research assistant at the Katholieke Universiteit Leuven, Belgium. Their research is supported by

Research Council KUL: GOA-Mefisto 666, IDO/99/003 and IDO/02/009 (Predictive computer models for medical classification problems using patient data and expert knowledge), several PhD/postdoc & fellow grants;
Flemish Government: FWO: PhD/postdoc grants, projects, G.0078.01 (structured matrices), G.0407.02 (support vector machines), G.0269.02 (magnetic resonance spectroscopic imaging), G.0270.02 (nonlinear Lp approximation), research communities (ICCoS, ANMMM); **IWT:** PhD Grants;
Belgian Federal Science Policy Office: IUAP P5/22 (Dynamical Systems and Control: Computation, Identification & Modelling);
EU: eTUMOUR, PDT-COIL,BIOPATTERN.

Address: D.M. Sima, S. Van Huffel, ESAT-SCD-SISTA, K.U. Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

E-mail: `diana.sima@esat.kuleuven.ac.be`,
`sabine.vanhuffel@esat.kuleuven.ac.be`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

SIMULATION AND MODELLING OF VEHICLE'S DELAY AT SEMI-ACTUATED SIGNALIZED INTERSECTIONS

Lurdes Simoes, P.M. Oliveira and A. Pires da Costa

Key words: Traffic simulation, queueing theory, server vacations, longitudinal analysis.

COMPSTAT 2004 section: Simulations, Optimization.

Abstract: Our aim is to analyse if the value of a given unitary increment to the green time, in intersections regulated by semi-actuated traffic signals, may result in significant changes in drivers' average delay, which is an important factor in the optimization of intersection performance. A longitudinal data analysis was applied to urban traffic simulations. Three scenarios with different unitary increment values for the green time were considered. A mixed-effects model relating the response to the average delay and the covariates has been established, the covariates being functional relations of the experimental variable. The latter represents the hourly debit of vehicles in the main street. We conclude, that models resulting from a combination between a second degree polynomial and an exponential term, are adequate. The unitary increment of green was shown to have a significant effect in the response, both in less intense flow rates and when the flows approach the maximum capacity of the intersection leg. Some numerical results and a brief comment on the use of queues with server vacations in this context is included at the end.

1 Introduction

In a longitudinal study, the individuals are observed during some periods of time. The data used in this work come from the simulation of road traffic in an urban intersection, between a secondary street and a main street. Mixed-effects models are primarily used to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors. The random effects in these models represent deviations of the individual parameters from the fixed effects. In some applications, these deviations arise from unexplained intergroup variation but, frequently, they can be at least partially explained by differences in covariate values among groups.

The most common application of mixed-effects models is for repeated data, in particular, longitudinal data [2]. At one level [5], the jth observation on the ith group is modeled as:

$$y_{ij} = f(\phi_{ij}, \nu_{ij}) + \epsilon_{ij}, i = 1, \ldots, M, j = 1, \ldots, n_i,$$

where M is the number of groups, n_i the number of observations on ith group, f a differentiable function (in $\mathbf{R}$) of a group specific parameter vector ϕ_{ij}, and a covariate vector ν_{ij}, and (ϵ_{ij}) is a normally distributed within-group error term. The function f is modeled as:

$$\phi_{ij} = A_{ij}\beta + B_{ij}b_i, \qquad b_i \sim \mathcal{N}(0, \Psi),$$

where β is a p-dimensional vector of fixed effects and b_i is a q-dimensional random effects vector associated with the ith group (not varying with j). The matrices A_{ij} and B_{ij} depend on the group and possibly on the values of some covariates, at the jth observation. This model allows the incorporation of covariates in the fixed effects or the random effects of the model.

We are interested in model the average delays suffered by drivers when crossing an intersection regulated by semi-actuated traffic signals [4]. We are especially interested in investigating a possible relation between the average delay suffered by each vehicle and the extension of green time previously regulated on the traffic signal. The representation of estimates of the random effects can be used to explore the relation between the variation of the delay and experimental factors in the regulation of traffic signals.

The data are balanced, the average delays of vehicles are measured for the same effective flow for the 90 simulations of 1 hour duration (individuals). Figure 1 shows the delay curves, when the flow in the secondary street is 200 veh/hour. The values of the flow in the main street adopted vary from 300 to 1200 veh/hour. These conditions of operation of semi-actuated intersection are frequently found in real world situations.

The data revealed a familiar pattern in plots of grouped data: the delay curves have a similar shape for different extensions of green but differ among individuals.

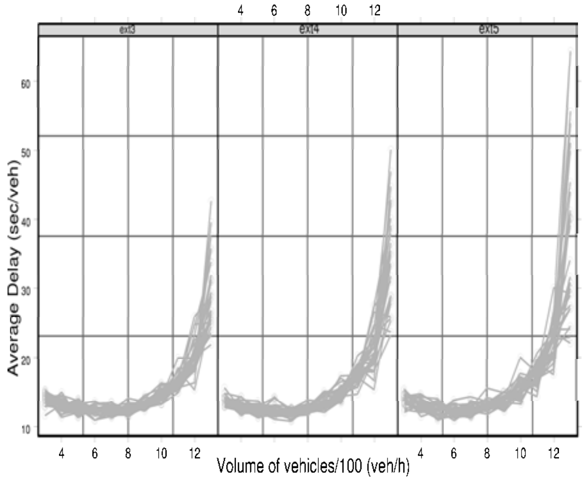


Figure 1: Average delay of the vehicles *v*ersus volume of vehicles in the main street, for extensions of green time of 3, 4 and 5 seconds, respectively.

2 Modelling of vehicles delay

One of the models suggested for the delay is of the form: $Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 e^t$, where Y_t represents the average delay suffered by vehicles when the volume of vehicles is t.

We will temporarily ignore the effect of extension green on grouping the average delays for each simulation and fit a single model to all the data. We express the average delay, y_{ij}, in individual i at flow t_j by

$$y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 t_j^2 + \beta_3 e^{t_j} + \epsilon_{ij}, \tag{1}$$

where the error terms, ϵ_{ij}, are assumed to be independently distributed as $\mathbf{N}(0, \sigma^2)$. These model is fit by nonlinear least squares and the information about the parameter estimates is presented in Table 1. Initial estimates are reasonable and the algorithm converges. The standard error for the parameters estimates are relatively small. Probably the most important drawback of using model (1) with grouped data is that it prevents us from understanding the true structure of the data and from considering different sources of variability that are of interest in themselves. To fit a separate model of this type to each group thus allowing the individual effects to be incorporated in the parameter estimates, we express the model as:

	Value	**Std. Error**	**t-value**
β_0	17.1997	0.81673	21.0594
β_1	-1.49256	0.25554	-5.8408
β_2	1.1506e-1	1.823e-2	6.312
β_3	3.7257e-5	1.736e-6	21.4611

Table 1: Parameter estimates for model (1).

$$y_{ij} = \beta_{0i} + \beta_{1i} t_j + \beta_{2i} t_j^2 + \beta_{3i} e^{t_j} + \epsilon_{ij}. \tag{2}$$

We can see from Table 2 that there is some variability in the parameter estimates of each group. The residual standard error is now 2.78 which is less than the residual obtained for model (1), of 2.97.

Extension (sec)	β_0	β_1	β_2	β_3
3	17.98445	-1.708505	0.1280751	2.731209e-5
4	17.02708	-1.479087	0.1147959	3.840112e-5
5	16.58434	-1.288872	0.1021925	4.610446e-5

Table 2: Parameter estimates for each group using model (2).

Naturally, model (2) is at the other extreme of flexibility compared to model (1). It uses 12 coefficients to represent the profiles of the average and does not take account of the obvious similarities among the groups. A model of type (2) is useful when one is interested in modelling the behaviour of a particular fixed set of individuals, but is no more adequate when the observed individuals are to be treated as a sample from a population of similar individuals, which constitutes the majority of applications involving grouped data. In this case, the interest is in estimating of average behaviour of a group of individuals in the population and the variability among and within groups, which is precisely what mixed effects models are designed for.

In order to include the fixed and random effects components in a linear model of *mixed-effects*, it is usual to rewrite model (2) in the form:

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j + (\beta_2 + b_{2i})t_j^2 + (\beta_3 + b_{3i})e^{t_j} + \epsilon_{ij}. \qquad (3)$$

The fixed effects β_0, β_1, β_2 and β_3 represent the mean values of the parameters in the population. The individual deviations are represented by the random effects b_{0i}, b_{1i}, b_{2i} and b_{3i}, which are assumed to be normally distributed with mean 0 and covariance matrix ψ. Random effects corresponding to different groups are assumed to be independent. The within-group errors, ϵ_{ij}, are assumed independently distributed as $\mathbf{N}(0, \sigma^2)$ and independent of the random effects. Model (3) gives a compromise between model (1) and the overparametrized model (2). It accommodates individual variations through the random effects but ties the different individuals together through the fixed effects and the covariance matrix ψ.

	Fixed Effects				**Random Effects**
	Value	Std.Error	t-value	p-value	StdDev
β_0	17.19811	0.7629162	22.54260	$<$ .0001	2.651385e-3
β_1	-1.49192	0.2387043	-6.25009	$<$ .0001	3.263414e-4
β_2	0.11500	0.0170275	6.75401	$<$ .0001	3.033456e-5
β_3	0.00004	0.0000046	8.17029	$<$ .0001	7.370581e-6

Table 3: Parameter estimates for model (3).

A crucial step in model building of mixed-effects models is to decide which of the coefficients in the model need random effects to account for their between-subject variation which can be treated purely as fixed effects. Plots of individual confidence intervals are used many times with this intention. One of the strategies consists of starting with a model with random effects for all parameters and then examine the fitted object, to decide which, if any, of the random effects can be eliminated from the model. The results are presented in Table 3. The standardized residuals versus the fitted values are plotted in Figure 2, showing a pattern of increasing variability for the within-group errors. Apparently the variability inside the groups increases with the volume of vehicles. We model the within group heterocedasticity using the exponential variance function presented in [5]. We will call model (4) the modification of model (3)to include the heterocedasticity component. Variance functions are used to model the variance structure of the within-group errors using covariates. Adequacy of the exponential variance function can be assessed by again plotting the standardized residual against the fitted values (Figure 3). A reasonable homogeneous pattern of variability for the standardized residuals is apparent. A primary question of interest is the

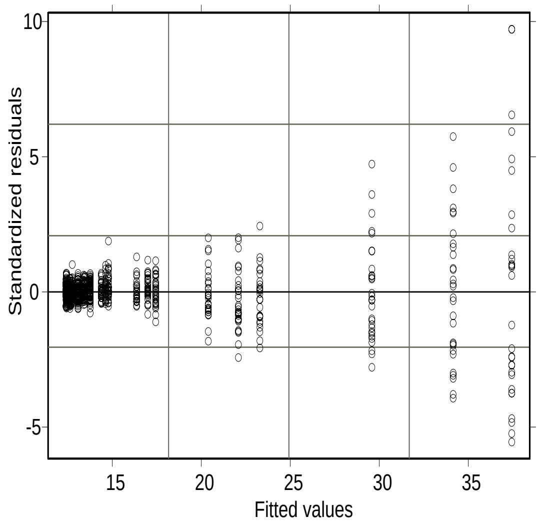


Figure 2: Scatter plot of standardized residuals *vs* fitted values for model (3).

Model	df	AIC^a	BIC^b	*logLik*	Test	L.Ratio	p-value
3	9	4846.31	4890.37	-2414.15			
4	10	3057.78	3106.74	-1518.89	3 vs 4	1790.53	$< .0001$

[a] *Akaike Information Criterion* [b] *Bayesian Information Criterion*

Table 4: Results of application of the ANOVA test to models (3) and (4).

possible relationship between the growth pattern of the delays and the experimental factor: extension of green time (group). To explore this relationship we will analyse the estimates of the random effects.

The normal plot of the within-group residuals for heterocedastic model (Figure 4) does not indicate severe violations of the assumption of normality for the within-group errors. Applying *ANOVA* test to models (3) and (4) we can confirm the importance of the heterocedasticity component (Table 4). The p-value is extremely small indicating that the more general model (4) is definitely a better fit. The *AIC* and *BIC* criteria confirm this. In brief, the heterocedastic model provides a much better representation of the data. The estimated standard deviation for the β_1 random effect in the fit of model (4) is only of 1.798e-5, corresponding to a negligible estimated coefficient of variation with respect to the β_1 fixed effect. This suggests that β_1 can be treated as a purely fixed effect. When we refit the model dropping β_1 random effect, we get a p-value of 0.0004 in the likelihood ration test. It often happens that creating a better fit for the fixed effects, by including their dependence on covariates reduces the need for random-effect terms.

Modeling the dependence of the four parameters according to the group

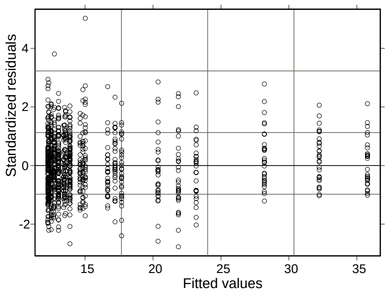


Figure 3: Plot of standardized residuals *versus* fitted values for model (4).

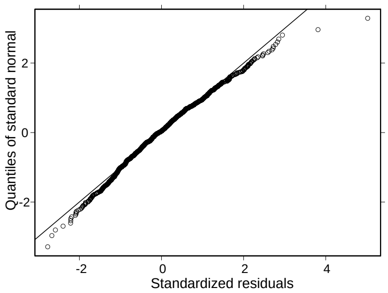


Figure 4: Normal plot of standardized residuals for model (4).

(model (5)), in order to quantify the relation between the average delay and the extension of green time, we get significant results. The significance of the group for the mixed effects is analysed by using ANOVA test (Table 5) and by comparing the values of AIC, BIC and *Log-Likelihood* criteria. Proceeding sequentially in the model building process by examining plots of the estimated random effects against the experimental factors, testing for the inclusion of covariates and for the elimination of random effects, we end up with a model, in which the only random effect is that for β_0. Analyzing the significance of the estimates of parameters and criteria of error analysis of the residuals, we found the best model, whose parameters are described in Table 6. The augmented predictions are plotted in Figure 5. The model provides a good representation of the average delays in the data.

Parameters	numDF	denDF	F-value	p-value
$\beta_0.(Intercept)$	1	978	205155.3	$< .0001$
$\beta_0.group$	2	978	5.0	0.0071
$\beta_1.(Intercept)$	1	978	34.7	$< .0001$
$\beta_1.group$	2	978	7.0	0.0010
$\beta_2.(Intercept)$	1	978	915.8	$< .0001$
$\beta_3.(Intercept)$	1	978	103.2	$< .0001$

Table 5: Results of application of the ANOVA test to model (5).

Parameters	**Value**	**Parameters**	**Value**
$\beta_0.ext3$	18.9075	$\beta_1.ext3$	-2.0118
$\beta_0.ext4$	18.0386	$\beta_1.ext4$	-1.8823
$\beta_0.ext5$	17.8637	$\beta_1.ext5$	-1.81704
β_2	0.1499766	β_3	0.0000284183

Table 6: Parameters estimates for model (5).

Although not suggested in the traffic analysis literature, the application of the theory of queues with server vacations was explored. The first dif-

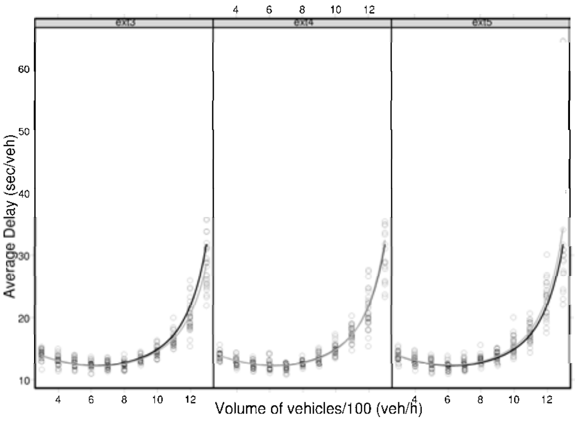


Figure 5: Forecasts for the population, forecasts inside the groups and observed delays *versus* number of vehicles per hour.

ficulty arises when parameters of the server vacations have to be given, as the traffic cycle is not fixed. Figure 2 illustrates the vehicles average delay (customers waiting time), when this theory is applied to an intersection regulated with prefixed time, as well as the results of *Webster*'s formula [7] for comparison. The curves are similar for any saturation degree but inferior values are obtained when *Webster*'s expression is applied. When comparing the results with those from the simulation, we find again that for a saturation degree larger than 0.7 both the expression developed by the theory of queues with server vacations and *Webster*'s formula underestimate the real waiting times. Since advantages in using theory of queues with server vacations for

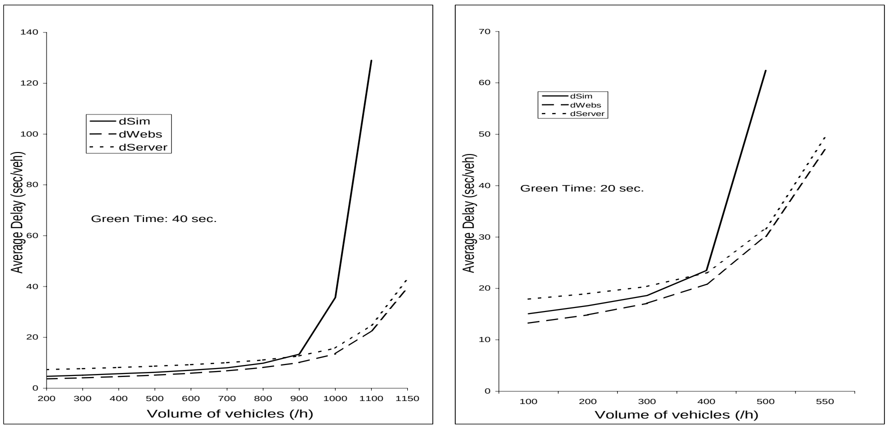


Figure 6: Average delays obtained by simulation (dSim) and by application of the theory of queues with (dServer) and without (dWebs) server vacations, with a cycle of 60 sec.

high traffic flows in the simplest case of intersections regulated with prefixed cycle are not found, the investigation of the generalization of this theory to include traffic signals with semi-actuated regulation is devoted to insuccess. Note that, in this type of regulation it is difficult to estimate the green time (idle time) as well as the red time (busy time).

Comparison with real world data is being made by using field measurements collected in an intersection located in metropolitan area of Porto (Portugal). Until now contradictions were not found between simulations and the real situation.

3 Conclusion

The average delay suffered by drivers in urban intersection can be described by a mixed effects linear model that results of the combination of a second degree polynomial with an exponential term. The increment of green in the secondary street has a significant effect in the response variable, both in less intense flow rates and when the flows approach the maximum capacity of the intersection leg. It is possible to optimize the regulation of this intersection by regulating the extension of green time as a function of the flow detected in the secondary street. The use of queues with server vacations's theory is not adequate to treat this problem.

References

[1] Davidian M., Giltinan D. M. (1995). *Nonlinear models for repeated measurement data.* Monographs on Statistics and Applied Probability **62**.

[2] Diggle P., Liang K.-Y., Zeger S. (1998). *Analysis of longitudinal data.* Oxford Statistical Science Series **13**.

[3] Laird N.M., Ware J.H. (1982). *Random-effects models for longitudinal data.* Biometrics **38**, 963 – 974.

[4] Lin F.-B. (1991). *Knowledge base on semi-actuated traffic-signal control.* Journal of Transportation Engineering **117** (4), 398 – 417.

[5] Pinheiro J.C., Bates D.M. (2000). *Mixed-effects models in S and S-Plus.* Statistics and Computing, Springer-Verlag.

[6] Rouphail N.M., Anwar M., Fambro D.B., Sloup P., Perez C.E. (1997). *Validation of generalized delay model for vehicle-actuated traffic signals.* Transportation Research Record **1572**, 105 – 111.

[7] Webster F. V. (1958). *Traffic signal settings.* Road Research Laboratory, Technical Paper **39**, HMSO, London.

Acknowledgement: Partially supported by PRODEP, CEC and CITTA

Address: L. Simoes, P.M. Oliveira, A. Pires da Costa, Department of Civil Engineering, Faculty of Engineering, University of Porto, R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

E-mail: lurdes.simoes@fe.up.pt

COMPSTAT'2004 Symposium

OPTIMUM ALLOCATION FOR BAYESIAN MULTIVARIATE STRATIFIED SAMPLING USING SEMIDEFINITE PROGRAMMING

Marcin Skibicki

Key words: Sample allocation, Bayes stratified sampling, semidefinite programming.

COMPSTAT 2004 section: Bayesian methods.

Abstract: The paper concerns problem of sample sizes optimization in Bayesian stratified sampling when there are many variables. Under assumption of exchangeability, minimization of the total survey cost subject to a fixed Bayes risk is considered. This problem, also for the other sampling designs appears in several papers, e.g. Khan [4], Mohd [5], Sekkappan [6]. Here the use of the semidefinite programming is proposed.

1 Bayesian stratified sampling

We assume a finite population U of N units partitioned into H strata $U_1, \dots, U_H$. Let N_h be the number of units in stratum U_h. There is a vector $Y = [Y_1, \dots, Y_r]$ of the variables of interest. Let Y_{ik} be the value of i-th variable for k-th unit of the population. Moreover let μ_{ih} and σ_{ih}^2 be the population mean and variance of variable Y_i in the h-th stratum, i.e.

$$\mu_{ih} = \frac{1}{N_h} \sum_{k \in U_h} Y_{ik} \quad \text{and} \quad \sigma_{ih}^2 = \frac{1}{N_h} \sum_{k \in U_h} (Y_{ik} - \mu_{ih})^2.$$

The overall population mean of the Y_i is

$$\mu_i = \sum_{h=1}^{H} \frac{N_h}{N} \mu_{ih}.$$

We suppose that a priori in each stratum U_h and for all i, the Y_{ik} are exchangable with means m_{ih}, variances v_{ih} and covariances c_{ih}. Exchangeability within h-th stratum means that all the $N_h!$ permutations of variables Y_{ik} (for $k \in U_h$) have the same joint probability. We also assume that Y_{ik} and Y_{ij} are uncorrelated if k and j are in different strata. We select a sample s of n population units and observe $Y_{ik} = y_{ik}$ for $k \in s$. The sample consists n_h units from h-th stratum. Let $\overline{Y}_{ih}$ be the mean of the i-th variable in the sample from h-th stratum.

Using above assumptions we can obtain the following results (see Ericson [2]):

$$E(\mu_{ih}) = E(\overline{Y}_{ih}) = m_{ih}, \quad E(\sigma_{ih}^2) = \frac{(N_h - 1)(v_{ih} - c_{ih})}{N_h} \tag{1}$$

$$V(\mu_{ih}) = \frac{v_{ih} + (N_h - 1)c_{ih}}{N_h}, \quad V(\overline{Y}_{ih}) = \frac{v_{ih} + (n_h - 1)c_{ih}}{n_h}. \tag{2}$$

The conditional means and variances of $\overline{Y}_{ih}$ are following:

$$E(\overline{Y}_{ih}|\mu_{ih}, \sigma^2_{ih}) = \mu_{ih}, \quad V(\overline{Y}_{ih}|\mu_{ih}, \sigma^2_{ih}) = \frac{(N_h - n_h)\sigma^2_{ih}}{(N_h - 1)n_h}.$$

The linear Bayes estimator of μ_i is given by

$$\widetilde{\mu}_i = \sum_{h=1}^{H} \frac{N_h \left[\overline{Y}_{ih} V(\mu_{ih}) + m_{ih} EV(\overline{Y}_{ih}|\mu_{ih}, \sigma^2_{ih})\right]}{N \left[V(\mu_{ih}) + EV(\overline{Y}_{ih}|\mu_{ih}, \sigma^2_{ih})\right]} \tag{3}$$

where $V(\mu_{ih})$ is given in (2) and

$$EV(\overline{Y}_{ih}|\mu_{ih}, \sigma^2_{ih}) = \frac{(N_h - n_h)E(\sigma^2_{ih})}{(N_h - 1)n_h} = \frac{(N_h - n_h)(v_{ih} - c_{ih})}{N_h n_h}.$$

Using results obtained by Ericson [3] we get

$$E_S V(\mu_i|s) \leqslant \sum_{h=1}^{H} \frac{N_h^2 \left[V(\mu_{ih})\, EV(\overline{Y}_{ih}|\mu_{ih}, \sigma^2_{ih})\right]}{N^2 \left[V(\mu_{ih}) + EV(\overline{Y}_{ih}|\mu_{ih}, \sigma^2_{ih})\right]} \tag{4}$$

Let $\mathbf{n} = [n_1, \ldots, n_H]$ be the vector of sample sizes. Suppose that the total cost of the survey is of the form

$$f_c(\mathbf{n}) = \sum_{h=1}^{H} k_h n_h$$

where k_h is a fixed cost of surveying one unit from h-th stratum.

Let us consider problem of optimum sample allocation formulated as the determination of the $\mathbf{n}$ that minimize the total cost $f_c(\mathbf{n})$ subject to a fixed Bayes risk $E_S V(\mu_i|s)$ for all $i = 1, \ldots, r$. This problem can be written

$$\begin{cases} f_c(\mathbf{n}) = \max \\ \varphi_i(\mathbf{n}) \leqslant e_i\,, & r = 1, \ldots, r \\ 1 \leqslant n_h \leqslant N_h\,, & h = 1, \ldots, H \end{cases} \tag{5}$$

where $\varphi_i(\mathbf{n})$ is the right-hand side of the inequality (4) and e_i are fixed admissible risk values. Substituting from (1) and (2) we obtain

$$\varphi_i(\mathbf{n}) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{(N_h - n_h)(v_{ih} - c_{ih})(v_{ih} + (N_h - 1)c_{ih})}{n_h c_{ih} + v_{ih} - c_{ih}} = \sum_{h=1}^{H} \frac{b_{ih}}{n_h + d_{ih}} - a_i$$

where

$$d_{ih} = \frac{v_{ih} - c_{ih}}{c_{ih}}, \quad b_{ih} = \frac{(N_h - d_{ih})d_{ih}\,[v_{ih} + (N_h - 1)c_{ih}]}{N^2}$$

$$a_i = \sum_{h=1}^{H} \frac{d_{ih}\,[v_{ih} + (N_h - 1)c_{ih}]}{N^2}$$

2 Semidefinite program

A semidefinite program is a convex optimization problem of minimizing a linear function of a variable $x \in \mathbb{R}^H$ subject to a matrix inequality (see Vandenberghe, Boyd [7]):

$$\begin{cases} c^T\mathbf{x} = \min \\ F(\mathbf{x}) \geqslant 0\,, \end{cases} \tag{6}$$

where

$$F(\mathbf{x}) = F_0 + \sum_{h=1}^{H} x_h F_h\,.$$

The vector $c \in \mathbb{R}^H$ and symmetric matrices $F_0, \ldots, F_H \in \mathbb{R}^{p\times p}$ are given. The matrix inequality $F(\mathbf{x}) \geqslant 0$ means that $F(\mathbf{x})$ is positive semidefinite, i.e. $z^T F(\mathbf{x})z \geqslant 0$ for all $z \in \mathbb{R}^p$.

Semidefite programming problems arise in a number of applications, also in the survey sampling. Most of sample allocation problems in survey sampling can be cast as semidefinite programs. Recently developed interior-point methods for semidefinite programming (see Alizadeh [1]) are very efficient even for large-scale problems.

Let us formulate the problem (5) as the semidefinite program (6). The constraints from (5) must be expressed as a linear matrix inequality in the variable $\mathbf{n}$. This matrix will be block-diagonal of the form

$$A(\mathbf{n}) = \begin{bmatrix} A_1(\mathbf{n}) & 0 \\ 0 & A_2(\mathbf{n}) \end{bmatrix}$$

where $A_1(\mathbf{n}) \geqslant 0$ represents inequalities $\varphi_i(\mathbf{n}) \leqslant e_i$ and $A_2(\mathbf{n}) \geqslant 0$ represents inequalities $1 \leqslant n_h \leqslant N_h$. The $A_2(\mathbf{n})$ is diagonal of size $2H \times 2H$ given by

$$A_2(\mathbf{n}) = \operatorname{diag}(n_1 - 1, \ldots, n_H - 1, N_1 - n_1, \ldots, N_H - n_H).$$

Matrix $A_1(\mathbf{n})$ can be written as block-diagonal

$$A_1(\mathbf{n}) = \operatorname{diag}(A_{11}(\mathbf{n}), \ldots, A_{1r}(\mathbf{n}))$$

where $A_{1i}(\mathbf{n}) \geqslant 0$ is equivalent to the i-th constraint $\varphi_i(\mathbf{n}) \leqslant e_i$. This matrix will be of the form

$$A_{1i}(\mathbf{n}) = \begin{bmatrix} e_i + a_i & 1 & 1 & \cdots & 1 \\ 1 & \frac{n_1+d_{i1}}{b_{i1}} & 0 & \cdots & 0 \\ 1 & 0 & \frac{n_2+d_{i2}}{b_{i2}} & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 1 & 0 & \cdots & 0 & \frac{n_H+d_{iH}}{b_{iH}} \end{bmatrix}.$$

For $H = 1$ one can check easily that $A_{1i}(\mathbf{n}) \geqslant 0$ is equivalent to inequalities $e_i + a_i \geqslant 0$ and $e_i \geqslant \frac{b_{i1}}{n_{i1}+d_{i1}}$. Restrictions e_i should be fixed so first of the inequalities is fulfilled.

For number $H > 1$ strata the $\varphi_i(\mathbf{n}) \leqslant e_i$ can be written as

$$a_i + e_i - \sum_{h=1}^{H} \frac{b_{ih}}{n_h + d_{ih}} \geqslant 0$$

or

$$(a_i + e_i) \prod_{h=1}^{H} \frac{n_h + d_{ih}}{b_{ih}} - \sum_{h=1}^{H} \prod_{\substack{j=1 \\ j \neq h}} \frac{n_j + d_{ij}}{b_{ij}} \geqslant 0.$$

The left-hand side of the above inequality can be expressed as

$$\frac{n_H + d_{iH}}{b_{iH}} \left[(a_i + e_i) \prod_{h=1}^{H-1} \frac{n_h + d_{ih}}{b_{ih}} - \sum_{h=1}^{H-1} \prod_{\substack{j=1 \\ j \neq h}} \frac{n_j + d_{ij}}{b_{ij}} \right] - \prod_{h=1}^{H-1} \frac{n_h + d_{ih}}{b_{ih}} =$$

$$= \frac{n_H + d_{iH}}{b_{iH}} \det A_{1i}^{(H-1)}(\mathbf{n}) + (-1)^{H+1} \det B_i(\mathbf{n}) = \det A_{1i}(\mathbf{n}).$$

The last equality results from evaluation of $\det A_{1i}(\mathbf{n})$ by the last row. If the $A_{1i}(\mathbf{n}) \geqslant 0$ has to be equivalent to the $\varphi_i(\mathbf{n}) \leqslant e_i$ then $a_i + e_i - \sum_{h=1}^{p} \frac{b_{ih}}{n_h+d_{ih}} \geqslant 0$ for all $p < H$ have to be fulfilled. Since $a_i + e_i \geqslant 0$, this holds if $\frac{b_{ih}}{n_h+d_{ih}} \geqslant 0$ for all $h = 1, \ldots, H$. However in most applications we have $c_{ih} > 0$.

Finally, the problem (5) can be cast as semidefinite program

$$\begin{cases} f_c(\mathbf{n}) = \max \\ A(\mathbf{n}) \geqslant 0 \end{cases} \tag{7}$$

where $A(\mathbf{n})$ is a block-diagonal matrix of size $p \times p$ and $p = r(H+1) + 2H$.

Concluding, in order to solve the problem (5) we can compute matrix $A(\mathbf{n})$ and use an interior-point algorithm for semidefinite programs. Several software packages including these algorithms exist, e.g. CSDP (primal-dual path following algorithm), SDPA (primal-dual path following algorithm), SP (primal-dual potential reduction algorithm). The interior-point methods are competitive with other (nonlinear programming) methods for problem (5) and even faster for large-scale problems ($r > 100$ and $H > 1000$). With properties of the interior-point methods results that the effort requaried to solve the problem (5) to a given accuracy grows with a polynomial of the number of strata H and the number of variables r. The global optimum of $\mathbf{n}$ is usually obtained in about 5-30 iterations of the interior-point algorithm.

References

[1] Alizadeh F. (1995). *Interior-point methods in semidefinite programming with applications to combinatorial optimization.* SIAM J. on Optimization **5** (1), 13 – 51.

[2] Ericson W. (1969). *Subjective Bayesian models in sampling finite populations.* J. Roy. Statist. Soc. B **31**, 195 – 233.

[3] Ericson W. (1970). *On the posterior mean and variance of a population mean.* J. Amer. Statist. Assoc. **65**, 649 – 652.

[4] Khan M. Z. (1976). *Optimum allocation in Bayesian stratified two phase sampling when there are m attributes.* Metrika **23**, 211 – 220.

[5] Mohd Z. K. (1976). *Optimum allocation in Bayes stratified two phase samples.* J. Indian Statist. Assoc. **14**, 65 – 74.

[6] Sekkappan M. R. (1981). *Subjective Bayes multivariate stratified sampling for finite populations.* Metrika **28**, 123 – 132.

[7] Vandenberghe L., Boyd S. (1996). *Semidefinite programming.* SIAM Review **38**, 49 – 95.

Address: M. Skibicki, Akademia Ekonomiczna, ul. 1 Maja 50,
PL-40-287 Katowice, Poland

E-mail: skibi@ae.katowice.pl

MULTIVARIATE BILINEAR GARCH MODELS

Giuseppe Storti

Key words: Multivariate bilinear GARCH models, volatility, EM algorithm.

COMPSTAT 2004 section: Time series analysis.

Abstract: The class of Multivariate Bilinear GARCH (MBL-GARCH) models is proposed. The MBL-GARCH model allows to account for asymmetric effects in the conditional variances of the marginal univariate processes as well as for time varying conditional covariances. These are modeled as linear functions of interactions between past volatilities and returns. In this way it is possible to introduce asymmetric components not only in the conditional variances but also in the conditional covariances. MBL-GARCH models are based on a parsimomious parameterization. For a k-dimensional process, the total number of parameters is $n_p = (5k^2 + 3k)/2$. Under the assumption of conditional multivariate normality, the unknown model parameters can be estimated by Maximum Likelihood. To this purpose, an EM type algorithm for the maximizazion of the likelihood function is derived. An important feature of the proposed algorithm is that it returns estimates which naturally satisfy the constraints required for the positive semi-definiteness of the estimated conditional covariance matrix. The results of an application to real stock market data are presented.

1 Introduction

Both empirical and theoretical evidence suggest the presence of comovements in the volatilities associated to different assets and markets. This consideration has motivated the great attention which in the last years has been dedicated to the analysis of multivariate GARCH specifications allowing to model the time varying conditional variances as well as the conditional covariances between different assets. Namely, for a k-dimensional multivariate time series of returns y_t, a set of recursions is needed for the $(k\times k)$conditional variance matrix:

$$H_t = var(y_t|I^{t-1})$$

where $I^{t-1} = \{y_0, y_1, y_2, \dots, y_{t-1}\}$. Several important issues must be faced when choosing an adequate parameterization for the time varying conditional variance matrix H_t. First, the need for parsimonious model specifications must be considered. In order to restrict the number of potential parameters, adequate constraints on the dynamics of H_t must be imposed. Obviously, the adequacy of a specific set of constraints depends on the nature of the problem to be analyzed. Second, the chosen parameterization must be able to guarantee the positive semi-definiteness of the conditional covariance matrix. To this purpose, further constraints on the admissible parameter space could be necessary. In general, at the model selection stage, it is then important to determine if the required constraints impose important untested characteristics on the conditional variance dynamics which are not supported by empirical or theoretical evidence.

In this paper a new class of multivariate conditionally heteroskedastic models is presented. The proposed class of models allows to account for the presence of asymmetric effects in the conditional variance as well as in the conditional covariance dynamics. Namely, the conditional covariances are modelled as linear functions of cross-interactions between past returns and volatilities yielding time varying conditional correlations between different assets. A relevant property of the model is that the conditional covariance matrix H_t is positive semi-definite by construction. Furthermore, under the assumption of conditional multivariate normality, the likelihood function can be maximized with respect to the unknown parameters by means of an EM-type algorithm. By its own definition, this directly returns estimates that naturally satisfy the requirements for the positive semi-definiteness of H_t, with no need for arbitrary parameter constraints. The proposed model is called a Multivariate BiLinear GARCH model (MBL-GARCH) since it can be considered as a generalization of the univariate BL-GARCH model discussed in Storti and Vitale [5], [6].

The paper is structured as follows. Section 2 introduces the class of MBL-GARCH models while an EM algorithm for obtaining maximum likelihood estimates of the model parameters is illustrated in Section 3. Section 4 presents the results of an application to real stock market data while Section 5 concludes.

2 The class of MBL-GARCH models

Let $y_t = [y_{1t} \dots y_{kt}]'$ be a $k \times 1$ vector stochastic process such that $(y_t | I^{t-1}) \sim (0, H_t)$, where the $(i, j)th$ element of H_t denotes the conditional covariance between y_{it} and y_{jt}. The multivariate BL-GARCH of order (1,1) is given by the linear state space model:

$$\begin{aligned} y_t &= C_t x_t && (1) \\ x_t &= u_t && (2) \end{aligned}$$

with u_t $(2k \times 1)$$\sim$ $MVN(0, V)$, independent of I^{t-1}, and

$$C_t = [I_k \quad B_t] \qquad (3)$$

where I_k is an identity matrix of order k and B_t is a $(k \times 2k)$ matrix such that

$$B_t = \begin{bmatrix} S_1 & 0_{1,2} & \dots & 0_{1,2} & 0_{1,2} \\ 0_{1,2} & S_2 & \dots & 0_{1,2} & 0_{1,2} \\ \vdots & \dots & \dots & \dots & \vdots \\ 0_{1,2} & \dots & \dots & S_{k-1} & 0_{1,2} \\ 0_{1,2} & \dots & \ddots & 0_{1,2} & S_k \end{bmatrix} \tag{4}$$

and $0_{1,2}$ is a (1×2) vector of zeros and

$$S_i = [y_{i,t-1} \qquad \sqrt{h_{ii}(t-1)}] \tag{5}$$

If the random coefficients vector u_t is partitioned as:

$$u_t = \{u'_{1,t} \quad u'_{2,t}\}' \tag{6}$$

with $u_{1,t}$ being of dimension $(k \times 1)$, the covariance matrix V can then be rewritten as:

$$\begin{Bmatrix} R & 0_{k,2k} \\ 0_{2k,k} & Q \end{Bmatrix}$$

where $R = var(u_{1t})$ is a $(k \times k)$ covariance matrix. The conditional covariance matrix of y_t is hence defined as:

$$H_t = C_t V C'_t = B_t Q B'_t + R \tag{7}$$

For a k-dimensional process, the total number of parameters is equal to $n_p = (5k^2 + 3k)/2$. Application of this formula gives n_p=13,27,46 for k=2,3,4, respectively, leading to a parsimonious parameterization of the the volatility dynamics. For a bidimensional process (k=2) we have:

$$H_t \quad = \quad \begin{Bmatrix} h_{11,t} & h_{12,t} \\ h_{21,t} & h_{22,t} \end{Bmatrix} \tag{8}$$

and

$$C_t = \begin{bmatrix} 1 & 0 & y_{1t} & \sqrt{h_{11}(t)} & 0 & 0 \\ 0 & 1 & 0 & 0 & y_{2t} & \sqrt{h_{22}(t)} \end{bmatrix} \tag{9}$$

Letting $y_{1t} = [y_{1,t} y_{2,t}]'$, we can then derive the following recursions for the elements of H_t:

$$h_{11}(t) = R_{11} + Q_{11} y_{1,t-1}^2 + Q_{22} h_{11}^2(t-1) + 2Q_{12} y_{1,t-1} \sqrt{h_{11}(t-1)} \tag{10}$$

$$h_{22}(t) = R_{22} + Q_{33} y_{1,t-1}^2 + Q_{44} h_{11}^2(t-1) + 2Q_{43} y_{2,t-1} \sqrt{h_{22}(t-1)} \tag{11}$$

$$\begin{aligned} h_{12}(t) &= R_{21} + Q_{13} y_{1,t-1} y_{2,t-1} + Q_{23} y_{2,t-1} \sqrt{h_{11}(t-1)} + \\ &+ Q_{14} y_{1,t-1} \sqrt{h_{22}(t-1)} + Q_{24} \sqrt{h_{11}(t-1)} \sqrt{h_{22}(t-1)} \end{aligned} \quad (12)$$

The conditional variances $h_{11}(t)$ and $h_{22}(t)$ follow two univariate BL-GARCH models of order (1,1) while the conditional covariance $h_{12}(t)$ is time varying and expressed as a linear combination of cross-terms given by the interactions between 1) past returns 2) past volatilities 3) past returns and volatilities of the two series. By considering interactions between past returns and volatilities in the recursion for $h_{12}(t)$, it is also possible to account for the presence of asymmetric effects in the conditional covariance dynamics.

Finally, the time varying conditional correlation coefficient can be computed as:

$$\rho_{12}(t) = \frac{h_{12}(t)}{\sqrt{h_{11}(t) h_{22}(t)}}.$$

3 Maximum likelihood estimation

The unknown parameters in the multivariate BL-GARCH model described in the previous section are concentrated in the matrix V. Under the assumption of conditional normality, $y_t \sim MVN(0, H_t)$, the log-likelihood function of the observed data can be expressed in the prediction error decomposition form as:

$$\log L(y; \theta) = -\frac{1}{2} \sum_{t=1}^{T} \log |H_t| - \frac{1}{2} \sum_{t=1}^{T} y_t^{'} H_t^{-1} y_t \quad (13)$$

with $\theta = [\text{vech}(R)' \text{vech}(Q)']$ where $vech(.)$ denotes the operator that stacks the lower triangular portion of a $(N \times N)$ matrix as a $N(N+1)/2 \times 1$ vector.

Numerical maximization of the above log-likelihood function can be performed by means of a multivariate generalization of the algorithm proposed by Storti and Vitale [6] for univariate BL-GARCH models. The algorithm allows to obtain estimates of the model parameters which, by construction, ensure the positive semi-definiteness of the estimated conditional variance.

The MBL-GARCH model is defined as a linear state space model with observation equation given by (1) and transition equation given by (2). Hence, as shown by Shumway and Stoffer [4], the EM algorithm [1] can be used to maximize the log-likelihood in (13). Since the observation equation of the model does not include an error component, the log-likelihood of the observed plus the unobserved data $\{y_1, \ldots, y_T, x_1, \ldots, x_T\}$ in equation (3) of Shumway and Stoffer [4] reduces to the log-likelihood of the unobserved data:

$$\log L(x; \theta) = -\frac{T}{2} \log |V| - \frac{1}{2} \sum_{t=1}^{T} x_t^{'} V^{-1} x_t \quad (14)$$

At the Expectation step, the Kalman filter can be used to calculate

$$E(\log L(x;\theta)|I^T) = -\frac{T}{2}\log|V| - \frac{1}{2}\mathrm{tr}(V^{-1}\Gamma) \tag{15}$$

where $\Gamma = \sum_{t=1}^{T}[P_t^T + x_t^T(x_t^T)']$ with $P_t^T = var(x_t|I^T)$ and $x_t^T = E(x_t|I^T)$. Because of the lack of serial dependence in the state vector series, it can be easily shown that $x_t^T = x_t^t$ and $P_t^T = P_t^t$. This means that equation (15) can be evaluated by means of the ordinary Kalman filter [3] with no need to apply the more demanding Fixed Interval Smoothing algorithm.

At the Maximization step, the expected log-likelihood can be then iteratively maximized with respect to θ. The algorithm needs to be initialized by means of an adequately chosen starting point $V^{(0)}$. Moving from the initial guess, the expected log-likelihood is then evaluated and analytically maximized by taking the first derivative of (15) with respect to θ. The procedure is repeated until convergence. The estimate of V at the $(i+1)$-th iteration is:

$$\hat{V}^{(i+1)} = T^{-1}\Gamma^{(i)}$$

where $\Gamma^{(i)}$ is the estimate of Γ based on $\hat{V}^{(i)}$. The algorithm, as shown by Wu et al. [7], allows to incorporate linear constraints on the elements of V and also allows to easily deal with situations in which some of the elements of V are constrained to be equal to zero. In this case, at each iteration, an estimate of θ can be obtained updating only the non-zero elements of V. In general, it can be shown that elementwise updating can be applied whenever V is diagonal or block-diagonal, as it is the case for multivariate BL-GARCH models.

The EM algorithm does not allow to directly calculate the standard errors associated to the estimated parameters. However, their asymptotic value can be approximated by evaluating the observed Information Matrix $\tilde{I}(\theta)$ at the maximum likelihood estimate i.e. for $\theta = \hat{\theta}$. The ij-th element of the Information Matrix can be shown to be given by:

$$\tilde{I}_{ij}(\theta) = \frac{1}{2}\sum_{t=1}^{T}\left\{\mathrm{tr}\left[H_t^{-1}\frac{\partial H_t}{\partial\theta_i}H_t^{-1}\frac{\partial H_t}{\partial\theta_j}\right]\right\} \tag{16}$$

The derivatives in (16) can be recursively calculated as [2]:

$$\frac{\partial H_t}{\partial\theta_i} = \frac{\partial C_t}{\partial\theta_i}VC_t^{'} + C_t\frac{\partial V}{\partial\theta_i}C_t^{'} + C_tV\frac{\partial C_t'}{\partial\theta_i}$$

4 An application to stock market data

In this section a bivariate BL-GARCH model of order(1,1) is applied to three different bivariate time series of stock market returns. The first dataset is given by the daily returns on the Dow Jones Industrials and NASDAQ Composite (DJ-NQ) stock market indexes observed from January 4th 1995

	DJ-NQ		CAC-DAX		MIB-DAX	
	coeff.	z-stat.	coeff.	z-stat.	coeff.	z-stat.
R_{11}	3.764$^{(*)}$	4.941	2.862$^{(*)}$	3.929	3.560$^{(*)}$	4.155
R_{21}	1.896$^{(*)}$	2.799	1.138$^{(*)}$	1.386	0.993$^{(*)}$	0.895
R_{22}	3.460$^{(*)}$	7.051	3.427$^{(*)}$	5.434	3.878$^{(*)}$	4.937
Q_{11}	0.098	4.213	0.070	3.843	0.124	5.039
$Q_{12}^{(A)}$	-0.068	-3.824	-0.066	-4.459	-0.031	-2.020
Q_{13}	0.033	1.490	0.051	2.926	0.037	1.720
$Q_{14}^{(B)}$	-0.053	-2.188	-0.050	-2.272	-0.073	-2.895
Q_{22}	0.559	6.785	0.763	13.811	0.711	12.534
$Q_{23}^{(B)}$	-0.072	-2.921	-0.085	-3.639	-0.047	-1.860
Q_{24}	0.506	10.607	0.610	11.206	0.529	8.251
Q_{33}	0.158	6.401	0.098	4.970	0.106	4.598
$Q_{34}^{(A)}$	-0.105	-6.991	-0.102	-6.143	-0.083	-4.848
Q_{44}	0.755	23.423	0.727	16.010	0.696	12.457

Table 1: Maximum Likelihood Estimates of BL-GARCH model parameters and associated z-statistics. Key to table: $^{(*)}$ the value has been multiplied by 10^5 $^{(A)}$ parameters controlling asymmetry in the conditional variance $^{(B)}$ parameters controlling asymmetry in the conditional covariance.

to February 9th 2001, for a total of 1542 observations. Returns have been calculated as the logarithmic first difference of the price series. The second dataset considered is a bivariate time series of returns on the CAC40 and DAX30 stock market indexes observed from December 1st 1995 to February 5th 2001, for a total of 1352 observations. Finally, the last dataset is given by the daily returns on the MIB30 and DAX30 over the same period. The maximum likelihood estimates returned by the EM algorithm described in section 3 and the associated z-statistics have been reported in Table 1. In all cases, the results suggest the presence of asymmetric effects in the conditional variance (Q_{12}, Q_{34}) as well as in the conditional covariances (Q_{23}, Q_{14}). For both the conditional variance and the conditional covariance the estimated coefficients are negative and significantly different from zero. This implies that if, for example, negative returns are observed at time $(t-1)$ for both series, i.e. $y_{1,t-1} < 0$ and $y_{2,t-1} < 0$, a positive quantity will be added to $h_{11}(t)$, $h_{22}(t)$ and $h_{12}(t)$ while, for $y_{1,t-1} > 0$ and $y_{2,t-1} > 0$, the same quantity will be subtracted. The time paths of the conditional covariances and correlations for the three datasets have been reported in figure 1.

5 Concluding remarks

A new class of multivariate conditional heteroskedastic time series models has been presented. The proposed model is based on a parsimonious param-

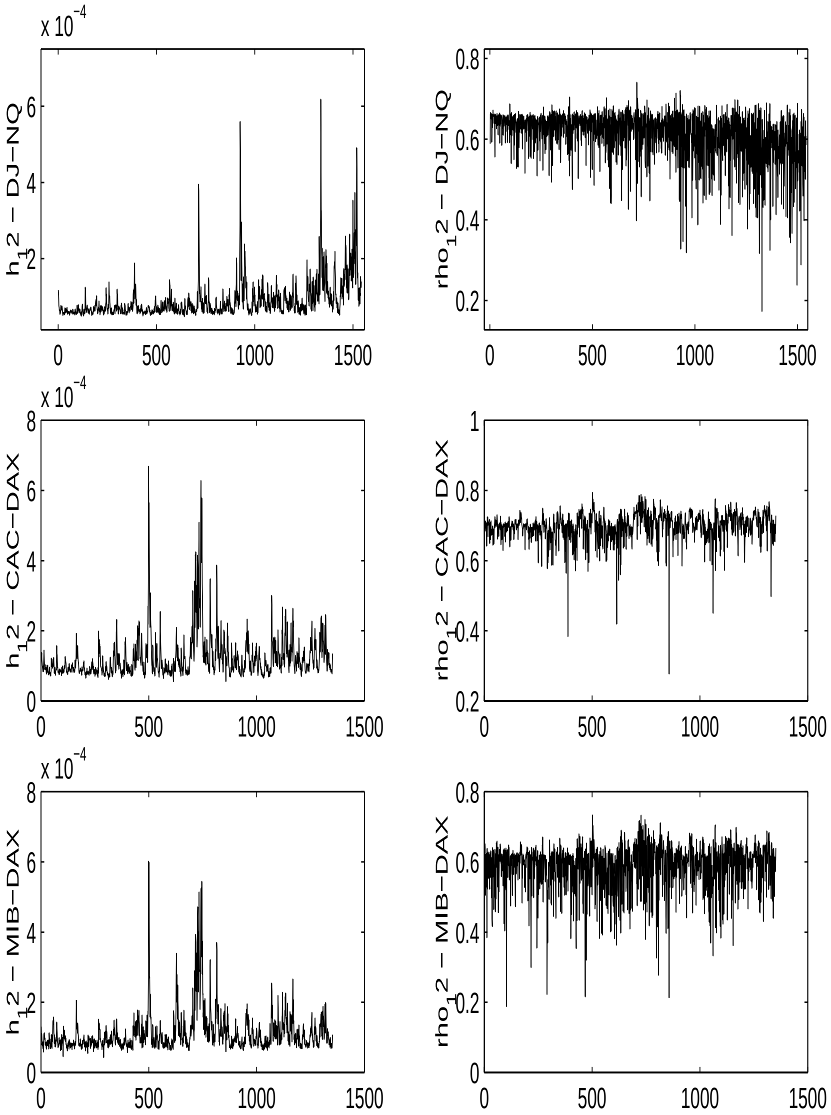


Figure 1: From top to bottom, conditional covariances (left) and correlations (right) for the datasets: 1) DJ-NQ 2) CAC-DAX 3) MIB-DAX.

eterization of the conditional covariance matrix. Nevertheless, it still allows to account for asymmetry in both the conditional variances and covariances and for time varying correlations. However, it has to be recognized that the MBL-GARCH specification does not allow to test for causality in variance. This is due to the fact that, for the sake of parsimony, the conditional variance processes have been modelled as univariate BL-GARCH models of order (1,1). In order to overcome this limitation a straightforward extension of the MBL-GARCH model could be considered but this would lead to a substantial increase in the number of parameters to be estimated. Finally, an attractive feature of the model is of course given by the availability of likelihood inference based on the EM algorithm.

Projects for future research include a formal investigation of the statistical properties of the proposed model extending to a multivariate settings the results already obtained for univariate BL-GARCH models (see [5], [6].

References

[1] Dempster, A.P., Laird, N.M., Rubin, D.B.(1977). *Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion).* Journal of the Royal Statistical Society (Series B) **39**, 1 – 38.

[2] Harvey, A. C. (1989). *Forecasting, structural time series, models and the Kalman filter.* Cambridge University Press.

[3] Kalman, R. E. (1960). *A new approach to linear filtering and prediction problems.* Journal of Basic Engineering (ASME) **82D**, 35 – 45.

[4] Shumway, R.H, Stoffer, D.S. (1982). *An approach to time series smoothing and forecasting using the EM algorithm.* Journal of Times Series Analysis **3**, 253 – 264.

[5] Storti G., Vitale, C. (2003a). *BL-GARCH models and asymmetries in volatility*, Statistical Methods and Applications (Journal of the Italian Statistical Society) **12**, 19 – 40.

[6] Storti G., Vitale, C. (2003b). *Likelihood inference in BL-GARCH models.* Computational Statistics **18**, 387 – 400.

[7] Wu, L.S., Pai, J.S., Hosking, J. R. M. (1996). *An algorithm for estimating parameters of state-space models.* Statistics and Probability Letters **28**, 99 – 106.

Address: G. Storti, Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Ponte Don Melillo, 84084, Fisciano (SA), Italy.

E-mail: storti@unisa.it

COMPSTAT'2004 Symposium

INFLUENCE ANALYSIS IN COX PROPORTIONAL HAZARDS MODELS WITH EMPHASIS ON MULTIPLE-CASE DIAGNOSTICS

Jimin Sung and Yutaka Tanaka

Key words: Cox regression, local influence, influential subsets.

COMPSTAT 2004 section: Robustness.

Abstract: In the present paper we propose a method of multiple-case diagnostics in Cox regression models with censored observations based on the above general procedure. A numerical example is analyzed to show the performance of the proposed method.

1 Introduction

Methods of influence analysis have been proposed in Cox proportional hazards and related models by Cain and Lange [2], Stoker and Crowley [8], Wei and Korosok [12] among others. The former two derived influence functions for regression coefficients in the proportional hazards and relative risk models, and proposed methods of single-case diagnostics, while the last one proposed a method of multiple-case diagnostics based on pairwise deletion and pairwise differentiation. For general statistical modeling Tanaka and his coworkers (see, e.g., Tanaka [9], Tanaka and Zhang [10] proposed a general procedure of influence analysis including multiple-case as well as single-case diagnostics. Their method is to make use of the additivity property of the influence measured with influence functions, try to reduce the dimension by applying PCA with metric $\mathbf{V}^{-1}$ to the influence functions, where $\mathbf{V}$ indicates an asymptotic covariance matrix of the estimated parameters, and search for individuals which are located far from and on similar directions from the origin in the space of principal component scores for detecting influential subsets of observations. It is known that their method of multiple-case diagnostics is closely related to Cook's [3] local influence.

2 Cox proportional hazards models

To analyze survival data Cox [4] proposed a regression model which is called proportional hazards model. It is described as below. Let $(t_i, \delta_i, \mathbf{Z}_i)$ be an observation vector of individual i for $i = 1, \cdots, n$, where t_i indicates the death or censored time, δ_i the dummy variable to denote death $(\delta_i = 1)$ or censored $(\delta_i = 0)$, and $\mathbf{Z}_i = [Z_{i1}, \ldots, Z_{ip}]^T$ the covariate values of individual i. Based on the assumption of proportional hazards the hazard function is expressed by

$$h(t) = h_0(t)\exp(\beta^T\mathbf{Z}),$$

where $h_0(t)$ indicates the so-called baseline hazard and it is assumed that the baseline hazard function is the same for all individuals in the study. It is called proportional hazards because, if we look at two individuals with covariate values $\mathbf{Z}$ and $\mathbf{Z}^*$, the ratio of their hazards rates is constant without depending on time t.

3 Parameter estimation

A method of maximum partial likelihood is proposed by Cox [4], where the partial likelihood is formed by multiplying conditional probabilities P (individual i dies at t_i | one death at t_i), which do not contain the baseline hazard function. When there are ties between death times, they are incorporated into partial likelihood using Breslow's [1] method. It is known that Breslow's method provides good approximation and this likelihood is implemented in most statistical packages [6]. To develop influence analysis we introduce case-weight w_i for individual i, $i = 1, \cdots, n$, that is, weight vector $\mathbf{w}$ is defined as $\mathbf{w} = \mathbf{w_o} = (1, 1, \cdots, 1)^T$ for the unperturbed condition and $\mathbf{w} = (w_1, w_2, \cdots, w_n)^T$ for the perturbed condition. Then the partial log-likelihood is expressed as

$$\ell(\beta) = \sum_{i=1}^{D}\sum_{k=1}^{p}\beta_k \sum_{l\in\Delta_i} w_l Z_{lk} - \sum_{i=1}^{D}\sum_{l\in\Delta_i} w_l \log[\sum_{j\in R_i} w_j \exp(\sum_{k=1}^{p}\beta_k Z_{jk})] \qquad (1)$$

where $t_{(i)}$'s ($t_{(1)} < t_{(2)} < \cdots < t_{(D)}$) indicate the distinct death times, d_i the number of individuals who died at time $t_{(i)}$, Δ_i the subset of individuals who died at $t_{(i)}$, and R_i the risk set at $t_{(i)}$. The regression coefficients can be obtained by solving the likelihood equation

$$U_h(\beta) = \partial\ell(\beta)/\partial\beta_h = 0, \quad h = 1, 2, \cdots, p,$$

using an appropriate iterative procedure such as the Newton-Raphson's method, and the precision of the estimates can be evaluated with the observed information matrix $\mathbf{I}(\hat{\beta})$ defined as

$$\mathbf{I}(\hat{\beta}) = [I_{gh}(\hat{\beta})]_{p\times p} \quad , \quad I_{gh}(\beta) = \partial^2\ell(\beta)/\partial\beta_g\partial\beta_h, \quad g, h = 1, 2, \cdots, p.$$

4 Influence analysis

As described in Section 1 Tanaka and his coworkers (see, e.g., Tanaka [9], Tanaka and Zhang [11] proposed a general procedure of influence analysis which deals with multiple-case diagnostics as well as single-case diagnostics. In the present paper, we propose a method of influence analysis in Cox regression models with censored observations based on their general procedure, which utilizes the additive property of the influence measured with influence functions. The general procedure is described as follows.

(1). Compute the influence function vectors $\partial\hat{\beta}/\partial w_i$, for $i = 1, 2, \cdots, n$.

(2). (Single-case diagnostics) Summarize the influence function vectors into scalar influence measures, from various aspects such as the influence on the estimate, on its precision and on the goodness-of-fit. Find individuals with large values of the measures.

(3). (Multiple-case diagnostics) Search for subsets of individuals whose members are individually relatively influential and have similar influence patterns by using PCA with metric V^{-1} (or V^{-}), where V indicates an estimated asymptotic covariance matrix of $\hat{\beta}$.

(4). Confirm the influence of single or multiple individuals by reanalyzing the sample without a subset of specified individuals.

4.1 Influence function

The influence of each individual on the estimate $\hat{\beta}$ can be evaluated with the partial differential coefficient of $\hat{\beta}$ with respect to w_j, i.e., $\partial\hat{\beta}/\partial w_j, j = 1, \cdots, n$, and it provides an approximation to $\hat{\beta} - \hat{\beta}_{(j)}$, where the partial differential coefficients are computed at w_0, and $\beta_{(j)}$ indicates the estimate for β based on the sample without individual j. We shall call this partial differential coefficient the empirical influence function, EIF or simply influence function. Application of the differentiation of implicit function yields

$$\frac{\partial\hat{\beta}}{\partial w_j} = \Big[- \frac{\partial \mathbf{U}}{\partial\hat{\beta}}\Big]^{-1}_{w_0} \frac{\partial \mathbf{U}}{\partial w_j},$$

where $\mathbf{U}$ is the so-called score vector and the term in brackets is the observed information matrix. As shown in Cain and Lange [2] the differential coefficient $\partial\mathbf{U}/\partial w_j$ is given by

$$\begin{aligned} \frac{\partial U_h}{\partial w_j}\Big|_{w_0} &= \delta_j \Big[Z_{jh} - \sum_{i=1}^{D} \hat{E}(Z_h|R_i) \Big] \\ &- \sum_{\{i|j\in R_i\}} d_i \frac{\exp(\sum_{m=1}^p \hat{\beta}_m Z_{jm})}{\sum_{k\in R_i} \exp(\sum_{m=1}^p \hat{\beta}_m Z_{km})} [Z_{jh} - \hat{E}(Z_h|R_i)], \end{aligned}$$

where

$$\hat{E}(Z_h|R_i) = \frac{\sum_{k\in R_i} Z_{kh} \exp(\sum_{m=1}^p \hat{\beta}_m Z_{km})}{\sum_{k\in R_i} \exp(\sum_{m=1}^p \hat{\beta}_m Z_{km})}.$$

The equation of $\partial\mathbf{U}/\partial w_j$ shows that the change in the score vector $\mathbf{U}$ due to changes in w_j consists of the sum of two components. The first component, which is called the partial residual of Schoenfeld [7], is included only if individual j died and is the difference between the covariates for case j and the weighted average of covariates for all individuals in the risk set R_i. The second component measures the combined effect that changes in w_j have upon all the risk sets that include individual j.

4.2 Single-case diagnostics

For single-case diagnostics, we compute Cook's D for each individual to study the influence on the regression coefficients. We can regard the individuals with large values of D as individually influential observations. The Cook's D is defined by

$$D_i = (\partial\hat{\beta}/\partial w_i)^T \mathbf{V}^{-1} (\partial\hat{\beta}/\partial w_i),$$

where $\mathbf{V}$, an estimate for the asymptotic covariance matrix of $\hat{\beta}$, can be given by the inverse of the observed information matrix. If we wish to evaluate the influence on the precision of the estimate, we may use a COVRATIO-like measure defined as

$$|\mathbf{I}(\hat{\beta})|/|\mathbf{I}(\tilde{\beta}_{(i)})|,$$

where $\tilde{\beta}_{(i)}$ indicates an approximate for $\hat{\beta}_{(i)}$ defined as

$$\tilde{\beta}_{(i)} = \hat{\beta} - \frac{\partial\hat{\beta}}{\partial w_i}.$$

4.3 Multiple-case diagnostics

Here we apply the general procedure of influence analysis (see, e.q., Tanaka [9], Tanaka and Zhang [11], which utilizes the additive property of the influence measured with influence functions. Let $\hat{\beta}_{(A)}$ be the estimate for β based on the sample without subset A of m individuals, i.e., $A = \{i_1, \cdots, i_m\}$. Then the estimate $\hat{\beta}_{(A)}$ can be approximated as

$$\hat{\beta}_{(A)} \cong \tilde{\beta}_{(A)} = \hat{\beta} - \sum_{i \in A} \frac{\partial\hat{\beta}}{\partial w_i},$$

by using up to the first order terms of the Taylor series expansion. Thus, if we can detect subsets of individuals with large values of the summation of influence functions $\sum_{i \in A}(\partial\hat{\beta}/\partial w_i)$, we may regard them candidates for influential subsets. Those subsets are characterized in such a way that the individuals within a subset are located far from and on similar directions from the origin in the space of influence functions. When the dimension of β is small, we may be able to find such candidates easily by inspecting directly the scatter plot of $\partial\hat{\beta}/\partial w_i, i = 1, \cdots, n$. When the dimension of β is large, however, we need to reduce the dimension. But, since the elements of $\hat{\beta}$ are mutually correlated, we should apply principal component analysis (PCA) with metric V^{-1} to the influence functions $\partial\hat{\beta}/\partial w_i$ instead of the ordinary

PCA. More precisely, we compute eigenvalues λ_j and associated eigenvectors $\mathbf{a}_j$ of the eigenproblem

$$\Big[\frac{1}{n}\sum_{k=1}^{n}(\frac{\partial\hat{\beta}}{\partial w_k})(\frac{\partial\hat{\beta}}{\partial w_k})^T - \lambda\mathbf{V}\Big]\mathbf{a} = \mathbf{0},$$

then compute the principal component (PC) scores $u_{jk} = \mathbf{a}_j^T\left(\partial\hat{\beta}/\partial w_k\right)$, $j = 1, \cdots, p; k = 1, \cdots, n$, draw scatter plots of dominent PC scores, and search for such subsets of individuals described above in the space of PC scores. It is known that there exists the relationship between the Cook's D and the PC scores such as

$$D_i = u_{1i}^2 + u_{2i}^2 + \cdots + u_{pi}^2, i = 1, \cdots, n.$$

This means that the PC scores provide the information of the multidimensional structure of the influence summarized in Cook's D. Also it is noted that the first PC scores give the information of the maximum curvature direction of Cook's [3] local influence (see, e.g., Tanaka and Zhang [11]). In this sense our general procedure provides a statistical interpretation of Cook's geometrical criterion.

5 Numerical example

For illustration we analyze a data set for 137 patients with lung cancer on a randomized clinical trial conducted by the Veteran's Administration, which is taken from Kalbfleisch and Prentice [5, pp. 223-224] In this data set 128 observations are uncensored and 9 are censored. The data set consists of the information on survival time, an indicator for censoring, prior therapy(prior), Karnofsky performance status(kps), age, months since diagnosis(diag), cell-type and treatment of test or standard chemotherapy(therapy). The cell type is composed of four categories such as squamous, small, adeno and large. These four types of cell are expressed by using dummy variables. When a Cox regression model with all covariates is fitted to the data set, three covariates kps, small and adeno are significant at 0.10 level. The model with these three covariates is also obtained as the model with the minimum AIC, when we apply the forward selection procedure of variable selection. Thus we select this three variables model as the final model. The fitted model is shown

Variable	Coef	exp(Coef)	se(Coef)	Z	p-value
Kps	-0.0305	0.97	0.00512	-5.95	0.000
Small	0.5740	1.78	0.21562	2.66	0.008
Adeno	1.0082	2.74	0.25830	3.90	0.000

Table 1: Results of model fitting.

in Table 1. Then our influence analysis is applied to this selected model to investigate if there exist individually or jointly influential observations. Figures 1 shows the index plot of Cook's D and Figure 2 gives the scatter plot of the first and second principal components obtained by PCA with metric $\mathbf{V}^{-1}$ of the influence functions, where the eigenvalues are 1.89, 0.96, 0.56, in order of their magnitudes. Looking these figures we can find that there are candidates for two individually influential observations, i.e., observations #131 and #52, because #131 and #52 are located far from the origin, but they are not on similar directions from the origin. However, we could not find any candidate for a subset of jointly influential observations. The actual influence of these two observations are shown in Table 2. For the purpose of checking the goodness of approximation of EIF to SIF the scatter plot is drawn for Cook's D's which are computed in two different manners.

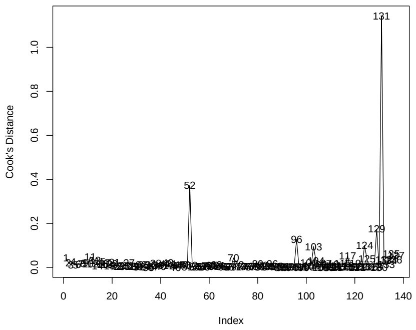


Figure 1: Index plot of Cook's D (Single-case diagnostics)

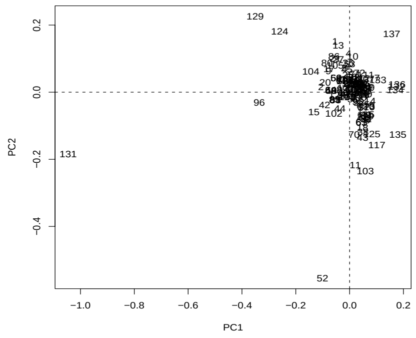


Figure 2: Scatter plot of two PCs of influence functions (Multiple-case diagnostics)

Case deleted	$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$	$(\hat{\beta} - \hat{\beta}_{(A)}/SE)$
None	(-0.03047,0.57401,1.00816)	
#131	(-0.03049,0.59136,1.03201)	(0.00283,-0.08051,-0.09234)
#52	(-0.03005,0.57355,1.03098)	(-0.08222,0.00210,-0.08837)
#52,#131	(-0.03006,0.59102,1.05541)	(-0.08078,-0.07890,-0.18295)

Note. The significance of the three regression coefficients are all $p < 0.01$

Table 2: Changes of the estimated regression coefficients.

One is based on $\partial\hat{\beta}/\partial w_i$ or EIF$_i$, and the other based on sample influence function(SIF$_i$), where

$$SIF_i = \hat{\beta} - \hat{\beta}_{(i)}, \quad i = 1, 2, \cdots, n,$$

$\hat{\beta}_{(i)}$ indicating the estimate based on the sample without the i-th individual. The resulting scatter plot is given in Figure 3. Looking this figure it can be said that the differential coefficient $\partial\hat{\beta}/\partial w_i$ gives a good approximation to the actual change $\hat{\beta} - \hat{\beta}_{(i)}$, and therefore it is expected that the scatter plots of PC scores of influence functions provide relatively accurate information of the influence of multiple observations.

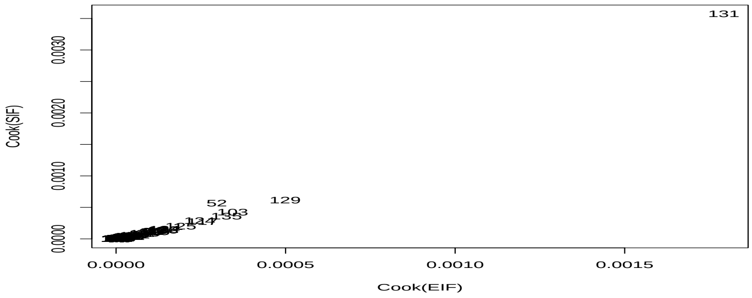


Figure 3. Scatter plot of Cook's D based on EIF and SIF

6 Concluding remarks

In the present paper we proposed a method of influence analysis in Cox regression with censored observations. It deals with not only single-case but also multiple-case diagnostics. Our method of multiple-case diagnostics is based on PCA with metric $\mathbf{V}^{-1}$ of the influence functions proposed by Tanaka and his coworkers, and it is known that the first principal component scores gives the information of the maximum curvature direction in Cook's local influence. In this sense it is closely related to Cook's local influence. It has an advantage compared to Cook's local influence that it can be applied to estimation methods other than the maximum likelihood method. In the numerical example in section 5 we could find two individually influential observations but no jointly influential one. It seems that so-called masking effects are not serious in this example. In our experiences the proposed method is effective when masking effects are small or intermediate. When masking effects are serious, we have to robustify the procedure as in Tanaka and Watadani [10]. It is a future work for us to study the robust procedure with an appropriate method of identifying a set of "inlying observations".

References

[1] Breslow N.E. (1974). *Covariance analysis of censored survival data.* Biometrics **30**, 89 – 99.

[2] Cain K.C., and Lange N.T. (1984). *Approximate case influence for the proportional hazards regression model with censored data.* Biometrics **40**, 493 – 499.

[3] Cook R.D. (1986). *Assessment of local influence.* J. R. Statist. Soc. **B48**, 133 – 169.

[4] Cox D.R. (1972). *Regression models and life-tables (with discussion).* J. R. Statist. Soc. **B34**, 187 – 220.

[5] Kalbfleisch J.D., Prentice R.L. (1980). *The statistical analysis of failure time data.* Wiley.

[6] Klein J.P., Moeschberger M.L. (2003). *Survival analysis.* Springer.

[7] Schoenfeld D. (1982) *Partial residuals for the proportional hazards regression model.* Biometrika **69**, 239 – 241.

[8] Storer B.E., Crowley J. (1985) *A diagnostic for Cox regression and general conditional likelihoods.* J. Am. Statist. Ass. **80**, 139 – 147.

[9] Tanaka Y. (1994) *Recent advance in sensitivity analysis in multivariate methods.* J. Jpn. Soc. Comp. Statist. **7**, 1 – 25.

[10] Tanaka Y., Watadani S. (1994). *Unmasking influential observations in multivariate methods.* Compstat 1994, Physica-Verlag, 292 – 297.

[11] Tanaka Y., Zhang F. (1999). *R-mode and Q-mode influence analyses in statistical modelling: Relationship between influence function approach and local influence approach.* Comp. Statist. & Data Analysis **31**, 2325 – 2347.

[12] Wei H.W., Kosorok R.M. (2000). *Masking unmasked in the proportional hazards model.* Biometrics **56**, 991 – 995.

[13] Wei H.W. and Su J.S. (1999). *Model choice and influential cases for survival studies.* Biometrics **55**, 1295 – 1299.

Acknowledgement: This research was partly supported by Japan Society of Promotion of Science (Grant-in-Aid for Scientific Research (C)13680374).

Address: J. Sung, Graduate School of Natural Science and Technology, Okayama University, Tsushima, Okayama 700-8530, Japan
Y. Tanaka, Faculty of Environmental Science & Technology, Okayama University, Tshusima, Okayama 700-8530, Japan

E-mail: `sungjm@ems.okayama-u.ac.jp`
`tanaka@ems.okayama-u.ac.jp`

COMPSTAT'2004 Symposium

EXISTENCE AND UNIQUENESS OF MINIMIZATION PROBLEMS WITH FOURIER BASED STABILIZERS

Terezie Šidlofová

Key words: Neural networks, minimization of functionals, regularization theory, stabilizers, Fourier transform.

COMPSTAT 2004 section: Neural networks and machine learning.

Abstract: We study minimization of regularized empirical error functional with a Fourier-based stabilizer. We prove existence and uniqueness of the solution. We also describe the shape of the minimizing function and show that it is in the form of a one-hidden layer feed-forward neural network with activation functions derived from the regularization part. Practical applications based on the idea have been studied performing best on tasks with lower input dimension or suitable conceptual characteristics (e.g. financial fields or image classification). These are unfortunately out of scope of the paper, so we kindly ask the reader to refer e.g. to [5].

1 Introduction

Learning from data usually means to fit a function to a set of data $z = \{(x_i, y_i);\ i = 1, \ldots, N\} \subseteq \boldsymbol{R}^d \times \boldsymbol{R}$. The problem is what type of functions will we use for the fitting, because there are infinitely many ways to go through the given points. And even if we have a reasonable set of functions (admissible set) to pick from, there is no guarantee that the problem will have a solution and that the solution will be unique.

Typically it is not necessary that the function fits the data exactly, we approximate. Thus nice functions (smooth, continuous) come into question and we also gain generalization (see [4]). Some of these properties are easily expressed by the set of admissible functions, but we might have more complicated (global) external information (a-priori knowledge) about the problem and want to add it, too.

Mathematical expression of these ideas lies in formulating a functional that would among admissible functions pick the one, that is reasonably close to the data and also agrees with global property assumptions ([1], [3], [11], [13], [14]). Existence and uniqueness of such a solution can be secured by minimizing a functional over a corresponding set of functions.

This article deals with a stabilizer based on Fourier transform proposed in [4]. In [3] it was suggested that regularization with Fourier based stabilizer can be reformulated in terms of Reproducing Kernel Hilbert Spaces. Taking advantage of these ideas we present construction of the admissible set (RKHS) and derive existence, uniqueness and the form of the solution of our minimization problem.

2 Preliminaries

A *Hilbert space* is a Banach space in which the norm is given by an inner product $\langle .,. \rangle$, that is $\|x\| = \langle x, x \rangle^{1/2}$. Sequences of elements of spaces are denoted by $\{x_n\}$ meaning $n \in \boldsymbol{N}_+$, where $\boldsymbol{N}_+$ is the set of positive integers.

The Banach space X^* of bounded (real-valued) linear functionals on X is called the dual space. It defines *weak convergence* on X. A sequence $\{x_n\} \in X$ converges weakly to x ($x_n \rightharpoonup x$) if and only if $\lim_{n\to\infty} |f(x_n) - f(x)| = 0$ for each fixed $f \in X^*$. Let X, Y be Banach spaces and $\mathcal{F} : X \to Y$ a mapping from X to Y. We define the *derivative of* $\mathcal{F}$ *in* f *in direction* h as $\mathrm{D}_h\mathcal{F}(f) = \lim_{t\to 0} \frac{\mathcal{F}(f+th)-\mathcal{F}(f)}{t}$. If $\mathrm{D}_h\mathcal{F}(f)$ is linear continuous and the limit is uniform in h, $\|h\| = 1$, we call the derivative the *Fréchet derivative.* We can analogously define the second and so on derivatives.

Let d, k be positive integers, $\Omega \subseteq \boldsymbol{R}^d$. We denote by $(C(\Omega), \|.\|_C)$ the space of continuous functions on Ω with the maximum norm. C_k will denote all functions with continuous Fréchet derivative up to order k. C_∞ denotes infinitely differentiable functions. We say that $f \in C_\infty$ belongs to the *Schwartz space* $\mathcal{S}(\boldsymbol{R}^n)$ if $p \cdot D^\alpha f$ is a bounded function for any multiindex $\alpha = (\alpha_1, \ldots, \alpha_n)$ and any polynomial $p = \sum_i c_{\beta_i} x_1^{\beta_{i_1}} \cdots x_n^{\beta_{i_n}}$ on $\boldsymbol{R}^n$ (where $D^\alpha(f) = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \ldots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n}$). For the sake of this article let us define the *normalized Lebesgue measure* m_d on $\boldsymbol{R}^d$ as $\mathrm{d}m_d(x) = (2\pi)^{-d/2}dx$ (introduced in [12]). The Lebesgue space $(\mathcal{L}_p(\Omega), \|.\|_p)$ of p-times integrable functions on Ω will be renormed: $\|f\|_p = \left\{\int_\Omega |f|^p \mathrm{d}m_d\right\}^{1/p}$. This will simplify the use of *Fourier transform* $\hat{f}$ of the function $f \in \mathcal{L}^1(\boldsymbol{R}^d)$: $\hat{f}(t) = \int_{\boldsymbol{R}^d} f(x)e^{-it\cdot x}\mathrm{d}m_d$, where $t \in \boldsymbol{R}^d$ and $t \cdot x = t_1x_1 + \cdots + t_dx_d$.

For a functional $\mathcal{F} : X \to (-\infty, +\infty]$ we write $\mathrm{dom}\,\mathcal{F} = \{f \in X : \mathcal{F}(f) < +\infty\}$ and call this set the *domain* of $\mathcal{F}$. *Continuity* of $\mathcal{F}$ in $f \in \mathrm{dom}\,\mathcal{F}$ is defined as usual. A functional is *sequentially lower semicontinuous* if and only if the convergence of $\{f_n\}$ to f implies $\mathcal{F}(f) \le \liminf_{n\to\infty} \mathcal{F}(f_n)$. Functional $\mathcal{F}$ is *weakly sequentially lower semicontinuous* if and only if $f_n \rightharpoonup f$ implies $\mathcal{F}(f) \le \liminf_{n\to\infty} \mathcal{F}(f_n)$.

A functional $\mathcal{F}$ is *convex* on a convex set $E \subseteq \mathrm{dom}\,\mathcal{F}$ if for all $f, g \in E$ and all $\lambda \in [0,1]$, $\mathcal{F}(\lambda f + (1-\lambda)g) \le \lambda\mathcal{F}(f) + (1-\lambda)\mathcal{F}(g)$. Functional $\mathcal{F}$ is *(strongly) quasi-convex* if for all $f, g \in E, f \neq g$ it holds: $\mathcal{F}\left(\frac{1}{2}f + \frac{1}{2}g\right)\ (<) \le \max\{\mathcal{F}(f), \mathcal{F}(g)\}$. Set E is *weakly sequentially compact* if any sequence in E has a weakly converging subsequence.

A symmetric real valued function $K(x,y)$ on X is *(strictly) positive definite* if for any $x_1, \ldots, x_d \in X$ and for any real $a_1, \ldots, a_d$ such that not all of them are zero, the sum $\sum_{i,j=1}^d a_i a_j K(x_i, x_j)$ is (positive) nonnegative.

A *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}(X)$ is a Hilbert space of functions $f : X \to \boldsymbol{R}$ (X is a nonempty set), where for all $x \in X$ the evaluation functionals $\mathcal{F}_x : f \mapsto f(x)$, are linear and bounded (i.e. continuous). Thus by Fréchet-Riesz Theorem [9, p. 19] we can define a unique *kernel* $K(.,.)$

corresponding to our RKHS as follows: $\mathcal{F}_x(f) = f(x) = \langle K(x,.), f(.)\rangle \quad \forall f \in \mathcal{H}$, ($\langle .,. \rangle$ is scalar product on $\mathcal{H}$). K is symmetric positive definite and defines a dot product (and norm) on $\mathcal{H}$. For any positive definite symmetric K we can construct an RKHS with K as a kernel ([14, p.1-3]).

3 Regularized empirical error functional

The task to find an optimal solution to the setting of approximating a data set $z = \{(x_i, y_i)\}_{i=1}^N \subseteq \boldsymbol{R}^d \times \boldsymbol{R}$ by a function from a general function space X is ill-posed. A standard method to cope with ill-posed problems is to impose additional (regularization) conditions on the solution ([4]). These are typically things like a-priori knowledge, or some smoothness constraints. The solution f_0 has to minimize a functional $\mathcal{F} : E \to \boldsymbol{R}$ that is composed of the error part and the "smoothness" part:

$$\mathcal{F}(f) = \mathcal{E}_z(f) + \gamma\Phi(f),$$

where $\mathcal{E}_z$ is the error functional depending on the data $z = \{(x_i, y_i)\}_{i=1}^N \subseteq \boldsymbol{R}^d \times \boldsymbol{R}$, Φ is the regularization part — the so called stabilizer and γ is the regularization parameter giving the trade-off between the two terms of the functional to be minimized.

An error functional is usually of the form $\mathcal{E}_z(f) = \sum_{i=1}^N V(f(x_i), y_i)$. A typical example of the empirical error functional is the classical mean square error:

$$\mathcal{E}_z(f) = \frac{1}{N}\sum_{i=1}^N (f(x_i) - y_i)^2.$$

In [4] a special stabilizer based on the Fourier Transform was proposed:

$$\Phi_G(f) = \int_{\boldsymbol{R}^d} \frac{|\hat{f}(s)|^2}{\hat{G}(s)} \mathrm{d}m_d(s),$$

where $\hat{G} : \boldsymbol{R}^d \to \boldsymbol{R}_+$ is symmetric ($\hat{G}(s) = \hat{G}(-s)$) function tending to zero as $\|s\| \to \infty$ (the last holds for any $G \in \mathcal{L}_1$). That means $1/\hat{G}$ is a high-pass filter.

Now we can define the functional $\mathcal{F}_G$ that is to be minimized:

$$\mathcal{F}_G(f) = \mathcal{E}_z(f) + \Phi_G(f) = \frac{1}{N}\sum_{i=1}^N (f(x_i) - y_i)^2 + \gamma \int_{\boldsymbol{R}^d} \frac{|\hat{f}(s)|^2}{\hat{G}(s)} \mathrm{d}m_d(s).$$

4 Existence and uniqueness of the solution

To minimize the functional $\mathcal{F}_G$ above we need to specify the set X (of admissible functions) over which we are minimizing and thus construct an optimization problem $(X, \mathcal{F}_G)$. We will build a special set of admissible functions $\mathcal{H}$

(RKHS) and obtain existence and uniqueness of solution to the minimization problem $(E, \mathcal{F}_G)$ for $E \subset \mathcal{H}$ satisfying some mild conditions. Observe that by using an RKHS we operate on functions defined pointwise (kernel property) and thus the error part $\mathcal{E}_z$ of $\mathcal{F}_G$ is well defined.

Let us first suppose existence and show uniqueness. For this purpose we will employ Reproducing Kernel Hilbert Spaces. We build an RKHS corresponding to the regularization part of our functional (so far the only conditions on G were $G \in \mathcal{L}_1, \hat{G}$ symmetric, positive):

Let us define

$$G^\dagger(x,y) = G(x-y) = \int_{\boldsymbol{R}^d} \hat{G}(t) e^{it.x} e^{-it.y} \mathrm{d}m_d(t).$$

For $G^\dagger \in \mathcal{S}(\boldsymbol{R}^{2d})$ symmetric positive definite we obtain an RKHS $\mathcal{H}$ (using the classical construction, see [3], [13],[14]). We put $\langle f, g\rangle_{\mathcal{H}} = \int_{\boldsymbol{R}^d} \frac{\hat{f}(s)\hat{g}^*(s)}{\hat{G}(s)}$ $\mathrm{d}m_n(s)$ and obtain the norm $\|f\|^2_{\mathcal{H}} = \int_{\boldsymbol{R}^d} \frac{|\hat{f}(s)|^2}{\hat{G}(s)} \mathrm{d}m_n(s)$, for $\mathcal{H} = \overline{\text{span}\ \{G^\dagger(x,.), x \in \boldsymbol{R}^d\}}$, where $\overline{\{\dots\}}$ denotes closure of the set $\{\dots\}$ and a^* means complex conjugate of a. It is easy to check the reproducing property of G on $\mathcal{H}$, that is $\langle f(x), G(x-y)\rangle_{\mathcal{H}} = f(y)$.
Now we will take advantage of a theorem mentioned for example in [2, p.15]:

Lemma 4.1 (Da71). *A strongly quasi-convex functional* $\geq$ *can achieve its minimum over a convex set* C *at no more than one point.*

Now we will show strong quasi-convexity for the functional $\mathcal{F}_G$:

Lemma 4.2. *With the notation from section 3, functional* $\mathcal{E}_z$ *is convex and functional* Φ_G *is strongly quasi convex on RKHS* $\mathcal{H}$. *Hence, also* $\mathcal{F}_G$ *is strongly quasi convex on* $\mathcal{H}$.

Proof: For the first part, $\mathcal{E}_z(f)$ is a sum of N elements, each of which is a convex functional, as (real) function $z \mapsto \frac{1}{N}(z-y_i)^2$ is convex.

To deal with the other functional, we observe that $\Phi_G(f) = \|f\|^2_{\mathcal{H}}$. We will prove that in any Hilbert space the norm $\|.\|$ satisfies strong quasi convexity: $\|\frac{1}{2}x + \frac{1}{2}y\|^2 < \max\{\|x\|^2, \|y\|^2\} \quad \forall x, y \in \mathcal{H}$. We will use the parallelogram law to show the fact. In any Hilbert space it holds, that $\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2)\ \forall x, y \in H$, and so we get: $\frac{1}{4}\|x+y\|^2 = \frac{2}{4}(\|x\|^2 + \|y\|^2) - \frac{1}{4}\|x-y\|^2$. Hence $\|\frac{1}{2}x + \frac{1}{2}y\|^2 \leq \frac{1}{2}(2\max\{\|x\|^2, \|y\|^2\}) - \frac{1}{4}\|x-y\|^2$. Since for $x \neq y$ we have $\|x-y\|^2 > 0$ and we get: $\|\frac{1}{2}x + \frac{1}{2}y\|^2 < \max\{\|x\|^2, \|y\|^2\}$ as proposed.

So we have $\mathcal{F}_G$ a sum of a convex and a strongly quasi convex functional and so clearly $\mathcal{F}_G$ is strongly quasi convex as claimed. □

Theorem 4.1. *If the problem* $(\mathcal{F}_G, \mathcal{H})$ *has a solution then it is unique for any* $G^\dagger \in \mathcal{S}(\boldsymbol{R}^{2d})$ *symmetric positive definite.*

Proof: By Lemma 4.2 we have strong quasi convexity of the problem and by Lemma 4.1 (since any space is convex) we obtain uniqueness. □

So we have proven uniqueness of the solution to the minimization problem $(\mathcal{F}_G, \mathcal{H})$. To prove existence we use two basic results of approximation theory, see for example [2, p. 7-13]:

Theorem 4.2 (Da71). *A weakly sequentially lower semicontinuous functional $\mathcal{F}$ defined on a weakly sequentially compact set E attains its minimum in f_0 such that $\mathcal{F}(f_0) = \inf_{f\in E} \mathcal{F}(f) = \min_{f\in E} \mathcal{F}(f)$.*

Weak lower sequential semicontinuity of a functional can be secured by several means, as for example by:

Theorem 4.3 (Da71). *A convex functional $\mathcal{F}$ that has first and second Fréchet derivatives at all points of an open convex set E is weakly sequentially lower semicontinuous in E.*

To apply Theorem 4.3 we have to prove the derivatives of $\mathcal{F}_G$ to exist.

Theorem 4.4. *Let $G : \boldsymbol{R}^d \to \boldsymbol{R}$ such that $G \in \mathcal{L}_1$ and $\hat{G}$ symmetric positive. Then functional $\mathcal{F}_G$ is weakly sequentially lower semicontinuous on $\mathcal{H}$ (or on any open convex subset of $\mathcal{H}$).*

Proof: Let us have a look at the regularization part. We compute the first derivative:

$$
\begin{aligned}
&\mathrm{D}_h\Phi_G(f) = \\
&= \lim_{t\to 0} \int_{\boldsymbol{R}^d} \frac{\left(\int_{\boldsymbol{R}^d}[f(x)+th(x)]e^{-ixs}\mathrm{d}m_d(x)\right)\left(\int_{\boldsymbol{R}^d}[f(\check{x})+th(\check{x})]e^{-i\check{x}s}\mathrm{d}m_d(\check{x})\right)^*}{t\hat{G}(s)} \\
&\quad - \frac{\left(\int_{\boldsymbol{R}^d} f(x)e^{-ixs}\mathrm{d}m_d(x)\right)\left(\int_{\boldsymbol{R}^d} f(\check{x})e^{-i\check{x}s}\mathrm{d}m_d(\check{x})\right)^*}{t\hat{G}(s)}\mathrm{d}m_d(s) \\
&= \int_{\boldsymbol{R}^d} \frac{\int_{\boldsymbol{R}^d}\int_{\boldsymbol{R}^d}\left(f(x)h(\check{x})^* + h(x)f(\check{x})^*\right)e^{-ixs}e^{i\check{x}s}\mathrm{d}m_d(x)\mathrm{d}m_d(\check{x})}{\hat{G}(s)}\mathrm{d}m_d(s) \\
&= \int_{\boldsymbol{R}^d} \frac{2\Re\left(\widehat{f}(s)\widehat{h}(s)^*\right)}{\hat{G}(s)}\mathrm{d}m_d(s)
\end{aligned}
$$

where $\mathrm{D}_h\Phi_G(f)$ means the first derivative of Φ_G in f in direction h.

Now we compute the second derivative:

$$
\begin{aligned}
\mathrm{DD}_{h,k}\Phi_G(f) = \lim_{t\to 0}\int_{\boldsymbol{R}^d} \frac{2\Re\left(\widehat{f+tk}(s)\widehat{h}^*(s)\right)}{t\hat{G}(s)}\mathrm{d}m_d(s) - \\
-\frac{2\Re\left(\widehat{f}(s)\widehat{h}^*(s)\right)}{t\hat{G}(s)}\mathrm{d}m_d(s) = \int_{\boldsymbol{R}^d}\frac{2\Re\left(\widehat{k}(s)\widehat{h}^*(s)\right)}{\hat{G}(s)}\mathrm{d}m_d(s)
\end{aligned}
$$

where $\mathrm{DD}_{h,k}\Phi_G(f)$ is the second derivative of Φ_G in f in directions h, k.

Now we will need also the error part derivative (recall the error part is of the form $\mathcal{E}_z(f) = \frac{1}{N}\sum_{i=1}^N (f(x_i) - y_i)^2$):

$$\mathrm{D}_h \mathcal{E}_z(f) = \frac{1}{N} \lim_{t\to 0} \frac{\sum_{i=1}^N (f(x_i) + th(x_i) - y_i)^2 - \sum_{i=1}^N (f(x_i) - y_i)^2}{t} =$$

$$= \frac{1}{N} \sum_{i=1}^N (2f(x_i)h(x_i) - 2h(x_i)y_i)$$

The second derivative is:

$$\mathrm{DD}_{h,k} \mathcal{E}_z(f) = \frac{1}{N} \lim_{t\to 0} \frac{\sum_{i=1}^N (2(f+tk)(x_i)h(x_i) - 2h(x_i)y_i)}{t} -$$

$$\frac{\sum_{i=1}^N (2f(x_i)h(x_i) - 2h(x_i)y_i)}{t} = \frac{1}{N} \sum_{i=1}^N 2k(x_i)h(x_i)$$

We see, that all directional derivatives are linear continuous uniform in h, resp. h, k, so we have Fréchet derivatives. By Theorem 4.3 $\mathcal{F}_G$ is weakly sequentially lower semicontinuous. □

Theorem 4.5. *The problem $(E, \mathcal{F}_G)$ has a solution for any $G^\dagger \in \mathcal{S}(\boldsymbol{R}^d)$ symmetric positive definite, $E \subset \mathcal{H}$ bounded weakly closed.*

Proof: Every bounded weakly closed subset of a reflexive space is weakly sequentially compact (see [10]). Since any Hilbert space is reflexive, we obtain the second condition of theorem 4.2 and using 4.4 the first condition comes and we conclude. □

5 The form of the solution

We can describe the shape of the solution using a basic theorem from mathematical analysis.

Theorem 5.1. *Let the functional $\mathcal{F}$ defined on a set E in a Banach space X be minimized at a point $f_0 \in E$, with f_0 an interior point in the norm topology. If $\mathcal{F}$ has a derivative $\mathrm{D}\mathcal{F}_{f_0}$ at f_0, then $\mathrm{D}\mathcal{F}_{f_0} = 0$.*

Existence and uniqueness of the solution have been proven, so we can use Theorem 5.1 to derive the form of the solution. Similar results have been sketched for example in [4] but without taking advantage of RKHS. The theorem presented is a modification of the well known Representer Theorem.

Theorem 5.2. *Let $G^\dagger : \boldsymbol{R}^{2d} \to \boldsymbol{R}$ be a positive definite symmetric function from $\mathcal{L}_1$ and let $G : \boldsymbol{R}^d \to \boldsymbol{R}$ have symmetric positive Fourier transform (with the notation from section 3). Then the unique minimizing function $f_0 \in E$ of the problem $(E, \mathcal{F}_G)$ ($E \subset \mathcal{H}$ bounded weakly closed) is of the form*

$$f_0(x) = \sum_{i=1}^N c_i G(x - x_i),$$

where x_i are the data points.

Proof: We have existence and uniqueness of f_0 from section 4.

The derivative of $\mathcal{F}_G$ in f in direction h is:

$$\mathrm{D}_h\mathcal{F}_G(f) = 2\frac{1}{N}\sum_{i=1}^{N}\Big(f(x_i)h(x_i) - h(x_i)y_i\Big) + \\ +\gamma\int_{\boldsymbol{R}^d}\frac{\widehat{f}(s)\widehat{h}(s)^* + \widehat{f}^*(s)\widehat{h}(s)}{\hat{G}(s)}\mathrm{d}m_d(s)$$

Now we put $\mathrm{D}_h\mathcal{F}_G(f_0) = 0$ (f_0 is the minimizing function). This has to hold for all $h \in \mathcal{H}$. Let us take $h(s) = G(s - x)$. We obtain (using symmetry of G and $\hat{G}$ and reproducing property of G) $0 = \mathrm{D}_h\mathcal{F}_G(f_0) = 2\frac{1}{N}\sum_{i=1}^{N}\Big(f_0(x_i)G(x_i - x) - G(x_i - x)y_i\Big) + 2\gamma f_0(x)$ and thus we have $\gamma f_0(x) = \frac{1}{N}\sum_{i=1}^{N} G(x - x_i)(f_0(x_i) - y_i)$. So we see that the solution must be in the form $f_0(x) = \sum_{i=1}^{N} c_i G(x - x_i)$. □

The solution to the problem $(E, \mathcal{F}_G)$ is very nice, since it is in the form of a neural network with G as the activation functions shifted to the data points x_i. On the other hand the shape isn't too surprising, since we are operating on $\mathcal{H}$ which is derived from functions of exactly this type. Anyway it is nice to know the minimizing function is one of the "basic" functions and not a wild one from the closure. The problem of the number of hidden units being too large to be implemented can be solved by variable basis approximation using the obtained shape of the activation functions (see [8]).

6 Conclusion

We have derived existence and uniqueness of the solution to the problem of finding a function close to the given data and simultaneously reasonably smooth (in terms of its Fourier transform). We showed that the solution is in the form of a one-hidden-layer feedforward neural network with activation functions depending on the form of the stabilizer.

The drawback of this approach is that the obtained neural network has too many hidden units (as much as the number of data), however algorithms using this approach work in practice. The problem of too many units can be dealt with by variable basis approximation limiting the number of hidden units apriori. Nice approximation properties have been proven. This is unfortunately out of the scope of this article, so we kindly ask the reader to refer for example to work [7], [15].

References

[1] Cucker F., Smale S. (2001). *On the mathematical foundations of learning.* Bulletin of the American Mathematical Society **39**, 1 – 49.

[2] Daniel J. W. (1971). *The approximate minimization of functionals.* Prentice-Hall, Inc.

[3] Girossi F. (1998). *An equivalence between sparse approximation and support vector machines.* Neural Computation **10**, 1455-1480, MIT. (A.I. Memo No. 1606, MIT, 1997).

[4] Girossi F., Jones M., Poggio T. (1995). *Regularization theory and neural networks architectures.* Neural Computation, **7**, 219 – 269.

[5] Kudová P. (2004). *Comparison of kernel based regularization networks and RBF neural networks.* Submitted to Compstat 2004, Prague.

[6] Kůrková V. (2003). *High-dimensional approximation by neural networks.* In Advances in Learning Theory: Methods, Models and Applications, Stuykens J. et. al. Ed., 69 – 88, Amsterdam, IOS Press.

[7] Kůrková V., Sanguinetti M. (2002). *Error estimates for approximate optimization by the extended Ritz method.* Research Report ICS-2002-882, Institute of Computer Science, Prague, submitted to SIAM Journal on Optimization.

[8] Kůrková V., Sanguinetti M. (2003). *Learning with generalization capability by kernel methods with bounded complexity.* Research Report ICS-2003-901, Institute of Computer Science, Prague, submitted to Journal of Complexity.

[9] Lukeš J. (2002). *Zápisky z funkcionální analýzy.* Karolinum, UK Praha.

[10] Lukeš J., Malý J. (1995). *Measure and integral.* Matfyzpress, Praha, 1995.

[11] Poggio T., Smale S. (2003). *The mathematics of learning: delaing with data.* Notices of the AMS **50** (5), 536 – 544.

[12] Rudin W. (1991). *Functional analysis*, 2nd Edition. McGraw-Hill, NY.

[13] Schölkopf B., Smola A. J. (2002). *Learning with kernels.* MIT Press, Cambridge, Massachusetts.

[14] Wahba G. (1990). *Spline models for observational data.* Series in Applied Mathematics **59**, SIAM, Philadelphia.

[15] Zoppoli R., Sanguinetti M., Parisini T. (2002). *Approximating networks and extended Ritz method for the solution of functional optimization problems.* J. of Optimization Theory and Applications **112**, 403 – 440.

Acknowledgement: Thanks to my advisor Věra Kůrková for useful guidance. This work was partially supported by GA ČR Grant 201/02/0428.

Address: T. Šidlofová, Institute of Computer Science, Academy of Sciences of CR, Pod vodárenskou věží 2, P.0. Box 5, 182 07 Prague 8, Czech Republic, Charles University, Faculty of Mathematics and Physics, Ke Karlovu 3, 121 16 Prague 2, Czech Republic.

E-mail: `terka@cs.cas.cz`

FITTING THE GENERALIZED LAMBDA DISTRIBUTION TO INCOME DATA

Agostino Tarsitano

Key words: Distribution fitting, grouped data, quantile function.

COMPSTAT 2004 section: Model selection.

Abstract: This paper proposes the generalized lambda distribution (GLD) as a model for describing the distribution of income over a population. Performances of various methods of fitting the GLD to grouped income data are evaluated. Of the estimators considered it is concluded that the unweighted least squares regression on group means should be used.

1 Introduction

There has been an increased interest in describing the distributions of personal income for the last several decades. A number of monographs have been published in the area, including those by Dagum [1], Kleiber and Kotz [4]. The study of income distributions usually provide a mathematical description F for the cumulative distribution of incomes and use it to summarize in a small number of parameters the peculiarities one discovers in empirical distributions. Also, F can be employed to smooth out irregularities in the histogram of observed data and to compute summary measures that can be compared spatially and temporally.

A wide variety of functional forms have been considered as possible models for incomes. One approach is to view the income density function as the outcome of a stochastic process (*e.g.* the Champernowne model). A second approach exploits the connections between income and aptitudes (*e.g.* the lognormal model). Also, the model is derived from a differential equation designed to capture a stable structure of observed distributions of income (*e.g.* Singh-Maddala model). Another approach is the search of a flexible analytic form, which ensures a satisfactory goodness of fit (*e.g.* the generalized beta model). Other approaches can no doubt be suggested.

The generalized lambda distribution (GLD) is a flexible and manageable tool for modeling empirical and theoretical distributions. The GLD is primarily specified by the quantile function

$$X_p(p;\lambda) = \lambda_1 + \lambda_2^{-1}\left[p^{\lambda_3} - q^{\lambda_4}\right] \quad 0 \leq p \leq 1,\ q = 1-p;\ \lambda_2 \neq 0 \tag{1}$$

Where λ_1 is a location parameter, λ_2 is a linear parameter related to (though not only to) the scale of X and λ_3, λ_4 are exponential parameters determining the shape of the quantile function. The following conditions are imposed:

$$If\ \lambda_2 \to \infty\ then\ \lambda_3,\ \lambda_4 > -\infty;\ If\ \lambda_3, \lambda_4 \to \infty\ then\ |\lambda_2| > 0 \tag{2}$$

Although there is scarcely a need for another model to fit the distribution of income the flexibility and the adaptability offered by the GLD legitimate its advancement in this context.

The basic proposition of this paper is that personal income distributions can be adequately described by using the quantile function (1). The content of the paper is organized as follows: in Section 2 the properties of GLD are described and its analytical and statistical peculiarities are summarized. Section 3 contains a discussion of several estimation procedures in the case of grouped data paying special attention to the extension of these methods to a random variable defined by its quantile function. The goodness-of-fit statistics assessing their usefulness are also considered. The results of an application to a real data set are exposed in Section 4 providing information about the relative merits of the different estimation techniques.

2 Shape, moments, and Lorenz curve of the GLD

The support of the GLD random variable is bounded $(\lambda_1 - 1/\lambda_2, \lambda_1 + 1/\lambda_2)$ if $\lambda_3, \lambda_4 > 0$ and is the real line when $\lambda_3, \lambda_4 < 0$. Hence, the extremes of $X(p, \lambda)$ are finite or infinite according to the sign of the exponential parameter. Analytic expression for the cumulative distribution function $F(x, \lambda)$ is in general not available. However, the fact that the GLD is not invertible is not a serious drawback because the same is true for many popular models such as lognormal and generalized beta. The limiting form of (1) as λ_3 diverges to ∞ is the Pareto distribution.

The probability density function of a GLD random variable is defined by the density quantile function, that is the density expressed in terms of $p = F(x, \lambda)$

$$\frac{1}{\frac{dX(p;\lambda)}{dp}} = h\left[X(p;\lambda)\right] = \frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 q^{\lambda_4 - 1}} \tag{3}$$

If $\lambda_3 = \lambda_4$ then (3) is symmetric about the pole $X = \lambda_1$. When scale and location are changed we transform the variable $Y = a + bX$. The transformed distribution is another member of the GLD family with λ_1, λ_2 replaced by $a + b\lambda_1$ and $b\lambda_2$ respectively. Expression $h[X(p, \lambda)]$ represents a legitimate probability density function if and only if it is nonnegative and integrates to one. The latter condition follows directly from (3). A good summary of the regions in which the GLD is well defined is given in Karian and Dudewicvz [3].

The ordinates of the density quantile function at the extremes of the range of variation are $(\lambda_2/\lambda_4, \lambda_2/\lambda_3)$ if $\lambda_3, \lambda_4 \geq 1$ and zero for $\lambda_3, \lambda_4 < 1$. The parameters λ_3 and λ_4 determine the type of tails of the GLD (provided that the sign of λ_2 ensures that (3) is a valid density function). For example, if $\lambda_3, \lambda_4 > 0$ then (3) has increasingly peakedness and short tails; if $\lambda_3, \lambda_4 < 0$ the tails have increasingly heaviness. The density tends to zero both as p goes to 0 and as p goes to 1 if, respectively, $\lambda_3 < 1$ and $\lambda_4 < 1$. On the other hand, if $\lambda_4 \geq 1(\lambda_3 \geq 1)$ then the density has truncated left (right) tail. The density (3) is unimodal if $\lambda_3, \lambda_4 > 2$, if $0 < \lambda_3, \lambda_4 < 1$ or if $0 < \lambda_3, \lambda_4 < 0$. It is zeromodal if $1 < \lambda_3, \lambda_4 < 2$. The arithmetic mean and the median of a GLD are

$$\mu = \lambda_1 + \lambda_2^{-1}\left\lfloor \frac{1}{(\lambda_3+1)} - \frac{1}{(\lambda_4+1)} \right\rfloor;\ M_e = \lambda_1 + \lambda_2^{-1}\left(0.5^{\lambda_3} - 0.5^{\lambda_4}\right) \quad (4)$$

Consider the linear transformation $Z = X - \lambda_1$. Then

$$E(Z^i) = \sum_{j=0}^{i} \binom{j}{i} (-1)^j \lambda_2^{-i} B(\lambda_3(i-j)+1, \lambda_4 j + 1);\ i = 1, 2, \cdots \quad (5)$$

Where $B(x, y)$ denotes the complete beta function. The i-th moment of the GLD exists if and only if $\min(\lambda_3, \lambda_4) > -i^{-1}$. Since $Z - E(Z) = X - E(X)$ the central moments of X coincide with the central moments of Z. The degree of skewness can be measured by

$$\begin{aligned} \frac{\mu - M_e}{S_{Me}} &= b(\lambda) \\ &= \frac{(\lambda_4+1)\left[1-(\lambda_3+1)0.5^{\lambda_3}\right] - (\lambda_3+1)\left[1-(\lambda_4+1)0.5^{\lambda_4}\right]}{(\lambda_4+1)\left[1-0.5^{\lambda_3}\right] + (\lambda_3+1)\left[1-0.5^{\lambda_4}\right]} \end{aligned} \quad (6)$$

where S_{Me} is the mean deviation about the median. From (6) it easily checked that (3) has a positive skewness if $\lambda_3 < \lambda_4$. The practical advantage of using $X(p, \lambda)$ instead of $F(x, \lambda)$ depends on having the $X(p, \lambda)$ in closed form. First, the Lorenz curve and other characteristics are handled simply.

$$L(p; \lambda) = \mu^{-1}\left\{\lambda_1 p + \lambda_2^{-1}\left[(\lambda_3+1)^{-1} p^{\lambda_3+1} + (\lambda_4+1)^{-1}\left(q^{\lambda_4+1} - 1\right)\right]\right\} \quad (7)$$

The condition $\lambda_2\lambda_3\lambda_4 \geq 0$ suffices to ensure the convexity of the Lorenz curve as long as the mean exists and $h[X(p, \lambda)]$ is a valid density function. Sarabia [7] used this model to define a hierarchy of Lorenz curves. Maddala and Singh [5] employed a version of (7) obtaining good results in terms of fitting. The use of (7) can be done analytically and not numerically. For instance, the Lorenz orderings can be obtained by a direct comparison of involved curves.

Second, several measures of inequality can be written as $\int J(p)X(p, \lambda)dp$ with $\int J(p)dp = 0$ where $J(.)$ is a monotone weight function. The following formulae express three well-known measures of income inequality.

Gini

$$\mu^{-1}\left\{\lambda_1 - \mu + 2\lambda_2^{-1}\left[(\lambda_3+1)^{-1} - (\lambda_4+1)^{-1}(\lambda_4+2)^{-1}\right]\right\} \quad (8)$$

Bonferroni

$$\mu^{-1}\left\{\mu-\lambda_1+\lambda_2^{-1}\left[(\lambda_4+1)^{-1}(\lambda+\psi(\lambda_4+2))-(\lambda_3+1)^{-2}\right]\right\} \tag{9}$$

Pietra-Ricci

$$\mu^{-1}\left\{(\mu-\lambda_1)p_\mu+\lambda_2^{-1}\left[(\lambda_3+1)^{-1}p_\mu^{\lambda_3+1}+(\lambda_4+1)^{-1}q_\mu^{\lambda_4+1}\right]\right\} \tag{10}$$

Where γ is the Eulero's constant and $\psi(.)$ is the digamma function.

Finally, the expected value of the i-th order statistic exists in closed form for each i

$$E(X_{i:n})=\lambda_1+\lambda_2^{-1}\left[\frac{B(n+1,\lambda_3)}{B(i,\lambda_3)}-\frac{B(n+1,\lambda_4)}{B(n-i+1,\lambda_4)}\right];\ i=1,\cdots,n \tag{11}$$

3 Parameter estimation

Suppose that n ordered incomes have been grouped (preserving the ordering) into k intervals where the boundaries are $(L_i, U_i],\ i=1,2,\cdots,k$. The number of values in the i-th interval is n_i with $\Sigma n_i = n$. The mean income is $m_i, f_i = n_i/n$ denotes the relative frequency, N_i and p_i are, respectively, the cumulative absolute and relative frequency of incomes not exceeding X_i. Clearly, the grouping scheme may significantly affect the parameter estimation and the variance of estimators. For instance, if the observations cluster significantly around particular values producing multimodal distributions, no GLD can give an acceptable agreement with this behavior.

Karian and Dudewicz [3, p. 155] considered the following system

$$S1: r_3=\frac{A_1-A_2}{A_3-A_1};\ r_4=\frac{A_4-A_5}{A_3-A_2};\ S2:\ r_2=\lambda_2^{-1}(A_3-A_2);\ r_1=\lambda_1+\lambda_2^{-1}A_1 \tag{12}$$

Where $A_i=(\alpha_i)^{\lambda_3}-(1-\alpha_i)^{\lambda_4},\ i=1,2,\cdots,5;\ \alpha_2<\alpha_1,\ \alpha_2<\alpha_3,\ \alpha_5<\alpha_4;\ \alpha_I$ is an observed percent point and r_i is its sample counterpart. The subsystem formed by the first two equations is free of $\lambda_1 and \lambda_2$. Now, given a solution $(\lambda_3,\lambda_4) of S1$, one can rapidly determine the best companion choice for (λ_1,λ_2) by solving the linear system $S2$. The roots of $S1$ can be obtained by a Newton method. This, however, should be preceded both by a trial and error search over the relevant range values of (λ_3,λ_4) and a direct search like the Nelder-Mead simplex algorithm to establish a reasonable starting point.

The method of quantiles has the advantage of being operative without the necessity of knowing every measurement. Moreover, the outliers are given less weight than in the moment estimates; in fact, (12) can be still be applied when the moments do not exist. The choice of α, however, involves an inherent arbitrariness. If the alfa's favor the central part of the distribution, then the X_i's around the mode are efficiently estimated, but at the cost of underestimating higher incomes. If the alfa's were selected in the tails then the most frequent incomes would be neglected. Karian and Dudewicz [3, p. 158] suggest: $\alpha=(0.5,0.1,0.9,0.75,0.25)$ which is quite unsatisfactory for income distributions that are typically skewed to the right. The estimates determined by equating four percentage points seem to be a valid alternative to (12). However, all the $C_{k-1,4}$ combinations should be investigated (supposing that at least one of the non linear four equations systems will give permissible values) to establish an optimal choice. The difficulties of applying this method for large k are such that it would probably be better to abandon it.

The method of moments has been advocated because of its widespread use in practice. The first step is the solution, following closely that of $S1$, of a nonlinear system that depends solely on (λ_3, λ_4)

$$\gamma_1 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{x}}{s} \right)^3 \quad \gamma_2 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{x}}{s} \right)^4 \tag{13}$$

Once the best values for $(\lambda_3,\ \lambda_4)$ have been attained, the values of $(\lambda_1,\ \lambda_2)$ are given by $\lambda_2 = \pm(b - a^2)^{0.5}/s,\ \lambda_1 = -a/\lambda_3,\ a = (1+\lambda_3)^{-1} - (1+\lambda_4)^{-1},\ b = (1+2\lambda_3)^{-1} - (1+2\lambda_4)^{-1} - 2B(1+\lambda_3, 1+\lambda_4),\ \min(\lambda_3, \lambda_4) \geq -0.25$.

The method of moments is inadequate. Its use is restricted to distributions possessing their first four moments, but the heavy tail usually observed in empirical income distributions does not support such a premise. Furthermore, when the available data are grouped, a correction for grouping should be considered and if L_1 and/or U_k were left unspecified, the moments cannot be estimated without making arbitrary assumptions. On the other hand (11) is cryptic: the GLD density is symmetric for $\lambda_3 = \lambda_4$ but $\gamma_1 = 0$ even if $\lambda_3 \neq \lambda_4$ and it is far from clear which characteristic is being measured by γ_2 in skewed distributions. Finally, for some data sets, the iterative process might converge to (λ_3, λ_4) for which GLD has no finite moments. The method of quantiles and the method of moments do not appear to be very convenient for income data at the present. The ordinary least squares estimates of λ can be obtained by minimizing

$$\begin{aligned}
S(\lambda) &= \sum_{i=1}^{k} \left[y_i - \lambda_1 - \beta_2 g_i(\lambda_3, \lambda_4) \right]^2 f_i; \quad \beta_2 = \lambda_2^{-1} \\
M1 &: \ y_i = U_i;\ g_i = p_i^{\lambda_3} - q_i^{\lambda_4};\ i = 1, 2, \cdots, k-1 \\
M2 &: \ y_i = U_i;\ g_i = \frac{B(n+1, \lambda_3)}{B(N_i, \lambda_3)} - \frac{B(n+1, \lambda_4)}{B(n - N_i + 1, \lambda_4)}, \\
&\quad\ i = 1, 2, \cdots, k-1 \\
M3 &: \ y_i = m_i;\ g_i = \frac{p_i^{\lambda_3+1} - p_{i-1}^{\lambda_3+1}}{f_i(\lambda_3+1)} + \frac{q_i^{\lambda_4+1} - q_{i-1}^{\lambda_4+1}}{f_i(\lambda_4+1)}, i = 1, 2, \cdots, k
\end{aligned}$$

$M1$ defines the estimators that minimize the sum of squared differences between predicted and observed quantiles. $M2$, based on (9), is an extension to grouped data of the method proposed by Oztürk and Dale [6]. $M3$ suggests itself because of the importance of the group means for measuring income inequality. This new approach is more demanding since it requires knowledge of the mean of each income group, but has the advantage of using more information than the other methods. Since λ_1 and λ_2 are in linear form, they can be replaced by their least squares estimates given (λ_3, λ_4)

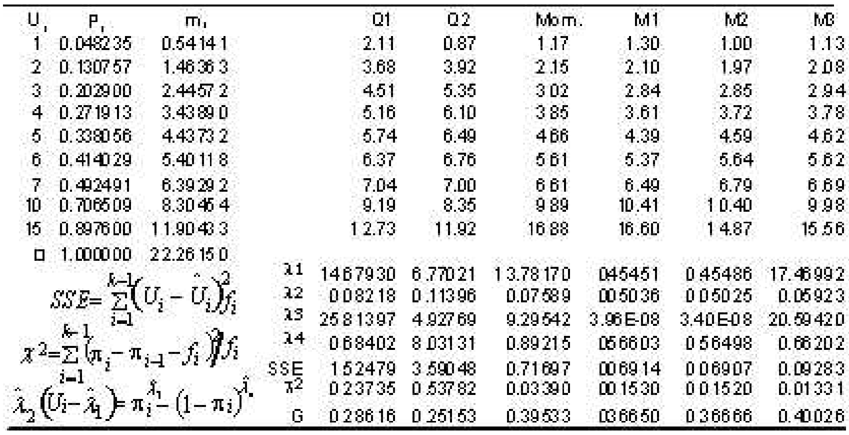
$$\begin{aligned} \hat{\lambda_1} &= \bar{y} - \hat{\lambda_2}^{-1}\bar{g}. \\ \hat{\lambda_2} &= \frac{1}{\hat{\beta_2}} = \frac{\sum_{i=1}^{k}(g_i - \bar{g})^2 f_i}{\sum_{i=1}^{k}(y_i - \bar{y})(g_i - \bar{g}) f_i} \\ &\Rightarrow S(\lambda_3, \lambda_4) = (1 - r_{yg}^2)\sum_{i=1}^{k}(y_i - \bar{y})^2 f_i \end{aligned} \tag{14}$$

Where r_{yg} is the correlation coefficient between y and g and r_yg does not depend on λ_1, β_2. Therefore, the pair (λ_3, λ_4) that minimizes $[1-(r_{yg})^2]$ also minimizes $S(\lambda_3, \lambda_4)$. It should be remarked that $S(\lambda_3, \lambda_4)$ in (14), like $S1$ and (12), can have multiple solutions or no solution for some data sets. Even when a solution exists, the numerical procedure devoted to its search may fail to find it because of convergence failure. Moreover, the observed y_i will not have equal variance nor will they be uncorrelated. Since this drawback is, at least in theory, serious further studies (e.g. in the line of generalized least squares) are needed to assess the effectiveness of GLD for income data.

4 Parameter estimation

Gastwirth [2] gives an income distribution in ten classes. The Gini index for the entire sample is 0.4014 and the crude bounds within which the index must lie are $(0.3883, 0.4083)$. Table 1 reveals the relative merit of five distinct estimators of λ.

Since the α_i have not been reached, the r_i's in (12) were computed by using linear interpolation on the given values ($Q1$) and the closest observed quantiles ($Q2$). It is easily seen that the quantile estimates depend markedly on the particular choice of percentage points. The moments have been calculated by assuming that all incomes in the i-th interval equal the average income m_i whereas, the solutions of (14), were obtained by using the Nelder-Mead simplex procedure. According to the SSE there is a sufficiently close agreement between observed and estimated percentiles with the exception of the two methods based on quantiles. As a general result, the fit of GLD is reasonable good in the middle part, but is poor in describing both the upper and the lower tails. The best performance has been obtained by $M2$ with $M1$ close competitor. $M3$ has an unduly bad fit in the last class. The Chi-squared criterion confirms the ranking of the six techniques determined by SSE. However, only the method of moments and the method of least squares on group means were able to provide an estimated Gini index (reported in



U_i	P_i	m_i		Q1	Q2	Mom.	M1	M2	M3
1	0.048235	0.54141		2.11	0.87	1.17	1.30	1.00	1.13
2	0.130757	1.46363		3.68	3.92	2.15	2.10	1.97	2.08
3	0.202900	2.44572		4.51	5.35	3.02	2.84	2.85	2.94
4	0.271913	3.43890		5.16	6.10	3.85	3.61	3.72	3.78
5	0.338056	4.43732		5.74	6.49	4.66	4.39	4.59	4.62
6	0.414029	5.40118		6.37	6.76	5.61	5.37	5.64	5.62
7	0.492491	6.39292		7.04	7.00	6.61	6.49	6.79	6.69
10	0.706509	8.30464		9.19	8.35	9.89	10.41	10.40	9.98
15	0.897600	11.90433		12.73	11.92	16.88	16.60	14.87	15.56
□	1.000000	22.26150							
$SSE=\sum_{i=1}^{k-1}\left(U_i-\hat{U}_i\right)^2 f_i$			$\lambda1$	1467930	6.77021	13.78170	045451	045486	17.46992
			$\lambda2$	008218	0.11396	0.07589	005036	005025	0.05923
$\chi^2=\sum_{i=1}^{k-1}\left(\pi_i-\pi_{i-1}-f_i\right)^2/f_i$			$\lambda3$	2581397	4.92769	9.29542	3.96E-08	3.40E-08	20.59420
			$\lambda4$	068402	8.03131	0.89215	056603	056498	0.66202
			SSE	152479	3.59048	0.71697	006914	006907	0.09283
$\hat{\lambda}_2\left(U_i-\hat{\lambda}_1\right)=\pi_i^{\hat{\lambda}_1}-\left(1-\pi_i\right)^{\hat{\lambda}}$.			χ^2	023735	0.53782	0.03390	001530	001520	0.01331
			G	028616	0.25153	0.39533	036650	036666	0.40026

Table 1: Observed and estimated quantiles of income data.

the last row of Table 1) lying inside the prescribed bounds. In this sense $M3$ carries the gold medal.

References

[1] Dagum C. (1990). *Generation and properties of income distribution functions.* In C. Dagum, M. Zenga (eds.), Income and Wealth Distribution, Inequality and Poverty. Springer-Verlag Berlin Heidelberg, 1 − 17.

[2] Gastwirth J.L. (1971). *The estimation of the Lorenz curve and Gini index.* The Review of Economics and Statistics **54**, 306 − 316.

[3] Karian Z.A., Dudewicz E.J. (2000). *Fitting statistical distributions.* The Generalized lambda distribution and gereralized bootstrap methods. CRC, Boca Raton (FL).

[4] Kleiber C., Kotz S. (2003). *Statistical size distributions in economics and actuarial sciences.* John Wiley &, Sons, New York.

[5] Maddala G.S., Singh A.K. (1977). *A flexible functional form for Lorenz curves.* Economie Appliquée, **30**, 481 − 486.

[6] Oztürk A., Dale R.F. (1985). *Least squares estimation of the parameters of the generalized lambda distributions.* Technometrics, **27**, 81 − 84.

[7] Sarabia J. M. (1996). *A hierarchy of Lorenz curves based on generalized Tukey's lambda distribution.* Econometric Reviews, **16**, 305 − 320.

Address: A. Tarsitano, Dipartimento di Economia e Statistica. Universita della Calabria, 87030 Arcavacata di Rende (Cs). Italy

E-mail: `agotar@unical.it`

COMPSTAT'2004 Symposium

AN APPLICATION OF CORRESPONDENCE ANALYSIS TO THE CLASSIFICATION OF CAUSES OF DEATH AMONG JAPANESE HEMOPHILIACS WITH HIV-1

Shinobu Tatsunami, Masaschi Taki, Rie Kuwabara and Kazutoschi Yamada

Key words: Correspondence analysis, medical application, HIV-1.

COMPSTAT 2004 section: Applications.

Abstract: Correspondence analysis was utilized to study whether there were new trends in the causes of death over time among Japanese hemophiliacs infected with HIV. The disease codes were used as row variables, while the year of death of the patients with the disease was used as a column variable. The hazard function computed for the present hemophiliacs showed a remarkable decline in 1997, when protease inhibitors were approved officially in Japan. However, it has been fluctuating perceptibly after 1999. Major factors contributing to the increase of the hazard were hemorrhage and liver disease. Regarding other diseases, Japanese frequencies of Kaposi's sarcoma and non-Hodgkin lymphoma looked different from those observed in the United States. The most important clinical implication is to focus on the prevention of liver diseases.

1 Introduction

We analyzed postmortem data among Japanese hemophiliacs with HIV-1 infection who had died of AIDS-related or other diseases. The diagnosis of AIDS in Japan is based on the onset of one or more of 23 AIDS-defining diseases. Therefore, many independent variables are necessary to identify all of the diseases that may cause death. To this end, we used correspondence analysis [1] to study whether there were new trends in the causes of death over time. The onset rates of AIDS-defining diseases as well as other diseases may change independently of one another. Understanding the trends requires proper classification of their patterns of occurrence. If we observe the frequency of occurrence of these diseases by aligning the disease and the year of death as row and column variables, correspondence analysis becomes a powerful analytical tool to classify the pattern of occurrence over time. This will help to identify the central problems that are associated with death in hemophiliacs in the current era of HIV therapy.

2 Methods

Variables for the analyses presented here were obtained from a national surveillance of coagulation disorders in Japan dated at the end of May 2003. Among the 1407 hemophiliacs registered by the end of May 2003, a total of 554 deaths (Hemophilia A: 428; Hemophilia B: 126) had occurred. Causes of death were unreported in 28 cases, leaving 526 cases for analysis.

The AIDS-defining diseases were termed *def01, def02, ...* through *def23* (Table 1). The number of patients with each disease was recorded in the interval from the beginning of 1983 to the end of May 2003. The number of cases with hemorrhage *(hem)*, liver disease *(lid)*, and other diseases *(oth)* reported as causes of death were summarized as well.

Code	Disease	Code	Disease
def01	Candidiasis (esophagus, bronchia, bronchial tree, lung)	*def13*	Cytomegalovirus infection
		def14	Herpes simplex virus infection
def02	Cryptococcosis	*def15*	Progressive multifocal leukoencephalopathy
def03	Coccidioidomycosis		
def04	Histoplasmosis	*def16*	Kaposi's sarcoma
def05	*Pneumocystis carinii* pneumonia	*def17*	Primary brain lymphoma
def06	Toxoplasma encephalitis	*def18*	Non-Hodgkin lymphoma
def07	Cryptosporidiosis	*def19*	Cervical cancer, invasive
def08	Isosporiasis	*def20*	Pneumonia, recurrent
def09	Pyogenic bacterial infections (younger than 13 years old)	*def21*	Lymphatic interstitional pneumonia / Hyperplasia (younger than 13 years old)
def10	Salmonellosis (recurrent, except salmonella typhi)		
def11	Tuberculosis	*def22*	HIV-associated encephalopathy
def12	Atypical mycobacterial infection	*def23*	HIV-associated wasting syndrome

Table 1: AIDS-defining diseases in 1999 diagnosis criteria.

The disease codes from *def01* to *def23*, *hem*, *lid*, and *oth* were used as row variables, while the year of death of the patients with the disease was used as a column variable. That is, the table f_{ij} defined as the frequency of i-th disease within j-th calendar year was forwarded to analysis. Computations were performed using SPSS version 11.5J (SPSS Japan, Tokyo, Japan).

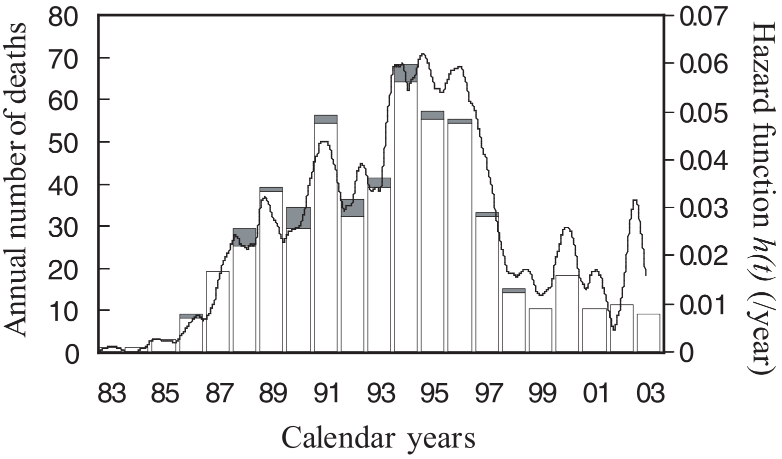


Figure 1: Changes in the annual number of deaths among Japanese hemophiliacs infected with HIV-1. A total of 554 deaths had been reported by the end of May 2003 among 1407 hemophiliacs with HIV-1. The 28 cases for whom the cause of death was unknown are shown in the shadowed area. The hazard function $h(t)$ computed for the present hemophiliacs is illustrated simultaneously.

In addition, hazard function $h(t)$ in the present hemophiliacs was computed by kernel-smoothed method using the Nelson-Aalen estimator [3]. A fixed bandwidth 0.5 year was used in the computation.

3 Results

3.1 Changes in the number of deaths and hazard function

Changes in the number of deaths annually from the beginning of 1983 to the end of May 2003 among 554 hemophiliacs are illustrated in Figure 1. The 28 cases for whom the cause of death was unknown are shown in the shadowed area in Figure 1, but were not included in the present correspondence analysis.

The hazard function $h(t)$ computed in the present hemophiliacs is illustrated simultaneously in Figure 1. It rose the most remarkably between 1984 and 1994, remained almost constant between 1994 and 1996, and then decreased markedly in 1997. It remained very low in two years (mostly below 0.02/year), however, tended to fluctuate after 1999.

3.2 Frequency of disease occurrence

The frequency f_{ij} observed in the present subjects is summarized in Figure 2. Because there were no occurrences of *def03, def08, def19*, and *def21*,

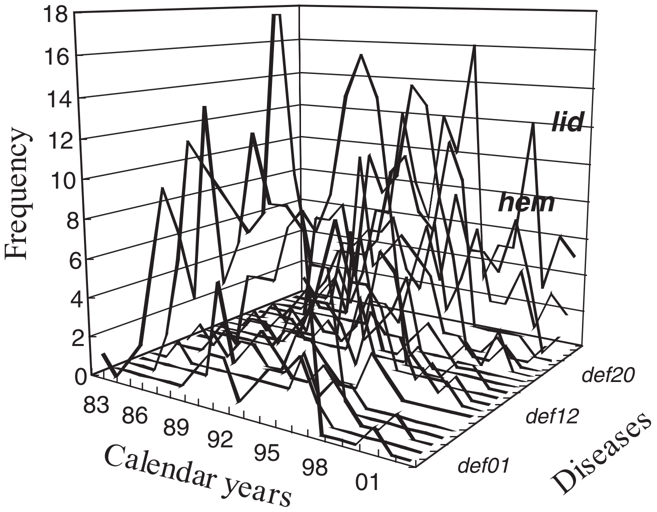


Figure 2: Observed frequency of occurrence of AIDS-defining diseases (*def01 def23*, except 03,04,08,09,10,19,21), hemorrhage *(hem)*, liver disease *(lid)*, and other diseases *(oth)* as causes of death from the beginning of 1983 to the end of May 2003.

and only one or two reports of *def04* (in 1993), *def09* (in 1989 and 1991) and *def10* (in 1996), these diseases were excluded from the analysis. Therefore, the frequencies of the remaining diseases, f_{ij} $(i \neq 3, 4, 8, 9, 10, 19, 21)$, were subsequently analyzed.

3.3 Results of correspondence analysis

The value of χ^2 in table f_{ij} was 525.4 with 360 degrees of freedom, indicating that the row and column variables were significantly associated ($p < 0.001$). Thus, the application of correspondence analysis to the table f_{ij} will be statistically meaningful.

The inertia, proportion of inertia and its cumulative obtained from correspondence analysis by symmetrical normalization are summarized in the first five dimensions (Table 2). Inertia represents the fraction of variance (weighted sum of squared distance to the origin in the full dimension) explained by each dimension, reflecting the relative importance of each dimension. In the present table, f_{ij}, the maximum number of dimensions is 18 and the total sum of inertia (from dimension 1 to dimension 18) was 0.564. Thus, the first five dimensions display 76.0% of the total inertia.

Although the cumulative proportion of inertia in dimension 1 and 2 is 47.6% (Table 2) of the total inertia, the relationships between diseases and year of death are easy to understand quantitatively in the two-dimensional

Dimension	Inertia	Proportion of inertia* (%)	Cumulative (%)
1	0.181	32.0	32.0
2	0.088	15.6	47.6
3	0.067	11.8	59.5
4	0.055	9.7	69.2
5	0.039	6.8	76.0

*Sum of inertia from dimension 1to dimension 18: 0.564

Table 2: Inertial, proportion of inertia and its cumulative in the first five dimensions.

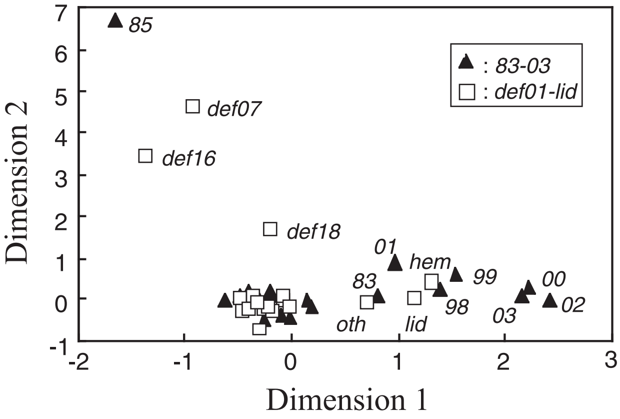


Figure 3: Simultaneous plots of row (disease) and column (year of death) scores in two dimensions. Codes of disease and year of death are indicated by open squares and closed triangles, respectively.

plot (Figure 3). The interpretation of the plot is fairly simple, namely, points that are closer together are more alike than points that are far apart.

Points representing almost all the diseases are concentrated near the origin (0, 0) in Figure 3, indicating similar or identical appearance in time. However, *def07* (cryptosporidiosis), *def16* (Kaposi's sarcoma), *def18* (non-Hodgkin lymphoma), *hem* (hemorrhages), *lid* (liver diseases) and *oth* (other diseases) resulted in separate plots, indicating that their temporal pattern of occurrence is not same as the other diseases.

The year of death, interpreted as a column variable, is plotted simultaneously in Figure 3. All the diseases other than *def07, def16, def18, hem, lid* and *oth* are plotted with calendar years from 84 to 97 in Figure 3. This was because of the synchronized appearance of the diseases; most of them arose between 1984 and 1996 and decreased abruptly in 1997. Reports of AIDS-defining diseases causing death have been rare after 1997.

On the other hand, the appearance of *hem* (hemorrhage) and *lid* (liver diseases) was clearly distinguishable from other diseases. Their appearance did not become negligible even after 1997 and in fact showed a slight increase between 1997 and 2000 (Figure 2). As a result, they are plotted near from 98 to 03 in Figure 3.

Regarding diseases *def07* (cryptosporidiosis) and *def16* (Kaposi's sarcoma), their appearance looked random, and they were reported only in a small number of patients (5 and 7, respectively). Therefore, they are plotted separately from the other diseases. Although reports of *def18* (non-Hodgkin lymphoma) were not so rare as *def07* or *def16*, its appearance also looked random except for its peak in 1991 (6 cases in 1991). Thus, def18 is located between the boundary of the other AIDS-defining diseases area and *def16* in Figure 3.

All three of the diseases mentioned above appeared simultaneously in1985. This resulted in a plot for the year 85 that was quite separate from the other years and the other AIDS-defining diseases.

4 Discussions

Our study demonstrates that the most notable diseases after 1997 are *hem* and *lid* as illustrated in Figure 3. The hazard function $h(t)$ computed for the present hemophiliacs showed a remarkable decline in 1997, when protease inhibitors were approved officially in Japan. However, it tended to fluctuate after 1999 (Figure 1). The major factors contributing to increase of $h(t)$ are *hem* and *lid*.

Liver diseases such as liver cirrhosis, liver failure and hepetocellular carcinoma are attributable to the high prevalence of hepatitis C virus (HCV) co-infection in hemophiliacs. In fact, the prevalence of HCV infection in Japanese hemophiliacs with HIV was higher than 98% [9]. Additionally, adverse hepatic effects of antiretroviral therapy may also be a contributing factor to liver disease.

Hemorrhage is naturally a possible cause of death among hemophiliacs [7]. In addition, a tendency for bleeding is sometimes reported in protease inhibitor therapy [10]. Therefore, the largest contribution of *hem* in dimension 1 as illustrated in Figure 3 should be noted. Further, fatal rupture of the esophageal varix is often caused by liver cirrhosis. As a result, associated with the increase of critical liver diseases, more hemorrhage will appear as one of the causes of death. Therefore *lid* might be one of the causes of the outstanding position of *hem* in dimension 1 in Figure 3.

The co-infection with HCV has been already known as a crucial factor that affect onset of opportunistic infections and survival among injection drug users [4]. However, the close relation between *hem* and *lid* demonstrated in Figure 3 is a unique characteristic in hemophiliacs.

Regarding other diseases, present correspondence analysis revealed that several diseases that have different patterns of appearance within the spectrum of AIDS-defining diseases. Almost all of the AIDS-defining diseases showed similar increases over time before the introduction of protease inhibitors but then declined after 1997. However, the time of occurrence of *def07, def16 and def18* appeared to be randomly distributed throughout the observation period. The rarity of *def16* (Kaposi's sarcoma) throughout the observation period among Japanese hemophiliacs should be noted because until 1996 it was the most commonly reported cancer among patients dying with HIV-1 in the United States [8]. Japanese hemophiliacs were infected with HIV-1 through non-heat-treated clotting factor concentrates imported from the United States. Therefore, the difference in the frequency of Kaposi's sarcoma cannot be explained by the difference of viral types but from the difference in host properties.

Patients with insufficient immunologic and virological responses after highly active antiretroviral therapy (HAART) are reported to be at highest risk of non-Hodgkin lymphoma [2]. However, the present analysis did not show an increase in the incidence of non-Hodgkin lymphoma after widespread of HAART in Japanese hemophiliacs.

Comments will be needed to use calendar years in the present analysis in-stead of the time from the infection with HIV. The exact date of individual infection was not discernible in our data. However, tentative mean date of seroconversion in hemophiliacs has been proposed as April 1983 [6]. Therefore, almost 14 years had passed after the infection at the time of official approval of protease inhibitors in Japan in 1997.

In conclusion, this study provides valuable information on causes of death in HIV-positive hemophiliacs in the period after protease inhibitors became available. The most important clinical implication is to focus on the prevention of liver diseases. Appropriate treatments for hepatitis [5] should also be undertaken to achieve survival benefits.

References

[1] Greenacre M.J. (1984). *Theory and application of correspondence analysis.* Academic Press, London.

[2] Kirk O., Pedersen C., Cozzi-Lepri A., et al. (2001). *Non-Hodgkin lymphoma in HIV-infected patients in the era of highly active antiretroviral therapy.* Blood **98**, 3406 – 3412.

[3] Klein J.P., Moeschberger M.L. (1997). *Survival analysis.* Techniques for censored and truncated data. Springer-Verlag, New York.

[4] Klein M.B., Lalonde R.G., Suissa S. (2003). *The impact of hepatitis C virus coinfection on HIV progression before and after highly active antiretro-viral therapy.* Journal of Acquired Immunodeficiency Syndromes **33**, 365 – 372.

[5] Landau A., Batisse D., Van Huyen J.P., et al. (2000). *Efficacy and safety of combination therapy with interferon-alpha2b and ribavirin for chronic hepatitis C in HIV-infected patients.* AIDS **14**, 839 – 844.

[6] Mimaya J., Meguro T., Tatsunami S., et al. (1992). *Natural history committee report.* In: Annual Report of Research Committee on Prevention of Developing Illnesses and Therapy for HIV Infected Patients 1991, 9 – 16.

[7] Ragni M.V. (1998). *Progression of HIV in haemophilia.* Haemophilia **4**, 601 – 609.

[8] Selik R.M., Byers, Jr. R.H., Dworkin M.S. (2002). *Trends in diseases reported on U.S. death certificates that mentioned HIV infection, 1987-1999.* Journal of Acquired Immunodeficiency Syndromes **29**, 378 – 387.

[9] Taki M., Tatsunami S., Shirahata A., et al. (2003). *Prevalence of hepatitis C virus infection in coagulation disorders in Japan.* International Journal of Hematology **77**, 528 – 529.

[10] Wilde J.T. (2000). *Protease inhibitor therapy and bleeding.* Haemophilia **6**, 485 – 490.

Address: S. Tatsunami, M. Taki, R. Kuwabara, K. Yamada, Unit of Medical Statistics, Faculty of Education and Culture, St. Marianna University School of Medicine. 2-16-1 Sugao, Miyamae-ku, Kawa-saki, Japan 216-8511

E-mail: s2tatsu@marianna-u.ac.jp

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

DOUBLE MONTE-CARLO SIMULATIONS IN FOOD RISK ASSESSMENT

Jessica Tressou

Key words: Risk assessment, contaminant, incomplete generalized U-statistics, bootstrap, jackknife, ochratoxin A.

COMPSTAT 2004 section: Nonparametrical statistics.

Abstract: In this paper the asymptotic properties of some incomplete generalized U-statistics well suited for risk assessment of the exposure to contaminants, when both contamination data and individual consumptions are available, are studied.

1 Motivations

Food may be naturally contaminated by some chemical components which may become toxic for the human organism if the total amount ingested through food consumption exceeds a certain tolerable dose. For example, Ochratoxin A (OTA) is a natural mycotoxin produced by fungi of the *Aspergillus* and *Penicillium* families, which has been classified as a genotoxic carcinogen in 1998 by the European Scientific Committee for Food. It may be detected in many products including cereals, grapefruit, dry fruits or vegetables, wine, coffee, beer, or pork and poultry meat.

An important toxicological concept to measure the medical impact of a contaminant is the so called Provisional Tolerable Weekly Intake (PTWI) expressed in terms of nanogram per body weight per week (ng/kgbw/wk in the following). It is fixed in Europe at 35 ng/kgbw/wk for OTA. This quantity is the scientifically and medically recognized level over which a permanent excess may be considered as potentially dangerous for the human health (without any distinction between individuals except their body weight). Even though its value may not be the same for different countries, this quantity generally serves as the basis to decide whether or not there is a specific public health problem related to a particular contaminant and to plan food regulatory programs. In particular, an important issue is to evaluate whether the (complete or partial) suppression of the contaminated products or the reduction of the contamination in some product (for instance by imposing a maximal limit to certain commercialized items) may have a significant impact on the global exposure of the individuals.

Our approach in this study consists in evaluating the probability that the individual exposure over a week exceeds the PTWI. Estimators of this quantity are actually the main risk indicator which is currently used in international committee (see [2]. Estimating precisely its value and giving confidence intervals is thus of prime importance.

The most realistic method actually seems the one based on fully nonparametric Monte Carlo simulations sometimes called bootstrap method (although it is not really bootstrap). It consists in independently randomly drawing a large number B of consumption vectors and contamination values in order to obtain B exposure values to get an empirical distribution of exposure. Then, an easy way to evaluate the probability of interest is to consider the frequency of simulations exceeding the PTWI among the simulated data. The purpose of this paper is to validate such a method and give some asymptotically correct methods to construct confidence intervals (CI). These CI are useful to statistically compare populations or to measure the impact of the introduction of a maximum limit (ML) on a particular product. Technical results are detailed in [1].

One should notice that the ideas developed here may also be useful in toxicology, environmental research or in other fields, when there are several sources of pollution, with rates that may also be random. However to better fix the main ideas, we decided to keep the framework of food contamination.

2 Estimating the probability of the exposure to exceed the PTWI

As explained in our introduction, food risk due to a contaminant will be evaluated by estimating the probability of exposure to exceed a fixed deterministic level d. To estimate this probability, two types of data are available if P food items are assumed to be contaminated:

- Contamination data: $q^p_{j_p}$ is the contamination value obtained for the j^{th}_p analysis of the food item p with $j_p = 1 \ldots L(p)$; We assume that the $(q^p_{j_p})_{j_p=1\ldots L(p)}$ are i.i.d. realizations of a random variable Q^p with probability distribution $\mathcal{Q}_p$, $p = 1, \ldots, P$.
- Normalized consumption data (also called individual contaminated baskets): $c^i = \left(c^i_1, \ldots, c^i_p, \ldots, c^i_P\right)$ is the vector of consumptions of individual i observed during a week, standardized by the respective individual weights for $i = 1, \ldots, n$; we assume that these are i.i.d. realizations of a multidimensional r.v. $C = (C_1, \ldots, C_P)$ with probability distribution $\mathcal{C}$.

All consumers are supposed to be independent and the consumption and contaminated data are assumed to be independent. Moreover, contamination observations for the P food items are generally independent. These assumptions are quite reasonable and correspond to what we practically observe in our data.

Let $\mathcal{D} = \mathcal{C} \times \prod_{p=1}^{P} \mathcal{Q}_p$ denote the joint probability distribution of the consumption and the contamination r.v.'s. The individual exposure $D = \sum_{p=1}^{P} Q^p C_p$ has a distribution entirely characterized by $\mathcal{D}$. In this framework, our parameter of interest is a functional of $\mathcal{D}$ defined by

$$\theta_d(\mathcal{D}) = \mathbb{P}_{\mathcal{D}}(D > d) = \mathbb{P}_{\mathcal{D}}\left(\sum_{p=1}^{P} Q^p C_p > d\right).$$

Let $\widehat{\mathcal{C}}_n$ and $\widehat{\mathcal{Q}}_{p,L(p)}$ $p = 1, \ldots, P$ be the empirical probability distribution functions based on our data. The empirical distribution of $\mathcal{D}$ is given by $\mathcal{D}_n = \widehat{\mathcal{C}}_n \times \prod_{p=1}^{P} \widehat{\mathcal{Q}}_{p,L(p)}$.

The natural plug-in estimator of $\theta(\mathcal{D})$ is given by:

$$\theta_d(\mathcal{D}_n) = \mathbb{P}_{\mathcal{D}_n}\left(\sum_{p=1}^{P} Q^p C_p > d\right) = \frac{1}{\Lambda} \sum_{i=1}^{n} \sum_{j_1=1}^{L(1)} \cdots \sum_{j_P=1}^{L(P)} \mathbb{1}\left\{\sum_{p=1}^{P} q_{j_p}^p c_p^i > d\right\}.$$

where $\Lambda = n \times \prod\limits_{p=1}^{P} L(p)$.

Intuitively, $\theta_d(\mathcal{D}_n)$ is the percentage of exceedings of d calculated over all possible combinations of consumption vectors and contamination values drawn with replacement. It is thus an unbiased estimator of $\theta_d(\mathcal{D})$.
The quantity $\theta_d(\mathcal{D}_n)$ may thus be seen as a generalized U-statistic of degrees $k_0 = 1, k_1 = 1, \ldots, k_P = 1$, with kernel $\psi(c^i, q^1, \ldots, q^P) = \mathbb{1}\left\{\sum_{p=1}^{P} q^p c_p^i > d\right\}$, where $c^i = (c_p^i)_{p=1,\ldots,P} \in \mathbb{R}^P$ see definition in [5]

Results on the asymptotic behavior of generalized U-statistics presented in Lee (p. 141), can be generalized under the assumption that the sample sizes in each independent samples are typically of the same order. In our framework, this is certainly not the case: in particular, consumption survey are generally based on large population whereas analytical data are generally obtained thanks to a smaller number of experiences.

Using the Hoeffding decomposition of the generalised U-Statistic, the two following versions of the asymptotic behavior of $\theta_d(\mathcal{D}_n)$ are obtained, depending on assumptions concerning the size of the samples.

Theorem 2.1. *(V1) If* $N = n + \sum_{p=1}^{P} L(p)$, $\frac{n}{N} \to \eta > 0$, $\frac{L(p)}{N} \to \beta_p > 0$ *for* $p = 1, \ldots, P$, *and if at least one of the sample from the* $\mathcal{Q}_p$, $p = 1, \ldots, P$ *or from* $\mathcal{C}$ *has a non zero variance then* $N^{1/2}(\theta_d(\mathcal{D}_n) - \theta_d(\mathcal{D})) \underset{N\to\infty}{\longrightarrow} \mathcal{N}(0, S^2)$ *,where* $S^2 = \frac{1}{\eta}\mathbb{V}_{\mathcal{C}} + \sum_{p=1}^{P} \frac{1}{\beta_p}\mathbb{V}_{\mathcal{Q}_p}$.

These $\mathbb{V}_{\mathcal{C}}$ *and* $\mathbb{V}_{\mathcal{Q}_p}$, $p = 1, \ldots, P$, *typically are the variance of the gradients of the generalized U-Statistics.*

(V2) If $N^* = \min\limits_{p=1,\ldots,P}\left\{L(p),\ \text{such that } 0 < \mathbb{V}_{\mathcal{Q}_p} < \infty\right\}$, $\frac{L(j)}{N^*} \to \beta_j^* > 1$ *and* $\frac{N^*}{n} \to 0$ *then* $N^{*1/2}(\theta_d(\mathcal{D}_n) - \theta_d(\mathcal{D})) \underset{N\to\infty}{\longrightarrow} \mathcal{N}(0, S^{*2})$, *where* $S^{*2} = \sum_{j=1}^{P} \frac{1}{\beta_j^*}\mathbb{V}_{\mathcal{Q}_j}$.

Complete proof of this theorem and details about the computation of the gradients are available in [1].

3 Approximating the estimator by incomplete U-statistics

From a practical point of view, it is generally not possible to construct the generalized U-statistic $\theta_d(\mathcal{D}_n)$, since it is the average of $\Lambda = n \times \prod_{p=1}^{P} L(p)$ terms. We rather use incomplete U-statistic defined by:

$$\theta_{d,B}(\mathcal{D}_n) = B^{-1} \sum_{(i,j_1,\ldots,j_p)\in\mathcal{L}_B} \mathbb{1}\left\{\sum_{p=1}^{P} q^p_{j_p} c^i_p > d\right\},$$

where $\mathcal{L}_B$ is a subset of $\{1,\ldots,n\} \times \{1,\ldots,L(1)\} \times \cdots \times \{1,\ldots,L(P)\}$ of size B much smaller than Λ.

More precisely, $\mathcal{L}_B$ is defined as a random subset of cardinality $\#\mathcal{L}_B = B$ selected with replacement, that is:

$$\mathcal{L}_B = \left\{ \begin{array}{c} \left(i,j^i_1,\ldots,j^i_P\right) \in \left\{1,\ldots,n\right\} \times \left\{1,\ldots,L(1)\right\} \times\cdots\times \left\{1,\ldots,L(P)\right\}, \\ \left\{ \begin{array}{c} i \text{ randomly chosen in } \left\{1,\ldots,n\right\}, \\ j^i_1 \text{ randomly chosen in } \left\{1,\ldots,L(1)\right\}, \\ \vdots \\ j^i_P \text{ randomly chosen in } \left\{,\ldots,L(P)\right\} \end{array} \right\} \text{such that } \#\mathcal{L}_B{=}B \end{array} \right\}.$$

Intuitively, it consists in drawing (with replacement) independent samples of consumption vectors and contamination values in order to obtain B exposure values. $\theta_{d,B}(\mathcal{D}_n)$ is the percentage of values exceeding d among the B corresponding calculated values.

This technique damages the variance of the estimator. However, if B is large enough, the induced distortion is negligible compared to the initial estimator. Indeed, it can be shown using arguments similar to [5, page 193], that $\mathbb{V}(\theta_{d,B}(\mathcal{D}_n)) = O\left(\frac{1}{B}\right) + \left(1 - \frac{1}{B}\right)\mathbb{V}(\theta_d(\mathcal{D}_n))$.

The asymptotic behavior of the incomplete U-statistic $\theta_{d,B}(\mathcal{D}_n)$ depends on the asymptotic behavior of the associated complete U-statistic $\theta_d(\mathcal{D}_n)$ according to the chosen hypotheses (see V1 and V2 of the asymptotic behavior of $\theta_d(\mathcal{D}_n)$). The larger B is, the nearer the two asymptotic distributions are, as shown in [1] (Theorem 3.1).

For the construction of CI, estimators of the asymptotic variances are needed. However the plug-in estimators of $\mathbb{V}_{\mathcal{C}}$ and $\mathbb{V}_{\mathcal{Q}_p}$, $p = 1,\ldots,P$ are not easily computable, since they are also defined as a sum of approximately Λ terms.

The estimation of the variance of U-statistics is generally based on jackknife or bootstrap techniques [5]. These methods are described for unidimensional U-statistics and unidimensional incomplete U-statistics in the case of random selection with replacement. For generalized U-statistics, the use of the jackknife method can not be easily transposed to the multidimensional case. Indeed, in that case, several definitions for the ”leave one out” may be possible (coordinate by coordinate or vector by vector). However this method can be used to estimate the variance of each term in the Hoeffding decomposition, that is equivalent to estimating each $\mathbb{V}_{\mathcal{C}}$ and $\mathbb{V}_{\mathcal{Q}_p}$, $p = 1, \ldots, P$.

We first propose to use a simple bootstrap estimator of the variance which allows to construct asymptotic CI as well as basic percentile CI [4]. These CI are denoted "Asymptotic" and "Percentile"; M is the number of boostrap resamplings in the application and $\overline{\theta_{B,M}}$ is the mean risk estimator over the M resamples.

Then we develop an approximate jackknife variance estimator for S^2 (V1) and S^{2*} (V2) that will serve as a basis for bootstrapping an asymptotically pivotal standardized U-statistics: these t-percentile methods, denoted "t-percentile (V1)" and "t-percentile (V2)" in the application, enjoy much more interesting second order properties than the first one. Much more details about these estimations are available in [1].

4 Application: exposure to OTA

As explained in the introduction, this method was developed to quantify precisely the risk related to OTA exposure. In this application, we particularly focus on the feasibility of the method and compare all the proposed CI. We also use this method to compare the exposure of different sub-populations and to test the impact of a new maximum limit ML on a specific food item. We answer a particular current issue, whether or not new maximum limits on OTA in wine have an impact on the exposure to OTA in France.

In this study we use as consumption data, the INCA survey on individual consumptions of 3003 French consumers (see for details [3]. The contamination analyses have been collected from different French institutions (INRA, DGAL, DGCCRF and ONIVINS for wine).

These analyses are strongly left censored because of the limit of detection (LoD) and/or quantification of the laboratories. To avoid this problem, we apply here the generally used treatment that consists in repeating the evaluation under three different specifications: the censored values are replaced by the LoD (case 1), by the LoD divided by two (case 2) or by zero (case 3). We are currently developing a model using the Kaplan-Meier estimator of the cdf to avoid these simplifications which have a great impact on the final risk level evaluation, as we shall see later.

Our parameter of interest is here defined as the probability for the exposure to exceed the $PTWI$, which, in Europe, is equal to 35 ng/kgbw/wk.

First, we give a few indications on the size of our data set. We consider $P = 9$ food item groups: *Wine*, *Pork and poultry meat*, *Cereal-based products*, *Cereals*, *Coffee, Fruit and vegetable products*, *Dry fruits and vegetables*, *Rice and semolina*, *Beer*. We can build up to $\Lambda = n \times \prod_{p=1}^{9} L(p) \simeq 4 \times 10^{21}$ different exposure values. It explains why we need to use incomplete U-statistics. The convergence rates of (V1) and (V2) depend on $N = n + \sum_{p=1}^{9} L(p) = 3003 + 2708 = 5711$ and $N^* = \min_{p=1,\dots,9} \left\{ L(p); 0 < \mathbb{V}_{\mathcal{Q}_p} < \infty \right\} = 43$, which is the smallest number of analyses realized for the category *"Rice and Semolina"*.

The results are given for different values of the following tuning parameters: B the size of the simulated distributions of the exposure, M the number of bootstrap resamples, B_C and the B_{Q_j} the subsampling size used in the jackknife variance approximation, see [1] for details.

Table 1 gives the 95%−CI for different values of M, B, B_C and the B_{Q_j}.

Comparing the applications of the two versions of the asymptotic behavior of $\theta_d(\mathcal{D}_n)$, we observe that, even though the standard error from (V2) is slightly lower than the one corresponding to (V1), both methods lead to very similar CI. In order to balance the computation times and the accuracy of the results, the parameter values can be chosen as follows: $B = 6000$, $M = 200$ and $B_C = B_{Q_j} = 500$,for all j. Reading Table 1 horizontally, we observe that the CI are very close to each other, so that there is (a posteriori) no real need to use the improved t-percentile method. The asymptotic and bootstrap percentile CI give similar results. In the following, we will use the "Percentile" method.

Parameters			Risk	95%-CI							
B	M	B_C, B_{Q_j}	$\overline{\theta}_{B,M}$	Percentile		Asymptotic		t-percentile (V1)		t-percentile (V2)	
5000	200	100	36.2	32.9	39.3	32.8	39.6	32.9	40.2	32.9	40.1
6000	200	200	36.1	33.3	39.5	32.7	39.6	32.5	39.0	32.5	38.9
3000	200	300	36.0	32.1	39.7	32.1	40.0	32.4	39.8	32.4	39.7
5000	200	300	36.3	32.7	39.7	32.9	39.6	32.6	39.7	32.5	39.7
6000	200	300	35.9	33.6	39.8	32.9	39.6	32.5	38.7	32.5	38.7
10000	200	300	36.0	32.5	39.2	32.5	39.6	32.7	39.4	32.6	39.4
6000	200	400	36.2	31.8	38.0	32.3	38.7	32.7	39.2	32.6	39.2
4000	200	500	36.0	32.6	39.2	32.2	39.7	32.5	39.9	32.4	39.8
5000	200	500	36.2	32.6	39.4	32.6	39.7	32.9	39.7	32.9	39.6
6000	300	300	36.2	32.6	39.5	32.8	39.5	32.6	39.4	32.5	39.3
5000	400	300	36.2	32.4	39.5	32.7	39.7	32.5	39.8	32.5	39.8
6000	400	300	36.1	31.4	37.9	32.1	38.8	33.0	39.5	33.0	39.5

Table 1: Comparison of the confidence intervals for different values of B, M, B_C and B_{Q_j},$j = 1, \dots, P$;
Contaminant: OTA; $PTWI = 35$ ng/kgbw/wk; Censorship case 1.

Table 2 illustrates the great impact of the censorship treatment, an issue that will be considered in the future. In any case, the risk related to OTA exposure is non negligible. Indeed, even if we use the lower bound given in "Case 3", the probability to exceed the $PTWI$ is between 9.1% and 15.8%.

Table 3 focuses on some particular points.

An important application of our results is that they allow to statistically evaluate the impact of new regulations for instance on the maximum limit of (contaminant) residual allowed on the market. To give some insight on the

Censorship	Risk estimator, $\overline{\theta_{B,M}}$	95%-CI	
Case 1	36.3%	32.9%	40.0%
Case 2	20.1%	15.8%	23.8%
Case 3	12.4%	8.4%	16.2%

Table 2: Comparison of the risk estimators and confidence intervals for the three censorship treatments;
Contaminant: OTA; $PTWI = 35$ ng/kgbw/wk; $B = 6000$, $M = 200$ and $B_C = B_{Q_j} = 500, j = 1, \dots, P$.

Assumption/population	Risk estimator, $\overline{\theta_{B,M}}$	95%-CI	
ML(Vin) = 1 μ g/L	12.2%	8.7%	15.5%
3-10 years old%	18.7%	15.0%	23.8%
over 11 years old	10.4%	7.7%	13.8%
male	13.6%	9.7%	16.5%
female	11.3%	8.2%	14.5%

Table 3: Impact of new ML on wine, comparison of population;
Contaminant: OTA; $PTWI = 35$ ng/kgbw/wk; $B = 5000$, $M = 200$ and $B_C = B_{Q_j} = 300, j = 1, \dots, P$.

importance of the problem, we consider the particular case of wine, for which a new European regulation is under study. At the present time, there is no maximum limit. We briefly investigate the impact of imposing a maximum limit for OTA of 1 μ g/L, which has recently been suggested. First, repeating the same calculation as Case 1 of Table 2 without taking into account the wine analyses that exceed 1 μ g/L allows to measure the impact of the introduction of a new ML on OTA in wine (assuming that all the corresponding wine will be withdrawn from the market). The comparison with Case 1 of Table 2 shows that the impact of such a new norm is negligible. This is clearly explained by the fact that cereal is the main factor of contamination. An exhaustive study of this regulation problem will be given in a forthcoming paper.

Considering Case 1 censorship treatment, we evaluate the risk for different sub-populations: it shows in particular, that, on the one hand, the children are overexposed to OTA compared to older people and, on the other hand, women's risk is lower than men's risk.

5 Conclusion

In this paper, we explore the asymptotic properties of some incomplete generalized U-statistics well suited for risk assessment of the exposure to contaminants, when both contamination data and individual consumptions are

available. We show that the estimator of the probability for the exposure to exceed some safe fixed level is asymptotically gaussian and we derive its asymptotic variance. We propose several methods for estimating the variance and we obtain CI for the exposure using i) a standard bootstrap method (percentile confidence and asymptotic intervals), a jackknife method (for estimating the variance) and ii) a bootstrap after jackknife procedure (to built t-percentile intervals). These theoretical results are applied to risk assessment of the exposure to Ochratoxin A (OTA). Some basic comparisons show that the naive Bootstrap and the percentile method give very good CI for this estimation problem. The main conclusion concerning OTA is that the risk is non negligible in France according to our data. We also show how these results may be used to study the impact of new acceptable limits on certain products. In particular, it is shown that the new regulations on the maximum limits of OTA in wine proposed by the European commission are not sufficient to significantly decrease the risk of exposure. We also point out that the risk of exposure is very high for children. This is clearly explained by the fact that cereals are the main source of contamination for this contaminant.

References

[1] Bertail P., Tressou J. (2003). *Incomplete generalized U-Statistics for food risk assessment.* Technical report. Série des Documents de Travail du CREST (Centre de recherche en Economie et Statistique).

[2] Codex Alimentarius (website). Official standards. `http://www.codexalimentarius.net`.

[3] CREDOC-AFFSA-DGAL (1999). *Enquête INCA (individuelle et nationale sur les consommations alimentaires).* TEC&DOC ed.. Lavoisier, Paris. (Coordinateur : J.L. Volatier).

[4] Efron B. (1979). *Bootstrap methods: another look at the jackknife.* Annals of Statistics **7**, 1 − 26.

[5] Lee A.J. (1990). *U-Statistics: Theory and Practice.* Vol. 110 of Statistics: textbooks and monographs. Marcel Dekker, Inc. New York, USA.

Acknowledgement: This study has benefited from financial supports from INRA and ONIVINS. I would like to thank Patrice Bertail, the Mét@Risk and CORELA teams. All errors remains mine.

Address: J. Tressou, INRA-Mét@Risk, Unité Méthodologies d'analyse des risques alimentaires, INA P-G, 16 rue Claude Bernard, 75005 Paris, France.

E-mail: `jessica.tressou@inapg.inra.fr`

FORECASTING THE LONDON METAL EXCHANGE WITH A DYNAMIC MODEL

Kostas Triantafyllopoulos and Giovanni Montana

Key words: Dynamic models, time series, MCMC, London Metal Exchange, Bayesian forecasting, state space models, Kalman filtering.

COMPSTAT 2004 section: Time series analysis.

Abstract: We propose a Bayesian dynamic linear model for a multivariate time series of aluminium official prices available from the London Metal Exchange. Both the observation and transition variances of the model are assumed unknown, and their estimation is performed by Markov chain Monte Carlo (MCMC) simulation. The statistical analysis shows that the model captures well the volatility of the series and has a good forecasting ability. An outlier detection analysis is also illustrated and provides further insights on this data set.

1 Introduction

Multivariate time series have been successfully applied to international exchange rates data [12], the dynamic linear model (DLM), in particular, provides a sound framework which can be adapted to real data analysis [20]. Over the last 10 years there has been a noticeable interest in modelling the London Metal Exchange (LME) market. Slade [14] and Meyer [8] discuss the problems of pricing and hedging in the non-ferrous metal market. The efficiency of the LME is examined in Sephton and Cochrane [13], Agbeyegbe [1] and Moore and Cullen [9]. The relationship of future and spot prices and how the futures can be used to predict the spot prices are considered in Gilbert [3] and in Heaney [5]. The important subject of volatility for the LME market is discussed by Hall [4] and McKenzie et al. [10]. Panas [11] suggests long memory and chaos analyses to assess the metal prices when nonlinear structure is evident. An excellent review of the London Metal Exchange literature can be found in Watkins and McAleer [19].

In this paper, we propose a dynamic linear model for aluminium official prices. It is our belief that these data are interesting for two reasons: they have strong similarities with international exchange rates data, for which there is an enormous literature, and they are not often discussed in the literature. This article presents a novel application of multivariate dynamic linear models showing that the LME data can be successfully modelled with the aid of modern Markov chain Monte Carlo (MCMC) techniques.

The paper is structured as follows. Section 2 gives a description of the metal market and the data set. Section 3 presents the linear dynamic model and describes the MCMC algorithm. The statistical analysis and comments are reported in Section 4.

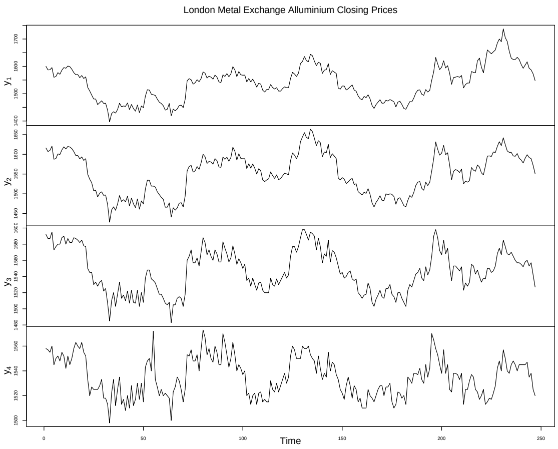


Figure 1: London Metal Exchange aluminium closing prices $\boldsymbol{y}_t$.

2 Description of the data

The London Metal Exchange (LME) is the world's premier non-ferrous metals market, with highly liquid contracts. Its trading customers may be metal industries or individuals (sellers or buyers). LMEX, the London Metal Exchange index, is a base metals index comprising the six primary non-ferrous metals traded on the Exchange: aluminium, copper grade A, standard lead, primary nickel, tin, and zinc. More details about the LME may be found on its web site: `http://www.lme.co.uk`.

In this paper we concentrate on aluminium official prices. We have 4 variables of interest collected in the observation vector $\boldsymbol{y}_t = (y_{1t}, y_{2t}, y_{3t}, y_{4t})'$. Each one of them comprises the official price per tonne of aluminium: y_{1t} is the spot variable, which indicates the daily/current closing price (ask) per tonne of aluminium; the remaining three variables are the relevant future contracts at 3, 15 and 27 months, respectively. The data are collected for every trading day from March 2000 to February 2001, and are plotted in Figure 1. After excluding week-ends and bank holidays, there are $T = 246$ trading days. The data have been kindly provided by Aluminium of Greece S.A.I.C, a member of the Pechiney Group (`http://www.pechiney.com/`).

3 Local level dynamic model and MCMC estimation

In efficient markets (see, for instance, Sephton and Cochrane [13]) a common practice for commodity price short-term forecasting is to take as one-step forecast for time t just the current observed value at time $t-1$ ($t = 2, \ldots, 246$). However, possible shocks can be identified by modelling the difference $\boldsymbol{y}_t - \boldsymbol{y}_{t-1}$ instead of the actual series $\boldsymbol{y}_t$, and we follow this approach here. We

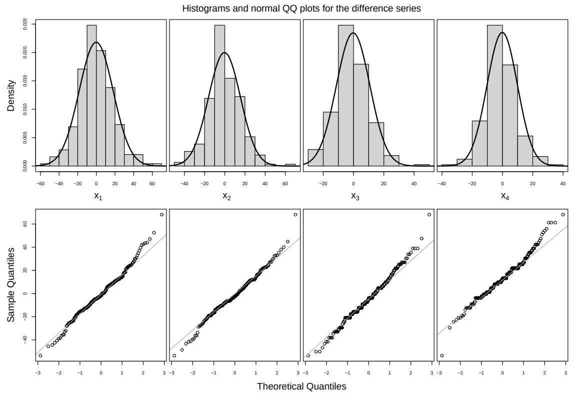


Figure 2: Histogram and normal QQ plots for the difference series $\boldsymbol{x}_t$.

define the difference series $\boldsymbol{x}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$ for $2 \leq t \leq 246$. Given $\boldsymbol{y}_{t-1}$, the difference $\boldsymbol{x}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$ is expected to have zero mean ($\boldsymbol{y}_t$ is predicted as $\boldsymbol{y}_{t-1}$) and some unknown variance. Histograms and normal probability plots for the 4−dimensional multivariate series $\{\boldsymbol{x}_t\}_{t=2}^{246}$ are shown in Figure 2. Such summary plots validate the assumption of a normal distribution for the first difference. Accordingly, we assume that given a state $\boldsymbol{\theta}_t$ and a variance matrix $\boldsymbol{\Sigma}$, we have

$$\boldsymbol{x}_t | \boldsymbol{\theta}_t, \boldsymbol{\Sigma} \sim N_4(\boldsymbol{\theta}_t, \boldsymbol{\Sigma}),$$

where $N_4(\cdot)$ denotes the 4-dimensional multivariate normal distribution.

The level of the series $\boldsymbol{\theta}_t$ may be considered constant, in which case we have a static model with $\boldsymbol{\theta}_1 = \boldsymbol{\theta}$ and $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$, for $t \geq 2$. It seems more reasonable to assume that $\boldsymbol{\theta}_t$ is *expected* to be static, but there is some uncertainty associated with it. This reasoning leads to the adaptation of a local level dynamic model defined by

$$\boldsymbol{x}_t = \boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t | \boldsymbol{\Sigma} \sim N_4(\mathbf{0}, \boldsymbol{\Sigma}) \qquad \text{(observation equation)} \tag{1a}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t | \boldsymbol{\Omega} \sim N_4(\mathbf{0}, \boldsymbol{\Omega}) \qquad \text{(transition equation)} \tag{1b}$$

where the 4 × 1 random vectors $\boldsymbol{\nu}_t$ and $\boldsymbol{\omega}_t$ are mutually and internally independent. The level of the series is the unobserved signal $\boldsymbol{\theta}_t$ and the 4 × 4 observation and transition variance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are assumed unknown, but fixed. The information set is $\boldsymbol{x}^t = \{\boldsymbol{x}_2, \dots, \boldsymbol{x}_t\}$, for $2 \leq t \leq 246$. Full details and applications of the local level model can be found in Durbin and Koopman [2, ch. 3] and West and Harrison [20, ch. 2].

Conditional on known $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$, and given a prior normal distribution for $\boldsymbol{\theta}_1$, the Kalman filter [2, p. 65] could be used to calculate the one-step forecast distribution of $\boldsymbol{x}_t|\boldsymbol{x}^{t-1}$ and the posterior distribution of $\boldsymbol{\theta}_t|\boldsymbol{x}^t$. In this paper both the variance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ are unknown, and we need to resort to Markov chain Monte Carlo simulation to estimate them. West and Harrison [20, ch. 16] state that multivariate dynamic models with unknown variances lack conjugate analyses apart from a handful of special cases; Triantafyllopoulos and Pikoulas [15] and Triantafyllopoulos and Montana [16] give a detailed discussion of this problem from both analytic and computational viewpoints.

We place the following prior densities on $\boldsymbol{\theta}_1$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$:

$$\boldsymbol{\theta}_1 \sim N_4\{(1,1,1,1)', \boldsymbol{I}_4\},$$
$$\boldsymbol{\Sigma} \sim IW_4(n+8, n\widehat{\boldsymbol{S}}), \qquad \boldsymbol{\Omega} \sim IW_4(n+8, n\widehat{\boldsymbol{W}}),$$

where $\boldsymbol{I}_4$ is the 4×4 identity matrix, $IW_4(\cdot)$ indicates inverted Wishart distributions with $n+8$ degrees of freedom and parameter matrices $\widehat{\boldsymbol{S}}$ and $\widehat{\boldsymbol{W}}$. The inverted Wishart densities are given by

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(n+8)/2} \exp\{-n\text{trace}(\widehat{\boldsymbol{S}}\boldsymbol{\Sigma}^{-1})/2\}, \tag{2a}$$
$$p(\boldsymbol{\Omega}) \propto |\boldsymbol{\Omega}|^{-(n+8)/2} \exp\{-n\text{trace}(\widehat{\boldsymbol{W}}\boldsymbol{\Omega}^{-1})/2\}, \tag{2b}$$

and they are the distributions routinely adopted as plausible priors for the observation and transition variance matrices (see e.g. West and Harrison [20, p. 568]).

We have implemented a *forward filtering, backward sampling* algorithm in the C++ language[1] which can be used for the estimation of general multivariate dynamic linear models with fixed but unspecified variances. The algorithm involves iterative sampling. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_{246})'$ and $\boldsymbol{\theta}_{-t}$ consist of all $\boldsymbol{\theta}_j$, excluding the current $\boldsymbol{\theta}_t$, i.e. $j = 2, \dots, t-1, t+1, \dots, 246$. Initially, the state vectors are updated by sampling from the full conditional posterior of each $\boldsymbol{\theta}_t|\boldsymbol{\theta}_{-t}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}, \boldsymbol{x}^{246}$, for all $t = 2, \dots, 246$ (the forward filtering part), which are all multivariate normals; then, the variances are updated by sampling from the full conditional of $\boldsymbol{\Sigma}|\boldsymbol{\Theta}, \boldsymbol{x}^{246}$ and $\boldsymbol{\Omega}|\boldsymbol{\Theta}, \boldsymbol{x}^{246}$ (the backward sampling part), which under the priors (2) are inverted Wishart distributions. This specific implementation allows us to avoid convergence problems which are typical of highly-correlated dynamical systems; discussions of such issues can be found in West and Harrison [20, ch. 15] and Gamerman (1997). The details of this algorithm are found in Triantafyllopoulos and Montana [16].

Much discussion is available in the literature for prior specification and initial values setting from a Bayesian standpoint (see e.g. Leonard and Hsu [7, ch. 6]. In our simulation we have set $n = 1$ and sampled initial values of the observation and transition variances from their prior distributions with initial parameters $\widehat{\boldsymbol{S}} = \boldsymbol{I}_4$ and $\widehat{\boldsymbol{W}} = 2\boldsymbol{I}_4$, respectively. We also notice that n should

[1] The C++ code for MCMC estimation used in this paper has originally been written by G. Montana for general multivariate time series applications including genomics.

be chosen not to be very close to 0 so as to avoid that the corresponding Wishart distributions of $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Omega}^{-1}$ become singular. A burn-in period of 10000 iterations was considered sufficient to reach convergence, and we used the same number of iterations for the Monte Carlo approximation to the posterior density of the variances. The computation took approximately 22 minutes on a Pentium 4 2.80GHz.

4 Statistical analysis

Fitting model (1) provides one-step ahead forecasts, as well as estimates for the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$. Figure 3 shows the one-step forecasts against the actual values; the full lines refer to the original data whereas the darker dashed lines refer to the forecasted values. Overall, the system captures very well the features of the data with accurate forecasts over the four series.

Let $\boldsymbol{S}$ and $\boldsymbol{W}$ be the estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$, respectively. The estimated variances are

$$\boldsymbol{S} = \{s_{ij}\}_{i,j=1,2,3,4} = \begin{bmatrix} 51.314 & 27.419 & \text{-}6.150 & \text{-}33.356 \\ 27.419 & 27.615 & 0.945 & \text{-}23.068 \\ \text{-}6.150 & 0.945 & 13.108 & 4.846 \\ \text{-}33.356 & \text{-}23.068 & 4.846 & 28.244 \end{bmatrix}$$

and

$$\boldsymbol{W} = \{w_{ij}\}_{i,j=1,2,3,4} = \begin{bmatrix} 574.752 & 490.489 & 355.324 & 290.497 \\ 490.489 & 419.483 & 303.603 & 248.283 \\ 355.324 & 303.603 & 220.373 & 180.008 \\ 290.497 & 248.283 & 180.008 & 147.637 \end{bmatrix}$$

The large estimates of $\boldsymbol{\Omega}$ allow us to capture the quite volatile series $\boldsymbol{x}_t$ and provide good forecasts, as in Figure 3. Had the elements of $\boldsymbol{W}$ been smaller, we would have smoother forecasts which would not be able to capture the shocks. Note that Gamerman (1997, pp. 147-150) suggests that the elements of $\boldsymbol{W}$ should be small compared with those of $\boldsymbol{S}$. This is usually desirable when smooth forecasts are on demand. However, when the modeller needs to model time series with volatile behaviour, as in our data, then we suggest that $\boldsymbol{W}$ should have large values like in our results. In such a case the posterior variance, $\text{var}(\boldsymbol{\theta}_t|\boldsymbol{y}^t)$, will be very large implying a model possibly not very stable for estimating the signal $\boldsymbol{\theta}_t$. However, usually the modeller will be able to control the forecast variance by allowing small values in $\boldsymbol{S}$ and large values in $\boldsymbol{W}$. We note that the initial variance setting $\widehat{\boldsymbol{S}} = \boldsymbol{I}_4$ and $\widehat{\boldsymbol{W}} = 2\boldsymbol{I}_4$ follows an uninformative prior setting so that the estimates $\boldsymbol{S}$ and $\boldsymbol{W}$ are determined only from the data. Estimation of the posterior densities of the signal $\boldsymbol{\theta}_t$ is also obtained via our MCMC algorithm. Due to space limitations these results are not shown here.

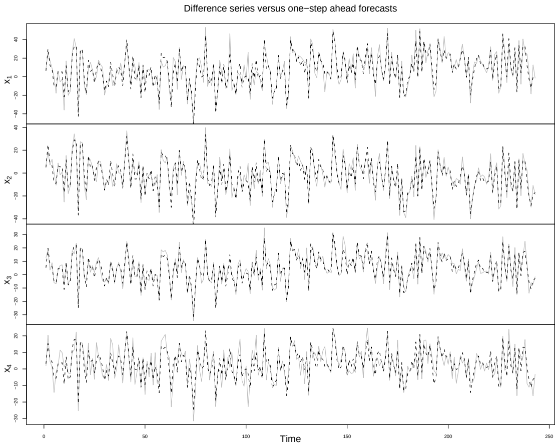


Figure 3: Plot of time series $\boldsymbol{x}_t$ versus one-step forecasts.

Next we performed a simple outlier analysis. With the observations $\boldsymbol{x}_t = (x_{1t}, x_{2t}, x_{3t}, x_{4t})'$, the one-step forecast mean $\boldsymbol{f}_t = E(\boldsymbol{x}_t|\boldsymbol{x}^{t-1}) = (f_{1t}, f_{2t}, f_{3t}, f_{4t})'$, the one-step forecast variance $\boldsymbol{Q}_t = \text{var}(\boldsymbol{x}_t|\boldsymbol{x}^{t-1}) = (q_{ij,t})$, $(i, j = 1, 2, 3, 4)$ and conditioning upon $\boldsymbol{S}$ and $\boldsymbol{W}$, we obtain the marginal standardized errors

$$z_{it} = \frac{x_{it} - f_{it}}{\sqrt{q_{ii,t}}} \qquad \text{for } t = 2, \dots, 246 \text{ and } i = 1, 2, 3, 4$$

which approximately have a standard normal distribution $N_1(0, 1)$. Figure 4 shows the standardized errors z_{it} plotted against ± 1.96 confidence intervals for capturing the marginal outliers. The round points indicate possible outliers and warning should be issued for these points.

Some comments are in order. First we note that there are 15 outliers for x_{1t} which correspond to 6.1% of all 245 observations. The analogous proportions for series x_{2t}, x_{3t} and x_{4t} reduce to $5.7\%, 4.9\%$ and 4.5% respectively. These are rather moderate figures for such kind of data. Warning should be issued for all outliers, although we must note that, for example, 6 outliers for the series x_{1t}, 5 outliers for x_{2t}, 7 outliers for x_{3t} and 3 outliers for x_{4t} are marginal. Decisions should be made in accordance to the requirements of each company and senior consultants should be involved. For example some outliers should be removed while others should not. Following Tsay et al. [18] and Tsay [17, pp. 413-418] outliers of one series may affect outliers of the other series and therefore a more detailed analysis could be required; the direct multivariate approach of Tsay et al. [18] would be a possibility.

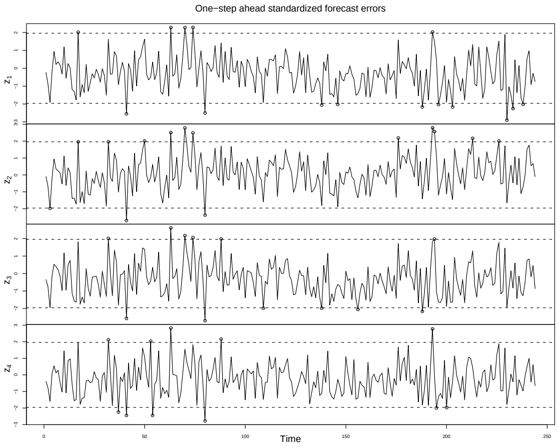


Figure 4: Standardized one-step forecast errors and indication of outliers.

References

[1] Agbeyegbe T.D. (1992). *Common stochastic trends: evidence from the London Metal Exchange*. Bulletin of Economic Research **44**, 141 – 151.

[2] Durbin J., Koopman S.J. (2001). *Time series analysis by state-space methods*. Oxford University Press, Oxford.

[3] Gilbert C.L. (1997). *Manipulation of metals futures: lessons from Sumitomo*. Centre for Economic Policy Research **26**, Discussion Paper: 1537.

[4] Hall S.G. (1991). *An Application of the stochastic GARCH-in-mean model to risk premia in the London Metal Exchange*. Manchester School of Economic and Social Studies (Supplement) **59**, 57 – 71.

[5] Heaney R. (2002). *Does knowledge of the cost model improve commodity futures price forecasting ability? A case study using the London Metal Exchange lead contract*. International Journal of Forecasting **18**, 45 – 65.

[6] Laulajainen R. (1995). *The geographical reach of a commodity exchange - the London Metal Exchange and beyond*. Resource Policy **21**, 133 – 141.

[7] Leonard T., Hsu S.J. (1999). *Bayesian methods*. Cambridge University Press, Cambridge.

[8] Meyer T.O. (1994). *The difficulty in cross-hedging London Metal Exchange spot-price risk using U.S. metal and British pound futures*. Journal of Multinational Financial Management **4**, 141 – 153.

[9] Moore M.J., Cullen U. (1995). *Speculative efficiency on the London Metal Exchange*. Manchester School of Economic and Social Studies **63**, 235 – 256.

[10] McKenzie M., Michell H., Brooks R.D., Faff R.W. (2001) *Power ARCH modelling of commodity futures data on the London Metal Exchange.* European Journal of Finance **7**, 22 – 38.

[11] Panas E. (2001). *Long memory and chaotic models of prices on the London Metal Exchange.* Resources Policy **27**, 235 – 246.

[12] Quintana J.M., West M. (1987). *An analysis of international exchange rates using multivariate DLMs.* The Statistician **36**, 275 – 281.

[13] Sephton P.S. and Cochrane D.K. (1990). *A note on the efficiency of the London Metal Exchange.* Economics Letters **33**, 341 – 345.

[14] Slade M.E. (1988). *Pricing of metals.* CRS Monograph series **22**, Queens University Centre for Resource Studies, Kingston Ontario.

[15] Triantafyllopoulos K., Pikoulas J. (2002). *Multivariate Bayesian regression applied to the problem of network security.* Journal of Forecasting **21**, 579 – 594.

[16] Triantafyllopoulos K., Montana G. (2003). *Variance estimation for multivariate normal dynamic linear models.* Research Report STA03-07, School of Mathematics and Statistics, University of Newcastle, URL: `http://www.ncl.ac.uk/math/research/publications/statistics/STA03-7.pdf`.

[17] Tsay R.S. (2002). *Analysis of financial time series.* Wiley, New York.

[18] Tsay R.S., Pena D., Pankratz A.E. (2000). *Outliers in multivariate time series.* Biometrika **87**, 789 – 804.

[19] Watkins C., McAleer M. (2004). *Pricing of non-ferrous metals futures on the London Metal Exchange.* Journal of Economic Surveys (to appear).

[20] West M., Harrison P.J. (1997). *Bayesian forecasting and dynamic models.* Springer-Verlag, 2nd edn., New York.

Acknowledgement: We are grateful to S. Ikonomakos and A. Papagergiou from the Commercial Department of Aluminum of Greece S.A.I.C for providing the LME data. We should also like to thank P.J. Harrison for useful suggestions on an earlier draft of the paper. K. Triantafyllopoulos was supported in part by Nuffield Foundation grant NAL/00642/G.

Address: K. Triantafyllopoulos, Lecturer in Statistics, School of Mathematics and Statistics, Merz Court, University of Newcastle, Newcastle upon Tyne, NE1 7RU, U.K.
G. Montana, Research Associate, Department of Human Genetics, University of Chicago, 920 E. 58th Street CLSC 507 Chicago, IL 60637 U.S.A.

E-mail: `kostas.triantafyllopoulos@ncl.ac.uk`,
`gmontana@genetics.bsd.uchicago.edu`

EVALUATING THE CDF OF THE KOLMOGOROV STATISTICS FOR NORMALITY TESTING

Wai Wan Tsang and Jingbo Wang

Key words: Goodness-of-fit tests, normality testing.

COMPSTAT 2004 section: Statistical software.

Abstract: A C program that evaluates the CDF of the Kolmogorov statistic for normality testing with unknown mean and variance is described. The method exploits a relation between the statistic for nomality testing with unknown parameters and that of known parameters. It computes the CDF of the former from that of the latter. The resulting program is accurate to the third digit for $35 \leq n \leq 8000$. A comparison of the critical values computed by the program and those from two formulas in literature is included. The results show that the maximum errors in previous formulas ranged from 2.3% to 1.4% whilst ours are bounded by 0.1%.

1 Introduction

The Kolmogorov-Smirnov tests are the most common goodness-of-fit tests for real samples. Consider an ordered set of samples, $x_1 {\leq} x_2 {\leq} \ldots {\leq} x_n$, with purported distribution $F(x)$, the empirical distribution function is defined as

$$F_n(x) = \begin{cases} 0, & x \ < x_1, \\ k/n, & x_k < x < x_{k+1}, k = 1, 2, \ldots, n-1, \\ 1, & x_n < x. \end{cases}$$

The Kolmogorov statistic, D_n= max$|\ F_n(x) - F(x)\ |$, measures the maximum absolute distance between $F(x)$ and $F_n(x)$ [3]. (Two other less comprehensive statistics, ${D_n}^+$ and ${D_n}^-$, were suggested by Smirnov.) The test results are often determined by looking up tables of critical values of D_n obtained from simulation. Recently we have found a matrix formula of the CDF derived by Durbin [1] that is efficient and computationally stable. We coded the method in C. The resulting program evaluates the CDF with 13-digit accuracy for $2 \leq n \leq 16000$ [5].

The Kolmogorov test is readily applicable for normality testing when the mean and variance are known. In practice, however, these parameters are often estimated from the same samples for the testing. To distinguish the two cases, let us denote the Kolmogorov statistic for the latter as ${D_n}^*$. The distribution of ${D_n}^*$ is different from that of D_n (see Figure 1). In particular, the value of ${D_n}^*$ tends to be noticeably smaller. The computation of the CDF of ${D_n}^*$ is harder than that of D_n. Lilliefors has published critical values of ${D_n}^*$ for $4 \leq n \leq 30$ and that of ${D_n}^*\sqrt{n}$ for $n > 30$ in 1967 [4]. The values

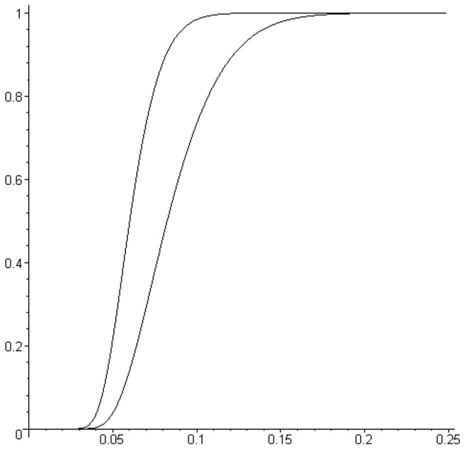


Figure 1: The left curve is the CDF of ${D_{100}}^*$. The right one is of D_{100}.

were obtained using simulations of 1,000 samples. Stephens gave critical values of ${D_n}^*(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$ for $n \geq 20$ obtained using simulations of 10,000 samples in 1974 [7].

Durbin has derived a method for computing the critical values of the Kolmogorov-Smirnov statistics when parameters are estimated in 1975 [2]. The method requires inversion of the Fourier transform of a density. Difficulties with convergence of the Fourier series were encountered when applying the method for the tests of exponentiality, where the Fourier transforms were 1-d functions. The convergency problem is expected more serious for normality tests as the corresponding Fourier transforms are 2-d functions.

We provide here a C program that computes the CDF of ${D_n}^*$. Let $G(n, d)$ be the CDF of ${D_n}^*$ and $K(n, d)$ be that of D_n. Our method exploits an inherent relation between $G(n, d)$ and $K(n, d)$. The relation is almost independent of n so that we can express $G(n, d) = h(K(n, d)) + R(n, d)$. The function, $h(u)$, and the small residue term, $R(n, d)$, are estimated based on simulation results using 10^8 samples. The resulting C program is simple and efficient. It computes $G(n, d)$ with 3-digit accuracy for $35 \leq n \leq 8000$. Such accuracy and scope of n are sufficient for most applications.

The next section describes how to approximate $h(u)$ with a rational polynomial function. The residue term, $R(n, d)$, is expressed as a product of two functions, one on $K(n, d)$ and the other on n. A complete C program that computes $G(n, d)$ is provided. Section 3 analyzes errors in simulations and approximations. The accuracy of the program is verified using additional simulation results from experiments of 10^9 samples. A comparison between the critical values computed by the program and those published by Lilliefors and Stephens is included. A discussion on future work is given in the last section.

2 Computing the CDF of ${D_n}^*$ from the CDF of D_n

Given a set of testing samples, ${D_n}^*$ and D_n are computed in the same way except that the mean and variance in computing the former are estimated from the samples. It is intuitive that their CDFs are related in a similar way for different n. Figure 2 shows the curves of plotting $G(n, d)$ against $K(n, d)$ for n = 35, 60, 125, 250, 500, 1000 and 2000. $K(n, d)$ is evaluated using our program published in [5] and $G(n, d)$ is estimated using simulations of 10^8 samples. The curves are so similar that they are almost identical. Such similarity inspired us to express $G(n, d)$ as a function of $K(n, d)$ plus a small residue term, i.e., $G(n, d) = h(K(n, d)) + R(n, d)$.

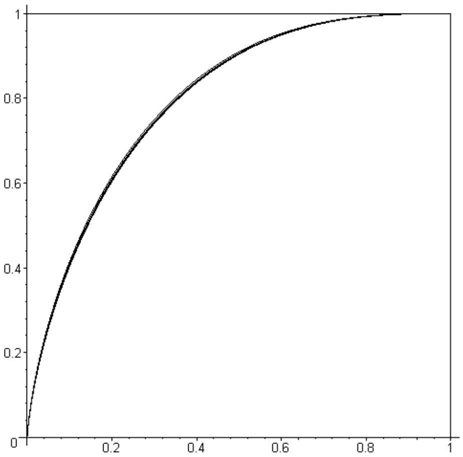


Figure 2: Curves of $G(n, d)$ versus $K(n, d)$ for n = 35, 60, 125, 250, 500, 1000 and 2000.

The function $h(u)$, that is independent of n, can be any function that approximates a curve in Figure 2. Subject to minimizing the maximum absolute errors from the curve of $n = 2000$, we found a rational polynomial function for $h(u)$ with absolute errors bounded by 0.00011. Note that $h(0) = 0$, $h(1) = 1$ and the denominator has no roots in [0,1].

$$h(u) = \frac{u(1.98094u^4 - 4.633u^3 + 3.6683u^2 + 0.3888u + 0.00206)}{u^5 - 1.6746u^4 + 0.7585u^3 + 1.257u^2 + 0.06599u + 0.00021}, 0 \le u \le 1.$$

The difference between $h(u)$ and the curve of $G(2000, d)$ versus $K(2000, d)$ is shown in Figure 3.

The residue term $R(n, d) = G(n, d) - h(K(n, d))$. Figure 4a shows the residue against $K(n, d)$ for various n's. As $h(u)$ is estimated from the curve of $n = 2000$ in Figure 2, the highest residue curve belongs to $n = 35$, the other end of the spectrum. The second highest belongs to $n = 60$, and so on and so forth. Despite the heights, these curves actually look very much alike.

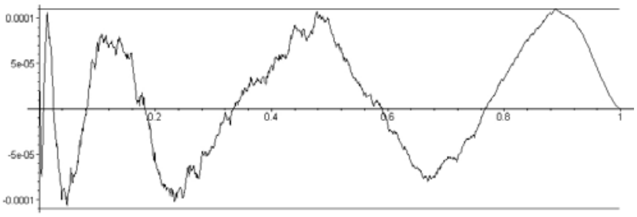


Figure 3: Errors between $h(u)$ and the curve of $G(2000, d)$ versus $K(2000, d)$.

Figure 4b shows the curves after scaling up the lower ones to the height of the highest. The humps in the curves are magnification of the errors in the approximation of $h(u)$.

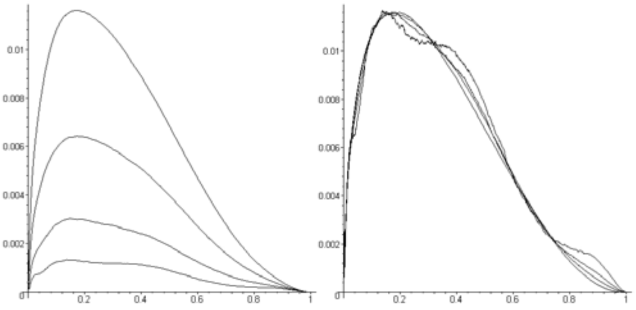


Figure 4: (a) The residue curves for n = 35, 60, 125, 250. (b) The residue curves after scaling.

The similarity in the shapes of the residue curves enables a simple way to compute them. Let $R(n, d) = r(K(n, d)) \times s(n)$, where $r(u)$ is a function approximates the residue curve of $n = 35$ and $s(n)$ is a decreasing function that scales down $r(u)$ to match with the smaller residue curves. $r(u)$ is found as a rational function using the same approach in finding $h(u)$. The absolute errors are bounded by 0.00015.

$$r(u) = \frac{u(0.0109u^5 - 0.02407u^4 + 0.014265u^3 + 0.001135u^2 + 3.9942 \times 10^{-5}u + 5.8 \times 10^{-8})}{u^6 - 1.22u^5 + 0.868u^4 - 0.0472u^3 + 5.366 \times 10^{-4}u^2 + 9.26 \times 10^{-5}u + 2.3 \times 10^{-9}}$$

The scaling function, $s(n)$, has to be accurate when n is small, where the largest residues occur. When n is large, the accuracy of $s(n)$ can be loosened. The following simple function serves the purpose.

$$s(n) = \begin{cases} 41n^{-1.04} - 0.0172, & n < 1700, \\ 0, & n \geq 1700. \end{cases}$$

Overall, the CDF of ${D_n}^*$ is computed with the formula $G(n, d) = h(K(n, d)) + r(K(n, d)) \times s(n)$, where $h(u)$, $r(u)$ and $s(n)$ are defined as above.

$K(n, d)$ is computed using the program described in [5]. The C program that computes $G(n, d)$ is shown below.

```
double G(int n, double d) {
double u, h, r, s, g;
u = K(n, d);
h=u*(u*(u*(u*(u*1.98094-4.633)+3.6683)+0.3888)+0.00206)/
  ( u*(u*(u*(u*(u-1.6746)+0.7585)+1.257)+0.06599)+0.00021 );
r=u*(u*(u*(u*(u*(u*0.0109-0.02407)+0.014265)-0.001135)+3.9942e-5)+5.8e-8)
 /( u*(u*(u*(u*(u*(u-1.22)+0.868)-0.0472)+5.366e-4)+9.26e-5)+0.23e-8);
if  (n < 1700) s = 41. / pow(n,1.04) - 0.0172;
else s = 0.;
g = h + r * s;
if (g > 1.) return 1.;
if (g < 0.) return 0.;
return  g;
}
```

3 Accuracy

The main errors in our approach are induced by the empirical distribution functions (EDFs), and by the approximations of the functions, $h(u)$, $r(u)$ and $s(n)$. Rounding errors accumulated in the computations are negligible. Subsection 3.1 shows that the maximum absolute errors (MAEs) in the EDFs are smaller than 0.00019 unless we encountered an event that is more than four standard deviations away from the mean. Subsection 3.2 shows that the overall MAEs, induced by the EDFs and the approximations, are less than 0.0005. This error bound was verified with extensive simulation results. Subsection 3.3 compares the critical values computed using our program with the values computed using the formulas of Lilliefors and Stephens. The results show that their values may be deviated by as much as 2.3% (Lilliefors) and 1.4% (Stevens) while ours are accurate beyond 0.1% . The Lilliefors and Stephens formulas were devised from simulation results. Given the computing power circa late sixties or earlier seventies, their compact formulas are actually quite good.

Our approach works well for large n but we can only verify its accuracy up to $n = 8000$. Therefore, we only claim that our program computes the CDF of ${D_n}^*$ for $35 \leq n \leq 8000$ with 3-digit of accuracy.

3.1 Errors in simulations

In our simulations, uniform random numbers are generated using the 64-bit generator described in [6]. They are converted into normal variates using the Ziggurat method. Suppose that an EDF is obtained using m samples. It is interesting to find out that the MAE in the EDF is indeed the Kolmogorov statistic with parameters complete specified, i.e., D_m. Thus, the MAE has the same distribution as D_m. Using the limiting form of the distribution of D_m given in [3], the mean and standard deviation of MAE are $0.87/\sqrt{m}$ and

$0.26/\sqrt{m}$, respectively. In our simulation experiments, $m = 10^8$. The mean plus four standard deviations is 0.00019. Therefore, it is very safe to state that the MAE in the EDF is less than 0.00019.

We have estimated the errors in the EDF of $G(35, d)$ by comparing its values with the 'true' values obtained using simulation of 10^{10} samples. The MAE is less than 0.00006.

3.2 Overall errors and verifications

Our program uses the formula $G(n, d) = h(K(n, d)) + r(K(n, d)) \times s(n)$ to compute the CDF of ${D_n}^*$. $h(u)$ is obtained from the EDF of $n = 2000$. The errors in it are taken into account in the approximation of $r(u)$. These erratic noises make the approximation of $r(u)$ more difficult.

$r(u)$ was obtained from the EDF of $n = 35$. The MAE of the EDF is unlikely to be larger than 0.00019, established in last subsection. The MAE in the approximation is 0.00015. As $s(n)$ is close to 1 when $n = 35$, the MAE in the formula is less than 0.00034 when $n = 35$.

The errors in the formula for other values of n are harder to assess. On one hand, when n increases, the residue curve shifts farther away from $r(u)$. On the other hand, as $s(n)$ is a decreasing function, the errors are scaled further down when n becomes larger. Simulation experiments show that the downward scaling is more dominating. The error bound of $n = 35$ (0.00034) actually caps the MAEs of $35 \leq n \leq 2000$. To verify, we have checked the values computed by the program with the EDFs obtained using 10^9 samples. The MAEs are 0.00008 ($n = 35$), 0.0002 ($n = 50$), 0.00012 ($n = 60$), 0.00023 ($n = 100$), 0.00016 ($n = 125$), 0.00013 ($n = 200$), 0.00016 ($n = 250$), 0.00019 ($n = 400$), 0.00015 ($n = 500$), 0.00021 ($n = 800$), 0.00012 ($n = 1000$), 0.00025 ($n = 1500$), and 0.00011 ($n = 2000$). All of them are well below 0.00034.

For $n > 2000$, $s(n) = 0$. The accuracy solely depends on the difference between $h(u)$ and the curve of $G(n, d)$ versus $K(n, d)$. Figure 5 shows the differences of the neighboring curves in Figure 2. The highest curve is the difference between the curve of $n = 35$ and that of $n = 60$. The second highest is the difference between those of $n = 60$ and $n = 125$, and so on and so forth. These curves demonstrate that the curves in Figure 2 convert quickly to the asymptotic one. From the trend of the maximum differences, we estimate that the maximum difference between the curve of $n = 2000$ and that of $n = 8000$ is less than 0.0002. As $h(u)$ is estimated from the curve of $n = 2000$ with overall MAE of 0.0003 (0.00019 from simulation and 0.00011 from approximation), the MAE in the values computed by the program for $2000 < n \leq 8000$ is less than 0.0005. This bound is verified for $n = 4000$ (MAE = 0.00032) and $n = 8000$ (MAE = 0.00033), with EDFs obtained using 10^9 and 10^8 samples, respectively.

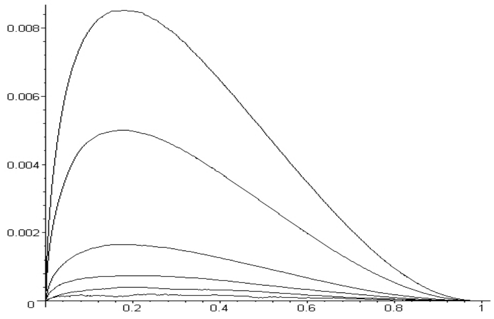


Figure 5: The differences of the neighboring curves in Figure 2.

3.3 Comparisons

To demonstrate the accuracy, we compare the critical values of ${D_n}^*$ computed using the formula of Lilliefors, that of Stephens, our program, and the EDFs obtained from simulations of 10^9 samples. The results are shown in the tables below. The header row specifies the proportions of the critical values. The 'true' proportion of each critical value, computed from the EDF, is shown in parentheses. The results show that the Lilliefors formula deviates by as much as 2.3% in the comparison. The Stephens formula is more accurate. The worst case deviates by 1.4%, whilst our program computes the critical values with 4 accurate digits after the decimal point.

$n = 35$	**0.85**	**0.9**	**0.95**	**0.99**
Lilliefors	0.12982(0.864)	0.13607(0.903)	0.14976(0.956)	0.17427(0.992)
Stephens	0.12810(0.852)	0.13538(0.899)	0.14794(0.951)	0.17108(0.989)
Ours	0.12786(0.850)	0.13557(0.900)	0.14760(0.950)	0.17192(0.990)
Simulation	0.12787	0.13557	0.14761	0.17193

$n = 250$	**0.85**	**0.9**	**0.95**	**0.99**
Lilliefors	0.048573(0.837)	0.050913(0.881)	0.056036(0.944)	0.065206(0.988)
Stephens	0.048880(0.843)	0.051655(0.893)	0.056449(0.947)	0.065278(0.988)
Ours	0.049220(0.850)	0.052163(0.900)	0.056770(0.950)	0.066174(0.990)
Simulation	0.049214	0.052153	0.056765	0.066151

$n = 2000$	**0.85**	**0.9**	**0.95**	**0.99**
Lilliefors	0.017173(0.827)	0.018000(0.873)	0.019812(0.940)	0.023054(0.987)
Stephens	0.017326(0.836)	0.018301(0.887)	0.020009(0.944)	0.023139(0.987)
Ours	0.017566(0.850)	0.018609(0.900)	0.020243(0.950)	0.023581(0.990)
Simulation	0.017565	0.018607	0.020243	0.023578

4 Future work

The program described here computes the CDF of ${D_n}^*$ to the third digit. To achieve higher accuracy, we need to increase the number of samples used in obtaining the EDFs and to find better approximations for $h(u)$, $r(u)$, and

$s(n)$. It is likely that we will achieve 4-digit accuracy if we increase the sample size to 10^{10}, estimate the $h(u)$ from the EDF of $n \geq 4000$, and reduce the MAEs in the approximations of $h(u)$ and $r(u)$ to 0.00003. The high accuracy in approximations can be achieved, in the worst case, using splines. We will, of course, first investigate the possibilities of using other simpler forms of functions.

Another aspect of improvement on the current method is to widen the scope of n, say, for any $n \geq 20$. The extension of the program for the computation for $20 \leq n < 35$ is just laborious work. On the other hand, to extend the program for n beyond 8000 requires computing $K(n, d)$ for $n > 8000$. The approach described in [5] is inefficient for $n > 16000$. We need to develop a new program that computes $K(n, d)$ for very large n.

$h(u)$ is ideally approximated from the curve of $G(n, d)$ versus $K(n, d)$, when n approaches infinite. The formula of the asymptotic $K(n, d)$ is given in [3]. The hurdle is to find a practical way for computing the asymptotic $G(n, d)$.

References

[1] Durbin J. (1968). *The probability that the sample distribution function lies between two parallel straight lines.* Journal of the American Statistical Association **39**, (2), 398 – 411.

[2] Durbin J. (1975). *Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings.* Biometrika **62**, (1), 5 – 22.

[3] Kolmogorov A. (1933). *Sulla determinazione empirica ei una legge di distributione.* Giornale dell' Istituto Italiano degli Attuari **4**, 83 – 91.

[4] Lilliefors H.W. (1967). *On the Kolmogorov-Smirnov test for normality with mean and variance unknown.* Journal of the American Statistical Association **60**, (318), 399 – 402.

[5] Marsaglia G., Tsang W.W., Wang J. (2003). *Evaluating Kolmogorov's distribution.* Journal of Statistical Software **8**, (18), 1 – 4 (Both the paper and the program are available at `http://www.jstatsoft.org/`).

[6] Marsaglia G., Tsang W.W. (2004). *The 64-bit universal RNG.* Letters in Statistics and Probability **6**, (2), 183 – 187.

[7] Stephens M.A. (1974). *EDF statistics for goodness of fit and some comparisons.* Journal of the American Statistical Association **69** (347), 730 – 737.

Acknowledgement: Research supported by the Hong Kong Research Grants Council, Grant HKU 7046/02E.

Address: W.W. Tsang, J. Wang, Department of Computer Science and Information Systems, The University of Hong Kong, Pokfulam Road, Hong Kong

E-mail: {`tsang, jbwang`}`@csis.hku.hk`

MAKING STATISTICAL ANALYSIS EASIER

I. Tsomokos, K.X. Karakostas and V.A. Pappas

Key words: Teaching computational statistics, S-plus, R system.

COMPSTAT 2004 section: Statistical software.

Abstract: Analyzing observed or measured data is an important step in applied sciences. The recent increase in computer capacity has resulted in a revolution both in data collection and data analysis. An increasing number of scientists, researchers and students are venturing into statistical data analysis; hence the need for more guidance in this field is obvious. The user having decided upon the statistical method to be used for his/her data set can then use the present program to get a complete analysis and a final report containing all the appropriate conclusions. The aim of the present program is to facilitate office workers to use the statistical techniques at a minimum effort and time.

1 Introduction

The collection and analysis of statistical data is a part of our every day life. A major contribution to this fact it is mainly due to the tremendous spread of the electronic computers and the associated development of a large number of statistical software. More and more people from various disciplines are entering into the process of collecting and analyzing data.

The first step, after collecting a set of data on a number of variables, is to decide upon the appropriate statistical method (for example hypothesis testing, regression, analysis of variance or covariance e.t.c.) to be used. This is a quite crucial decision which we have to make up. A wrong selection will end up with wrong conclusions. The first two authors have presented a work on this topic in the DSI (Decision Sciences Institute) 5th International conference [3] held at Athens. More on that can be found on (for those who can read Greek) the URL address `http://daedalus.math.uoi.gr/assa/index.`

The second step is to use a statistical package to implement the statistical method that we have decided upon in the previous step and write down the conclusions. This step is also a critical one. Usually the use of most statistical methods depends upon the validity of some assumptions. Even if a part of those assumptions are not satisfied then the use of the specific statistical method may lead us to wrong conclusions. For example the simple t-test for the mean of one population can be used if the following assumptions are satisfied: i) our sample is random and ii) it follows a normal distribution. If even one of these assumptions is not satisfied then the t-test should not be used. Almost all of the well known statistical packages available in the market provide the user with the appropriate techniques to test those assumptions. This means that the user is aware of the restrictions under which the specific statistical method is valid and knows how to check their validity. Moreover the user must be able to select the appropriate transformation for his/hers data set in order to correct, if possible, any violation on those assumptions. For example in the simple t-test, mentioned previously, if the assumption of normality it is not satisfied, then one can use the Box-Cox transform (see e. g. Madansky [1]) in order to correct the problem. Other considerations are also possible, for example using the Central Limit Theorem. However, in this case, the user must be aware that the so obtained p-value is approximate only. From all the above it becomes clear that those who want to implement any one statistical method must know, in a good depth, the corresponding statistical theory. Obviously this it is not the case for an office worker.

The present work aims exactly at the second step having in mind mainly an office worker, a small company (and why not a bigger one) that does not prefer to spend money to analyze the data collected. Those users are going to find the present program quite helpful. But also a statistician (either a theoretical or an applied) can be benefit from the use of this program by getting the analysis in a greatly reduced time compared to that needed to do the same analysis manually.

In the next paragraph we give a brief description of the program together with some examples. In paragraph 3 we discuss what it is known to the authors to be available and is on the same direction with the present program. A final conclusion is presented in the last paragraph.

2 Description of the program

As it was pointed out earlier the aim of the present work is twofold. The first is to do the specified analysis, and the second to provide the user with a report of conclusions. As such our effort was concentrated on implemented the statistical work and give a final report rather than writing down the various necessary statistical routines needed for it. Although a large number of works exist on statistical algorithms (see e.g. Thisted [2]) trying to implement it, it is a bit tricky. For that reason we decided to use well known and established routines. Such ones are the routines provided by the R language. This is almost identical with the S one (on which it is based the well known S-Plus statistical package) and can be obtained free from the following address `http://www.r-project.org/`

The general structure of the program is the following:

(1). *It reads the variables that are to be included in the specified analysis.* For example for the two samples t-test it reads two vectors. One qualitative, containing the values for the two groups (that these values can be numbers or characters) and one quantitative, containing the values for the study variable. There is also the possibility the two samples to be in different columns.

(2). *It reports the assumptions (if any) that need to be satisfied in order to use the specified statistical method.* In the two sample t-test case e. g. those assumptions are a) there are no outliers in the data, b) the two samples are independent, c) each sample follows a normal distribution, and d) the two population variances are equal.

(3). *It checks if those assumptions are satisfied or if the can be satisfied after a transformation.* For the two sample t-test, for example, it checks for outliers. If a big number of outliers exist, say more than 10% of the category's sample size, and then it tries to correct the problem by taking e.g. the logarithm of all observations. In the same situation if the normality assumption is rejected (using e.g. the Shapiro-Wilk test) (see e. g. Madansky [1]) then the program uses the Box-Cox transform to correct it. (Note that before any kind of transformation observations that were excluded as outliers, are now included in the analysis)

(4). *Finally the program gives a report relative to the statistical method used and a final conclusion.* In the case of the two samples t-test the program reports about the process used (i.e. checking for outliers and normality and possible transformations, if they were needed) and gives a conclusion on rejecting or not rejecting the initial hypothesis.

(5). *Various other utilities are available depending on the statistical method used.* For example, some descriptive statistics, the power, the least significance difference, e.t.c.. More on that can be seen in the following examples. The examples are the output of the program in R environment.

(6). In the front end of the program the user can find the "explain" and "why" windows. The first of theme explains the various statistical terms used, and the second one tries to inform the user why the program follows the given procedure.

Example 1 (t-test for one population). This example is about the mean age of the bears in an area. The data set contain a number of outliers and do not follow a normal distribution. However the log transform corrects the problem at 1 % significance level. In this as well in the second example the italics are the output of our program.
Read 1 items (Information on how many variables have been red)
Read 83 items (Information on how many records have been red)

—————————–
| *REPORT* |
—————————–

We would like to examine if the mean of the variable Age equals to 56. Our problem is that of testing if the mean of a population is equal to a specified value. In order to use the t-test for one population we need to check the following.

----

1. - The sample is random.
2. - There exist no outliers in our data.
3. - Our data follow a normal distribution.

The first of the above assumptions, that is the sample is random, can be verified from the way we have selected it.

The outlier box plot showed that there were outliers in our data. More specifically, the observations with number(s), in order of appearance, 34 14 13 12 11 are outliers. Since the total number of outliers does not exceed the 10% of our sample size, which in our case equals to 83, they can be removed from our data. After removing these observations there were no more outliers in our data.

The Shapiro-Wilk test showed that the hypothesis that our data follow a normal distribution, should be rejected (p <0.0001). To correct this problem we made use of the Box-Cox transformation. (This method gives us the ability to transform our data so as their distribution gets closer to the normal one). (Note that in this case observations that were probably discarded as outliers are now included in the sample). The transformation suggested for our data is log(x) (lamda= 0) Transforming our data we find that there are no outliers. Using the Shapiro-Wilk test for the transformed data, we find p= 0.04184099. This means that the hypothesis that our transformed data follow a normal distribution can not be rejected, for significance level 1%. According the above results, we can use the t- test for testing our initial hypothesis. The value of the t statistic equals to - 6.382168, while the corresponding p-value is <0.0001. The hypothesis that the mean value of the variable equals to 56 should be rejected. A 99 % confidential interval for the mean value of the transformed (log(x)) variable is:(3.264659 , 3.709492)
Please make your selection from the table below:

1. Boxplots (Initial/Modified data)
2. Boxplots (Transformed/Modified data)
3. Normal Probability plots
4. Descriptive Statistics (Initial/Modified data)
5. Descriptive Statistics (Transformed/ Modified data)
6. Exit

The selections in the above table give us extra information about our data. So, selections 1 and 2 give, respectively, the outlier box plots and the normal probability plots for the initial and/or the modified data. Selections 3 and 4 give descriptive statistics for the initial and/or modified data. Depending on the type of data used (initial or modified) to do the t-test the appropriate selection shows the power of the test together with graphs for the least significance number (LSN) and least significance difference (LSD).

Example 2 (t-test for two populations, independent samples). In this example we test about the equality of the means of two populations. There are two groups denoted as zero (0) and (1). The second group has two outliers which can be removed. The normality assumption can not be accepted for both groups and the Box-Cox transform does not correct the problem. So finally we make use of a non parametric test
Read 2 items (Information on how many variables have been red)
Read 178 records (Information on how many records have been red)

————————–

| REPORT |

————————–

We would like to check if the mean value of variable COLOR is the same at the two categories of variable CARTOON1 .This is a problem of testing for the equality of the means of two populations. To use the well known t test for two populations the following assumptions must be satisfied.

---

1. - The two samples are independent
2. - There exist no outliers in both samples
3. - Both samples follow a normal distribution
4. - The variances of the two populations are equal

---

The first of the above assumptions, that is the two samples are independent, can be verified from the way we have selected them. For the other three assumptions the analysis goes as follows:

OUTLIERS

The random sample of category 0 has no outliers. The random sample of category 1 has the following outliers: 8 32. Since the number of outliers in category 1 is less than 10% of sample size (which is 89), we can remove those outliers.

NORMALITY ASSUMPTION

The random sample of category 0 doesn't follow a normal distribution: p-value <0.0001. In order to correct the problem we will use the Box-Cox transformation. (Note that in this case observations that were probably discarded as outliers are now included in the sample).

NOTE: All possible no positive values will be removed from our data. The random sample of category 1 doesn't follow a normal distribution: p-value <0.0001. In order to correct the problem we will use the Box-Cox transformation. Note that in this case observations that were probably discarded as outliers are now included in the sample).

NOTE: All possible no positive values will be removed from our data. The Box-Cox transformation did not correct the normal distribution problem in category 0, (after using Box-Cox transformation lamda=2), we get p-value <0.0001.

The Box-Cox transformation did not correct the normal distribution problem in category 1, (after using Box-Cox transformation lamda=2), we get p-value <0.0001.

So we will use a no parametric test in order to test that the means of the two populations are or do not are equal. (Note that in this case observations that were probably discarded as outliers are now included in the sample). Using the Mann- Whitney rank sum test we get that the value of the test statistic is, 4490 and the corresponding p-value equals to 0.1181558. This test is actually testing if the medians of the two populations are equal. In order to use it for testing equality of means the distributions, in the two samples, must be as symmetrical as possible. From that p-value we conclude that the hypothesis that the means of the two populations, are equal can not be rejected for significance level a=0.05.

```
———————————————————————————-
Please select a category:
———————————————————————————-
For category 0 press number 1
For category 1 press number 2
For exit press number 3
———————————————————————————-
```

The selections available in the above table are the same as those in one mean t test, for each category separately.

3 Comparison with other programs

To find similar work we first looked among the leading statistical packages. Among them the packages that include work analog to that presented here are the S.A.S, with the module SAS/Lab, the NCSS, itself and its module PASS, and the Minitab with the StatGuide. A general remark for all the above statistical packages is that what they offer it is program oriented. That means that what ever the suggestions are, for any statistical analysis, they are based on the abilities of the specific package used.

The basic idea in the StatGuide of Minitab is to offer a specific example on the selected statistical method. Based on this example, which has nothing to do with the real data analyzed, the program interprets the output given by the program. Following that interpretation the user could interpret his/hers data.

The PASS and LAB modules of NCSS and SAS, respectively, are both to the right direction with the SAS/LAB to be better than NCSS/PASS. However much of the decision are left to the user (e.g. possible transformations of the original variables). This means that the user should more or less have some knowledge of statistical theory.

To compare each one of them with our work we are going to use two sets of data. The first is a t test for the mean of one population. The second one is the t test for testing equality of the means of two independent populations. The comparison will include the programs NCSS and SAS with ours. The analysis of both data sets is presented in the previous examples using our program. For the first data set SAS/LAB can not be used since the one sample t-test it is not available. The procedure in NCSS/PASS a) does not check for possible outliers, b) it correctly rejects the normality assumption, but does not try to correct it so c) it uses a non parametric test. For the second data set the SAS/LAB procedure does not point out the two outliers in category 1. It identifies the problem with the normality assumption but can not correct it. Hence its suggestion, of no differences between the two groups, it is not correct. The NCSS procedure a) does not mention the existence of any outliers; b) it correctly rejects the normality assumption, but does not try to correct it so c) it uses a non parametric test.

In the internet we located the following address. However, this address is rather on a theoretical base than on a practical one, in the sense of given practical advices for the specific data at hand.

`http://www.graphpad.com/articles/interpret/principles/stat_principles.htm`

4 Conclusions

From the above short comparison on the specific statistical techniques it is clear that proposed program is more complete and more instructive than SAS/LAB and NCSS/PASS. Further more it is friendlier to the user and especially for the non statistician user as it is an office worker.

References

[1] Madansky A. (1980). *Prescriptions for working statisticians.* Springer-Verlag.

[2] Thisted R. A. (1991). *Elements of statistical computing.* Chapman and Hall.

[3] Tsomokos I., Karakostas K. X. (1999). *5th International Conference of Decision Sciences Institute*, Athens, Greece, 4-7 July, 1919 – 1920.

[4] MINITAB and the MINITAB logo are registered trademarks of Minitab Inc.

[5] NCSS and PASS are trade marks of NCSS Statistical Software.

[6] S.A.S and SAS/LAB are trade marks of SAS Institute Inc., Cary NC, USA/

Address: I. Tsomokos, K.X. Karakostas, V.A. Pappas, Department of Mathematics, University of Ioannina

E-mail: `gtsomok@cc.uoi.gr`, `kkarakos@cc.uoi.gr`,
`vasileios_p@yahoo.gr`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

SYMMETRIC NORMAL MIXTURES

Michael Turmon

Key words: Symmetry constraint, algebraic group, EM algorithm.

COMPSTAT 2004 section: Clustering.

Abstract: We consider mixture density estimation under the symmetry constraint $x \overset{\mathcal{D}}{=} Ax$ for an orthogonal matrix A. This distributional constraint implies a corresponding constraint on the mixture parameters. Focusing on the gaussian case, we derive an expectation-maximization (EM) algorithm to enforce the constraint and show results for modeling of image feature vectors.

1 Introduction

We consider a simple constraint which captures underlying symmetry in density estimation problems. In particular, we are interested in cases where the target random variable $x \in R^d$ satisfies

$$x \overset{\mathcal{D}}{=} Ax \tag{1}$$

for a known linear transform A. It is immediate that A is nonsingular: otherwise Ax would concentrate in a proper subspace of R^d, and the law of x would fail to have a density with respect to Lebesgue measure on R^d. Indeed, $|A| = 1$ (writing $|A|$ for absolute value of the determinant) since

$$1 = \int p(x)\, dx = \int p(Ax)\, dx = |A|^{-1} \int p(y)\, dy = |A|^{-1} \ .$$

There are no general restrictions on A through its singular values. For example, consider for any orthonormal U the symmetry $A = U \left[\begin{smallmatrix} 0 & 2 \\ 1/2 & 0 \end{smallmatrix}\right] U^{\mathsf{T}}$. Choosing $x \sim N(0, \Sigma)$ where $\Sigma = U \left[\begin{smallmatrix} 2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right] U^{\mathsf{T}}$ implies $A \Sigma A^{\mathsf{T}} = \Sigma$ so that $x \overset{\mathcal{D}}{=} Ax$.

Iterating (1), noting $|A| \neq 0$, shows $x \overset{\mathcal{D}}{=} A^p x$ for any integer p. For clarity we confine this paper to cyclic symmetries: $A^P = I$ for some period P. The set of symmetries $G = \{I, A, \ldots, A^{P-1}\}$ is then isomorphic to the cyclic group of order P. This can be relaxed in various ways. Some multiple symmetries are encoded by finite groups that are not cyclic. Also, continuous (e.g., scale) or aperiodic (e.g., translation) invariances are important in applications.

Mixture estimation problems for image data having symmetries motivated this work; see the figure on the last page. The upper left plot shows bivariate feature vectors taken from pairs of synchronized solar images from the MDI imager on the SoHO spacecraft. These densities have symmetry with respect to changing the sign of the magnetic flux, corresponding to $A = \left[\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Similar data are gathered by other solar observatories. Taking A as a general rotation matrix can encode a variety of similar geometric constraints.

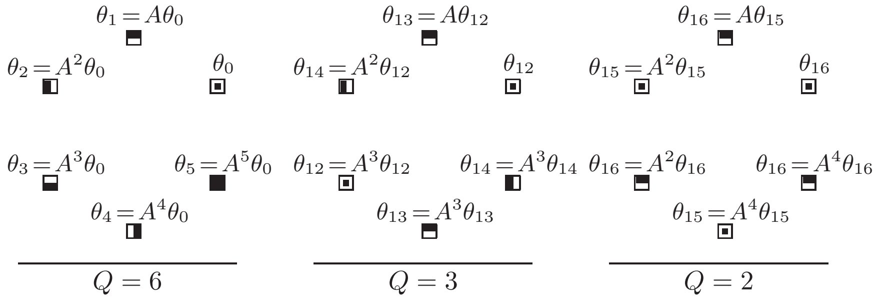

$\theta_1 = A\theta_0$
$\theta_2 = A^2\theta_0$
θ_0
$\theta_3 = A^3\theta_0$
$\theta_5 = A^5\theta_0$
$\theta_4 = A^4\theta_0$
$Q = 6$
$\theta_{13} = A\theta_{12}$
$\theta_{14} = A^2\theta_{12}$
θ_{12}
$\theta_{12} = A^3\theta_{12}$
$\theta_{14} = A^3\theta_{14}$
$\theta_{13} = A^3\theta_{13}$
$Q = 3$
$\theta_{16} = A\theta_{15}$
$\theta_{15} = A^2\theta_{15}$
θ_{16}
$\theta_{16} = A^2\theta_{16}$
$\theta_{16} = A^4\theta_{16}$
$\theta_{15} = A^4\theta_{15}$
$Q = 2$

Figure 1: Schematic rendering of three cycles in a system with $P = 6$. All P versions of the component are shown; the P/Q aliases have the same markers.

Taking A as a permutation enforces within-feature-vector distributional constraints. For complex x, $A = \sqrt{-1}I$ gives real-imaginary symmetry.

As density models for x we use finite normal mixtures [6]:

$$p(x) = \sum_{k=0}^{K-1} \gamma_k N(x;\, \mu_k, \Sigma_k) \tag{2}$$

where $\sum_{k=0}^{K-1} \gamma_j = 1$, the constituent mean vectors μ_k are arbitrary, and the covariance matrices Σ_k are symmetric positive-definite. We require that the (μ_k, Σ_k) be distinct to preserve identifiability. The free parameters

$$\Theta = \{(\gamma_k, \mu_k, \Sigma_k)\}_{k=0}^{K-1} \tag{3}$$

are chosen using training data $X = \{x_n\}_{n=1}^{N}$ and maximum likelihood:

$$\Theta_{\mathrm{ML}} = \arg\max_{\Theta \in \boldsymbol{\Theta}} \log p(X; \Theta) \ . \tag{4}$$

To estimate these parameters, we use the well-known EM (Expectation-Maximization) algorithm, which leaves room to accommodate key physical constraints like (1). Constraints also ameliorate the problem of local maxima — which is especially troublesome in mixture estimation.

The sequel is organized as follows. In the next section we lay out the structure of the parameter constraints implied by the symmetry constraint in the context of normal mixtures, briefly examining related work. We then derive the EM algorithm for the general solution. Implementation issues and some representative results follow this derivation.

2 Constrained mixture parameters

Suppose x is governed by a normal mixture $\Theta = \{(\gamma_k, \mu_k, \Sigma_k)\}_{k=0}^{K-1}$. Then the constraint (1) is satisfied if and only if

$$(\gamma, \mu, \Sigma) \in \Theta \Rightarrow (\gamma, A\mu, A\Sigma A^{\mathsf{T}}) \in \Theta \ . \tag{5}$$

Henceforth, for short: when $\theta = (\gamma, \mu, \Sigma) \in \Theta$, write $A\theta$ for $(\gamma, A\mu, A\Sigma A^\mathsf{T})$.

To see (5) suffices for (1), first note that (5) implies existence of a permutation π of $\{0, \ldots, K-1\}$ mapping mixture components according to A:

$$\pi(k) = \arg \min_{l:\theta_l = A\theta_k} (l-k) \bmod K \; . \tag{6}$$

To see that π is a permutation, note that the set of l satisfying the condition is guaranteed to be nonempty by (5) so π is a well-defined function on $\{0, \ldots, K-1\}$. And, the inverse exists:

$$\pi^{-1}(l) = \arg \min_{k:\theta_l = A\theta_k} (k-l) \bmod K \tag{7}$$

which has the effect of counting down from l, looking for the first matching parameter tuple, while π counts up. Now π and $|A| = 1$ establish (1):

$$p(Ax) = \sum_{k=0}^{K-1} \gamma_k N(Ax;\, \mu_k, \Sigma_k) = \sum_{k=0}^{K-1} \gamma_{\pi(k)} N(x;\, \mu_{\pi(k)}, \Sigma_{\pi(k)}) = p(x) \; .$$

The reverse implication, which is not so important for our purposes, follows from the linear independence of Gaussian functions [8].

The domain of π can be partitioned into cycles, each of the form $\mathcal{C} = (k_1, \ldots, k_Q)$ for some length Q. Cycles are the minimal subsets of the domain which are fixed by the permutation: $\pi(k_i) = k_{i+1}$ and $\pi(k_Q) = k_1$. Listing the cycles of π uniquely determines and succinctly describes its structure. This decomposition will prove key to compactly specifying the form of the mixture to be fit to X, e.g. section 4.

The cycles correspond to structural properties of the mixture. They partition the components, so write $[k]$ for the equivalence class of bump k under π. For instance, a component θ_k might itself satisfy $A\theta_k = \theta_k$, and $\pi(k) = k$: a cycle of length $Q = 1$. At the other end, a chain of $Q = P$ intermediate components, each having no symmetry properties, lead back to θ_k. Such a group is shown at left in figure 1, which takes $P = Q = 6$ and schematically represents application of A to some θ_k as rotation by 60°, and distinct components θ_l, $l \in [k]$ as different markers. (The figure shows them in sequence, although that is not true in general.) Cycles of length $Q > P$ cannot occur: otherwise, both θ and $\theta' = A^P\theta$ would exist as distinct members of Θ. Since $A^P = I$, this would violate identifiability.

More generally, cycles of $1 \leq Q \leq P$ components occur if and only if $Q \mid P$ (i.e., Q divides P). The middle panel of the figure shows the $Q = 3$ case where $\mathcal{C} = (12, 13, 14)$; there are only three distinct markers because $\theta_{12} = A^3\theta_{12}$. At right, $Q = 2$ and $\mathcal{C} = (15, 16)$. These diagrams illustrate why Q must divide P. Formally, this is just Lagrange's theorem applied to the cyclic group of order P: all its subgroups are cyclic and of an order dividing P.

Cycle	Mixture Indexes	Q	P'	Internal Constraint	Shared Parameter Constraint
1	0–5	6	1	none: $A^6 = I$	$\theta_5 = A\theta_4 = \cdots = A^5\theta_0$
2	6–11	6	1	none: $A^6 = I$	$\theta_{11} = A\theta_{10} = \cdots = A^5\theta_6$
3	12–14	3	2	$\theta_{12} = A^3\theta_{12}$	$\theta_{14} = A\theta_{13} = A^2\theta_{12}$
4	15–16	2	3	$\theta_{15} = A^2\theta_{15}$	$\theta_{16} = A\theta_{15}$
5	17–18	2	3	$\theta_{17} = A^2\theta_{17}$	$\theta_{18} = A\theta_{17}$
6	19	1	6	$\theta_{19} = A\theta_{19}$	—

Within this restriction, many component structures may coexist in Θ; we establish conventions for their ordering. A K-bump mixture corresponds to a tuple K_s, with entries summing to K, each giving the number of mixture components devoted to cycles of each possible length Q such that $Q \mid P$. For instance, if $P = 6$, a symmetry of $K_s = (12, 3, 4, 1)$ implies $K = 20$ and

$$\pi = ((0, 1, 2, 3, 4, 5)(6, 7, 8, 9, 10, 11)(12, 13, 14)(15, 16)(17, 18)(19)).$$

The table above itemizes the parameters, and figure 1 shows parameters corresponding to the first, third, and fourth cycles of π.

Suppose a given cycle contains Q components. In the conventional ordering, the components have a *shared parameter constraint*

$$\theta_{k+1} \equiv A\theta_k, \ldots, \theta_{k+Q-1} \equiv A^{Q-1}\theta_k\ . \tag{8a}$$

Furthermore, each component also satisfies an *internal constraint*

$$(\forall l \in [k])\, \theta_l = A^Q\theta_l\ . \tag{8b}$$

We will use Lagrange multipliers to enforce (8b). The Lagrangian term for $\mu = A\mu$ is $l_\mu = \lambda^\mathsf{T}(\mu - A\mu)$ for a vector λ to be determined. Enforcing $\Sigma = A\Sigma A^\mathsf{T}$ calls for a matrix Λ, one for each entry of $D = \Sigma - A\Sigma A^\mathsf{T}$:

$$l_\Sigma = \sum\nolimits_{i,j} \Lambda_{ij} D_{ij} = \operatorname{tr} D^\mathsf{T}\Lambda = \operatorname{tr}(\Sigma - A\Sigma A^\mathsf{T})\Lambda = \operatorname{tr}\Sigma(\Lambda - A\Lambda A^\mathsf{T}) \tag{9}$$

where we have used $\Sigma = \Sigma^\mathsf{T}$ and the trace identity, $\operatorname{tr} AB = \operatorname{tr} BA$. The constraint on Σ is equivalent to the same constraint on Σ^{-1}, so we use instead the more convenient $l_{\Sigma^{-1}} = \operatorname{tr}\Sigma^{-1}(\Lambda - A\Lambda A^\mathsf{T})$.

Earlier work on constrained mixtures imposes structure to compactly parameterize the covariance. Some structured covariances (e.g., $\Sigma_k = \sigma_k^2 I$) can be trivially handled in the EM algorithm. This idea has been extended using the eigendecomposition $\Sigma_k = \lambda_k H_k D_k H_k^\mathsf{T}$ where the H_k are orthogonal, $\lambda_k D_k$ is the diagonal eigenvalue matrix, and $|D_k| = 1$; a family of EM algorithms results [2], [3] from various parameter-sharing schemes. A "semi-tied" covariance model has been used in output modeling for hidden Markov models (HMMs) [4]. This parameterizes a subset $\mathcal{K} \subset \{0, \ldots, K-1\}$ of covariances by sharing H. Other subsets $\mathcal{K}'$ could have different structuring matrices H. Mixtures of factor analyzers [5] are another twist: covariance models of the form $\Sigma_k = H_k H_k^\top + D_k$, with low-rank H_k and diagonal D_k. The constraints we consider give rather different structure to the covariance, and affect the means and weights as well. The structure imposed on the Gaussian distribution (i.e., $K = 1$) by symmetry expressed as an algebraic group has been deeply elucidated [1, App. A]. For concreteness, we have specialized in this paper to the cyclic group, while treating the more general class of K-component mixtures with a more computational viewpoint.

3 Normal mixture solution

Following the standard approach to fitting a mixture distribution via EM (e.g., [6, sec. 3.2]), define for each x_n a corresponding sequence of indicator variables $Z_n = (z_{n,0}, \dots, z_{n,K-1})$. Exactly one of these indicators equals one, signaling which component of (2) generated x_n. We correspondingly denote $Z = \{Z_n\}_{n=1}^N$, and the pair (X, Z) becomes the complete-data of the EM algorithm. The log probability of the complete-data decouples as

$$\log p(X, Z) = \sum_{k=0}^{K-1} \sum_{n=1}^{N} z_{n,k} \log[\gamma_k N(x_n;\, \mu_k, \Sigma_k)]$$

and its expectation given the observation is

$$Q(\Theta) = E[\log p(X, Z) \,|\, X] = \sum_{k=0}^{K-1} \sum_{n=1}^{N} \tau_{n,k} \log[\gamma_k N(x_n;\, \mu_k, \Sigma_k)] \tag{10}$$

where the weights are regarded as known:

$$\tau_{n,k} := E[z_{n,k} \,|\, x_n] = \gamma_k N(x_n; \mu_k, \Sigma_k) \Big/ \sum_{l=0}^{K-1} \gamma_l N(x_n; \mu_l, \Sigma_l)\ . \tag{11}$$

The quantity $\tau_{n,k}/N$ is a joint pmf. It is convenient to also define $\tau_k = \sum_{n=1}^N \tau_{n,k}$ and $\tau_{n|k} = \tau_{n,k}/\tau_k$. The latter is a correctly normalized conditional distribution. We maximize $Q(\Theta)$ at every EM iteration to update the parameters. We use the parameter ordering convention described above.

The update for the weights can be derived separately because the terms of Q involving γ_k separate out. Including the Lagrangian term for the unit mass constraint on the weights, the function to be maximized is

$$Q_C(\gamma_0, \dots, \gamma_{K-1}) = \sum\nolimits_{k=0}^{K-1} \tau_k \log \gamma_k + \lambda\Big(1 - \sum\nolimits_{k=0}^{K-1} \gamma_k\Big)\ .$$

To find γ_k, recall from (8a) that all of the weights γ_l, $l \in [k]$, are in fact the same parameter. Differentiating reveals the optimal weight is

$$\hat{\gamma}_k = \big(1/\#[k]\big) \sum\nolimits_{l \in [k]} \tau_l / N \tag{12}$$

where $\#[k]$ is the cardinality of the cycle. This is just the average class-membership in the cycle containing k, normalized to sum to unity.

Using the trace identity, the terms of (10) involving means and covariances are conventionally written via the weighted sufficient statistics:

$$Q(\mu_0, \ldots, \mu_{K-1}, \Sigma_0, \ldots, \Sigma_{K-1}) =$$

$$-\frac{1}{2} \sum_{k=0}^{K-1} \tau_k \left[\log |\Sigma_k| + (m_k - \mu_k)^\mathsf{T} \Sigma_k^{-1} (m_k - \mu_k) + \operatorname{tr} \Sigma_k^{-1} S_k(m_k)\right] \tag{13}$$

$$m_k := \sum_{n=1}^{N} \tau_{n|k} x_n \quad \text{and} \quad S_k(\eta) := \sum_{n=1}^{N} \tau_{n|k} (x_n - \eta)(x_n - \eta)^\mathsf{T} \,. \tag{14}$$

The k subscript indicates weighting by the conditional probabilities $\tau_{n|k}$.

It is immediate from the sum in (13) that, in the usual unconstrained mixture problem, parameter updates for $(\hat{\mu}_k, \hat{\Sigma}_k)$ decouple across k. In the constrained case, differentiating with respect to μ_k or Σ_k will involve all components in $[k]$, but no others: components within a cycle are tied via (8a). In the remainder of this section, we suppose the cycle is indexed as $[k] = \{0, \ldots, Q-1\}$ to cut down on superfluous notation.

To enforce the shared parameter constraint (8a), let μ_0 be a free parameter and write $\mu_l = A^l \mu_0$, $0 < l < Q$, and similarly for the covariances. Use the Lagrangian mechanism to account for the internal constraint (8b), namely

$$\mu_l = A^Q \mu_l, \quad \Sigma_l = A^Q \Sigma_l A^{\mathsf{T}Q}, \quad 0 \le l < Q, \tag{15}$$

which is of course accomplished by constraining (μ_0, Σ_0) only. With this way of writing the parameters, the cycle-k terms of (13) are

$$Q(\mu_0, \Sigma_0) = -\frac{\tau_{[0]}}{2} \sum_{k=0}^{Q-1} \bar{\tau}_k \left[\log |\Sigma_0| + (A^{\mathsf{T}k} m_k - \mu_0)^\mathsf{T} \Sigma_0^{-1} (A^{\mathsf{T}k} m_k - \mu_0) + \right.$$

$$\left. \operatorname{tr} \Sigma_0^{-1} A^{\mathsf{T}k} S_k(m_k) A^k \right] \tag{16}$$

where $\tau_{[0]} := \sum_{k=0}^{Q-1} \tau_k$ and $\bar{\tau}_k = \tau_k / \tau_{[0]}$, a pmf on $\{0, \ldots, Q-1\}$.

Collapsing Q parameters to one makes, e.g., $m_0, \ldots, m_{Q-1}$ informative about μ_0. It aids understanding to write (16) with new sufficient statistics

$$\bar{m} := \sum_{k=0}^{Q-1} \bar{\tau}_k A^{\mathsf{T}k} m_k \quad \text{and} \quad \bar{S} := \sum_{k=0}^{Q-1} \bar{\tau}_k A^{\mathsf{T}k} S_k(A^k \bar{m}) A^k \,. \tag{17}$$

Intuitively, the cycle's statistics are transformed back to the (μ_0, Σ_0) coordinates and averaged there. Formally, $\bar{m}$ arises by completing the square in the quadratic form involving μ_0 in (16). With this definition, and including Lagrangian terms, the objective function simplifies to

$$Q_C(\mu_0, \Sigma_0) = -\log |\Sigma_0| - (\bar{m} - \mu_0)^\mathsf{T} \Sigma_0^{-1} (\bar{m} - \mu_0) - \operatorname{tr} \Sigma_0^{-1} \bar{S} +$$

$$2\lambda^\mathsf{T} (\mu_0 - A^Q \mu_0) + \operatorname{tr} \Sigma_0^{-1} (\Lambda - A^Q \Lambda A^{\mathsf{T}Q}) \tag{18}$$

Differentiating with respect to μ_0 gives the necessary condition

$$\hat{\mu}_0 = \bar{m} + \Sigma_0 (I - A^Q)^{\mathsf{T}} \lambda$$

To satisfy the constraint, note that the average of $P' = P/Q$ transformed means $\hat{\mu}_0, A^{\mathsf{T}Q}\hat{\mu}_0, \ldots, A^{\mathsf{T}(P'-1)Q}\hat{\mu}_0$ telescopes to:

$$\hat{\mu}_0 = (1/P') \sum\nolimits_{r=0}^{P'-1} A^{\mathsf{T}Qr} \bar{m} \; . \tag{19}$$

Substituting $\hat{\mu}_0$ into the Lagrangian (18) and differentiating with respect to the elements of Σ_0^{-1} reveals a necessary condition

$$\hat{\Sigma}_0 - \bar{S} - (\bar{m} - \hat{\mu}_0)(\bar{m} - \hat{\mu}_0)^{\mathsf{T}} + (\Lambda - A^Q \Lambda A^{\mathsf{T}Q}) = 0$$

Enforcing the constraint with the averaging method reveals

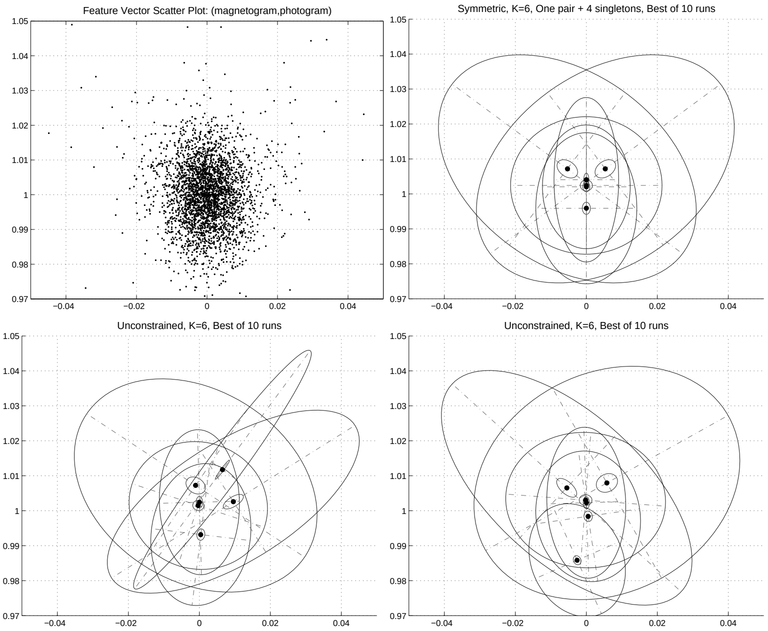
$$\hat{\Sigma}_0 = (1/P') \sum\nolimits_{r=0}^{P'-1} A^{\mathsf{T}Qr} \big[\bar{S} + (\bar{m} - \hat{\mu}_0)(\bar{m} - \hat{\mu}_0)^{\mathsf{T}} \big] A^{Qr} \; ; \tag{20}$$

compare [1, Thm. A.2] for the $K = Q = 1$ case. To sum up, the parameters are updated with a nested average of transformed sufficient statistics. The inner averages (17) are across Q terms, one for each linked component in the cycle. The outer averages, in (19) and (20), sum over the symmetries in the order-P' cyclic subgroup of G to enforce invariance with respect to A^Q.

4 Implementation and results

The new information needed is A and the symmetry vector K_s giving π: how many bumps to allocate to each symmetric configuration. (Unconstrained EM has $K_s = K$, $A = I$.) Standard EM finds $(m_k, \Sigma_k)_{k=0}^{K-1}$ as in (14). The new procedure follows these E and M steps with a constraint step which loops over each cycle of π, performing a $P = QP'$-fold averaging as in (17), (19), and (20). This takes $O(Kd^3)$ operations, dwarfed by the $O(NKd^3)$ in each ordinary EM step. If all cycles have $Q = P$, the constrained algorithm is equivalent to copying each $x \in X$, P times $(x, Ax, ..., A^{P-1}x)$ plus unconstrained EM, but requires P times less computation.

On the next page we compare unconstrained versus constrained methods with $K = 6$ on $N = 15032$ feature vectors from MDI images (top left). Each run selects the highest-likelihood model after ten, 1000-update EM sequences. The unconstrained models are unstable from run to run; the bottom panels show concentration ellipses and centers of two typical best-of-ten models. The constrained model (top right) uses $K_s = (2, 4)$. It does not have run-to-run instability, and its decomposition provides interpretable information: the symmetric pair is due to the chromospheric network, a small brightening distributed across the solar disk. One-bump cycles (i.e., $Q = 1$) are needed: models with $K_s = (K, 0)$ do not coalesce paired bumps and converge very slowly to inferior models. Three similar mixtures with K_s of $(4, 4)$, $(12, 2)$, and $(4, 2)$ are used operationally to identify three types of solar activity [7]. The constraint proved essential to estimate these more complex models.



References

[1] Andersson S., Madsen J. (1998). *Symmetry & lattice conditional independence in a multivariate normal distribution.* Ann. Statist. **26** (2),525 – 572.

[2] Celeux G., Govaert G. (1995). *Gaussian parsimonious clustering models.* Pattern Recognition **28**, 781 – 793.

[3] Fraley C., Raftery A.E. (2002). *Model-based clustering, discriminant analysis, and density estimation.* JASA **97**, 611 – 631.

[4] Gales M.J.F. (1999). *Semi-tied covariance matrices for hidden Markov models.* IEEE Trans. Speech and Audio Processing **7** (3), 272 – 281.

[5] Ghahramani Z., Hinton G.E. (1997). *The EM algorithm for mixtures of factor analyzers.* Technical Report CRG-TR-96-1, U. of Toronto.

[6] McLachlan G., Peel D. (2000). *Finite mixture models.* Wiley.

[7] Turmon M., Pap J., Mukhtar S. (2002). *Statistical pattern recognition for labeling solar active regions.* Astrophysical Journal **568** (1), 396 – 407.

[8] Yakowitz S.J., Spregins J.D. (1968). *On the identifiability of finite mixtures.* Ann. Math. Statist. **39**, 209 – 214.

Acknowledgement: This research was carried out for the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA.

Address: M. Turmon, M/S 126-347, Jet Propulsion Laboratory, Pasadena, CA 91109

E-mail: turmon@jpl.nasa.gov

COMPARISON OF ALGORITHMS FOR NONLINEAR REGRESSION ESTIMATES

Josef Tvrdík and Ivan Křivý

Key words: Global optimization, evolutionary algorithms, heuristics, nonlinear regression.

COMPSTAT 2004 section: Algorithms.

Abstract: This paper deals with the use of some stochastic algorithms for solving the problem of global optimization of nonlinear regression models. The algorithms were applied to estimating the parameters of eight nonlinear regression models taken from the Nonlinear Least Squares Datasets of the National Institute of Standards and Technology. The experiments showed that the stochastic algorithms under consideration provide more reliable results as compared with programs built into well-known statistical packages. The results obtained with the individual stochastic algorithms were compared and the revealed differences are briefly discussed.

1 Introduction

Least square estimation of parameters in non-linear regression models is the problem of finding the global minimum. The global optimization problem is usually defined as follows: For a given objective function $f : D \to \mathcal{R}$, $D \subset \mathcal{R}^d$, the point $\mathbf{x}^*$ is to be found such that $\mathbf{x}^* = \arg\min_{\mathbf{x}\in D} f(\mathbf{x})$. The point $\mathbf{x}^*$ is called the global minimum and D is a searching space. In practice the searching space D is often connected, defined as $D = \prod_{i=1}^{d} \langle a_i, b_i \rangle$, $a_i < b_i$, $i = 1, 2, \ldots, d$, and the objective function is computable, i.e. there is an algorithm capable to evaluate $f(\mathbf{x})$ with sufficient accuracy at any point $\mathbf{x} \in D$. The objective function may be multimodal. It is known that this problem cannot be effectively solved by deterministic algorithms in general, for details see e.g. [2]. But in last decades it was found that there are stochastic algorithms which are relatively successful in searching for the global minimum (or maximum) of complicated objective functions.

Ideas of combining different heuristics when searching for the global optimum appeared in the last decade, see e.g. [6, 3]. In this paper we deal also with algorithms based on the competition of different stochastic heuristics, the probability of the selection of a given heuristics in the current step being proportional to its successfulness in preceding steps. We hope that this is the way how to develop the algorithms searching for the global extreme with high self-adaptation and a low number of input parameters which have to be set by user. It promises that such a kind of algorithms will be useful in the statistical procedures where the global optimization of a potentially multimodal objective function is needed, standard local optimizers are not reliable enough and specialized algorithms are not available.

2 Nonlinear regression models for testing

The models taken from the Nonlinear Least Squares Datasets of the National Institute of Standards and Technology are listed in Table 1.

Dataset name	Model	d	n	R^2
Bennett5	$\beta_1(\beta_2 + x)^{-1/\beta_3}$	3	154	0.99998939
BoxBOD	$\beta_1(1 - \exp(-\beta_2 x))$	2	6	0.88046777
Eckerle4	$\frac{\beta_1}{\beta_2} \exp(\frac{-(x-\beta_3)^2}{2\beta_2^2})$	3	35	0.99706429
MGH09	$\frac{\beta_1(x^2+\beta_2 x)}{x^2+\beta_3 x+\beta_4}$	4	11	0.99405700
MGH10	$\beta_1 \exp(\frac{\beta_2}{x+\beta_3})$	3	16	0.99999988
Rat42	$\frac{\beta_1}{1+\exp(\beta_2-\beta_3 x)}$	3	9	0.99826670
Rat43	$\frac{\beta_1}{(1+\exp(\beta_2-\beta_3 x))^{1/\beta_4}}$	4	15	0.99183768
Thurber	$\frac{\beta_1+\beta_2 x+\beta_3 x^2+\beta_4 x^3}{1+\beta_5 x+\beta_6 x^2+\beta_7 x^3}$	7	37	0.99950790

Table 1: List of regression models.

The columns d, n and R^2 of Table 1 show the number of parameters, the number of observations and the index of determination, respectively. Additional information including source of data, starting values of parameters, certified values of parameters and the corresponding standard deviations are also summarized in [8].

All the models represent optimization tasks of higher level of difficulty. The tasks are not easy for standard algorithms, see [4]. Our results when several statistical packages were used are shown in Table 2. We used several standard packages for least squares estimation of parameters, namely NCSS 2001, where Levenberg-Marquardt algorithm is used, S-PLUS 4.5 using Gauss-Newton (GN) algorithm, SPSS 10.0 with modified Levenberg-Marquardt algorithm and SYSTAT 8.0 where both a modified GN algorithm and the simplex method are implemented.

Identification of the tasks is the same as in N.I.S.T Datasets. Both variants of starting values of estimates recommended in [8] were used for all tasks and packages. The results for the first set of starting values are in the columns 1, for the second set of starting values in the columns 2. The result marked "OK" in Table 2 means that agreement in sum of residuals squares is at least in four digits, "F" stands for failure (no solution found or the algorithm stopped at local minimum) and numerical data mean the level of agreement with certified value in [8] giving the number of identical valid digits. We see in Table 2 that no package was completely successful in spite of the fact that the recommended starting values are very near to the certified values of estimates.

	NCSS		SYST GN		SYST S		S-Plus		SPSS	
Task	1	2	1	2	1	2	1	2	1	2
Bennett5	2	1	OK	OK	F	F	OK	OK	OK	OK
BoxBOD	F	F	OK	OK	F	F	OK	OK	F	OK
Eckerle4	F	3	OK	OK	F	F	F	OK	OK	OK
MGH09	F	F	F	OK	OK	OK	F	2	OK	OK
MGH10	F	F	OK	OK	F	OK	F	OK	F	F
Rat42	OK	OK	OK	OK	F	F	F	OK	OK	OK
Rat43	F	3	OK	OK	F	F	F	OK	OK	OK
Thurber	F	F	OK	OK	OK	OK	F	F	F	F

Table 2: Reliability of statistical packages.

3 Stochastic algorithms

Five stochastic algorithms were applied to estimate the parameters in the tasks mentioned above. Four of them (marked CRS, MCRS, ALT and COMP1 in the text) are a generalization of controlled random search, which is described as follows:

Algorithm 1. Generalized controlled random search

```
generate $\mathcal{P}$ (population of $N$ points in $D$ at random);
find $\mathbf{x}_{\text{worst}}$ (the point $\mathcal{P}$ with the highest value of objective function);
repeat
    apply a heuristic to generate a new trial point $\mathbf{y} \in D$
    if $f(\mathbf{y}) < f(\mathbf{x}_{\text{worst}})$ then
        $\mathbf{x}_{\text{worst}} := \mathbf{y}$;
        find $\mathbf{x}_{\text{worst}}$;
    endif
until stopping condition is true;
```

The CRS (Controlled Random Search) algorithm was originally proposed by W. L. Price [7]. It starts from the population $\mathcal{P}$ of N ($d \ll N$) points chosen arbitrarily from the searching space D. At each iteration $d+1$ points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{d+1}$ are taken at random from $\mathcal{P}$ forming a simplex S in d - space. The new trial point $\mathbf{y}$ is defined as the point obtained by reflection of the point $\mathbf{x}_H$ with respect to the centroid $\mathbf{g}$ of remaining d points, i. e.

$$\mathbf{y} = 2\mathbf{g} - \mathbf{x}_H, \tag{1}$$

where $\mathbf{x}_H$ is the point with the highest objective function value in the simplex S.

When using the MCRS (Modified Controlled Random Search) algorithm [5], the new trial point $\mathbf{y}$ is generated from the simplex S by the relation

$$\mathbf{y} = \mathbf{g} - Y(\mathbf{x}_H - \mathbf{g}). \tag{2}$$

The multiplication factor Y is not constant like in [7] but a random variable. In this implementation we used Y distributed uniformly in $\langle 0, \alpha)$ with $\alpha = 2$. Thus the mean value $EY = 1$ and for the mean value (2) gives (1).

ALT and COMP1 algorithms do not use only one heuristic for generating a new trial point $\mathbf{y}$ but they choose among several ones. The heuristic in generalized controlled random search is any non-deterministic rule which gives a new point $\mathbf{y} \in D$. Let us say that we have h heuristics at disposal. In both ALT and COMP1 a heuristic is chosen at random with the probability $q_i,\ i = 1, 2, \ldots, h$. In the ALT procedure the probabilities q_i are constant and the same through searching process ($q_i = 1/h$). In the COMP1 procedure the probabilities are changing according to the successfulnes of heuristics in preceding steps of the searching process. The strategy used here is based on simply counting the number of successful insertions of new trial points, n_i,

$$q_i = \frac{n_i + n_0}{\sum_{j=1}^{h}(n_j + n_0)} \ ,$$

where $n_0 > 0$ is a constant. Setting $n_0 > 1$ prevents a dramatic change in q_i by one random successful use of the i-th heuristics. In order to avoid the degeneration of the evolutionary process the current values of q_i are reset to their starting values ($q_i = 1/h$) when any probability q_i decreases bellow a lower limit δ. We applied eleven heuristics in the implementation:

- four heuristics based on randomized reflection in the simplex [5],
- four heuristics coming from differential evolution [9],
- two heuristics derived from evolutionary strategy (see e.g. [2]),
- uniform random search.

The list of the eleven heuristics and setting their parameters is given in [10], [11]. The values of additional parameters used in the experiments were $\delta = 0.02$ and $n_0 = 2$.

The DER (differential evolution) algorithm [9], unlike CRS and MCRS, attempts to replace all N points of the population by new points at each generation. For each point $\mathbf{x}_i, i = 1, 2, \ldots, N$, a new trial point $\mathbf{y}_i$ is found from two parents, the point $\mathbf{x}_i$ and the point $\mathbf{u}_i$ obtained by using mutation. In order to determine the mutant point $\mathbf{u}_i$ it is necessary to select randomly three distinct points $\mathbf{x}_{p1}, \mathbf{x}_{p2}$ and $\mathbf{x}_{p3}$ from $\mathcal{P}$ (not coinciding with the current $\mathbf{x}_i$) and to add the weighted difference of any two points of them to the third point, which can be described as

$$\mathbf{u}_i = \mathbf{x}_{p1} + F(\mathbf{x}_{p2} - \mathbf{x}_{p3}), \tag{3}$$

where $F > 1$ is a scaling factor. Generating the point $\mathbf{u}_i$ according to (3) is called *rand* strategy in evolutionary-algorithms community. Finally, the components y_{ij} of trial point $\mathbf{y}_i$ is found from its parents $\mathbf{x}_i$ and $\mathbf{u}_i$ using the following crossover rule:

$$y_{ij} = \begin{cases} u_{ij} & \text{if } R_j \le C \quad \text{or } j = I_i \\ x_{ij} & \text{if } R_j > C \quad \text{and } j \ne I_i, \end{cases}$$

where I_i is a randomly chosen integer from $\{1, 2, \ldots, d\}$, $R_j \in (0, 1)$ taken at random for each j, and $C \in \langle 0, 1\rangle$ is a constant. The input parameters of the DER algorithm are the population size N, scaling factor F and the parameter C influencing the number of elements to be exchanged by crossover. The values of parameters used in the experiments were $F = 0.5$ and $C = 0.5$.

		b_1	b_2	b_3	b_4	b_5	b_6	b_7
Bennett5	min	-5000	0	0.1				
	max	-1000	500	10				
BoxBOD	min	1	0.1					
	max	1000	2					
Eckerle4	min	0	1	400				
	max	10	10	500				
MGH09	min	0	0	0	0.01			
	max	100	100	100	100			
MGH10	min	0	0	0				
	max	100	1e+6	1e+5				
Rat42	min	0	0	0				
	max	1000	10	1				
Rat43	min	0	0	0	0.1			
	max	1000	100	1	10			
Thurber	min	0	0	0	0	0	0	0
	max	1e+4	5000	5000	1000	10	10	10

Table 3: Searching spaces.

4 Experimental results

All the algorithms were implemented in Matlab. Common input parameters for all the five stochastic algorithms were set as follows:

- Population size $N = 10\,d$ where d is the number of regression parameters
- Stopping condition was based on the difference in the R^2 between the best and the worst points in population, the process was stopped when the difference was less then 1×10^{-12} or when the number of objective function evaluations exceeds $40000\,d$.

	CRS			MCRS			DER			ALT			COMP1		
Task	R	$\overline{\text{NE}}$	vc	R	$\overline{\text{NE}}$	vc	R	$\overline{\text{NE}}$	vc	R	$\overline{\text{NE}}$	vc	R	$\overline{\text{NE}}$	vc
Bennett5	100	45352	33.6	100	54807	32.97	0			24	78645	37.7	88	78281	39.7
BoxBOD	100	1091	18.4	100	737	7.3	100	2507	7.2	100	1300	8.5	100	1131	9.0
Eckerle4	100	3381	4.7	100	1420	4.8	100	3911	5.0	100	2358	5.8	100	2033	5.7
MGH09	100	13102	7.3	100	8549	16.2	27	133644	15.3	100	21311	14.5	100	13562	12.7
MGH10	100	24415	4.6	100	23523	15.0	0			100	70928	3.5	100	41765	6.6
Rat42	100	3374	4.9	100	1468	5.5	99	9792	8.0	100	3495	6.6	100	2734	7.6
Rat43	100	6289	3.3	100	1996	6.0	100	23762	5.6	100	5708	5.6	100	3567	6.4
Thurber	95	26447	2.6	74	9161	19.9	0			74	22122	6.8	73	12569	10.8

Table 4: Overall results.

- The searching space D was also the same for all the algorithms, the settings of D for the tasks are given in Table 3.

One hundred of independent runs were carried out for all tasks and algorithms. The number of objective function evaluations (NE) and the success in finding a point very near to the global minimum were the most substantial variables evaluated in each run. The run was considered successful if the agreement of residual sum of squares was at least to four digits with the certified value in [8]. The reliability R was measured by the percentage of the successful runs, the convergence rate was measured by the average number of objective function evaluations $\overline{\text{NE}}$. The averages of function evaluation $\overline{\text{NE}}$ and standard deviations s were computed only from the successful runs. The overall results for all the stochastic algorithms are shown in Table 4. The variability of NE expressed as the variation coefficient ($100\ s\ /\ \overline{\text{NE}}$) is given in the columns marked vc. The mean CPU time needed for one objective function evaluation varied from 2 to 4 milliseconds on standard PC (667 MHz). depending on the task and algorithm. The algorithms are significantly different both in reliability and convergence rate. With respect to reliability the algorithms are ordered in decreasing sequence: CRS, MCRS, COMP1, ALT, DER. The DER algorithm completely failed in three tasks when no successful search occurred. The convergence rate is substantially worse than in the case of other algorithms for remaining tasks. The average NEs which are not significantly different by Scheffe's multiple comparisons are underlined in Table 4.

When we drop the DER algorithm and compare the convergence rate of the remaining algorithms on the the tasks where they were completely successful (Bennett5 and Thurber excluded), the fastest algorithm is MCRS with average rank 1.08, followed by COMP1 (average rank 2.50), CRS (average rank 2.83) and ALT (average rank 2.58). The ranks of underlined values was replaced with their average ranks.

5 Conclusions

The stochastic algorithms CRS, MCRS, DER, ALT and COMP1 were applied to optimizing eight nonlinear regression models of higher level of difficulty. The algorithms proved to be generally more reliable as compared with commercial algorithms built in standard statistical packages but rather time-consuming. The CRS algorithm showed to be the most reliable while the MCRS was the fastest among the tested algorithms. The DER algorithm completely failed when optimizing three of the tested models, which is surprising with respect to the very good reputation of differential evolution, see also [1]. The bad convergence rate of the DER algorithm used in the experiments is perhaps due to its sensitivity to parameter setting, namely to the size of population. The algorithm with competing heuristics (COMP1) did not meet our expectations; it is not adaptive enough when solving difficult optimization tasks. Our experimental results indicate the validity of the no free lunch theorem [12] claiming that there is no stochastic algorithm which generally outperforms the others.

References

[1] Ali M. M., Törn A. (2004). *Population set based global optimization algorithms: Some modifications and numerical studies.* Computers and Operations Research **31**, 1703 – 1725.

[2] Bäck T. (1996). *Evolutionary algorithms in theory and practice.* Oxford University Press, New York.

[3] Gacôgne L. (2002). *Steady-state evolutionary algorithm with operator family.* Intelligent Technologies - Theory and Applications. IOS Press, Amsterdam, 173 – 182.

[4] McCullough B.D., Wilson B. (1999). *On the accuracy of statistical procedures in Microsoft Excel 97*. Comput. Statist. and Data Anal. **31**, 27 – 37.

[5] Křivý I. and Tvrdík J. (1995). *The controlled random search algorithm in optimizing regression models*. Comput. Statist. and Data Anal. **20**, 229 – 234.

[6] Kvasnička V. and Pospíchal J. (1997). *A hybrid of simplex method and simulated annealing*. Chemometrics and Intelligent Laboratory Systems **39**, 161 – 173.

[7] Price W.L. (1977). *A controlled random search procedure for global optimization*. Computer J. **20**, 367 – 370.

[8] Statistical Reference Datasets. *Nonlinear regression*. NIST Information Technology Laboratory. http://www.itl.nist.gov/div898/strd/. December 2001.

[9] Storn R., Price K. (1997). *Differential evolution – a simple and efficient heuristic for global optimization*. J. Global Optimization **11**, 341 – 359.

[10] Tvrdík J., Mišík L. and Křivý I. (2002). *Competing heuristics in evolutionary algorithms*. Intelligent Technologies - Theory and Applications, IOS Press, Amsterdam, 159 – 165.

[11] Tvrdík J., Křivý I. and Mišík L. (2002). *Evolutionary algorithm with competing heuristics in computational statistics*. Compstat 2002, Physica-Verlag, Heidelberg, 349 – 354.

[12] Wolpert D. H., Macready W.G. (1997). *No free lunch theorems for optimization*. IEEE Transactions on Evolutionary Computation, **1**, 67 – 82.

Acknowledgement: This work was supported by the institutional research grant no. CEZ:J09/98:179000002. We thank to Radek Krpec for his help with processing data for the Table 2.

Address: Department of Computer Science, University of Ostrava, 30. dubna 22, 701 03 Ostrava, Czech Republic

E-mail: `Josef.Tvrdik@osu.cz; Ivan.Krivy@osu.cz`

ROBUST CLASSIFICATION OF HIGH-DIMENSIONAL DATA

Karlien Vanden Branden and Mia Hubert

Key words: Robustness, classification, high-dimensional data, principal component analysis.

COMPSTAT 2004 section: Classification.

Abstract: In this paper we introduce a robust method to classify high-dimensional observations. We consider the SIMCA (Soft Independent Modelling of Class Analogies) approach to obtain a classification rule. Although a primary goal of this method was to detect outlying observations, the method itself is not robust. Therefore we will consider a robust classification method that shares the same ideas as the SIMCA approach. This robust method is based on a robust method for principal component analysis for high-dimensional data.

1 Introduction

We will use the SIMCA approach [11], [1], [9] to construct a model such that future observations are most likely to be correctly classified. Such a classification rule is developed based on a *training* data set that contains observations which are drawn out of m different populations. Let us denote the p-dimensional observations by $\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \dots, \boldsymbol{x}_{n_i}^i$ for $i = 1, 2, \dots, m$ and let n_i represent the number of observations in the ith group. For each observation we thus know to which group it belongs which is a natural assumption in classification or discriminant analysis. Note that we write column vectors in bold and the transpose of a matrix or a vector is represented with a $'$.

Many classification rules have been proposed in the past. The SIMCA method is very popular in chemometrics as it is in particular very useful to classify high-dimensional data, as there are spectra or micro-array data. This makes SIMCA also to be a very interesting method in applications based on gene expression data (e.g. the classification of cancer tumors) or in image analysis. More classical methods such as linear and quadratic discriminant analysis on the other hand are limited to situations where the number of observations in each group is at least as large as the dimension p.

Because the class membership of each observation in the training set is known, the SIMCA method starts by performing a Principal Component Analysis (PCA) in each of the m groups separately. Hereby the original p-dimensional observation $\boldsymbol{x}_j^i$ is transferred into a k_i-dimensional score vector $\boldsymbol{t}_j^i$. Note that we use the notation k_i meaning that for each group of observations the retained number of significant principal components can differ. This analysis thus provides information on the shape and the center of each group which will be used while constructing a classification rule. We will discuss the SIMCA method and the evaluation of the classification rule in more detail in Section 2.

However, if the data set does not only contain clean observations, classical PCA can give bad estimates for data reduction. Therefore, robust PCA methods for high-dimensional data have recently been developed (e.g. [5], [6]). In Section 3 we will consider a robust SIMCA method, RSIMCA, that incorporates a robust PCA method and consequently is resistant towards outlying samples. We will also construct and evaluate various classification rules.

A small simulation study in Section 4 will illustrate that our proposed method is robust and that it surpasses SIMCA in the case of contaminated data. Also an example with real data is provided in Section 5 to illustrate how contaminated data effect the SIMCA method while it does not influence the RSIMCA approach. We conclude this paper with some remarks in Section 6.

2 The SIMCA approach

As mentioned in the introduction, the SIMCA method consists of two important stages. First the data set is split in m separate groups according to the membership of each observation. Each group contains n_i p-dimensional observations $\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \ldots, \boldsymbol{x}_{n_i}^i$. Let X^i denote the data matrix for the ith group with as jth row element $(\boldsymbol{x}_j^i)'$, so X^i is a $n_i \times p$ matrix. Then a dimension reduction in each group is obtained by means of PCA:

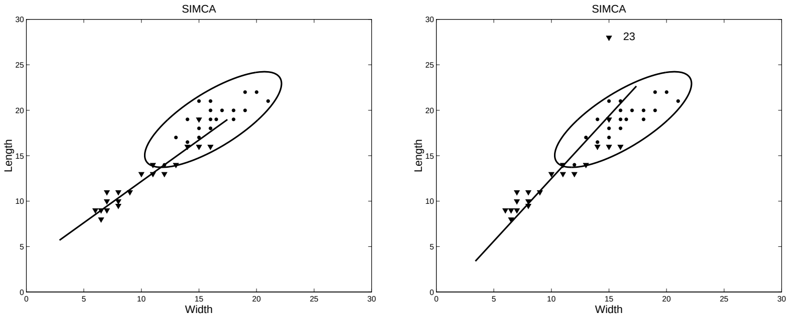
$$T_{n_i,k_i}^i = (X^i - \mathbf{1}_{n_i}(\bar{\boldsymbol{x}}^i)')P_{p,k_i}^i. \tag{1}$$

Here P_{p,k_i}^i denotes the matrix containing the first k_i principal components of the observations in group i, i.e. the first k_i dominant eigenvectors of the variance-covariance matrix S of the data points. The mean of the observations is denoted by $\bar{\boldsymbol{x}}^i$ and T_{n_i,k_i}^i, the score matrix, represents the coordinates of the projected observations. So the projected observation $(\boldsymbol{t}_j^i)'$ can be found on the jth row of T_{n_i,k_i}^i. The dimension k_i, which is often determined based on cross-validation, indicates how strongly the p-dimensional space is decreased. Again notice that in each group a different number of principal components can be retained.

In the second step of the method the information from the PCA stage is used to obtain and evaluate a classification rule. This classification rule is developed by means of two distances. In the SIMCA method one looks at the orthogonal distance of an observation to the space spanned by the k_i most important principal components of a particular group, and to the distance of an observation to a box surrounding the observations in that group. This box is constructed based on the scores of a group. The sum of these two squared distances is then converted such that an F-test is appropriate. An observation is then addressed to the ith group if its experimental value for the ith group is lower than a critical bound. For more information on this bound, we refer to [1], [9].

A drawback of the SIMCA approach is the large effect one single outlying observation can exert on the classification rule. Let us therefore look at a simple two-dimensional example. The *Jellyfish* data set, which is available at `http://www.statsci.org/data/`, consists of measurements on the width and the length of two types of jellyfish. For the first group of jellyfish 22 observations are available. As the first principal component already accounts for 97.54% of the total variation, we retain $k_1 = 1$ component for this group. The second sample consists of 24 observations. Here we retain $k_2 = 2$ components as one single component explains only 87.75% of the total variation, so no dimension reduction is performed. In Figure 1(a) the two populations of jellyfish are plotted together with the first principal component of the first group and the 97.5% tolerance ellipse of the second group for the classical SIMCA method. This ellipse is defined as the set of vectors in $\mathbb{R}^2$ whose Mahalanobis distance is equal to $\chi^2_{2;0.975}$, the 0.975 quantile of the chi-squared distribution with two degrees of freedom. Applying the SIMCA classification rule that is implemented in the PLS toolbox [10] and which is roughly explained at the end of Section 3, results in 10.87% of misclassifications (number of misclassified observations divided by total observations) for the original clean data.

If we create one outlier in the first group (a new observation 23 is added to group 1) we obtain Figure 1(b). The first principal component is clearly twisted towards the outlier and the classification rule with respect to this first group is heavily damaged. Consequently, the misclassification percentages are also effected by this single outlier. The total percentage of misclassifications is raised to 17.39%. Note that we did not evaluate the classification result of the outlier.



(a) (b)

Figure 1: The effect of one outlier on the *Jellyfish* data: (a) the original data; (b) the contaminated data (observation 23 added to group 1). A ‘▼’ represents the observations in the first group and a ‘•’ the observations in the second group.

3 Robust classification

This previous example clearly illustrates the need for a Robust SIMCA method, RSIMCA. To obtain a 'robust' classification method that shares the same ideas as the SIMCA approach, we first perform a robust PCA method. A fast and recently developed robust PCA method for high-dimensional data is the ROBPCA method [6]. This method combines the ideas of projection pursuit and of robust covariance estimation. First the data points of the data matrix X^i are projected on a k_i-dimensional subspace. This subspace is defined by means of a measure of outlyingness which is computed for each data point. This measure is obtained by projecting the high-dimensional data points on many univariate directions $\boldsymbol{v}$. For every direction a robust center and scale of the projected data points $(\boldsymbol{x}_j^i)'\boldsymbol{v}$ is computed, namely the univariate Minimum Covariance Determinant (MCD) estimator [8] of location $\hat{\boldsymbol{\mu}}^i_{\text{MCD}}$ and scale $\hat{\sigma}^i_{\text{MCD}}$. The outlyingness of a data point $\boldsymbol{x}_j^i$ is then measured by means of:

$$\text{outl}(\boldsymbol{x}_j^i) = \max_{\boldsymbol{v}} \frac{|(\boldsymbol{x}_j^i)'\boldsymbol{v} - \hat{\boldsymbol{\mu}}^i_{\text{MCD}}|}{\hat{\sigma}^i_{\text{MCD}}}.$$

By performing PCA on the h points with the smallest outlyingness the k_i-dimensional subspace is determined. In the next step of the ROBPCA method the covariance matrix and the mean of the observations in this k_i-dimensional space are robustly estimated by means of the MCD estimator [8]. The obtained robust center and the robust covariance estimate are then transformed to the original p-dimensional space and finally we obtain a similar decomposition as in (1), but now with robust estimates.

In practical situations the optimal value for k_i can be determined through cross-validation. We apply a leave-one-out approach and take that value k_i for which the robust PRESS value is minimized. Its definition and a description of a fast algorithm for its computation are discussed in [3].

Next, we will develop a classification rule based on two distances that are very common in PCA. Assume a new observation $\boldsymbol{x}$ needs to be assigned to one of the m groups. As in SIMCA we also take into account the orthogonal distance of this observation to the PCA space of the ith group, OD^i. This distance is defined as:

$$\text{OD}^i = \|\boldsymbol{x} - \hat{\boldsymbol{\mu}}^i - P^i_{p,k_i}\boldsymbol{t}^i\|$$

with $\hat{\boldsymbol{\mu}}^i$ the robust center of the observations in the ith group and $\boldsymbol{t}^i = (t_1^i, t_2^i, \ldots, t_{k_i}^i)'$ the projection of $\boldsymbol{x}$ in the k_i-dimensional PCA subspace of the ith group. Next, we consider the score distance SD^i which represents the distance inside the PCA space of group i taking into account the covariance structure of the data. More formally this distance is defined by:

$$\mathrm{SD}^i = \sqrt{\sum_{l=1}^{k_i} \frac{(\boldsymbol{t}_l^i)^2}{l_l^i}}$$

with l_l^i the eigenvalues obtained with the robust PCA method.

In the next section we will evaluate different classification rules. All these rules are based on a combination of these two distances. For each new observation $\boldsymbol{x}$ to be classified we compute the orthogonal distance to and the score distance within each group. The assignment of this observation is then based on a linear combination of the distances or on a linear combination of the squared distances. To allow one of the distances to exert a larger influence on the classification rule, we insert a tuning parameter $\lambda \in [0, 1]$. More precisely, we assign $\boldsymbol{x}$ to the jth group if

$$\text{(R1):} \qquad \lambda \mathrm{OD}^i + (1-\lambda)\mathrm{SD}^i, \qquad i = 1, \ldots, m$$

is minimal for $i = j$. As a second possibility we consider the assignment to be based on the squared distances:

$$\text{(R2):} \qquad \lambda (\mathrm{OD}^i)^2 + (1-\lambda)(\mathrm{SD}^i)^2, \qquad i = 1, \ldots, m.$$

We also look at the above classification rules for the standardized distances. These distances are obtained by dividing the original distance by a cutoff value c_v^i or c_h^i such that if $\mathrm{OD}_j^i / c_v^i > 1$ observation j of the ith group can be regarded as an outlier in this group. The same holds for SD_j^i / c_h^i. Details about these cutoff values are given in [6]. In this way the orthogonal distance and the score distance receive equal importance in the assignment, at least when $\lambda = 0.5$. We will refer to these classification rules as (R3) and (R4). This last classification rule (R4) coincides more or less with the rule implemented in the PLS toolbox [10], although the standardization of the two distances is based on a different cutoff value.

4 Simulation results

In this section we compare the different classification rules by means of a small simulation experiment. We consider two high-dimensional samples drawn from a normal population. The first set of 30 observations is simulated from a multivariate normal distribution with mean $(2, 10, \mathbf{0}_{98}')'$ with $\mathbf{0}_{98}' = (0, 0, \ldots, 0)$ and variance-covariance matrix $\mathrm{diag}(5, 3, 0.01 : -0.0001 : 0.0003)$ with 'diag' a diagonal matrix, so $p = 100$. The second set of 50 observations comes from a multivariate normal population with mean $(5, 2, \mathbf{0}_{98}')'$ and variance-covariance matrix $\mathrm{diag}(3, 5, 1, 0.01 : -0.0001 : 0.0004)$. To evaluate the classification rule, a clean test set is constructed consisting of 20 observations (10 observations in each group). We repeated this experiment 100 times and report mean misclassification percentages in the next tables. If we

	RSIMCA				SIMCA			
λ	(R1)	(R2)	(R3)	(R4)	(R1)	(R2)	(R3)	(R4)
0	2.6	2.6	2.7	2.7	2.4	2.4	2.5	2.5
0.3	2.3	2.4	**2.4**	**2.3**	2.2	2.3	**2.0**	**1.9**
0.5	2.1	2.3	2.6	2.5	1.9	2.2	2.1	2.1
0.7	**2.0**	**2.2**	2.7	2.7	**1.7**	**2.0**	2.4	2.4
1	4.5	4.5	5.6	5.6	6.1	6.1	7.8	7.8

Table 1: Misclassification percentages for the test data based on the uncontaminated training data.

construct the classification rules based on an uncontaminated training set, we obtain the results in Table 1 for $\lambda \in \{0, 0.3, 0.5, 0.7, 1\}$. The percentages represent the number of misclassifications divided by the total number of observations in the test data. In bold the lowest values for each classification rule are shown. We first note that there is only a small difference between the results of SIMCA and RSIMCA. Also the difference between the various classification rules is negligible. All the rules with $\lambda = 1$ (only orthogonal distances are considered) are clearly not to be preferred for this simulation setting. However, if we introduce 10% of contamination in the second group of the training data, i.e. 5 observations are replaced by observations from a multivariate normal distribution with mean $(-1, 18, \mathbf{0}_{98})'$ and variance-covariance matrix $0.01\,\mathrm{diag}(3, 5, 1, 0.01 : -0.0001 : 0.0004)$, the misclassifications for the test data increase slightly for RSIMCA, but the results for SIMCA are largely affected as can be seen in Table 2. For the robust results we again see little differences between the different classification rules, although similarly as in the uncontaminated case, $\lambda = 1$ should be discarded.

	RSIMCA				SIMCA			
λ	(R1)	(R2)	(R3)	(R4)	(R1)	(R2)	(R3)	(R4)
0	2.6	2.6	2.9	2.9	18.6	18.6	23.9	23.9
0.3	2.5	2.6	**2.8**	**2.7**	18.7	18.7	29.4	38.4
0.5	**2.2**	2.4	**2.8**	2.8	19.3	18.9	37.5	45.3
0.7	**2.2**	**2.2**	3.6	3.6	19.9	19.5	46.7	48.2
1	8.5	8.5	10.2	10.2	40.4	40.4	49.7	49.7

Table 2: Misclassification percentages for the test data based on the contaminated training data.

5 Example

In this section we will apply the RSIMCA method with the various classification rules on a data set from image analysis and compare the results with the results from SIMCA. The *Image Segmentation* data are available on the UCI Repository [2] and contain information from instances randomly drawn from a database of seven outdoor images. The data can thus be split in seven categories (brickface, sky, foliage, cement, window, path and grass) from which 30 observations are available to obtain a classification rule. For each instance $p = 19$ properties are measured. Also a large test set consisting of 2100 instances is available (300 observations per group). Applying ROBPCA on the data already revealed some outlying samples i.e. observations with a very large orthogonal distance and/or a very large score distance. Only in group 7 we did not detect any. We choose $k_1 = k_2 = k_4 = \ldots = k_7 = 3$ and $k_3 = 2$ based on a robust decision criterion developed in [3]. We then applied the different classification rules to the large test data. The results are summarized in Table 3.

	RSIMCA				SIMCA			
λ	(R1)	(R2)	(R3)	(R4)	(R1)	(R2)	(R3)	(R4)
0	23.7	23.7	22.0	22.0	64.1	64.1	60.7	60.7
0.3	**5.6**	**6.7**	6.6	7.9	13.7	15.7	26.3	22.2
0.5	6.4	7.2	**5.3**	**6.1**	15.2	16.5	16.9	15.9
0.7	7.8	7.7	7.4	6.7	16.4	16.8	13.4	13.3
1	14.7	14.7	16.2	16.2	16.9	16.9	30.3	30.3

Table 3: Misclassification percentages for the image segmentation data.

The calculation of the misclassification percentages is slightly different from the one used in the simulation experiment. Since outlying observations, indicated by the ROBPCA method as observations with an abnormal orthogonal distance or an unusual high score distance, will influence these percentages, we did not take these observations into account while calculating the misclassification percentages. So not all 300 observations of each of the seven sets were considered. To be able to compare the results from RSIMCA with the results of SIMCA, we computed the misclassification percentages of SIMCA on the same set of observations.

We see in Table 3 that the percentages for RSIMCA are much lower than for SIMCA. The differences between the four classification rules are very small, but it is clear that a classification rule that is only based on the score distance ($\lambda = 0$) or the orthogonal distance ($\lambda = 1$) has a very poor performance.

6 Conclusions

We have illustrated that the SIMCA method for classifying high-dimensional observations is affected by outlying samples. A robust method RSIMCA is shown to improve the SIMCA method in the presence of outliers. Some further research is needed to make a more thorough comparison between the difference classification rules and an optimal choice for λ. The robust RSIMCA method that gives equal importance to the orthogonal and score distance ($\lambda = 0.5$) seems a logical and good choice for practical applications. Moreover we will also investigate whether the PLS-DA method [4] can also be robustified using a robust PLS method [7].

References

[1] Beebe K.R., Pell R.J., Seasholtz M.B. (1998). *Chemometrics: A Practical Guide*. Wiley, New York.

[2] Blake C.L., Merz C.J. (1998). *UCI Repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/MLRepository.html]*. Irvine, CA: University of California, Department of Information and Computer Science.

[3] Engelen S., Hubert M. (2004). *Fast cross-validation in robust PCA*. Submitted to the Proceedings of COMPSTAT2004. Available at *http://www.wis.kuleuven.ac.be/stat/robust.html*.

[4] Eriksson L., Johansson E., Kettaneh-Wold N., Wold S. *Multi- and megavariate data analysis – principles and applications; chapter 8: classification and discrimination.* 2001, Umetrics AB.

[5] Hubert M., Rousseeuw P.J., Verboven S. (2002). *A fast robust method for principal components with applications to chemometrics*. Chemometrics and Intelligent Laboratory Systems **60**, 101 − 111.

[6] Hubert M., Rousseeuw P.J., Vanden Branden K. (2004). *ROBPCA: a new approach to robust principal component analysis*. To appear in *Technometrics*. Available at *http://www.wis.kuleuven.ac.be/stat/robust.html*.

[7] Hubert M., Vanden Branden K. (2003). *Robust methods for partial least squares regression*. Journal of Chemometrics **17**, 537 − 549.

[8] Rousseeuw P.J. (1985). *Multivariate estimation with high breakdown point*. In Mathematical Statistics and Applications, (edited by W. Grossmann, G. Pflug, I. Vincze and W. Wertz), Reidel Publishing Company, Dordrecht, 283 − 297.

[9] Sharaf M.A., Illman D.L., Kowalski B.R. (1986). *Chemometrics*. Wiley, New York.

[10] Wise B.M., Gallagher N.B. *PLS_Toolbox 2.1 for use with MATLAB*, manual of the PLS toolbox. Available at *http://www.eigenvector.com*.

[11] Wold S. (1976). *Pattern recognition by means of disjoint principal components models*. Pattern Recognition **8**, 127 − 139.

Address: K.V. Branden, M. Hubert, Katholieke Universiteit Leuven, Department of Mathematics, W. de Croylaan 54, B-3001 Leuven, Belgium

E-mail: {karlien.vandenbranden,mia.hubert}@wis.kuleuven.ac.be

AN ADJUSTED BOXPLOT FOR SKEWED DISTRIBUTIONS

Ellen Vanderviere and Mia Huber

Key words: Boxplot, skewness, medcouple.

COMPSTAT 2004 section: Graphics.

Abstract: The boxplot is a very popular graphical tool to visualize the distribution of continuous univariate data. First of all, it shows information about the location and the spread of the data by means of the median and the interquartile range. The length of the whiskers on both sides of the box and the position of the median within the box are helpful to detect possible skewness in the data. Finally, observations that fall outside the whiskers are pinpointed as outliers, hence the boxplot also includes information from the tails. However, when the data are skewed, usually too many points are classified as outliers. This is because the outlier rule is solely based on measures of location and scale, and the cutoff values are derived from the normal distribution. We present a generalization of the boxplot that includes a robust measure of skewness in the determination of the whiskers. We show with several simulation results that this adjusted boxplot gives a more accurate representation of the data and of possible outliers.

1 Introduction

One of the most frequently used graphical techniques for analyzing a univariate data set is the *boxplot*, proposed by Tukey [6].
If $X_n = \{x_1, x_2, \ldots, x_n\}$ is a univariate data set, the boxplot is constructed by

- putting a line at the height of the sample median Q_2
- drawing a box from the first quartile Q_1 to the third quartile Q_3
- classifying all points outside the interval

  $$[Q_1 - 1.5 \text{ IQR} \; ; \; Q_3 + 1.5 \text{ IQR}] \qquad (\text{with IQR } = Q_3 - Q_1)$$

  as outlier and marking them on the plot
- drawing the whiskers (i.e. the lines that go from the ends of the box to the most remote points that are no outliers)

This construction implies that a boxplot gives information about the location, spread, skewness and tails of the data.

However, in some cases the information about the tails that is given by the boxplot, is not reliable. As was already mentioned in Hoaglin, Mosteller and Tukey [3], too many points are classified as outlier when the data are sampled from a skewed distribution. Consider e.g. the chi-squared distribution with *df* degrees of freedom, which is very skewed for small *df*. Table 1 lists the theoretical lower whisker and the upper whisker for $df = 1, 5$ and 20. The last column gives the probability of a type I error, which we define as the probability to exceed the whiskers, or equivalently, to mark regular observations as outliers. We see that this probability is about 7.6% for χ^2_1 which is very high, and decreases with *df*. For the normal distribution on the other hand, the expected percentage of outliers is only 0.7%.

Distr.	Lower outlier cutoff	Upper outlier cutoff	Total % outliers
χ^2_1	-1.730	3.155	7.58
χ^2_5	-3.252	12.552	2.80
χ^2_{20}	2.888	36.392	1.39
N(0,1)	-2.698	2.698	0.70

Table 1: Theoretical lower and upper outlier cutoff values for several distributions, and the expected percentage of classified outliers according to the boxplot rule.

The large discrepancy between these percentages is caused by the fact that the outlier rule is solely based on measures of location and scale, and the cutoff values are derived from the normal distribution. We present a generalization of the boxplot that includes a robust measure of skewness in the determination of the whiskers. To construct this adjusted boxplot we will derive new outlier rules at the *population* level. To draw the boxplot at a particular data set, we then just need to plug in the finite-sample estimates.

2 Generalization of the boxplot

2.1 A robust measure of skewness

To measure the skewness of a continuous distribution F, we will use the *medcouple*, which we denote as MC. It is defined as

$$\mathrm{MC}(F) = \underset{x_1 < m_F < x_2}{\mathrm{med}} h(x_1, x_2)$$

with x_1 and x_2 sampled independently from F, m_F the median of F and the kernel function h given by

$$h(x_i, x_j) = \frac{(x_j - m_F) - (m_F - x_i)}{x_j - x_i}.$$

From the definition we see that the medcouple always lies between -1 and 1. A distribution that is skewed to the right has a positive value for the medcouple, whereas it becomes negative at a left skewed distribution. Finally, a symmetric distribution has a zero medcouple. As shown in Brys, Hubert and Struyf [1] this robust measure of skewness has a bounded influence function and a breakdown value of 25%. Besides, the MC turned out to be the overall winner when comparing it with two other robust skewness measures which are solely based on quantiles, namely the QS (quartile skewness) and the OS (octile skewness). The MC combines the strengths of OS and QS: it has the sensitivity of OS to detect skewness and the robustness of QS towards outliers. For the computation of the MC, a fast algorithm of $\mathrm{O}(n \log n)$ time has been constructed, and Matlab and S-PLUS functions are available.

2.2 Possible models

We generalize the original boxplot by introducing functions $h_l(\mathrm{MC})$ and $h_r(\mathrm{MC})$ in the cutoff values to classify the outliers. Thus instead of using the interval

$$[Q_1 - 1.5 \text{ IQR} \; ; \; Q_3 + 1.5 \text{ IQR}]$$

for the regular observations, we propose the boundaries of the interval to be defined as

$$[Q_1 - h_l(\mathrm{MC}) \text{ IQR} \; ; \; Q_3 + h_r(\mathrm{MC}) \text{ IQR}].$$

We additionally require that $h_l(0) = h_r(0) = 1.5$ in order to obtain the original boxplot at symmetric distributions. Note that by using different functions h_l and h_r we allow to obtain whiskers of different length. Moreover the boundaries are location and scale equivariant due to the location and scale invariancy of the medcouple.

We studied three different models, which are easy and which do not contain too many parameters, namely a

(1). *linear model*: $h_l(\mathrm{MC}) = 1.5 + a \text{ MC}$, $h_r(\mathrm{MC}) = 1.5 + b \text{ MC}$.

(2). *quadratic model*: $h_l(\mathrm{MC}) = 1.5 + a_1 \text{ MC} + a_2 \text{ MC}^2$,
$h_r(\mathrm{MC}) = 1.5 + b_1 \text{ MC} + b_2 \text{ MC}^2$

(3). *exponential model*: $h_l(\mathrm{MC}) = 1.5 \; e^{a\mathrm{MC}}$, $h_r(\mathrm{MC}) = 1.5 \; e^{b\mathrm{MC}}$.

2.3 Defining the constants

In order to determine the constants in the models mentioned above, we require that the expected percentage of marked outliers is 0.7%, which coincides with the outlier rule of the original boxplot at the normal distribution. If we use for example the linear model, this implies that the constants a and b should satisfy $Q_1 - (1.5 + a \text{ MC}) \text{ IQR} = Q_\alpha$ and $Q_3 + (1.5 + b \text{ MC}) \text{ IQR} = Q_\beta$ where in general Q_p denotes the pth quantile of the distribution, $\alpha = 0.0035$

and $\beta = 0.9965$. The previous system can be rewritten as $\frac{Q_1 - Q_\alpha}{\text{IQR}} - 1.5 = a$ MC, and $\frac{Q_\beta - Q_3}{\text{IQR}} - 1.5 = b$ MC. Linear regression without intercept can then be used to obtain estimates of the parameters a and b. The parameter derivation in case of the quadratic or the exponential model is analogous to that of the linear case. For the exponential model for example, we obtain the linear system

$$\begin{cases} \ln\left(\frac{2}{3}\ \frac{Q_1 - Q_\alpha}{\text{IQR}}\right) = a \text{ MC} \\ \ln\left(\frac{2}{3}\ \frac{Q_\beta - Q_3}{\text{IQR}}\right) = b \text{ MC} \end{cases}$$

To derive the constants we used 12,605 distributions from the family of Γ, χ^2, F, Pareto and G_g-distributions [4]. More precisely, we used $\Gamma(\beta, \gamma)$ distributions with scale parameter $\beta = 0.1$ and shape parameter $\gamma \in [0.1; 10]$, χ^2_{df} distributions with $df \in [1; 30]$, F_{m_1,m_2} distributions with $(m_1, m_2) \in [1; 100] \times [1; 100]$, Pareto distributions $Par(\alpha, c)$ with $c = 1$ and $\alpha \in [0.1; 20]$, and G_g-distributions with $g \in [0; 1]$.

The parameters of the distributions were always selected such that the medcouple did not exceed 0.6. Doing so, we retain a large collection of distributions that are not extremely skewed. It appeared that constructing one good and easy model that also includes the cases with MC > 0.6 is hard to find. Hence, we currently only concentrated on the more common distributions with moderate skewness. Note that we only considered symmetric and right-skewed distributions, as the boundaries just need to be switched for left-skewed distributions. To obtain the population values of the medcouple and the quartiles at all these distributions, we generated 10,000 observations from each of them, and used their finite-sample estimates as the true values.

In Figure 1(a) we show for the parameter b the fitted regression curves, after applying (robust) reweighted LTS regression [5] for each model. Note that the regression results are based on the whole set of distributions we considered. On the vertical axis we have set the response value for the exponential model. Figure 1(b) only displays the G_g distributions (with the same fits superimposed).

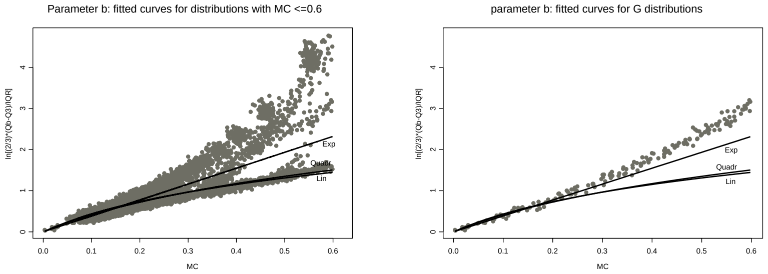


Figure 1: Regression curves for the linear, quadratic and exponential model.

2.4 The adjusted boxplot

From Figure 1 we see that the exponential model is the most appropriate one. Although the fitted line will produce an underestimate of Q_β for some distributions, the same quantile will be overestimated for others. So it gives a good compromise for the whole set of distributions we considered. The linear and quadratic model on the other hand give underestimates for a large group of distributions. For the estimate of the left tail, we have not included the figures because of lack of space. Here it could be seen that the linear model fails completely to estimate Q_α, whereas the exponential and quadratic model perform much better. As the exponential model is appropriate for both the left and the right tail and as it only includes one parameter (on each side), we will use the *exponential model* in the definition of our adjusted boxplot. To make the model easier, we rounded off the estimated values of $a = -3.79$ and $b = 3.87$ to $a = -3.5$ and $b = 4$. We will investigate other possibilities such as $-a = b = 4$ and $-a = b = 3$ in further research.

To summarize, the adjusted boxplot marks the observations that fall outside the interval

$$[Q_1 - 1.5\, e^{-3.5\ \mathrm{MC}}\ \mathrm{IQR}\ ;\ Q_3 + 1.5\, e^{4\ \mathrm{MC}}\ \mathrm{IQR}] \qquad (1)$$

3 Simulation study

To compare our adjusted boxplot with the original one, the percentage of left and right outliers (observations that fall outside the boundaries defined by (1)), together with the total percentage of outliers, is computed at samples of different size n from several distributions. For each distribution, 100 samples of size n were considered. The average percentages of outliers are reported in Table 2. Standard errors are below 0.2% for most of the entries. The superscript $*$ means a standard error between 0.2% and 0.5%, $**$ between 0.5% and 0.9%.

At the normal distribution, we notice that, slightly remarkably, the adjusted boxplot classifies more observations as outliers than before. This is because the finite-sample medcouple is not exactly zero, hence the adjusted whiskers are slightly different from the original ones. At larger sample sizes, we see that the total percentage of outliers is again close to 0.7%.

Much more pronounced differences can be seen at the skewed distributions. At the χ^2_1 distribution for example the average number of marked outliers is less than 0.18%. The adjusted boxplot of the Pareto(3,1) distribution now yields at most 1.23% outliers, as opposed to more than 8% at the original boxplot. We also included two distributions that were not used in the calibration of the exponential model, namely the Pareto(1,3) and the G_3 distribution. Also here, we see that our model highlights much fewer outliers than before.

As we see, the improvements differ somewhat over the distributions. The overall improvement is mainly due to a substantial increase of the right

		original boxplot			adjusted boxplot		
Distr	n	% L	% R	Tot %	% L	% R	Tot %
$N(0,1)$	100	0.600	0.700	1.300	1.180	0.800	1.980*
	500	0.358	0.402	0.760*	0.462*	0.634*	1.096*
	1000	0.335	0.362*	0.697*	0.484*	0.445*	0.929*
χ^2_1	100	0.000	7.350*	7.350*	0.000	0.180	0.180
	500	0.000	7.940**	7.940**	0.000	0.032	0.032
	1000	0.000	7.726**	7.726**	0.000	0.015	0.015
χ^2_{20}	100	0.060	1.360	1.420	0.880	0.780	1.660
	500	0.002	1.478*	1.480*	0.400*	0.392*	0.792*
	1000	0.002	1.456**	1.458**	0.382*	0.311*	0.693*
$\Gamma(0.1, 0.5)$	100	0.000	7.960*	7.960*	0.000	0.410	0.410
	500	0.000	7.716**	7.716**	0.000	0.030	0.030
	1000	0.000	7.708**	7.708**	0.000	0.019	0.019
Pareto(3,1)	100	0.000	8.130*	8.130*	0.280	0.950	1.230
	500	0.000	8.350**	8.350**	0.034	0.620*	0.654*
	1000	0.000	7.943**	7.943**	0.000	0.558*	0.558*
$F(90, 10)$	100	0.000	5.210*	5.210*	1.480*	0.960	2.440*
	500	0.000	5.000**	5.000**	0.584*	0.636*	1.220*
	1000	0.000	5.230**	5.230**	0.485**	0.714*	1.199**
Pareto(1,3)	100	0.000	12.250*	12.250*	0.710*	2.490	3.200*
	500	0.000	12.338**	12.338**	0.000	2.314*	2.314*
	1000	0.000	12.461**	12.461**	0.000	2.166*	2.166*
G_3	100	0.000	16.300*	16.300*	0.000	3.290	3.290
	500	0.000	16.516*	16.516*	0.000	2.966*	2.966*
	1000	0.000	16.408**	16.408**	0.000	3.028**	3.028**

Table 2: For different distributions and samples sizes, the mean percentage of left outliers (% L), right outliers (% R) and the mean total percentage of outliers (Tot %) are reported, resulting from the original boxplot and the adjusted boxplot. The superscript $*$ means a standard error between 0.2% and 0.5%, $**$ between 0.5% and 0.9%. No superscript is set if the standard error is smaller than 0.2%.

whiskers. The lower whiskers are often still somewhat too small, yielding zero percentages of marked outliers. We don't consider the latter as a too serious problem as it is mainly the outlyingness to the right which is of importance at right-skewed distributions. However, to improve the fits, several modifications could be made (we thank a referee for pointing out several of them). The most natural one is to include tail information of the distributions as well. We could for example try to construct a model which includes robust measures of left and right tail, such as those proposed in Brys et al. [2]. We see however several disadvantages of such a procedure. First of all, the model will become more complex with more estimators and parameters. The robustness will decrease as the tail measures have a lower breakdown value, and the variability of the whisker's length will increase, due to the variability of the tail measures.

If we have a priori information of the distribution, for example, we know that it belongs to the class of G_g distributions, it is clear from Figure 1 that a more specific model could be constructed, for example by including results from extreme value theory. Another possibility is to vary the quantiles $\alpha = 0.0035$ and $\beta = 0.9965$ within one (or all) distributions to see if each of its tails is being modeled appropriately. Or different functional forms could be considered for the two tails. This will certainly be considered in our future research.

4 Example

In this section we consider a sample of size $n = 200$ from a G_1-distribution (which is exactly the lognormal distribution) and apply both the original and our adjusted boxplot. As we did not yet implement a graphical representation, we summarize the results as in Figure 2. On the plot with the data versus their index, we have drawn full lines at the median, and at the first and third quartile of the data. Next, we have drawn dashed lines for the original boxplot, and dash-dotted lines at the boundaries of the adjusted boxplot.

We see that by introducing the medcouple in our definition, both the left and the right boundary have shifted upwards. We notice a significant decrease from 10% to 3.5% of right outliers. The lower bound lies much closer to the data points. This might yield more left outliers, but it also better reflects the shorter left tail.

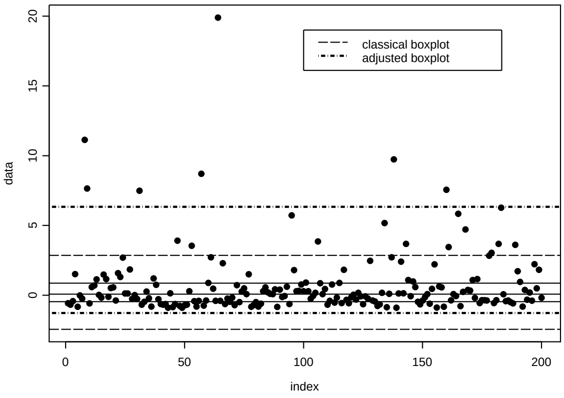


Figure 2: A sample of 200 observations from a lognormal distribution with the boundaries to classify outliers based on the original and the adjusted boxplot.

5 Conclusion

A frequently used graphical tool to analyze a univariate data set is the boxplot. Unfortunately, when drawing this boxplot to a skewed distribution, the tail information is not reliable. Therefore, we have presented a generalization of the boxplot, that takes the skewness factor into account.

To measure skewness of the data, the medcouple has been used and different models for generalizing the original boxplot have been studied. The overall winner seems to be an exponential model. Finally, some simulation results and a graphical representation have been given to indicate the gain of accuracy, achieved by using the adjusted boxplot at skewed distributions.

Note that in this paper we have studied the adjusted boxplot with respect to its type I error, which we have defined as the probability to wrongly declare regular observations as outliers. In the future we will also study its behavior at data sets which contain real outliers.

References

[1] Brys G., Hubert M., Struyf A. (2003). *A robust measure of skewness.* Journal of Computational and Graphical Statistics, to appear.

[2] Brys G., Hubert M., Struyf A. (2004). *Robust measures of tail weight.* Submitted.

[3] Hoaglin D.C., Mosteller F., Tukey J.W. (1983). *Understanding robust and exploratory data analysis.* Wiley, New York.

[4] Hoaglin D.C., Mosteller F., Tukey J.W. (1985). *Exploring data tables, trends and shapes.* Wiley, New York.

[5] Rousseeuw P.J. (1984). *Least median of squares regression.* Journal of the American Statistical Association **79**, 871–880.

[6] Tukey, J.W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Address: E. Vanderviere, University of Antwerp, Department of Mathematics and Computer Science, Middelheimlaan 1, B-2020 Antwerp, Belgium
M. Huber, Katholieke Universiteit Leuven, Department of Mathematics,
W. de Croylaan 54, B-3001 Leuven, Belgium

E-mail: `ellen.vandervieren@ua.ac.be`,
`mia.hubert@wis.kuleuven.ac.be`

MATLAB SOFTWARE FOR ROBUST STATISTICAL METHODS

Sabine Verboven and Mia Hubert

Key words: Outlier detection, Robust classification, Robust regression, MATLAB Toolbox, Calibration.

COMPSTAT 2004 section: Statistical software.

Abstract: Since MATLAB is very popular in industry and academia, and is frequently used by statisticians, chemometricians, chemists, and engineers, we introduce a MATLAB library of robust statistical methods. Those methods were developed because their classical alternatives produce unreliable results when the data set contains outlying observations. The robust calibration toolbox mainly contains implementations of methods that have been developed at our research groups. It currently contains functions for location and scale estimation [15], regression (FAST-LTS [14], MCD-regression [12]), covariance estimation (FAST-MCD [13]), classification (RDA [8]), principal component analysis (RAPCA [5], ROBPCA [6]), principal component regression (RPCR [9]) and partial least squares (RSIMPLS [7]). The toolbox also provides many graphical tools to detect and classify the outliers. By means of an example we show how to use the toolbox for robust estimation and outlier detection.

1 Introduction

The need and effectiveness of robust methods has been described in many paper and books, see e.g. [4], [3], [11]. Robust methods are developed because atypical observations in a data set heavily affect the classical estimates. Outlying values can e.g. occur by mistake (misplacement of a comma), through a malfunction of the machinery, or the production of a “bad” sample.

Over the years, many robust methods are already implemented in two of the leading statistical software packages SAS and S-PLUS. Since MATLAB is well known in industry and academia, we started collecting robust methods in a MATLAB library.

This toolbox mainly contains implementations of methods that have been developed at our research groups. It currently contains functions for location and scale estimation, regression (FAST-LTS, MCD-regression), covariance estimation (FAST-MCD), classification (RDA), principal component analysis (RAPCA, ROBPCA), principal component regression (RPCR) and partial least squares (RSIMPLS). As it includes many functions for calibration, we call it the ‘Matlab Toolbox for Robust Calibration’. The library also provides many graphical tools to detect and classify the outliers. It can be freely downloaded from the websites `http://www.agoras.ua.ac.be` and `http://www.wis.kuleuven.ac.be/stat/robust.html` for non-commercial use.

By means of an example in Section 2 we demonstrate how to work with the toolbox and we explain some of the implemented diagnostic plots to visualize and classify the outliers. Section 3 contains a list of the main functions.

2 Using the toolbox

In this section we demonstrate how to use our toolbox by analyzing the Octane data set [2]. This data set contains 39 NIR-spectra of gasoline measured at 251 wavelengths. It was known that samples 25, 26 and 36-39 contained added alcohol. Since the data set contains more variables than observations, we perform a robust principal component analysis by means of the ROBPCA method [6]. Typing

```
>> out=robpca(X)
```

at the MATLAB command line executes the 'robpca.m' program with its default settings. When we ask

```
>> out=robpca(X,'plots',0,'classic',1)
```

plots are not automatically shown on the screen, but classical PCA is performed as well. Note that we have chosen another way of assigning input arguments than in standard MATLAB functions. Doing so, we create much more flexible input arguments in the function calls. Their order is of no importance, and they can be omitted (which implies that the defaults are used). For example, the commands

```
>> out=robpca(X,'classic',1,'plots',0)
```

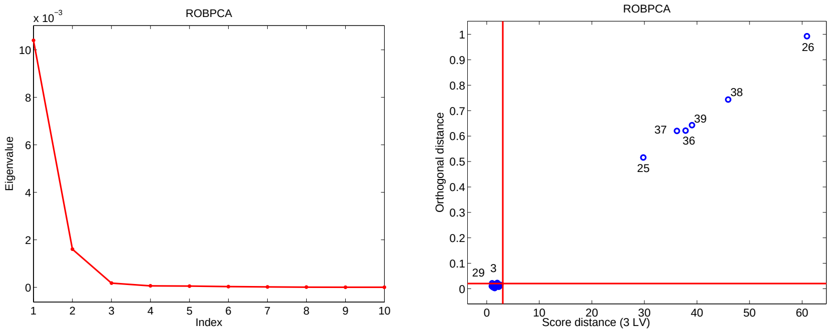
or

```
>> out=robpca(X,'plots',0,'classic',1,'alpha',0.75)
```

would produce the same result. Here, the input argument 'alpha' is a lower bound for the number of non-outliers, and should be between half the sample size $n/2$ and n.

ROBPCA yields a screeplot (Figure 1a) from which we decide to retain $k = 3$ principal components. To identify contaminated samples in the data the output of 'robpca' also lists two robust distances: the score distance (SD) and the orthogonal distance (OD). The SD tells how far away a particular spectrum is from the center of the regular spectra inside the PCA-subspace whereas the OD is the distance of the spectrum to the PCA-subspace. These two distances are summarized in a score outlier map, which displays the OD_i for each observation $i\,(1 \leq i \leq n)$ against its SD_i.

In general, four groups of data points can then be distinguished: regular data (with small SD and small OD), good PCA-leverages points (with large



(a) (b)

Figure 1: Robust PCA analysis of the Octane data set : (a) Screeplot; (b) Score outlier map.

SD and small OD), orthogonal outliers (with small SD and large OD) and bad PCA-leverage points (with large SD and large OD). Details about the horizontal and vertical cutoff values on this outlier map can be found in [6]. On the outlier map of the Octane data in Figure 1 we immediately spot the samples 25, 26, 36-39 as bad PCA-leverage points.

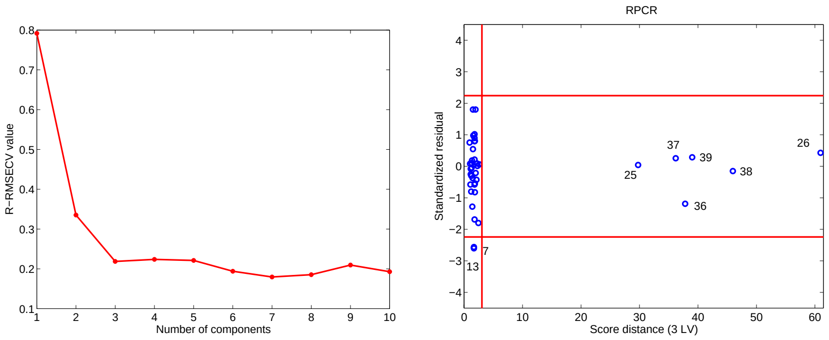
In the next step of the analysis we would like to predict the octane number y from the spectra. A regression model thus has to be formulated. Since here the number of variables p is larger than the number of observations n, there is no unique solution to the least squares problem. Moreover, in such a high dimensional data set the variables are necessarily intercorrelated, which is called multicollinearity.

Two popular regression methods to analyze this type of high-dimensional data are Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). They both find estimates for the regression parameters by first projecting the x-variables (here, the spectra) onto a lower dimensional space, yielding scores t_i. The response variable is then regressed onto these scores using linear regression.

Robust alternatives for classical PCR and PLS have recently been developed [9], [7]. Here we will concentrate on RPCR. First, ROBPCA is applied to the x-variables, and in the second stage, the robust LTS regression method [10] is performed.

To select the optimal number of principal components, the cross-validated Root Mean Squared Error (RMSECV) can be computed at the model with $k = 1, \ldots, k_{\max}$ components. For a formal definition, see [7]. This is a rather time-consuming approach, but faster methods for its computation have been developed and will become part of the toolbox [1]. The function call in MATLAB then becomes:

```
>> out=rpcr(X,y,'rmsecv',1)
```



(a) (b)

Figure 2: Analysis of the Octane data set: (a) robust RMSECV curve; (b) regression outlier map.

On the robust RMSECV-curve of Figure 2(a), we see that the first important valley is attained at $k = 3$, hence we select 3 components. This corresponds with our decision based on the screeplot (Figure 1a) from the ROBPCA analysis.

The output of the RPCR function also includes the score and orthogonal distances from the robust PCA analysis. So by default the PCA-outlier map will again be plotted (Figure 1b). In addition, a residual outlier map is generated. It displays the standardized LTS residuals versus the score distances, and can be used to classify the observations according to the regression model. We distinguish regular data (small SD and small residual), good leverage points (large SD but small residual), vertical outliers (small SD and large residual), and bad leverage points (large SD and large residuals). The vertical outliers and bad leverage points are the most harmful for classical least squares regression as they disturb the linear relationship.

Figure 2(b) shows de regression outlier map from the octane data set. We see that the outlying spectra 25, 26, and 36-39 all belong to the group of the good leverage points. This can be explained by the fact that the added alcohol also raises the octane number in those samples and therefore produces regular samples which only lie further away from the majority. Cases 7 and 13 are vertical outliers but as their standardized residual is not too far from the cutoff value, they can be considered as borderline cases.

Note that the RPCR output is a structure which includes a field called ‘flag’. This is a binary vector that indicates for each observation whether or not it exceeds the cutoff values on the vertical axis on either the PCA or the regression outlier map. Regular observations, good PCA and good regression leverage points receive a flag 1, while the others obtain a flag equal to zero.

3 Toolbox content

Until now the toolbox contains about fifty m-files. Some of the MATLAB functions only serve as help-functions. The most important functions are listed below. Appropriate references are included in the help-files of the programs.

(1). Robust estimators of location/scale/skewness.

- mlochuber - M-estimator of location, with Huber psi-function.
- mloclogist - M-estimator of location, with logistic psi-function.
- hl - Hodges-Lehmann location estimator.
- mad - Median absolute deviation with finite sample correction factor (scale).
- mscalelogist - M-estimator of scale, with logistic psi-function.
- qn - Q_n-estimator of scale. Available as DLL or M-file.
- mc - Medcouple, a robust estimator of skewness. Available as DLL.

(2). Robust multivariate analysis.

- l1median - L1 median of multivariate location.
- mcdcov - MCD estimator of multivariate location and covariance.
- rapca - Robust principal component analysis (based on projection pursuit).
- robpca - Robust principal component analysis (based on projection pursuit and MCD estimation).
- rda - Robust linear and quadratic discriminant analysis.
- robstd - Columnwise robust standardization.

(3). Robust regression methods.

- ltsregres - Least Trimmed Squares Regression.
- mcdregres - Multivariate MCD regression.
- rpcr - Robust principal component regression.
- rsimpls - Robust partial least squares regression.

(4). Classical multivariate analysis and regression.

- classSVD - Singular value decomposition if $n > p$.
- kernelEVD - Singular value decomposition if $n < p$.
- mlr - Multivariate multiple linear regression.
- cpca - Classical principal component analysis.
- cpcr - Classical principal component regression.
- csimpls - Partial least squares regression (SIMPLS).
- cda - Classical linear and quadratic discriminant analysis.

References

[1] Engelen S., Hubert M. (2004). *Fast cross-validation in robust PCA.* Proceedings of COMPSTAT 2004, edited by J. Antoch, Springer, Physica-Verlag.

[2] Esbensen K.H. (2001). *Multivariate data analysis in practice.* Camo Process AS (5th edition).

[3] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust statistics: the approach based on influence functions.* Wiley, New York.

[4] Huber P.J. (1981). *Robust statistics.* Wiley, New York.

[5] Hubert M., Rousseeuw P.J., Verboven S. (2002). *A fast robust method for principal components with applications to chemometrics.* Chemometrics and Intelligent Laboratory Systems **60**, 101 – 111.

[6] Hubert M., Rousseeuw P.J., Vanden Branden K. (2004). *ROBPCA: a new approach to robust principal component analysis.* To appear in Technometrics. Available at *http://www.wis.kuleuven.ac.be/stat/robust.html*.

[7] Hubert M., Vanden Branden K. (2003). *Robust methods for partial least squares regression.* Journal of Chemometrics **17**, 537 – 549.

[8] Hubert M., Van Driessen K. (2004). *Fast and robust discriminant analysis.* Computational Statistics and Data Analysis **45**, 301 – 320.

[9] Hubert M., Verboven S. (2003). *A robust PCR method for high-dimensional regressors.* Journal of Chemometrics **17**, 438 – 452.

[10] Rousseeuw P.J. (1984). *Least median of squares regression.* Journal of the American Statistical Association **79**, 891 – 880.

[11] Rousseeuw P.J., Leroy A. (1987). *Robust regression and outlier detection.* John Wiley, New York.

[12] Rousseeuw P.J., Van Aelst S., Van Driessen K., Agulló A. (2004). *Robust multivariate regression.* To appear in Technometrics. Available at *http://www.agoras.ua.ac.be.*

[13] Rousseeuw P.J., K. Van Driessen (1999). *A fast algorithm for the minimum covariance determinant estimator.* Technometrics **41**, 212 – 223.

[14] Rousseeuw P.J., Van Driessen K. (2000). *An algorithm for positive-breakdown methods based on concentration steps.* In Data Analysis: Scientific Modeling and Practical Application, (W. Gaul, O. Opitz, and M. Schader, eds.), Springer-Verlag, New York, 335 – 346.

[15] Rousseeuw P.J., Verboven S. (2002). *Robust estimation in very small samples.* Computational Statistics and Data Analysis **40**, 741 – 758.

Address: S. Verboven, University of Antwerp, Department of Mathematics and Computer Science, Middelheimlaan 1, B-2020 Antwerpen, Belgium
M. Huber, Katholieke Universiteit Leuven, Department of Mathematics, W. de Croylaan 54, B-3001 Leuven, Belgium

E-mail: `sabine.verboven@ua.ac.be, mia.hubert@wis.kuleuven.ac.be`

COMPSTAT'2004 Symposium

ROBUSTIFYING INSTRUMENTAL VARIABLES

Jan Ámos Víšek

Key words: Robust regression, orthogonality condition, controllable level of robustness.

COMPSTAT 2004 section: Robustness.

Abstract: A modification of the Method of Instrumental Variables for the Least Weighted Squares is proposed. It appears that all solutions of the corresponding normal equations are bounded in probability (so the first step for proving consistency is passed).

1 Introduction of basic framework

Let us consider the linear regression model

$$Y_i = X_i^T \beta^0 + e_i, \quad i = 1, 2, \dots, n. \tag{1}$$

We shall assume that:
C1 *The sequence $\left\{ (X_i^T, e_i)^T \right\}_{i=1}^{\infty} \subset R^{p+1}$ is sequence of independent and identically distributed random variables (i.i.d. r.v.'s) with absolutely continuous distribution function $F_{X,e}(x,v)$. Moreover, the existence of second moments is assumed, the density $f_{X,e}(x,v)$ is bounded, say by U, and the marginal d.f. $F_X(x)$ of vectors X_i's have a bounded support, i.e. putting $M = \sup\{\|x\| \ : \ f_X(x) > 0\}$, we have $M < \infty$.*

Remark 1.1. *We shall consider the model with intercept, i.e. we shall assume that the first coordinate of X_i is degenerated and equal to 1. Without loss of generality, we may then assume that $\mathbb{E} X_{ij} = 0$ for $j = 2, 3, \dots, p$. Notice please that we have not assumed independence between the explanatory variables X_i's and the disturbances e_i's.*

Prior to continuing in the explanation let us give basic notations. The set of all positive integers will be denoted by N and p-dimensional Euclidean space by R^p. For any $\beta \in R^p$ $\quad r_i(\beta) = Y_i - X_i^T \beta$ denotes the i-th residual and $r_{(h)}^2(\beta)$ the h-th order statistic among the squared residuals, i.e. we have

$$r_{(1)}^2(\beta) \le r_{(2)}^2(\beta) \le \ \dots \ \le r_{(n)}^2(\beta). \tag{2}$$

Without loss of generality we may assume that $\beta^0 = 0$ (otherwise we should write in what follows $\beta - \beta^0$ instead of β).

2 Reasons for instrumental variables

It is well known that in the case that the orthogonality condition $\mathbb{E}\left\{e_i|X_i\right\} = 0$ is broken, the ordinary least squares are not consistent due to the fact that

$$\hat{\beta}^{(OLS,n)} = \beta^0 + \left(\frac{1}{n}\sum_{k=1}^{n} X_k X_k^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i e_i$$

and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i e_i \neq 0 \quad \text{in probability.}$$

The best known example of the situation when the orthogonality condition fails, is the model assuming that the explanatory variables are measured with random error. Another example considers the lagged response variable as explanatory one, see Judge et al. [11] or Víšek [21].

The problem is usually treated by means of the *Method of Instrumental Variables* (definition given below). In nineties the method became a standard tool in many case studies of dynamic regression model since the correlation of explanatory variables and disturbances frequently appeared. Moreover, many papers considering possibilities how to select the instruments for explanatory variables brought applicable results (including also easy available implementations), see e.g. Arellano, Bond [1], Arellano, Bover [2] or Sargan [16] (and for examples of implementation - TSP or Stata).

3 Why the least weighted squares

In 1992, Hettmansperger and Sheather, utilizing Engine Knock Data [13] demonstrated that a small change of data may cause a large change of the *Least Median of Squares* estimator [14]. Later their (at the first glance) surprising result appeared to be due to the software, they used. In [18] the application of software based on the simplex method [4] corrected the result. Nevertheless, employing Engine Knock Data and evaluating the *Least Trimmed Squares* [9] by the algorithm which gave precise value of the estimator (due to the small number of observations – 16 cases – it was possible to apply the *Ordinary Least Squares* on the all subsamples of size 11), we have confirmed that a small change of data can really cause a large change of the estimate [18]. In Víšek [20] and [23] instructive academic examples demonstrated why an arbitrarily small shift of one observation may change the estimator as much as you want. It means: ***Robust, especially high breakdown point estimators can be very sensitive to a very small change of data.***

On the other hand, Víšek [17], [19] and [28] revealed that for the M-estimator with discontinuous ψ-function, deletion of even one observation may cause very large change of the estimate. Víšek [24] established the same result for the *Least Trimmed Squares*. It means: ***The sensitivity of robust estimators with respect to the deletion of even one point from data can be very high*** (see also Chatterjee, Hadi [6]).

Both these unpleasant features of some robust estimators have had one denominator, namely that the estimators relay "to much" on a group of observations (considering them to be "clean" or "proper", as you want) while the others are assumed to be contamination. However, it can be in some situation an advantage, see e.g. Víšek [22] or Benáček, Víšek [3] where this "feature" of LTS allowed to decompose the Czech economy on two samples of industries, one already market-economy oriented, another still oriented on a planned economy and hence relaying on a help of state.

Taking all these circumstances into account we offer a proposal of the *Least Weighted Squares* in the form (Víšek [25], see also [26], [27]

$$\hat{\beta}^{(LWS,n,w)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r_{(i)}^2(\beta) \tag{3}$$

where w is a weight function with following properties[1]:

C2 *Weight function $w : [0,1] \to [0,1]$ is absolutely continuous and nonincreasing, with the derivative $w'(\alpha)$ bounded from below by L, $w(0) = 1$.*

For any $i \in \{1, 2, \ldots, n\}$ let us put $\pi(\beta, i) = j \in \{1, 2, \ldots, n\}$ so that $r_i^2(\beta) = r_{(j)}^2(\beta)$ (notice that $\pi(\beta, i)$ is r.v.). Then we have

$$\hat{\beta}^{(LWS,n,w)} = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) r_i^2(\beta). \tag{4}$$

Finally, let for any $n \in N$ $\mathcal{P}_n$ be the set of all permutations of the indices $\{1, 2, \ldots, n\}$ and denote π_i the i-th coordinate of the vector $\pi \in \mathcal{P}_n$. Denoting $\pi(\beta) = (\pi(\beta,1), \pi(\beta,2), \ldots, \pi(\beta,n))^T$, we have $\pi(\beta) \in \mathcal{P}_n$ (keep in mind that $\pi(\beta)$ is however r.v.). Now, taking into account **C2**, (3) and (4), we conclude that for any $\pi \in \mathcal{P}_n$

$$\sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) r_i^2(\hat{\beta}^{(LWS,n,w)}) \leq \sum_{i=1}^{n} w\left(\frac{\pi_i-1}{n}\right) r_i^2(\hat{\beta}^{(LWS,n,w)})$$

so that for any $\omega \in \Omega$ there is some $\pi \in \mathcal{P}_n$ such that for the vector of weights $w^* = (w(n^{-1}(\pi_1 - 1)), w(n^{-1}(\pi_2 - 1)), \ldots, w(n^{-1}(\pi_n - 1)))$ we have $\hat{\beta}^{(LWS,n,w)} = \hat{\beta}^{(WLS,n,w^*)}$, i.e. (in words) the *Least Weighted Squares* estimator is equal to the (classical) *Weighted Least Squares* estimator (with the weights w^*) at given, fixed $\omega \in \Omega$. Since $\hat{\beta}^{(WLS,n,w^*)}$ is the solution of corresponding normal equations, considering successively all $\omega \in \Omega$, we verify that $\hat{\beta}^{(LWS,n,w)}$ is one of solutions of *normal equations*

$$I\!N\!E_{X,n}(\beta) = \sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) X_i \left(Y_i - X_i^T\beta\right) = 0. \tag{5}$$

[1] See also Čížek [7] where the estimator is called the *Smoothed Least Trimmed Squares*.

4 Instrumental variables for the least weighted squares

The inconsistency of the *Least Squares* which is due to the failure of the orthogonality condition (as we recalled it in INTRODUCTION), takes place generally also for the *Least Weighted Squares*. That is why we define an estimator which will be an analogy of the estimator obtained by the *Method of Instrumental Variables* but which will weight down the residuals of those observations which seem to be atypical.

Definition 4.1. *For any sequence of random vectors $\{Z_i\}_{i=1}^{\infty} \subset R^p$ the solution(s) of the (vector) equation*

$$I\!N E_{Z,n}(\beta) = \sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) Z_i\left(Y_i - X_i^T\beta\right) = 0 \tag{6}$$

will be called the Instrumental Least Weighted Squares estimator and denoted by $\hat{\beta}^{(ILWS,n,w)}$.

For any $\beta \in R^p$ the distribution of the absolute value of residual will be denoted $F_\beta(v)$. In other words,

$$F_\beta(v) = P(|Y_1 - X_1^T\beta| < v) = P(|e_1 - X_1^T\beta| < v). \tag{7}$$

(remember, we have assumed $\beta^0 = 0$). Similarly, for any $\beta \in R^p$ the empirical distribution of the absolute value of residual will be denoted $F_\beta^{(n)}(v)$. It means that, denoting the indicator of a set A by $I\{A\}$, we have

$$F_\beta^{(n)}(v) = \frac{1}{n}\sum_{j=1}^{n} I\{|r_j(\beta)| < v\} = \frac{1}{n}\sum_{j=1}^{n} I\left\{|e_j - X_j^T\beta| < v\right\}. \tag{8}$$

It is straightforward that then

$$F_\beta^{(n)}(|r_i(\beta)|) = \frac{\pi(\beta,i)-1}{n}$$

and so (6) can be written as

$$\sum_{i=1}^{n} w\left(F_\beta^{(n)}(|r_i(\beta)|)\right) Z_i\left(Y_i - X_i^T\beta\right) = 0. \tag{9}$$

The classical regression analysis, to be able to prove consistency of the estimator obtained by the *Method of Instrumental Variables*[2], accepted assumption that $I\!E Z_1 X_1^T$ is regular [5] or [11] and, of course, $I\!E\{e_1|Z_1\} = 0$. It corresponds to the assumption which we adopt for the *Ordinary Least Squares*, namely that the matrix $I\!E X_1 X_1^T$ is regular (and hence positive definite) and $I\!E\{e_1|X_1\} = 0$. Now, transforming the variables so that we put $\tilde{X}_{11} = X_{11}$ and for any $j = 2, 3, \ldots, p$

$$\tilde{X}_{1j} = X_{1j} - \sum_{k=1}^{j-1} \lambda_{jk}\tilde{X}_{1k}$$

[2] The estimator is defined as in (6) but with the weight function $w \equiv 1$.

where λ_{jk} are selected so that $\mathrm{cov}(\tilde{X}_{1j}, \tilde{X}_{1k}) = 0$ for $j \neq k$, we have the matrix $I\!E\tilde{X}_1\tilde{X}_1^T$ diagonal and the model for transformed data, namely $Y_i = \tilde{X}_i^T\tilde{\beta} + u_i$ has the same "explanatory" abilities as (1). New explanatory variables $\left\{\tilde{X}_i\right\}_{i=1}^{\infty}$ would not allow presumably so direct (physical, biological, economic etc.) interpretation, nevertheless they have also at least one advantage. The signs of the estimates of the regression coefficients really indicate the positive (negative) influence of given explanatory variable on the response one.

Let us make following, a bit academic considerations. Assuming that we shall look for a sequence of instrumental variables $\left\{\tilde{Z}_i\right\}_{i=1}^{\infty}$ for the sequence of transformed explanatory variables $\left\{\tilde{X}_i\right\}_{i=1}^{\infty}$, we would like to find it so that also $I\!E\tilde{Z}_1\tilde{X}_1^T$ is regular and diagonal. Then of course $\tilde{Z}_{1j}$ is correlated only with $\tilde{X}_{1j}$ and hence we may assume that $I\!E\tilde{Z}_{1j}\tilde{X}_{1j} > 0$ (otherwise we take instead of $\tilde{Z}_{1j}$ the instrumental variable -$\tilde{Z}_{1j}$). Then however $I\!E\tilde{Z}_1\tilde{X}_1^T$ is positive definite. Hence in what follows we shall assume:

C3 *The instrumental variables $\{Z_i\}_{i=1}^{\infty} \subset R^p$ are independent and identically distributed with distribution function $F_Z(z)$. Moreover, they are independent from the sequence $\{e_i\}_{i=1}^{\infty}$. Finally, $I\!EZ_1X_1^T$ is positive definite*[3] *and for any $\beta \in R^p$*

$$\beta^T \left[\int w(F_e(r + X_1^T\beta) - F_e(-r + X_1^T\beta))zx^T \mathrm{d}P(r,z,x)\right]\beta \\ \geq \beta^T\left[\int w(F_e(r) - F_e(-r))zx^T \mathrm{d}P(r,z,x)\right]\beta.$$

Then we can prove:

Lemma 4.1. *Let the conditions* **C1**, **C2** *and* **C3** *be fulfilled. Then any sequence $\left\{\hat{\beta}^{(ILWS,n,w)}\right\}_{n=1}^{\infty}$ of the solutions of normal equations $I\!N E_{Z,n}$ $(\hat{\beta}^{(ILWS,n,w)}) = 0$ is bounded in probability* [4].

Remark 4.1. *The fact that for any i and any $\omega \in \Omega$ the matrix $X_iX_i^T$ is positive semidefinite allows to prove the same assertion (i.e. that all solutions of the normal equations are bounded in probability) for the Least Weighted Squares in significantly simpler way, see Mašíček [12].*

5 Concluding remarks

We have added a small pebble (of mosaic) to equip the *Least Weighted Squares* by additional (or alternative, if you want) methods (similarly as the classical

[3] Compare C3 with Víšek [21] where we considered instrumental M-estimators and the discussion of assumptions for M-instrumental variables was given.

[4] The proof of the lemma is included in the *"large"* version of paper which can be obtained on request

(Ordinary) Least Squares are equipped) to be able to cope with situations when the basic assumptions are broken or when the "main" method is not suitable[5]. The lack of such tools and of course of easy available and reliable implemantations of robust methods hamper a wide (or at least wider than the present) employment of robust methods. On the other hand, not using robust methods along with the classical ones we take a risk of obtaining misleading results of case studies under presence of even slight contamination.

Appendix

Theorem 5.1. *(Glivenko [8]) Let $F(v)$ and $F^{(n)}(v)$ be a d.f. and corresponding empirical d.f. of a sequence of i.i.d. r.v.'s, respectively. Put $D_n = \sup_v \left|F^{(n)}(v) - F(v)\right|$, then*

$$P(\lim_{n\to\infty} D_n = 0) = 1.$$

Lemma 5.1. [6]
For any $\lambda > 0$, $\varepsilon > 0$ and $\delta > 0$ there is $n_0 \in N$ so that for any $n > n_0$ (for $F_\beta^{(n)}(v)$ and $F_\beta(v)$ see (7) and (8))

$$P\left(\left\{\omega \in \Omega : \sup_{\|\beta\|=\lambda} \sup_v \left|F_\beta^{(n)}(v) - F_\beta(v)\right| \le \delta\right\}\right) > 1 - \varepsilon.$$

Theorem 5.2. [7]
Recalling that we have denoted by $\mathcal{P}_n$ the set of all permutations of the indices $\{1, 2, \ldots, n\}$, we have for any $\beta \in R^p$ and $\pi \in \mathcal{P}_n$

$$P\left(\{\pi(\beta,1), \pi(\beta,2), \ldots, \pi(\beta,n)\} = \pi\right) = \frac{1}{n!}.$$

Remark 5.1. *The previous assertion says that, due to the fact that $r_i^2(\beta)$, $i = 1, 2, \ldots, n$ (for any $\beta \in R^p$) represent sequence of i.i.d. r.v.'s, any permutation of indices in (2) has the same probability. Hence mimicking the proof of assertion we obtain:*

Corollary 5.1. *For any $\beta \in R^p$, any $\pi \in \mathcal{P}_n$ and any set $C_n \subset \Omega$ of positive probability which is "permutation-free" with respect $\left\{(X_i^T, e_i)\right\}_{i=1}^\infty$ and $\left\{(Z_i^T)\right\}_{i=1}^\infty$ (i.e. which is defined by means of r.v.'s X_i, Z_i, e_i's but does not depend on the order of them in the sequence $\left\{(X_i, e_i)^T\right\}_{i=1}^\infty$) we have*

$$P(\{\pi(\beta,1), \pi(\beta,2), \ldots, \pi(\beta,n)\}|C_n) = \frac{1}{P(C_n)\cdot n!}.$$

References

[1] Arellano M., Bond S. (1991). *Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations.* Review of Economic Studies **58**, 277–297.

[5] For the latter case let us mention the situation of discrete response variable.

[6] The proof of the lemma is also included in the *"large"* version of paper which can be obtained on request.

[7] The same as in 7.

[2] Arellano M., Bover O. (1995). *Another look at the instrumental variables estimation of error components models.* Journal of Econometrics **68** (1), 29 – 52.

[3] Benáček V., Víšek J.A. (2003). *Determining factors of trade specialization and growth of a small economy in transition. Impact of the EU opening-up on Czech exports and imports.* IIASA, Austria, IR series, no. IR-03-001, 1 – 41.

[4] Boček P., Lachout P. (1993). *Linear programming approach to LMS-estimation.* Memorial volume of Comput. Statist. & Data Analysis **19** (1995), 129 – 134.

[5] Bowden R.J., Turkington D.A. (1984). *Instrumental variables.* Cambridge: Cambridge University Press.

[6] Chatterjee S., Hadi A.S. (1988). *Sensitivity analysis in linear regression.* New York: J. Wiley & Sons.

[7] Čížek P. (2002). *Robust estimation with discrete explanatory variables.* COMPSTAT 2002, Berlin, 509 – 514.

[8] Glivenko V.I. (1933). *Sulla determinazione empirica delle leggi di probabilita.* Giorn. Ist.Ital. Attuari **4**, (92).

[9] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986). *Robust statistics – the approach based on influence functions.* New York: J.Wiley & Sons.

[10] Hettmansperger T.P., Sheather S.J. (1992). *A cautionary note on the method of least median squares.* The American Statistician **46**, 79 – 83.

[11] Judge G.G., Griffiths W.E., Hill R.C., Lutkepohl H., Lee T.C. (1985). *The theory and practice of econometrics.* New York: J.Wiley & Sons(second edition).

[12] Mašíček L. (2003). *Diagnostika a sensitivita robustních odhadů.* (Diagnostics and sensitivity of robust estimators, in Czech.) Disertační práce.

[13] Mason R.L., Gunst R.F., Hess J.L. (1989). *Statistical design and analysis of experiments.* New York: J.Wiley & Sons.

[14] Rousseeuw P.J. (1984). *Least median of square regression.* Journal of Amer. Statist. Association **79**, 871 – 880.

[15] Rousseeuw P.J., Leroy A.M. (1987). *Robust regression and outlier detection.* New York: J.Wiley & Sons.

[16] Sargan J.D. (1988). *Testing for misspecification after estimating using instrumental variables.* In Massouumi E. (ed.) Contribution to Econometrics: John Denis Sargan **1**, Cambridge University Press.

[17] Víšek J.Á. (1992). *Stability of regression model estimates with respect to subsamples.* Computational Statistics **7**, 183 – 203.

[18] Víšek J.Á. (1994). *A cautionary note on the method of the Least Median of Squares reconsidered.* Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, 254 – 259.

[19] Víšek J.Á. (1996). *Sensitivity analysis of M-estimates.* Annals of the Institute of Statistical Mathematics **48**, 469 – 495.

[20] Víšek J.Á. (1996). *On high breakdown point estimation.* Computational Statistics **11**, Berlin, 137 – 146.

[21] Víšek J.Á. (1998). *Robust instruments.* Robust'98 J. Antoch & G. Dohnal (eds), published by Union of Czechoslovak Mathematicians and Physicists, 195 – 224.

[22] Víšek J.Á. (1999). *The robust regression and the experiences from its application on estimation of parameters in a dual economy.* Proceedings of the Conference "Macromodels'99", Poland 1999, ISBN 83-86840-95-1, 424 – 445.

[23] Víšek J.Á. (2000). *On the diversity of estimates.* Computational Statistics & Data Analysis **34**, 67 – 89.

[24] Víšek J.Á. (2000). *A new paradigm of point estimation.* Proceedings of Data Analysis 2000/II, Modern Statistical Methods - Modelling, Regression, Classification and Data Mining, ISBN 80-238-6590-0, 195 - 230.

[25] Víšek J.Á. (2000). *Regression with high breakdown point.* Robust 2000 J. Antoch & G. Dohnal (eds), published by Union of Czechoslovak Mathematicians and Physicists), ISBN 80-7015-792-5, 324 - 356.

[26] Víšek J.Á. (2002). *The least weighted squares I. The asymptotic linearity of normal equations.* Bulletin of the Czech Econometric Society **9** (15), 31 – 58.

[27] Víšek J.Á. (2002). *The least weighted squares II. Consistency and asymptotic normality.* Bulletin of the Czech Econometric Society **9** (16), 1 – 28.

[28] Víšek J.Á. (2002). *Sensitivity analysis of M-estimates of nonlinear regression model: Influence of data subsets.* Annals of the Institute of Statistical Mathematics **54**, 261 – 290.

Acknowledgement: Research was supported by grant of GA ČR number 402/03/0084.

Address: J.Á. Víšek, Faculty of Social Sciences, Charles University & Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Smetanovo nábřeží 6, 110 01 Praha 1, the Czech Republic

E-mail: `visek@mbox.fsv.cuni.cz`

SIMULTANEOUS OPTIMIZATION OF SELECTION-MASTERY DECISIONS

Hans J. Vos

Key words: Bayesian decision theory, simultaneous optimization, monotonicity conditions.

COMPSTAT 2004 section: Bayesian methods.

Abstract: The purpose of this paper is to show that multiple decisions in networks can also be optimized *simultaneously* using the framework of Bayesian decision theory.

1 Introduction

Combinations of elementary test-based decisions in psychology and education arise when one decision problem leads to another, which, in turn, may lead to a third one. An example is test-based decision making in computer-aided instruction (CAI) in education, for instance, a selection decision for a treament followed by a mastery decision. Such systems can be described as instructional networks with individual routes for the students. The question is raised how such networks of decisions should be optimized. An obvious approach is to address each decision separately, optimizing its decision rule on the basis of test data exclusively gathered for this individual decision. This approach is common in current design of instructional systems.

The advantages of a simultaneous approach are twofold. First, data gathered earlier in the network can be used to optimize later decisions. Second, the option is now available to define utility functions (one of the basic elements of Bayesian decision making) on the ultimate success criterion in the complete network instead of on intermediate criteria measuring the success on individual treatments.

2 The selection-mastery problem

An example of the selection-mastery problem is an instructional module with a pretest and a posttest. The pretest is administered to select students for the module. It is assumed that the possible actions are to admit or to reject the student for the module. The posttest is used to decide whether or not the students have mastered the objectives of the module. Typically, the posttest is not perfectly reliable, and the criterion is supposed to be a threshold on the true score underlying the test. The possible actions are to classify a student as a master or a nonmaster.

The following notation is needed. For a randomly sampled student, the observed scores on the selection and mastery tests (i.e., selection and mastery test) are continuous random variables denoted by X and Y, with realizations x and y, respectively. The criterion considered is the classical test theory true score [2] underlying the mastery test. For a randomly sampled student, the true score is denoted by a continuous random variable T with realization t. It is assumed that the standard denoting true mastery is a threshold value t_c on T. Further, it will be assumed that the relation between X, Y, and T can be represented by a joint density function $f(x, y, t)$.

3 Simultaneous decision rules

Let each of the actions denoted by a_{ij} (i, j = 0,1), where i = 0,1 stand for the actions of rejecting and accepting a student and j = 0,1 for the actions of retaining and advancing an accepted student. Since for a rejected student no further mastery decisions are made, the index j will be dropped for $i = 0$. Generally, a decision rule specifies for each possible realization (x, y) of (X, Y) which action a_{ij} is to be taken.

3.1 Weak and strong rules

The decision rule for the mastery decision may or may not depend on the score X on the selection test. Intuitively, one would expect a more lenient mastery rule for a student with high performance on the selection test because this prior information implies that a possible low score on the mastery test is more likely due to measurement error than to a true low performance. Simultaneous rules in which decisions are a function both of the current test score and previous test scores will be called weak rules in this paper. If decisions are only a function of current test scores, the rules will be called strong (simultaneous) rules.

For our selection-mastery network a weak simultaneous rule δ can be defined as:

$$\begin{aligned}
\{(x, y) : \delta(x, y) = a_0\} &= A \times R \\
\{(x, y) : \delta(x, y) = a_{10}\} &= A^C \times B(x) \\
\{(x, y) : \delta(x, y) = a_{11}\} &= A^C \times B^C(x),
\end{aligned}$$

where A and A^C are the sets of x values for which a student is rejected or admitted for the module, and $B(x)$ and $B^C(x)$ are the sets of y values for which a student fails or passes the mastery test. R represents the set of real numbers. With strong rules, the sets $B(x)$ and $B^C(x)$ are independent of x.

3.2 Monotone and nonmonotone rules

Decision rules can take a monotone or a nonmonotone form. A decision rule is monotone if cutting scores are used to partition the sample space into regions for which different actions are taken. All other possible rules are nonmonotone. It will be shown later on under what conditions optimal rules will be monotone for our example.

For our decision problem, a weak monotone rule δ can be defined as:

$$\delta(X,Y) = \begin{cases} a_0 & \text{for } X < x_c \text{ and } Y \in R \\ a_{10} & \text{for } X \geq x_c \text{ and } Y < y_c(x) \\ a_{11} & \text{for } X \geq x_c \text{ and } Y \geq y_c(x), \end{cases}$$

with $y_c(x)$ being the cutting score on Y.

4 Additive utility structure

Generally speaking, a utility function $u_{ij}(t)$ evaluates the total costs and benefits of all possible decision outcomes for a student whose true score is t. Here, the utilities involved in the combined decision problem are defined as the following additive structure:

$$u_{ij}(t) = w_1 u_i^{(s)}(t) + w_2 u_j^{(m)}(t),$$

where $u_i^{(s)}(t)$ and $u_j^{(m)}(t)$ represent the utility functions for the separate selection and mastery decisions, respectively, and w_1 and w_2 are nonnegative weights. Since utility is supposed to be measured on an interval scale, the weights can always be rescaled as follows:

$$u_{ij}(t) = w u_i^{(s)}(t) + (1-w) u_j^{(m)}(t),$$

where $0 \leq w \leq 1$.

Since no mastery decisions are made for rejected students, it is assumed that such students do not contribute to the utility. Hence, it follows that $u_{0j}(t)$ is equal to $wu_0^{(s)}(t)$ for all j.

It should be noted that the first term in the right-hand side of the above expression is a function of t and not, for example, of a true score underlying X. This fact illustrates one of the advantages of a simultaneous approach to decision making, namely, that there is no need to resort to intermediate criteria of success but that for all decisions utility can be defined as a function of the ultimate criterion in the network.

Methods for establishing empirical utility functions for test-based decisions have been studied in Vrijhof, Mellenbergh, and van den Brink [3]. The dominant conclusion from the series of studies of empirical utilities in selection and mastery decisions in this reference is that a choice from the family of linear utility functions is often realistic for both types of decisions. This choice will be made in the empirical example below.

5 Expected utility in the simultaneous approach

For the weak (simultaneous) rules and utility structure defined above, the expected utility for the combined decision problem is equal to:

$$E\left[U_{sim}\left(A^C, B^C(x)\right)\right] \equiv \int_A \int_R \int_R wu_0^{(s)}(t) f(x,y,t)dtdydx +$$
$$\int_{A^C} \int_{B(x)} \int_R u_{10}(t) f(x,y,t)dtdydx + \int_{A^C} \int_{B^C(x)} \int_R u_{11}(t) f(x,y,t)dtdydx.$$

In a Bayesian fashion, the expected utility defined above will be taken as the criterion of optimality in this paper (e.g. [1]).

Taken expectations, completing integrals, and rearranging terms, the above expression can be written as:

$$E\left[U_{sim}\left(A^C, B^C(x)\right)\right] = wE\left[u_0^{(s)}(T)\right] + \int_{A^C} \left\{E\left[u_{10}(T) - wu_0^{(s)}(T) \mid x\right] + \right.$$
$$\left.\int_{B^C(x)} E\left[u_{11}(T) - u_{10}(T) \mid x, y\right] h(y \mid x)dy\right\} q(x)dx,$$

where $q(x)$ and $h(y \mid x)$ denote the p.d.f.'s of X and Y given $X = x$.

6 Sufficient conditions for monotone rules

To find the conditions for monotonicity, first an upper bound is derived to $E\left[U_{sim}\left(A^C, B^C(x)\right)\right]$ by using the well-known theorem that any integral is maximal for those values of the integration variable for which the integrand is nonnegative. Applying this theorem to the inner integral in the right-hand side of the expression for $E\left[U_{sim}\left(A^C, B^C(x)\right)\right]$ with respect to y, and using $q(x) \geq 0$, it follows that for all $B^C(x)$ and an arbitrary but fixed A^C :

$$E\left[U_{sim}\left(A^C, B^C(x)\right)\right] \leq wE\left[u_0^{(s)}(T)\right] + \int_{A^C} \left\{E\left[u_{10}(T) - wu_0^{(s)}(T) \mid x\right] + \right.$$
$$\left.\int_{B_0^C(x)} E\left[u_{11}(T) - u_{10}(T) \mid x, y\right] h(y \mid x)dy\right\} q(x)dx,$$

with

$$B_0^C(x) \equiv \left\{y : E\left[u_{11}(T) - u_{10}(T) \mid x, y\right] \geq 0\right\}.$$

Again, applying the theorem to the outside integral in the right-hand side of the above inequality with respect to x, it follows that for all A^C :

$$E\left[U_{sim}\left(A^C, B_0^C(x)\right)\right] \leq wE\left[u_0^{(s)}(T)\right] + \int_{A_0^C} \left\{E\left[u_{10}(T) - wu_0^{(s)}(T) \mid x\right] + \right.$$
$$\left.\int_{B_0^C(x)} E\left[u_{11}(T) - u_{10}(T) \mid x, y\right] h(y \mid x)dy\right\} q(x)dx,$$

with

$$A_0^C \equiv \Big\{x : E\left[u_{10}(T) - wu_0^{(s)}(T) \mid x\right] + \int_{B_0^C(x)} E\left[u_{11}(T) - u_{10}(T) \mid x, y\right] h(y \mid x)dy \geq 0\Big\}.$$

For weak monotone rules, the sets A_0^C and $B_0^C(x)$ take the form $[x_c, \infty)$ and $[y_c(x), \infty)$, respectively. It follows that optimal rules take weak monotone forms if the left-hand sides of the inequalities for A_0^C and $B_0^C(x)$ are increasing functions in x and in y for all x, respectively.

7 Calculation of optimal simultaneous rules

Assuming the conditions for weak monotonicity are satisfied, optimal (weak) cutting scores can now be obtained for those values of x_c and $y_c(x)$ for which the left-hand sides of the inequalities for A_0^C and $B_0^C(x)$ turn into equalities. Since the sets A_0^C and $B_0^C(x)$ are defined for all x, and thus for x_c , first the optimal (weak) cutting score on the selection test X can be found by solving simultaneously the two equalities for x_c and $y_c(x_c)$. Then, for each $x \geq x_c$, $y_c(x)$ is obtained by putting the left-hand side of the inequality for $B_0^C(x)$ equal to zero and solving for y.

Since no analytical solutions for the system of equations could be found, Newton's method for solving nonlinear systems was used for the calculation of all cutting scores in the empirical example below. The method was implemented in a computer program called NEWTON. Another program, UTILITY, was written to analyze differences in expected utility for the simultaneous and separate rules.

8 Optimal separate rules

It is observed that optimal rules for the separate decisions can easily be found by imposing certain restrictions on $E\left[U_{sim}\left(A^C, B^C(x)\right)\right]$. First, substituting $w = 1$ into the expression for $E\left[U_{sim}\left(A^C, B^C(x)\right)\right]$, the expected utility for the separate selection decision $E[U^{(s)}(A^C)]$, can be written as:

$$E\left[U^{(s)}(A^C)\right] = E\left[u_0^{(s)}(T)\right] + \int_{A^C} E\left[u_1^{(s)}(T) - u_0^{(s)}(T) \mid x\right] q(x)dx.$$

Next, substituting $w = 0$, $A^C = R$ (i.e., accepting all students for the instructional treatment), and $B^C(x) = B^C$ into the expression for the expected utility in the simultaneous approach gives the following result for the expected utility of the separate mastery decision:

$$E\left[U^{(m)}(B^C)\right] = E\left[u_0^{(m)}(T)\right] + \int_{B^C} E\left[u_1^{(m)}(T) - u_0^{(m)}(T) \mid y\right] s(y)dy,$$

where $s(y)$ denotes the p.d.f. of Y.

Analogous to the simultaneous approach, applying again the theorem stated above, it can easily be verified that the monotonicity conditions for the separate decisions boil down to the integrands in the right-hand expressions for $E\left[U^{(s)}(A^C)\right]$ and $E\left[U^{(m)}(B^C)\right]$ being increasing functions in x and y, respectively. Similarly, assuming that these conditions are satisfied, optimal cutting scores for the separate selection and mastery decisions, say x_c^* and y_c^*, can be obtained by putting these integrands equal to zero and solving for x and y, respectively.

9 An empirical example

Optimal rules were calculated for a selection-mastery decision problem consisting of a CAI module on elementary medical knowledge preceded by a selection test for the instructional module and followed by a mastery test at the end of the module. Both tests consisted of 21 items and had possible test scores ranging from 0-100. Data were available for a sample of 76 freshmen in a medical program. The instructors in the program considered students as having mastered the module if their true scores were larger than 55. Therefore, t_c was fixed at this value. All students in the program were admitted to the instructional module, therefore the samples of the score distributions involved did not suffer from any restriction of range.

9.1 Score distributions and monotonicity conditions

It was assumed that (X, Y, T) followed a trivariate normal distribution. Under this assumption, the bivariate distribution of (X, Y) is normal and the regression function $E(Y \mid x)$ is linear. The two observable consequences were tested against the data using a chi-square and a t-test. The probabilities of exceedance were 0.219 and 0.034, showing a satisfactory fit which confirmed our visual inspection of various plots of the distributions.

The means, standard deviations, reliabilities, and correlations of X, Y, and (X, Y) were estimated as follows: $\mu_X = 50.679$, $\mu_Y = 62.436$, $\sigma_X = 8.781$, $\sigma_Y = 9.456$, $\rho_{XX'} = 0.773$, $\rho_{YY'} = 0.802$, and $\rho_{XY} = 0.751$. Standard results from classical test theory [2] were used to express the conditional expectations and variances of T given x and/or y as functions of these observable quantities. Substituting the conditional expectations and variances of T given x and/or y into the sets A_0^C and $B_0^C(x)$, it then turned out that the conditions for weak monotonicity were satisfied for all values of x and y in the range from 0-100. In addition, it turned out that the monotonicity conditions for the separate decisions were satisfied.

9.2 Separate utility functions

The following choice was made for the functions $u_{(i)}^{(s)}(t)$ and $u_{(j)}^{(m)}(t)$ in the expression for the additive utility structure of the combined problem:

$$u_i^{(s)}(t) = \begin{cases} b_0^{(s)}(t_c - t) + d_0^{(s)} & \text{for } i = 0 \\ b_1^{(s)}(t - t_c) + d_1^{(s)} & \text{for } i = 1 \end{cases}$$

$$u_j^{(m)}(t) = \begin{cases} b_0^{(m)}(t_c - t) + d_0^{(m)} & \text{for } j = 0 \\ b_1^{(m)}(t - t_c) + d_1^{(m)} & \text{for } j = 1 \end{cases}$$

where $b_i^{(s)}, b_j^{(m)} > 0$ $(i,\ j = 0, 1)$. The parameters $d_i^{(s)}$ and $d_j^{(m)}$ can represent, for example, the fixed amount of costs involved in following an instructional module and testing the examinees. The condition $b_0^{(s)}, b_1^{(s)} > 0$ states that utility be a decreasing function for the rejection decision, but an increasing function for the acceptance decision. Similarly, the condition $b_0^{(m)}, b_1^{(m)} > 0$ expresses that the utilities associated with failing and passing the mastery test be decreasing and increasing functions in t, respectively.

9.3 Results for the simultaneous and separate rules

For several values of the utility parameters and the weights, optimal weak cutting scores were calculated using the program NEWTON. For instance, with $b_0^{(s)} = 2$, $b_1^{(s)} = 4$, $b_0^{(m)} = 1$, $b_1^{(m)} = 2$, $d_0^{(s)} = -2$, $d_1^{(s)} = -3$, $d_0^{(m)} = -4$, $d_1^{(m)} = -5$, and $w = 0.3$, the values for the optimal weak cutting scores on the selection and mastery test, x_c and $y_c(x_c)$, were respectively 41.83 and 55.38. Optimal cutting scores x_c^* and y_c^* were also computed for the separate selection and mastery decisions. For the same values of the utility parameters and same weight, x_c^* and y_c^* were respectively 41.70 and 53.58.

9.4 Compensatory character of optimal weak rules

It turned out that $y_c(x)$ was decreasing in x. As already explained, the weak rules introduce an element of compensation in the decision procedure. A quantitative estimate of this effect could be calculated by substituting the estimated regression plane $E(T \mid x, y)$ into the left-hand side of the inequality for $B_0^C(x)$ and solving for y. For our values of the utility parameters and value of the weight, it appeared that optimal mastery scores on the mastery test, $y_c(x)$, had to be lowered by 0.675 for each score point above the optimal weak cutting score $x_c = 41.83$ on the selection test.

A consequence of this compensatory character of optimal weak rules is that only students who were just accepted for the instructional module have to compensate their rather low cutting scores on the selection test with relatively high scores on the mastery test to reach the mastery status. However, the decreasing character of $y_c(x)$ in x implies that students with selection scores far above the optimal weak cutting score $x_c = 41.83$ on the selection test do need rather low weak cutting scores $y_c(x)$ on the mastery test to reach the mastery status.

9.5 Comparison of the expected utilities

For the simultaneous approach a gain in expected utility relative to the separate approach was expected. To check whether this expectation could be confirmed, the weighted sum of the expected utilities for the optimal separate rules was compared with the expected utilities for the optimal weak monotone rules. The results indicated that the expected utilities for the optimal weak monotone rules yielded the largest values for all values of the utility parameters and all weights. For instance, for the same values of the utility parameters and same weight again, the expected utilities for the weak monotone and separate approach were respectively 19.14 and 17.58.

10 Concluding remarks

Although the area of individualized instruction is a useful application of simultaneous decision making, it should be emphasized that the optimization models advocated in this paper have a larger scope of application. For any situation in which subjects are accepted for a certain treatment on the basis of their scores on a selection test with attainments evaluated by a mastery test, the optimal rules presented in this paper can improve the decisions. An example is psychotherapy where clients accepted have to pass a success criterion before being dismissed from the therapy.

A final note is appropriate. In the current paper, it was demonstrated how a very simple instructional network consisting of a combined selection-mastery decision can be optimized simultaneously. It should be noticed, however, that much more complicated instructional networks consisting of many decisions as nodes can be optimized simultaneously using the same procedures. Doing so, the weak monotone approach actually provides us with some 'artificial intelligence' for setting optimal weak cutting scores. The more test data for each student comes available, the better the optimal weak cutting scores for each student can be set by taking optimal advantage of each student's preceding (test) history in the CAI-system.

References

[1] Lehmann E.L. (1959). *Testing statistical hypotheses.* New York: Wiley.

[2] Lord F.M., Novick M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

[3] Vrijhof B.J., Mellenbergh G.J., van den Brink W.P. (1983). *Assessing and studying utility functions in psychometric decision theory.* Applied Psychological Measurement **7**, 341 – 357.

Address: H.J. Vos, University of Twente, Faculty of Behavioral Sciences, Department of Research Methodology, Measurement, and Data Analysis. P.O. Box 217, 7500 AE Enschede, the Netherlands.

E-mail: `Vos@edte.utwente.nl`

ON OPTIMAL DESIGN FOR DISCRIMINATION AND ESTIMATION

T.H. Waterhouse, J.A. Eccleston and S.B. Duffull

Key words: T-optimal designs, D-optimal designs, compound designs, statistical computing.

COMPSTAT 2004 section: Design of experiments.

Abstract: Methods of constructing experimental designs using compound optimality criteria are described in which the aim is to produce designs which are efficient in terms of both model discrimination and parameter estimation.

1 Introduction

Research into the optimal design of experiments has in the main concentrated on optimisation with respect to parameter estimation. An experimental design is 'optimised', that is a choice of runs, experimental units, etc. is derived/found through the use of an optimality criterion. The most commonly used criterion is D-optimality, which minimises the determinant of the variance-covariance matrix of the parameter estimates (see Atkinson and Donev [2] for an excellent presentation on optimal design). In comparison, relatively little research has been carried out into criteria and techniques for discriminating between competing models for an experimental design. Atkinson and Fedorov [3],'[4] introduced T-optimality as a criterion for discriminating between the models. T-optimality has not been widely used, due in part to the computational complexity involved in the computing for even relatively simple models, but also due to the theoretical challenges it presents (see Bogacka and Uciński [6]).

The use of optimisation techniques and modern powerful computers has lead to the use of more complex and realistic models being used to model data. However, the majority of research in optimal design remains focussed on linear models, while in fact nonlinear models are commonly used in many areas such as chemistry and pharmacology. The major difference between the design of experiments for linear and nonlinear models is that nonlinear models require estimates of the model parameters to be known for an optimal design to be constructed. Typically initial parameter estimates are based on the results of previous studies or 'expert' guesses. Given the parameter values a D-optimal design can be readily derived or constructed. A T-optimal design for discriminating between two models requires one model to be assumed correct and the parameters to be known. It is unclear as to whether T-optimal designs are good D-optimal designs and vice versa.

It would be very useful and advantageous to be able to have designs that perform both model discrimination and model parameter estimation simultaneously. Compound criteria, such as the product of D-optimality criteria (Atkinson and Cox [1], Walter and Pronzato [7], and Waterhouse et al. [8]) have been found to be useful.

In this paper criteria and methods are introduced which are aimed at achieving designs that efficiently discriminate between models and yield efficient parameters estimates. Product and conditional optimality criteria are developed and applied to both linear and nonlinear models.

2 Compound optimality criteria

Compound or composite criteria can be useful if they combine two (say) design properties that we are interested in achieving. As mentioned above it is not known whether a design can be simultaneously T- and D-optimal. Two compound criteria are proposed, both aimed at achieving some degree of effective model discrimination and efficient parameter estimation. Each of the proposed compound criteria used either or both T- or D-optimality objective functions as defined below.

The model under consideration is

$$y_i = \eta(x_i, \theta) + \varepsilon_i \quad (i = 1, \ldots, N) \tag{1}$$

for some nonlinear functional $\eta : \mathbb{R}^{m+p} \to \mathbb{R}$, where the $y_i \in \mathbb{R}$ are the univariate responses, $x_i \in \mathcal{X} \subset \mathbb{R}^m$ are the explanatory variables, and $\theta \in \Theta \subset \mathbb{R}^p$ is the vector of parameters. The additive error terms ε_i are assumed to be independent identically distributed normal random variables with zero mean and constant variance.

A continuous design ξ for this model with n support points $(n \le N)$ is defined as

$$\xi = \begin{Bmatrix} x_1 & x_2 & \cdots & x_n \\ w_1 & w_2 & \cdots & w_n \end{Bmatrix},$$

where the x_i are the support points, with weights $w_i \in [0, 1]$ summing to 1. Methods for choosing such a design (including the number of support points) in an 'optimal' way are described below.

2.1 T-optimality

T-optimal designs for discriminating between two candidate models $\eta_1(x, \theta_1)$ and $\eta_2(x, \theta_2)$ with $x \in \mathcal{X}$ and $\theta_r \in \Theta_r$, as described by Atkinson and Fedorov [3, 4], depend on the assumption that one of these models is true, and that its parameter values are known. If the first model is arbitrarily chosen as true (and it is therefore assumed that θ_1 is known), write $\eta_t(x) = \eta_1(x, \theta_1)$, and the T-optimal design can be defined, using notation similar to that of Bogacka and Uciński [6], as

$$\xi_T^* = \arg\max_{\xi \in \Xi} \left\{ \min_{\theta_2 \in \Theta_2} J(\xi, \theta_2) \right\}, \tag{2}$$

where

$$J(\xi,\,\theta_2) = \sum_{i=1}^{n} w_i \|\eta_t(x_i) - \eta_2(x_i,\,\theta_2)\|^2, \tag{3}$$

and $\|\cdot\|$ is the Euclidean norm. As noted by Atkinson and Fedorov [3] in their first example, there are certain pairs of models (e.g. nested models) for which it is meaningless to design these types of experiments when parameter values are unrestricted. In such cases it is necessary to place further constraints on θ_2.

To the best of the author's knowledge, a comparison of designs relating to the T-optimality criterion has not yet been considered. We therefore propose two possible methods for making such a comparison. An obvious measure for a design ξ is the ratio

$$\frac{J(\xi,\,\theta_2)}{J(\xi_T^*,\,\theta_2^*)},$$

where $\theta_2^* = \arg\min_{\theta_2 \in \Theta_2} J(\xi_T^*,\,\theta_2)$. However, the choice of θ_2 in the numerator presents a problem. The T-optimal design gives the greatest separation between the response of the two models for the 'worst-case scenario' of θ_2, so it makes sense to assess the new design ξ in a similar manner, i.e. let $\theta_2 = \arg\min_{\theta_2 \in \Theta_2} J(\xi,\,\theta_2)$. On the other hand, the competing D-optimal designs for nonlinear models (discussed later) rely on the prior specification of $\theta_2 = \theta_2^*$, the same value used in the T-optimal design. One may argue that in this case it is fair to judge the design using the same assumption of parameter values as was used to find the design, similarly to the calculation of D-efficiencies.

In this light, we have opted to include both methods of T-efficiencies in this paper:

$$T_{\text{eff}}^a = \frac{J(\xi,\,\hat{\theta}_2)}{J(\xi_T^*,\,\theta_2^*)}, \quad T_{\text{eff}}^b = \frac{J(\xi,\,\theta_2^*)}{J(\xi_T^*,\,\theta_2^*)}, \tag{4}$$

where $\hat{\theta}_2 = \arg\min_{\theta_2 \in \Theta_2} J(\xi,\,\theta_2)$.

2.2 D-optimality

The value of a design ξ in terms of estimation of the parameters of the rth model can be addressed by its efficiency compared to the D-optimal design, i.e. its D-efficiency,

$$D_{\text{eff}}^r(\xi) = \left\{ \frac{|M_r(\xi)|}{\left|M_r(\xi_{D_r}^*)\right|} \right\}^{1/p_r} \qquad (r = 1,\,2), \tag{5}$$

where $M_r(\xi)$ is the information matrix, p_r is the number of parameters and $\xi_{D_r}^*$ is the D-optimal design, which maximises $|M_r(\xi)|$. If the r^{th} model is linear in its parameters, i.e. expressible in the form $E(Y) = X_r\theta_r$, where

$E(Y)$ is the vector of expected responses for the n support points and X_r is the $n \times p$ extended design matrix, then the information matrix does not depend on the parameters θ_r, and is given by

$$M_r(\xi) = X_r^T W X_r, \tag{6}$$

where $W = \mathrm{diag}(w_1, \ldots, w_n)$. In the more general case, where the model may be nonlinear in its parameters, the information matrix is given by

$$M_r(\xi) = F_r^T W F_r, \tag{7}$$

where F is the $n \times p$ matrix of partial derivatives:

$$F = \left\{ \frac{\partial \eta_r(x_1, \theta_r)}{\partial \theta_r} \ \ldots \ \frac{\partial \eta_r(x_n, \theta_r)}{\partial \theta_r} \right\}^T \tag{8}$$

2.3 Product optimality

If the objective of our experiment is not model discrimination, but efficient estimation of both sets of parameters, then, as suggested by Atkinson and Cox [1], one option is to maximise the product of the two D-optimality criteria, scaled for the number of parameters.

$$\xi^*_{D_1,\, D_2} = \arg\max_{\xi} |M_1(\xi)|^{1/p_1} |M_2(\xi)|^{1/p_2} \tag{9}$$

If η_1 and/or η_2 are nonlinear in their parameters, the product design will depend on the values of θ_1 and θ_2. Since we aim to compare these designs to the T-optimal designs, let θ_1 take the value assumed to be known in the construction of the T-optimal design, and let $\theta_2 = \theta_2^*$.

2.4 Conditional optimality

Since the construction of T-optimal designs does not take parameter estimation into account, the D-efficiencies of the designs can be quite poor in practice, which is demonstrated in the following examples. In order to find designs which are useful for both parameter estimation *and* model discrimination, we propose a method which involves the extension of the T-optimal design to include additional support points so that the overall design maximises the product criterion in equation 9.

Suppose that we have a T-optimal design via equations 2 and 3,

$$\xi_T^* = \left\{ \begin{matrix} x_1^* & x_2^* & \ldots & x_n^* \\ w_1^* & w_2^* & \ldots & w_n^* \end{matrix} \right\}. \tag{10}$$

We wish to find k further support points which aid in parameter estimation to add to the existing design. Of course, the weights $w_1^*, \ldots, w_n^*$ will need to be adjusted in order to 'make room' for the additional points. To do this, simply multiply each weight by a factor α, proportional to the importance the experimenter places on model discrimination. The construction of the 'conditional' optimal design (conditional on the fixed T-optimal design) then involves finding the additional support points $x_{n+1}, \ldots, x_{n+k}$ and weights $w_{n+1}, \ldots, w_{n+k}$ so that the complete design

$$\xi_\alpha = \left\{ \begin{matrix} x_1^* & \ldots & x_n^* & x_{n+1} & \ldots & x_{n+k} \\ \alpha w_1^* & \ldots & \alpha w_n^* & w_{n+1} & \ldots & w_{n+k} \end{matrix} \right\} \tag{11}$$

maximises the product criterion as described in equation 9. As already noted, for non-linear models these information matrices will depend on the parameter values θ_1 and/or θ_2. In other words, the conditional optimal design will be locally optimal. As this is an extension of the T-optimal design, we let θ_1 take the value already assumed to be known. For the second model, we also let the parameters be dictated by the T-optimal design: let $\theta_2 = \theta_2^*$.

3 Construction of designs and examples

The procedures outlined in Section 2 were implemented in Matlab to obtain the results in the following examples. A search routine employing simulated annealing was used to optimise each criterion. The minimisation involved in equation 2 at each iteration of the annealing algorithm was implemented using Matlab's Optimisation Toolbox. The functions `fminsearch` and `lsqnonlin` were used for this purpose, with varying success, depending on the model structure and current design. A heuristic approach was taken, using both functions to perform the minimisation, after which the smallest minimum was selected. The initial number of support points was chosen to be larger than was expected to be required for the design. In most cases, the design 'collapsed' to fewer support points as the algorithm progressed, i.e. some weights approached zero or two or more support points approached the same value. In the cases where collapsing didn't occur, the algorithm was rerun with a larger number of initial support points. For further details on computation and additional examples, please refer to Waterhouse et al. [9].

3.1 Nonlinear example

For this example, the nonlinear models used in the case study of §3.13 of Bates and Watts [5] are considered. The first model is the three-parameter quadratic Michaelis-Menten type model

$$\eta_1(x,\,\theta_1) = \frac{\theta_{11}x}{\theta_{12} + x + \theta_{13}x^2} \quad (x \geq 0,\, -\infty \leq \theta_{1j} \leq \infty), \tag{12}$$

and the second is the exponential difference model

$$\eta_2(x,\,\theta_2) = \theta_{21}\left(e^{-\theta_{23}x} - e^{-\theta_{22}x}\right) \quad (x \geq 0,\, -\infty \leq \theta_{2j} \leq \infty). \tag{13}$$

	Efficiency			
Design, $\xi^* = \begin{Bmatrix} x^* \\ w^* \end{Bmatrix}$	D^1_{eff}	D^2_{eff}	T^a_{eff}	T^b_{eff}
T-optimal				
$\xi^*_T = \begin{Bmatrix} 0.1730 & 1.131 & 5.104 & 28.08 \\ 0.0187 & 0.1216 & 0.2163 & 0.6431 \end{Bmatrix}$	0.2428	0.2335	1	1
D-optimal, product				
$\xi^*_{D_1,\,D_2} = \begin{Bmatrix} 0.2577 & 1.142 & 6.416 \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}$	0.9277	0.9511	0	0.8620
Conditional, $\alpha = 0.75$				
$\xi^*_{\alpha=0.75} = \begin{Bmatrix} x^*_T & 0.2594 & 1.127 \\ 0.75w^*_T & 0.1683 & 0.0817 \end{Bmatrix}$	0.5541	0.5117	0.7825	0.9609
Conditional, $\alpha = 0.5$				
$\xi^*_{\alpha=0.5} = \begin{Bmatrix} x^*_T & 0.2589 & 1.141 & 6.914 \\ 0.5w^*_T & 0.2290 & 0.1710 & 0.1000 \end{Bmatrix}$	0.6770	0.6633	0.5254	0.9158
Conditional, $\alpha = 0.25$				
$\xi^*_{\alpha=0.25} = \begin{Bmatrix} x^*_T & 0.2583 & 1.143 & 6.581 \\ 0.25w^*_T & 0.2814 & 0.2525 & 0.2161 \end{Bmatrix}$	0.8013	0.8088	0.2635	0.8864

Table 1: Efficiencies of near-optimal designs for two competing models for rise and decay.

To find the T-optimal design, the first model is assumed true, and we arbitrarily select $\theta_{11} = \theta_{12} = \theta_{13} = 1$. The T-optimal design, product design and conditional designs (for a range of α) and their D- and T-efficiencies are given in Table 1. It should be noted that with respect to T^b_{eff}, the product optimal and conditional optimal designs are very similar. However the D-efficiencies of the product design are much higher. The efficiencies for a range of conditional optimal designs, including the product design ($\alpha = 0$) and the T-optimal design ($\alpha = 1$) are shown graphically in Figure 1. There is an obvious increasing trend in D-efficiencies of the conditional designs as more importance is placed on parameter estimation (i.e. as α decreases) and an almost linear increase in both T-efficiencies as we favour model discrimination (i.e. as α increases). A reasonable trade-off between the two objectives would be at around $\alpha = 0.6$.

3.2 Linear example

For completeness, we consider the two linear models of Example 20.3 of Atkinson and Donev [2]:

$$\eta_1(x,\,\theta_1) = \theta_{11} + \theta_{12}e^{x} + \theta_{13}e^{-x} \quad (-1 \le x \le 1,\, -\infty \le \theta_{1j} \le \infty), \tag{14}$$

and

$$\eta_2(x,\,\theta_2) = \theta_{21} + \theta_{22}x + \theta_{12}x^2 \quad (-1 \le x \le 1,\, -\infty \le \theta_{2j} \le \infty). \tag{15}$$

As in Atkinson and Donev [2], we assume the first model to be true, with parameter values $\theta_{11} = 4.5$, $\theta_{12} = -1.5$ and $\theta_{13} = -2$. The same process as the previous example was followed, but the table of results has been

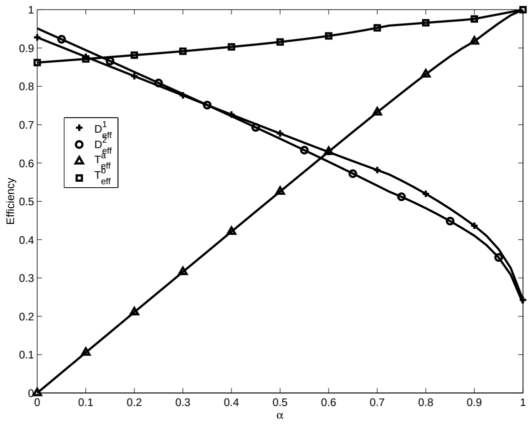


Figure 1: Efficiency of conditional designs, Example 3.1.

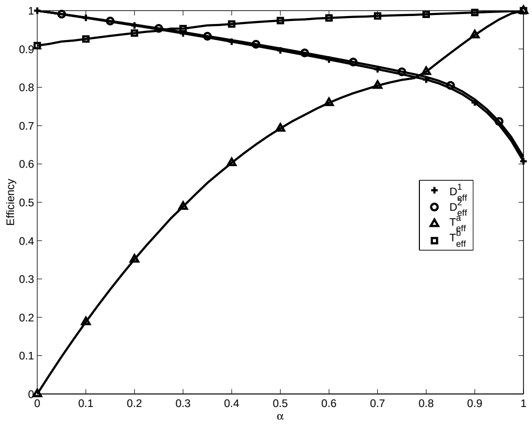


Figure 2: Efficiency of conditional designs, Example 3.2.

omitted for brevity. However, Figure 2 has been included which describes the efficiencies of a range of conditional designs, with the product design ($\alpha = 0$) and the T-optimal design ($\alpha = 1$). Similar trends are shown here, with high values of α (around 0.8) giving good efficiencies for both parameter estimation and model discrimination.

4 Conclusion

We have seen that although T-optimal designs can be inefficient for parameter estimation, and conversely D-optimal designs may be unsuitable for model discrimination, we are able to construct designs using compound criteria which offer a comfortable compromise between the two objectives. Through the conditional designs described here, the trade-off between estimation and discrimination can be simply specified by the choice of α.

Further results for additional models (see Waterhouse et al. [9]) show that D-optimal designs (including product designs) with few support points offer little value in terms of discrimination. In these cases the additional support points in the conditional designs give a dramatic improvement in T-efficiencies. It would appear that for models whose D-optimal designs have several support points, a product optimal design is efficient for both model discrimination and parameter estimation. This is a topic of ongoing research.

References

[1] Atkinson A.C., Cox D.R. (1974). *Planning experiments for discriminating between models (with discussion).* J. R. Statist. Soc. B **36**, 321 − 48.

[2] Atkinson A.C., Donev A.N. (1992). *Optimum experimental designs.* Oxford University Press, Oxford.

[3] Atkinson A.C., Fedorov V.V. (1975). *The design of experiments for discriminating between two rival models.* Biometrika **62**, 57 − 70.

[4] Atkinson A.C., Fedorov V.V. (1975). *Optimal design: experiments for discriminating between several models.* Biometrika **62**, 289 − 303.

[5] Bates D.M., Watts D.G. (1998). *Non-linear regression analysis and its applications.* Wiley, New York.

[6] Bogacka B., Uciński D. (2002). *Construction of T-optimum designs for multiresponse dynamic models.* Proceedings in Computational Statistics. 15th Symposium COMPSTAT 2002, Berlin, August 2002, W. Härdle, B. Rönz (eds), Physica-Verlag, 267 − 272.

[7] Walter E., Pronzato L. (1997). *Identification of parametric models from experimental data.* Springer, Berlin.

[8] Waterhouse T.H., Duffull S.B., Eccleston J.A. (2003). *Optimal design for model discrimination.* Poster presentation, The Twelfth Meeting of the Population Approach Group in Europe. (Available from the authors.)

[9] Waterhouse T.H., Eccleston J.A., Duffull S.B. (2004). *On optimal design for discrimination and estimation.* Research Report #106, Centre for Statistics, University of Queensland.

Address: T.H. Waterhouse, J.A. Eccleston, S.B. Duffull, Department of Mathematics, The University of Queensland, Brisbane QLD 4072, Australia

E-mail: `thw@maths.uq.edu.au`

ENCYCLOPEDIA OF STATISTICAL GRAPHICS

Adalbert F.X. Wilhelm and Rüdiger Ostermann

Key words: Statistical graphics, learning environment.

COMPSTAT 2004 section: Graphics, Teaching statistics.

Abstract: Despite the recent improvements in statistical software and availability of statistical graphics software, visual analysis techniques still play a minor role in statistical education. The major reason for this might be that there is no extensive and thorough collection of statistical graphics and their use available, neither in standard book form nor using modern computerized formats like a website or a multimedia DVD. This paper shall present a recently started project to fill this void. In this project we are currently developing a web-based encyclopedia used for teaching that will be available in the intranet at International University Bremen and Fachhochschule Münster.

1 Statistical graphics

Graphical displays have a long tradition in statistical analyses and some forms can be traced back to 3800 B.C. [8], [2]. The modern history of statistical graphs as we know them goes back to the middle of the 18th century and is closely related with state statistics or Staatenkunde. Influenced by the work of the late John W. Tukey statistical graphics have experienced a strong resurgence as a cornerstone in exploratory data analysis since the early 70s of the last century, see Chambers et al. [3] and Tukey [14]. As many publications show, statistical graphics are also extremely useful for checking model assumptions and data reliability, especially when used as diagnostic plots, see for example Cook and Weisberg [5]. The shift from using graphics primarily as a result presenting device to an analytic tool was accompanied by the invention of new complex plots and techniques, for example *box plots*, *parallel coordinate plots* and *projection pursuit*, see Tukey [13], Wegman [15], and Furnas and Buja [9]. Dynamic plots like *3-D rotating plots* and the *Grand Tour* [1], could not have been imagined without the rapid developments in computer science. It was only in the last fifteen years that interaction with graphical displays could be put into practice. As a consequence plots changed their character from formerly being a final product to now being a temporary tool that can be modified and adapted according to the situation by simple mouse clicks or keyboard commands. Information overload that can result from exceedingly detailed static plots, is replaced by simpler plots with multiple querying options, facilitating an organized, and focused approach to extracting information from data.

2 Structure of the encyclopedia

As the name 'encyclopedia suggests, the project aims to cover as many aspects of statistical graphics as possible. The basic setup for each plot will provide information on the historical development, the basic construction, application, extensions and modifications of the plot. In addition, for each plot the static version will be explained as well as interactive variants. An interactive environment such as the internet helps to make the graphics as vivid as possible and eases also communication of concepts and techniques of visual analysis. Our approach can be differentiated from three other, currently common offerings. It differs from already available internet collections of statistical graphics, like M. Friendly's pages (Gallery of Visualisation: http://hotspur.psych.yorku.ca/SCS/Gallery/ and Data Vis/Stat Graphics: http://www.math.yorku.ca/SCS/StatResource.html) in the way that we do not focus on particular highlights of statistical visualization but that we try to give a broad persepctive. In contrast to newly developed internet teaching courses (e.g. EMILeA-stat: http://www.emilea.de) we put the graphics in the center of our description. The third offering we like to differentiate is the one accompanying software for statistical graphics (e.g. MANET: http://stats.uni-augsburg.de, or GGobi: http://www.ggobi.org). We do not restrict to one particular software or implementation, but we focus on the conceptual framework of the graphics.

3 Example: bar chart

To illustrate our approach to the encyclopedia we present a detailed outline of the section for bar charts. Each topic in the encyclopedia will have a historical background, details of construction, applications, modifications, extensions, and ways to define direct manipulation on the plot.

3.1 Historical background

William Playfair, a political economist, is credited for being inventor of the bar chart and the pie chart. Actually, in the first edition of his Commercial and Political Atlas he apologized for inventing the bar chart, because he only created it due to lack of data. He aimed to create a time-line chart as Joseph Priestley had established them in 1765 but Playfair's problem was that in one instance, for the trade data of Scotland, he had only data for one year so he was unable to portray a time series. At that time, such data was typically displayed in the form of a statistical map, see Crome [6]. Playfair, however, split the data by the trade partners and represented Scotlands imports and exports by a single bar for each of the 17 trading partners.Without knowing it, Playfair had opened the way for visual comparison of discrete quantitative data. Up to this time all visual displays of quantitative data have been limited to those that located data either in space or in time. Now it was possible to visually compare data that was organized by any discrete categorization. By 1801 when he published his first pie chart he had recognized the importance of his invention.

3.2 Construction

The bar chart represents counts by area and reduces this two-dimensional information to a visual comparison of scalars by simply keeping one dimension fixed for all categories. Thus, in the bar chart counts in each category are represented by a rectangle (bar, tile) with common width such that the areas of the tiles only vary with the height of the bars. It is much easier for the human eye and brain to compare heights of bars than to compare areas of rectangles. This makes the bar chart easy to use and interpret.

Thus, for constructing a bar chart one needs

- the common width
- the individual heights of a bar
- and the order of the categories (and hence the bars)

The order of the categories is an important task. Neighboring bars can be more easily compared than bars that are far apart. The standard approach in statistical software is either to use the order of appearance of the categories in the data set or the lexicographic ordering of the category labels.

3.3 Application

The bar chart can be used for any kind of discrete data in the way that the area of a bar represents the frequency of the corresponding category. However, most users are not aware of the area principle that is inherent to the bar chart by definition and only go along with the comparison of heights of the bars, relying on that a common width has been used for all bars.

Today, the bar chart is second to the pie chart the most popular statistical graphic. On election day it will be used to show the results, the gains and losses of the individual political parties. The easy comparison of lengths makes the bar chart very appealing and understandable to almost everyone.

3.4 Modifications

The spine plot [11] is one modification of the bar chart. In a standard bar chart the area of a bar represents the frequency of the corresponding category, and typically, bars are drawn with a common width such that the frequency can easily be read from the height of the bar. The disadvantage of this procedure appears with the use of linked subsetting. If selected cases are highlighted in bar charts it is usually hard to compare relative proportions with each other. In a spine plot, each bar is chosen with a constant height, so that now the width varies according to the number of data points in a category. Comparison of highlighted proportions is now straightforward.The spine plot as described above can be seen as a one-dimensional mosaic plot. In its static form the mosaic plot was first used by v. Mayr [12] (see Figure 1) and re-invented by Hartigan and Kleiner [10].

Mosaic Plots are recursively defined and can — at least theoretically —be generalized to arbitrary numbers of variables.

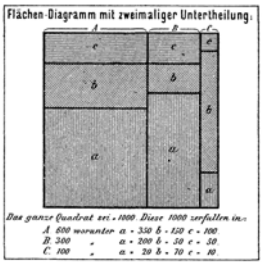


Figure 1: Mosaic diagram by Georg von Mayr [12] representing a two-way classification.

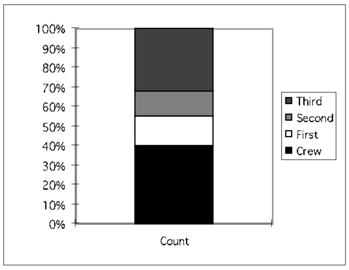


Figure 2: Stacked bar charts representing a two-way classification.

3.5 Extension

Stacked bar charts are an alternative to pie charts (see Figure 2). Their main idea is to draw one bar for all cells and split this bar horizontally in slices according to the counts in the categories.Shading or coloring is used to distinguish the resulting blocks. As well as in pie charts it isdifficult to compare proportions of two groups that only differ a little. Also the type of shading or coloring influences the human perception and might lead to wrong conclusions.The splitting idea of stacked bar chart is identical to the one used in spine plots. However, the layout is rotated by 90 degrees. This splitting principle is recursively extended in mosaic plots to higher dimensions. If the grouping variable is ordinal stacked bar chart relates to empirical distribution functions. The comparison of two different groups (males and females for example) can be visually underlined and enhanced by the use of juxtaposed bar charts (sometimes also called back-to-back bar charts). Corresponding bars are then positioned opposite to each other and are pointing in different directions. Here again, a constant width allows to simply compare the lengths of the bars. The age pyramid is typically a particular example for such a back-to-back bar chart.

Mayr [12] has already discussed the possibility to deviate from the rectangular basis of bars and to introduce them in a triangular form. Being aware that people tend to ignore the area principle in bar charts and are simply comparing the heights of the bars he argued that for cases in which a small but important minority is compared to a large but less important majority, a triangular chart can show both aspects within one diagram, see Figure 3.

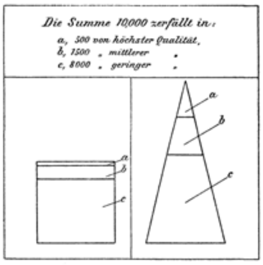


Figure 3: Triangular bar chart by Georg von Mayr [12] to enforce the importance of a small minority compared with a less important majority.

3.6 Interactive variation

How does linked highlighting and conditioning look in bar charts. The standard view of a bar chart shows the counts of cases that fall into each class. Usually, all bars in a bar chart have the same width thus displaying the counts not only by the area but also by the height of the bars. To illustrate these concepts let us look at the passenger data of the Titanic disaster, as presented by Dawson [7]. Drawing one bar chart for the survival status and one for the class membership, we can now start to condition on a subgroup. For example, let us select the survivors in the Titanic disaster. The highlighted bar chart shows three different distributions at the same time. First of all the total heights of the bars are still displaying the number of cases that fall into that particular class, i.e. the marginal distribution. The heights of the highlighted areas reflect the counts for the cross-classification in the variable "Class", i.e. they show $\mid \{\omega : Class(\omega) = \cdot, Surv(\omega) = yes\} \mid$. Dividing these counts by the total number of survivors yields the conditional probabilities $P(Class \mid Survival = yes)$. Since the denominator is constant for all classes it can be ignored whence the counts and the highlighted areas respectively can be taken as representation of the conditional distribution $P(Class = \cdot \mid Survival = yes)$. At the same time we can interpret all areas (highlighted parts and those that are not highlighted) as graphical representations of the joint distribution $P(Class, Survival)$. This can be seen more clearly when we rearrange the areas to obtain a standard bar chart for the 8 classes of the cross-classification, see Figure 5. In terms of conditional probabilities we compare in Figure 6 $P(\text{Survived} = \text{yes}|\text{class} = \text{crew})$ to $P(\text{Survived} = \text{yes}|\text{class} = \text{first})$ and $P(\text{Survived} = \text{yes}|\text{class} = \text{second})$ and $P(\text{Survived} = \text{yes}|\text{class} = \text{third})$. To get this graphically we have to select the category "yes" in the bar chart for "Survived" and switch the "Class" bar chart to a spine plot. From the graph we conclude that the survival rate was highest for the first class, substantially lower for the second class, and much the same for third class and crew.

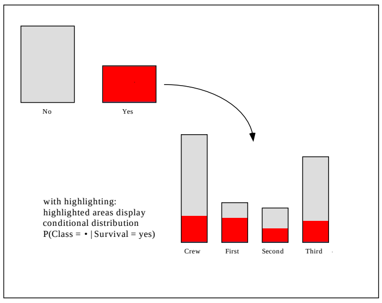


Figure 4: Conditioning a bar chart via linked highlighting represents three different aspects of the data: marginal distribution, conditional distribution, and joint distribution.

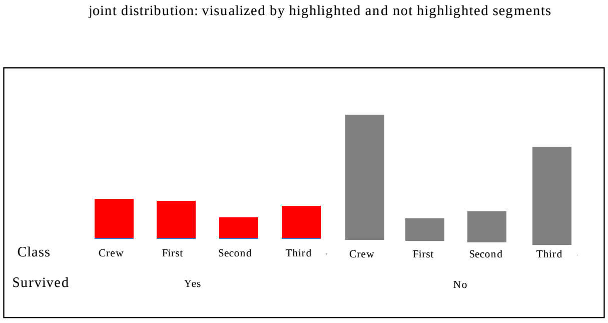


Figure 5: Rearranging the areas of a bar chart with highlighting yields a graphical representation of the joint distribution.

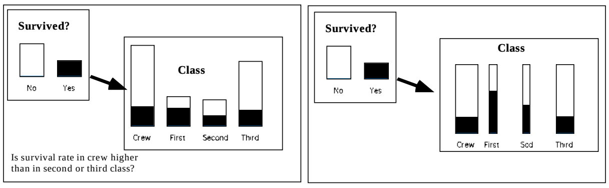


Figure 6: In the linked bar charts on the left it is hard to compare survival rates. Switching the bar chart for variable "Class" to a spine plot eases the comparison substantially.

4 Conclusion

This recently started project to develop an encyclopedia of statistical graphics aims in providing a digital platform of statistical graphic.As we have shown for the bar chart , we will cover the following topics for each graphical method: basic construction, historical development, applications, extensions and modifications, static and interactive variants as well as connections and/or differences to similar graphical methods. In the first phase we anticipate developing material on the following topics: Barcharts, histograms, scatterplots, pie charts, mosaic plots, boxplots, line and area diagrams, parallel coordinate plots. This encyclopedia will be based on a web technology to allow for an extended and easy link between various display types and concepts. Thus, we can flexibly organize the different sorts of plots, point out the connexions between the displays, and present the conceptual similarities. Moreover, a digital version instead of a print version will offer inexpensive use of colors.

References

[1] Asimov D. (1985). *The Grand Tour: A Tool for viewing multidimensional sata.* SIAM Journal Scientific and Statistical Computing **6**, 128–143.

[2] Beniger, J.R. and Robin, D.L. (1978). *Quantitative graphics in statistics: A brief history.* The American Statistician **32**, 1 − 11.

[3] Chambers J.M., Cleveland W.S., Kleiner B., Tukey P.A. (1983). *Graphical methods for data analysis.* Chapman & Hall, New York.

[4] Cleveland W.S., McGill M.E. (1988). *Dynamic graphics for statistics.* Wadsworth & Brooks/Cole, Pacific Grove, CA.

[5] Cook D.R., Weisberg S. (1994). *An introduction to regression graphics.* Wiley, New York.

[6] Crome A.F.W. (1782). *Europens Produkte. Zum Gebrauch der neuen Produkten-Karte von Europa.* Dessau.

[7] Dawson R.J. (1995). *The "unusual episode" data revisited.* Journal of Statistics Education **3**.

[8] Funkhouser H.G. (1937). *Historical development of the graphical representation of statistical data.* Osiris **3**, 269 – 404.

[9] Furnas G.W., Buja A. (1994). *Prosection views: dimensional inference through sections and projections.* Journal of Computational and Graphical Statistics **3**, 323 – 385.

[10] Hartigan J., Kleiner B. (1981). *Mosaics for contingency tables.* Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, 268 – 273, Springer-Verlag, New York.

[11] Hummel J. (1996). *Linked bar charts: analyzing categorical data graphically.* Computational Statistics **11**, 36 – 44.

[12] Mayr G. von (1877). *Die Gesetzmäßigkeit im Gesellschaftsleben.* Oldenbourg, München.

[13] Tukey J.W. (1972). *Some graphic and semigraphic displays.* In: T.A. Bancroft (ed.): Statistical Papers in Honor of George W. Snedecor, Iowa State University, 293 – 316.

[14] Tukey J.W. (1977). *Explorative data analysis.* Addison–Wesley, Reading, MA.

[15] Wegman E.J. (1990).*Hyperdimensional data analysis using parallel coordinates.* Journal of the American Statistical Association **85**, 664 – 675.

Address: A.F.X. Wilhelm, International University Bremen, Campus Ring 1, P.O. Box 750541, D-28725 Bremen, Germany

R. Ostermann, University of Applied Sciences / Department of Nursing, Leonardo Campus 8, D - 48149 Münster, Germany

E-mail: `a.wilhelm@iu-bremen.de`

A FAST BOOTSTRAP METHOD FOR THE MCD ESTIMATOR

Gert Willems and Stefan Van Aelst

Key words: Minimum covariance determinant, bootstrap, robust inference.

COMPSTAT 2004 section: Robustness.

Abstract: We introduce and investigate a fast bootstrap method for Rousseeuw's Minimum Covariance Determinant (MCD). While the classical bootstrap approach requires the time-consuming recalculation of the MCD, we propose to use a fast approximation in the resamples instead. In this way the bootstrap becomes a practical inference tool for MCD and an alternative to the asymptotic variance approach. Extensive simulations show that the method performs quite well in estimating the variability of the MCD, as well as in other inference procedures. Furthermore this fast bootstrap has a robustness benefit over the classical bootstrap.

1 Introduction

It is well known in multivariate statistical analysis that the classical mean and covariance matrix are extremely sensitive to outliers in the data. Therefore, several robust alternatives have been studied in the literature. Among the most widely used is the Minimum Covariance Determinant (MCD) estimator of Rousseeuw [5]. It estimates the location and scatter by searching for the subset (of a given size) whose empirical covariance matrix has the smallest determinant. Besides its intuitively appealing definition, its advantages include asymptotic normality, affine equivariance, a bounded influence function and a high breakdown point which can be up to 50%.

When it comes to inference about the unknown population parameters estimated by MCD, one usually turns to asymptotic variance results based on normality. The latter are derived in [1] for the location and [2] for the scatter. However, asymptotic estimates may be inaccurate for small sample sizes, and often are inappropriate in situations where robust estimators are most needed. Resampling methods such as the bootstrap [4] constitute an obvious alternative to the asymptotic estimates. Some drawbacks arise though when using the classical bootstrap on MCD. The most serious of these is the computational cost of the procedure. Indeed, as with most robust estimators the MCD algorithm is time-consuming. Therefore, recalculating the MCD in each resample is not practical or not even feasible, especially for large samples in high dimensions. Another typical problem that occurs when bootstrapping robust estimators is that the breakdown point of the bootstrap variance and quantile estimates is lower than that of the estimator itself. This robustness problem has been investigated by Singh [8] and by Stromberg [9] who specifically addressed the case of the MCD estimator. Both authors suggest straightforward robustifications of the bootstrap such as Winsorization [8] or 'limited replacement bootstrap' [9]. The problem of computability however remains. Recently, a fast and robust bootstrap method was introduced [7], [10] for estimators that can be represented as a solution of a smooth fixed-point equation, such as MM- and S-estimators. For less smooth estimators such as MCD it is more difficult to obtain theoretical results on bootstrap methods. Nevertheless we will show that it is possible to construct an approximating bootstrap procedure for MCD that is both fast and robust, and that works well. The idea is to draw bootstrap samples just as in the classical bootstrap, but instead of computing the actual MCD in each resample, we use a short-cut that makes use of outlier information gathered from the MCD solution in the original sample. Through an extensive simulation study we will investigate the performance of this method. We will show that it is a robust inference method which yields fairly accurate results and often outperforms the approach based on the asymptotic variance.

In the next section we will give some more details on the MCD and its computation. Section 3 describes the fast bootstrap method, while in Section 4 we will present the simulation results. Finally Section 5 concludes.

2 Minimum covariance determinant estimator

Given a sample $X_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^p$ the MCD estimates the location vector $\boldsymbol{\mu}$ and the scatter matrix Σ. The estimate is determined by selecting that subset of size h (where $\lfloor (n+p+1)/2 \rfloor \leq h \leq n$) with minimal determinant of its empirical covariance matrix, among all possible subsets of size h. The MCD location estimate is then the mean of that subset and the MCD scatter is a multiple of its covariance matrix. The multiplication factor consists of a consistency factor and an additional finite-sample correction factor, to obtain unbiasedness at the normal model (see [4]). The robustness obviously depends on the value of h. The choice $h = \lfloor (n+p+1)/2 \rfloor$ yields the highest possible breakdown point, which asymptotically equals 50%. There is however a trade-off between robustness and efficiency. Therefore we prefer to use $h \approx 0.75n$ which still yields an asymptotic breakdown value of 25% and has far better efficiency.

Until recently the computation time was a major drawback of the MCD. Nowadays there are computationally efficient algorithms available such as the FAST-MCD algorithm [6]. While still relatively time-consuming, it makes the MCD applicable as a routine tool, even in high dimensions. The FAST-MCD algorithm aims to find the h-subset which yields the smallest objective value, being the determinant of its covariance matrix. It is not an exact algorithm, but typically finds a sort of local minimum that is close to the global minimum. A key element is the fact that starting from any h-subset, it is possible to construct another h-subset that yields a lower determinant of its covariance matrix. This so-called *C-step* works as follows:

- suppose we have an h-subset H_0 with corresponding empirical mean and covariance estimates $\overline{X}_0$ and S_0
- compute the distances $d_0(i) = ((\boldsymbol{x}_i - \overline{X}_0)^t S_0^{-1}(\boldsymbol{x}_i - \overline{X}_0))^{\frac{1}{2}}$, $i = 1, \ldots, n$
- set $H_1 := \{\, h$ observations with smallest distance $d_0(i)\}$.

With S_1 the covariance matrix corresponding to H_1 we then have that $|S_1| \leq |S_0|$. Hence applying C-steps successively to an initial subset yields subsets that become more and more concentrated (C stands for "concentration"). The basic idea of the FAST-MCD algorithm is then to take many initial h-subsets, apply C-steps to each of these until convergence and keep the solution with the smallest determinant.

3 A fast bootstrap method

The inference part of the MCD has received little attention in the literature. As noted in section 1, asymptotic variance results are well-known but are certainly not always appropriate for estimating the finite-sample variance of the MCD or for constructing confidence intervals. Bootstrap can be expected to give better results in many situations. Its general idea is to generate a large number B of samples by randomly drawing observations with replacement from the original sample, and to recalculate the estimator for each bootstrap sample. The empirical distribution of the B recalculated estimates is then assumed to be an approximation to the true sample distribution of the estimator. Besides variance estimation the main applications include confidence intervals for the estimands as well as hypothesis tests. These can be constructed using appropriate quantiles of the marginal bootstrap distributions.

In spite of the computational efficiency of the FAST-MCD algorithm, the MCD still has to be regarded as a computer-intensive estimator. Recalculating the algorithm for many bootstrap samples may not be practical, especially for large samples in high dimensions. For example, recalculating the default FAST-MCD algorithm in Matlab for $B = 1000$ samples with $n = 500$ and $p = 10$ would take approximately 1 hour (on a Pentium IV, 1.9 Ghz).

The FAST-MCD algorithm typically starts from up to 500 initial subsets. The main reason for this large number is to have a very high probability that at least one initial subset will not be contaminated with any outliers. Since bootstrap samples are drawn from the original dataset, they consist of observations also present in the original dataset. Often it will be the case that observations identified as outliers by the original MCD estimates, are also outliers when they appear in bootstrap samples. Hence, when we want to compute the MCD estimate in a bootstrap sample we could label those observations as outliers in advance. In this way, a large number of initial subsets is unnecessary. Moreover, we argue that given the effectiveness of C-steps, one initial "clean" solution is sufficient to obtain a good approximation to the MCD solution. In particular we propose the following resampling procedure which intends to mimic the classical bootstrap. After computing the

MCD estimates $\widehat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$ in the original sample, label those observations $\boldsymbol{x}_i$ for which $(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^t \widehat{\Sigma}^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}) > \chi^2_{p,0.975}$ as outliers. Then, as in the classical procedure, draw B bootstrap samples from the complete original data, and for each sample use the following steps to compute an approximate MCD solution:

- Compute the mean and covariance matrix of the observations that were not labeled as outliers in the original dataset (denote by $\widehat{\boldsymbol{\mu}}^*_0$ and $\widehat{\Sigma}^*_0$).
- Compute for all observations in the bootstrap sample the distances w.r.t. $\widehat{\boldsymbol{\mu}}^*_0$ and $\widehat{\Sigma}^*_0$. Use the h observations with the smallest distances to form an initial h-subset.
- Apply further C-steps on this h-subset until convergence to obtain the approximate MCD solution. Note that observations labeled as outliers are allowed to be included in the C-step process.
- In case the number of non-outlying observations h' in the resample is less than h, then work with h'-subsets instead of h-subsets in the previous two steps for this particular resample.

Hence, to compute an MCD approximation in each bootstrap sample, we essentially start with the classical mean and covariance of the observations thought to be non-outlying, and attempt to move from there to the MCD solution through successive C-steps. Often the approximation will be quite rough, but as will be seen in the next section, this method succeeds in accurately mimicking the variability of the MCD estimator. Further on in this paper we will refer to our method as the "short-cut" bootstrap for MCD.

Note also the robustness benefit over the classical bootstrap procedure. When one chooses to apply the MCD, it is assumed that the number of outliers in the dataset is smaller than $n - h$ so that the estimator is able to resist the contamination. When subsequently performing bootstrap, resamples may emerge however that contain more recurrences of the outliers than $n - h$, in which case the recalculated MCD will be adversely affected. This problem has been addressed in general in [8], [9]. The short-cut bootstrap handles such highly contaminated resamples by changing h into h', the number of observations not labeled as outliers. In practice the short-cut bootstrap can be considered to be as robust as the MCD itself. Theoretically it could happen that a bootstrap sample emerges that has too few non-outlying observations to obtain a non-singular covariance matrix. Such situations however will hardly occur in practice and can be handled appropriately if necessary.

4 Simulations

We will show through simulations that in spite of its simple and approximating nature, the method generally performs well. We will compare the short-cut bootstrap with the asymptotic variance results on three inference aspects. The first is variance estimation of the MCD estimates, secondly we will investigate the coverage and the length of derived confidence intervals and finally we will consider an MCD based version of Hotelling's T^2 test for the mean of the data.

Simulations were performed for sample sizes $n = 50, 200$ and 500, dimensions $p = 2, 5$ and 10, considering the following cases:

- multivariate normal $N(\mathbf{0}, I_p)$
- multivariate Student with 3 d.f. (T_3)
- contaminated normal: 85% $N(\mathbf{0}, I_p)$, 15% $N(5\sqrt{\chi^2_{p,0.99}}[1\ldots1]', I_p)$.

For each case 1000 random datasets were generated. On each dataset we computed the MCD and subsequently performed the short-cut bootstrap with $B = 1000$. We considered both the 25% breakdown $(h \approx 0.75n)$ and the 50% breakdown $(h \approx 0.50n)$ MCD. Due to limited space we only report the results for the 25% breakdown MCD. Also, we focus here on the MCD location and omit results for the MCD scatter. The simulations showed that the bootstrap yields nice results concerning the MCD scatter in normal data, but unbiased robust estimation of the scale remains difficult and therefore the bootstrap does not work well for the MCD scatter in contaminated data.

A straightforward application of the bootstrap is the estimation of the variance of the estimators. Table 1 lists the bootstrap estimates and the asymptotic estimates, averaged over the 1000 samples, for the different cases. The values shown are in turn averages over the p components. Note that for data from $N(\boldsymbol{\mu}, \Sigma)$, the asymptotic variance for the i-th component of the MCD location is given by $\kappa\Sigma_{ii}$, where the factor κ depends on p and (the limit of) h/n, and can be found in [1]. For the asymptotic estimates (ASV) we used this κ and replaced Σ by the MCD scatter estimate. The bootstrap and the asymptotic estimates are compared to the Monte Carlo estimates over the 1000 samples. It can be seen that the short-cut bootstrap yields quite accurate variance estimates in all cases considered. This is somewhat surprising since one might expect an inherent underestimation. Suppose we would not perform C-steps and instead use the initial approximation $\widehat{\boldsymbol{\mu}}^*_0$ in each sample, then the simulations showed that smaller variance estimates would emerge, as expected. The C-steps are meant to increase the variance by improving the approximations to the MCD in the bootstrap samples. Apparently those C-steps succeed in enough improvement such that the bootstrap variance becomes an accurate estimate of the true variance of the MCD. Table 1 also shows that the asymptotic estimates are only accurate for large samples in case of normal data.

We used the short-cut bootstrap to construct 95% confidence intervals for the components of the population mean. In particular we considered percentile intervals formed by the 2.5% and 97.5% empirical quantiles of the recalculated estimates, for each component. The intervals associated with the ASV were based on a normal approximation, using the asymptotic estimates for the variance. Table 2 shows the percentage of the simulated datasets for which the corresponding intervals contained the value 0 (the population

	$p=2$			$p=5$			$p=10$		
n	MC	Boot	ASV	MC	Boot	ASV	MC	Boot	ASV
normal data									
50	43.5	39.1	55.6	33.7	33.1	44.4	28.1	27.1	43.1
200	12.5	11.0	12.8	8.91	9.13	9.89	7.98	8.67	8.97
500	4.82	4.67	5.05	3.61	3.70	3.84	3.32	3.38	3.44
Student data (T_3)									
50	40.3	40.7	74.7	33.6	33.2	59.3	32.0	32.0	57.7
200	9.73	9.94	17.2	8.47	8.74	12.8	7.55	7.42	11.2
500	3.91	3.97	6.74	3.28	3.43	4.97	3.07	2.97	4.31
normal data w/ 15% outliers									
50	31.9	34.2	74.1	27.6	29.4	52.1	26.5	26.5	47.2
200	9.50	9.24	16.9	7.96	8.15	11.5	7.36	7.87	9.87
500	3.81	3.72	6.66	3.17	3.25	4.49	2.93	3.11	3.82

Table 1: Variance estimates (×1000) for 25% breakdown MCD.

mean). The respective lengths of the intervals are also given. Results for the Student data were similar to those for the contaminated normal and are omitted here, as are the results for $n = 200$. We see that the coverage for the bootstrap intervals is generally close to the nominal value, although the intervals are somewhat conservative. The asymptotic variance intervals in each case are longer than the bootstrap intervals, which is in accordance with the overestimation of the variance as seen in Table 1. In some cases (e.g. $p = 2$ for normal data) we have that bootstrap intervals have higher coverage than the ASV intervals while still being significantly shorter. This can be understood by noting again that the approximations in the bootstrap samples each time start from $\widehat{\boldsymbol{\mu}}_0^*$, the mean of the non-outlying observations, which is a more efficient estimator of the population mean than the MCD itself. In fact, if we were to use $\widehat{\boldsymbol{\mu}}_0^*$ (omitting the C-steps), we would obtain intervals that are even much shorter. The simulations showed however that the coverage of those intervals is too low, especially for small samples.

Finally we consider a hypothesis test for the mean $\boldsymbol{\mu}$ of the underlying distribution. The well-known Hotelling T^2 test uses the classical mean $\overline{X}$ and covariance S to construct the test statistic $T^2 = n(\overline{X} - \boldsymbol{\mu}_0)^t S^{-1}(\overline{X} - \boldsymbol{\mu}_0)$. For normal data the distribution of T^2 under the null hypothesis $H_0 : \boldsymbol{\mu}_0 = \boldsymbol{\mu}$ is a multiple of an F-distribution, not depending on $\boldsymbol{\mu}$. A robust version of Hotelling's test can be obtained by replacing the classical estimators by the MCD. The distribution of the test statistic under H_0 can then be determined through the bootstrap, assuming that this distribution does not depend on $\boldsymbol{\mu}$. For example, the 5% critical value for the test would be given by the 95% quantile of the recalculated statistics $n(\widehat{\boldsymbol{\mu}}^* - \widehat{\boldsymbol{\mu}})^t(\widehat{\Sigma}^*)^{-1}(\widehat{\boldsymbol{\mu}}^* - \widehat{\boldsymbol{\mu}})$, where $\widehat{\boldsymbol{\mu}}$ is the MCD estimate of the original sample and $\widehat{\boldsymbol{\mu}}^*$ and $\widehat{\Sigma}^*$ denote the recalculated estimates of the short-cut bootstrap. ¿From the asymptotic normality

	$p = 2$		$p = 5$		$p = 10$	
n	Boot	ASV	Boot	ASV	Boot	ASV
normal data						
50	96.2 (0.766)	95.7 (0.910)	95.8 (0.708)	96.1 (0.817)	94.0 (0.642)	97.5 (0.807)
500	95.8 (0.268)	94.9 (0.278)	96.6 (0.239)	95.7 (0.243)	96.5 (0.228)	95.2 (0.230)
normal data w/ 15% outliers						
50	96.7 (0.720)	99.1 (1.056)	95.4 (0.669)	98.7 (0.888)	94.4 (0.636)	98.7 (0.846)
500	95.3 (0.239)	98.8 (0.320)	96.0 (0.224)	98.0 (0.262)	96.2 (0.219)	97.4 (0.242)

Table 2: Coverage and length of univariate 95% confidence intervals based on 25% breakdown MCD.

	$p = 2$		$p = 5$		$p = 10$	
n	Boot	ASV	Boot	ASV	Boot	ASV
normal data						
50	6.3 - 1.9	6.6 - 1.7	6.2 - 0.9	11.1 - 4.7	4.9 - 0.7	18.0 - 10.0
500	5.4 - 1.0	6.2 - 1.8	4.3 - 0.8	5.0 - 1.5	4.2 - 0.8	6.3 - 1.5
normal data w/ 15% outliers						
50	3.9 - 0.5	0.6 - 0.1	2.3 - 0.2	1.6 - 0.2	0.7 - 0.1	4.6 - 1.4
500	4.9 - 1.0	0.6 - 0.0	3.9 - 0.9	1.1 - 0.1	2.7 - 0.8	1.4 - 0.1

Table 3: T^2 test based on 25% breakdown MCD; percentage of erroneous rejections of H_0 (levels 5% - 1%)

of the MCD location and the consistency of the MCD scatter it follows that for large samples the MCD based T^2 under H_0 should be distributed approximately as $\kappa\chi^2_p$. As before κ is obtained from the asymptotic variance as given in [1]. Table 3 lists the observed probability that the robust T^2 statistic yields a value above the 5% and 1% critical value. The short-cut bootstrap is again compared to the ASV approach ($\kappa\chi^2_p$ distribution). For normal data the results show that the critical values of the test are estimated fairly accurately by the bootstrap. Using the asymptotic distribution, the critical values are somewhat too low in case of normal data, while on the other hand they are far too high for the contaminated data. The bootstrap performs much better than the ASV in case of contamination, although results for small samples are not too good. Results for the Student data again were similar to those for the contaminated data.

5 Conclusion

Asymptotic estimates are not always appropriate to obtain inference for the MCD. Alternatively, classical bootstrap is often extremely time-consuming and has some robustness problems. Therefore we investigated a bootstrap procedure which is computationally simple as well as robust. Simulations showed that the method accurately estimates the variance of the MCD, gives relatively short confidence intervals with good coverage, and can be used to perform hypothesis tests for the mean. The method outperforms the approach based on asymptotic estimates for all these inference procedures.

References

[1] Butler R.W., Davies P.L., Juhn M. (1993). *Asymptotics for the minimum covariance determinant estimator.* Ann. Statist. **21**, 1385 – 1400.

[2] Croux C., Haesbroeck G. (1999). *Influence function and efficiency of the minimum covariance determinant scatter matrix estimator.* J. Multivariate Anal. **71**, 161 – 190.

[3] Efron B. (1979). *Bootstrap methods: another look at the jackknife.* Ann. Statist. **7**, 1 – 26.

[4] Pison G., Van Aelst S., Willems G. (2002). *Small sample corrections for LTS and MCD.* Metrika **55**, 111 – 123.

[5] Rousseeuw P.J. (1985). *Multivariate estimation with high breakdown point.* Mathematical Statistics and Applications , Reidel Publishing Company, Dordrecht, 283 – 297.

[6] Rousseeuw P.J., Van Driessen K. (1999). *A fast algorithm for the minimum covariance determinant estimator.* Technometrics **41**, 212 – 223.

[7] Salibian-Barrera M., Zamar R. (2002). *Bootstrapping robust estimates of regression.* Ann. Statist. **30**, 556 – 582.

[8] Singh K. (1998). *Breakdown theory for bootstrap quantiles.* Ann. Statist. **26**, 1719 – 1732.

[9] Stromberg A.J. (1997). *Robust covariance estimates based on resampling.* J. Statist. Plann. Inference **57**, 321 – 334.

[10] Van Aelst S., Willems G. (2002). *Robust bootstrap for S-estimators of multivariate regression.* Statistics in Industry and Technology: Statistical Data Analysis, 201 – 212.

Address: G. Willems, Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium
S. Van Aelst, Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium
E-mail: `gert.willems@ua.ac.be, stefan.vanaelst@ugent.be`

CONFIDENCE REGION FOR PARAMETERS IN REPLICATED ERRORS IN VARIABLES MODEL

Gejza Wimmer, Viktor Witkovský and Alexandr Savin

Key words: Errors in variables (EIV) regression model, measurements with errors, univariate calibration, Kenward-Roger approximation.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: We consider replicated regression model with errors in variables (EIV model), also known as the measurement error model. The calibration problem is closely related to the considered model. In this model, we have derived the Kenward-Roger type of the confidence region for the unknown regression parameters, which is suggested to be used with small sample sizes. Our simulation study approved good statistical properties of the proposed confidence region.

1 Introduction

Here we consider the errors in variables model (EIV model), also known as the measurement error model (see [2]), which is a generalization of simple linear regression

$$Y_i = a + bx_i + \varepsilon_i$$

where we do not assume that the xs are known. Instead, we observe only realizations of random variables X_i, $i = 1, \dots, n$, whose mean equals to the (unknown parameter) μ_i. This model is also closely related to the calibration curve (identification of the calibration curve), which roughly speaking, expresses the relation between the results of measuring the same object (quantity) by two measuring devices A and B, respectively.

Under the term *calibration problem* we will understand here the task of construction of the calibration curve, in particular, the construction of the confidence region for the unknown regression parameters, see e.g. [6]. In more general, the calibration problem was formulated in e.g. [10] and [1], as a problem of predicting one set of variables from another set.

We are looking for solution to above mentioned problem under the following circumstances:

(i) The measurement result x_i obtained by the measuring device A is a realization of normally distributed random variable X_i, i.e. $X_i \sim N(\mu_i, \sigma_x^2)$, $i = 1, \dots, n$, where the mean value $\mathcal{E}(X_i) = \mu_i$ is the errorless (ideal) measurement result made by the measuring device A and σ_x^2 is the (unknown) dispersion of X_i, for all $i = 1, \dots, n$ (σ_x is the standard uncertainty of the measuring device A).

(ii) The measurement result y_i obtained by the measuring device B is a realization of normally distributed random variable Y_i, i.e. $Y_i \sim N(\nu_i, \sigma_y^2)$, $i = 1, \ldots, n$, where the mean value $\mathcal{E}(Y_i) = \nu_i$ is the errorless (ideal) measurement result made by the measuring device B and σ_y^2 is the (unknown) dispersion of Y_i for all $i = 1, \ldots, n$ (σ_y is the standard uncertainty of the measuring device B).

(iii) All measurements are mutually independent.

(iv) In typical range of values of μ and ν (the range of interest) we assume that the true, however unknown, calibration curve is a linear function, i.e. $\nu = a + b\mu$, with (unknown) parameters a, b.

(v) For estimation of the parameters of the calibration curve and for obtaining the confidence region of the parameters we accomplish a pre-planned calibration experiment with measurements made by both measuring devices, A and B, respectively, on a set of $n \geq 3$ suitably chosen objects (quantities of interest) $v_1, v_2, \ldots, v_n$, repeated $m \geq 2$ times for each object (quantity).

2 The replicated and linearized EIV model

Let us denote the vectors of errorless measurement results made by the measuring devices A and B, respectively, by $\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$ and $\nu = (\nu_1, \nu_2, \ldots, \nu_n)'$. Let the vector of measurements made by the measuring device A be $X_{n,1} \sim N(\mu; \sigma_x^2 I_{n,n})$. Let the vector of measurements made by the measuring device B be $Y_{n,1} \sim N(\nu; \sigma_y^2 I_{n,n})$. This yields a model

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu \\ \nu \end{pmatrix} \begin{pmatrix} \sigma_x^2 I & 0 \\ 0 & \sigma_y^2 I \end{pmatrix}\right], \tag{1}$$

with nonlinear constraints on the parameters

$$\nu = a 1_{n,1} + b\mu, \tag{2}$$

where $1_{n,1} = (1, 1, \ldots, 1)'$. First, we will linearize the model using the Taylor series expansion about $\mu_0 = (\mu_{01}, \mu_{02}, \ldots, \mu_{0n})'$ and b_0 (some values close to the true parameters μ and b). So, $\mu = \mu_0 + \delta\mu$, $b = b_0 + \delta b$ and the new parameters are $\delta\mu = (\delta\mu_1, \delta\mu_2, \ldots, \delta\mu_n)'$, ν, a, δb, σ_x^2, σ_y^2. We obtain the model

$$\begin{pmatrix} X - \mu_0 \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \delta\mu \\ \nu \end{pmatrix}, \begin{pmatrix} \sigma_x^2 I & 0 \\ 0 & \sigma_y^2 I \end{pmatrix}\right], \tag{3}$$

with linear constraints

$$b_0\mu_0 + (b_0 I \vdots -I) \begin{pmatrix} \delta\mu \\ \nu \end{pmatrix} + (1, \mu_0) \begin{pmatrix} a \\ b_0 + \delta b \end{pmatrix} = 0. \tag{4}$$

According to (v) we repeat independently the experiment (3) m times. If $\xi_1, \ldots, \xi_m$ are the independent replications of

$$\xi = \begin{pmatrix} X - \mu_0 \\ Y \end{pmatrix},$$

i.e. the vector $\underline{\xi} = (\xi_1', \ldots, \xi_m')'$ is the vector of all measurements, then the replicated model is

$$\underline{\xi} \sim N\left[(1_{m,1} \otimes I_{2n,2n}) \begin{pmatrix} \delta\mu \\ \nu \end{pmatrix}, I_{m,m} \otimes (\sigma_x^2 V_1 + \sigma_y^2 V_2)\right], \tag{5}$$

with constraints (4), where

$$V_1 = \begin{pmatrix} I_{n,n} & 0 \\ 0 & 0 \end{pmatrix}, \quad V_2 = \begin{pmatrix} 0 & 0 \\ 0 & I_{n,n} \end{pmatrix},$$

and $\otimes$ means the Kronecker product.

It can be shown (see e.g. [6], [8]) that the BLUEs (best linear unbiased estimators) of the parameters $\mu = \mu_0 + \delta\mu$, ν, a, δb in the linearized model (5) with constraints (4), assuming that the model holds true, are

$$\hat{\mu} = \bar{X} + \frac{b_0\sigma_x^2}{b_0^2\sigma_x^2 + \sigma_y^2} M_{1,\mu_0}(\bar{Y} - b_0\bar{X}), \tag{6}$$

$$\hat{\nu} = \bar{Y} - \frac{\sigma_y^2}{b_0^2\sigma_x^2 + \sigma_y^2} M_{1,\mu_0}(\bar{Y} - b_0\bar{X}), \tag{7}$$

$$\begin{pmatrix} \hat{a} \\ \hat{\delta b} \end{pmatrix} = \begin{pmatrix} n & 1'\mu_0 \\ \mu_0'1 & \mu_0'\mu_0 \end{pmatrix}^{-1} \begin{pmatrix} 1'(\bar{Y} - b_0\bar{X}) \\ \mu_0'(\bar{Y} - b_0\bar{X}) \end{pmatrix}, \tag{8}$$

with associated covariance matrices

$$\begin{aligned} cov\,(\hat{\mu}) &= \frac{\sigma_x^2}{m} I - \frac{b_0^2\sigma_x^4}{m(b_0^2\sigma_x^2 + \sigma_y^2)} M_{1,\mu_0}, \\ cov\,(\hat{\nu}) &= \frac{\sigma_y^2}{m} I - \frac{\sigma_y^4}{m(b_0^2\sigma_x^2 + \sigma_y^2)} M_{1,\mu_0}, \\ cov\,(\hat{\mu}, \hat{\nu}) &= \frac{b_0\sigma_x^2\sigma_y^2}{m(b_0^2\sigma_x^2 + \sigma_y^2)} M_{1,\mu_0}, \\ cov\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} &= \frac{b_0^2\sigma_x^2 + \sigma_y^2}{m} \begin{pmatrix} n & 1'\mu_0 \\ \mu_0'1 & \mu_0'\mu_0 \end{pmatrix}^{-1}, \end{aligned} \tag{9}$$

$$cov\left[\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}, \begin{pmatrix} \hat{\mu} \\ \hat{\nu} \end{pmatrix}\right] = -\frac{1}{m} \begin{pmatrix} n & 1'\mu_0 \\ \mu_0'1 & \mu_0'\mu_0 \end{pmatrix}^{-1} \begin{pmatrix} b_0\sigma_x^2 1' & -\sigma_y^2 1' \\ b_0\sigma_x^2\mu_0' & -\sigma_y^2\mu_0' \end{pmatrix},$$

where

$$M_{1,\mu_0} = I_{n,n} - (1, \mu_0) \begin{pmatrix} n & 1'\mu_0 \\ \mu_0'1 & \mu_0'\mu_0 \end{pmatrix}^{-1} \begin{pmatrix} 1' \\ \mu_0' \end{pmatrix},$$

$\hat{b} = b_0 + \hat{\delta b}$ and $\bar{X} = \frac{1}{m}\sum_{j=1}^m X_j$, $X_j = (X_{j1}, \ldots, X_{jn})'$, $\bar{Y} = \frac{1}{m}\sum_{j=1}^m Y_j$, $Y_j = (Y_{j1}, \ldots, Y_{jn})'$.

However, the above estimators and covariance matrices depend on the unknown variance components (σ_x^2, σ_y^2) which should be estimated. Here we propose to use (iterated) $(\sigma_{x0}^2, \sigma_{y0}^2)$-MINQUE ($(\sigma_{x0}^2, \sigma_{y0}^2)$-locally minimum norm quadratic unbiased estimator), see e.g. [9], [6], of σ_x^2 and σ_y^2, respectively, and its covariance matrix. According to [6], see also [9] and [7], the $(\sigma_{x0}^2, \sigma_{y0}^2)$-MINQUE of $(\sigma_x^2, \sigma_y^2)'$ in linear model (5) – (4) is given by

$$\begin{pmatrix} \hat{\sigma}_x^2 \\ \hat{\sigma}_y^2 \end{pmatrix} = \frac{1}{n(m-1)} \left[I_{2,2} - c_0 \begin{pmatrix} b_0^4 \sigma_{x0}^4 & b_0^2 \sigma_{x0}^4 \\ b_0^2 \sigma_{y0}^4 & \sigma_{y0}^4 \end{pmatrix} \right] \begin{pmatrix} \hat{\kappa_1} \\ \hat{\kappa_2} \end{pmatrix}, \tag{10}$$

where

$$\begin{aligned} c_0 &= \frac{n-2}{(b_0^4 \sigma_{x0}^4 + \sigma_{y0}^4)(mn-2) + 2b_0^2 \sigma_{x0}^2 \sigma_{y0}^2 (m-1)n}, \\ \hat{\kappa_1} &= \sum_{j=1}^{m} (X_j - \bar{X})'(X_j - \bar{X}) + m(\bar{X} - \hat{\mu})'(\bar{X} - \hat{\mu}), \\ \hat{\kappa_2} &= \sum_{j=1}^{m} (Y_j - \bar{Y})'(Y_j - \bar{Y}) + m(\bar{Y} - \hat{\nu})'(\bar{Y} - \hat{\nu}). \end{aligned}$$

The covariance matrix (locally at $(\sigma_{x0}^2, \sigma_{y0}^2)$) of this estimator is

$$\begin{aligned} W &= \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} = cov \left(\begin{pmatrix} \hat{\sigma}_x^2 \\ \hat{\sigma}_y^2 \end{pmatrix} \middle| \sigma_{x0}^2, \sigma_{y0}^2 \right) \\ &= \frac{2}{n(m-1)} \left[I_{2,2} - c_0 \begin{pmatrix} b_0^4 \sigma_{x0}^4 & b_0^2 \sigma_{x0}^4 \\ b_0^2 \sigma_{y0}^4 & \sigma_{y0}^4 \end{pmatrix} \right] \begin{pmatrix} \sigma_{x0}^4 & 0 \\ 0 & \sigma_{y0}^4 \end{pmatrix}. \end{aligned} \tag{11}$$

Of course, all the estimators strongly depend on the chosen initial values μ_0, b_0, σ_{x0}^2, and σ_{y0}^2 and on the quality of linearization of the model at these selected initial values. If there is no specific prior information on the true values of the parameters, a natural choice of the initial values, estimated from the measured data, could be the following:

$$\begin{aligned} \mu_0 &= \bar{X}, \\ b_0 &= \frac{n\bar{X}'\bar{Y} - (1'\bar{X})(1'\bar{Y})}{n\bar{X}'\bar{X} - (1'\bar{X})^2}, \\ \sigma_{x0}^2 &= \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ji} - \bar{X}_i)^2, \\ \sigma_{y0}^2 &= \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (Y_{ji} - \bar{Y}_i)^2, \end{aligned}$$

with $\bar{X}_i = \frac{1}{m} \sum_{j=1}^{m} X_{ij}$ and $\bar{Y}_i = \frac{1}{m} \sum_{j=1}^{m} Y_{ij}$. Further we compute $\hat{a}$, $\hat{b}$ from (8), $\hat{\mu}$ from (6), $\hat{\nu}$ from (7), $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ from (10).

The estimation procedure could be iterated in such a way until convergence is reached: For given estimates $\hat{\mu}^{(k)}$ and $\hat{b}^{(k)}$ estimate $\hat{b}^{(k+1)}$ from (8). Using this $\hat{b}^{(k+1)}$ and $\hat{\mu}^{(k)}$, $\hat{\sigma}_x^{2(k)}$, $\hat{\sigma}_y^{2(k)}$, estimate $\hat{\mu}^{(k+1)}$ from (6) and $\hat{\nu}^{(k+1)}$

from (7). Finally, using $\sigma_x^{2(k)}$, $\hat{\sigma}_y^{2(k)}$, and $\hat{b}^{(k+1)}$, $\hat{\mu}^{(k+1)}$, $\hat{\nu}^{(k+1)}$, estimate $\sigma_x^{2(k+1)}$ and $\hat{\sigma}_y^{2(k+1)}$ from (10). Typically, this iteration procedure converges very quickly, and after few iterations the estimates settle down at stable values (in our simulation study the maximamum number of iterations was 7).

3 Confidence region for the regression parameters $(a, b)'$

Note that, based on the linearized model (5) – (4), we have

$$\bar{Y} - b_0\bar{X} \sim N\left[(1, \mu_0)\begin{pmatrix} a \\ \delta b \end{pmatrix}, \frac{b_0^2\sigma_x^2 + \sigma_y^2}{m} I_{n,n}\right]. \tag{12}$$

From (12), and also from (8) and (9), it immediately follows that

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} \sim N\left[\begin{pmatrix} a \\ b \end{pmatrix}, \frac{b_0^2\sigma_x^2 + \sigma_y^2}{m}\begin{pmatrix} n & 1'\mu_0 \\ \mu_0'1 & \mu_0'\mu_0 \end{pmatrix}^{-1}\right]. \tag{13}$$

Model (12) is a special case of Gaussian linear regression model with linear covariance structure depending on two variance components σ_x^2 and σ_y^2.

For construction of the confidence region for the regression parameters $(a, b)'$ we suggest to utilize a method for small sample inference for fixed effects proposed by Kenward and Roger in [5]. We apply the procedure on model (12), using the MINQUE-type estimator of variance components, see (10) and its covariance matrix (11), to obtain the scaled Wald-type statistic and the F-approximation to its sampling distribution:

$$F = \frac{1}{2}\begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix}' \hat{\Phi}_A^{-1}\begin{pmatrix} \hat{a} - a \\ \hat{b} - b \end{pmatrix}. \tag{14}$$

Here by $\Phi = \Phi(\sigma_x^2, \sigma_y^2)$ we denote the true (or asymptotic) covariance matrix of $(\hat{a}, \hat{b})'$ given by (13), and $\hat{\Phi}_A$ is an adjusted estimator of the small sample variance-covariance matrix of $(\hat{a}, \hat{b})'$ based on estimated variance components, (10) and (11),

$$\hat{\Phi}_A = \hat{\Phi} - \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij}\frac{\partial^2\Phi}{\partial\sigma_i^2\partial\sigma_j^2}$$

where $\hat{\Phi} = \Phi(\hat{\sigma}_x^2, \hat{\sigma}_y^2)$, $w_{ij} = \{W\}_{ij}$, W given by (11) and $\sigma_1^2 = \sigma_x^2, \sigma_2^2 = \sigma_y^2$. In model (12) we obtain

$$\sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij}\frac{\partial^2\Phi}{\partial\sigma_i^2\partial\sigma_j^2} = 0,$$

so $\hat{\Phi}_A = \hat{\Phi}$.

Further, following the considerations in [5], we approximate F in (14) in such a way that λF has Fisher-Snedecor F-distribution with 2 and u degrees of freedom. In model (12) we obtain

$$\lambda = 1 \tag{15}$$

and

$$u = (mn-2) + \frac{2n(m-1)b_0^2\hat{\sigma}_x^2\hat{\sigma}_y^2}{b_0^4\hat{\sigma}_x^4 + \hat{\sigma}_y^4}, \tag{16}$$

where b_0 is replaced by its estimated value $\hat{b}^{(k-1)}$ if the k-steps of the iterated algorithm were used. Note that the estimated degrees of freedom u range from minimal value $mn-2$ up to the maximum $2mn-n-2$ if $b_0^2\hat{\sigma}_x^2 = \hat{\sigma}_y^2$. So, the $(1-\alpha)-$confidence region for $(a,b)'$ is given by

$$\mathcal{C}_{(1-\alpha)} = \left\{ \begin{pmatrix} a \\ b \end{pmatrix} : F \leq F_{2,u}(1-\alpha) \right\}, \tag{17}$$

where F is given by (14) and $F_{r,s}(1-\alpha)$ is the $(1-\alpha)-$quantile of the Fisher-Snedecor F-distribution with r and s degrees of freedom.

4 Simulation study

In order to check the basic statistical properties of the proposed confidence region for the regression parameters $(a,b)'$ of the EIV model (1) we have performed the following simulation study.

Assuming the model (1) – (2) we have used the following values of the parameters in the simulation study:

- $\sigma_y^2 = 1$ and $\sigma_x^2 \in \{0.001, 0.01, 0.1, 1\}$.
- For each $n \in \{3, 5, 10\}$ we have used $\mu = [1 : +2 : 2n-1]$.
- For $a = 0$ and $b \in \{\sqrt{1000}, 10, \sqrt{10}, 1, \frac{1}{2}\}$ we have calcutated $\nu = a1_{n,1} + b\mu$. Except the smallest value $b = \frac{1}{2}$, the parameter b was chosen in such a way that in combination with particular σ_x^2 and σ_y^2, the equality $b^2\sigma_x^2 = \sigma_y^2$ holds true.
- The independent measurements were replicated $m \in \{2, 5, 20\}$ times.

For each combination of the parameters we generated $N = 10000$ realizations of $\underline{\xi}$ given by (5) and by using the iterated algorithm for calculation of the F-statistic and u — the approximate degrees of freedom, we calculated the empirical probability that the true parameter $(a,b)'$ was covered by the nominal 95%-confidence region (17).

The results are presented in Figure 1. For each combination of the variances σ_x^2 and σ_y^2 and for each number of replications m we have plotted the empirical coverage probabilities in groups of three sets (for $n = 3$, $n = 5$, and $n = 10$) of five values (the coverage probabilities calculated for different b-s). We can see that for $n > 3$ the relative error is less than one per-cent in all considered situations.

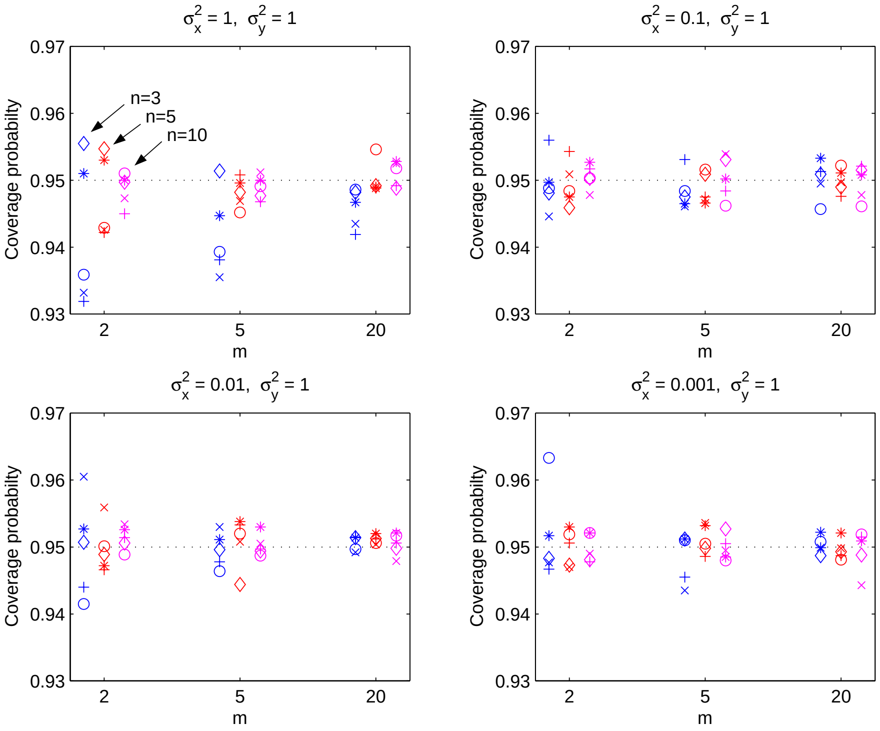


Figure 1: The empirical coverage probabilities of the 95% confidence regions (17) calculated with the iterated MINQUE estimates of variance components, based on 10,000 Monte Carlo runs for each specific design. Here we use the symbol $\circ$ for designs with $b = \sqrt{1000}$, $\times$ for designs with $b = 10$, $+$ for designs with $b = \sqrt{10}$, $*$ for designs with $b = 1$, and $\diamond$ for designs with $b = 1/2$. The dotted line shows the nominal 95% level.

5 Conclusions

In this paper we have suggested to use the Kenward-Roger type of small sample inference to construct the confidence region for the regression parameters in the replicated errors in variables (EIV) linear regression model. The simulation results imply that the empirical coverage probabilities of this confidence region are very close to the nominal level if n, the number of measured objects (quantities) is greater than 3, even if the number of replications is very small, e.g. $m = 2$. During the simulation study we observed that the quality of the confidence region strongly depended on the quality of linearization of the model, which on the other hand depended on the distribution of the true values of μ and the variance component σ_x^2.

References

[1] Brown P.J. (1960). *Measurement, regression, and calibration.* Clarendon Press, Oxford.

[2] Casella G., Berger R.L. (1999). *Statistical inference.* (Second edition). Duxbury Advanced Series, Belmont CA.

[3] Harville D.A., Jeske D.R. (1992). *Mean square error of estimation or prediction under a general linear model.* Journal of the American Statistical Association **87**, 724 – 731.

[4] Kackar A.N., Harville D.A. (1984). *Approximations for standard errors of estimation for fixed and random effects in mixed linear models.* Journal of the American Statistical Association **79**, 853 – 862.

[5] Kenward M.G., Roger J.H. (1997). *Small sample inference for fixed effects from restricted maximum likelihood.* Biometrics **53**, 983 – 997.

[6] Kubáček L., Kubáčková L. (2000). *Statistika a metrologie.* (In Czech). Univerzita Palackého v Olomouci, Olomouc.

[7] Kubáček L., Kubáčková L., Volaufová J. (1995). *Statistical models with linear structures.* VEDA, Bratislava.

[8] Kubáčková L. (1992). *Foundations of experimental data analysis.* CRC Press, Boca Raton – Ann Arbor – London – Tokyo.

[9] Rao C. R., Kleffe J. (1988). *Estimation of variance components and applications.* North-Holland, Amsterdam – New York – Oxford-Tokyo.

[10] Scheffé H. (1973). *A statistical theory of calibration.* The Annals of Statistics **1**, 1 – 37.

Acknowledgement: The research was supported by the grants 1/0264/03 and 2/4026/04 of the Slovak Scientific Grant Agency VEGA.

Address: G. Wimmer, Faculty of Science, Masaryk University, Janáčkovo nám. 2a, 662 95 Brno, Czech Republic, and Institute of Mathematics and Computer Science, Banská Bystrica, Slovak Republic

V. Witkovský, Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovak Republic

A. Savin, Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovak Republic

E-mail: `wimmer@mat.savba.sk, witkovsky@savba.sk`
`savin@savba.sk`

COMPSTAT'2004 Symposium

MATLAB ALGORITHM TDIST: THE DISTRIBUTION OF A LINEAR COMBINATION OF STUDENT'S T RANDOM VARIABLES

Viktor Witkovský

Key words: Distribution function, cdf, pdf, quantiles, characteristic function, Student's t random variables, linear combination, convolution, numerical inversion, Matlab algorithm.

COMPSTAT 2004 section: Algorithms.

Abstract: The Matlab algorithm *TDIST* computes the cumulative distribution function (cdf), the probability density function (pdf), the quantile function (qf), and the characteristic function (chf) of a linear combination of independent central Student's t random variables with small degrees of freedom ($1 \leq df \leq 100$) and/or standard normal random variables. The algorithm is available from the author or from the web page: *http://www.mathworks.com*, see the section *Matlab Central > File Exchange > Statistics and Probability > TDIST*.

1 Introduction

One of the desired goals of nowadays computational probability and statistics is according to me to have a powerful tool (calculator) to evaluate the distribution of any reasonable function of arbitrary random variables. The need for such a tool comes directly from the modern theory of statistics, see e.g. Weerahandi [10], where the inferential procedures are frequently based on methods which lead to the problem of evaluation of the distribution of a well specified function of several random variables.

From theoretical point of view, the ideal solution would be to have the exact (closed form, if it exists) distribution of the function of random variables. On a good way to be such a tool to automate the probability calculations is the *APPL – A Probability Programming Language*, based on computer algebra system Maple, see Glen et al. [5] and Glen et al. [6], and the *mathStatica*, based on the system Mathematica, see Rose and Smith [8]. However, the above mentioned tools based on computer algebra systems have some drawbacks and limitations. Namely, the result (if any) is often expressed as a complex combination of special functions and/or unevaluated integrals.

From statistical applications point of view, a fast and precise numerical procedure to compute the distribution of the function of random variables would be preferred. Here we present one particular result: The Matlab algorithm *TDIST* for computing the cumulative distribution function (cdf), the probability density function (pdf), the quantile function (qf) (also known as the inverse distribution function), and the characteristic function (chf) of an arbitrary linear combination of independent central Student's t random variables with small degrees of freedom ($1 \leq df \leq 100$) and/or standard normal (Gaussian) random variables. The algorithm is based on the method discussed by Witkovský [12], and is closely related to the method for computing the distribution of a linear combination of independent chi-square random variables suggested by Imhof [7], see also Davies [2], and with the method for computing the distribution of a linear combination of independent inverted gamma variables suggested by Witkovský [13]. All the above procedures are based on the numerical inversion of the characteristic function, for more details see Gil-Pelaez [4]. The algorithm *TDIST* requires evaluation of the modified Bessel function of the second kind (Bessel K function), which is available in Matlab through the integrated package for Bessel functions due to Amos [1].

The distribution of a linear combination of independent Student's t random variables is important for many statistical applications. For instance, Fairweather [3] discussed and suggested the method of obtaining an exact confidence interval for the common mean of several normal populations which is based on pivotal quantity $T = \sum_{i=1}^{k} \lambda_i t_{\nu_i}$, which is a weighted linear combination of independent Student's t random variables with ν_i, $i = 1, \ldots, k$, degrees of freedom. The Behrens-Fisher distribution of the test statistic for testing equality of the means of two normal populations, without making any assumption about the variances, is that of a linear combination of two independent Student's t random variables, for more details see e.g. Weerahandi [10] and Witkovský [14].

The well known method how to approximate the distribution of the linear combination of independent Student's t random variables, say

$$T = \sum_{i=1}^{k} \lambda_i t_{\nu_i}, \tag{1}$$

is to approximate it by the distribution of the c multiple of single Student's t random variable with ν degrees of freedom, say ct_ν, where ν and $c > 0$ are to be determined by equating the second and fourth moments of ct_ν to those of T. In particular, if we additionally assume that the degrees of freedom $\nu_i > 4$ for all $i = 1, \ldots, k$, then we get

$$\nu = 4 + \frac{1}{\sum_{i=1}^{k} \lambda_i^2/(\nu_i - 4)}, \quad c = \sqrt{\frac{\nu - 2}{\nu \sum_{i=1}^{k} (\nu_i - 2)/\nu_i}}, \tag{2}$$

for more details see also Welch [11] and Fairweather [3].

An alternative approach was suggested by Walker and Saw [9] who derived the exact distribution of an arbitrary linear combination of Student's t random variables with *odd degrees of freedom*. The method of expressing this distribution as a mixture of t distributions (cdfs) relies on the fact that the characteristic functions of such variables are proportional to the polynomial functions.

2 The method

Consider the random variable $T = \sum_{i=1}^{k} \lambda_i t_{\nu_i}$, a linear combination of independent Student's t_{ν_i} random variables with ν_i, $i = 1, \ldots, k$, degrees of freedom and let $\phi_i(t)$ denote the characteristic function of t_{ν_i}. The characteristic function of T is

$$\phi_T(t) = \phi_1(\lambda_1 t) \cdots \phi_k(\lambda_k t). \tag{3}$$

where

$$\phi_i(\lambda_i t) = \frac{1}{2^{\frac{\nu_i}{2}-1}\Gamma(\frac{\nu_i}{2})} \left(\nu_i^{\frac{1}{2}}|\lambda_i t|\right)^{\frac{\nu_i}{2}} K_{\frac{\nu_i}{2}} \left\{\nu_i^{\frac{1}{2}}|\lambda_i t|\right\}, \tag{4}$$

and $K_\alpha\{z\}$ denotes the modified Bessel function of the second kind. For more details see Witkovský (2001a). Note that the characteristic function of the Student's t random variable is a real function.

The distribution function (cdf) $F_T(x)$ of T, $F_T(x) = \Pr\{T \leq x\}$, is according to the inversion formula due to Gil-Pelaez [4] given by

$$\begin{aligned} F_T(x) &= \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \Im\left(\frac{e^{-itx}\phi_T(t)}{t}\right) dt \\ &= \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{\sin(tx)\phi_T(t)}{t}\, dt, \end{aligned} \tag{5}$$

and the probability density function (pdf) $f_T(x)$ of T is given by

$$\begin{aligned} f_T(x) &= \frac{1}{\pi}\int_0^\infty \Re\left(e^{-itx}\phi_T(t)\right) dt \\ &= \frac{1}{\pi}\int_0^\infty \cos(tx)\phi_T(t)\, dt. \end{aligned} \tag{6}$$

Note that $\Pr\{T \leq x\} = \Pr\{cT \leq cx\}$ for any $c > 0$. During the calculations, the algorithm *TDIST* normalizes the linear combination T by $c = (\sum_{i=1}^{k} \lambda_i^2)^{-1/2}$ in order to stabilize the typical shape and nonzero range of the characteristic function of the random variable T. For any chosen x *TDIST* evaluates the integrals in (5) and (6) by multiple p-points Gaussian quadrature over the real interval $t \in (0, 10\pi)$. For higher values of t the normalized characteristic function of T is essentially equal to zero, see Figure 1. The whole interval is divided into m subintervals given by pre-specified limits and the integration over each subinterval is done with an p-points Gaussian quadrature which involves base points b_{ij}, and weight factors w_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, m$. So,

$$F_T(x) \approx \frac{1}{2} + \frac{1}{\pi}\sum_{j=1}^{m}\sum_{i=1}^{p} \frac{\sin(b_{ij}x)}{b_{ij}} w_{ij}\phi_T(b_{ij})$$

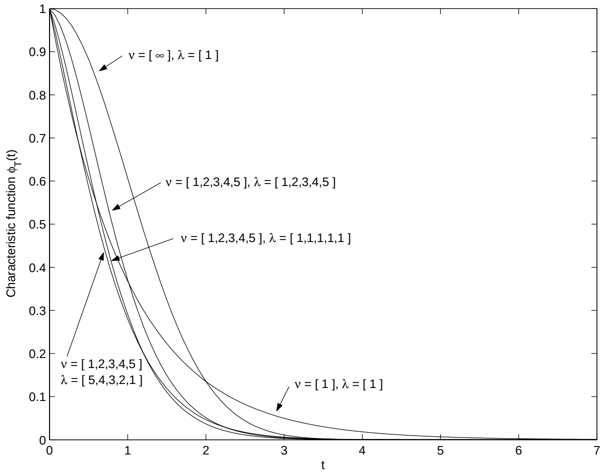


Figure 1: The characteristic function $\phi_T(t)$, over the interval $t \in (0, 7)$, of the normalized linear combination $T = c\sum_{i=1}^{k} \lambda_i t_{\nu_i}$, with $c = (\sum_{i=1}^{k} \lambda_i^2)^{-1/2}$, of independent Student's t_{ν_i} random variables for different sets of the degrees of freedom ν_i and the coefficients λ_i.

$$= \frac{1}{2} + \frac{1}{\pi}\sum_{j=1}^{m}\sum_{i=1}^{p} \frac{\sin(b_{ij}x)}{b_{ij}} W_{ij}, \tag{7}$$

$$\begin{aligned} f_T(x) &\approx \frac{1}{\pi}\sum_{j=1}^{m}\sum_{i=1}^{p} \cos(b_{ij}x) w_{ij} \phi_T(b_{ij}) \\ &= \frac{1}{\pi}\sum_{j=1}^{m}\sum_{i=1}^{p} \cos(b_{ij}x) W_{ij}, \end{aligned} \tag{8}$$

where $W_{ij} = w_{ij}\phi_T(b_{ij})$. Notice, that for evaluation of $F_T(x)$ and $f_T(x)$ in many different points the algorithm requires only one evaluation of the weights W_{ij}, for $i = 1, \ldots, p$ and $j = 1, \ldots, m$, which directly depend on the characteristic function $\phi_T(\cdot)$ and do not depend on x.

The algorithm *TDIST* calculates the α-quantile q, $\alpha \in (0, 1)$, of the distribution of T by the iterative Newton's method:

$$q^{(n+1)} = q^{(n)} - \frac{F_T(q^{(n)}) - \alpha}{f_T(q^{(n)})}, \tag{9}$$

with $q^{(0)} = 0$. It is not obvious that the Newton method will converge in all theoretical situations. By default, the algorithm allows 100 iterations before ending with a warning message. However, according to our testing results, this should never happen in typical statistical applications. Note that during the iteration process repeated evaluation of the cdf and pdf is required. However, only one evaluation of the weights W_{ij}, for $i = 1, \ldots, p$ and $j = 1, \ldots, m$, is needed.

The theoretical evaluation of the numerical error of the algorithm *TDIST* is not known yet and is subject of further research.

3 TDIST overview

The algorithm *TDIST* is distributed as simple Matlab m-file `tdist.m`. The algorithm is available on request from the author or from the Matlab home page: `http://www.mathworks.com`, (*The MathWorks, Inc.*), see the *Matlab Central*, *File Exchange*, and the *Statistics and Probability* sections.

`tdist` computes the distribution of a linear combination of independent Student's t random variables (with degrees of freedom restricted to be less than 100) and/or standard normal (Gaussian) random variables. The calling syntax inside the Matlab environment is:

```
[yfun,xfun,iserr] = tdist(funx,df,lambda,funtype,points),
```

where the input and output arguments are as follows:
Inputs:

- `funx` - the vector of function input values. If `funx=[]` then the output vector `xfun` is generated automatically, otherwise `xfun` equals to the input vector `funx`.
- `df` - the vector of degrees of freedom of the linear combination T, given by (1), of independent Student's t random variables. Use `Inf` as degrees of freedom for the standard normal (Gaussian) random variable.
- `lambda` - the vector of coefficients of the linear combination T, given by (1), of independent Student's t random variables.
- `funtype` - the type of the evaluated function. The default value is 1 (i.e., the algorithm calculates the cdf). The following are legible values of `funtype`:

  0: `tdist` calculates cdf and pdf. The output argument `yfun` consists of two columns of the same length as `funx`, i.e. `yfun=[cdf,pdf]`.
  1: `tdist` calculates the cumulative distribution function, cdf at given values `funx` and the output argument is a column of the same length as `funx`, i.e. `yfun=cdf`.
  2: `tdist` calculates the probability density function, pdf at given values `funx` and the output argument is a column of the same length as `funx`, i.e. `yfun=pdf`.
  3: `tdist` calculates the characteristic function, chf, at given values `funx` and the output argument is a column of the same length as `funx`, i.e. `yfun=chf`.

4: `tdist` calculates the quantile function qf (inverse distribution function) at given values `funx` and the output argument is a column of the same length as `funx`, i.e. `yfun=qf`.

- `points` - is the number of points of the p-points Gaussian quadrature. The default value is `points=14`. For many practical purposes, fast and reasonably precise results are for choices of `points` as small as 3.

Outputs:

- `yfun` - the column vector or two columns matrix with calculated function values. The result depends on chosen input argument `funtype`. If `funtype=0`, `yfun` has two columns, one for cdf and one for pdf.
- `xfun` - the column vector of function abscissas of evaluated function. Typically `xfun=funx`. However, if `funx=[]`, then `xfun` is generated automatically by the algorithm.
- `iserr` - The error status message. If `iserr=1`, some problem has occurred during the calculation process, see the warning messages which appeared on the screen during the run of the algorithm. If `iserr=0`, the algorithm ended correctly.

3.1 Examples

(1). Plot the pdf of the convolution of the standard normal random variable and the two independent Student's t random variables with 1 and 10 degrees of freedom. Use an integration method based on 6-points Gaussian quadrature. The Matlab command is:

```
funx=[];df=[Inf 1 10];lambda=[1 1 1];funtype=2;points=6;
[yfun,xfun,iserr] = tdist(funx,df,lambda,funtype,points);
plot(xfun,yfun);
```

(2). Plot the cdf of the linear combination of the standard normal random variable and the two independent Student's t random variables with 1,2,3 and 4 degrees of freedom. Use an integration method based on 6-points Gaussian quadrature. The Matlab command is:

```
funx=[];df=[Inf 1 2 3 4];lambda=[1 1 2 3 4];funtype=1;
cdf = tdist(funx,df,lambda,funtype);
plot(funx,cdf);
```

(3). Table the 0.9-, 0.95- and 0.99-quantile of the average of standard normal random variable and the Student's t random variable with 1 degree of freedom:

```
prob=[0.9 0.95 0.99]';
quantiles=tdist(prob,[1 Inf],[1 1]/2,3);
disp([prob quantiles]);
```

(4). Plot the characteristic function of the t variable with 1 degree of freedom:

```
[chf,t] = tdist(linspace(0,7),1,1,4);
plot(t,chf);
```

4 Numerical results

The theoretical evaluation of the numerical error of the algorithm *TDIST* is not known to the author and is subject of further research. For illustration, however, we present here some numerical results of the algorithm *TDIST* together with the known exact results derived according to Walker and Saw [9]. Their method was illustrated by the following example: Let

$$T = \frac{5}{12}t_1 + \frac{1}{3\sqrt{3}}t_3 + \frac{1}{4\sqrt{5}}t_5, \tag{10}$$

then the exact cdf of T is given by

$$\Pr(T \leq x) = \frac{60}{144}F_{t_1}(x)+\frac{42}{144}F_{t_3}(\sqrt{3}x)+\frac{27}{144}F_{t_5}(\sqrt{5}x)+\frac{15}{144}F_{t_7}(\sqrt{7}x), \tag{11}$$

where F_{t_ν} is the cdf of the Student's t distribution with ν degrees of freedom. The following sequence of Matlab commands will generate the results:

```
df=[1 3 5]; lambda=[5/12 1/(3*sqrt(3)) 1/(4*sqrt(5))];
[cdftdist,x] = tdist([0:4:40],df,lambda,1);
exact=(60/144)*tcdf(x,1)+(42/144)*tcdf(sqrt(3)*x,3)+ ...
  (27/144)*tcdf(sqrt(5)*x,5)+(15/144)*tcdf(sqrt(7)*x,7);
disp([x cdftdist exact])
```

```
     0.0   0.50000000000000   0.50000000000000
     4.0   0.96658151813966   0.96658151813966
     8.0   0.98338730341311   0.98338730341311
    12.0   0.98893740341903   0.98893740341903
    16.0   0.99170637784325   0.99170637784325
    20.0   0.99336633852020   0.99336633852021
    24.0   0.99447250947928   0.99447250947928
    28.0   0.99526244109506   0.99526244109506
    32.0   0.99585480082848   0.99585480082848
    36.0   0.99631547903494   0.99631547903494
    40.0   0.99668399586331   0.99668399586363
```

5 Conclusion

The Matlab algorithm *TDIST* is a reasonably fast and precise numerical algorithm for statistical applications which require computing the distribution and quantiles of a linear combination of independent Student's t random variables.

References

[1] Amos D.E. (1986). *A portable package for Bessel functions of complex argument and nonnegative order.* ACM Transactions on Mathematical Software **12**, 265 – 273.

[2] Davies R.B. (1980). *Algorithm AS 155: The distribution of a linear combinations of χ^2 random variables.* Applied Statistics **29**, 232 – 333.

[3] Fairweather W.R. (1972). *A method of obtaining an exact confidence interval for the common mean of several normal populations.* Applied Statistics **21**, 229 – 233.

[4] Gil-Pelaez J. (1951). *Note on the inversion theorem.* Biometrika **38**, 481 – 482.

[5] Glen A.G., Evans D.L., Leemis L.M. (2001). *APPL: A probability programing language.* The American Statistician **55**, 156 – 166.

[6] Glen A.G., Leemis L.M., Drew J.H. (2004). *Computing the distribution of the product of two continuous random variables.* Computational Statistics & Data Analysis **44**, 451 – 464.

[7] Imhof J.P. (1961). *Computing the distribution of quadratic forms in normal variables.* Biometrika **48**. 419 – 426.

[8] Rose C., Smith M.D. (2002). *Mathematical statistics with mathematica.* Springer-Verlag, New York.

[9] Walker G.A., Saw J.G. (1978). *The distribution of linear combinations of t-variables.* Journal of The American Statistical Association **73**, 876 – 878.

[10] Weerahandi S. (1995). *Exact statistical methods for data analysis.* Springer-Verlag, New York.

[11] Welch B.L. (1947). *The generalization of Student's problem when several different population variances are involved.* Biometrika **34**, 28 – 35.

[12] Witkovský V. (2001a). *On the exact computation of the density and of the quantiles of linear combinations of t and F random variables.* Journal of Statistical Planning and Inference **94**, 1 – 13.

[13] Witkovský V. (2001b). *Computing the distribution of a linear combination of inverted gamma variables.* Kybernetika **37**, 79 – 90.

[14] Witkovský V. (2002). *On the Behrens-Fisher distribution and its generalization to the pairwise comparisons.* Discussiones Mathematicae Probability and Statistics **22**, 73 – 104.

Acknowledgement: The research was supported by the grants 1/0264/03 and 2/4026/04 of the Slovak Scientific Grant Agency VEGA.

Address: V. Witkovský, Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 841 04 Bratislava, Slovak Republic

E-mail: witkovsky@savba.sk

COMPSTAT'2004 Symposium

PARALLEL COMPUTING IN A STATISTICAL SYSTEM JASP

Yoshikazu Yamamoto, Junji Nakano, Takeshi Fujiwara and Ikunori Kobayashi

Key words: Java, parallel computing, statistical system.

COMPSTAT 2004 section: Statistical software.

Abstract: Modern statistical computation often requires powerful computers. Parallel computing is a technique to realize them. Existing software technologies for parallel computing, however, are not easy for statisticians to use, as they are accustomed to ”high-level” statistical languages. In this paper, we describe parallel computing functions in a statistical analysis system Jasp, which are implemented for ease of use.

1 Introduction

Recently, the amounts of data that statistics calculations must handle have become huge, partly because automatic and continuous data acquisition systems have become popular. As data generating mechanisms are sometimes very complicated, simple classical statistical models are not sufficient to describe them. Thus, several computer intensive statistical techniques have been developed, including simulation and resampling techniques and data mining techniques. These methods depend essentially on the huge calculation abilities of modern computers. To execute these statistical methods, we require much faster computers than those we have now, although they are almost as powerful as the “super computers” of several years ago.

One of the promising techniques for realizing powerful computers is parallel computing. This technology uses more than one CPU simultaneously to execute computation. Now, cheap “computer cluster” systems, which are groups of independent computers connected by networks, are available for parallel computing. To support this hardware, we have parallel computing software such as Parallel Virtual Machine (PVM) [6] and Message Passing Interface (MPI) [4]. These libraries were designed and developed mainly for Fortran and C, which have been used for scientific numerical calculations for a long time.

Recently, many statisticians have become accustomed to using “high-level” statistical languages, which have been developed to express complicated statistical ideas as easily as possible by a few functions or commands. As many of them are reluctant to use Fortran or C, it is difficult for them to write programs for PVM or MPI.

Therefore, a parallel computing statistical system that is easy to use and learn is required. In this paper, we describe the parallel computation abilities of an experimental general purpose statistical system called Jasp [5].

Jasp was originally designed for use in network environments and adopted a client/server structure, which is a simple distributed computing technology, where a client program and a server program may be placed on different computers and collaborate to perform a single job. We have added new functions for realizing parallel computing abilities.

In the next section, we consider several requirements for parallel computing abilities in statistical calculations. Section 3 describes design principles and the implementation for parallel computing by using Jini and JavaSpaces technologies. We explain the functions for parallel computing in Jasp and show examples in Section 4. In the last section, we offer some concluding remarks.

2 Parallel computing and statistical computing

Historically, parallel computing has been mainly developed for executing scientific calculations in physics or earth sciences, where huge calculations are required. The target users are professional scientists who are willing to write complicated programs in compiler languages such as Fortran and C for their research purposes. They are not reluctant to write detailed programs to divide matrix calculations into small pieces to execute them simultaneously, if the calculation speed is improved by such efforts. PVM and MPI were developed for them, as libraries of Fortran and C routines. They are designed to support "low-level" programming such as matrix calculations, and are not easy to use.

At present, MPI is the most widely used parallel computing library, partly because many hardware makers provide their own MPI libraries to utilize their hardware most efficiently. To use MPI effectively, users must consider every detail of a calculation while writing a program. These programs require precise knowledge of their algorithms and the process of parallel computing. In general, statisticians do not wish to write such programs. Many statistical languages have been developed to conceal details of algorithms from users, and to let them perform complex calculations by using simple functions or commands.

In parallel computing, if a job can be divided into small tasks that can be calculated almost independently without much communication among them, it is most efficiently executed. Such problems are sometimes referred to as "embarrassingly parallel" computations. Fortunately, many statistical jobs have embarrassingly parallel natures. For example, in data analysis, we often wish to perform the same statistical calculations on many data sets. Each calculation for a data set is performed completely independently from other data sets. Another example is a simulation or a resampling calculation, in which new data sets are generated by a random number mechanism based on a given data set or parameters. We calculate some statistics for these data sets, repeat the calculations many times and summarize the results to show the characteristics of the statistics from the empirical distribution. In this example, all calculations can be performed simultaneously except the last part. The last example is a grid search procedure for estimating parameter values by the maximum likelihood method. We calculate values of a likelihood function by changing parameter values gradually, and choose the parameter value that maximizes the likelihood function. Maximum likelihood calculations for each parameter value can be performed simultaneously and independently. In these statistical problems, calculations are naturally embarrassingly parallel. As this is a favorable characteristic for parallel computing, parallel computing environments can be useful for statisticians if they are easy to use.

Furthermore, many statisticians can access several computers that run various operating systems and are connected by networks. It often happens that they wish to use them simultaneously for a single job to reduce computing time.

In this situation, some statistical analysis systems have parallel computing abilities. RPVM [3], Rmpi [11] and the Simple Network of Workstations (SNOW) [7] of the R [9] system are examples. RPVM is a wrapper of the PVM functions on R. Rmpi is an R interface to the MPI. With these packages, R users can invoke other executable programs written in native languages such as C, C++ or Fortran as child tasks or spawn separate R processes. SNOW provides a sophisticated and unified interface to Rmpi, RPVM and socket functions. Users freely choose one of them to implement parallel computing, and can use their abilities by simple functions of R without knowing details of them. Another example is a system TISAS [10], which is able to perform classical linear time series analysis on several remote computers through TkPVM, an implementation of PVM in the Tcl/Tk language.

3 Design and implementation of parallel computing in Jasp

We designed Jasp's parallel computing abilities mainly for a "computer cluster", which is a group of independent computers connected by a network, and is one of the major architectures for parallel computing called "distributed memory" systems. They are easily accessible for many statisticians, because they are cheaper than "shared memory" parallel computing systems, in which several CPUs share a single memory address space. Another design goal of our system is ease of use. For this reason, we decided to add just a few functions for realizing parallel computing.

In parallel computing, we must be careful to distribute tasks to remote computers appropriately, because the job will be finished when all the remote computers finish their tasks. It is important to deliver small amounts of calculations to slow computers and large amount to fast computers so as not to have idle computers. This assignment problem is called "load balancing" and is crucial for performing efficient parallel computing.

The basic components of Jasp are a client and a server. The client realizes the user interface and the server calculates statistics. We extend the model in such a way that one server can communicate with other servers for parallel computing. We call the server to which the client is connected the main server. The main server has all the data sets for the analysis at the beginning, and sends functions and data for a calculation to remote servers, then gathers their results from remote servers.

Jasp is written in the Java language and is easily extended by using Java libraries [2]. We used the Jini and JavaSpaces libraries to implement features for parallel computing. Jini was developed for connecting all computer resources including personal computers, workstations, and even computerized electric devices such as TVs and refrigerators, in the Java environment [8]. JavaSpaces is a Java class library to support distributing tasks on the Java network environment, and is implemented using Jini technology. JavaSpaces provides is an object like a whiteboard, to which clients write a Java object to be executed, and from which servers take one of those Java objects and run it. The results of calculations by a Java server are written to JavaSpaces and are received by the client that writes the Java object. It is not necessary to specify the name of a Java server for executing a task, because an available server automatically receives a task from JavaSpaces.

The use of JavaSpaces can realize a simple load balancing mechanism. The main server writes many small tasks to JavaSpaces objects and each available remote server takes one of them. In this way, a fast server performs many tasks and a slow server performs few tasks. This mechanism works well on heterogeneous computer cluster systems.

4 Jasp functions for parallel computing

To support parallel statistical computing, we must distribute programs and data to remote computers, execute them on the remote computers and gather results from them to the local computer, then summarize the final result by incorporating the remotely calculated results. In Jasp, we introduce four functions for executing these parallel computing tasks. Table 1 shows them. The argument `function` indicates a function that is executed on a remote server computer specified by the argument `server`. The argument `variable` denotes a variable name on the local computer to which the result of the function executed on a remote server is assigned.

The function `remoteFunctions` declares the names of functions that are executed on remote servers by using a list.

The function `setRemoteAssignment` sets a function that is executed on a remote server and specifies a local variable to which the result is assigned. This function can have four arguments, of which the last two are optional. The first argument specifies a local variable to which the result of the function indicated by the second argument is assigned. The third argument shows the name of a remote computer on which the function specified by the second argument is to be executed. If this argument is omitted, Jasp chooses an

```
remoteFunctions
      remoteFunctions([function1, function2, ...])
setRemoteAssignment
      setRemoteAssignment(variable, function)
      setRemoteAssignment(variable, function, server)
      setRemoteAssignment(variable, function, server, index)
      setRemoteAssignment(variable, function, "", index)
executeRemote
      executeRemote()
      executeRemote(index)
executeRemoteFunction
      executeRemoteFunction(variable, function, server)
      executeRemoteFunction(variable, function)
```

Table 1: Parallel computing functions in Jasp.

appropriate remote server. The last argument specifies an index that is used for grouping remote function executions. If it is omitted, the function execution does not belong to any group. Note that this function does not start the remote execution of the function.

Remote calculations are started by the function `executeRemote`. The first argument of this function specifies the group named in the `setRemoteAssignment`. If the argument is not specified, it executes remote functions that do not have any index. If the argument is "`all`", all remote functions are executed. The function `executeRemote` starts sending each function specified by `setRemoteAssignment` to remote servers and executing them. When all the results are returned from remote servers, this function completes the execution.

The function `executeRemoteFunction` executes the specified function on a remote server, and is designed mainly for distributed computing. The calculation result on the remote server is returned to a local variable. This function use similar arguments to `setRemoteAssignment`.

We now show several examples and experimental results.

In the first example, we generate many random numbers and calculate their mean. We divide the job into 10 sub-tasks in which smaller numbers of random numbers are generated. All the results of sub-tasks are returned to the local computer and summarized to calculate the total mean of the all generated random numbers. This work can be performed by the program shown in Listing 1. The Jasp function `randomMean(length)` generates `length` uniform random numbers between `0` and `100`, and calculates their mean (the program is omitted). As we wish to execute `randomMean` on remote servers, it is specified by `remoteFunctions`. The first `for` loop is used to set functions for remote executions `R` times by using a function `setRemoteAssignment`. The function `setRemoteAssignment` indicates that `randomMean(N[i])` is executed on a remote computer, and the result is assigned to the local variable `tm[i]`. We note that variables in arguments of the remote function indicated in `setRemoteAssignment` are evaluated on the local computer before they are sent to remote servers. After we set all the parallel computing functions, a function `executeRemote` starts the functions executing simultaneously. Final results are summarized by the second `for` loop on the local computer.

```
R = 10
N0 = 50000
tm = Matrix(R)
N = Matrix(R)
remoteFunctions([randomMean])
for (i = 1; i <= R; i++) {
  N[i] = N0 + i*10000
  setRemoteAssignment(tm[i], randomMean(N[i]))
}
executeRemote()
sum = 0.0
Nsum = 0
for(i = 1; i <= R; i++) {
    sum = sum + N[i] * tm[i]
    Nsum = Nsum + N[i]
}
mean = sum / Nsum
mean
```

Listing 1: An example of parallel computing in Jasp.

The next example, Listing 2, is a bootstrap example [1]. The variable `R` is the number of resamplings, `nt` is the number of remote servers and `ratio` is used to store the results of bootstrap calculation results. As functions `nonparamBootstrap` and `meanRatio` are executed on remote servers, we indicate them by the function `remoteFunctions`. We set remote executions by the function `setRemoteAssignment` in the `for` loop. The function `nonparamBootstrap` executes the function specified by the second argument repeatedly for the number of times in the third argument. The first argument of the function identifies a target data set. In this example, we execute the function `meanRatio` on the data `bigcity` for `R` times.

Table 2 shows the calculation times of this example for some situations. We can use 70 remote servers from a client computer. Each server computer has a Pentium 4 1.6 GHz CPU and 512 MB of memory, and the client computer has a Pentium III 800 MHz CPU and 384 MB of memory. They are connected by a 100-Mbps Ethernet. We change the number of servers and always keep the total number of bootstrap resamplings as 350,000.

The results show that parallel computing is useful to reduce total execution times. In this example, however, the effects of using many servers are reduced for the 35- and 70-server cases. This happens because many remote

```
bigcity = read("bigcity.dat")
ratioOriginal = meanRatio( bigcity )
R = 70000
nt = 5
ratio = Matrix(R * nt)
remoteFunctions([nonparamBootstrap, meanRatio])
for (i = 1; i <= nt; i++) {
  setRemoteAssignment(ratio[ R * (i - 1) + 1 ],
                      nonparamBootstrap(bigcity, meanRatio, R))
}
executeRemote()
xBS = ratio - ratioOriginal
plot = plotMeanRatio( xBS )
plot
```

Listing 2: A bootstrap example using Jasp parallel computing.

Number of computers	The number of resamplings for one computer	Time (seconds)
1	350,000	845
5	70,000	219
7	50,000	173
35	10,000	129
70	5,000	199

Table 2: Caltulation time of a bootstrap example.

servers finish their calculation and return their results to the main server almost at the same time and the load of the main server temporarily becomes very heavy.

We note that transfer of information on the network is slower than calculations by CPUs and it is recommended that we keep data transfer among computers as small as possible. In our implementation, all data sets are sent as arguments of the `setRemoteAssignment` function implicitly, and we have no explicit data transfer function.

5 Concluding remarks

We designed the Jasp parallel computing abilities to utilize several heterogeneous or homogeneous computers connected by networks. Jasp is suitable for implementing these features because it is written in the Java language. We note that the Jasp system and parallel computing programs written in the Jasp language can be executed on any computer that supports the Java virtual machine.

We designed the parallel computing abilities mainly considering ease of use. This may damage the efficiency and the resulting system may lack some functions usually supplied in parallel computing environments. Our experiment shows that if the number of servers is too high, the calculation time becomes even longer. We decided not to provide “low level” functions, for example, a separate function for sending data to remote servers. Many statistical calculations are divided rather simply and naturally into independent sub-tasks, and it is sufficient to send data implicitly as arguments of the function executed on a remote server.

We are still in the experimental stage of designing the parallel computing ability of Jasp. We plan to implement several statistical procedures using our parallel computing functions, check their performance and improve them.

References

[1] Davison A., Hinkley, D. (1997). *Bootstrap methods and their application.* Cambridge University Press.

[2] Fujiwara T., Nakano J., Yamamoto Y., Kobayashi I. (2001). *An implementation of a statistical language based on java.* Journal of the Japanese Society of Computational Statistics **14** (1), 59 – 67.

[3] Li N., Rossini A. (2001). *RPVM: Cluster statistical computing in R.* R News **1** (3), 4 – 7.

[4] Message Passing Interface (MPI) Forum. Message passing interface (MPI) forum home page. `http://www.mpi-forum.org/`.

[5] Nakano J., Fujiwara T., Yamamoto Y., Kobayashi I. (2000). *A statistical package based on pnuts.* In J. G. Bethlehem and P. G. M. van der Heijden, (eds.), Proceedings in Computational Statistics, Heidelberg, Physica-Verlag, 361 – 366.

[6] PVM Project Members. PVM: Parallel virtual machine. `http://www.csm.ornl.gov/pvm/pvm_home.html`.

[7] Rossini A., Tierney L., Li N. (2003). *Simple parallel statistical computing in R.* UW Biostatistics working paper series, Paper 193, University of Washington. (`http://www.bepress.com/uwbiostat/paper193`).

[8] Sun Microsystems Inc. Jini^{tm} Network Technology. `http://wwws.sun.com/software/jini/`.

[9] The R Development Core Team. The R project for statistical computing. `http://www.r-project.org/`.

[10] Yamamoto Y., Nakano J. (2002). *Distributed processing functions of a time series analysis system.* Journal of the Japanese Society of Computational Statistics **15** (1), 65 – 77.

[11] Yu H. (2002). *Rmpi: Parallel statistical computing in R.* R News **2** (2), 10 – 14.

Address: Y. Yamamoto, J. Nakano, T. Fujiwara, I. Kobayashi, Tokushima Bunri University, 1314-1, Shido, Sanuki, Kagawa, Japan

E-mail: `yamamoto@is.bunri-u.ac.jp`

DandD CLIENT SERVER SYSTEM

Daisuke Yokouchi and Ritei Shibata

Key words: XML, Java, DandD, DandDServer.

COMPSTAT 2004 section: E-statistics.

Abstract: DandD Client Server System is a basic support system of softwares providing a comfortable environment for data manipulation, data understanding and data modelling, based on a given DandD instance. DandD (Data and Description) instance is an XML document written in accordance with DandD rule, in which data and its enough descriptions are stored to the extent of semantics. DandDServer is a key software in DandD Client Server System. Following to the request of client program, DandDServer returns necessary information to the client in an appropriate format by interpreting the underlying DandD instance. Various client programs have been also developed, for example, DandDBrowser, DandDR, and DandDGenerator. Such softwares work independently but work together to provide a comfortable unified environment to the user in every stage of data collection, data cleaning, data organising, data analysis, model building and its utilisation.

1 DandD

Environment of statistics has been drastically changed due to highly developed computers and Internet. It becomes much easier to collect enormous amount of data at once and store it as files or as a database, and some of them can be freely browsed through the Internet. However, it is rather rare that such data is accompanied with enough description of data to understand the meaning and the background. Even personally, it may take a long time to remember what the data meant and what was the background, for example, after a year.

One of solutions to such a problem is to establish a rule to describe data. There would be various ways of defining such a rule and some of works already have been done. DandD rule is one of such rules but different from others in various ways. We don't explain it into detail to avoid duplication to another invited talk by Shibata [3]. It would be enough to say that DandD rule is a general machine readable rule which is intended to cover any field of science and business. The aim of this paper is to show an outline of the rule, called DandD rule, and the implementation of its support softwares in DandD Client Server System.

1.1 DandD rule

DandD rule is a rule to describe data together with its background, implemented by eXtensible Markup Language (XML). Under the DandD rule, any data are decomposed into a set of vectors which are merely sequences of numbers, called **DataVector**. Each DataVector has several attributes which are specific to the sequence, for example, LongName which makes the DataVector self explanatory, Code which gives a coding for categorical data or Length which gives the number of elements. Given data are organised into several structures, by combining such DataVectors. To explain the rule a little bit more deeply, let us give a simple example of DandD instance in the following.

Example1

```
<Title Title="etitle gtitle" Language="English German"/>

...

<Data>
 <Relational Id="r1">
   <Value RefId="height"/>
   <Value RefId="weight"/>
   <Value RefId="sex"/>
 </Relational>
</Data>
<DataBody>
  <DataVector Id="height" LongName="elh">
  187 168 177 167
  </DataVector>
  <DataVector Id="weight"  LongName="elw">
  84 58 78 65
  </DataVector>
  <DataVector Id="sex" Code="ec1 gc1" LongName="els">
  1 2 2 1
  </DataVector>

...

</DataBody>
<Appendix>
  <Text Id="etitle">An Example of DandD Instance</Text>
  <Text Id="gtitle">Beispiel von DandD instance</Text>
  <Text Id="elh">Height of Each Person</Text>
  <Text Id="elw">Weight of Each Person</Text>
  <Text Id="els">Gender of Each Person</Text>
  <Code Id="ec1">"male" "female"</Code>
  <Code Id="gc1">"Mann" "Frau"</Code>
</Appendix>
```

As you can easily understand from this example, the tag `<Title ··· />` gives us the title of the DandD instance. However, several languages can be used for writing the title. In fact, two languages are used in this example. The attribute `Title="etitle gtitle"` indicates that the title is given in two languages. The `etitle` and `gtitle` are I.D.s of `<Text>` tags placed in the `<Appendix>`. The `<Text>` tag with Id `etitle` gives us an English title and that with Id `jtitle` gives us a German title in this example. Another attribute `Language` is explicitly giving the names of languages used in order. The order of languages given in this attribute is also applied to any attributes in a DandD instance, as far as internationalisation is necessary. For example, the Code attribute of DataVector `sex` is ”ec1 gc1” which means that the code `ec1` should be used for English, and the code `gc1` is used otherwise. Therefore, if English is chosen as a language, the number 1 and 2 should be coded to ”male” and ”female”, respectively, but otherwise coded differently. On the other hand, the LongName attribute of each DataVector consist of a single Id, so that the corresponding text is used in common for different languages. The same principle is applied to Introduction or Reference in BackGround which are omitted in this example. We learned the need of such mechanism in our execution of DandD project, since most of data can be exchanged internationally but it would be natural to describe the attributes or the background in each language. Then, the user can choose own language to easily understand the meaning of the data.

The role of the tag `<Data>` is to define several structures by referring DataVectors in `<DataBody>`. In this example, a structure is defined by the tag `<Relational>`, where three DataVectors specified by the tag `<Value>` are organised as a relational scheme, in other words, a relational table is defined by this tag with such DataVectors as its columns. Any number of relations or arrays can be defined in `<Data>`, so that any complicated relations among relations and arrays can be constructed.

All DataVectors are placed in `<DataBody>`. The body of DataVector can be empty if the sequence of numbers can be obtained from outside. Then, various attributes of DataVector tell where and how the sequence can be obtained. See Shibata [3] for details. For any other details of the DandD rule, see DandD.dtd [1].

1.2 DandD instance

DandD instance is an XML document, whose body is written in accordance with the DandD rule as has been described. Advantage of XML document is that it is easy to exchange such documents over the Internet, since it takes a form of text file. However, a caution is needed for the character code. In view of internationalisation, the use of Unicode [6] is indispensable. We have chosen UTF-16 as a standard character code of DandD instance because of its generality. This seems a minor point, but it is in fact an important point to make smooth exchange of DandD instance over the Internet. Another point is the suffix to the name of DandD instance as a file. Default choice is xml since DandD instance is an XML instance, but we have chosen a suffix `.dad` to DandD instance file, to be able to invoke automatically DandD Client programs from the DandD instance file at least on Windows.

2 DandDServer

2.1 Why client-server system?

Various support softwares should have been developed to provide a comfortable environment for different needs of user with data. However, it would be laborious to develop such softwares one by one. Rather, it was better to develop a server which works for other softwares to support common procedures, for example, loading the instance, finding an object, extracting its attributes or the body. Other advantages of the client-server system are:

Easiness of Programming Support Softwares Any programming language can be used to develop a client software, as far as the language supports socket handling. Also, the programming becomes simpler because it is enough to invoke a request and wait for the response from server.

Flexibility Even when the DandD rule is modified, the client program works as same as before in most cases, since the server can absorb such modification. Any server can be used as far as it is connected through network or the Internet.

Mobility The size of client program can be reduced. Therefore it is installable on low ability machine like cellular phone or PDA.

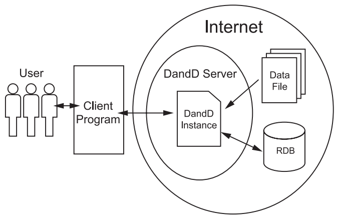


Figure 1: DandD Client Server System

2.2 DOM

We adopted Document Object Model (DOM) Level1 API [7] as a basic language of communication between the DandDServer and its client. DOM is a general definition of Markup Language including HTML or XML, together with all necessary method definition to manipulate the instance. Although DOM covers most of all operations on an XML document, loading and flushing of the document is not defined because it is assumed in DOM that the document has been already loaded on memory. What we needed to introduce is the two extra methods, **loadDocumet** and **flushDocumet**.

DOM is updated time to time. In fact, W3C is extending DOM up to Level3 and drafting Load and Save Specification. However it would not cause any serious problem. It is enough to modify DandDServer to catch up such updates, because of our organisation of support softwares, client server system.

2.3 Communication with client

To begin communication with DandDServer, it is necessary for the client to open a socket with its IP address together with the port number 11009. Then the server waits for the request from the client. The client can send any DOM sentence to the server as a text. The line feed symbol signals the end of the sentence. The server will return the result of the execution of the DOM sentence. If the execution were successful, the header to the result is just 1 when the result is empty, otherwise 2 and the number of characters of the result to be sent. If it were unsuccessful, just 0 will be sent to the client. Any communication between the server and the client is in the form of text of UTF-16, so that it is free from differences of machine architectures.

2.4 Java

The main reason why we adopted Java [4] as a language for implementing DandDServer is simple. Available DOM APIs were limited to the implementations on Java at the time of the start of DandDserver development. The currently used API is Xerces for Java2 [5].

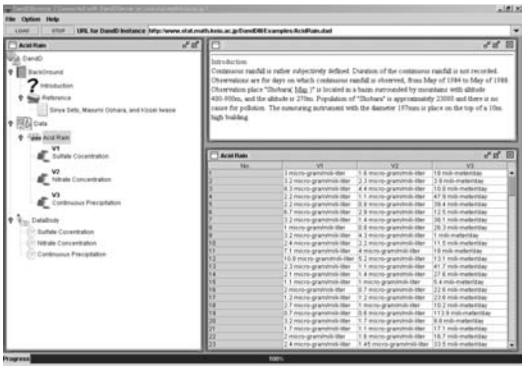
As has been explained, the body of DataVector is not necessarily included in DandD instance, so that DandDServer should have a functionality of getting the body by accessing other files or databases through network in most cases. We could make use of `java.i.o` and `JDBC`, which are both implemented on Java for such needs. This functionality is implemented as a method **getDataVector** in DandDServer. Simply by invoking this method, the body of various kinds of External DataVector can be obtained automatically reflecting the attributes of the DataVector.

DandDServer can run on any computer as far as Java Virtual Machine (JVM) has been installed. This would be an advantage of using Java as a language of implementation of DandDServer. DandDServer can be downloaded with JVM from DandD Project Home Page freely [1], together with other client softwares.

3 DandD client softwares

3.1 DandDBrowser

DandDBrowser is a program to help user to understand quickly what kind of data are described in the underlying DandD instance. Currently it is implemented by Java. DandDBrowser can access each element of the DandD instance by traversing the DOM tree expanded on DandDServer with DOM methods. Figure.2 shows a display of DandDBrowser. The graphical interface is nothing special. The tree structure of DandD instance is shown in the left internal frame. By a double-left-clicking of a node, the node is further expanded unless it is a terminal node, otherwise the body will be displayed on a right panel. To see the attributes of each node, it is enough to do right-clicking, then a popup-menu appears to select what you want to see.



Introduction

Continuous rainfall is rather subjectively defined. Duration of the continuous rainfall is not recorded. Observations are for days on which continuous rainfall is observed, from May of 1984 to May of 1988. Observation place "Shobara(Map)" is located in a basin surrounded by mountains with altitude 400-900m, and the altitude is 270m. Population of "Shobara" is approximately 23000 and there is no cause for pollution. The measuring instrument with the diameter 197mm is place on the top of a 10m high building

No	V1	V2	V3
1	1 micro-gram/mili-liter	1.6 micro-gram/mili-liter	18 mili-meter/day
2	3.2 micro-gram/mili-liter	2.3 micro-gram/mili-liter	3.9 mili-meter/day
3	6.3 micro-gram/mili-liter	4.4 micro-gram/mili-liter	10.8 mili-meter/day
4	2.2 micro-gram/mili-liter	1.1 micro-gram/mili-liter	47.9 mili-meter/day
5	2.2 micro-gram/mili-liter	0.8 micro-gram/mili-liter	39.4 mili-meter/day
6	6.7 micro-gram/mili-liter	2.9 micro-gram/mili-liter	12.5 mili-meter/day
7	3.2 micro-gram/mili-liter	1.4 micro-gram/mili-liter	38.1 mili-meter/day
8	1 micro-gram/mili-liter	0.8 micro-gram/mili-liter	26.3 mili-meter/day
9	3.2 micro-gram/mili-liter	4.3 micro-gram/mili-liter	1 mili-meter/day
10	2.4 micro-gram/mili-liter	2.2 micro-gram/mili-liter	11.5 mili-meter/day
11	7.1 micro-gram/mili-liter	4 micro-gram/mili-liter	19 mili-meter/day
12	10.8 micro-gram/mili-liter	5.2 micro-gram/mili-liter	13.1 mili-meter/day
13	2.3 micro-gram/mili-liter	1.1 micro-gram/mili-liter	41.7 mili-meter/day
14	2.1 micro-gram/mili-liter	1.4 micro-gram/mili-liter	27.6 mili-meter/day
15	1.1 micro-gram/mili-liter	1 micro-gram/mili-liter	5.4 mili-meter/day
16	2 micro-gram/mili-liter	0.7 micro-gram/mili-liter	22.6 mili-meter/day
17	1.2 micro-gram/mili-liter	1.2 micro-gram/mili-liter	23.6 mili-meter/day
18	2.7 micro-gram/mili-liter	1 micro-gram/mili-liter	10.2 mili-meter/day
19	0.7 micro-gram/mili-liter	0.8 micro-gram/mili-liter	113.9 mili-meter/day
20	3.2 micro-gram/mili-liter	1.7 micro-gram/mili-liter	9.8 mili-meter/day
21	1.7 micro-gram/mili-liter	1.1 micro-gram/mili-liter	17.1 mili-meter/day
22	2 micro-gram/mili-liter	1.9 micro-gram/mili-liter	16.7 mili-meter/day
23	2.4 micro-gram/mili-liter	1.45 micro-gram/mili-liter	23.5 mili-meter/day

Figure 2: DandDBrowser

To support the use of different languages, DandDBrowser asks the user to choose one of possible languages at the time of loading a new DandD instance. As an option, the user can set a limit of the number of elements of DataVector displayed, since it is unrealistic to display enormous amount, for example, more than 100,000, of data as a whole.

If a DandD instance were on the local machine, DandDServer has no way to access to the instance by security reasons. In this case, the user can select "send DandD instance" from "File" menu to specify the DandD instance sent to the DandDServer. The same "File" menu provides another item "save DandD instance" to save the current DandD instance on the local machine.

3.2 DandDR

DandDBrowser is just for browsing data and its description. Our principle of the design of support softwares is to avoid duplication of the functions of each softwares. Users can invoke simultaneously several support softwares and get around them according to their need. It is not expensive to invoke simultaneously several support softwares, because all DandD support softwares have been developed in the style of client-server system. In a primary use of support softwares, besides the DandDBrowser, it would be natural to have a good software to support Data Analysis, which communicates with DandDServer. To make it possible, it is enough to provide an interface to each existing softwares. DandDR is an example of such interfaces, an interface to R [2].

Import The function DandD on R invokes the interface DandDR which imports the DandD instance onto R as a list. The function then put together DataVectors organised in the `<Data>` of the DandD instance and attaches it as an element of the search list. The value of each DataVector is not evaluated yet in this stage. All DataVectors are stored in a form of delayed evaluation object, so that the evaluation occurs only when the value is really needed. This is quite useful for saving time of loading large DataVectors. The function DandDR returns the imported list, so that any description written in the DandD instance can be immediately extracted from the saved value of this function. Several functions for extracting necessary information from the saved object have been developed.

Encoding DandDR converts character code from UTF-16 to an appropriate code which the current R can understand.

Although DandDR is an example of interface to R, it can be used for interfaces to other softwares by a small modification, since it is a simple interface written by C and encoding of character code is implemented by `iconv` package in GNU library. We are developing further supporting softwares on R, including an automatic visualisation of the data described in DandD instance. All such interfaces and R functions are open to public on the DandD Home Page [1].

3.3 DandDGenerator

It is possible to make or modify a DandD instance by hand, since it is not only machine readable but human readable. However, it would be tedious and tend to produce an erroneous instance. This is the need of third type of support softwares in DandD Client Softwares. DandDGenerator is to help user to create a DandD instance from scratch. The raw data can be given to DandDGenerator beforehand, or comes from instrument time to time. In any case, user can choose the closest template of DandD instance to the data from a library and modify it. DandDGenerator helps user to choose one of them, reflecting the given data or the environment of the data collection. The basic mechanism of this software is the same as that of DandDBrowser or of DandDR. The DandD instance is generated on DandDServer through communication from DandDGenerator. After the completion, the DandDServer returns the whole instance to the DandDGenerator. In this frame work, there is little difference between creation of a new DandD instance and modification of an existing DandD instance, so that DandDGenerator can be used also for update or modification of DandD instance. This software is under development at the time of writing this paper, but soon available for public.

4 Practical applications

DandD Server Client system not only provides those who are working with data a comfortable environment, but also leads them to thinking of the meaning of data more seriously. This always results in constructing a powerful model in various practical applications. It is particularly important for the case when any statistician is involved in the project as a consultant. The consulting process becomes more smooth if the data is described well in the form of a DandD instance. We have already had such a collaboration with physician or with athlete. One is with the Institute of Rheumatism at Tokyo Woman Medical University to organise data of questionnaire to patients, another is with Japan Swimming Foundation to improve their swimming records. Such a successful collaboration based on DandD Server Client System shows a real need of the description together with the well developed support softwares.

However, the field of the application of DandD is not limited to such academic activities. It makes possible a smooth communication among companies in a less expensive way, because no large unified database nor specific system is necessary. Any DandD server on the Internet can be used freely and it is enough for the client to have a small socket interface. The return is not only a better understanding of the data but also an exhaustive use of the data.

References

[1] DandD Project. (2004). *DandD Home Page.* http://www.stat.math.keio.ac.jp/DandD/.

[2] Ihaka R. and Gentleman R.(1996). *R:A language for data analysis and graphics.* Journal of Computational and Graphical Statistics, **5**(3), 299–314.

[3] Shibata R. (2004). *InterDatabase and DandD.* Compstat 2004, Physica-Verlag, Prague.

[4] Sun Micro Systems. (2004). *Java Technology.* http://java.sun.com/.

[5] THE APACHE XML PROJECT. (2003). *xml.apache.org.* http://xml.apache.org/.

[6] The Unicode Consortium. (2004). *Unicode Home Page.* http://www. unicode.org/ .

[7] W3C. (2003). *Document Object Model.* http://www.w3.org/DOM/.

Acknowledgement: This work was partially supported by the 21st Century COE Program; Integrative Mathematical Sciences: Progress in Mathematics Motivated by Social and Natural Sciences.

Address: D. Yokouchi, R. Shibata, Department of Mathematics, Keio University, 3-14-1 Hiyoshi Koho- ku-ku Yokohama 223-8522 Japan

E-mail: `yokouchi@stat.math.keio.ac.jp`

COMPSTAT'2004 Symposium © Physica-Verlag/Springer 2004

ON UNBIASEDNESS OF SOME EBLU PREDICTOR

Tomasz Zadlo

Key words: Model approach in survey sampling, general linear model, general mixed linear model, BLU and EBLU predictors.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: We consider two BLU predictors - the BLU predictor proposed by Henderson [3] (following Rao [6]) and the BLU predictor proposed by Royall [7]. We show that the BLU predictor proposed by Royall [7] is a generalisation of the BLU predictor proposed by Henderson [3]. Formulae of both BLU predictors includes unknown elements of the variance-covariance matrix of random variables. If the elements in formula of BLU predictor proposed by Henderson [3] are replaced by some type of estimators, we will obtain the two-stage predictor called EBLU predictor which is model-unbiased [4]. In the paper we prove model-unbiasedness of the EBLU predictor which is based on formula of BLU predictor proposed by Royall [7]. The proof may be treated as a generalisation of the results received by Kackar and Harville [4].

1 Superpopulation models

The finite population consists of N units, each of which has a value of a target variable y associated with it. The population vector of y's is $\mathbf{y} = [y_1, y_2, \cdots, y_N]^T$ and is treated as the realization of a random vector $\mathbf{Y} = [Y_1, Y_2, \cdots, Y_N]^T$. From the population of N units, a sample s of n units is selected, and the y values of the sample units are observed. For any sample s we can reorder the population vector y so that the first n elements are those in the sample: $\mathbf{y} = [\mathbf{y}_s^T, \mathbf{y}_r^T]^T$ where $\mathbf{y}_s$ is the n-vector of observed values and $\mathbf{y}_r$ the N_r-vector of unobserved values where $N_r = N - n$. Hence, vector $\mathbf{Y}$ we can reorder as follows: $\mathbf{Y} = [\mathbf{Y}_s^T, \mathbf{Y}_r^T]^T$. We study the general linear model:

$$\begin{cases} E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \\ D^2(\mathbf{Y}) = \mathbf{V} \end{cases} \tag{1}$$

where $\mathbf{X}$ is an $N \times p$ matrix of values of p auxiliary variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and $\mathbf{V}$ is positive definite variance-covariance matrix depending on some variance parameters $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_q]^T$. We assume that all values of auxiliary variables are known for each unit in the population. If the population elements are rearranged so that the first n elements of $\mathbf{Y}$ are those in the sample, and the first n rows of $\mathbf{X}$ are for units in the sample, then $\mathbf{X}$ and $\mathbf{V}$ can be expressed as $\mathbf{X} = \begin{bmatrix} \mathbf{X}_r \\ \mathbf{X}_s \end{bmatrix}$, $\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$ where $\mathbf{X}_s$

is $n \times p$, $\mathbf{X}_r$ is $N_r \times p$, $\mathbf{V}_{ss}$ is $n \times n$, $\mathbf{V}_{rr}$ is $N_r \times N_r$, $\mathbf{V}_{sr}$ is $n \times N_r$ and $\mathbf{V}_{rs} = \mathbf{V}_{sr}^T$. We assume that $\mathbf{V}_{ss}$ is positive definite.

The general mixed linear model with the following assumption is a special case of the general linear model with the assumption (1):

$$\begin{cases} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \\ E(\mathbf{v}) = E(\mathbf{e}) = \mathbf{0} \\ D^2\begin{bmatrix} \mathbf{v} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \end{cases} \tag{2}$$

where $\mathbf{Z}$ is known $N \times h$ matrix, and random vectors $\mathbf{v}$ and $\mathbf{e}$ are $h \times 1$ and $N \times 1$ respectively. If the population elements are rearranged so that the first n elements of $\mathbf{Y}$ are those in the sample, and the first n rows of $\mathbf{Z}$ are for units in the sample, then $\mathbf{e}$, $\mathbf{Z}$ and $\mathbf{R}$ can be expressed as $\mathbf{e} = \begin{bmatrix} \mathbf{e}_s \\ \mathbf{e}_r \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_s \\ \mathbf{Z}_r \end{bmatrix}$, $\mathbf{R} = \begin{bmatrix} \mathbf{R}_{ss} & \mathbf{R}_{sr} \\ \mathbf{R}_{rs} & \mathbf{R}_{rr} \end{bmatrix}$ where $\mathbf{e}_s$ is $n \times 1$, $\mathbf{e}_r$ is $N_r \times 1$, $\mathbf{Z}_s$ is $n \times h$, $\mathbf{Z}_r$ is $N_r \times h$, $\mathbf{R}_{ss}$ is $n \times n, \mathbf{R}_{rr}$ is $N_r \times N_r$, $\mathbf{R}_{sr}$ is $n \times N_r$ and $\mathbf{R}_{rs} = \mathbf{R}_{sr}^T$. Under (2) we can express variance-covariance matrix of $\mathbf{Y}$ as

$$D^2(\mathbf{Y}) = \mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T \tag{3}$$

and variance-covariance matrix of $\mathbf{Y}_s$ as

$$D^2(\mathbf{Y}_s) = \mathbf{V}_{ss} = \mathbf{R}_{ss} + \mathbf{Z}_s\mathbf{G}\mathbf{Z}_s^T \tag{4}$$

We assume that matrices $\mathbf{R}$ and $\mathbf{G}$ (and hence matrix $\mathbf{V}$) depend on some variance parameters $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_q]^T$.

2 BLU predictors

In this paragraph we present two theorems which give formulae of BLU predictors and its MSEs.

Theorem 2.1. *(Royall [7]). Assume that population data obey the general linear model (see equation (1)). Among linear, model-unbiased predictors $\hat{\theta} = \mathbf{g}_s^T\mathbf{Y}_s$ of linear combination of random variables $\theta = \boldsymbol{\gamma}^T\mathbf{Y}$ (where $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_s^T, \boldsymbol{\gamma}_r^T]^T$) the MSE is minimized by:*

$$\hat{\theta_{BLU}} = \boldsymbol{\gamma}_s^T\mathbf{Y}_s + \boldsymbol{\gamma}_r^T\left[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_S - \mathbf{X}_s\hat{\boldsymbol{\beta}})\right], \tag{5}$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{Y}_s. \tag{6}$$

The MSE of $\hat{\theta}_{BLU}$ is given by

$$\begin{aligned} MSE(\hat{\theta_{BLU}}) = {} & \boldsymbol{\gamma}_r^T(\mathbf{V}_{rr} - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr})\boldsymbol{\gamma}_r + \\ & + \boldsymbol{\gamma}_r^T(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s)(\mathbf{X}_s\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}(\mathbf{X}_r - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^T\boldsymbol{\gamma}_r \end{aligned} \tag{7}$$

The proof of the theorem is presented in details for example by Valliant, Dorfman, Royall (2000) pp. 29-30.

Theorem 2.2. *(Henderson [3], following Rao [6]). Assume that population data obey the general mixed linear model (see equation (2)). Among linear, model-unbiased predictors $\hat{\theta}^s = \mathbf{a}^T \mathbf{Y}_s + b$ of linear combination of $\boldsymbol{\beta}$ and the realization of $\mathbf{v}$ given by $\theta^s = \mathbf{l}^T\boldsymbol{\beta} + \mathbf{m}^T\mathbf{v}$ the MSE is minimized by:*

$$\hat{\theta}^s_{BLU} = \mathbf{l}^T\hat{\boldsymbol{\beta}} + \mathbf{m}^T\hat{\mathbf{v}} \tag{8}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{Y}_s \tag{9}$$

$$\hat{\mathbf{v}} = \mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}\left(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}\right) \tag{10}$$

The MSE of $\hat{\theta}^s_{BLU}$ is given by

$$\begin{aligned} &MSE_{\xi}(\hat{\theta}^s_{BLU}) = \mathbf{m}^T\left(\mathbf{G} - \mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}\mathbf{Z}_s\mathbf{G}\right)\mathbf{m} + \\ &+ \left(\mathbf{l}^T - \mathbf{m}^T\mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)\left(\mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\left(\mathbf{l}^T - \mathbf{m}^T\mathbf{G}\mathbf{Z}_s^T\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^T \end{aligned} \tag{11}$$

The proof of the theorem is presented in details for example by Rao [6, pp. 112-113].

The BLU predictor (5) of θ is a generalisation of the BLU predictor (8) of θ^S because of two reasons. First, it is derived for the general linear model which is the generalisation of the general linear mixed model. Second, for the general mixed linear model the linear combination of random variables $\mathbf{Y}$ denoted by θ is the generalisation of linear combination of $\boldsymbol{\beta}$ and the realization of $\mathbf{v}$ denoted by θ^S because:

$$\theta = \boldsymbol{\gamma}^T\mathbf{Y} = \boldsymbol{\gamma}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}^T\mathbf{Z}v + \boldsymbol{\gamma}^T\mathbf{e} = \theta^S + \boldsymbol{\gamma}^T\mathbf{e} \tag{12}$$

where

$$\boldsymbol{\gamma}^T\mathbf{X} = \mathbf{l}^T \tag{13}$$

$$\boldsymbol{\gamma}^T\mathbf{Z} = \mathbf{m}^T \tag{14}$$

We should also underline that forms of the predictors in both theorems, i.e. $\hat{\theta} = \mathbf{g}_s^T\mathbf{Y}_s$ and $\hat{\theta}^s = \mathbf{a}^T\mathbf{Y}_s + b$, are equivalent because the equality $b = 0$ is one of conditions of model-unbiasedness in the proof of the theorem 2.2.

3 EBLU predictors

The BLU predictors $\hat{\theta}_{BLU}$ and $\hat{\theta}^s_{BLU}$ depend on the variance parameters $\boldsymbol{\delta}$ which are unknown in practical applications. Replacing $\boldsymbol{\delta}$ by an estimator $\hat{\boldsymbol{\delta}}$, we obtain a two-stage predictors called EBLU predictors. They will be denoted by $\hat{\theta}_{EBLU}$ and $\hat{\theta}^s_{EBLU}$.

Theorem 3.1. *(Kacker i Harville [4]) Assume that population data obey the general mixed linear model. If (i) $E(\hat{\theta}^s_{EBLU})$ is finite; (ii) $\hat{\boldsymbol{\delta}}$ is any even, translation-invariant estimator of $\boldsymbol{\delta}$, that is $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s) = \hat{\boldsymbol{\delta}}(-\mathbf{Y}_s)$ and $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s - \mathbf{X}_s b = \hat{\boldsymbol{\delta}}(\mathbf{Y}_s)$ for all $\mathbf{Y}_s$ and $\mathbf{b}$; (iii) the distributions of $\mathbf{v}$ and $\mathbf{e}_s$ are both symmetric around $\mathbf{0}$ (not necessarily normal), then the two-stage estimator $\hat{\theta}^s_{EBLU}$ remains model-unbiased.*

We should underline that many standard procedures for estimating $\boldsymbol{\delta}$ including maximum likelihood and restricted maximum likelihood yield even, translation-invariant estimators [4].

In the paper we consider more general EBLU predictor $\hat{\theta}_{EBLU}$ based on the formula of BLU predictor $\hat{\theta}_{BLU}$ proposed by Royall [7]. For the general linear mixed model it is given by:

$$\hat{\theta}_{EBLU} = \boldsymbol{\gamma}_s^T \mathbf{Y}_s + \boldsymbol{\gamma}_r^T [\mathbf{X}_r \tilde{\boldsymbol{\beta}} + \hat{\mathbf{V}}_{rs} \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \tilde{\boldsymbol{\beta}})], \tag{15}$$

where

$$\hat{\mathbf{R}}_{ss} = \mathbf{R}_{ss}(\hat{\boldsymbol{\delta}}),\ \hat{\mathbf{R}}_{rs} = \mathbf{R}_{rs}(\hat{\boldsymbol{\delta}}),\ \hat{\mathbf{R}}_{rr} = \mathbf{R}_{rr}(\hat{\boldsymbol{\delta}}), \hat{\mathbf{G}} = \mathbf{G}(\hat{\boldsymbol{\delta}}),$$

$$\hat{\mathbf{V}}_{rs} = \mathbf{V}_{rs}(\hat{\boldsymbol{\delta}}) = \hat{\mathbf{R}}_{rs} + \mathbf{Z}_r \hat{\mathbf{G}} \mathbf{Z}_s^T,\ \hat{\mathbf{V}}_{ss} = \mathbf{V}_{ss}(\hat{\boldsymbol{\delta}}) = \hat{\mathbf{R}}_{ss} + \mathbf{Z}_{ss} + \mathbf{Z}_s \hat{\mathbf{G}} \mathbf{Z}_s^T$$

$$\hat{\mathbf{V}}_{rr} = \mathbf{V}_{rr}(\hat{\boldsymbol{\delta}}) = \hat{\mathbf{R}}_{rr} + \mathbf{Z}_r \hat{\mathbf{G}} \mathbf{Z}_r^T,\ \tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\delta}}) = \left(\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{Y}_s,$$

For the following proof of the model-unbiasedness of the predictor $\hat{\theta}_{EBLU}$ assumption (1) is too general because the information on the structure of random components is needed. Hence, we assume that the equation (2) holds.

Theorem 3.2. *Assume that population data obey the general mixed linear model. If (i) $E(\hat{\theta}_{EBLU}$ is finite; (ii) $\hat{\boldsymbol{\delta}}$ is any even, translation-invariant estimator of $\boldsymbol{\delta}$, that is $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s) = \hat{\boldsymbol{\delta}}(-\mathbf{Y}_s)$ and $\hat{\boldsymbol{\delta}}(\mathbf{Y}_s - \mathbf{X}_s b) = \hat{\boldsymbol{\delta}}(\mathbf{Y}_s)$ for all $\mathbf{Y}$ and $\mathbf{b}$; (iii) the distributions of $\mathbf{v}$ and $\mathbf{e}$ are both symmetric around $\mathbf{0}$ (not necessarily normal), then the two-stage estimator $\hat{\theta}_E BLU$ is model-unbiased.*

The following proof may be treated as a generalisation of the proof of Theorem 4 presented by Kackar and Harville [4].
Proof of Theorem 3.2 First, consider following lemma.

Lemma 3.1. *(Kacker i Harville [4]). If* $\mathbf{z}$ *is a random vector which has a symmetric distribution around zero in the sense that* $\mathbf{z}$ *and* $-\mathbf{z}$ *are identically distributed and* $f(\mathbf{z})$ *is a random variable that is an odd function of* $\mathbf{z}$ *in the sense that* $f(-\mathbf{z}) = -f(\mathbf{z})$*, then* $f(\mathbf{z})$ *has a symmetric distribution around zero.*

Proof of the lemma is presented by Kacker and Harville [4].

To prove model-ubiasedness of the predictor $\hat{\theta}_{EBLU}$ we must show that $E(\hat{\theta}_{EBLU} - \theta) = 0$. Using (15) we have that

$$\hat{\theta}_{EBLU} - \theta = \boldsymbol{\gamma}_s^T \mathbf{Y}_s + \boldsymbol{\gamma}_r^T [\mathbf{X}_r \tilde{\boldsymbol{\beta}} + \hat{\mathbf{V}}_{rs} \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \tilde{\boldsymbol{\beta}})] - \boldsymbol{\gamma}^T \mathbf{Y} \tag{16}$$

Using second assumption of the Theorem 4 (that $\hat{\boldsymbol{\delta}}$ is any even, translation-invariant estimator of $\boldsymbol{\delta}$) we obtain

$$\hat{\boldsymbol{\delta}}(\mathbf{Y}_s) = \hat{\boldsymbol{\delta}}(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}) = \hat{\boldsymbol{\delta}}(\mathbf{Z}_s \mathbf{v} + \mathbf{e}_s) = \hat{\boldsymbol{\delta}}(-\mathbf{Z}_s \mathbf{v} - \mathbf{e}_s) \tag{17}$$

Hence:

$$\hat{\mathbf{V}}_{ss} = \mathbf{V}_{ss}(\hat{\boldsymbol{\delta}}(\mathbf{Y}_s)) = \mathbf{V}_{ss}(\hat{\boldsymbol{\delta}}(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta})) = \mathbf{V}_{ss}(\hat{\boldsymbol{\delta}}(\mathbf{Z}_s \mathbf{v} + \mathbf{e}_s)) = \mathbf{V}_{ss}(\hat{\boldsymbol{\delta}}(-\mathbf{Z}_s \mathbf{v} - \mathbf{e}_s)) \tag{18}$$

$$\hat{\mathbf{V}}_{rs} = \mathbf{V}_{rs}(\hat{\boldsymbol{\delta}}(\mathbf{Y}_s)) = \mathbf{V}_{rs}(\hat{\boldsymbol{\delta}}(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta})) = \mathbf{V}_{rs}(\hat{\boldsymbol{\delta}}(\mathbf{Z}_s \mathbf{v} + \mathbf{e}_s)) = \mathbf{V}_{rs}(\hat{\boldsymbol{\delta}}(-\mathbf{Z}_s \mathbf{v} - \mathbf{e}_s)) \tag{19}$$

Assuming that population data obey the general mixed linear model the equation (16) may be written as follows

$$\hat{\theta}_{EBLU} - \theta = \boldsymbol{\gamma}_r^T [\mathbf{X}_r \tilde{\boldsymbol{\beta}} + \hat{\mathbf{V}}_{rs} \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \tilde{\boldsymbol{\beta}})] - \boldsymbol{\gamma}_r^T \mathbf{Y}_r =$$

$$= -\boldsymbol{\gamma}_r^T (\mathbf{Y}_r - \mathbf{X}_r \tilde{\boldsymbol{\beta}}) + \boldsymbol{\gamma}_r^T \hat{\mathbf{V}}_{rs} \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \tilde{\boldsymbol{\beta}}) = \tag{20}$$

$$= -\boldsymbol{\gamma}_r^T (\mathbf{X}_r \boldsymbol{\beta} + \mathbf{Z}_r \mathbf{v} + \mathbf{e}_r - \mathbf{X}_r \tilde{\boldsymbol{\beta}}) + \boldsymbol{\gamma}_r^T \hat{\mathbf{V}}_{rs} \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{X}_s \boldsymbol{\beta} + \mathbf{Z}_s \mathbf{v} + \mathbf{e}_s - \mathbf{X}_s \tilde{\boldsymbol{\beta}}) =$$

$$-\boldsymbol{\gamma}_r^T \mathbf{X}_r (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - \boldsymbol{\gamma}_r^T (\mathbf{Z}_r \mathbf{v} + \mathbf{e}_r) + \boldsymbol{\gamma}_r^T \hat{\mathbf{V}}_{rs} \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \boldsymbol{\gamma}_r^T \hat{\mathbf{V}}_{rs} \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{Z}_s \mathbf{v} + \mathbf{e}_s)$$

Consider expected values of the elements of the sum (20). The first element of the sum (20) may be written as follows:

$$-\boldsymbol{\gamma}_r^T \mathbf{X}_r \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right) = -\left(\boldsymbol{\gamma}_r^T \mathbf{X}_r \boldsymbol{\beta} - \boldsymbol{\gamma}_r^T \mathbf{X}_r \left(\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{X}_s \boldsymbol{\beta} + \mathbf{Z}_s \mathbf{v} + \mathbf{e}_s)\right)$$

$$= -\left(\boldsymbol{\gamma}_r^T \mathbf{X}_r \boldsymbol{\beta} - \boldsymbol{\gamma}_r^T \mathbf{X}_r \boldsymbol{\beta} - \boldsymbol{\gamma}_r^T \mathbf{X}_r \left(\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{Z}_s \mathbf{v} + \mathbf{e}_s)\right) =$$

$$= \boldsymbol{\gamma}_r^T \mathbf{X}_r \left(\mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \hat{\mathbf{V}}_{ss}^{-1} (\mathbf{Z}_s \mathbf{v} + \mathbf{e}_s) \tag{21}$$

From (18), the expression (21) is an odd function of the random variable $\mathbf{Z}_s \mathbf{v} + \mathbf{e}_s$. Thus, it follows from the lemma 1 that the random variable (21) has a distribution which is symmetrically distributed around zero. Hence, for the first element of the sum (20) we receive that $E(-\boldsymbol{\gamma}_r^T \mathbf{X}_r (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})) = 0$.

The third assumption of the theorem 3.2 (that the distributions of $\mathbf{v}$ and $\mathbf{e}$ are both symmetric around $\mathbf{0}$) implies for the second element of the sum (20) that $E\left(\boldsymbol{\gamma}_r^T(\mathbf{Z}_r\mathbf{v}+\mathbf{e}_r)\right)=0$.

To analyse expected value of the third element of the sum (20) we use the following transformation:

$$\boldsymbol{\gamma}_r^T\hat{\mathbf{V}}_{rs}\hat{\mathbf{V}}_{ss}^{-1}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)=\boldsymbol{\gamma}_r^T\hat{\mathbf{V}}_{rs}\hat{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\left(\mathbf{X}_s^T\hat{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s^T\hat{\mathbf{V}}_{ss}^{-1}\left(\mathbf{X}_s\boldsymbol{\beta}-\mathbf{Y}_s\right)$$

$$=-\boldsymbol{\gamma}_r^T\hat{\mathbf{V}}_{rs}\hat{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\left(\mathbf{X}_s^T\hat{\mathbf{V}}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s^T\hat{\mathbf{V}}_{ss}^{-1}\left(\mathbf{Z}_s\mathbf{v}+\mathbf{e}_s\right) \tag{22}$$

From (18) and (19), the expression (22) is an odd function of random variable $\mathbf{Z}_s\mathbf{v}+\mathbf{e}_s$. Thus, it follows from the lemma 1 that the random variable (22) has a distribution which is symmetrically distributed around zero. Hence, for the third element of the sum (22) we receive that $E\left(\boldsymbol{\gamma}_r^T\hat{\mathbf{V}}_{rs}\hat{\mathbf{V}}_{ss}^{-1}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)\right)=0$.

From (18) and (19), the fourth element of the sum (20) is also an odd function of the random variable $\mathbf{Z}_s\mathbf{v}+\mathbf{e}_s$. Thus, it follows from the lemma 3.1 that the fourth element of the sum (20) has a distribution which is symmetrically distributed around zero. Hence, for the fourth element of the sum (20) we receive that $E\left(\boldsymbol{\gamma}_r^T\hat{\mathbf{V}}_{rs}\hat{\mathbf{V}}_{ss}^{-1}\left(\mathbf{Z}_s\mathbf{v}+\mathbf{e}_s\right)\right)=0$.

We thus conclude that expected values of all elements of the expression $\hat{\theta}_{EBLU}-\theta$ are equal zero. Hence $E\left(\hat{\theta}_{EBLU}-\theta\right)=0$, provided that $E\left(\hat{\theta}_{EBLU}\right)$ is finite (the first assumption of the theorem 3.2).

4 Example

In the example we compute values of EBLU predictor (15) based on the data on Swedish municipalities from Särndal, Swensson, Wretman [8]. Revenues from the 1985 municipal taxations in millions of kronor in Swedish municipalities are the variable of interest. The 1975 population size in thousands of people is the auxiliary variable. The largest three municipalities are treated as outliers and they are excluded from the analysis. The population of 281 municipalities is divided into eight regions which are treated as domains of interest. Number of domains will be denoted by D. The purpose of the survey is prediction of domain totals of revenues from 1985 municipal taxations based on sample which consists of 56 elements.

Consider the domain $d(d=1,\cdots,D)$ of size N_d. Because we study the problem of prediction of the total in the domain d, the element i of the vector $\boldsymbol{\gamma}$ in the equation (15) equals 1 if the element i of the population belongs to the domain d or it equals 0 otherwise. We assume that population data obey the nested-error regression model [1] with one auxiliary variable and no intercept, which is a special case of the general mixed linear model with the assumption (2). For the element i of the domain d it is assumed that:

$$\mathbf{Y}_{id}=x_{id}\boldsymbol{\beta}+\nu_d+\mathbf{e}_{id} \tag{23}$$

where $i = 1, \cdots, N_d$ and $d = 1, \cdots, D$. The random errors ν_d are assumed to be independent with variance σ_ν^2 , independent of the errors e_{ij} which are assumed to be independent with variance σ_e^2 . Hence $\boldsymbol{\delta} = \lfloor \sigma_\nu^2, \sigma_e^2 \rfloor$. The parameters σ_ν^2 and σ_e^2 are estimated using maximum likelihood method under normality assumption.

In the table values of the EBLU predictors of domain totals are computed based on the equation (15) and the assumption (23). To estimate $MSE(\hat{\theta}_{EBLU})$ we approximate it by $MSE(\hat{\theta}_{BLU})$ and then substitute $\boldsymbol{\delta}$ by $\hat{\boldsymbol{\delta}}$. This approach for estimation of $MSE(\hat{\theta}^s_{EBLU})$ is called “naive approach” (e.g. Rao [6, p. 104]. For the predictor $\hat{\theta}^s_{EBLU}$ other estimators of MSE are also studied in the literature (e.g. Prasad and Rao [5], Datta and Lahiri [2]). Values of the estimated relative root MSE presented in the table are computed by dividing values of the estimated root MSEs by values of EBLU predictors.

Results in the table show that the EBLU predictor $\hat{\theta}_{EBLU}$ based on the formula of Royall's BLU predictor may be a valuable method for prediction of domain total in practical applications.

Domain (region) d	**Value of EBLU predictor**	**Estimated root MSE**	**Estimated relative root MSE (in %)**
1	7685,6028	112,2124	1,4600
2	10752,3560	238,0463	2,2139
3	5866,7760	219,0579	3,7339
4	7316,4939	293,6900	4,0141
5	8458,4318	219,3256	2,5930
6	355,2387	268,3472	4,2225
7	3012,9580	129,3590	4,2934
8	3871,9508	200,4267	5,1764

5 Conclusion

In the paper we consider the BLU predictor proposed by Royall [7] and we show that it may be treated as the generalisation of the BLU predictor proposed by Henderson [3]. We prove that EBLU predictor based on the formula of the Royall's BLU predictor is model-unbiased as EBLU predictor based on the formula of the Henderson's BLU predictor. The proof may be treated as the generalisation of the results received by Kackar and Harville [3]. Additionally, we present short example based on data from Särndal, Swensson, Wretman [8]. In the example we predict domain totals assuming that population data obey a special case of the general mixed linear model.

References

[1] Battese G. E., Harter R.M., Fuller W.A. (1988). *An error-components model for prediction of county crop areas using survey and satellite data.*

Journal of the American Statistical Association **83**, 28 – 36.

[2] Datta G.S., Lahiri P. (2000). *A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems.* Statistica Sinica **10**, 613 – 627.

[3] Henderson C.R. (1950). *Estimation of genetic parameters* (Abstract). Annals of Mathematical Statistics **21**, 309 – 310.

[4] Kackar R.N., Harville D.A. (1981). *Unbiasedness of two-stage estimation and prediction procedures for mixed linear models.* Communications in Statistics, Series A **10**, 1249 – 1261.

[5] Prasad N.G.N., Rao J.N.K. (1990). *The estimation of mean the mean squared error of small-area estimators.* Journal of the American Statistical Association **85**, 163 – 171.

[6] Rao J.N.K. (2003). *Small area estimation.* John Wiley & Sons, New York.

[7] Royall R.M. (1976). *The linear least squares prediction approach to two-stage sampling.* Journal of the American Statistical Association **71**, 657 – 473.

[8] Särndal C.E., Swensson B., Wretman J. (1992). *Model assisted survey sampling*, Springer-Verlag, New York.

[9] Valliant R., Dorfman A.H., Royall R.M. (2000). *Finite population sampling and inference. A prediction approach.* John Wiley & Sons, New York.

Address: T. Zadlo, Department of Statistics, University of Economics, Bogucicka 14, 40-226 Katowice, Poland

E-mail: `zadlo@ae.katowice.pl`

A GRAPHICAL PROCEDURE TO ASSESS THE UNCERTAINTY OF SCORES IN PRINCIPAL COMPONENT ANALYSIS

Manuel Zarzo

Key words: Principal Component Analysis, cross-validation, clustering.

COMPSTAT 2004 section: Multivariate analysis.

Abstract: This paper describes a graphical procedure to assess the uncertainty of scores that consists of displaying in a plot the deviations suffered by scores if any observation is left out and afterwards projected over the model. This chart can be useful to detect influential observations, identify clusters and to determine if a certain component provides relevant information. Applying this procedure to a matrix of 16 observations by 20 variables, only the first component can be considered statistically significant according to cross-validation. Anyway, in the second and third ones, scores suffer moderate shifts when an observation is removed, and several clusters have been established among the observations after checking the uncertainty regions of scores. Thus, it can be concluded that both components also provide significant information. This example reveals the utility of the method to identify clusters and as a graphical tool complementary to cross-validation aimed at diagnosing if a component is significant.

1 Introduction: Principal Component Analysis

Principal Component Analysis (PCA) is one of the multivariate statistical techniques wider spread. It was first proposed in statistics by Pearson [6], and ample literature can be found about its theoretical fundamentals and applications [3], [4], [9]. One of the main advantages of this technique is that it handles matrixes with more variables than individuals, contrary to other multivariate classical methods. To understand the fundamentals of PCA let us consider a matrix $\mathbf{X}$ formed by n variables (in columns) measured for a set of m observations or individuals (in rows). If all variables have the same variance and are not correlated (independent), the display of observations in the n-dimensional space of variables would have the shape of a hyperspherical cloud of points, with no preferential directions of variability. But in practice a certain correlation very often exists among variables, and then observations form certain directions in the space of variables. The first principal component (PC) is the line that, crossing the origin of co-ordinates, better adjusts the observations according to the criterion of minimum squares. It is the direction of maximum inertia or variance of data so that when observations are projected over that direction, the variance of those projections is maximum. These projections are called *scores*, and the *score vector* $\mathbf{t}$ contains the projections of a certain component. The first PC is the direction that crosses the origin of co-ordinates and the centre of gravity (average value) of observations. In most cases this average value does not contain information, and for this reason data are frequently centred before carrying out a PCA, so that the mean of all variables becomes null. Thus, with centred data, principal components are those directions of maximum variance of the observations that cross their centre of gravity. The direction associated to a certain component is determined by a *loading vector* **p**, that contains the weights or contributions of variables in the formation of that component. The display of scores corresponding to two different components in a scatter plot (denominated *score plot*) allows the analysis of relationships among observations: identification of clusters (groups of individuals with a similar pattern), classification of observations and detection of outliers. The equivalent charts for weights (*loading plots*) allow the analysis of the correlation structure of variables and provide a practical interpretation of the components.

Different methods can be used to carry out a PCA. Some of them calculate all components at once, like the singular value decomposition method (SVD), but they have computational problems when are applied to matrixes with a large number of variables highly correlated, which are common in practice. This problem is overcome by sequential algorithms that extract components one by one, like the NIPALS algorithm, developed by Wold [7]. Moreover, the first components are the most important, since they model the data variability and provide relevant information, while the last ones are often related with random disturbances.

For the first PC, once known the vectors **t** and **p**, a residual matrix $\mathbf{E}_1$ is calculated with equation 1. The second PC is the direction that remains orthogonal to the first one and accounts for the maximum inertia of $\mathbf{E}_1$, and again a residual matrix is calculated with equation 2. This procedure can be conducted sequentially until **E**=0, what implies the total decomposition of **X** according to the generalisation of equation 2. However, this extraction is usually stopped when considered conveniently, and the rest of components not calculated are included in a matrix of errors or residuals as stated in equation 3, that describes the decomposition of an **X** matrix in j principal components. The number of principal components that can be extracted from **X** coincides with its range, that working with centred variables is the minimum of m-1 and n.

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1 \cdot \mathbf{p}_1^T \tag{1}$$

$$\mathbf{E}_2 = \mathbf{X} - \mathbf{t}_1 \cdot \mathbf{p}_1^T - \mathbf{t}_2 \cdot \mathbf{p}_2^T \tag{2}$$

$$\mathbf{X} = \mathbf{t}_1 \cdot \mathbf{p}_1^T + \mathbf{t}_2 \cdot \mathbf{p}_2^T + \cdots + \mathbf{t}_j \cdot \mathbf{p}_j^T + \mathbf{E} \tag{3}$$

When principal components are obtained sequentially from a matrix, the variance of a certain score vector is smaller than the one from the previous component. When this variance becomes too low, it reflects that probably the component does not provide relevant information about data, so that the extraction of new components should stop. But it is not always easy to determine how many components should be calculated to model the data variability. Working with standardised variables, a common criterion is extracting all components with an eigenvalue bigger than one. Anyway, perhaps the criterion applied more often is cross-validation [1], [5], [8].

One PC is considered that provides significant information if it does not change much when several observations are removed. The extraction of components is stopped when the capability of the model to predict the discarded observations does not increase any more. To determine if a certain principal component j is statistically significant, the goodness of fit for that component obtained by cross-validation, $Q^2(j)$, can be used, according to equation 6. The residual sum of squares (RSS) is calculated with equation 4, as the summation of the squared distance between the values observed for an individual i and the ones predicted by the PCA model with j components. The prediction is obtained with equation 3 considering $\mathbf{E}$=0. PRESS (PRediction Error Sum of Squares) is similar to RSS, but the predicted values for the i-th observation are calculated with a model that does not contain it, as indicated in expression 5 with the subscript (-i). A certain component j provides systematic information and can be considered statistically significant if the residual sum of squares obtained by cross-validation with a model including that component ($PRESS_j$) is smaller than the residual sum of squares with one component less (RSS_{j-1}). This is equivalent to say $Q^2(j) > 0$.

$$RSS = \sum_{i=1}^{i=m} [(x_{i1} - \hat{x}_{i1})^2 + \ldots + (x_{in} - \hat{x}_{in})^2] \tag{4}$$

$$PRESS = \sum_{i=1}^{i=m} [(x_{i1} - \hat{x}_{(-i)1})^2 + \ldots + (x_{in} - \hat{x}_{(-i)n})^2] \tag{5}$$

$$Q_j^2 = 1 - \frac{PRESS_j}{RSS_{j-1}} \tag{6}$$

Cross-validation is suitable for a reduced number of observations, what quite often occurs in practice. In these situations it seems reasonable to leave out one observation at every round. But with a high number of observations, the model will suffer a negligible change when removing a certain observation, unless it is an influential one. Thus, it will be more reasonable to calculate the PRESS eliminating several observations simultaneously, constructing the model and afterwards obtaining the predicted values for those observations.

If a component is not significant according to cross-validation ($Q_j^2 < 0$) sometimes it is due to the presence of one or more influential observations or outliers. Possibly the PC might become significant if these observations are discarded. If a component is not significant and there are no outliers, this situation corresponds in theory to a residual matrix nearly hyperspherical, with no dominant or preferential directions of variability, so that when some observations are left out and afterwards projected over the model, the directions of maximum inertia change considerably. But in practice no graphical procedure has been proposed in order to evaluate how and how much these directions change. A certain component will be significant if the removal of any observation provokes a slight or moderate shift in scores. The analysis of these shifts displayed in a score plot could be useful for the classification of observations, specially in those cases when clusters appear quite close to each other. If two components are not significant, in theory their score plot will change quite much. Thus, score plots are graphical tools that could also be used to evaluate if a component can be considered significant.

The objective of this paper is to propose a graphical method for assessing the uncertainty of scores, that is, how much they are modified if any observation is left out, in order to:

- Identify influential observations or outliers.
- Facilitate the establishment of clusters.
- Assess the significance of components.

2 Data, method and results

The proposed procedure is applied to an **X** data matrix formed by 16 European countries and 20 variables, that are the relative consumption of 20 common foodstuffs ranging between 0 and 100%. This matrix is included in the user's guide of the software SIMCA-P [2]. It uses the NIPALS algorithm and cross-validation. A PCA has been conducted with this software, with data centred and scaled to unit variance. The characteristics of the first 5 components are shown in table 1. The eigenvalue is λ, and R_x^2 is the proportion of the data variance explained by a component. Q^2 has been calculated with a cross-validation number of rounds $CVr = 16$ (equal to the number of observations) and with 7 rounds (the default considered by the software). The Q^2 limit to consider the component as significant, according to the software, is also shown.

PC	R_x^2	R_x^2(cum)	λ	$Q^2_{CVr=16}$	$Q^2_{CVr=7}$	limit
1^{st} comp.	0.317	0.317	5.07	0.108	0.065	0.107
2^{nd} comp.	0.192	0.510	3.08	0.018	-0.106	0.113
3^{rd} comp.	0.138	0.648	2.21	-0.016	-0.029	0.120
4^{th} comp.	0.081	0.729	1.29	-0.500	-0.490	0.128
5^{th} comp.	0.062	0.791	0.99	-0.542	-0.463	0.137

Table 1: Results of the Principal component Analysis applied to **X**.

To check if the results in table 1 are conditioned by influential observations, different score plots have been analysed, but no outliers appear with

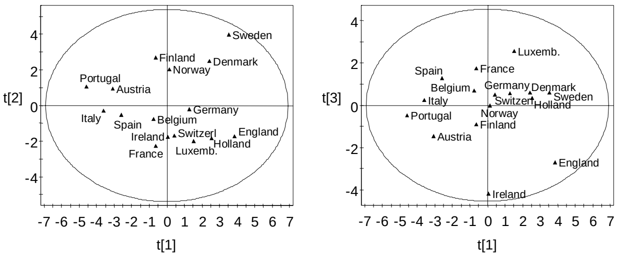


Figure 1: Score plots of the principal components: first versus second (left) and first versus third (right), calculated with all observations.

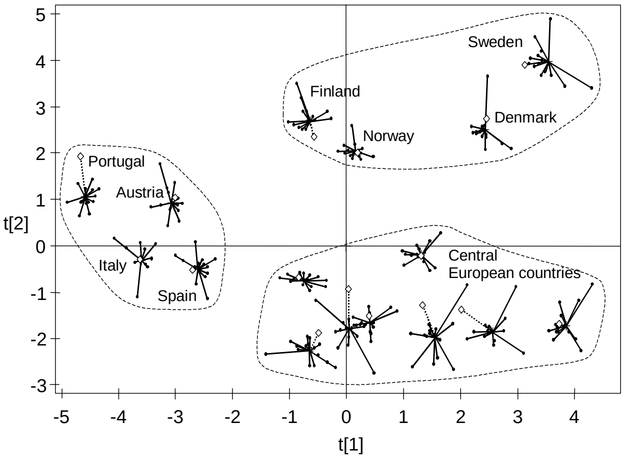


Figure 2: Score plot with uncertainties for the first and second components.

anomalous high scores. Also there are no outliers regarding the distance of observations to the model fitted with the first component. Aimed at obtaining clusters among countries with a similar pattern according to the eating habits, two score plots are shown in figure 1.

In the second PC appears a cluster formed by 4 countries, and the third PC shows differences regarding two countries. But in order to establish the clusters properly, it would be interesting to know the uncertainty of these scores. For this purpose, first a PCA has been conducted without one country, obtaining the scores corresponding to the three first components. Doing the same with every country, 16 score plots have been plotted with the first and second components. Figure 2 shows the results if all of them are overlapped in a same graph, including also the one calculated with all observations. Afterwards every country is projected over the model fitted without that country, and another score plot is obtained, that has also been overlapped in figure 2 (with the scores shown as ♢). Figure 3 is equivalent, with the first and third components. As these charts contain too many points, to facilitate their interpretation the scores corresponding to a same observation have been joined with the initial score obtained with the model using all observations, that occupies a central position. Thus, every line represents the displacement or shift that suffers a score if an observation is removed from the model.

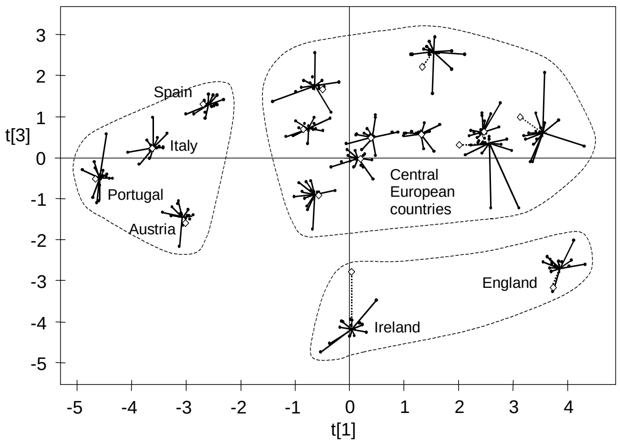


Figure 3: Score plot with uncertainties for the first and third components.

3 Discussion

Although 15 principal components can be extracted from this matrix, only the first five components are enough to explain 79 % of the data variance, that according to table 1 have an eigenvalue similar or greater than one. The software considers that a component is significant if Q^2 is larger than a significance limit, shown in table 1. The first PC has a positive Q^2 considering 7 or 16 cross-validation rounds (CVr) and then it can be regarded as significant. In the second component Q^2 is slightly higher than zero considering CVr=16, but lower if CRv=7. So, it seems that only the first PC provides systematic information. The rest ones do not appear to be significant, since $Q^2 < 0$. In this situation intuitively one imagines that when an observation is left out for cross-validation, the second and third components will change too much. Therefore, one should expect that their score plot would be modified substantially when removing observations.

In figures 2 and 3 the set of points joined with the initial score define an uncertainty region where the scores are displaced when any observation is removed. It can be observed that the variations on the first component are smaller than the ones registered for the second one, and these are slightly smaller than for the third one. But these regions are not so big as might be expected for the second and third components, whose Q^2 is negative. On the other hand, when one observation is left out and afterwards projected over the model (scores indicated with the symbol $\diamondsuit$), most of them are located approximately inside the uncertainty region defined by the rest of scores, what reflects that these components are consistent.

Considering the uncertainty regions, a cluster can be established in figure 2 with the four countries characterised by the most negative scores in the first component: Portugal, Italy, Spain and Austria. Three of them are Mediterranean countries. It is possible to form another cluster with the Scandinavian countries (Finland, Norway, Denmark and Sweden), that present the highest values in the second component. The rest are Central European countries. As these clusters correspond with geographic differences, the second component can be considered statistically significant.

Ireland and England do not differ from the rest of Central European countries in the first and second components, but in the third one they reflect a different pattern, as stated in figure 3. Both countries form a cluster that again corresponds with geographical differences, what could explain a different habit in the food consumption. Taking into account that the uncertainty regions in figure 3 are not overlapped with the ones of the rest of Central European countries, it can be concluded that the third component provides relevant information and can be considered significant, despite $Q^2 = -0.016$.

The four clusters detected reveal groups of countries with a similar pattern regarding the food consumption. In order to identify the foodstuffs that characterise each cluster, the next step would be to analyse the loadings in the three principal components. Similar to the procedure followed to obtain figures 2 and 3, uncertainty regions for the weights could also be calculated overlapping the loading plots corresponding to the 17 PCA models. This could be useful to form clusters among variables in matrixes with a low number of variables, otherwise the charts might become difficult to interpret.

4 Conclusions

This paper presents a graphical procedure to assess the uncertainty of scores aimed at detecting influential observations, facilitating the formation of clusters and determining if a certain component provides relevant information. It consists of displaying graphically in a score plot the shifts registered by scores if any observation is left out of the model. This procedure has been applied to one example where only the first PC seems to be significant according to cross-validation. But with this procedure it can be concluded that also the second and third components provide relevant information. This method has the advantage that the results are displayed graphically, the implementation in software is simple and it can constitute a complementary tool in cross-validation to decide whether a component is significant. This procedure is useful with any number of variables, but requires a reduced set of observations in order to avoid the overlapping of many uncertainty regions, what would make more difficult the interpretation of plots.

References

[1] Eastment H.T., Krzanowski W.J. (1982). *Cross-validatory choice of the number of components from a principal component analysis.* Technometrics **24**, 73 – 77.

[2] Eriksson L., Johansson E., Kettaneh-Wold N., Wold S. (1999). *Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS).* Umetrics, Sweden, http://wwww.umetrics.com.

[3] Jackson J.E. (1991). *A user's guide to Principal Components.* John Wiley & Sons, New York.

[4] Jolliffe I.T. (1986). *Principal component analysis.* Springer-Verlag, New York.

[5] Krzanowski W.J. (1987). *Cross-validation in principal component analysis.* Biometrics **43**, 575 – 584.

[6] Pearson K. (1901). *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine **2**, 559 – 572.

[7] Wold H. (1966). *Estimation of principal components and related models by iterative least squares.* In *Multivariate Analysis*, Krishnaiah, P.R. (ed.). Academic Press, New York, 391 – 420.

[8] Wold S. (1978). *Cross-validatory estimation of the number of components in factor and principal components models.* Technometrics **20**, 397 – 405.

[9] Wold S., Esbensen K., Geladi P. (1987). *Principal component analysis.* Chemom. Intell. Lab. Syst. **2**, 37 – 52.

Acknowledgement: The author is grateful to Miguel Angel Edo and Eva Vallada for their valuable help in the use of LaTeX for preparing this paper.

Address: M. Zarzo, Department of Applied Statistics, Operations Research and Quality; Polytechnic University of Valencia; Camino de vera s/n, Edificio I-3, 46022, Valencia, Spain

E-mail: mazarcas@eio.upv.es

Author Index

COMPSTAT 2004 Section Index

Bayesian Methods

Biostatistics

Classification

Design of Experiments

Dimensional Reduction

Historical Keynote

Model Selection

Multivariate Analysis

Neural Networks and Machine Learning

Nonparametrical Statistics

Numerical Methods for Statistics

Official Statistics

Optimization

Partial Least Squares

Robustness

Time Series Analysis