P. Brito (Editor)

Proceedings of COMPSTAT'2008

International Conference on Computational Statistics

Porto - Portugal, August 24th-29th 2008

Contributed Papers

Physica-Verlag

A Springer Company

Preface

The 18th Conference of IASC-ERS, COMPSTAT'2008, is held in Porto, Portugal, from August 24th to August 29th 2008, locally organised by the Faculty of Economics of the University of Porto.
COMPSTAT is an initiative of the European Regional Section of the International Association for Statistical Computing (IASC-ERS), a section of the International Statistical Institute (ISI). COMPSTAT conferences started in 1974 in Wien; previous editions of COMPSTAT were held in Berlin (2002), Prague (2004) and Rome (2006). It is one of the most prestigious world conferences in Computational Statistics, regularly attracting hundreds of researchers and practitioners, and has gained a reputation as an ideal forum for presenting top quality theoretical and applied work, promoting interdisciplinary research and establishing contacts amongst researchers with common interests. COMPSTAT'2008 is the first edition of COMPSTAT to be hosted by a Portuguese institution.

Keynote lectures are addressed by Peter Hall (Department of Mathematics and Statistics, The University of Melbourne), Heikki Mannila (Department of Computer Science, Faculty of Science, University of Helsinki) and Timo Teräsvirta (School of Economics and Management, University of Aarhus). The conference program includes two tutorials: "Computational Methods in Finance" by James Gentle (Department of Computational and Data Sciences, George Mason University) and "Writing R Packages" by Friedrich Leisch (Institut für Statistik, Ludwig-Maximilians-Universität). Each COMPSTAT meeting is organised with a number of topics highlighted, which lead to Invited Sessions. The Conference program includes also contributed sessions in different topics (both oral communications and posters).

The Conference Scientific Program Committee includes Paula Brito (University of Porto, Portugal), Helena Bacelar-Nicolau (University of Lisbon, Portugal), Vincenzo Esposito-Vinzi (ESSEC, France), Wing Kam Fung (The University of Hong Kong, Hong Kong), Gianfranco Galmacci (University of Perugia, Italy), Erricos Kontoghiorghes (University of Cyprus, Cyprus), Carlo Lauro (University of Naples Federico II, Italy), Alfredo Rizzi (University "La Sapienza", Roma, Italy), Esther Ruiz-Ortega (University Carlos III, Spain), Gilbert Saporta (Conservatoire National des Arts et Métiers, France), Michael Schimek (Medical University of Graz, Austria), Antónia Turkman (University of Lisbon, Portugal), Joe Whittaker (University of Lancaster, UK), Djamel A. Zighed (University Lumière Lyon 2, France) and Edward Wegman (George Mason University, USA), who were responsible for the Conference Scientific Program, and whom the

organisers wish to thank for their invaluable cooperation and permanent availability. Special thanks are also due to Tomas Aluja, Chairperson of the IASC-ERS and Jaromir Antoch, IASC President, for their continuous support and collaboration.

Due to space limitations, the Book of Proceedings includes keynote speakers' papers and invited sessions speakers' papers only, while the CD-Rom, which is part of it, includes all accepted papers, as well as the tutorials' support texts. The chapters of this volume hence correspond to contributed sessions, as follows:

Biostatistics, Genomics and Micro-Array Analysis
Categorical Data Analysis
Classification and Discrimination
Clustering
Computational Methods for Industry
Computational Methods in Official Statistics
Data Mining and Machine Learning
Econometrics
Finance and Insurance
Functional Data Analysis
Graphical Models and Bayes Nets
Image and Signal Processing
Multivariate Data Analysis and Dimensionality Reduction
Non-Parametric Statistics and Smoothing
Optimization and Random Search Algorithms
Partial Least Squares and Structural Equations Models
Resampling
Robustness
Simulation

The papers included in this volume present new developments in topics of major interest for statistical computing, constituting a fine collection of methodological and application-oriented papers that characterize the current research in novel, developing areas. Combining new methodological advances with a wide variety of real applications, this volume is certainly of great value for researchers and practitioners of computational statistics alike.

First of all, the organisers of the Conference and the editors would like to thank all authors, both of invited and contributed papers and tutorial texts, for their cooperation and enthusiasm. We are specially grateful to all colleagues who served as reviewers, and whose work was crucial to the scientific quality of these proceedings. We also thank all those who have contributed to the design and production of this Book of Proceedings, Springer

Verlag, in particular Dr. Martina Bihn and Irene Barrios-Kezic, for their help concerning all aspects of publication.

The organisers would like to express their gratitude to the Faculty of Economics of the University of Porto, who enthusiastically supported the Conference from the very start, and contributed to its success, and all people there who worked actively for its organisation. We are very grateful to all our sponsors, for their generous support. Finally, we thank all authors and participants, without whom the conference would not have been possible.

The organisers of COMPSTAT'2008 wish the best success to Gilbert Saporta, Chairman of the 19th edition of COMPSTAT, which will be held in Paris in Summer 2010. See you there!

Porto, August 2008

Paula Brito
Adelaide Figueiredo
Ana Pires
Ana Sousa Ferreira
Carlos Marcelo
Fernanda Figueiredo
Fernanda Sousa
Joaquim Pinto da Costa
Jorge Pereira
Luís Torgo
Luísa Canto e Castro
Maria Eduarda Silva
Paula Milheiro
Paulo Teles
Pedro Campos
Pedro Duarte Silva

Acknowledgements

The Editors are extremely grateful to the reviewers, whose work was determinant for the scientific quality of these proceedings. They were, in alphabetical order :

Andres M. Alonso
Russell Alpizar-Jara
Tomás Aluja-Banet
Conceição Amado
Annalisa Appice
Helena Bacelar-Nicolau
Susana Barbosa
Patrice Bertrand
Lynne Billard
Hans-Hermann Bock
Carlos A. Braumann
Maria Salomé Cabral
Jorge Caiado
Margarida Cardoso
Nuno Cavalheiro Marques
Gilles Celeux
Andrea Cerioli
Joaquim Costa
Erhard Cramer
Nuno Crato
Guy Cucumel
Francisco De A. T. De Carvalho
José G. Dias
Jean Diatta
Pedro Duarte Silva
Lutz Edler
Ricardo Ehlers
Lars Eldén
Vincenzo Esposito Vinzi
Nuno Fidalgo
Fernanda Otília Figueiredo
Mário Figueiredo
Peter Filzmoser
Jan Flusser
Roland Fried
Fabio Fumarola
Wing K. Fung
Gianfranco Galmacci
João Gama
Ivette Gomes
Esmeralda Gonçalves
Gérard Govaert
Maria Do Carmo Guedes
André Hardy
Nick Heard
Erin Hodgess
Sheldon Jacobson
Alípio Jorge
Hussein Khodr
Guido Knapp
Erricos Kontoghiorghes
Stéphane Lallich
Carlo Lauro
S.Y. Lee
Friedrich Leisch
Uwe Ligges
Corrado Loglisci
Rosaria Lombardo
Nicholas Longford
Donato Malerba
Jean-François Mari
J. Miguel Marin
Leandro Marinho
Geoffrey McLachlan
Paula Milheiro-Oliveira
Isabel Molina Peralta
Yuichi Mori
Irini Moustaki
Maria Pilar Muñoz Gracia
Amedeo Napoli
Manuela Neves
João Nicolau

Monique Noirhomme
M. Rosário de Oliveira
Francesco Palumbo
Rui Paulo
Ana Pérez Espartero
Jorge Pereira
Isabel Pereira
Ana Pires
Mark Plumbley
Pilar Poncela
Christine Preisach
Gilbert Ritschard
Alfredo Rizzi
Paulo Rodrigues
J. Rodrigues Dias
Julio Rodriguez
Fernando Rosado
Patrick Rousset
Esther Ruiz
Gilbert Saporta
Radim Sara
Pascal Sarda
Michael G. Schimek
Lars Schmidt-Thieme
Luca Scrucca
Maria Eduarda Silva
Giovani Silva
Artur Silva Lopes
Carlos Soares
Gilda Soromenho
Fernanda Sousa
Ana Sousa Ferreira
Elena Stanghellini
Milan Studeny
Yutaka Tanaka
Paulo Teles
Valentin Todorov
Maria Antónia Turkman
Kamil Turkman
Antony Unwin
Michel Van De Velden
Maurizio Vichi
Philippe Vieu
Jirka Vomlel
Rafael Weissbach
Joe Whittaker
Peter Winker
Michael Wiper
Djamel A. Zighed

Sponsors

We are extremely grateful to the following institutions whose support contributes to the success of COMPSTAT'2008:

ORGANIZERS:

Contents

Part I

Biostatistics, Genomics and Micro-Array Analysis

Searching for Powerful Tests in Shape Analysis

Chiara Brombin[1] and Luigi Salmaso[2]

[1] Department of Statistics, University of Padova
Via Cesare Battisti 241-243, 35121 Padova, Italy, *chiara.brombin@stat.unipd.it*
[2] Department of Management and Engineering, University of Padova
Stradella S. Nicola 3, 36100 Vicenza, Italy, *salmaso@gest.unipd.it*

Abstract. Traditional approaches for the statistical analysis of shape involve methods assessing the difference between configurations of landmarks optimally superimposed using a least-squares procedure or methods based on interlandmark distances. All these methods are based on strong assumptions, like equality of covariance matrices, independence, multivariate normal model for landmarks. Moreover, in almost all real applications, researchers have to cope with few individuals and many landmarks, implying over-dimensioned spaces and loss of power. For these reasons we suggest a nonparametric permutation approach to shape analysis. Focussing on the two independent sample case, through a simulation study, we evaluate the behaviour of some nonparametric permutation tests and we show that the proposed tests are very powerful, both in for balanced and unbalanced sample sizes.

Keywords: morphometric methods, NPC methodology, shape analysis

1 Introduction

Statistical shape analysis relates to the study of random objects, where the concept of shape correspond to some geometrical information that is invariant under translation, rotation and scale effects. According to Kendall (1977) a shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object. In many biological and biomedical investigations, the most effective way to analyze the forms of whole biological organs or organisms is by recording and locating a finite number of points on the outline, i.e. landmarks. A landmark is a point of correspondence on each object that matches between and within population (Dryden and Mardia (1998)). So landmark points do not only have their own locations but also have the same locations in every other form of the study in the average of the all forms of the data set (Bookstein (1991)). The main strategies to analyze databases of landmark locations are multivariate morphometrics and deformation analysis (Bookstein (1986)). Only recently, in the late 1980s, various authors, included Fred Bookstein and Jim Rohlf, proposed a synthesis of these two experiences called geometric morphometrics, that is a collection of approaches for the multivariate statistical analysis of Cartesian coordinate data, usually (but not always) limited to landmark point locations.

A key benefit to use the geometric morphometric methods, instead of *traditional* morphometric methods, is that since all the geometric information is retained throughout a study, results of high-dimensional multivariate analyses can be mapped back into physical space to achieve appealing and informative visualizations that are frequently not possible with alternative methods (Slice (2005)).

2 Inference and shape analysis

Statistical shape analysis is considered a cross-disciplinary field, allowing for applications in biology, geology, medicine and many other sciences, since the theory and techniques are very flexible and potentially adaptable to any appropriate configuration matrices. The statistical community has shown an increased interest in shape analysis in the last decade.
Particular efforts have been addressed to the development of powerful statistical methods based on model for shape variation of entire configurations of point corresponding to the locations of morphological landmarks.
Rohlf (2000) reviews the main tests used in the field of shape analysis and compares the statistical power of various tests that have been proposed to test for equality of shape in two populations. Even if his work is limited to the simplest case of homogeneous, independent, spherical variation at each landmark and the sampling experiments emphasize the case of triangular shapes, it allows the user to choose the method that has the highest statistical power under a set of assumptions that are appropriate for the data. Through a simulation study, he found that Goodall's F-test had the highest power followed by T^2-tests using Kendall tangent space coordinates. Power for T^2-tests using Bookstein shape coordinates was good if the baseline was not the shortest side of the triangle. The Rao and Suryawanshi shape variables had much lower power when triangles were not close to being equilateral. Power surfaces for the `EDMA-I` T statistic revealed very low power for many shape comparisons including those between very different shapes. Power surface for the `EDMA-II` Z statistic depended strongly on the choice of baseline used for size scaling.
All the above mentioned tests are based on strong assumptions. In particular, the tests based on the T^2 statistic (e.g. T^2-tests using Bookstein, Kendall tangent space coordinates, Rao and Suryawanshi shape variables, like Rao-d (1996) and Rao-a (1998)) require independent samples, homogeneous covariance matrices and shape coordinates distributed according to the multivariate normal law. We remark that Hotelling's T^2 test statistic is derived under the assumption of population multivariate normality and it may not be very powerful unless there are a large number of observations available (Dryden and Mardia (1998)). It is well known in the literature that Hotelling's T^2 test is formulated to detect any departures from the null hypothesis and therefore often lacks power to detect specific forms of departures that may arise in practice, i.e. the T^2 test fails to provide an easily implemented one-sided

(directional) hypothesis test (Blair et al. (1994)).
Goodall's F requires a restrictive isotropic model and assumes that the distributions of the squared Procrustes distances are approximately chi-squared distributions.
If we consider the methods based on interlandmark distances, `EDMA-I` T assumes independent samples and the equality of the covariance matrices in the two populations being compared (Lele and Cole (1996)), while `EDMA-II` Z assumes only independent samples and normally distributed variation at each landmark. As pointed out in Good (1994), the assumption of equal covariance matrices may be unreasonable in certain applications, the multinormal model in the tangent space may be doubted and sometimes there are few individuals and many landmarks, implying over-dimensioned spaces and loss of power for the Hotelling's T^2 test. Hence an alternative procedure is to consider a permutation approach. Permutation methods are distribution-free, allow us for quite efficient solutions when the number of cases is less than the number of covariates, may be tailored for sensitivity to specific treatment alternatives and provide one-sided as well as two-sided tests of hypotheses (Blair et al. (1994)). In the wake of these considerations, we propose an extension of the Nonparametric Combination (NPC) methodology (Pesarin (2001)), briefly summarized in the next section.

3 NPC approach to shape analysis

The method is based a) on a decomposition of the hypotheses into k, $k > 1$, sub-hypotheses, where for each sub-hypothesis, there exists a suitable partial permutation test statistic; b) on a simulation procedure, conditional on the set of observed data, which provides an estimate of the null multivariate permutation distribution of the whole set of statistics; c) on a combination of the partial tests into a second-order statistic whose null permutation distribution is estimated by using the same simulation results of the first step. With regard to point a), the k-dimensional hypothesis testing problem is processed in two phases: at first, we define a suitable set of k, with $k \geq 1$, unidimensional permutation tests called *partial tests*. Each partial test examines the marginal contribution of any single response variable in the comparison between several treatment groups. The second phase is the nonparametric combination of dependent tests in an overall *second order combined test*, which is suitable for testing possible global differences between the multivariate distributions of two or more groups. When there is a stratification variable, we expect two combination levels: the partial tests combination in s second order combined tests, $s \geq 1$, within each stratum, and a further combination of such tests into a single *third order combined test*. The NPC methodology is a conditional testing procedure that, under very mild and reasonable conditions, provided that exchangeability of data with respect to groups is satisfied in the null hypothesis, is found to be consistent and unbiased (Celant et al. (2000)).

Focussing on the two independent sample case, through a simulation study, we have compared the behaviour of traditional tests with that of nonparametric permutation tests. In particular we have implemented both global tests derived from the combination of all the partial p-values and global tests derived from the combination of predefined domains, as an extension of the standard NPC methodology. We refer to *domains* as subgroups of landmarks sharing anatomical, biological or locational features.
Let us consider a two-sample problem in which the side-assumptions for the problem are that the treatment may act on the first two moments of responses belonging to the first group. Moreover and without loss of generality, let us assume that dataset and response model behave as $X_{1i} = \mu + \Delta_{1i} + \epsilon_{1i}, X_{2i} = \mu + \epsilon_{2i}$, $i = 1, \ldots, n_j$, $j = 1, 2$, where μ is a population nuisance constant, ϵ_{ji} are exchangeable random errors such that $\mu + \epsilon_{ji} > 0$ in probability, and $\Delta_{1i} \geq 0$ are non-negative stochastic effects which may depend on $\mu + \epsilon_{1i}$, and in addition satisfy the second-order condition $(\mu + \Delta_{1i} + \epsilon_{1i})^2 \geq (\mu + \epsilon_{1i})^2$, $i = 1, \ldots, n_1$. Suppose the hypotheses are $H_0 : \{X_1 \stackrel{d}{=} X_2\}$ against $H_1 : \{X_1 \stackrel{d}{\neq} X_2\}$ and that we are essentially interest in the first two moments, so that the hypotheses become equivalent to $H_0 : \{(\mu_{11} = \mu_{12}) \bigcap (\mu_{21} = \mu_{22})\}$ and $H_1 : \{(\mu_{11} \neq \mu_{12}) \bigcup (\mu_{21} \neq \mu_{22})\}$, where $\mu_{rj} = \mathbb{E}(X_j^r)$ is the rth moment of the jth variable. A Multi Aspect (MA) approach deals with one partial permutation test to each current aspect, $T_1^* = \sum_i X_{1i}^*$ and $T_2^* = \sum_i X_{1i}^{*2}$, followed by their nonparametric combination (Pesarin (2001)). Hence, in order to include the MA procedure in NPC method we perform the above mentioned T_1^* and T_2^* aspect tests for each coordinate of a single landmark and we consider their combination (Salmaso and Solari (2005)). For all combining functions throughout the NPC procedure, we have selected either Liptak and Fisher combining functions. For the combining function choice see the practical guidelines described in Pesarin (2001).

3.1 Simulation setting

Let us assume that our samples are made of configurations of $p = 8$ landmarks in $m = 2$ dimensions characterized by slightly different means. Suppose to deal with male and female skull configurations of a particular animal *ad hoc* created, representing the two independent samples. Since sample means differ from each other in 6 over 16 coordinates (Table 1), we have generated data under the alternative hypothesis H_1 in order to evaluate the power of the competing tests.
In all the simulations we have set the number B of permutations equals to 1000 and the number CMC of Conditional Monte Carlo iterations equals to 1000. Focussing on the two independent sample case, we have carried out the simulation study in the same conditions of homogeneous, independent, spherical variation at each landmark, as described in Rohlf (2000). Hence, we

Table 1. Hypothetical configuration means.

#	Lnd. name	Male		Female	
		x	y	x	y
1	nasion	65.00	223.00	65.00	**222.85**
2	basion	54.00	−40.00	**53.75**	−40.00
3	staphylion	0.00	0.00	0.00	0.00
4	prosthion	0.00	35.00	0.00	**34.50**
5	nariale	19.00	121.00	**18.90**	121.00
6	bregma	70.00	203.00	70.00	203.00
7	lambda	110.00	112.00	**109.95**	112.00
8	opisthion	104.00	17.00	104.00	**16.88**

have considered the conditions in which parametric tests based on T^2 or F statistics and those based on interlandmark distances perform better.
Let n_i, $i = 1, 2$, be the sample size in the two samples.
In the first simulation we have fixed $n_1 = n_2 = 10$, in the second simulation study we have considered the unbalanced sample size case with $n_1 = 50$ and $n_2 = 20$, in the third and last simulation we have set $n_1 = n_2 = 50$.
Two different variances, i.e. $\sigma^2 = 0.25$ and $\sigma^2 = 0.50$, and three domains, i.e. baseline (nasion and basion), face (staphylion, prosthion anf nariale) and braincase (bregma, lambda and opisthion), have been examined. We denote with G the combination of all partial tests and with G_d the combination using domains. Fisher, Liptak are the possible combining functions used and MA, if present, denotes the Multi Aspect procedure.
In Table 2 we display 6 simulations: sim1: $n_1 = n_2 = 10$, $\sigma^2 = 0.25$; sim2: $n_1 = n_2 = 10$, $\sigma^2 = 0.50$; sim3: $n_1 = 50$, $n_2 = 20$, $\sigma^2 = 0.25$; sim4: $n_1 = 50$, $n_2 = 20$, $\sigma^2 = 0.50$; sim5: $n_1 = n_2 = 50$, $\sigma^2 = 0.25$; sim6: $n_1 = n_2 = 50$, $\sigma^2 = 0.50$. First of all, we wish to emphasize that combination with domains results in powerful tests. The greater is the proportion of global tests without domains that are smaller than the counterparts with domains, the smaller are the values assumed by global tests without domains, viz. they are less powerful since we are under H_1. This is an important finding since we can assess that using a NPC methodology we can acquire not only global information about the entire configuration of landmarks, like traditional tests known in literature, but we can also make local or regional assessments.

3.2 Results

We have compared, in terms of statistical power, traditional approaches for the statistical analysis of shape like Hotelling's T^2 test using approximate tangent space coordinates and Bookstein shape coordinates, Goodall's F test, `EDMA-I` and `EDMA-II` tests, T^2-test using Rao and Suryawanshi shape variables, Rao-a and Rao-d (only for $n_1 = n_2 = 50$) to the nonparametric

Table 2. Relative frequencies showing how many times global tests without domains are equal, greater or smaller than the counterparts with domains.

Simulation	sim1	sim2	sim3	sim4	sim5	sim6
G=G_d (Liptak)	0.0220	0.0160	0.1020	0.0380	0.1700	0.0630
G=G_d (Fisher)	0.0220	0.0180	0.1530	0.0550	0.3360	0.1170
G=G_d (Liptak,MA)	0.0250	0.0140	0.0780	0.0430	0.1710	0.0580
G=G_d (Fisher,MA)	0.0290	0.0140	0.1490	0.0480	0.3600	0.1030
G>G_d (Liptak)	0.3960	0.4230	0.3250	0.3710	0.3410	0.3870
G>G_d (Fisher)	0.2760	0.3310	0.1840	0.2010	0.1750	0.1900
G>G_d (Liptak,MA)	0.3770	0.4220	0.3370	0.3590	0.3450	0.3790
G>G_d (Fisher,MA)	0.2570	0.3120	0.1890	0.2200	0.1630	0.2040
G<G_d (Liptak)	0.5820	0.5610	0.5730	0.5910	0.4890	0.5500
G<G_d (Fisher)	0.7020	0.6510	0.6630	0.7440	0.4890	0.6930
G<G_d (Liptak,MA)	0.5980	0.5640	0.5850	0.5980	0.4840	0.5630
G<G_d (Fisher,MA)	0.7140	0.6740	0.6620	0.7320	0.4770	0.6930

permutation tests: nonparametric permutation Hotelling's T^2, nonparametric permutation global tests, with and without domains, using Liptak and Fisher combining functions, nonparametric permutation MA tests, with and without domains, considering location and scale aspects and using Liptak and Fisher combining functions. Because of the lack of space we report only the simulation with $n_1 = n_2 = 10$ (Table 3 and Table 4) and the complete simulation study is available at *www.gest.unipd.it/∼salmaso/COMPSTAT2008_sim.pdf*. Focussing on $\alpha = 0.05$ nominal level, we show that the Hotelling's T^2 permutation counterpart, the Goodall's F test, the global test obtained using Fisher combining function, in its standard, domain and MA versions, are the tests that perform better in almost all the simulations.

4 Conclusion

The proposed nonparametric permutation approach seems to perform better than the traditional tests used in shape analysis. Due to their nonparametric nature, the suggested tests may be computed even when the number of covariates exceeds the number of cases. Moreover they do not rely on a well-defined distributional model, they have the obvious advantage of not requiring the assumption of homogeneity of variance, they deal with one-sided as well as two-sided tests of hypotheses and allow a more flexible analysis even in terms of the nature of the variables (continuous, categorical or mixed) involved in the analysis. Through the simulation study, we have highlithed not only the power of these methods, but also that they enable the researcher to give local assessment using a combination with domains. Theoretical matters of future research concern with the analysis of the case of heterogeneous and depen-

Table 3. Simulation 1 ($n_1 = n_2 = 10$, B=CMC=1000, $\sigma^2 = 0.25$).

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T^2 Approx Proc Tg	0.0260	0.1050	0.2000	0.3670	0.4950	0.7300
T^2 perm	0.0990	0.2700	0.3920	0.5550	0.6580	0.8340
Bookstein shape coords, T^2	0.0220	0.0980	0.1860	0.3360	0.5280	0.7500
Goodall F-test	0.1060	0.2500	0.3980	0.5200	0.6220	0.8120
EDMA-I, Z	0.0580	0.1340	0.2660	0.4280	0.5720	0.8360
EDMA-II,T	0.0140	0.0280	0.0400	0.1300	0.2460	0.4320
G (Liptak)	0.0490	0.1710	0.2810	0.4610	0.5810	0.7660
G_d (Liptak)	0.0400	0.1680	0.2750	0.4360	0.5670	0.7430
G (Fisher)	0.0720	0.2330	0.3650	0.5440	0.6500	0.8100
G_d (Fisher)	0.0670	0.2160	0.3430	0.5120	0.6190	0.7870
G (Liptak, MA)	0.0570	0.1760	0.3030	0.4580	0.5860	0.7490
G_d (Liptak, MA)	0.0520	0.1770	0.2890	0.4420	0.5660	0.7390
G (Fisher, MA)	0.0750	0.2350	0.3650	0.5330	0.6590	0.8010
G_d (Fisher, MA)	0.0650	0.2110	0.3440	0.5080	0.6300	0.7800

Table 4. Simulation 2 ($n_1 = n_2 = 10$, B=CMC=1000, $\sigma^2 = 0.50$).

	α=0.01	α=0.05	α=0.10	α=0.20	α=0.30	α=0.50
T^2 Approx Proc Tg	0.0090	0.0710	0.1440	0.2910	0.4200	0.6240
T^2 perm	0.0450	0.1600	0.2630	0.4010	0.5240	0.6940
Bookstein shape coords, T^2	0.0120	0.0680	0.1580	0.3020	0.4340	0.5780
Goodall F-test	0.0520	0.1740	0.2260	0.3920	0.5380	0.6960
EDMA-I, Z	0.0340	0.1360	0.1780	0.3760	0.4820	0.6820
EDMA-II,T	0.0160	0.0240	0.0500	0.1060	0.2120	0.4300
G (Liptak)	0.0270	0.1080	0.1860	0.3390	0.4430	0.6480
G_d (Liptak)	0.0200	0.0970	0.1830	0.3330	0.4380	0.6300
G (Fisher)	0.0340	0.1340	0.2270	0.3690	0.4980	0.6890
G_d (Fisher)	0.0360	0.1200	0.2100	0.3560	0.4850	0.6750
G (Liptak, MA)	0.0220	0.1090	0.2060	0.3450	0.4460	0.6510
G_d (Liptak, MA)	0.0210	0.0980	0.1930	0.3340	0.4380	0.6470
G (Fisher, MA)	0.0320	0.1420	0.2420	0.3720	0.4930	0.6950
G_d (Fisher, MA)	0.0280	0.1260	0.2210	0.3560	0.4910	0.6780

dent variation at each landmark (nonzero covariance) and also the extension to the two dependent sample case (paired data).

References

BOOKSTEIN, F.L. (1986): Size and Shape Spaces for Landmark Data in Two Dimensions. *Statistical Science 1, 181-242.*

BOOKSTEIN, F.L. (1991): *Morphometric Tools For Landmark Data: Geometry and Biology.* Cambridge University Press, Cambridge.

CELANT G., PESARIN F., SALMASO L. (2000): Two sample permutation tests for repeated measures with missing values. *Journal of Applied Statistical Science 9, 291-304.*

DRYDEN I.L., MARDIA, K.V. (1998): *Statistical Shape Analysis.* John Wiley & Sons, Chichester, New York.

GOOD, P. (1994): *Permutation tests.* Springer-Verlag, New York.

KENDALL, D.G. (1977): The diffusion of shapes. *Advances in Applied Probability 9, 428-430.*

LELE S., COLE T.M. (1996): A new test for shape differences when variance-covariance matrices are unequal. *The Journal of Human Evolution 31, 193-212.*

PESARIN F. (2001): *Multivariate Permutation tests: with application in Biostatistics.* John Wiley & Sons: Chichester-New York, 2001.

RAO C.R., SURYAWANSHI S. (1996): Statistical analysis of shape of objects based on landmark data. *Proceedings of the National Academy of Sciences of the United States of America 93, 12132-12136.*

RAO C.R., SURYAWANSHI S. (1998): Statistical analysis of shape through triangulation of landmarks: a study of sexual dimorphism in hominids. *Proceedings of the National Academy of Sciences of the United States of America 95, 4121-4125.*

ROHLF, F.J. (1999): On the use of shape space to compare morphometric methods. *Hystrix, Italian Journal of Mammalogy (n.s.), 11 (1), 9-25.*

ROHLF, F.J. (2000): Statistical Power Comparisons Among Alternative Morphometric Methods. *American Journal of Physical Anthropology 111, 463-478.*

SALMASO, L., SOLARI, A. (2005): Multiple aspect testing for case-control designs. *Metrika 62, 331-340.*

SLICE, D.E., BOOKSTEIN, F.L., MARCUS, L.F., ROHLF, F.J. (1996): A glossary for geometric morphometrics. *Advances in Morphometrics 284, 531-551.*

SLICE, D.E. (2005): *Modern Morphometrics In Physical Anthropology* Springer-Verlag New York, LLC.

Estimating Markov and Semi-Markov Switching Linear Mixed Models with Individual-Wise Random Effects

Florence Chaubert-Pereira[1], Yann Guédon[1], Christian Lavergne[2], and Catherine Trottier[2]

[1] CIRAD, UMR DAP & INRIA, Virtual Plants,
Avenue Agropolis, TA A-96/02, 34398 Montpellier, France,
chaubert@cirad.fr, guedon@cirad.fr
[2] UM3, UMR I3M,
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France,
Christian.Lavergne@math.univ-montp2.fr, trottier@math.univ-montp2.fr

Abstract. We address the estimation of Markov (and semi-Markov) switching linear mixed models i.e. models that combine linear mixed models with individual-wise random effects in a (semi-)Markovian manner. A MCEM-like algorithm whose iterations decompose into three steps (sampling of state sequences given random effects, prediction of random effects given the state sequence and maximization) is proposed. This statistical modeling approach is illustrated by the analysis of successive annual shoots along Corsican pine trunks.

Keywords: Markov switching model, linear mixed model, MCEM algorithm, plant growth

1 Introduction

Lindgren (1978) introduced Markov switching linear models, i.e hidden Markov models (Cappé et al. (2005)) with linear models as output process; see Frühwirth-Schnatter (2006) for an overview of Markov switching models. In the literature, hidden Markov models with random effects in the output process have been used in a limited way. Chaubert et al. (2007) applied to forest tree growth data Markov switching linear mixed models (MS-LMM), i.e models that combine linear mixed models in a Markovian manner. These models broaden the class of Markov switching linear models by incorporating individual-wise random effects in the output process. Altman (2007) introduced Markov switching generalized linear mixed models (MS-GLMM) where the output process is supposed to belong to the exponential family, and applied these models to brain lesion counts observed on multiple sclerosis patients. Since covariates and individual-wise random effects are incorporated in the output process, the generalization of MS-LMM to hidden semi-Markov model (Guédon (2007)) is straightforward. The resulting models are called semi-Markov switching linear mixed models (SMS-LMM).

The remainder of this paper is organized as follows. MS-LMM are formally defined in Section 2. A Monte Carlo EM-like (MCEM) algorithm (McLachlan and Krishnan (2008)) whose iterations decompose into three steps (sampling of state sequences given random effects, prediction of random effects given state sequence and maximization) is presented in Section 3. This statistical modeling approach is illustrated in Section 4 by the analysis of successive annual shoots along Corsican pine trunks using SMS-LMM. Section 5 consists of concluding remarks.

2 Markov switching linear mixed models (MS-LMM)

Let $\{S_t\}$ be a Markov chain with finite-state space $\{1, \ldots, J\}$. This J-state Markov chain is defined by the following parameters:

- initial probabilities $\pi_j = P(S_1 = j),\ j = 1, \ldots, J$, with $\sum_j \pi_j = 1$,
- transition probabilities $p_{ij} = P(S_t = j | S_{t-1} = i),\ i, j = 1, \ldots, J$, with $\sum_j p_{ij} = 1$.

Let Y_{at} be the observation and S_{at} the non-observable state for individual a, $a = 1, \ldots, N$, at time t, $t = 1, \ldots, T_a$. Let $\sum_{a=1}^{N} T_a = T$. We denote by $Y_{a1}^{T_a}$ the T_a-dimensional vector of observations for individual a, and by Y_1^T the T-dimensional vector of all the observations; i.e. the concatenation of $Y_{a1}^{T_a}$; $a = 1, \ldots, N$. The vectors of non-observable states, $S_{a1}^{T_a}$ and S_1^T, are defined analogously.

A **Markov switching linear mixed model** can be viewed as a pair of stochastic processes $\{S_{at}, Y_{at}\}$ where the output process $\{Y_{at}\}$ is related to the state process $\{S_{at}\}$, which is a finite-state Markov chain, by the following linear mixed model:

$$Y_{at}|_{S_{at}=s_{at}} = X_{at}\beta_{s_{at}} + \tau_{s_{at}}\xi_{as_{at}} + \epsilon_{at}, \tag{1}$$

$$\xi_{as_{at}} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=s_{at}} \sim \mathcal{N}(0, \sigma^2_{s_{at}}),$$

where X_{at} is the Q-dimensional row vector of covariates. Given the state $S_{at} = s_{at}$, $\beta_{s_{at}}$ is the Q-dimensional fixed effect parameter vector, $\xi_{as_{at}}$ is the individual a effect, $\tau_{s_{at}}$ is the standard deviation for the random effect and $\sigma^2_{s_{at}}$ is the residual variance. For convenience, random effects are supposed to follow the standard Gaussian distribution. Including random effects in the output process relaxes the assumption that the observations are conditionally independent given the non-observable states. The observations are here assumed to be conditionally independent given the non-observable states and the random effects.

3 Maximum likelihood estimation

Altman (2007) proposed to estimate the MS-GLMM parameters by maximizing directly the observed-data likelihood. Her approach based on Gaussian quadrature and quasi-Newton methods is strongly sensitive to starting values and to the number of quadrature points. Since both the states of the underlying Markov chain and the random effects are non observable, the EM algorithm (McLachlan and Krishnan (2008)) is a natural candidate to estimate MS-LMM. Let us consider the complete-data log-likelihood where both the outputs y_1^T, the random effects $\xi_1^J = \{\xi_{a1}^J = (\xi_{aj})_{j=1,\ldots,J}; a = 1,\ldots,N\}$ and the states s_1^T of the underlying Markov chain are observed

$$\begin{aligned}\log f(y_1^T, s_1^T, \xi_1^J; \theta) &= \log f(s_1^T) + \log f(\xi_1^J) + \log f(y_1^T | s_1^T, \xi_1^J) \\ &= \sum_{a=1}^{N} \log \pi_{s_{a1}} + \sum_{a=1}^{N}\sum_{t=2}^{T_a} \log p_{s_{a,t-1},s_{a,t}} + \sum_{a=1}^{N}\sum_{j=1}^{J} \log \phi(\xi_{aj}; 0, 1) \\ &+ \sum_{a=1}^{N}\sum_{t=1}^{T_a} \log \phi(y_{at}; X_{at}\beta_{s_{at}} + \tau_{s_{at}}\xi_{as_{at}}, \sigma^2_{s_{at}}). \end{aligned} \quad (2)$$

where $\theta = (\pi, P, \beta, \tau, \sigma^2)$ is the set of parameters to be estimated and $\phi(y; \mu, \sigma^2)$ is the density of the Gaussian distribution with mean μ and variance σ^2.

The EM algorithm for hidden Markov chains cannot be transposed because the observations are not conditionally independent given the non-observable states; see Section 2. The EM algorithm for finite mixture of linear mixed models (Celeux et al. (2005)) cannot be adapted because the distribution of $\xi_1^J | Y_1^T = y_1^T$ cannot be analytically derived. Altman (2007) proposed a MCEM algorithm to estimate MS-GLMM where the random effects are sampled by Monte Carlo methods like Gibbs sampling. In the M-step, numerical methods like quasi-Newton routines are necessary to obtain updates for the parameter estimates. Altman (2007) noted the prohibitive computation burden due to the Monte Carlo and quasi-Newton methods, the slowness to converge and the sensitivity to starting values. Since sampling both a state sequence and random effects $\{s_1^T, \xi_1^J\}$ from their conditional distribution $S_1^T, \xi_1^J | Y_1^T = y_1^T$ is rather complicated, we propose here a MCEM-like algorithm where the Monte Carlo E-step is decomposed into two conditional steps:

- Monte Carlo Conditional E-step : given the random effects, state sequences are sampled for each individual a using a "forward-backward" algorithm (Chib (1996)).
- Conditional E-step : given the state sequence, the random effects are predicted.

In the M-step, the quantities $\sum_a \sum_t \log \phi(y_{at}; X_{at}\beta_{s_{at}} + \tau_{s_{at}}\xi_{as_{at}}, \sigma^2_{s_{at}})$, $\sum_a \log \pi_{s_{a1}}$ and $\sum_a \sum_{t=2}^{T_a} \log p_{s_{a,t-1},s_{a,t}}$ in Equation (2) can be maximized separately.

3.1 Forward-backward algorithm for sampling state sequences given the random effects

For each individual a, the state sequences are sampled from the conditional distribution $P(S_{a1}^{T_a} = s_{a1}^{T_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J)$.
For a Markov switching linear mixed model, since

$$P\big(S_{a1}^{T_a} = s_{a1}^{T_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big) = \Big\{ \prod_{t=1}^{T_a - 1} P\big(S_{at} = s_{at} | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big) \Big\} \times P\big(S_{aT_a} = s_{aT_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big),$$

the following conditional distributions should be used for sampling state sequences:

- final state (initialization) $P\big(S_{aT_a} = s_{aT_a} | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big)$,
- previous state $P\big(S_{at} = s_{at} | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big)$.

The forward-backward algorithm for sampling state sequences given the random effects can be decomposed into two passes, a forward recursion which is similar to the forward recursion of the forward-backward algorithm for hidden Markov chains, and a backward pass for sampling state sequences.

Forward recursion
The forward recursion is initialized for $t = 1$ by:

$$F_{aj}(1) = P\big(S_{a1} = j | Y_{a1} = y_{a1}, \xi_{a1}^J\big) = \frac{\phi(y_{a1}; X_{a1}\beta_j + \tau_j \xi_{aj}, \sigma_j^2)\pi_j}{N_{a1}} = \frac{G_{aj}(1)}{N_{a1}},$$

where $N_{a1} = P(Y_{a1} = y_{a1} | \xi_{a1}^J) = \sum_{j=1}^J G_{aj}(1)$ is a normalizing factor.
For $t = 2, \ldots, T_a$, the forward recursion is given by:

$$F_{aj}(t) = P\big(S_{at} = j | Y_{a1}^t = y_{a1}^t, \xi_{a1}^J\big) = \frac{\phi(y_{at}; X_{at}\beta_j + \tau_j \xi_{aj}, \sigma_j^2) \sum_{i=1}^J p_{ij} F_{ai}(t-1)}{N_{at}} = \frac{G_{aj}(t)}{N_{at}}.$$

The normalizing factor $N_{at} = P\big(Y_{at} = y_{at} | Y_{a1}^{t-1} = y_{a1}^{t-1}, \xi_{a1}^J\big) = \sum_{j=1}^J G_{aj}(t)$ is obtained directly during the forward recursion. The forward recursion can be used to compute the observed-data log-likelihood given the random effects for the parameter θ, as $\log P(Y_1^T = y_1^T | \xi_1^J; \theta) = \sum_a \sum_t \log N_{at}$.

Backward pass
The backward pass can be seen as a stochastic backtracking procedure. The final state s_{aT_a} is drawn from the smoothed probabilities

$$\Big(P\big(S_{aT_a} = j | Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big) = F_{aj}(T_a); j = 1, \ldots, J \Big).$$

For $t = T_a - 1, \ldots, 1$, the state s_{at} is drawn from the conditional distribution

$$\Big(P\big(S_{at} = j | S_{a,t+1}^{T_a} = s_{a,t+1}^{T_a}, Y_{a1}^{T_a} = y_{a1}^{T_a}, \xi_{a1}^J\big) = \frac{p_{j s_{a,t+1}} F_{aj}(t)}{\sum_{i=1}^J p_{i s_{a,t+1}} F_{ai}(t)}; j = 1, \ldots, J \Big).$$

3.2 Random effect prediction given the state sequence

The predicted random effects ξ_{a1}^J for each individual a is:

$$\xi_{a1}^J = \mathrm{E}[\xi_{a1}^J | y_{a1}^{T_a}] = \mathrm{E}\Big[\mathrm{E}[\xi_{a1}^J | y_{a1}^{T_a}, s_{a1}^{T_a}] | y_{a1}^{T_a}\Big] \approx \frac{1}{M}\sum_{m=1}^{M} \mathrm{E}[\xi_{a1}^J | y_{a1}^{T_a}, s_{a1}^{T_a}(m)], \quad (3)$$

with,

$$\mathrm{E}[\xi_{a1}^J | y_{a1}^{T_a}, s_{a1}^{T_a}(m)] = \Omega U_a^{(m)\prime}\Big(U_a^{(m)}\Omega^2 U_a^{(m)\prime} + \mathrm{Diag}\{U_a^{(m)}\sigma^2\}\Big)^{-1}\Big(Y_{a1}^{T_a} - \sum_{j=1}^{J} I_{aj}(m) X_a \beta_j\Big),$$

where:

- $s_{a1}^{T_a}(m)$ is the mth state sequence sampled for individual a,
- $\Omega = \mathrm{Diag}\{\tau_j; j = 1, \ldots, J\}$ is the $J \times J$ random standard deviation matrix,
- $U_a^{(m)}$ is the $T_a \times J$ design matrix associated with state sequence $s_{a1}^{T_a}(m)$, composed of 1 and 0 with $\sum_j U_a^{(m)}(t,j) = 1$ and $\sum_t \sum_j U_a^{(m)}(t,j) = T_a$,
- $\mathrm{Diag}\{U_a^{(m)}\sigma^2\}$ is the $T_a \times T_a$ diagonal matrix with $\{u_{at}^{(m)}\sigma^2; t = 1, \ldots, T_a\}$ on its diagonal,
- $u_{at}^{(m)} = \Big(\mathrm{I}(s_{at}(m) = 1) \cdots \mathrm{I}(s_{at}(m) = J)\Big)$ is the tth row of the design matrix $U_a^{(m)}$, I() is the indicator function,
- $\sigma^2 = (\sigma_1^2 \cdots \sigma_J^2)'$ is the J-dimensional residual variance vector,
- $I_{aj}(m) = \mathrm{Diag}\{I(s_{at}(m) = j); t = 1, \ldots, T_a\}$ is a $T_a \times T_a$ diagonal matrix,
- X_a is the $T_a \times Q$ matrix of covariates.

3.3 Extension to hidden semi-Markov models

Semi-Markov chains generalize Markov chains with the distinctive property of explicitly modeling the sojourn time in each state. Let $\{S_t\}$ be a semi-Markov chain defined by the following parameters:

- initial probabilities $\pi_j = P(S_1 = j)$, with $\sum_j \pi_j = 1$,
- transition probabilities
  - nonabsorbing state i: for each $j \neq i$, $\tilde{p}_{ij} = P(S_t = j | S_t \neq i, S_{t-1} = i)$, with $\sum_{j \neq i} \tilde{p}_{ij} = 1$ and $\tilde{p}_{ii} = 0$,
  - absorbing state i: $p_{ii} = P(S_t = i | S_{t-1} = i) = 1$ and for each $j \neq i$, $p_{ij} = 0$.

An occupancy distribution is attached to each nonabsorbing states:

$$d_j(u) = P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-2 | S_{t+1} = j, S_t \neq j), u = 1, 2, \ldots$$

As for the MS-LMM, the output process $\{Y_{at}\}$ of the semi-Markov switching linear mixed model (SMS-LMM) for individual a is related to the underlying semi-Markov chain $\{S_{at}\}$ by the linear mixed model (1). Since covariates and individual-wise random effects are incorporated in the output process,

the observations are assumed to be conditionally independent given the non-observable states and the random effects. The proposed MCEM-like algorithm can therefore be directly transposed to SMS-LMM. Given the random effects, the state sequences are sampled using the "forward-backward" algorithm proposed by Guédon (2007). Given a state sequence, the random effects are predicted as previously described. The underlying semi-Markov chain parameters and the linear mixed model parameters are obtained by maximizing the Monte Carlo approximation of the complete-data log-likelihood.

4 Application to forest trees

The use of SMS-LMM is illustrated by the analysis of forest tree growth. The data set comprised four sub-samples of Corsican pines: 31 6-year-old trees, 29 12-year-old trees, 30 18-year-old trees and 13 23-year-old trees. Tree trunks were described by annual shoot from the base to the top where the length (in cm) was recorded for each annual shoot. The annual shoot is defined as the segment of stem established within a year. The observed growth is mainly the result of the modulation of the endogenous growth component by climatic factors. The endogenous growth component is assumed to be structured as a succession of roughly stationary phases separated by marked change points (Guédon et al.(2007)). The length of successive annual shoots along tree trunks was previously analyzed using a hidden semi-Markov chain (Guédon et al.(2007)) and a MS-LMM (Chaubert et al. (2007)). In the first case, the influence of climatic factors and the inter-individual heterogeneity were not explicitly modeled while in the second case, the length of the successive growth phases was not explicitly modeled.

A "left-right" three-state SMS-LMM composed of two successive transient states followed by a final absorbing state was estimated. We chose to use an intercept and the centered cumulated rainfall during a period recovering one organogenesis period and one elongation period as fixed effects for each linear mixed model. The linear mixed model attached to the growth phase j is:

$$Y_{at}|_{S_{at}=j} = \beta_{j1}+\beta_{j2}X_t+\tau_j\xi_{aj}+\epsilon_{at}, \qquad \xi_{aj} \sim \mathcal{N}(0,1), \qquad \epsilon_{at}|_{S_{at}=j} \sim \mathcal{N}(0,\sigma_j^2),$$

where Y_{at} is the length of the annual shoot for individual a at time t, β_{j1} is the intercept, X_t is the centered cumulated rainfall at time t ($\mathrm{E}(X_t) = 0$) and β_{j2} is the cumulated rainfall parameter. As the cumulated rainfall is centered, the intercept represents the average length of successive annual shoots in each growth phase. The estimation algorithm was initialized with the parameter values π, P, β and σ^2 estimated without taking into account random effects (hence, $\xi_1^J = 0$). The algorithm converged in 62 iterations with $m = 100$ state sequences sampled for each tree at each iteration. The convergence of the algorithm was monitored using the log-likelihood of the observed sequences given the random effects, which is directly obtained as a byproduct of the forward recursion; see Section 3.1.

The marginal distribution of the linear mixed model attached to growth phase j is $\mathcal{N}(\mu_j, \Gamma_j^2)$ with $\mu_j = \beta_{j1} + \beta_{j2}\mathrm{E}_j(X)$ and $\Gamma_j^2 = \tau_j^2 + \sigma_j^2$ where $\mathrm{E}_j(X)$ is the mean of the cumulated rainfalls X in growth phase j. The marginal distributions of the linear mixed models attached to each growth phase are well separated (few overlapping between marginal distributions corresponding to two successive states); compare the mean difference $\mu_{j+1} - \mu_j$ between consecutive states and the standard deviations Γ_j and Γ_{j+1} in Table 1. The standard deviation of the cumulated rainfall effect was computed as $\beta_{j2} \times sd(X)$ for each state j where $sd(X)$ is the standard deviation of the cumulated rainfalls X. The standard deviation of the cumulated rainfall effect represents the average amplitude of the climatic fluctuations in each growth phase. The influence of the cumulated rainfall is weak in the first growth phase (of slowest growth) while it is strong in the last two growth phases (a little less in the second phase than in the third phase); see Table 1.

	State		
	1	2	3
Intercept β_{j1}	7.19	26.08	50.48
Cumulated rainfall parameter β_{j2}	0.0042	0.0171	0.0304
Cumulated rainfall effect standard deviation	0.56	2.23	3.97
Random variance τ_j^2	6.81	52.34	72.83
Residual variance σ_j^2	5.13	39.75	76.54
Part of inter-individual heterogeneity	57.04%	56.84%	48.76%
Marginal distribution (μ_j, Γ_j)	$7.05, 3.46$	$26.17, 9.60$	$50.55, 12.22$

Table 1. Intercepts, regression parameters, centered cumulated rainfall effect, variability decomposition and marginal distributions of the estimated SMS-LMM.

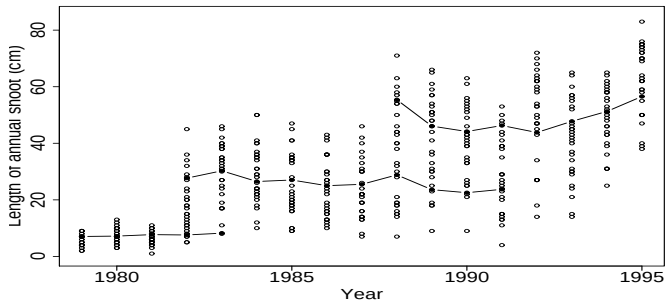


Fig. 1. 18-year-old Corsican pines: observed annual shoot lengths (points) and fixed part of the three observation linear mixed models (point lines).

The part of inter-individual heterogeneity, defined by the ratio between the random variance τ_j^2 and the total variance Γ_j^2 , is greater at the beginning of the plant life (first two growth phases with more than 56%) and decreases slightly in the last growth phase (near 49%). The most probable state sequence given the predicted random effects was computed for each observed sequence using a Viterbi-like algorithm. The fixed part of the three

observation linear mixed models (i.e. $\beta_{j1} + \beta_{j2} X_t$ for each growth phase j) for 18-year-old trees is represented in Figure 1. The growth phases are well separated with few overlapping.

5 Concluding remarks

SMS-LMM enables to separate and to characterize the different growth components (endogenous, environmental and individual components) of forest trees. The behavior of each tree within the population can be investigated on the basis of the random effects predicted for each growth phase.
An interesting direction for further research would be to develop the statistical methodology for semi-Markov switching generalized linear mixed models. Since the hidden semi-Markov chain likelihood cannot be written as a simple product of matrices, the MCEM algorithm proposed by Altman (2007) for the MS-GLMM cannot be directly extended to the semi-Markovian case. In our MCEM-like algorithm proposed for MS-LMM and SMS-LMM, the difficulty lies mainly in the prediction of the random effects.

References

ALTMAN, R.M. (2007) : Mixed hidden Markov models : An extension of the hidden Markov model to the longitudinal data setting. *J. Amer. Stat. Assoc., 102, 201-210.*

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005): *Inference in hidden Markov models.* Springer Series in Statistics. New York, NY: Springer. xvii, 652 p.

CELEUX, G., MARTIN, O. and LAVERGNE, C. (2005): Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Stat. Model., 5, 243-267.*

CHAUBERT, F., CARAGLIO, Y., LAVERGNE, C., TROTTIER, C. and GUÉDON, Y. (2007): A statistical model for analyzing jointly growth phases, the influence of environmental factors and inter-individual heterogeneity. Applications to forest trees. *Proceedings of the 5th International Workshop on Functional-Structural Plant Models, P43, 1-3.*

CHIB, S. (1996): Calculating posterior distributions and modal estimates in Markov mixture models. *J. Econometrics, 75, 79-97.*

FRÜHWIRTH-SCHNATTER, S. (2006): *Finite Mixture and Markov Switching Models.* Springer Series in Statistics. New York, NY: Springer. xix, 492 p.

GUÉDON, Y. (2007): Exploring the state sequence space for hidden Markov and semi-Markov chains. *Comput. Stat. and Data Anal., 51 (5), 2379-2409.*

GUÉDON, Y., CARAGLIO, Y., HEURET, P., LEBARBIER, E. and MEREDIEU, C. (2007): Analyzing growth components in trees. *J. of Theor. Biology, 248 (3), 418-447.*

LINDGREN, G. (1978): Markov regime models for mixed distributions and switching regressions. *Scand. J. Stat., Theory Appl., 5, 81-91.*

McLACHLAN, G.J. and KRISHNAN, T. (2008): *The EM algorithm and extensions. 2nd Edition.* Wiley Series in Probability and Statistics. New York, NY: John Wiley and Sons. xviii, 360 p.

Analysis of Association Between Genotype with Diplotype Configuration and Phenotype of Multiple Quantitative Responses

Noboru Hashimoto[1], Makoto Tomita[2], and Yutaka Tanaka[3]

[1] Department of Information Systems and Mathematical Sciences, Graduate School of Nanzan University, Japan, *m06mm003@nanzan-u.ac.jp*
[2] Corresponding Author: Department of Information Systems and Mathematical Sciences, Nanzan University, Japan, *tomita@nanzan-u.ac.jp*
[3] Department of Information Systems and Mathematical Sciences, Nanzan University, Japan, *ytanaka@nanzan-u.ac.jp*

Abstract. Recently the association analysis has been actively studied between the genotype information and the phenotype variables which are associated with a specific disease. For the case of one quantitative phenotype variable, an algorithm called *QTLhaplo* has been proposed. It is a general method of association analysis which can deal with dominant, resessive and additive models for the genotype-to-phenotype relationship. We consider a multivariate method of genome wide association studies (GWAs) and develop a program in R language/environment, which is an extension of *QTLhaplo* to the case of multivariate quantitative variables. An artificial data set, which is generated using normal random number generator in R, is analyzed to show the performance of the proposed method. The false discovery rate (FDR) is chosen to control the false positive error.

Keywords: multivariate analysis, quantitative responses, haplotype, likelihood ratio test

1 Introduction

A major goal of current human genome-wide studies is to identify the genetic basis of complex disorders. Haplotype-based methods offer powerful approaches to disease gene mapping, based on the association between causal mutations and the ancestral haplotypes on which these mutations arose. Variation in the human genomic sequence plays a powerful but poorly understood role in the etiology of common medical conditions. For linkage disequilibrium analysis, there are many methods for identifying LD blocks. (Tomita M. *et al*., 2008; etc..)

Recently the association has been studied between genotype and phenotype. Here 'genotype' means not only genotype itself but also haplotype and diplotype configurations that are estimated from genotype. In contrast, 'Phenotype' is a qualitative or quantitative variable which may be related

to a specific disease, as quantitative variable, called QTL (quantitative trait locus), it includes covaraites such as BMI, glucose level or others.

Some algorithms have been proposed so far to analyze the association between the genotype information and a quantitative phenotypic QTL. The algorithm *QTLhaplo* (Shibata *et al*., 2004) deals with the association between the genotype and univariate phenotype, assuming the normality of the conditional distribution of the phenotype given the genotype information. The likelihood is calculated on the basis of the frequencies of diplotype configurations (joint probability of haplotypes frequencies that compose the dipolotype) and the density function of a normal distribution. The algorithm *QTLmarc* (Kamitsuji and Kamatani, 2006) has been proposed for multivariate analysis of multiple quantitative responses, however, it can deal with only the case where the diplotype configuration is determined uniquely from the genotype. Therefore, it will be valuable to develop a general method of association analysis for multivariate quantitative responses. In the present paper we extend the algorithm *QTLhaplo* such that the association between the genotype and multivariate quantitative variables can be analyzed assuming the dominant, the resessive and the additive model.

2 Method

2.1 Univariate models

Shibata *et al*., (2004) describe the algorithm *QTLhaplo* as follows. Suppose that there are $\mathfrak{l}$ linked loci. As DNA is of double helix structure and each haplotype has a counterpart, the number of possible haplotypes is $\mathfrak{L} = 2^{\mathfrak{l}}$ in total. Let the relative frequencies of the haplotypes be given as $\Theta = (\theta_1, \cdots, \theta_j, \cdots, \theta_{\mathfrak{L}})$, where θ_j is the relative frequency of the jth haplotype, and $\theta_j \geq 0, \sum_{j=1}^{\mathfrak{L}} \theta_j = 1$. Each subject has a combination of two haplotypes sampled randomly from the multinominal distribution with parameters Θ, and therefore there are $\mathfrak{L}^2$ possible combinations $a_1, a_2, \cdots, a_{\mathfrak{L}^2}$. The probability that the ith subject has a diplotype configuration a_k of the $\mathfrak{l}$th and the mth haplotypes is given by $P(d_i = a_k|\Theta) = \theta_{\mathfrak{l}}\theta_m$, where d_i is a diplotype configuration for the ith subject. Also suppose that the ith subject has quantitative phenotype ψ_i with a probability density function f. Let us assume that sample of size N has been observed in an experiment. The phenotype for each diplotype configuration is assumed to follow a normal distribution with a common variance but with a mean which depends on the diplotype configuration. The outcome of the experiment can then be expressed as (Θ, D, Ψ), where $D = (d_1, \cdots, d_N)$ indicates the vector of the diplotype configuration and $\Psi = (\psi_1, \cdots, \psi_N)$ indicates the matrix of the quantitative phenotypes. Assume next, the data set is divided into the two groups of subjects with and without a specified haplotype h_p in the diplotype configurations, and the group means μ_k and the common variance σ^2 are estimated, respectively.

The problem is to test whether there exists any difference in the distribution of the phenotype between the two groups. Let D_+ denote the set of the diplotype configurations containing the haplotype h_p, and D_- denote the set if not, when considering a dominant model.

From the above assumptions we obtain that, for a diplotype $d_i \in D_+$, the distribution of phenotype is given by $N(\mu_1, \sigma^2)$ and for $d_i \in D_-$ it is given by $N(\mu_2, \Sigma)$. Denote the probability density functions by $f_{\mu_j}(x)$, $j = 1, 2$. Then the probability density function for ψ_i is defined by $f_{\mu_1}(\mathbf{x}) = f(\psi_i = \mathbf{x}|d_i \in D_+), \text{if} d_i \in D_+$ and $f_{\mu_2}(\mathbf{x}) = f(\psi_i = \mathbf{x}|d_i \in D_-), \text{if} d_i \in D_-$.

Let A denote a haplotype with a specified h_p and B denote a haplotype without h_p. Then every diplotype configuration is expressed as AA, AB, or BB, and sets D_+ and D_- can be defined. In the case of dominant models, AA and AB belong to D_+, while BB belongs to D_-. In the case of recessive models, AA belongs to D_+, while AB and BB belong to D_-. For additive models, the distributions of ψ_i for AA, BB and AB are given by $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ and $N(\mu_3, \sigma^2)$, respectively, where $\mu_3 = (\mu_1 + \mu_2)/2$.

2.2 Extension to multivariate models

We try to extend the above univariate model to a multivariate model. Suppose that the quantitative phenotype vector Ψ_i follows a multidimensional normal distribution with a common variance-covariance matrix but with different mean vectors corresponding to the groups defined by the diplotype configurations. In dominant/recessive models, the density function is given by,

$$f(\Psi_i = \mathbf{x}|d_i = a_k, \mu, \Sigma) = \begin{cases} (\frac{1}{(2\pi)^{p/2}|\Sigma|^{-\frac{1}{2}}}) e^{-\frac{1}{2}(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1)} & \text{if } a_k \in D_+, \\ (\frac{1}{(2\pi)^{p/2}|\Sigma|^{-\frac{1}{2}}}) e^{-\frac{1}{2}(\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2)} & \text{if } a_k \notin D_+, \end{cases} \quad (1)$$

whereas in an additive model it is given by,

$$f(\Psi_i = \mathbf{x}|d_i = a_k, \mu, \Sigma) = \begin{cases} (\frac{1}{(2\pi)^{p/2}|\Sigma|^{-\frac{1}{2}}}) e^{-\frac{1}{2}(\mathbf{x}-\mu_1)'\Sigma^{-1}(\mathbf{x}-\mu_1)} & \text{if } a_k \in D_{AA}, \\ (\frac{1}{(2\pi)^{p/2}|\Sigma|^{-\frac{1}{2}}}) e^{-\frac{1}{2}(\mathbf{x}-\mu_2)'\Sigma^{-1}(\mathbf{x}-\mu_2)} & \text{if } a_k \in D_{BB}, \\ (\frac{1}{(2\pi)^{p/2}|\Sigma|^{-\frac{1}{2}}}) e^{-\frac{1}{2}(\mathbf{x}-\frac{\mu_1+\mu_2}{2})'\Sigma^{-1}(\mathbf{x}-\frac{\mu_1+\mu_2}{2})} & \\ \text{if } a_k \in D_{AB}, & \end{cases} \quad (2)$$

where μ, Σ indicate the mean vector and the variance-covariance matrix, respectively, $\mathbf{x}$ is the vector of individual quantitative phenotypes, and p is the number of phenotype variables.

2.3 Likelihood function

The observed data consist of the genotype and quantitative phenotype of N subjects. Let $G_{obs} = (g_1, g_2, \cdots, g_N)$ and $\Psi_{obs} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_N)$ be

the vectors of the observed genotypes and the matrix of the quantitative phenotypes, respectively. Then the likelihood function is given by

$$L(\Theta, \mu, \Sigma) \propto \prod_{i=1}^{N} \sum_{a_k \in A_i} P(d_i{=}a_k|\Theta) f(\psi_i = \mathbf{w}_i | d_i = a_k, \mu, \Sigma),$$

where A_i denotes the set of diplotype configurations a_k, which is consistent with g_i, and f is the probability density function for $N(\mu, \Sigma)$, where μ depends on a_k. Under the null hypothesis it is assumed that the distribution of the phenotype does not depend on the haplotype, i.e., the mean vector μ is equal to a common vector μ_0. Under the alternative hypothesis, two multidimensional normal distributions, $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, are defined, if the model is dominant or recessive, and three multidimensional normal distributions, $N(\mu_1, \Sigma)$, $N(\mu_2, \Sigma)$, and $N((\mu_1 + \mu_2)/2, \Sigma)$, are defined, if the model is additive. The phenotype $\mathbf{x}$ of the i th subject follows a distribution given by equations (1) or (2), depending on the assumed model.

2.4 Estimation of parameters, likelihood ratio test

If the complete data of $d_1, d_2, \cdots, d_N$ and $\Psi_1, \Psi_2, \cdots, \Psi_N$ were available, the maximum likelihood estimators for μ, Σ and the haplotype frequencies $\Theta = (\theta_1, \theta_2, \cdots, \theta_L)$ would be obtained as

$$\hat{\theta_j} = n_j/(2N) \ (j = 1, 2, \cdots, L),$$
$$\hat{\mu_1} = \sum_{d_i \in D_+} \psi_i / N_+, \hat{\mu_2} = \sum_{d_i \notin D_+} \psi_i / N_-,$$
$$\hat{\Sigma} = \left[\sum_{d_i \in D_+} (\Psi_i - \mu_1)(\Psi_i - \mu_1)' + \sum_{d_i \notin D_+} (\Psi_i - \mu_2)(\Psi_i - \mu_2)' \right] / N,$$

where n_j counts how often the j^{th} haplotype appears among the N subjects, and N_+ or N_- denote the numbers of subjects who possess or do not possess haplotypes h_p, respectively. However, actually the complete data are not available and we can observe only genotypes and phenotypes of the subjects.

In a subject, two or more haplotypes might correspond to one genotype. Therefore, there are two or more candidates for the diplotype configuration of the subject, given the haplotype. Several procedures have been developed for estimating haplotype frequencies Θ using of *EM-algorithm*, *MCMC*, or others. However, as the purpose of our study is somewhat different, we omit here a detailed explanation of those algorithms.

Here we assume that the vector of frequencies of the haplotypes have been estimated using some appropriate software. Their results can be used to estimate the mean vectors and the variance-covariance matrix in our multivariate models.

In the case of the dominant and recessive models the log-likelihood function is expressed as $l = \log L(\Theta, \mu, \Sigma)$ and the maximum likelihood estimators for μ_1 and μ_2 are obtained by

$$\hat{\mu_1} = \frac{\sum_{i=1}^{N} \psi_i(u_b/u_0)}{\sum_{i=1}^{N}(u_b/u_0)} \qquad \text{and} \qquad \hat{\mu_2} = \frac{\sum_{i=1}^{N} \psi_i(v_b/v_0)}{\sum_{i=1}^{N}(v_b/v_0)},$$

where

$$\begin{aligned}
u_b &= \sum_{a_k \in D_+ \cap A_i} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu_1, \sigma), \\
u_0 &= \sum_{a_k \in A_i} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu, \sigma), \\
v_b &= \sum_{a_k \in A_i \cap D-} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu_2, \Sigma), \\
v_0 &= \sum_{a_k \in A_i} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu, \Sigma).
\end{aligned}$$

The variance-covariance matrix is estimated as

$$\hat{\Sigma} = \frac{1}{n}\Big[\sum_{i=1}^{N}(\psi_i - \mu_1)(\psi_i - \mu_1)^T(u_b/u_0) + \sum_{i=1}^{N}(\psi_i - \mu_2)(\psi_i - \mu_2)^T(v_b/v_0)\Big],$$

where n is $\sum_{i=1}^{N}(u_b/u_0) + \sum_{i=1}^{N}(v_b/v_0)$. D_+ is a set of the diplotype configurations which contain haplotype h_p in such a way that it is consistent with dominant/recessive models. In the case of additive models, the mean vectors and the variance-covariance matrix are estimated by solving the following equations.

$$\begin{aligned}
\Big(\textstyle\sum_{i=1}^{N} \frac{u_b}{u_0} + \frac{1}{4}\sum_{i=1}^{N} \frac{w_b}{w_0}\Big)\mu_1 + \frac{1}{4}\sum_{i=1}^{N} \frac{w_b}{w_0}\mu_2 &= \sum_{i=1}^{N}\Big(\frac{u_b}{u_0} + \frac{1}{2}(\frac{u_0}{u_b})\Big)\mathbf{x}_i \\
\frac{1}{4}\sum_{i=1}^{N} \frac{w_b}{w_0}\mu_1 + \Big(\textstyle\sum_{i=1}^{N} \frac{v_b}{v_0} + \frac{1}{4}\sum_{i=1}^{N} \frac{w_b}{w_0}\Big)\mu_2 &= \sum_{i=1}^{N}\Big(\frac{v_b}{v_0} + \frac{1}{2}(\frac{u_0}{u_b})\Big)\mathbf{x}_i
\end{aligned}$$

$$\begin{aligned}
\hat{\Sigma} = \frac{1}{n}\Big[&\sum_{i=1}^{N}(\psi_i - \mu_1)(\psi_i - \mu_1)^T(u_b/u_0) + \sum_{i=1}^{N}(\psi_i - \mu_2)(\psi_i - \mu_2)^T(v_b/v_0) \\
&+ \sum_{i=1}^{N}(\psi_i - (\mu_1 + \mu_2)/2)(\psi_i - (\mu_1 + \mu_2)/2)^T(w_b/w_0)\Big]
\end{aligned}$$

In the above equations, u's, v's and w's are defined by

$$u_b = \sum_{a_k \in A_i \cap AA} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu_2, \Sigma),$$
$$u_0 = \sum_{a_k \in A_i} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu, \Sigma),$$
$$v_b = \sum_{a_k \in A_i \cap BB} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu_2, \Sigma),$$
$$v_0 = \sum_{a_k \in A_i} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu, \Sigma),$$
$$w_b = \sum_{a_k \in A_i \cap AB} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu_2, \Sigma) \text{ and}$$
$$w_0 = \sum_{a_k \in A_i} P(d_i = a_k|\Theta) f(\psi_i|d_i = a_k, \mu, \Sigma).$$

The above two systems of equations can be solved by using iterative procedures with appropriate initial values, When converging, the solution provides maximum likelihood estimates.

Likelihood ratio tests can be applied to the association analysis between the haplotypes and the phenotypes. Let L_{0max} and L_{max} be the likelihood functions under the null and alternative hypotheses, respectively. It is known that under the null hypothesis the log-likelihood ratio $-2\log(L_{0max}/L_{max})$ asymptotically follows a χ^2 distribution. The degree of freedom is given by the difference of the distributions corresponding to the both hypotheses.

Table 1. Artificial data set of diplotype configurations and phenotype variables.

No	probability	pop.freq.	haplotype1	haplotype2	No	pheno1	pheno2
1	1	0.0625	112	112	1	142.3014	138.2740
2	1	0.0625	112	112	2	136.8866	122.6957
3	1	0.0625	112	112	3	138.7330	123.5078
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
99	1	0.0260	221	121	99	138.9166	122.1911
100	1	0.0260	221	121	100	139.5149	136.6850

3 Numerical example

Assume that there are three loci with two kinds of alleles as genotypes and two quantitative phenotype variables. Let us consider the seven haplotypes $\{112, 111, 121, 122, 212, 221, 222\}$, and let the haplotype 112 be considered as

Table 2. Result of bivariate analysis of phenotype1 and phenotype2.

haplotype	hap.freq.	Xsquared	Pvalue	qvalue	df
112	0.200	9.2527185	0.009790339	0.007142857	2
122	0.040	2.0111034	0.365842738	0.014285714	2
222	0.025	1.7200912	0.423142785	0.021428571	2
221	0.050	1.1772382	0.555093296	0.028571429	2
212	0.065	0.5182846	0.771713198	0.035714286	2
121	0.500	0.4878873	0.783531774	0.042857143	2
111	0.120	0.2182499	0.896618389	0.050000000	2

Table 3. Result of univariate analysis of phenotype1.

haplotype	hap.freq.	Xsquared	Pvalue	qvalue	df
112	0.200	1.93542878	0.1641657	0.007142857	1
222	0.025	1.70928751	0.1910778	0.014285714	1
122	0.040	1.09585964	0.2951765	0.021428571	1
221	0.050	0.21890184	0.6398779	0.028571429	1
111	0.120	0.20923284	0.6473694	0.035714286	1
121	0.500	0.09174249	0.7619735	0.042857143	1
212	0.065	0.00801758	0.9286520	0.050000000	1

Table 4. Result of univariate analysis of phenotype2.

haplotype	hap.freq.	Xsquared	Pvalue	qvalue	df
112	0.200	1.74675259	0.1862855	0.007142857	1
222	0.025	0.68852443	0.4066667	0.014285714	1
212	0.065	0.40496132	0.5245381	0.021428571	1
221	0.050	0.28021100	0.5965630	0.028571429	1
121	0.500	0.11617108	0.7332249	0.035714286	1
111	0.120	0.11546249	0.7340089	0.042857143	1
122	0.040	0.03062968	0.8610691	0.050000000	1

the target one. Also assume a dominant model in which the quantitative variables of the subjects with or without target haplotype 112 follow a bivariable normal distribution $N(\mu_{\mathbf{1}}, \Sigma)$ or $N(\mu_{\mathbf{2}}, \Sigma)$, where $\mu_{\mathbf{1}} = (140, 128.5)$, $\mu_{\mathbf{2}} = (138.5, 130)$ and $\Sigma = (\sigma_{11} = 5, \sigma_{12} = \sigma_{21} = 4.2, \sigma_{22} = 6)$. We have generated two samples with sizes 30 and 70 from $N(\mu_{\mathbf{1}}, \Sigma)$ and $N(\mu_{\mathbf{2}}, \Sigma)$ using a normal random number generating function in R. We fixed first 30 subjects to have the haplotype 112 and 70 subjects to have other haplotype. A part of the data set is given in Table 1. Then we applied the association analysis to this data set. The result of the bivariate analysis and those of univariate analysis are given in Table 2 and Table 3, respectively. The q-value is defined by B-H method (Benjamini and Hochberg, 2000). It controls the

false discovery rate (FDR), fore this multiple hypothesis testing problem. It is shown in Table 2 that the p-value for haplotype 112 is equal to is 0.0098, while the p-values for other cases are all larger than 0.36. With univariate analysis of individual variables 1 or 2 the p-values for haplotype 112 are 0.16 and 0.19, respectively as shown in Table 3. It is noted that the p-values are much smaller in the bivariate analysis than in univariate analysis.

4 Concluding remarks

In the present paper we proposed a method of multivariate association analysis and showed through a numerical example how the relationship was detected more clearly between the genotype information and the phenotype quantitative variables in the multivariate analysis compared to then univariate analysis based on the algorithm of *QTLhaplo*. Our method is an extension of *QTLhaplo*, where dominant, recessive and additive models can be applied. Although in our numerical example the diplotype configuration is determined uniquely from the genotype, our algorithm can be applied to the cases where the diplotype configuration is not determined uniquely from the genotype. It is a merit of our method compared to the *QTLmarc* (Kamitsuji *et al*., 2006) that can treat such cases whereas the latter can not. It is expected that our method will be useful evaluating association studies of complex diseases such as schizophrenia and autism where the causes of the diseases are not yet resolved and where multiple candidate responses exist.

5 Acknowledgment

The present study was supported by grants from Pache Research Subsidy I-A-2, Nanzan University (2008).

References

BENJAMINI, Y. and HOCHBERG, Y. (2000): On the adaptive control of the False Discovery Rate in multiple testing with independent statistics, *J. Educ. Behav. Statist.*, 25(1): 60-83.

KAMITSUJI, S. and KAMATANI, N. (2006): Estimation of haplotype associated with several quantitative phenotypes based on maximization of are under a receiver operating characteristic (ROC) curve. *J. Hum. Genet.*, 51(4):314-325.

SHIBATA, K., ITO, T., KITAMURA, Y., IWASAKI, N., TANAKA, H., KAMATANI, N. (2004): Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. *Genetics*, 168:525-539

TOMITA, M., HATSUMICHI, M. and KURIHARA, K. (2008): Identify LD Blocks Based on Hierarchical Spatial Data. *Comput. Statist. Data Anal.*, 52(4):1806-1820.

A Representation of the Transition Density of a Logistic Diffusion Process

Franz Konecny

BOKU - University of Natural Resources and Applied Life Sciences, Vienna
Muthgasse 18, A-1190 Vienna, Austria, *franz.konecny@boku.ac.at*

Abstract. We derive an integral representation of the transition probability density function of a logistic diffusion process given by the stochastic differential equation $dN_t = r(1 - N_t/K)N_t\,dt + \sigma N_t\,dW_t$, where $r,\ K, \sigma > 0$ and W is a standard Wiener process. The derivation is based on solving a Schrödinger-type equation in terms of the confluent hypergeometric function and of the classical Bromwich inversion integral of a Laplace transform. The computational relevance of this result lies in the fact that the transition density is represented in terms of a Bromwich integral for which efficient numerical procedures are available.

Keywords: logistic diffusion process, Schrödinger-type equation, Laplace transform

1 Problem formulation

The stochastic differential equation (SDE)

$$dN_t = r(1 - N_t/K)N_t\,dt + \sigma N_t\,dW_t \tag{1}$$

is used as a model for the growth of a population of size N_t in a stochastic crowded environment. The constant $K > 0$ is called the *carring capacity* of the environment, r is the *growth rate* per individual and σ is a noise parameter. A detailed discussion of the stochastic logistic growth model can be found in chapter 6 of the monograph of Gard [1988]. Modeling, analysis and discretization of the stochastic logistic equation and some meaningful generalizations of it, are the subjects of a paper of Schurz [2007]. In many computational and statistical studies of SDEs, (1) is used as a test problem (e.g. Higham [2001], Rimmer et al. [2005] and Beskos et al. [2006]).

Performing a substitution, (1) can be reduced to a linear equation (cf. Gard [1988]). It leads to the explicit solution

$$N_t = \frac{N_0 \exp[(r - \sigma^2/2)t + \sigma W_t]}{1 + N_0(r/K)\int_0^t \exp[(r - \sigma^2/2)s + \sigma W_s]\,ds} \tag{2}$$

for $t \geq 0$.

A problem of central importance in the study of diffusion processses is the computation of the transition probability density function (t.p.d.f.)

$p_N(t;x,z) = (d/dz)P(N_t \leq z|N_0 = x)$. Maximum likelihood estimation is based on the knowledge of the t.p.d.f. , when only discrete observation of the process are available. The t.p.d.f. is known to satisfy the forward and backward Kolmogorov equations (Gihman-Skorohod [1972]), but its solutions are not known analytically.

In this paper we shall derive an integral representation of the t.p.d.f. of (N_t). The derivation is based on results of Kac [1949] and Rosenblatt [1951] upon deriving and inverting a Laplace transform. This method was adopted by Benes and Karatzas [1987] for the study of two special diffusion processes arising in the theory of electronic circuits. By a logarithmic transformation $X_t = -\log(N_t)/\sigma$, equation (1) can be reduced to

$$dX_t = \left[\frac{\sigma}{2} - \frac{r}{\sigma} + \frac{r}{\sigma K}\exp(-\sigma X_t)\right] dt + dW_t, \tag{3}$$

a SDE with constant diffusion coefficient equal to one. Instead of tackling the t.p.d.f. of X_t, called $p_X(t;x,z)$, directly, Benes and Karatzas [1987] proposed to cast it in the form

$$p_X(t;x,z) = \exp\left(\int_x^z \beta(u)\,du\right) q(t;x,z), \tag{4}$$

where $\beta(.)$ is the drift of (3). A potential $V(.)$ is introduced by

$$V(z) := \beta'(z) + \beta^2(z). \tag{5}$$

The function $q(t;x,z)$ of (5) satisfies the parabolic PDE

$$\frac{\partial q}{\partial t} = \frac{1}{2}\frac{\partial^2 q}{\partial x^2} - V(z)q, \tag{6}$$

with

$$\lim_{t\to 0+} q(t;x,z) = \delta(z-x).$$

According to Kac [1949], Rosenblatt [1951] its Laplace transform

$$\psi(z) = \psi(s;x,z) = \int_0^\infty e^{-st} q(t;z,x) \tag{7}$$

is the continuous and bounded solution of the second-order ODE

$$\psi''(z) = [2s + V(z)]\psi(z), \quad z \neq x, \tag{8}$$

which is twice continuously differentiable for $z \neq x$ and satisfy the jump condition

$$\psi'(x+) - \psi'(x-) = -2. \tag{9}$$

A quick derivation of the condition (9) is given in Benes and Karatzas [1987].

2 The t.p.d.f. of (X_t)

In the SDE (3) under consideration, the potential (5) is given by

$$V(z) = a + be^{-\sigma z} + ce^{-2\sigma z} \tag{10}$$

with

$$a = \left(\frac{\sigma}{2} - \frac{r}{\sigma}\right)^2, \;\; b = -\frac{2r^2}{K\sigma^2}, \;\; c = \frac{r^2}{K^2\sigma^2}.$$

This potential is bounded below. The change of variables

$$\psi(z) = \phi(e^{-\sigma z}), \quad y = e^{-\sigma z} \tag{11}$$

transforms the equation (8) into

$$\sigma^2 y^2 \phi''(y) + \sigma^2 y \phi'(y) = (2s + a + by + cy^2)\phi(y), \; y > 0. \tag{12}$$

The substitution $\phi(y) = y^\kappa \chi(y)$ with $\kappa = \sqrt{2s+a}/\sigma$ reduces equation (12) to

$$\sigma^2 y \chi''(y) + \sigma^2(2\kappa + 1)\chi'(y) = (cy + b)\chi(y). \tag{13}$$

Let $w(\alpha, \gamma; \xi)$ be any solution of Kummer's equation

$$\xi w''(\xi) + (\gamma - \xi)w'(\xi) - \alpha w(\xi) = 0. \tag{14}$$

Then

$$\chi(y) = e^{2cy} w(\alpha, \gamma; \xi) \tag{15}$$

with

$$\alpha = \frac{2\kappa + 1}{2} - \frac{b}{4\sigma^2 c}, \;\; \gamma = 2\kappa + 1, \;\; \xi = -4cy$$

is a solution of (13), cf. Polyanin and Zaitsev [1994, p. 21,eq. 103]. Two linearly independent solutions of Kummer's equation are the confluent hypergeometric function

$$w_1 = M(\alpha, \gamma; \xi) = 1 + \frac{\alpha}{\gamma}x + \frac{\alpha(\alpha+1)}{\gamma(\gamma+1)}\frac{x^2}{2!} + ... \tag{16}$$

and

$$w_2 = \xi^{1-\gamma} M(\alpha - \gamma + 1, 2 - \gamma; \xi), \tag{17}$$

provided that $\gamma \neq 0, -1, -2, ...$ Therefore, the original equation (8) has in this case the linearly independent solutions

$$\begin{aligned} \psi_1(z) = {} & \exp(2ce^{-2\sigma z} - 2\kappa\sigma z) \\ & \cdot M\left(\frac{-b + 2\,c\,\sigma^2\,(1 + 2\,\kappa)}{4\,c\,\sigma^2}, 1 + 2\kappa, -4\,c\,e^{-2\sigma z}\right) \end{aligned} \tag{18}$$

and

$$\psi_2(z) = \exp(2ce^{-2\sigma z} + 2\kappa\sigma z) \cdot M\left(\frac{-b + 2\,c\,\sigma^2\,(1+2\,\kappa)}{4\,c\,\sigma^2} - 2\kappa, 1-2\kappa, -4\,c\,e^{-2\sigma z}\right) \tag{19}$$

In order to discuss the asymptotic behaviour for $z \to -\infty$, we use the known relation (cf. Abramowitz and Stegun [1964, p. 504])

$$M(\alpha,\gamma,-w) \sim \frac{\Gamma(\gamma)}{\Gamma(\gamma-\alpha)} w^{-\alpha}. \tag{20}$$

We find that ψ_2 is bounded for $t \to -\infty$, whereas ψ_1 is unbounded

$$\lim_{z\to-\infty} \psi_1(z) = \infty, \quad \lim_{z\to-\infty} \psi_2(z) = 0 \tag{21}$$

For $z \to \infty$ we get the asymptotic relations

$$\psi_1 \sim \exp(2ce^{-2\sigma z} - 2z\sqrt{a+2\,s}) \tag{22}$$

$$\psi_2 \sim \exp(2ce^{-2\sigma z} + 2z\sqrt{a+2\,s}) \tag{23}$$

resulting

$$\lim_{z\to\infty} \psi_1(z) = 0, \quad \lim_{z\to\infty} \psi_2(z) = \infty \tag{24}$$

The general solution of equation (6) is of the form

$$\psi(s;x,z) = \begin{cases} C_1\psi_1(z) + C_2\psi_2(z), & -\infty < z < x \\ D_1\psi_1(z) + D_2\psi_2(z), & x < z < \infty. \end{cases} \tag{25}$$

Boundedness considerations as $z \to \infty$ and $z \to -\infty$ yield $C_1 = 0$ and $D_2 = 0$. Continuity at $z = x$ gives

$$D_1\,\psi_1(x) - C_2\,\psi_2(x) = 0, \tag{26}$$

and the jump condition (9)

$$D_1\,\psi_1'(x) - C_2\,\psi_2'(x) = -2. \tag{27}$$

In terms of the Wronskian $W = \psi_1(x)\psi_2'(x) - \psi_2(x)\psi_1'(x)$ we get

$$D_1 = \frac{2\psi_2(x)}{W}, \quad C_2 = \frac{2\psi_1(x)}{W} \tag{28}$$

Since there is no first derivative term in equation (8), the Wronskian is constant in x. Hence, after some symbolic computations with *Mathematica*, we

obtain

$$W = \frac{2\,e\,\sigma}{4\,\kappa^2 - 1} \cdot \{\lambda(1-2\,\kappa)M(1+\lambda, 2+2\kappa, -1)\,M(\lambda-2\kappa, 1-2\kappa, -1) + (1+2\,\kappa))M(\lambda, 1+2\,\kappa, -1)[2\,\kappa\,(2\,\kappa-1)M(\lambda-2\,\kappa, -1) + (2\,\kappa-\lambda)M(\lambda+1-2\,\kappa, 2-2\,\kappa, -1)]\} \tag{29}$$

where

$$\lambda = \frac{-b + 2\,c\,\sigma^2\,(1+2\,\kappa)}{4\,c\,\sigma^2}.$$

We obtain the following closed-form expression for the Laplace transform (7)

$$\psi(s; x, z) = \begin{cases} \frac{\psi_1(x)\psi_2(z)}{W}, & -\infty < z < x \\ \frac{\psi_2(x)\psi_1(z)}{W}, & x < z < \infty \end{cases}$$

The classical *Bromwich inversion formula* expresses $q(t; x, z)$ exactly via the contour integral

$$q(t; x, z) = \frac{1}{2\,\pi\,i} \int_{\delta - i\infty}^{\delta + i\infty} e^{s\,t} \psi(s; x, z)\, ds, \tag{30}$$

for a fixed $\delta > 0$.

Proposition 1. The t.p.d.e. $p_X(t; x, z)$ for the diffusion process (X_t) is given by

$$p_X(t; x, z) = \exp\left(a\,(z-x) - \frac{b}{\sigma}(e^{-\sigma z} - e^{-\sigma z}\right) q(t; x, z). \tag{31}$$

A change of variable argument yields

Proposition 2. The t.p.d.f. $p_N(t; v, w)$ of the logistic diffusion process (N_t) can be expressed in terms of the transformed process (X_t):

$$p_N(t; v, w) = \frac{1}{\sigma w}\, p_X(t; -\frac{\log v}{\sigma}, -\frac{\log w}{\sigma}), \qquad \text{for all positive } t, v, w. \tag{32}$$

The problem of computing the t.p.d.e. of (N_t) is thus reduced to computing the Bromwich integral (30). Multi-precision numerical procedures for the numerical integration the Bromwich integral were developped (cf. Abate and Valko [2004]), which can be realized in a few lines using current computer algebra systems.

3 Conclusions

Transition densities play a crucial role in likelihood inference for discretely observed diffusion processes. This inference is complicated by the unavailability of the transition density, except for some rare cases. In the case of the stochastic logistic diffusion process the t.p.d.e. is now available via computing a Bromwich integral. Thus the problem is reduced to a computational one.

Typically, inference for diffusion processes is carried out by approximate likelihood-based methods, see the review of Sørensen [2004]. Frequently, authors used the stochastic logistic equation as a test problem (e.g. Rimmer et al. [2005], Beskos et al. [2006] and Beskos et al. [2008]). For such studies an explicit formula for the t.p.d.e. is certainly useful.

References

ABATE, J. and VALKO, P.P. (2004): Multi-precision Laplace transform inversion. *Internat. J. Numer. Methods Engnr. 60, 979-995.*

ABRAMOWITZ, M. and STEGUN, I.A. (1965): *Handbook of Mathematical Functions.* Dover Publ., New York.

BENES, V. E. and KARATZAS, I. (1987): Transition Probabilities for Some 'Special' Diffusions. *J. Appl. Prob. 24,888-898.*

BESKOS, A., PAPASPILIOPOULOS, O., ROBERTS, G.O. and FEARNHEAD (2006): Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Statist. Soc. B 68(2), 1-29.*

BESKOS, A., PAPASPILIOPOULOS, O. and ROBERTS, G.O. (2006): Monte Carlo Maximum Likelihood Estimation for Discretely Observed Diffusion Processes. *to appear in Ann. Statist.*

GARD, T.C. (1988):*Introduction to Stochastic Differential Equations.* M. Dekker Inc., New York and Basel.

GIHMAN, I.I. and SKOROHOD, A.V. (1972): *Stochastic Differential Equations.* Springer-Verlag, Berlin.

HIGHAM, D.J. (2001): An Introduction to Numerical Simulation of Stochastic Differential Equations. *SIAM Review 43(3), 525-546.*

KAC, M. (1949): On Distributions of Certain Wiener Functionals. *Trans. Amer. Math. Soc. 65, 1-13.*

POLYANIN, A.D. and ZAITSEV, V.F. (1996): *Handbuch der linearen Differentialgleichungen: exakte Lösungen.* Spektrum Akademischer Verlag, Heidelberg.

RIMMER, D., DOUCET, A. and FITZGERALD, W.J. (2005): Particle filters for stochastic differential equations of nonlinear diffusions. Technical Report. University of Cambridge: Department of Engineering, Cambridge, UK.

ROSENBLATT, M. (1951): On a Class of Markov Processes. *Trans. Amer. Math. Soc. 71, 120-135.*

SCHURZ, H. (2007): Modeling, Analysis and Discretization of Stochastic Logistic Equations. *Internat. J. Numer. Anal. Model. 4(2), 178-197*

SØRENSEN, H. (2004): Parametric Inference for Diffusion Processes Observed at Discrete Points of Time: a Survey. *Internat. Statist. Rev. 72(3), 337-354*

Estimation of Sample Size to Compare the Accuracy of Two Binary Diagnostic Tests in the Presence of Partial Disease Verification

José A. Roldán Nofuentes[1], Miguel Á. Montero Alonso[2] and Juan D. Luna del Castillo[3]

[1] Biostatistics, School of Medicine, University of Granada, Spain, *jaroldan@ugr.es*
[2] School of Social Sciences, Campus of Melilla, University of Granada Avd. Alfonso XIII s/n, 52006 Melilla, Spain, *mmontero@ugr.es*
[3] Biostatistics, School of Medicine, University of Granada, Spain, *jdluna@ugr.es*

Abstract. Calculating sample size to compare the accuracy of two binary diagnostic tests is an important question in the study of diagnostic statistical methods. In the presence of partial disease verification, the disease status of some patients in the sample is unknown, so that the calculation of sample size can be complicated. In this study, we propose a method to calculate sample size to compare the sensitivities and the specificities of two binary tests in the presence of partial disease verification.

Keywords: partial verification, sample size, sensitivity, specificity.

1 Introduction

In the comparison of the accuracy of two binary diagnostic tests an important question is the determination of the sample size necessary to carry out the study. When we want to compare the sensitivities (specificities) of two binary tests, researchers have to consider calculating the sample size necessary to estimate the difference between sensitivities (specificities) with a determined precision or to compare the sensitivities (specificities) of the two diagnostic tests to an error α and a power $1-\beta$. In the presence of partial disease verification, the calculation of sample size to evaluate the accuracy of a binary test and to compare the accuracy of two binary tests cannot be carried out applying traditional methods, since sensitivity and specificity cannot be estimated as binomial proportions (Zhou, 1998). In this study, we propose a method to calculate sample size to compare the sensitivities (specificities) of two binary tests in the presence of partial disease verification.

2 The method

Comparison of the accuracy of two binary diagnostic tests is one of the most important problems in the study of diagnostic statistical methods and has

been the subject of numerous studies. The equation used to calculate sample size (m) when trying to construct a two-tailed confidence interval for the difference between the two sensitivities or specificities is

$$m = \frac{z^2_{1-\alpha/2}}{L^2} V(\hat{\theta}_1 - \hat{\theta}_2), \tag{1}$$

where z_γ is the $100\gamma th$ percentile of the normal standard distribution, L is the precision of the estimation (difference between the sensitivities or specificities) and $V(\hat{\theta}_1 - \hat{\theta}_2)$ is the variance function (McCullagh and Nelder, 1989) of $\hat{\theta}_1 - \hat{\theta}_2$, when θ_i is the sensitivity (Se_i) or the specificity (Sp_i) of each diagnostic test. Therefore, in order to calculate the sample size it is necessary to previously determine $V(\hat{\theta}_1 - \hat{\theta}_2)$.

Let us consider two diagnostic binary tests which are applied independently to the same random sample of n patients. Let T_1 and T_2 be random variables which model the results diagnostic tests 1 and 2 respectively, in such a way that $T_h = 1$ when the result of the hth test ($h = 1, 2$) is positive and $T_h = 0$ when the result of the hth test is negative. Let V be the random variable which models the verification process, $V = 1$ when the patient is verified with the gold standard and $V = 0$ when the patient is not verified; and D be the random variable which models the result of the gold standard, $D = 1$ when the patient is diseased and $D = 0$ when the patient is non-diseased. Let $Se_h = P(T_h = 1|D = 1)$ and $Sp_h = P(T_h = 0|D = 0)$ be the sensitivity and the specificity of the hth diagnostic test ($h = 1, 2$), $p = P(D = 1)$ the disease prevalence, and $\lambda_{ijk} = P(V = 1|T_1 = i, T_2 = j, D = k)$ the probability of verifying a patient with results $T_1 = i$, $T_2 = j$ and $D = k$, with $i, j, k = 0, 1$. The application of the two diagnostic tests to all of the patients in a random sample sized n and the application of the gold standard to a part of the sample gives us Table 1. When the verification process only depends on the

Table 1. Frequencies observed when comparing two binary tests in the presence of partial disease verification.

		$T_1 = 1$		$T_1 = 0$		
		$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$V = 1$						
	$D = 1$	s_{11}	s_{10}	s_{01}	s_{00}	s
	$D = 0$	r_{11}	r_{10}	r_{01}	r_{00}	r
$V = 0$		u_{11}	u_{10}	u_{01}	u_{00}	u
	Total	n_{11}	n_{10}	n_{01}	n_{00}	n

results of the two binary tests and not on the result of the gold standard, it is verified that $\lambda_{ijk} = \lambda_{ij}$, $i, j, k = 0, 1$. This assumption is equivalent to supposing that the verification process is missing at random *(MAR)* (Rubin,

1976). Subject to the *MAR* assumption, let the probabilities be

$$\xi_{ij} = P(V = 1, D = 1, T_1 = i, T_2 = j),$$

$$\psi_{ij} = P(V = 1, D = 0, T_1 = i, T_2 = j), \zeta_{ij} = P(V = 0, T_1 = i, T_2 = j) \quad (2)$$

with $i, j = 0, 1$, and $\sum_{i,j=0}^{1} \xi_{ij} + \sum_{i,j=0}^{1} \psi_{ij} + \sum_{i,j=0}^{1} \zeta_{ij} = 1$. In general, and as happens in most practical situations, both diagnostic tests are conditionally dependent on the disease (Torrance-Rynard and Walter, 1997), so that the probabilities (2) are expressed in terms of sensitivities, specificities, prevalence and dependence factors such as (Roldán Nofuentes and Luna del Castillo, 2005)

$$\xi_{ij} = p\lambda_{ij}\{Se_1^i(1 - Se_1)^{1-i} Se_2^j(1 - Se_2)^{1-j} + \delta_{ij} Se_1 Se_2(\varepsilon_1 - 1)\},$$

$$\psi_{ij} = (1-p)\lambda_{ij}\{Sp_1^{1-i}(1-Sp_1)^1 Sp_2^{1-j}(1-Sp_2)^j + \delta_{ij}(1-Sp_1)(1-Sp_2)(\varepsilon_0-1)\}, \quad (3)$$

$$\zeta_{ij} = \frac{(1 - \lambda_{ij})}{\lambda_{ij}}(\xi_{ij} + \psi_{ij}),$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = -1$ if $i \neq j$, and the parameters ε_1 and ε_0 are the dependence factors between the two diagnostic tests. The dependence factor ε_1 (ε_0) is the covariance between both diagnostic tests when $D = 1$ ($D = 0$)(Berry at al, 2002), and it is verified that $1 \leq \varepsilon_1 \leq \dfrac{1}{\max Se_k}$ and $1 \leq \varepsilon_0 \leq \dfrac{1}{\max(1 - Sp_k)}$. If $\varepsilon_1 = \varepsilon_0 = 1$, both diagnostic tests are conditionally independent on the disease. When $D = 1$, the correlation between the two diagnostic tests is

$$\rho_1 = Corr(T_1 = 1, T_2 = 1|D = 1) = (\varepsilon_1 - 1)\sqrt{\frac{Se_1 Se_2}{(1 - Se_1)(1 - Se_2)}},$$

and when $D = 0$ the correlation is

$$\rho_2 = Corr(T_1 = 1, T_2 = 1|D = 0) = (\varepsilon_0 - 1)\sqrt{\frac{(1 - Sp_1)(1 - Sp_2)}{Sp_1 Sp_2}}.$$

Let $\omega = (\xi_{11}, \xi_{10}, \xi_{01}, \xi_{00}, \psi_{11}, \psi_{10}, \psi_{01}, \psi_{00}, \zeta_{11}, \zeta_{10}, \zeta_{01}, \zeta_{00})^T$. As ξ_{ij}, ψ_{ij} and ζ_{ij} are the probabilities of a multinomial distribution, the variance-covariance matrix of ω is $\sum = diag(\omega) - \omega^T\omega$.

Let $\phi_{ij} = \xi_{ij} + \psi_{ij} + \zeta_{ij}$ with $i, j = 0, 1$. The sensitivity and the specificity of each diagnostic test can be written in terms of the previous probabilities

as

$$Se_1 = \frac{\sum_{j=0}^{1} \frac{\xi_{1j}\phi_{1j}}{\xi_{1j}+\psi_{1j}}}{\sum_{i,j=0}^{1} \frac{\xi_{ij}\phi_{ij}}{\xi_{ij}+\psi_{ij}}} \qquad \text{and} \qquad Sp_1 = \frac{\sum_{j=0}^{1} \frac{\psi_{0j}\phi_{0j}}{\xi_{0j}+\psi_{0j}}}{\sum_{i,j=0}^{1} \frac{\psi_{ij}\phi_{ij}}{\xi_{ij}+\psi_{ij}}},$$

for diagnostic test 1, and

$$Se_2 = \frac{\sum_{i=0}^{1} \frac{\xi_{i1}\phi_{i1}}{\xi_{i1}+\psi_{i1}}}{\sum_{i,j=0}^{1} \frac{\xi_{ij}\phi_{ij}}{\xi_{ij}+\psi_{ij}}} \qquad \text{and} \qquad Sp_2 = \frac{\sum_{i=0}^{1} \frac{\psi_{i0}\phi_{i0}}{\xi_{i0}+\psi_{i0}}}{\sum_{i,j=0}^{1} \frac{\psi_{ij}\phi_{ij}}{\xi_{ij}+\psi_{ij}}},$$

for test 2. From these expressions, it holds that

$$Se_1 - Se_2 = \left\{ \frac{\xi_{10}\phi_{10}}{\xi_{10}+\psi_{10}} - \frac{\xi_{01}\phi_{01}}{\xi_{01}+\psi_{01}} \right\} \Bigg/ \sum_{i,j=0}^{1} \frac{\psi_{ij}\phi_{ij}}{\xi_{ij}+\psi_{ij}}$$

and

$$Sp_1 - Sp_2 = \left\{ \frac{\xi_{01}\phi_{01}}{\xi_{01}+\psi_{01}} - \frac{\xi_{10}\phi_{10}}{\xi_{10}+\psi_{10}} \right\} \Bigg/ \sum_{i,j=0}^{1} \frac{\psi_{ij}\phi_{ij}}{\xi_{ij}+\psi_{ij}},$$

and applying the delta method, it holds that

$$V(\widehat{\theta}_1 - \widehat{\theta}_2) = \left(\frac{\partial(\theta_1-\theta_2)}{\partial\omega} \right) \sum_{\omega} \left(\frac{\partial(\theta_1-\theta_2)}{\partial\omega} \right)^T$$

where θ is the sensitivity or specificity, and the partial derivates $\frac{\partial(\theta_1-\theta_2)}{\partial\omega}$ are calculated in the routine way. Therefore, if we know the conjectured sensitivity and specificity for each diagnostic test, the disease prevalence, the covariances ε_1 and ε_0, applying equation (1) it is easy to calculate the sample size to estimate the difference between the two sensitivities (specificities) with a precision L and a confidence of $100(1-\alpha)\%$.

The method which we propose to calculate sample size requires knowledge of the sensitivity and specificity, prevalence, verification probabilities and the covariances (correlations) between the two binary tests. In practice the calculation of sample size is difficult, since the covariances ε_1 and ε_0 (or the correlations ρ_1 and ρ_2) are difficult to estimate. Nevertheless, these covariances (correlations) can be estimated by applying the *EM* algorithm. In Appendix I, we show an *EM* algorithm which allows us to estimate the values of the covariances between the two diagnostic tests in the presence of verification bias. On the other hand, if for a concrete example we calculate the sample size to compare the sensitivities and the sample size to compare

the specificities, the sample size necessary to realize the study (comparison of two binary diagnostics tests) will be the maximum value of both sample sizes calculated.

3 Simulation study

We conducted simulation experiments to study the robustness of the method proposed to calculate sample size when estimating the difference between the two sensitivities (specificities). We generated 5000 random samples with multinomial distributions and probabilities given as (3). As sensitivities and specificities we took the values ($Se_1 = 0.90, Sp_1 = 0.85, Se_2 = 0.85, Sp_2 = 0.80$) and ($Se_1 = 0.90, Sp_1 = 0.85, Se_2 = 0.80, Sp_2 = 0.70$) which are values which appear with a certain frequency in clinical practice; as prevalence we considered the values 10%, 30%, 50%, 70% and 90%, and as verification probabilities we took $\lambda_{11} = 0.80, \lambda_{10} = \lambda_{01} = 0.40, \lambda_{00} = 0.10$. For 5000 samples with the same multinomial distribution we calculated the average sample size and the relative root mean squared error (RRMSE):

$$RRMSE = \sqrt{\sum_{i=1}^{5000} \frac{(\widehat{m}_i - m)^2}{4999}} \Bigg/ m,$$

where m is the sample size calculated from the values with which the multinomial samples were generated and $\widehat{m}_i$ is the estimated value of the sample size calculated from the maximum likelihood estimators of sensitivity, specificity, prevalence and verification probabilities obtained from each random sample.

The sample size was calculated through the method proposed in previous Section, and the maximum likelihood estimators for sensitivities, specificities, prevalence and correlations were obtained applying the EM algorithm described in Appendix I; the maximum likelihood estimators for the verification probabilities were obtained through the equations $\hat{\lambda}_{ij} = (s_{ij} + r_{ij})/n_{ij}$, with $i, j = 0, 1$. In Table 2 we show some of the results obtained for ($Se_1 = 0.90, Sp_1 = 0.85, Se_2 = 0.85, Sp_2 = 0.80$) and different values of ρ_1 and ρ_2. From the results obtained with these experiments we can deduce the following conclusions. For sensitivities, in general terms, when the correlation ρ_1 between the two binary tests is low or intermediate (independent of the value of the correlation ρ_2), the estimator of size to compare the two sensitivities has a relative root mean square error (RRMSE) lower than 25%; while if the correlation ρ_1 is high, its RRMSE is higher than 25%. Regarding specificities, in general terms, when the correlation ρ_2 between the two binary tests is low or intermediate, the estimator of simple size to compare the two specificities has an RRMSE lower than 25%; whilst if the correlation ρ_2 is high its RRMSE is higher than 25%. Therefore, the correlations ρ_1 and ρ_2 between the two diagnostic tests has an important effect on the method which we propose to calculate sample size, and we think that this is due to the minor

effect that the correlations have on the proportion of verified patients, since the increase of ρ_1 and/or ρ_2 does not have an important effect on the increase in the percentage of verified patients. Similar conclusions are obtained for $(Se_1 = 0.90, Sp_1 = 0.85, Se_2 = 0.80, Sp_2 = 0.70)$.

Table 2. Sample sizes $(L = 0.05, \alpha = 5\%)$.

Sample sizes to compare sensitivities									
$Se_1 = 0.90, Sp_1 = 0.85, Se_2 = 0.85, Sp_2 = 0.80, \delta = Se_1 - Se_2 = 0.05$									
	$\rho_1 = \rho_2 = 0$			$\rho_1 = 0.36 \rho_2 = 0.37$			$\rho_1 = 0.36 \rho_2 = 0.37$		
p	m	ASS	$RRMSE$	m	ASS	$RRMSE$	m	ASS	$RRMSE$
10%	7892	7772	0.09	5063	4984	0.16	2090	2336	0.36
30%	2341	2304	0.09	1483	1462	0.16	562	673	0.43
50%	1222	1200	0.10	773	763	0.16	292	356	0.47
70%	732	716	0.10	467	459	0.17	187	212	0.40
90%	444	439	0.10	287	297	0.18	127	153	0.46
Sample sizes to compare specificities									
$Se_1 = 0.90, Sp_1 = 0.85, Se_2 = 0.85, Sp_2 = 0.80, \delta = Sp_1 - Sp_2 = 0.05$									
	$\rho_1 = \rho_2 = 0$			$\rho_1 = 0.36 \rho_2 = 0.37$			$\rho_1 = 0.36 \rho_2 = 0.37$		
p	m	ASS	$RRMSE$	m	ASS	$RRMSE$	m	ASS	$RRMSE$
10%	550	573	0.09	350	385	0.17	148	199	0.56
30%	863	882	0.08	546	563	0.14	217	271	0.46
50%	1444	1459	0.08	906	915	0.13	337	414	0.47
70%	2848	2852	0.08	1780	1791	0.14	637	743	0.44
90%	10126	10125	0.08	6408	6442	0.16	2433	2712	0.41

p: prevalence; m: sample size; ASS: average sample size;
RRMSE: relative root mean squared error.

4 Discussion

In the presence of partial disease verification, the calculation of sample size cannot be carried out using the traditional methods, since a sub sample of patients does not have its disease status verified. In this study, we propose a method to calculate sample size when we compare the accuracy of two binary tests when not all of the patients are verified with the gold standard. The method which we propose to calculate sample size requires that the verification process be MAR. It is also necessary to know the values of sensitivity and specificity of each test, disease prevalence, covariances between the two diagnostic tests and the verification probabilities (values which can be conjectured through an initial study or a pilot sample), as well as the specifications to calculate sample size (confidence and precision). We carried out simulation experiments to study the robustness of the method which we have proposed to calculate the sample size. We have studied the effect that prevalence, verification probabilities and the correlations (ρ_1 and ρ_2) have on sample size to

estimate the difference in sensitivities (specificities) to a precision L with a confidence $100(1-\alpha)\%$. The results of the simulation experiments have shown that the correlations (ρ_1 and ρ_2) are the parameters which have the greatest effect on sample size and it holds that the method which we propose to calculate sample size to estimate the difference in sensitivities (specificities) is valid when the correlation $\rho_1(\rho_2)$ is low or intermediate, and is not a good method when the correlation $\rho_1(\rho_2)$ is high. These correlations can be estimated from a pilot sample, or from previous studies, applying the *EM* algorithm which we propose in Appendix I, and their corresponding standard errors can be estimated applying the *SEM* algorithm (Meng and Rubin, 1991).

On the other hand, the method which we propose to calculate the sample size is based on the normal approach to estimators. When this approach is not valid, the method cannot be applied. Another limitation of our method lies in the fact that the equations of sample size are based on Wald confidence intervals, so that when the values of the accuracy of the diagnostic test are very near to 0 or to 1, the method used to calculate the sample size is affected by the bad performance of the confidence interval. We relieve that a possible solution to these problems could be found by applying exact inference methods or multiple imputation.

Finally, the method that we propose to calculate the sample size can be applied when we compare the likelihood ratios or the weighted kappa coefficients (Roldán Nofuentes and Luna del Castillo, 2005a and 2005b) of two binary diagnostic tests in the presence of verification bias.

Appendix I

Let r_{ij} (s_{ij}) be the number of diseased (non-diseased) patients with $T_1 = i$ and $T_2 = j$, and u_{ij} the number of non-verified patients with $T_1 = i$ and $T_2 = j$. In the situation of partial verification, the missing information is the result of the gold standard of each non-verified patient. This information is reconstructed in step E of the algorithm, and in step M we impute the maximum likelihood estimators from the reconstructed data in the previous step. Let us suppose that from each frequency u_{ij} of non-verified patients x_{ij} patients are diseased and $u_{ij} - x_{ij}$ patients are non-diseased. Therefore, the data observed can be expressed in the form of a 2×4 table with $s_{ij} + x_{ij}$ frequencies for $D = 1$ and $r_{ij} + u_{ij} - x_{ij}$ for $D = 0$. Subject to the *MAR* assumption, the logarithm of the likelihood function of the data in this 2×4 table is $l \propto \sum_{i,j=0}^{1}(s_{ij} + x_{ij}) \log\{P(T_1 = i, T_2 = j, D = 1)\} + \sum_{i,j=0}^{1}(r_{ij} + u_{ij} - x_{ij}) \log\{P(T_1 = i, T_2 = j, D = 0)\}$.

Let Se_i and Sp_i be the sensitivity and specificity of each diagnostic test, p be the disease prevalence and ε_k the covariance between the two diagnostic tests when $D = k$. Subject to the *MAR* assumption, let $x_{ij}^{(k)}$ be the values of x_{ij} in the kth iteration of the *EM* algorithm. The values of the *MLEs* in the kth iteration are calculated through the expressions:

$$\widehat{Se}_1^{(k)} = \frac{\sum_{j=0}^{1}(s_{1j} + x_{1j}^{(k)})}{s + x^{(k)}}, \qquad \widehat{Sp}_1^{(k)} = \frac{\sum_{j=0}^{1}(r_{0j} + u_{0j} - x_{0j}^{(k)})}{r + u - x^{(k)}},$$

$$\widehat{Se}_2^{(k)} = \frac{\sum_{i=0}^{1}(s_{i1} + x_{i1}^{(k)})}{s + x^{(k)}}, \qquad \widehat{Sp}_2^{(k)} = \frac{\sum_{i=0}^{1}(r_{i0} + u_{i0} - x_{i0}^{(k)})}{r + u - x^{(k)}},$$

$$\hat{p}^{(k)} = \frac{s + x^{(k)}}{n}, \qquad \hat{\varepsilon}_1^{(k)} = \frac{(s + x^{(k)})(s_{11} + x_{11}^{(k)})}{\left\{\sum_{i=0}^{1}(s_{i1} + x_{i1}^{(k)})\right\}\left\{\sum_{j=0}^{1}(s_{1j} + x_{1j}^{(k)})\right\}},$$

$$\hat{\varepsilon}_0^{(k)} = \frac{(r + u - x^{(k)})(r_{11} + u_{11} - x_{11}^{(k)})}{\left\{\sum_{i=0}^{1}(r_{i1} + u_{i1} - x_{i1}^{(k)})\right\}\left\{\sum_{j=0}^{1}(r_{1j} + u_{1j} - x_{1j}^{(k)})\right\}},$$

when $s = \sum_{i,j=0}^{1} s_{ij}$, $r = \sum_{i,j=0}^{1} r_{ij}$, $u = \sum_{i,j=0}^{1} u_{ij}$, $x^{(k)} = \sum_{i,j=0}^{1} x_{ij}^{(k)}$ and $n = s + r + u$. The estimators in the following iteration are optained sustituyendo en las ecuaciones anteriores k por k+1, and where

$$x_{ij}^{(k+1)} = u_{ij}\frac{P^{(k)}(T_1 = i, T_2 = j, D = 1)}{P^{(k)}(T_1 = i, T_2 = j, D = 1) + P^{(k)}(T_1 = i, T_2 = j, D = 0)}$$

with $i, j = 0, 1$ and where $P^{(k)}(T_1 = i, T_2 = j, D)$ is $P(T_1 = i, T_2 = j, D)$ in the kth iteration of the *EM* algorithm.

References

BERRY G, SMITH, C.L., MACASKILL, P., IRWIG L., (2002): Analytic methods for comparing two dichotomous screening or diagnostic tests applied to two populations of differing disease prevalence when individuals negative on both tests are unverified. *Statistics in Medicine 21, 853-862.*

McCULLAGH, P., NELDER, J.A., (1989): *Generalized linear models.* Chapman and Hall, Boca Raton, FL.

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.

ROLDÁN NOFUENTES, J.A. and LUNA DEL CASTILLO, J.D. (2005a): Comparing the likelihood ratios of two binary diagnostic tests in the presence of partial verification. *Biometrical Journal* 47, 442-457.

ROLDÁN NOFUENTES, J.A. and LUNA DEL CASTILLO, J.D. (2005b): Comparing two binary diagnostic tests in the presence of verification bias. *Computational Statistics and Data Analysis* 50, 1551-1564.

RUBIN, D.B., (1976): Inference and missing data. *Biometrika 4, 73-89.*

TORRANCE-RYNARD, V.L., WALTER, S.D., (1997): Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine 16, 2157-2175.*

ZHOU, X.H., (1998): Comparing accuracies of two screening tests in a two-phase study for dementia. *Journal of Royal Statistical Society Series C Applied Statistics 47, 135-147.*

Goodness of Fit for Auto-Copulas: Testing the Adequacy of Time Series Models

Pál Rakonczai[1], László Márkus[1], and András Zempléni[1]

Eötvös Loránd University, Department of Probability Theory and Statistics, Budapest, Hungary, *paulo@math.elte.hu*

Abstract. Copula models proved to be powerful tools in describing the interdependence structure of multivariate data sets. The advantage of using copulas lays in the fact that - unlike Pearson-correlations - they represent nonlinear dependencies as well, and make it possible to study the interdependence of high (or low) values of the variables. So, it is very natural to extend the use of copulas to the interdependence structure of time series. To the analogy of the autocorrelation function the use of auto-copulas for the lagged series can reveal the specifics of the dependence structure in a much finer way. Therefore the fit of the corresponding auto-copulas can provide an important tool for evaluating time series models. For making comparisons among competing models the fit of copulas has to be measured. Our suggested goodness of fit test is based on the probability integral transformation of the joint distribution reducing this way the multivariate problem to one dimension. We apply the proposed methods for investigating the auto-dependence of river flow time series with particular focus on the synchronised appearance of high values and we also consider how the evaluating methods could be improved by choosing different weights for more efficient detecting.

Keywords: copulas, goodness of fit tests, probability integral transformation, river flow time series

1 Copula theory and GOF tests

1.1 Archimedean copulas

Due to Sklar's theorem we know that for any bivariate distribution function $H(x, y)$, with $F(x), G(y)$ univariate marginals there exits a copula C such that $H(x, y) = C(F(x), G(y))$. [1] This fact allows us to capture the dependence structure without specifying the marginal distributions. For the the purpose of this paper the Archimedean family was chosen because of its very convenient build up and its capabilty for handling the dependence structure between extremes as well.

[1] Moreover, C is unique if the marginal distributions are continuous.

Let us consider a copula generator function: $\phi_\theta(u) : [0,1] \to [0,\infty]$, which is continuous and strictly decreasing with $\phi_\theta(1) = 0$. Then the distribution function of the Archimedean copula is

$$C_{\phi_\theta}(u,v) = \phi_\theta^{-1}\Big(\phi_\theta(u) + \phi_\theta(v)\Big). \tag{1}$$

Since the focus of this paper is to introduce some appropriate methods for checking a given model's adequacy and not to consider many different copula models [2] , we review only the Gumbel family, which is eligible for our purpose. We should emphasize, however, that the presented methods can be adapted for more general copula models as well.

The generator function of the Gumbel copula has the form $\phi_\theta(u) = [-ln(u)]^\theta$, where $\theta \in [1, +\infty)$. Hence the Gumbel bivariate copula distribution function is given by

$$C_{Gumbel}(u,v) = e^{-([-ln(u)]^\theta + [-ln(v)]^\theta)^{\frac{1}{\theta}}}. \tag{2}$$

Simulations for Monte Carlo applications can be performed by general methods, such as conditional sampling, that can be computed quite easily with the help of the derivatives of the function $\phi_\theta^{-1}(t)$, for further details see Cherubini et al. (2004). Aside of the simulation, another relevant question is the parameter estimation which can be achieved easily based on the measures of association, e.g. for Kendall's τ it is known that $\tau_{Gumbel}(\theta) = 4\int_0^1 \frac{\phi_\theta(u)}{\phi'_\theta(u)}du + 1 = 1 - \frac{1}{\theta}$. Inverting the above statistics we get a consistent estimator for the parameter

$$\hat{\theta}_n = \tau^{-1}_{Gumbel}(\tau_n) = \frac{1}{1-\tau_n}, \tag{3}$$

where $\tau_n = -1 + \frac{4}{n(n-1)} \sum_{i \neq j} \mathbf{1}(X_i \leq X_j, Y_i \leq Y_j)$ is the sample version of Kendall's τ.

1.2 Goodness of fit tests based on the copula distribution function

The GoF statistics, we discuss, are based on the probability integral transformation (PIT) of the the copula distribution function. Let a random vector (X,Y) possess a bivariate copula model C_θ with unknown marginal distribution functions F, G and $(X_1,Y_1),...,(X_n,Y_n)$, $n \geq 2$ a random sample. The so called K-function $K(\theta,t)$ of a copula is the distribution function of the variable $H(X,Y)$, where H is the joint distribution function of (X,Y):

$$K(\theta,t) = P(H(X,Y) \leq t) = P(C_\theta(F(X), G(Y)) \leq t). \tag{4}$$

[2] The open-source R language includes an entire copula package with all of the known classes of copulas. There are options for estimating copula parameters and for simulating from a given copula model. For description see http://cran.r-project.org/doc/packages/copula.pdf

In the case of Gumbel copula family (4) can be computed as

$$K_{Gumbel}(\theta, t) = t - \frac{\phi_\theta(t)}{\phi'_\theta(t)} = t(1 - \frac{ln(t)}{\theta}), \tag{5}$$

where $t \in (0, 1]$.[3]

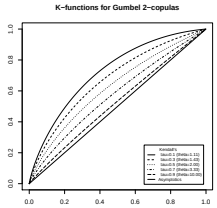


Fig. 1. K-functions for Gumbel copulas for different Kendall's τ.

Define the empirical version of the K-function as

$$K_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(E_{in} \leq t), t \in [0, 1], \tag{6}$$

where $E_{in} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}(X_k \leq X_i, Y_k \leq Y_i)$. In order to investigate the null-hypothesis of the Gumbel family, we first fit a copula model to the observations, then compute its K-function $K(\hat{\theta}_n, t)$, and compare it to the empirical counterpart $K_n(t)$. Known tests for the bivariate case (see Genest at el. 2006) use continuous functionals of Kendall's process $\kappa_n(t) = \sqrt{n}(K(\theta_n, t) - K_n(t))$ having favourable asymptotic properties. Our suggestion is based on the Cramer- von Mises[4] type statistics $S_n = \int_0^1 (\kappa_n(t))^2 dt$, but we propose to introduce appropriate weights to create two weighted test statistics as in the second column. By doing so, we expect these new tests to be more sensitive in detecting discrepancies, that occur near the tails (see Rakonczai, 2007).

Deviations	Weighted deviations
$S_1 = \sum_{t_i \in [0+\varepsilon, 1-\varepsilon]} \lvert K(\theta_n, t_i) - K_n(t_i) \rvert$	$S_3 = \sum_{t_i \in [0+\varepsilon, 1-\varepsilon]} \frac{(K(\theta_n, t_i) - K_n(t_i))^2}{K(\theta_n, t_i)}$
$S_2 = \sum_{t_i \in [0+\varepsilon, 1-\varepsilon]} (K(\theta_n, t_i) - K_n(t_i))^2$	$S_4 = \sum_{t_i \in [0+\varepsilon, 1-\varepsilon]} \frac{(K(\theta_n, t_i) - K_n(t_i))^2}{K(\theta_n, t_i)^2}$

where $(t_i)_{i=1}^n$ is an appropriately fine division of the interval $(0, 1)$. The empirical distribution of these statistics can be obtained by Monte Carlo simulation from the fitted model, and one can use some high quantiles as critical values for the hypothesis testing.

[3] For the general form for multivariate case we refer to Genest, et al. (2006).

[4] Other type of measures are also conceivable e.g. Kolmogorov-Smirnov type statistics $T_n = sup_{0 \leq t \leq 1} |\kappa_n(t)|$, but these are proved to be generally less powerful.

2 Interdependence structure of river flow series

An application of the aforementioned methods is aimed at the evaluation of models for Danube and Tisza Rivers in Hungary, fitted to daily river discharge data. Model construction is described in detail and is tested against extremal characteristics like the fit of quantiles, maxima, extremal index, time spent over thresholds (flood duration) and the aggregate excesses (flood volume) in Elek and Márkus (2008), Vasas et.al (2007). Here we address the adequacy of the interdependence structure of the time series simulated from those fitted models, as compared to the observed discharge series of Tisza River at Tivadar gauge. Analogously to the autocorrelation function, we characterise the interdependence structure of time series through the auto-copulas, that is the copulas of the lagged series. The empirical auto-copulas for the observed data are shown in Figure 2.

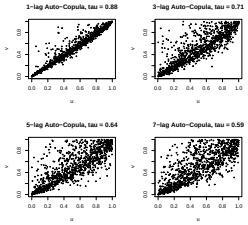


Fig. 2. Empirical auto-copulas and Kendall's correlation (tau) at fixed lags for River Tisza at Tivadar gauge.

First we briefly sketch the models given in the mentioned papers, where motivation and further analysis are also presented. The deseasonalised river flow series $X_t - c_t$ (with seasonal component c_t) has skewed, leptokurtic and light-tailed marginal distribution. The autocorrelated squares and absolute values of the innovations of a fitted ARMA filter and its non-seasonal periods of high and low variances point to conditionally heteroscedastic (GARCH-type in wide sense) modelling. Elek and Markus (2008) suggests:

$$\begin{aligned} X_t = c_t + \sum_{i=1}^{p} a_i (X_{t-i} - c_{t-i}) + \sum_{i=1}^{q} b_i \epsilon_{t-i} \\ \epsilon_t = \sigma(X_{t-1}) Z_t, \\ \sigma(x) = (\alpha_0 + \alpha_1 (x-m)_+)^{1/2}. \end{aligned}$$

with positive constants $a_i, b_i, \alpha_0, \alpha_1, m$, innovation ϵ_t and noise Z_t. The model differs from conventional ARMA-GARCH ones as the variance of innovations is conditioned on the lagged values of the *generated process* instead of the innovations themselves and is asymptotically *proportional* (instead of being a quadratic function) to its past values. Vasas et al. (2007) propose a regime

switching approach, where the dynamics is governed by two regimes, along which both the autoregressive coefficients and the innovation distributions are altering, moreover, the hidden regime indicator process is allowed to be non-Markovian. Assume that the discharge process, now denoted by Y_t for distinction, is governed by the hidden regime process I_t in the following way:

$$Y_t = Y_{t-1} + \epsilon_{1,t} \qquad \text{if} \qquad I_t = 0 \tag{7}$$

$$Y_t = a(Y_{t-1} - c) + c + \epsilon_{2,t} \qquad \text{if} \qquad I_t = 1 \tag{8}$$

where $\epsilon_{1,t}$ is an i.i.d. sequence distributed as $\Gamma(\alpha, \lambda)$ (i.e. as a gamma distribution with shape parameter α and scale parameter λ) and $\epsilon_{2,t}$ is an i.i.d. Gaussian sequence with zero mean and variance σ^2. α, λ and σ are positive real numbers and we assume that $0 < a < 1$. The duration of the $I_t = 0$ regime is distributed as negative binomial with parameters (b, p_0) and the duration of the $I_t = 1$ regime is geometrically distributed with parameter p_1, where $b > 0$ and $0 < p_i < 1$ $(i = 0, 1)$. The tail behaviour and extremal clustering of this regime switching model was investigated in Elek and Zempléni (2008).

3 Results and conclusions

After fitting the mentioned two models to the discharge series of Tivadar gauge we simulated artificial discharges from both. This allowed for comparison of the auto-copulas of the observed river flow data with that of the simulated series at fixed lags by means of the GOF test proposed in section 1. As a reference we also compare the Gumbel copulas fitted to the lagged observed series, with the corresponding empirical copula. The simulations from the fitted Gumbel model are illustrated in Figure 3. It is obvious that the

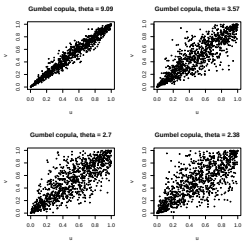


Fig. 3. Empirical auto-copulas at fixed lags for tthe simulation from the fitted Gumbel model.

symmetric copula model can not be perfect, because of the *asymmetry* arising from the dynamics of river flows. Even so, the simulated structure shows a visually quite appealing similarity with the observations (Fig. 2). Note that

near the tails the dependence structure is particularly well captured. Performing the tests for checking the hypothesis whether the observations can arise from the Gumbel model we found that the fit is acceptable. None of the mentioned tests (weighted and non-weighted) rejects the hypothesis. Figure 4 displays a diagnostic plot of differences for the K-functions at 3 days lag together with the 95% confidence bounds. The bounds were determined by simulations from the theoretical model (200 repetitions).

Sim. Quantiles of	$q = 0.9$	$q = 0.95$	$q = 0.975$	$q = 0.99$	Obs-Stat.	Obs-Quant.
S_1	3.839	4.128	4.298	4.696	3.575	0.778
S_2	0.058	0.068	0.077	0.086	0.047	0.754
S_3	0.189	0.212	0.22	0.232	0.214	0.944
S_4	1.868	2.178	2.42	2.518	2.497	0.976

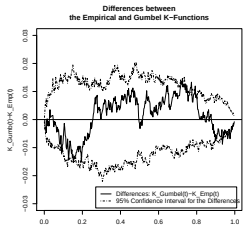


Fig. 4. Differences between K-functions at 3 days lag with 95% confidence bound.

Figure 5 and 6 shows the auto-copulas of the artificial discharges simulated from the two competing time series models. An asymmetry appears in the copulas, showing that the Gumbel model is not quite adequate in this respect and the dependence decays faster with the lags than in the observed series. Basically both copula fits are weaker than that of the Gumbel, but we have to emphasise here that the Gumbel copula is not associated with a dynamical model, and so it can not be utilised for simulating flow series.

Statistics for the 3 simulated models

Modell	S_1	S_2	S_3	S_4
Heterosc.	14.335	0.6367	2.7294	24.5510
Reg. Sw.	17.028	1.0220	4.5876	38.1918
Gumbel	3.579	0.0474	0.2140	2.4967

Figure 7 compares the auto-copula of the two dynamical models and the fitted Gumbel copula in terms of tail dependence. Although globally the Gumbel model outperforms by far the two dynamical ones, at the high values, so important for applications, i.e. at the quantile range of $0.8 - 1$ the regime switching model is better than the other two. This is very much in line with

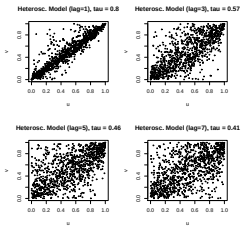


Fig. 5. Empirical auto-copulas at fixed lags for the simulation from the fitted Heteroscedastic Model.

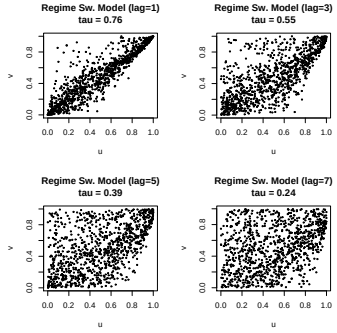


Fig. 6. Empirical auto-copulas at fixed lags for the simulation from the fitted Regime Switching Model.

the fact that the theoretical extremal index of the regime switching model is less than one - unlike the heteroscedastic model – indicating the clustering of high values. In the range of 0.9 – 1 all three models perform similarly. Perhaps more simulations would be needed for a further reaching analysis in this range.

Statistics at the quantile range of 0.8 - 1

Model	S_1	S_2	S_3	S_4
Heterosc.	1.083	0.0208	0.0267	0.0343
Reg. Sw.	0.449	0.0036	0.0047	0.0061
Gumbel	0.511	0.0052	0.0069	0.0092

Statistics at the quantile range of 0.9 - 1

Model	S_1	S_2	S_3	S_4
Heterosc.	0.249	0.0023	0.0026	0.0030
Reg. Sw.	0.155	0.0009	0.0010	0.0012
Gumbel	0.157	0.0010	0.0011	0.0012

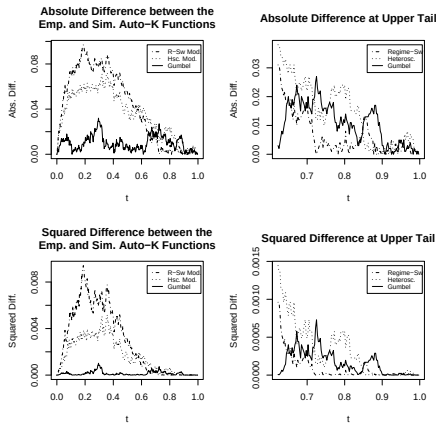


Fig. 7. Tail dependence for the different models.

References

CHERUBINI, U. LUCIANO, E. and VECCHIATO, W. (2004): *"Copula methods in Finance"*, WileyFinance, West Sussex, England.

ELEK, P. and ZEMPLÉNI, A. (2008): *"Tail behaviour and extremes of two-state Markov-switching autoregressive models"*, Computers and Mathematics with Applications

ELEK, P. and MÁRKUS, L. (2008): *"A light-tailed conditionally heteroscedastic model with applications to river flows"*, J. Time Series Analysis, Vol.29, No.1, pp.14-36.

GENEST, C. Quessy, J.-F. and RÉMILLIARD, B. (2006): *"Goodness-of-fit Procedures for Copula Models Based on the Integral Probability Transformation"*, Scandinavian J. of Statistics, 33, pp. 337-366.

GENEST, C. and FAVRE, A.-C. (2006): *"Everything you always wanted to know about copula modeling but were afraid to ask"*, Journal of Hydrologic Engineering.

RAKONCZAI, P. BOZSÓ, D. and ZEMPLENI, A. (2005): *"Goodness of fit in extreme value analysis and for copulas"*, Morgan Stanley Conference on Quantitative and Mathematical Finance, Budapest, Hungary.

RAKONCZAI, P. and ZEMPLÉNI, A. (2007): *"Copulas and goodness of fit tests"*, in Recent Advances in Stochastic Modelling and Data Analysis, World Scientific Publishing

VASAS, K. and ELEK, P. and MÁRKUS, L. (2007): *"Two-state regime switching autoregressive model with an application to river flow analysis"*, J. Stat. Planning and Inference, Vol 137, No. 10, pp. 3113-3126.

Visualizing Gene Clusters Using Neighborhood Graphs in R

Theresa Scharl[1,2] and Friedrich Leisch[3]

[1] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Haupstraße 8-10/1071, 1040 Vienna, Austria, *theresa.scharl@ci.tuwien.ac.at*
[2] Department of Biotechnology, University of Natural Resources and Applied Life Sciences, Muthgasse 18, A-1190 Vienna, Austria
[3] Department of Statistics, University of Munich, Ludwigstraße 33, D-80539 München, Germany, *friedrich.leisch@stat.uni-muenchen.de*

Abstract. The visualization of cluster solutions in gene expression data analysis gives practitioners an understanding of the cluster structure of their data and makes it easier to interpret the cluster results. Neighborhood graphs allow for visual assessment of relationships between adjacent clusters. The number of clusters in gene expression data is for biological reasons rather large. As a linear projection of the data into 2 dimensions does not scale well in the number of clusters there is a need for new visualization techniques using non–linear arrangement of the clusters. The new visualization tool is implemented in the open source statistical computing environment R. It is demonstrated on microarray data from yeast.

Keywords: cluster analysis, graphs, microarray data, R

1 Introduction

Gene expression microarray experiments yield large and complex multivariate datasets that consist of several thousands of genes at multiple states. A typical question during the analysis is to find groups of co–expressed genes in the data. Cluster analysis is commonly used to reduce the complexity of the data from multidimensional space to a single nominal variable, the cluster membership. In the analysis of microarray data clustering is used as vector quantization as no clear density clusters exist in the data. Genetic interactions are so complex that the definition of gene clusters is not clear. Additionally microarray data are very noisy and co–expressed genes can end up in different clusters. Therefore the set of genes is divided into artificial subsets where relationships between clusters play an important role.

Clusters of co–expressed genes can help to discover potentially co–regulated genes or association to conditions under investigation. Usually cluster analysis provides a good initial investigation of microarray data before actually focusing on functional subgroups of interest. In the literature numerous cluster algorithms for clustering gene expression data have been proposed. Besides traditional methods like hierarchical clustering, K–means, partitioning

around medoids (PAM, K–medoids) or self–organizing maps there are several algorithms dealing with time–course gene expression data (e.g., Heyer et al., 1999, De Smet et al., 2002, Ben-Dor et al., 1999).

The display of cluster solutions particularly for a large number of clusters is very important in exploratory data analysis. Visualization methods give practitioners an understanding of the relationships between segments of a partition and make it easier to interpret the cluster results. Neighborhood graphs (Leisch, 2006) can be used for visual assessment of the cluster structure of centroid–based cluster solutions. In this paper neighborhood graphs are used to display high–dimensional gene expression data which are usually separated into a lot of clusters (e.g., over 25 clusters in Heyer et al., 1999). A linear projection of the data into 2 dimensions using for example linear discriminant analysis (LDA) does not scale well in the number of clusters and the vast amount of information cannot be shown in the plane. In the following a new visualization technique is presented using non–linear arrangement of the clusters where the cluster structure can be displayed up to a very large number of clusters. The layout algorithms implemented in the open source graph visualization software Graphviz are used for non–linear arrangement of the clusters. The new visualization tool is currently available at the homepage of the first author (http://www.ci.tuwien.ac.at/∼scharl/Software/) and will be released as an R package (R Development Core Team, 2007, http://www.R-project.org) soon. The functionality is demonstrated on a publicly available data set from yeast.

2 Methods

2.1 Cluster algorithms

In this work we focus on centroid–based cluster algorithms like K–means and PAM or others where clusters can be represented by centroids (e.g., QT–Clust, Heyer et al., 1999). For a given data set $X_N = \{x_1, \ldots, x_N\}$ the distance between points x and y is given by $d(x, y)$, e.g., the Euclidean or absolute distance. $C_K = \{c_1, \ldots, c_K\}$ is a set of centroids and the centroid closest to x is denoted by

$$c(x) = \text{argmin}_{c \in C_K} d(x, c).$$

The set of all points where c_k is the closest centroid is given by

$$A_k = \{x_n | c(x_n) = c_k\}.$$

Minimizing the average distance between each data point and its closest centroid

$$D(X_n, C_K) = \frac{1}{N} \sum_{n=1}^{N} d(x_n, c(x_n)) \rightarrow \min_{C_K}$$

is the task of most cluster algorithms.

2.2 Neighborhood graphs

Neighborhood graphs (Leisch, 2006) use the idea of topology–representing networks (TRNs, Martinetz and Schulten, 1994) to count the number of data points a pair of centroids is closest and second–closest. In TRNs the counts are used as weights for the edges of the graph. Silhouette plots (Rousseeuw, 1987) are diagnostic plots revealing the goodness of a partition. The distance from each point to the points in its own cluster is compared to the distance to points in the second closest cluster. The larger the silhouette values the better a cluster is separated from the other clusters. But silhouette plots do not show the proximity of clusters. They only give an indicator how well-separated single points are from other clusters. Neighborhood graphs combine these two approaches and use the mean relative distances as edge weights in order to measure how separated pairs of clusters are. Hence they display the distance between clusters. In the graph each node corresponds to a cluster centroid and two nodes are connected by an edge if there exists at least one point that has these two as closest and second–closest centroid.

As described above the centroid closest to x is denoted by $c(x)$ and the second closest centroid to x is denoted by

$$\tilde{c}(x) = \text{argmin}_{c \in C_K \setminus \{c(x)\}} d(x, c).$$

Now the set of all points where c_i is the closest centroid and c_j is second–closest is given by

$$A_{ij} = \{x_n | c(x_n) = c_i, \tilde{c}(x_n) = c_j\}.$$

For each observation x we define

$$s(x) = \frac{2d(x, c(x))}{d(x, c(x)) + d(x, \tilde{c}(x))}.$$

$s(x)$ is small if x is close to its cluster centroid and close to 1 if it is almost equidistant between the two cluster centroids. The average s–value of all points where cluster i is closest and cluster j is second closest can be used as a proximity measure between clusters and as edge weight in the graph.

$$s_{ij} = \begin{cases} |A_i|^{-1} \sum_{x \in |A_{ij}|} s(x), & A_{ij} \neq \emptyset \\ 0, & A_{ij} = \emptyset \end{cases}$$

$|A_i|$ is used in the denominator instead of $|A_{ij}|$ to make sure that a small set A_{ij} consisting only of badly clustered points with large shadow values does not induce large cluster similarity.

3 Data

In this work a publicly available dataset from yeast was investigated, the seventeen time point mitotic cell cycle data (Cho et al., 1998) available at

http://genome-www.stanford.edu. The dataset was preprocessed adapting the instructions given by Heyer et al. (1999). The outlier time points 10 and 11 were removed from the original 17 variables. The gene vectors were standardized to have median 0 and MAD 1. Finally genes that were either expressed at very low levels or did not vary significantly over the time points were removed. This procedure yields gene expression data on $N = 2832$ genes (observations) for $T = 15$ time points (variables). The data was clustered using the K–means algorithm. In this example 15 clusters were selected.

4 Software and implementation

All cluster algorithms and visualization methods used are implemented in the statistical computing environment R. R package `flexclust` (Leisch, 2006) is a flexible toolbox to investigate the influence of distance measures and cluster algorithms. It contains extensible implementations of the K–centroids and QT–Clust algorithm and offers the possibility to try out a variety of distance or similarity measures as cluster algorithms are treated separately from distance measures. New distance measures and centroid computations can easily be incorporated into cluster procedures. The default plotting method for cluster solutions in `flexclust` is the neighborhood graph.

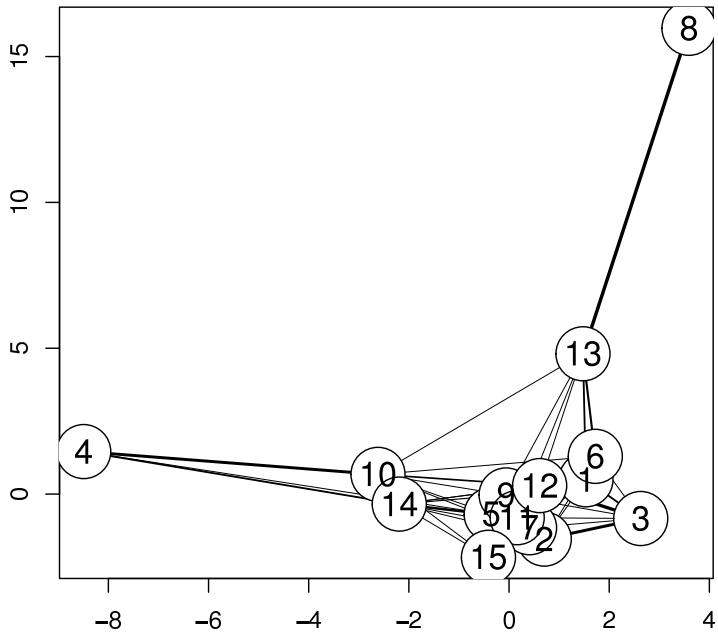


Fig. 1. Projection of neighborhood graph of a K–means cluster solution of the yeast data into 2 dimensions.

The visualization of partitioning cluster solutions is commonly accomplished by linear projection of the data into 2 dimensions using for example LDA. In Figure 1 the best possible separation using LDA is shown. The relationships between the centroids of 15 clusters can hardly be displayed in the plane. Therefore a new visualization tool is presented using non-linear arrangement of the nodes. Infrastructure for creating, manipulating, and visualizing graphs is provided in Bioconductor (Gentleman et al., 2005, http://www.bioconductor.org) package `graph`. The package contains functionality for data structure, classes and methods to manipulate graphs and enables efficient representations of very large graphs. An interface to the open source graph visualization software Graphviz (http://www.graphviz.org/) is provided in Bioconductor package `Rgraphviz` which returns the layout information for a graph object, x- and y–coordinates of the graph's nodes as well as the parameterization of the trajectories of the edges. Several layout algorithms can be chosen.

dot: hierarchical layout algorithm for directed graphs
neato and fdp: layout algorithms for large undirected graphs
twopi: radial layout
circo: circular layout

Additionally global and local properties (e.g., labels, shape, color, ...) can be assigned to both nodes and edges.

Using non–linear arrangement of clusters clearly improves the visualization of the cluster solution (Figure 2). In the new visualization method related clusters are not forced to lie next to each other. For example cluster 10 located at the bottom end of the graph is related to cluster 12 located at the right end of the graph. Additionally the graph is simplified by only drawing edges between nodes if the similarity of a cluster to another cluster is at least 10%. In Figure 2 the cluster structure of the cluster solution can easily be investigated. Cluster 15 is very different from the remaining clusters as no edge is drawn to cluster 15 and the similarities to connected clusters are very small. This indicates that the genes in cluster 15 are very different from the remaining genes. As cluster 4 is similar to clusters 10 and 14 which are also strongly connected the three clusters seem to be highly related.

4.1 Node methods

In the simple visualization of a neighborhood graph one single kind of node symbol is used for all nodes. By just looking at the graph no information about the different clusters is revealed. There are several possibilities how to include additional information in the representation of nodes. The most simple method is to use color coding, e.g., to color nodes by size or tightness of the corresponding clusters. Another possibility is to use different shapes or symbols for nodes representing clusters with specific properties.

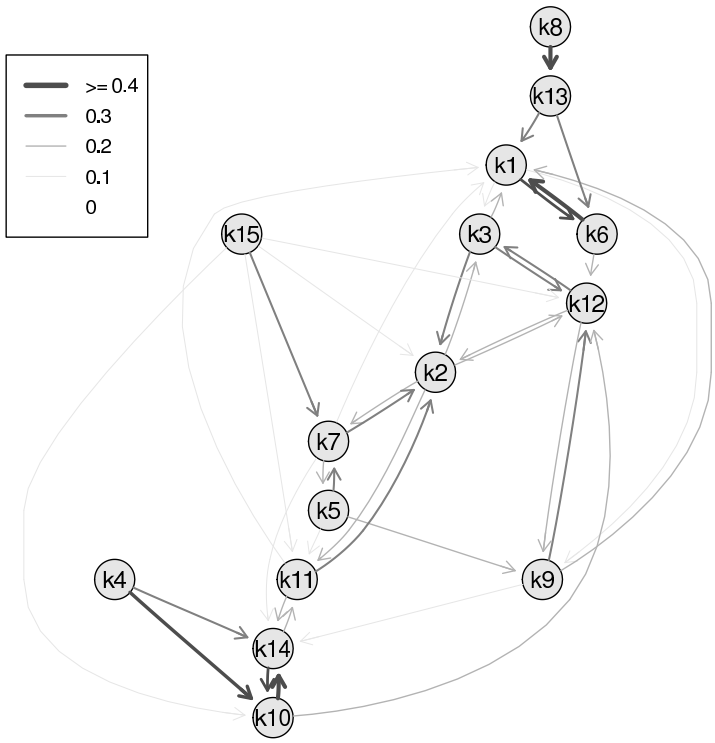


Fig. 2. Neighborhood graph of a K–means cluster solution for the yeast data using the *dot* layout algorithm.

In Figure 3 the cluster solution is shown with tight clusters highlighted, i.e., clusters with small average distance of the genes to the cluster centroid. In this example the tightest clusters are numbers 15 and 8 indicating genes very different from the rest of the genes (colored darkgrey) and numbers 13 and 4 (lightgrey).

4.2 Other features

The new visualization tool offers various possibilities for the analysis of microarray data which cannot be shown here due to space constrains. The neighborhood graph is a directed graph as the similarity of cluster 1 to cluster 2 is different from the similarity of cluster 2 to cluster 1. Besides plotting the original directed graph there are several possibilities how to plot edges taking into account for instance the mean, minimum or maximum of the similarities between two clusters.

The neighborhood graph is implemented in an interactive way and gene clusters can be investigated by clicking on the nodes. Plots of the expression profiles of the corresponding genes pop up as well as tables giving further

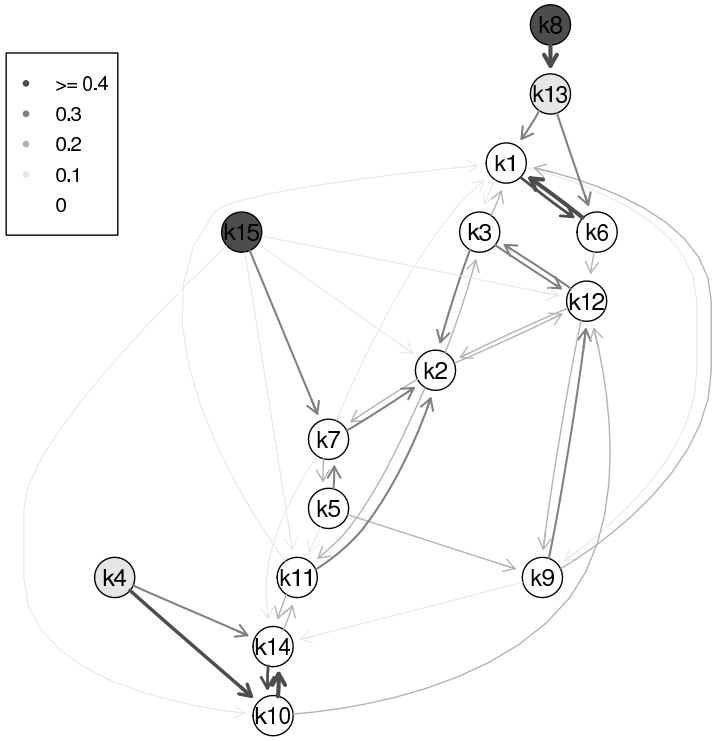


Fig. 3. The neighborhood graph with tight clusters highlighted.

information about the genes. Additional information about the gene clusters can be included in the neighborhood graph. External information from differential expression analysis or functional grouping can easily be included in the node representation, e.g., the accumulation of gene ontology (GO) categories in certain gene clusters.

5 Summary

Cluster analysis is commonly used in the analysis of gene expression data to find groups of co–expressed genes. But the definition of gene clusters is not very clear as genetic interactions are extremely complex. For this reason the relationship between clusters is very important as co–expressed genes can end up in different clusters. The neighborhood graph is a useful tool to visualize the underlying cluster structure. The visualization of partitioning cluster solutions is commonly accomplished by linear projection of the data into 2 dimensions. However, this is not recommended for high–dimensional data like microarray data and a large number of clusters. In our new visualization method layout algorithms for non–linear arrangement of the clusters

are used to display the relationships between clusters. Our interactive software tools for the analysis of gene expression data is very helpful not only for statisticians but also for practitioners. It was motivated by data of our cooperation partners at the University of Natural Resources and Applied Life Sciences in Vienna and is currently used to extract useful information for the further advancement of bioprocessing.

Acknowledgement

This work was supported by the Austrian K_{ind}/K_{net} Center of Biopharmaceutical Technology (ACBT).

References

BEN-DOR, A., SHAMIR, R. and YAKHINI, Z. (1999): Clustering gene expression patterns. *Journal of Computational Biology, 6 (3–4), 281–297.*

BICKEL, D.R. (2003): Robust cluster analysis of microarray gene expression data with the number of clusters determined biologically. *Bioinformatics, 19 (7), 818–824.*

CAREY, V.J., GENTLEMAN, R., HUBER, W. and GENTRY, J. (2005): Bioconductor Software for Graphs. In: R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry and S. Dudoit (Eds.): *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

CHO, R.J., CAMPBELLl, M.J., WINZELER E.A., STEINMETZ, L., CONWAY, A., WODICKA, L., WOLFSBERG, T.G., GABRIELIAN, A.E., LANDSMAN, D., LOCKHART, D.J. and DAVIS, R.W. (1998): A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell, 2/1, 65-73.*

DE SMET, F., MATHYS, J., MARCHAL, K., THIJS, G., DE MOOR, B. and MOREAU, Y. (2002): Adaptive quality-based clustering of gene expression profiles. *Bioinformatics, 18 (5), 735–746.*

GENTLEMAN, R., CAREY, V.J., HUBER, W., IRIZARRY, R.A. and DUDOIT, S. (Eds.) (2005): *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

HEYER, L.J., KRUGLYAK, S. and YOOSEPH, S. (1999): Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research, 9, 1106-1115.*

LEISCH, F. (2006): A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis 51 (2), 526–544.*

MARTINETZ, T. and SCHULTEN, K. (1994): Topology representing networks. *Neural Networks, 7 (3), 507–522.*

R DEVELOPMENT CORE TEAM (2007): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

ROUSSEEUW, P.J. (1987): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20, 53–65.*

Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance

Carolin Strobl[1] and Achim Zeileis[2]

[1] Department of Statistics, Ludwig-Maximilians-Universität München
Ludwigstraße 33, D-80539 München, Germany,
Carolin.Strobl@stat.uni-muenchen.de
[2] Department of Statistics and Mathematics, Wirtschaftsuniversität Wien
Augasse 2–6, A-1090 Wien, Austria, *Achim.Zeileis@wu-wien.ac.at*

Abstract. Random forests have become a widely-used predictive model in many scientific disciplines within the past few years. Additionally, they are increasingly popular for assessing variable importance, e.g., in genetics and bioinformatics. We highlight both advantages and limitations of different variable importance scores and associated testing procedures. For the test of Breiman and Cutler (2008), we investigate the statistical properties and find that the power of the test depends both on the sample size and the number of trees in an undesirable way that nullifies any significance judgments. Moreover, the specification of the null hypothesis of this test is discussed in the context of correlated predictor variables.

Keywords: feature selection, variable importance, permutation tests

1 Introduction

Within the past few years, random forests (Breiman (2001)) have become a popular and widely-used tool for non-parametric regression in many scientific areas such as genetics, bioinformatics, clinical medicine and psychology. Random forests are typically found to have high predictive accuracy and are applicable even in high dimensional problems, as well as problems involving correlated predictor variables and high-order interactions. Recently, their variable importance measures have also been suggested for the selection of relevant predictor variables in the analysis of microarray data, DNA sequencing and many other applications (cf. e.g., Lunetta et al. (2004), Arun and Langmead (2005), Bureau et al. (2005), Huang et al. (2005), Diaz-Uriarte and Alvarez de Andrés (2006), Qi et al. (2006) ,Ward et al. (2006)). Most random forest implementations offer two different variable importance measures (plus class-wise versions of the latter): the Gini importance, based on the Gini gain split selection criterion, and the permutation accuracy importance. However, Strobl et al. (2007) show that, when predictor variables vary in their scale of measurement or their number of categories, the Gini importance is biased

in favor of, e.g., predictor variables with many categories. As opposed to that, the permutation importance is reliable when the ensembles of trees are built on subsamples drawn without replacement instead of bootstrap samples drawn with replacement (Strobl et al. (2007)). Therefore, in the following we will only consider the permutation importance.

A key advantage of the random forest permutation importance, as compared to univariate screening methods, is that it covers the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables. For example, Lunetta et al. (2004) find that genetic markers relevant in interactions with other markers or environmental variables can be detected more efficiently by means of random forests than by means of univariate screening methods like Fisher's exact test. Random forests can also be applied when predictor variables are highly correlated.

Currently, most applications of the random forest permutation importance rely on a merely descriptive ranking of the potential predictor variables with respect to their importance: The few top-ranked predictors are selected for further exploration, where the number of selected variables is chosen arbitrarily or with respect to subject matter. A different approach for variable selection with random forests is introduced by Diaz-Uriarte and Alvarez de Andrés (2006), who suggest a backward elimination strategy based on the variable importance scores that takes under consideration the prediction accuracy: The underlying rationale is that the prediction accuracy will remain almost constant when irrelevant predictor variables are excluded, while it drops when relevant ones are excluded.

While in statistical modelling the aim may often be to select a model as sparse as possible, it is of equal interest in many applied sciences to be able to identify *all* predictor variables that are associated with the response, even if some of them are correlated. The question of interest here is to decide for each variable whether or not its importance is significantly greater than zero. A statistical test for this question is suggested by Breiman and Cutler (2008). At first sight it looks like this test could aid the decision which or how many of the top-ranked variables have significant importance and can be considered relevant. However, in the following we will present statistical reasoning and simulation results illustrating that the suggested test is not appropriate for statements of significance. Moreover, we will explore the unclear null hypothesis of the suggested test and give an outlook on a new permutation scheme for variable importance in random forests that better represents the null hypothesis of zero importance of a given variable.

2 Testing random forest variable importance

The rationale of the random forest permutation accuracy importance is the following: By randomly permuting the predictor variable X_j, its original association with the response Y is broken. When the permuted variable X_j,

together with the remaining non-permuted predictor variables, is used to predict the response for the out-of-bag observations, the prediction accuracy (i.e. the number of observations classified correctly) decreases substantially if the original variable X_j was associated with the response. Thus, a reasonable measure for variable importance is the difference in prediction accuracy before and after permuting X_j, averaged over all trees:

Let $\overline{\mathfrak{B}}^{(t)}$ be the out-of-bag sample for a tree t, with $t \in \{1, \ldots, ntree\}$. Then the variable importance for one tree is

$$VI^{(t)}(\mathbf{x}_j) = \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|} - \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_{i,\pi_j}^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|}$$

where $\hat{y}_i^{(t)} = f^{(t)}(\mathbf{x}_i)$ is the predicted classes for observation i before and $\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(\mathbf{x}_{i,\pi_j})$ is the predicted classes for observation i after permuting its value of variable j, i.e. with $\mathbf{x}_{i,\pi_j} = (x_{i,1}, \ldots, x_{i,j-1},\ x_{\pi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p})$. (Note that $VI^{(t)}(\mathbf{x}_j) = 0$ by definition, if variable X_j is not in tree t.) The raw variable importance score for each variable is then computed as the mean importance over all trees:

$$VI(\mathbf{x}_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(\mathbf{x}_j)}{ntree}$$

Because the individual importance scores $VI^{(t)}(\mathbf{x}_j)$ are computed from *ntree* independent bootstrap samples, a simple test for the relevance of variable X_j can be constructed based on the central limit theorem for the mean importance $VI(\mathbf{x}_j)$. If each individual variable importance $VI^{(t)}$ has standard deviation σ, the mean importance from *ntree* replications has standard error $\sigma/\sqrt{ntree}$. Therefore, under the null hypothesis of zero variable importance, the z-score

$$\widetilde{VI}(\mathbf{x}_j) = \frac{VI(\mathbf{x}_j)}{\frac{\hat{\sigma}}{\sqrt{ntree}}}$$

is asymptotically standard normal. Hence, when the z-score $\widetilde{VI}(\mathbf{x}_j)$ exceeds the α-quantile of the standard normal distribution, the null hypothesis of zero importance for variable X_j is rejected. This approach has been suggested by Breiman and Cutler (2008) for testing the variable importance. Note, however, that in the computation of the z-score averaging and scaling is not conducted with respect to the sample size n but to the number of trees in the ensemble *ntree* (cf. also Lunetta et al. (2004)).

2.1 Investigating the power of the current test

To investigate the power of the test suggested by Breiman and Cutler (2008), that is outlined in the previous section, a simulation study was conducted.

The experimental parameters that were varied are (a) the relevance of the predictor variable, (b) the sample size, and (c) the number of trees in the forest. For each combination of experimental parameters 1000 replications were run. In each replication a data set with the respective relevance and sample size was generated, a random forest with the respective number of trees was fit to the data, and the z-score was computed as described in the previous section. The test decision, i.e. whether or not the null hypothesis was rejected, was stored in every replication. The relative frequency of rejections of the null hypothesis (out of the 1000 replications) serves as an estimator for the power of the test in each combination of experimental parameters. In Figure 1 the empirical power is displayed as a function of the experimental parameters.

For a deeper understanding of the underlying mechanism we also display the curves for the unstandardized mean importance VI, the standard error of the mean and the z-score $\widetilde{VI}$ (all averaged over 1000 replications). In each iteration, a data set of sample size $n = 100, 200$ or 500 is generated that includes five predictor variables of which only one binary variable is relevant. Within the categories of this variable the binary response class is sampled from a binomial distribution with class probability $0.5 \pm \rho$, where ρ is the relevance parameter ($\rho = 0, 0.05, \ldots, 0.5$) as indicated on the abscissas of Figure 1. The parameter settings for the random forests were given by the varying number of trees ($ntree = 100, 200$ or 500) and a fixed number of two preselected variables per split. The simulation was conducted with the function *randomForest* (from the package of the same name by Breiman et al. (2007), Liaw and Wiener (2002) give an introduction), which is the reference implementation of random forests in the R system for statistical computing (R Development Core Team (2007)).

As depicted in the bottom row of Figure 1 the power of the test against the null hypothesis of zero importance shows the following irritating behavior: The power does increase with the relevance of the predictor variable as expected for any reasonable power curve. However, the power also does increase with the number of trees in the forest (the curves are shifted to the left, resulting in higher power for low relevance values), meaning that the power here depends on a tuning parameter that can be arbitrarily chosen by the user. This effect is due to the construction of the test statistic where, unlike in the standard test for the mean under normality, averaging and scaling is not with respect to a given sample size n but to the number of trees as outlined above. Even more dramatically, we find that the power does depend on the sample size—however not as expected for any reasonable test, where the power is supposed to increase with increasing sample size, but to the contrary: For large sample sizes (as compared to the number of trees) the power is zero.

To explore in more detail the mechanism responsible for this odd behavior we will follow the construction of the z-score, that is derived from the mean

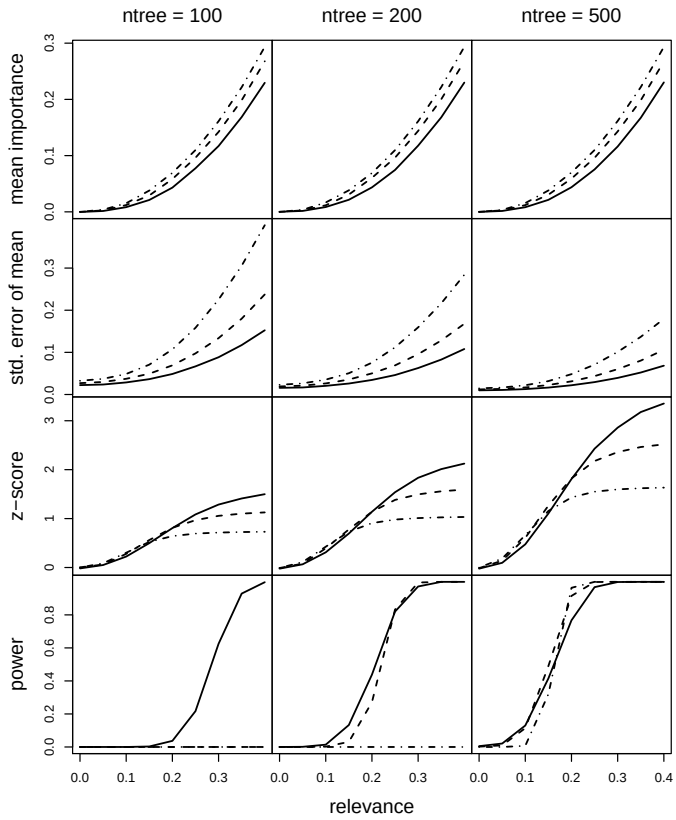


Fig. 1. Mean variable importance, standard error of mean, z-score and power as functions of relevance for sample size 100 (solid), 200 (dashed), and 500 (dash-dotted) and different numbers of trees.

importance by division through the standard error of the mean. The top row of Figure 1 shows that the unstandardized mean importance VI for one predictor variable increases with the relevance of the predictor variable and with the sample size as expected. There is no effect of the number of trees on the average importance—at least not when the number of trees is chosen sufficiently large to guarantee a stable estimate of the mean importance. This increase in the relevance and the sample size is desirable and exactly what we would have expected for any statistic to be employed in a test against the null hypothesis of zero importance. Therefore, the standard error of the mean, which is used for scaling, must be responsible for the odd behavior of the z-scores: The numerator of the fraction for the standard error of the

mean, the standard deviation, also increases with the relevance and with the sample size, and does not depend on the number of trees either. (The increase in the sample size is due to the resulting increase in the out-of-bag sample size that again extends the range of possible changes in the prediction accuracy induced by permuting the predictor variable. The dependence on the relevance is caused by a mechanism in the tree-building process: In many trees of the ensemble a variable with a low relevance may not be included at all, and produce an importance score of exactly zero, which diminishes the variation of the importance.) As a result of the division by the square root of the number of trees, however, an additional dependence on the number of trees is induced in the standard error of the mean, such that it decreases in the number of trees as depicted in the second row of Figure 1. Note also that the curves for the different sample sizes vary more strongly for the standard error of the mean than for the mean importance.

When finally the z-score is computed by means of standardizing the mean importance with the standard error of the mean, the rationale of this standardization is to account for the fact that the mean importance is an average over all trees in the ensemble—it does, however, not account for the effect of the sample size. The fact that the dependence of the mean importance on the sample size is less pronounced than that of its standard error causes an inversion of the importance pattern with respect to the sample size in the z-scores: We find in the third row of Figure 1 that the z-score decreases in the sample size but increases with the number of trees. This finally leads to the pattern for the power curves that we found in the bottom row of Figure 1: Only for high numbers of trees the overall level of the scaled importance is high enough for all sample sizes to ever reject the null hypothesis, while for lower numbers of trees the curves for the high sample sizes never exceed the threshold for rejecting the null hypothesis and result in a power of zero. This behavior is undesired and is an artefact of the scaling, that induces a dependence on the number of trees but at the same time inverts the dependence on the sample size. We therefore summarize the results of our simulation study that the mean variable importance VI shows the increase in the relevance and sample size that would be desired for a test for the null hypothesis of zero importance, while the scaled variable importance and the resulting test behave oddly.

2.2 Specifying the null hypothesis

Another issue when considering the test for the random forest permutation importance suggested by Breiman and Cutler (2008) is the very fundamental question: Exactly what null hypothesis is being tested? In the previous sections for simplicity we referred to the null hypothesis as "importance equal to zero". This implies some kind of independence between the predictor variable whose importance is being tested and the response. However, it is unclear

what kind of independence is being tested. The currently employed permutation scheme, where only the values of the variable of interest are permuted while the values of the response variable and the other predictors are held constant, does mimic the elimination of the predictor variable when predicting the response—however, at the same time it destroys all correlations between the variable of interest and the other covariates. Unlike standard permutation test of the global null hypothesis that the response is not correlated with any of the predictor variables, where the response is permuted against the complete predictor matrix and all associations within the predictor matrix are retained, the current random forest approach tests the rather unintuitive null hypothesis that the predictor of interest is not correlated with either one of the response or covariates. In cases where predictor variables may be correlated this permutation scheme might not reflect the actual null hypothesis of interest. This topic is investigated in more detail and a new, conditional permutation importance measure is suggested in Strobl et al. (2008).

3 Conclusion and outlook

We conclude that, in principle, a test for the random forest permutation importance could help identify relevant predictor variables. However, the results of our simulation studies also show that, in its current form, the test of Breiman and Cutler (2008) has prohibitively undesirable properties: The power of the test does not increase with the sample size, as would be expected for any reasonable statistical test, but rather remains zero for large sample sizes as compared to the number of trees. On the other hand the power does increase with the number of trees, which is a parameter that can be arbitrarily chosen by the user. This means that any statement of significance made with the current test for random forest variable importance is nullified.

Another issue, that is relevant in the context of correlated predictor variables, is the question whether the null hypothesis that is being tested in the current test is the one that reflects our understanding of the impact of a predictor variable on the response. A conditional permutation scheme that better reflects the null hypothesis of interest is suggested in Strobl et al. (2008).

Further research will address the issue of an adequate test statistic and rejection area for this null hypothesis. For high numbers of variables multiple testing issues will also have to be taken into consideration.

References

ARUN, K. and LANGMEAD, C.J. (2006): Structure based chemical shift prediction using random forests non-linear regression. In: T. Jiang, U. C. Yang, Y.-P. P. Chen and L. Wong (Eds.), *Proceedings of the Fourth Asia-Pacific Bioinformatics Conference*, Taipei, Taiwan, 317–326.

BREIMAN, L. (2001): Random forests. *Machine Learning 45 (1), 5–32*.

BREIMAN, L. and CUTLER, A. (2008): Random forests – Classification manual. URL `http://www.math.usu.edu/~adele/forests/`.

BREIMAN, L., CUTLER, A., LIAW, A. and WIENER, M. (2007): Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.5-22. URL `http://CRAN.R-project.org/`.

BUREAU, A., DUPUIS, J., FALLS, K., LUNETTA, K.L., HAYWARD, B., KEITH, T.P. and EERDEWEGH, P. V. (2005): Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology 28 (2), 171–182*.

DIAZ-URIARTE, R. and ALVAREZ DE ANDRES, S. (2006): Gene selection and classification of microarray data using random forest. *BMC Bioinformatics 7:3*.

HUANG, X., PAN, W., GRINDLE, S., HAN, X., CHEN, Y., PARK, S.J., MILLER, L. W. and HALL, J. (2005): A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics 6:205*.

LIAW, A. and WIENER, M. (2002): Classification and regression by randomForest. *R News 2 (3), 18–22*. URL `http://CRAN.R-project.org/ doc/Rnews/`.

LUNETTA, K.L., HAYWARD, L.B., SEGAL and J., EERDEWEGH, P.V. (2004): Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics 5:32*.

QI, Y., BAR-JOSEPH, Z. and KLEIN-SEETHARAMAN, J. (2006): Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins 63 (3), 490–500*.

R DEVELOPMENT CORE TEAM (2007): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

STROBL, C., BOULESTEIX, A.-L., ZEILEIS, A. and HOTHORN, T. (2007): Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8:25*.

STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T. and ZEILEIS, A. (2008): Conditional variable importance for random forests. *Technical Report 23*, Department of Statistics, Ludwig-Maximilians-Universität München, URL `http://epub.ub.uni-muenchen.de/2821/`.

WARD, M. m prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests. *Arthritis and Rheumatism 55 (1), 74–80*.

ZOU, H. and HASTIE, T. (2005): Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B 67 (2), 301–320*.

Part II

Categorical Data Analysis

Comparison of Mixture and Classification Maximum Likelihood Approaches in Poisson Regression Models

Susana Faria[1] and Gilda Soromenho[2]

[1] Department of Mathematics for Science and Technology, Research Centre Officina Mathematica, University of Minho, 4800-058 Guimarães, Portugal *sfaria@mct.uminho.pt*

[2] Faculty of Psychology and Sciences of Education, Research LEAD, University of Lisbon. Portugal *soromenhop@sapo.pt*

Abstract. In this work, we propose to compare two algorithms to compute maximum likelihood estimates of the parameters of a mixture Poisson regression models. To estimate these parameters, we may use the EM algorithm in a mixture approach or the CEM algorithm in a classification approach. The comparison of the two procedures was done through a simulation study of the performance of these approaches on simulated data sets in a target number of iterations. Simulation results show that the CEM algorithm is a good alternative to the EM algorithm for fitting Poisson mixture regression models, having the advantage of converging more quickly.

Keywords: maximum likelihood estimation, EM algorithm, classification EM algorithm, mixture Poisson regression models, simulation study

1 Introduction

Finite mixture models are a well-known method for modelling unobserved heterogeneity (see e.g. McLachlan et al. (2000) and Fruhwirth-Schnatter (2006) for a review). The study of these models is a well-established and active area of statistical research and mixtures of regressions have also been studied fairly extensively.

In this work, we study the procedure for fitting Poisson mixture regression models, which are commonly used to analyze heterogeneous count data (see Wedel et al. (1993)), by means of maximum likelihood. We apply two maximization algorithms to obtain the maximum likelihood estimates: the Expectation Maximization (EM) algorithm (see Dempster et al.(1977)) and the Classification Expectation Maximization (CEM) algorithm (see Celeux et al. (1992)).

The comparison of these two different approaches in a cluster analysis is well known in the mixture models literature (see Celeux et al. (1993) and Govaert et al. (1996)). Our goal is to compare the performance of these two approaches using samples drawn from mixtures of Poisson regression model.

The paper is organized as follows: in Section 2, we present the Poisson mixture regression models and the two maximization algorithms to obtain the maximum likelihood estimates. Section 3 provides a simulation study investigating the performance of the algorithms for fitting two and three component mixtures of Poisson regression models. In Section 4 the conclusions of our study are drawn.

2 Poisson mixture regression models

Let Y_i denote the $i-$th response variable, observed in reaction to a covariate $\mathrm{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$. It is assumed that the marginal distribution of Y_i follows a mixture of Poisson distributions,

$$Y_i \sim \sum_{j=1}^{J} \pi_j f_j(y_i|\lambda_{i|j}) \tag{1}$$

where

$$f_j(y_i|\lambda_{i|j}) = \frac{\exp^{-\lambda_{i|j}}(\lambda_{i|j})^{y_i}}{y_i!}, \;\; i = 1, \ldots, n, j = 1, \ldots, J \tag{2}$$

and $\lambda_{i|j} = exp(\mathrm{x}_i^T \beta_j)$, i.e., each mean depends on a set of covariates for the ith individual x_i^T and a vector of regression coefficients $\beta_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})$. The proportions π_j are the mixing probabilities ($0 < \pi_j < 1$, for all $j = 1, \ldots, J$ and $\sum_j \pi_j = 1$) and can be interpreted as the unconditional probabilities that an individual belongs to component j of the mixture.

Generally, the mixture parameters $\theta = (\pi_1, \ldots, \pi_J, \beta_1, \ldots, \beta_J)$ are estimated by maximizing the log-likelihood

$$L(\theta|\mathrm{x}_1, \ldots, \mathrm{x}_n, y_1, \ldots, y_n) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{J} \pi_j f_j(y_i|\lambda_{i|j}) \right) \tag{3}$$

2.1 The EM algorithm

The standard tool for finding maximum likelihood solution is the EM algorithm (Dempster et al. (1977)). In order to pose the procedure as an incomplete-data problem, an unobservable random vector z is introduced. For each observation y_i there is a corresponding $z_i = (z_{i1}, \ldots, z_{iJ})^T$ and its entries are all zero except for one, say z_{ij}, equal to unity indicating that y_i belongs to the jth component. The complete-data log-likelihood is given by

$$CL(\theta|\mathrm{x}_1, \ldots, \mathrm{x}_n, y_1, \ldots, y_n) = \sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} \log \left(\pi_j f_j(y_i|\lambda_{i|j}) \right) \tag{4}$$

The EM algorithm is easy to program and proceeds in two steps, E (for expectation) and M (for maximization).

2.2 The CEM algorithm

To fit Poisson mixture regression models, we also make use of a classification version of the EM algorithm, the so-called CEM algorithm. The CEM algorithm maximizes in θ and $z_1, \ldots, z_n$ the complete-data log-likelihood CL, where the missing component label z_i of each sample observation is included in the data set:

$$CL(\theta|z_1, \ldots, z_n, \mathrm{x}_1, \ldots, \mathrm{x}_n, y_1, \ldots, y_n) = \sum_{j=1}^{J} \sum_{\{i|z_i=j\}} \log\left(f_j(y_i|\lambda_{i|j})\right) \quad (5)$$

where $\{i|z_i = j\}$ is the set of observations arising from the jth mixture component.

The CEM algorithm incorpores a classification step (C-step) between the E- and M-steps of EM. This classification step involves assigning each observation to one of the J components.

3 Simulation study of algorithm performance

In order to compare the performance of the two algorithms in fitting Poisson mixture regression models, a simulation study was performed. The scope was limited to the study of two and three components. We used the freeware R to develop the simulation program.

3.1 Design of the study

In this study, the simulated data sets were generated according to the following factors:

Initial Conditions. In our simulation study, two different strategies of choosing initial values were considered. In the first strategy, the true values were used as the starting values. In the other strategy we ran the algorithm 20 times from random initial position and we selected the solution out of 20 runs which provided the best value of the optimized criterion (see Celeux et al. (1993)).

Stopping Rules. A rather strict stopping criterion for the two algorithms was used: iterations were stopped when the relative change in log-likelihood was smaller than 10^{-40}.

Number of Samples. For each type of simulated data set, we generated 100 samples of given sample size $n = 100$ and $n = 500$.

Data set. Each datum (x_i, y_i) was generated by the following scheme. First, a uniform $[0, 1]$ random number c_i was generated and its value is used to select a particular component j from mixture of regressions model. Next, x_i was randomly generated from a uniform $[x_L, x_U]$ distribution and then we have $\lambda_j = exp(\beta_{j0} + \beta_{j1}\ x_i)$. Finally, we simulate the value y_i from the Poisson distribution $P(\lambda_j)$.

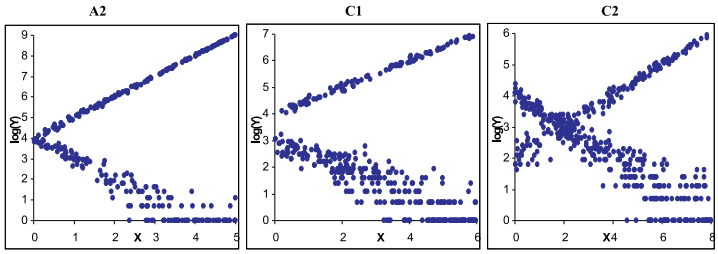


Fig. 1. Scatter plot of samples from 2 component models with $n = 500$.

Measures of Algorithm Performance: In order to examine the performance of two algorithms, the following criteria was used:

- the mean number of iterations required for convergence;
- the mean square error (MSE) of the parameter estimates over the 100 replications which is given by:

$$MSE(\hat{\theta}_j) = \frac{1}{100} \sum_{m=1}^{100} \left(\hat{\theta}_j^{(m)} - \theta_j \right)^2 \tag{6}$$

where $\theta_j = (\pi_j, \beta_j)$ and $\hat{\theta}_j^{(m)} = (\hat{\pi}_j^{(m)}, \hat{\beta}_j^{(m)})$, $j = 1, \ldots, J$.

The simulation process consists of the following steps:

1. Create a data set of size n.

2. Fit a mixture of Poisson regression model to the data using the EM and the CEM algorithms. Save the number of iterations required for convergence and the estimated parameters $\hat{\theta}$.

3. Repeat steps 1-2, for a total of 100 trials. Compute the mean number of iterations required for convergence and the mean square error (MSE) of the parameter estimates.

3.2 Simulation results: two component mixture of Poisson regressions

In our numerical experiments, for two component models ($J = 2$), we considered eight groups of different parameters $\theta = (\pi_1, \beta_{10}, \beta_{11}, \beta_{20}, \beta_{21})$, which generates different overlaps of the regression lines, studied in Yang et al. (2005). For illustration we show scatter plots of random samples of 500 points in Figure 1.

Samples of two different sizes n ($n = 100, 500$) were generated for each set of true parameter values $(\pi_1, \beta_{10}, \beta_{11}, \beta_{20}, \beta_{21})$ shown on Table 1.

Table 2 and 3 provide the MSE of the parameter estimates and the mean number of iterations required for convergence using the EM and CEM algorithm for fitting two component mixtures of Poisson regression models.

Cases	β_{10}	β_{11}	β_{20}	β_{21}	π_1	Cases	β_{10}	β_{11}	β_{20}	β_{21}	π_1	Cases	β_{10}	β_{11}	β_{20}	β_{21}	π_1
$A1$	3	-1	4	1	0.5	$B1$	4	-0.5	4	0.5	0.3	$C1$	3	-0.5	4	0.5	0.7
$A2$	4	-1	4	1	0.5	$B2$	4	-0.5	4	0.5	0.5	$C2$	4	-0.5	2	0.5	0.7
$A3$	5	-1	4	1	0.5	$B3$	4	-0.5	4	0.5	0.7						

Table 1. True parameter values for the essays with a two component mixture of Poisson regressions.

Cases	n	Algorithm	MSE					IT
			π_1	β_{10}	β_{11}	β_{20}	β_{21}	
A1		EM	2,41E-03	1,25E-02	8,28E-03	1,55E-04	5,76E-06	2,84
		CEM	2,41E-03	1,25E-02	8,28E-03	1,55E-04	5,76E-06	2,35
A2		EM	2,91E-03	2,05E-03	8,91E-04	1,81E-04	7,15E-06	9,17
		CEM	3,30E-03	1,89E-03	8,48E-04	1,78E-04	7,02E-06	5,09
A3		EM	2,95E-03	2,51E-03	1,24E-03	1,70E-04	6,98E-06	9,45
		CEM	3,34E-03	2,41E-03	1,22E-03	1,73E-04	7,12E-06	6,04
B1	100	EM	2,59E-03	6,65E-03	1,13E-03	4,75E-04	2,26E-05	13,04
		CEM	2,47E-03	7,62E-03	1,25E-03	4,64E-04	2,19E-05	5,45
B2		EM	2,76E-03	4,21E-03	7,80E-04	7,57E-04	3,61E-05	11,98
		CEM	3,32E-03	4,37E-03	7,24E-04	8,36E-04	4,02E-05	6,66
B3		EM	2,70E-03	1,84E-03	4,47E-04	1,19E-03	6,25E-05	10,75
		CEM	2,87E-03	1,89E-03	4,25E-04	1,24E-03	6,53E-05	6,86
C1		EM	2,17E-03	4,72E-03	8,66E-04	1,04E-03	5,25E-05	3,18
		CEM	2,17E-03	4,73E-03	8,65E-04	1,04E-03	5,25E-05	2,78
C2		EM	2,42E-03	1,92E-03	3,33E-04	9,08E-03	4,61E-04	12,31
		CEM	3,19E-03	1,97E-03	3,41E-04	8,64E-03	4,47E-04	6,52
A1		EM	4,84E-04	2,82E-03	1,27E-03	3,27E-05	1,28E-06	3,40
		CEM	4,84E-04	2,82E-03	1,27E-03	3,27E-05	1,28E-06	2,70
A2		EM	4,33E-04	1,14E-03	5,86E-04	3,34E-05	1,37E-06	9,64
		CEM	4,89E-04	1,50E-03	7,16E-04	3,35E-05	1,37E-06	5,60
A3		EM	5,66E-04	3,79E-04	2,19E-04	3,18E-05	1,17E-06	8,12
		CEM	6,58E-04	4,01E-04	2,27E-04	3,21E-05	1,17E-06	5,87
B1	500	EM	4,45E-04	1,31E-03	1,96E-04	8,72E-05	3,99E-06	12,04
		CEM	4,63E-04	2,05E-03	2,86E-04	9,45E-05	4,28E-06	6,40
B2		EM	4,61E-04	6,25E-04	1,18E-04	1,50E-04	6,90E-06	10,82
		CEM	6,01E-04	1,42E-03	1,89E-04	1,66E-04	7,55E-06	6,45
B3		EM	4,80E-04	4,08E-04	9,11E-05	2,50E-04	1,19E-05	9,63
		CEM	5,15E-04	5,61E-04	1,14E-04	2,28E-04	1,12E-05	6,10
C1		EM	4,65E-04	1,11E-03	2,23E-04	1,78E-04	9,34E-06	3,45
		CEM	4,65E-04	1,11E-03	2,23E-04	1,78E-04	9,34E-06	2,90
C2		EM	5,90E-04	4,48E-04	8,86E-05	2,05E-03	9,54E-05	11,35
		CEM	1,48E-03	4,52E-04	9,14E-05	2,15E-03	9,91E-05	6,57

Table 2. The mean number of iterations (IT) required for convergence and mean square error (MSE) of estimates based on 100 replications of the 2 component mixtures of Poisson regression models when the true values were used as the starting values.

In all cases, the mean number of iterations for convergence is smaller using the CEM algorithm rather than using the EM algorithm. It is evident that the EM and CEM estimates of the regression coefficients and the mixture proportion have relatively small MSE and, in generality, the MSE of EM estimates are slightly smaller than the MSE of CEM estimates. However, when the algorithms are initiated with the true parameter values and for $n = 100$, the CEM algorithm performs generally better. It also seems that the EM and CEM algorithm have practically the same behavior in situations where the overlap is small (A1, C1). The MSE of both estimates seems to depend strongly upon the mixing proportion value and the overlap of the regression models. Although only two sample sizes were considered, it seems that the MSE of both estimates tend to approach zero for greater samples.

Cases	n	Algorithm	MSE					IT
			π_1	β_{10}	β_{11}	β_{20}	β_{21}	
A1		EM	2,92E-03	9,07E-03	2,58E-03	1,28E-04	4,85E-06	3,81
		CEM	2,92E-03	9,07E-03	2,58E-03	1,28E-04	4,85E-06	3,73
A2		EM	2,70E-03	8,32E-03	3,67E-03	2,15E-04	7,99E-06	10,74
		CEM	2,82E-03	7,79E-03	3,44E-03	2,15E-04	8,04E-06	7,65
A3		EM	2,55E-03	1,48E-03	8,08E-04	1,34E-04	5,18E-06	11,38
		CEM	2,91E-03	2,12E-03	9,331E-04	1,34E-04	5,22E-06	8,65
B1	100	EM	1,93E-03	5,51E-03	8,51E-04	3,77E-04	1,81E-05	15,64
		CEM	2,02E-03	7,56E-03	9,55E-04	3,90E-04	1,84E-05	8,36
B2		EM	3,04E-03	3,76E-03	6,11E-04	6,54E-04	3,10E-05	12,54
		CEM	3,48E-03	3,93E-03	6,47E-04	7,51E-04	3,54E-05	7,90
B3		EM	2,01E-03	1,68E-03	6,33E-04	8,75E-04	4,39E-05	12,03
		CEM	2,05E-03	1,74E-03	6,81E-04	8,90E-04	4,61E-05	10,34
C1		EM	1,90E-03	4,86E-03	9,35E-04	9,67E-04	4,73E-05	4,42
		CEM	1,905E-03	4,86E-03	9,35E-04	9,67E-04	4,73E-05	4,38
C2		EM	2,72E-03	2,33E-03	4,88E-04	8,28E-03	3,99E-04	17,94
		CEM	3,82E-03	2,28E-03	4,81E-04	9,51E-03	4,46E-04	11,35
A1		EM	5,44E-04	2,98E-03	1,43E-03	3,36E-05	1,31E-06	4,48
		CEM	5,44E-04	2,98E-03	1,43E-03	3,36E-05	1,31E-06	4,18
A2		EM	4,33E-04	8,57E-04	5,28E-04	2,65E-04	1,16E-06	10,76
		CEM	4,63E-04	1,58E-03	6,09E-04	3,34E-05	1,29E-06	9,94
A3		EM	5,63E-04	3,23E-04	1,54E-04	3,89E-05	1,44E-06	10,32
		CEM	6,91E-04	3,38E-04	1,54E-04	3,81E-05	1,43E-06	9,51
B1	500	EM	6,28E-03	8,84E-04	1,82E-04	5,65E-04	4,69E-06	16,19
		CEM	3,82E-04	2,60E-03	3,75E-04	1,01E-04	4,87E-06	8,75
B2		EM	5,08E-04	7,49E-04	1,61E-04	2,30E-04	4,60E-06	12,11
		CEM	6,21E-04	1,37E-03	1,91E-04	1,19E-04	5,00E-06	10,42
B3		EM	3,51E-04	4,21E-04	8,35E-05	2,17E-04	1,02E-05	11,73
		CEM	4,32E-04	6,51E-04	1,05E-04	2,87E-04	1,12E-05	11,08
C1		EM	3,63E-04	1,52E-03	2,72E-04	1,94E-04	8,82E-06	5,18
		CEM	3,63E-04	1,50E-03	2,69E-04	1,93E-04	8,79E-06	5,00
C2		EM	5,67E-04	4,69E-04	6,61E-05	1,09E-03	2,99E-05	13,08
		CEM	1,08E-03	5,16E-04	7,12E-05	1,17E-03	3,15E-05	11,31

Table 3. The mean number of iterations (IT) required for convergence and mean square error (MSE) of estimates based on 100 replications of the 2 component mixtures of Poisson regression models when the second strategy was used as the starting values.

3.3 Simulation results: three component mixture of Poisson regressions

For three component models (J=3), samples of size $n = 100$ and $n = 500$ were generated for the five sets of parameter values (π, β) shown in Table 4. For illustration we show scatter plots of random samples of 500 points in Figure 2.

Cases	β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	π_1	π_2	Cases	β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	π_1	π_2
$D1$	3	-0.5	4	0.5	3	0.5	0.4	0.4	$D3$	4	-0.5	4	0.5	2	0.8	0.4	0.4
$D2$	4	-0.5	2	0.5	6	-0.5	0.4	0.3	$D4$	4	-0.5	4	0.5	2	0.8	0.4	0.3
									$D5$	4	-0.5	4	0.5	2	0.8	0.3	0.5

Table 4. True parameter values for the essays with a 3 component mixture of Poisson regressions.

Table 5 and 6 report the MSE of the parameter estimates and the mean number of iterations required for convergence using the EM and CEM algorithm for fitting three component mixtures of Poisson regression models. Also in all cases, the mean number of iterations for convergence is smaller using the CEM algorithm rather than using the EM algorithm. It is evident that the EM and CEM estimates of the regression coefficients and the mixture proportion have relatively small MSE, especially when the algorithms

	n	Alg	MSE									IT
			π_1	π_2	π_3	β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	
D1		EM	5,91E-05	6,53E-03	6,04E-04	1,32E-02	2,50E-03	2,76E-03	2,65E-03	7,61E-04	1,73E-03	15,28
		CEM	5,90E-05	6,97E-03	6,19E-04	1,50E-02	2,78E-03	2,89E-03	2,65E-03	7,61E-04	2,13E-03	7,02
D2		EM	1,65E-04	1,02E-03	1,86E-04	6,15E-03	2,83E-03	7,23E-04	2,91E-03	6,58E-03	2,79E-03	13,02
	100	CEM	1,74E-04	9,89E-04	1,75E-04	6,10E-03	3,40E-03	7,38E-04	3,86E-03	6,95E-03	3,42E-03	7,22
D3		EM	4,41E-05	6,94E-03	2,98E-04	6,21E-03	2,29E-03	1,10E-03	2,07E-03	8,62E-04	1,44E-03	14,49
		CEM	4,65E-05	6,91E-03	2,93E-04	6,46E-03	2,55E-03	9,83E-04	2,39E-03	9,06E-04	1,46E-03	7,10
D4		EM	4,26E-05	5,97E-03	2,41E-04	5,02E-03	2,71E-03	7,62E-04	1,28E-02	7,88E-04	1,27E-02	14,24
		CEM	4,27E-05	5,83E-03	2,37E-04	5,82E-03	3,50E-03	8,31E-04	1,32E-02	8,05E-04	1,48E-02	8,20
D5		EM	3,79E-05	5,71E-03	2,74E-04	6,73E-03	2,40E-03	1,12E-03	2,86E-03	6,81E-04	1,58E-03	17,45
		CEM	3,62E-05	5,83E-03	2,78E-04	7,63E-03	2,53E-03	1,18E-03	3,01E-03	6,55E-04	1,74E-03	6,93
D1		EM	5,29E-04	4,55E-04	3,52E-04	9,70E-04	1,63E-04	1,44E-04	7,12E-06	1,10E-03	4,38E-05	12,59
		CEM	7,96E-04	5,31E-04	5,20E-04	1,17E-03	1,78E-04	1,58E-04	7,39E-06	1,08E-03	4,51E-05	7,69
D2		EM	5,95E-04	4,47E-04	5,10E-04	8,70E-04	1,57E-04	1,91E-04	9,66E-06	9,70E-04	3,88E-05	12,71
	500	CEM	8,93E-04	4,79E-04	6,92E-04	1,34E-03	1,95E-04	2,22E-04	1,05E-05	9,74E-04	3,91E-05	7,95
D3		EM	4,14E-04	5,81E-04	4,42E-04	1,55E-03	2,47E-04	8,73E-05	4,92E-06	1,28E-03	5,07E-05	13,80
		CEM	4,85E-04	6,29E-04	5,15E-04	2,18E-03	3,23E-04	9,91E-05	5,43E-06	1,37E-03	5,37E-05	7,82
D4		EM	4,21E-04	5,37E-04	2,94E-04	1,85E-03	4,10E-04	1,12E-04	8,78E-06	1,16E-03	7,98E-05	12,06
		CEM	5,41E-04	5,376E-04	3,67E-04	3,23E-03	6,18E-04	1,13E-04	8,79E-06	1,78E-03	1,22E-04	6,48
D5		EM	4,01E-04	4,67E-04	4,59E-04	9,77E-04	1,28E-04	1,37E-03	3,21E-05	1,62E-04	2,28E-05	11,15
		CEM	6,76E-04	9,30E-04	5,60E-04	1,06E-03	1,41E-04	1,40E-03	3,30E-05	1,59E-04	2,24E-05	7,27

Table 5. The mean number of iterations (IT) required for convergence and mean square error (MSE) of estimates based on 100 replications of the 3 component mixtures of Poisson regression models when the true values were used as the starting values.

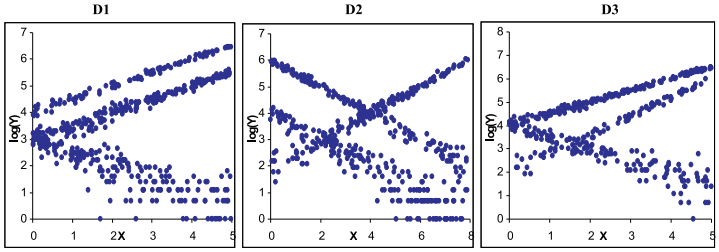


Fig. 2. Scatter plot of samples from 3 component models with $n = 500$.

are initiated with the true parameter values. In generality, EM outperforms CEM by producing estimates of the parameters that have smaller MSE. As in the case of the two component model, the MSE of both estimates seems to depend strongly upon the mixing proportion value and the overlap of the regression models. Although only two sample sizes were considered, it seems that the MSE of both estimates tend to approach zero for greater samples.

3.4 Conclusion

In this work, we compared the performance of two algorithms to compute maximum likelihood estimates of the parameters of a mixture Poisson regression models, the EM and the CEM algorithm.

We run a number of simulations, and in all of them the CEM algorithm converged in fewer iterations than the EM algorithm, which implies a reduction in the computational time to reach the parameter estimates.

Simulation results show that CEM algorithm is a good alternative to the EM algorithm for fitting Poisson mixture models, having the advantage of converging more quickly, and as the difference between the MSE of both

	n	Alg	MSE									IT
			π_1	π_2	π_3	β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	
D1		EM	2,34E-03	2,46E-03	1,84E-03	8,29E-03	1,71E-03	1,07E-03	9,02E-05	5,47E-03	3,81E-04	17,81
		CEM	2,57E-03	2,42E-03	2,12E-03	1,20E-02	1,95E-03	1,69E-03	3,84E-04	9,55E-03	4,43E-04	10,72
D2		EM	3,15E-03	3,57E-03	1,51E-03	8,53E-03	1,14E-03	2,60E-03	9,32E-03	1,26E-03	5,30E-03	21,56
	100	CEM	3,25E-03	5,15E-03	2,56E-03	4,34E-03	1,20E-03	9,55E-03	8,26E-03	7,05E-03	4,88E-03	13,55
D3		EM	2,80E-03	2,38E-03	1,56E-03	7,12E-04	1,32E-04	5,41E-04	3,98E-03	1,65E-03	9,02E-03	19,54
		CEM	3,51E-03	2,71E-03	1,74E-03	8,18E-04	5,51E-05	2,87E-04	5,57E-03	1,09E-03	6,83E-03	11,18
D4		EM	2,67E-03	1,10E-03	1,02E-03	8,50E-04	1,93E-04	7,40E-04	4,48E-03	1,86E-03	1,17E-03	19,73
		CEM	3,36E-03	1,16E-03	1,28E-03	9,44E-04	1,85E-04	1,05E-03	6,52E-03	1,97E-03	1,49E-03	11,63
D5		EM	1,06E-03	1,27E-03	1,36E-03	9,41E-04	1,08E-04	6,74E-04	4,38E-03	1,40E-03	1,07E-02	19,04
		CEM	1,14E-03	1,26E-03	1,26E-03	9,57E-04	5,72E-05	5,12E-04	4,93E-03	8,73E-04	8,34E-03	11,88
D1		EM	4,02E-04	5,04E-04	2,00E-03	5,92E-04	9,93E-05	2,70E-04	1,22E-05	2,90E-03	4,78E-05	17,10
		CEM	1,28E-03	1,23E-03	1,54E-02	2,85E-03	3,64E-04	3,53E-03	1,69E-05	2,03E-03	3,31E-04	11,65
D2		EM	4,97E-04	4,18E-04	4,80E-04	1,09E-03	1,93E-04	2,47E-04	1,33E-05	7,77E-04	3,26E-05	18,15
	500	CEM	9,43E-04	8,04E-04	9,98E-04	1,56E-03	2,30E-04	7,99E-04	1,22E-05	8,02E-04	3,38E-05	14,00
D3		EM	6,51E-04	4,72E-04	1,19E-03	1,50E-03	4,57E-04	2,02E-04	5,58E-06	8,89E-03	4,88E-05	21,23
		CEM	7,35E-04	1,79E-03	3,16E-03	2,11E-03	5,29E-04	4,16E-03	6,05E-05	1,17E-03	1,32E-05	10,49
D4		EM	4,39E-04	3,61E-04	3,33E-04	3,50E-03	5,68E-04	2,06E-04	1,40E-05	1,37E-03	9,57E-05	18,68
		CEM	4,89E-04	3,62E-04	4,06E-04	5,00E-03	9,01E-04	2,06E-04	1,39E-05	1,96E-03	1,29E-04	13,42
D5		EM	5,10E-04	4,35E-04	4,73E-04	1,21E-03	1,53E-04	1,27E-03	2,87E-05	1,95E-04	2,93E-05	22,65
		CEM	7,80E-04	9,68E-04	5,85E-04	1,25E-03	1,71E-04	1,29E-03	2,85E-05	1,96E-04	3,02E-05	15,60

Table 6. The mean number of iterations (IT) required for convergence and mean square error (MSE) of estimates based on 100 replications of the 3 component mixtures of Poisson regression models when the second strategy was used as the starting values.

approaches is so insignificant, so CEM algorithm seems to be preferred. This simulation study achieves basically the same conclusions already obtained on comparing EM and CEM algorithms for a mixture of linear gaussian regression models (see Faria et al. (2007))

References

CELEUX, G. and GOVAERT, G. (1992): A classification EM algorithm and two stochastic versions, *Computational Statistics & Data Analysis, 14, 315-332.*

CELEUX, G. and GOVAERT, G.(1993): Comparison of the mixture and the classification maximum likelihood in cluster analysis. *J. Statistical Computation and Simulation, 47, 127-146.*

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B.(1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B, 39, 1-38.*

FARIA, S. and SOROMENHO, G. (2007): Comparison of the mixture and the classification maximum likelihood in regression analysis. IN: *Proceedings of IASC07 - Statistics for Data Mining, Learning and Knowledge Extraction*, August.

FRUHWIRTH-SCHNATTER(2006): *Finite Mixture and Markov Switching Models*, Springer, Heidelberg.

GOVAERT, G. and NADIF, M. (1993): Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data. *Computational Statistics & Data Analysis, 23, 65-81.*

MCLACHLAN, G.J. and PEEL, D., (2000): *Finite Mixture Models*, Wiley, New York.

WEDEL, M., DESARBO, W.S., BULT, J.R. and RAMASWAMY, V.(1993): A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics, 8, 397 - 411.*

YANG, M.S. and LAI, C.Y. (2005): Mixture Poisson regression models for heterogenous count data based on latent and fuzzy class analysis. *Soft Computing, 9, 519-524*

ANOVA on the Textile Plot

Natsuhiko Kumasaka

Center for Genomic Medicine, RIKEN
4-6-1, Shirokane-dai, Minato-ku, Tokyo, 108-8639, Japan, *kumasaka@src.riken.jp*

Abstract. The textile plot is a parallel coordinate plot in which the ordering, locations and scales of the axes are simultaneously chosen so that the connecting lines, each of which represents a case, are aligned as horizontally as possible. This article shows the potential usefulness of the textile plot as an aid to the interpretation of the result of ANOVA for complete factorial design data. A practical example, the soldering data, is employed, which greatly aids the interpretation of the axis ordering, locations and scales of the textile plot in the context of ANOVA.

Keywords: complete factorial design data, horizontalisation criterion, parallel coordinate plot

1 Introduction

The complete factorial design data often arise in the fields of agriculture, biology, social sciences and manufacturing and other applications of engineering and the physical sciences. Such data consist of one or more responses which are made for changing values of several factors. ANOVA is a technique for modelling the variation in the response in the hope of understanding how it depends on the levels of the factors.

Several graphical representations have been developed for facilitating the interpretation of the result of ANOVA. Tukey's two-way display (Tukey (1977)) is designed to show the result of ANOVA for two way table with one observation per cell. That is to say, it can show the complete factorial design data with one response and two factors. As is described in Friendly et al. (2003), Biplot is also used as an alternative display for showing the result of ANOVA. The problem would be that these two displays are not easily extended to more than three factors.

For multiple factor data, the mean plot (Chambers and Hastie (1992)) is one of the simplest representations for giving a graphical summary of the relationship between the response and factors, showing the mean value of the response at each level of each factor. It suggests that useful information about the experiment sometimes can be seen without any formal modelling, particularly using graphical representations.

Parallel coordinate plots (Inselberg (1985), Wegman (1990)) would also be possible device with which to explore multiple factor data. The basic idea of the parallel coordinate plot is to place axes, representing each variable, in

parallel and its associated coordinates on adjacent axes are then connected by straight lines for a given data point observed in a high dimensional space. However, the problem of displaying multiple factor data on the parallel coordinate plot is that, the levels of each factor are usually assigned on the axis alphabetically at even intervals, and the assignment is not meaningful for understanding the phenomena behind the data.

Textile plot (Kumasaka and Shibata (2008)) improves the parallel coordinate plot so as to accentuate the differences between levels of factors based on the contribution to the response. This is due to the introduction of the horizontalisation criterion so that all connected lines are aligned as horizontally as possible. As a special bonus of introducing the criterion, ANOVA table can be seen on the textile plot for complete factorial design data. The result also provides another interpretation of the axis ordering of the textile plot, in which the further left axis has the bigger variation.

2 Example data

The soldering data (Chambers and Hastie (1992)) is employed as a practical example in this article because of its simplicity and familiarity, which is the results of an experiment varying five factors relevant to the wave-soldering procedure for mounting components on printed circuit boards. The response is a count of how many skips appeared to a visual inspection. The data consists of 720 observations of the response *Skips* in a balanced subset of all the experimental runs, with the corresponding values for five experimental factors:

Opening: ordered factor indicating the amount of clearance around the mounting pad (S < M < L);
Solder: ordered factor indicating the amount of solder used (Thin < Thick);
Mask: factor indicating which of 4 types of solder mask was used. The type and thickness of the material used for the mask were varied. The levels are A1.5, A3, B3, and B6;
PadType: factor indicating which of 10 mounting pads was used. The geometry and size of the mounting pad were varied. The levels are W4, D4, L4, D6, L6, D7, L7, L8, W9, and L9; and
Panel: factor indicating which of panels 1, 2 or 3 on a board is being counted.

Figure 1 shows a parallel coordinate plot of the data set. The first axis shows the number of *Skips*. The points on the axis are then connected to the intersection points on the other axes, which indicate levels of factor variables. Needless to say, it is hopeless to understand underlying phenomena of the data without any interactive methods or any design enhancements since all levels are uniformly connected by segments between two factors. The next section will describe an improvement of the parallel coordinate plot, called *Textile Plot* in conjunction with the horizontalisation criterion.

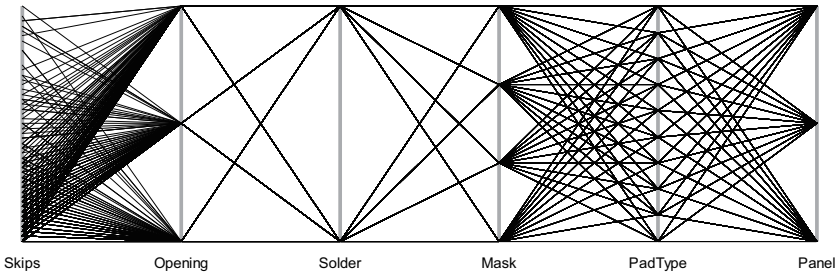


Fig. 1. Parallel coordinate plot for the solder balanced data.

3 Textile plot

Before showing the textile plot, we briefly review how to choose the locations and scales in the textile plot.

Let $\mathbf{x}_j$ denote the vector of n observations on variable j $(j = 1, \ldots, p)$. Then each row of the data matrix $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$ gives us a p-dimensional observation. If the data vectors $\mathbf{x}_j$ is numerical, then it is simply transformed into a coordinate vector

$$\mathbf{y}_j = \alpha_j \mathbf{1} + \beta_j \mathbf{x}_j \tag{1}$$

where $\mathbf{1}$ is a vector of ones, which results in a parallel coordinate plot with a common coordinate system. The vector $\mathbf{y}_j = (y_{1j}, \ldots, y_{nj})^T$ gives us the coordinates of the n observations on the jth axis.

If data vector $\mathbf{x}_j$ is a categorical data vector with q_j levels, the element of the coordinate vector $\mathbf{y}_j$ takes only q_j different values on the jth parallel coordinate axis. Then the coordinate vector in (1) can be re-defined as

$$\mathbf{y}_j = \alpha_j \mathbf{1} + \mathbf{X}_j \boldsymbol{\beta}_j \tag{2}$$

with $n \times (q_j - 1)$ data matrix $\mathbf{X}_j$ encoded by a set of contrasts (Chambers and Hastie (1992)) and $(q_j - 1)$-dimensional scale parameter vector $\boldsymbol{\beta}_j$. Thus, to cover cases where both numerical and categorical data vectors exist, we will use the matrix notation $\mathbf{X}_j$ in place of the numerical data vector $\mathbf{x}_j$ by letting $q_j = 2$.

For the texitle plot, the degree to which each connecting line on the parallel coordinate system is horizontal can be measured by the sum of squared deviations from a horizontal line at level ξ_i, that is $S_i^2 = \sum_{j=1}^{p} (y_{ij} - \xi_i)^2$ for the ith line connecting the points at the levels $y_{i1}, \ldots, y_{ip}$. Then our criterion would be to choose α_j and $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, so that

$$S^2 = \sum_{i=1}^{n} S_i^2 = \sum_{j=1}^{p} \|\mathbf{y}_j - \boldsymbol{\xi}\|^2 \tag{3}$$

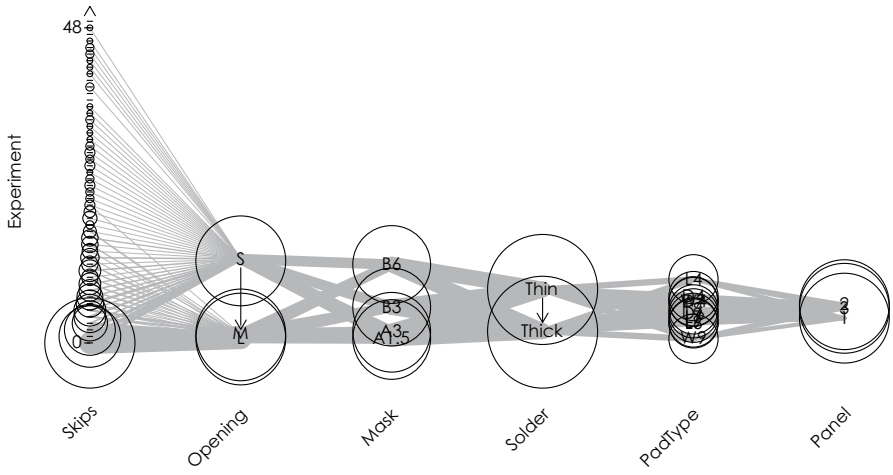


Fig. 2. Textile plot for the solder balanced data where the order of warps are rearranged by variance criterion.

is minimised. The vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$ also has to be chosen to minimise the sum of squares since the levels ξ_i, $i = 1, \ldots, n$ are unknown a priori.

Here we need a constraint on α_j, $\boldsymbol{\beta}_j$ and ξ_i so as to avoid trivial solutions like $\alpha_j = 0$, $\boldsymbol{\beta}_j = \mathbf{0}$ for $j = 1, \ldots, p$ and $\xi_i = 0$ for $i = 1, \ldots, n$. A natural constraint would be that the total dispersion of the points on the textile plot remains constant, that is

$$\sum_{j=1}^{p} \|\mathbf{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2 = np. \tag{4}$$

By using the matrix notations $\mathbf{A} = (\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_k;\ 1 \leq j, k \leq p)/p$ and $\mathbf{B} = \mathrm{diag}(\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_j;\ 1 \leq j \leq p)$, we have the following proposition for the solution of the minimisation problem, where $\tilde{\mathbf{X}}_j = \mathbf{X}_j - \mathbf{1}\mathbf{1}^T\mathbf{X}_j/n$. (the proof is e.g. given in Section 2.2 in Kumasaka and Shibata (2008)).

Proposition 1 *For the given numerical or categorical data vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_p$*, The solution which minimises* (3) *under the constraint* (4) *is given by* $\hat{\alpha}_j = \alpha_0 - \mathbf{1}^T\mathbf{X}_j\hat{\boldsymbol{\beta}}_j/n,\ j = 1, \ldots, p$*, for an arbitrary constant* α_0*,* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \ldots, \hat{\boldsymbol{\beta}}_p^T)^T$ *which is the eigenvector of* $\mathbf{A}$ *with respect to* $\mathbf{B}$ *associated with the largest eigenvalue such that* $\hat{\boldsymbol{\beta}}^T\mathbf{B}\hat{\boldsymbol{\beta}} = np$ *and the ideal coordinates* $\hat{\boldsymbol{\xi}} = \sum_{j=1}^{p} \mathbf{y}_j/p$*.*

Note here that the solution $\hat{\alpha}_j$ and $\hat{\boldsymbol{\beta}}_j$ are variant for the choice of the set of contrasts, the resulting coordinate vectors $\mathbf{y}_j$ are invariant for the choice of contrasts (see Section 2.2 in Kumasaka and Shibata (2008)).

The order of axes are also appropriately arranged in the textile plot. One possible way is that the further left axes has the bigger variation, that is

$\|\mathbf{y}_j - \bar{y}_{\cdot j}\mathbf{1}\| \geq \|\mathbf{y}_{j+1} - \bar{y}_{\cdot j+1}\mathbf{1}\|$ for $j = 1, \ldots, p-1$. If the minimum $\hat{S}^2$ of S^2 in (3) is less than $np/2$, then the further left axis is closer to the ideal coordinate vector $\hat{\boldsymbol{\xi}}$, since the following relationship

$$\left(1 - \frac{2\hat{S}^2}{np}\right)\|\mathbf{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2 = \left(n - \frac{\hat{S}^2}{p}\right) - \|\mathbf{y}_j - \hat{\boldsymbol{\xi}}\|^2 \tag{5}$$

always holds true. Therefore, the leftmost axes are considered to be the most important axes for the classification of cases, since $\hat{\boldsymbol{\xi}}$ essentially gives us a set of ideal coordinates for each case.

Figure 2 shows the textile plot of the soldering data, where the order of axes is rearranged so that the leftmost axes have bigger variations. The area of each circle is proportional to the frequency at the level, and the width of each segment between two levels is proportional to the conditional frequency. The arrow on *Opening* or *Solder* axis shows the order of levels in the category. We can see that the level of the small opening, the B6 mask or the thin solder causes much more skips, and the knot where all levels are shrunk at a point is produced on *Panel* axis, indicating the variable is essentially orthogonal to the other variables.

However, the structural difficulty in understanding complete factorial design data on the textile plot is that connected lines running through the levels of the factors cannot be horizontally aligned, since all levels between two factors are uniformly connected, as was mentioned in the parallel coordinate plot of Figure 1. The order of axes is, in fact, confusing right and left. That is to say, the rightmost axis is closer to the ideal coordinate vector. The next section will provide an interpretation of such textile plots in the context of ANOVA.

4 Relationship with ANOVA

Let $\mathbf{x}_1$ be a n-dimensional numerical data vector performed by complete factorial design with a combination of $p-1$ factors given by n-dimensional categorical data vectors $\mathbf{x}_j$, $j = 2, \ldots, p$, where each element of the categorical data vector $\mathbf{x}_j$ takes one of q_j levels, that is $n = \prod_{j=2}^{p} q_j$. In the context of ANOVA, the linear model of the response $\mathbf{x}_1$ is modelled by factors $\mathbf{x}_2, \ldots, \mathbf{x}_p$ so that

$$\mathbf{x}_1 = \alpha\mathbf{1} + \sum_{j=2}^{p} \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon} \tag{6}$$

with an intercept α, coefficients $\boldsymbol{\beta}_j$, $j = 2, \ldots, p$ and residuals $\boldsymbol{\varepsilon}$, where $\mathbf{X}_j$ is a $(q_j - 1) \times n$ encoded matrix of $\mathbf{x}_j$ with a set of contrasts (Chambers and Hastie (1992)). It is well known that the estimation of the intercept and coefficients which minimise the squared sum of residuals $\|\boldsymbol{\varepsilon}\|^2$ is given by

$$\hat{\alpha} = \mathbf{1}^T\mathbf{x}_1/n \quad \text{and} \quad \hat{\boldsymbol{\beta}}_j = (\mathbf{X}_j^T\mathbf{X}_j)^{-1}\mathbf{X}_j^T\mathbf{x}_1, \quad j = 2, \ldots, p, \tag{7}$$

Table 1. ANOVA table for soldering data.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
Opening	2	13449.02	6724.510	250.9723	0.000000000
Mask	3	8521.22	2840.405	106.0097	0.000000000
Solder	1	4445.17	4445.168	165.9027	0.000000000
PadType	9	2562.26	284.695	10.6254	0.000000000
Panel	2	329.20	164.601	6.1433	0.002265146
Residuals	702	18809.27	26.794		

if the set of contrasts used is orthogonal to the vector $\mathbf{1}$, that is $\mathbf{X}_j^T\mathbf{1} = \mathbf{0}$ for $j = 2, \ldots, p$ (e.g. the Helmert contrast). By introducing projection matrices $\mathbf{P}_j = \mathbf{X}_j(\mathbf{X}_j^T\mathbf{X}_j)^{-1}\mathbf{X}_j^T$, $j = 2, \ldots, p$, we can decompose the variance of $\mathbf{x}_1$ into

$$\|\tilde{\mathbf{x}}_1\|^2 = \sum_{j=2}^{p} \|\mathbf{X}_j\hat{\boldsymbol{\beta}}_j\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2 = \sum_{j=2}^{p} \|\mathbf{P}_j\tilde{\mathbf{x}}_1\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2, \tag{8}$$

where $\tilde{\mathbf{x}}_1 = \mathbf{x}_1 - \hat{\alpha}\mathbf{1}$ and $\hat{\boldsymbol{\varepsilon}}$ is the estimated residuals.

Table 1 shows an ANOVA table of the soldering data. A value on the column Sum of Sq indicates the size of $\|\mathbf{P}_j\tilde{\mathbf{x}}_1\|^2$ in (8). We can see that *Opening*, *Mask*, *Solder* and *PadType* are important factors, and *Panel*, of course, does not contribute to the total variance of *Skips* as was also shown in the textile plot. Here, one might say that the *Panel* factor is clearly significant since the p-value corresponding to *Panel* is considerably small. However, an absolute scale of the p-value (probability) is sometimes meaningless, and it is strongly recommended to compare p-values among factors. The textile plot will prevent misinterpretation of statistical analyses by visualising data.

4.1 Interpretation of locations and scales

The following theorem which follows from Proposition 1 shows the relationship between the locations and the scales of the textile plot and the estimated coefficients in (7). The proof is given in Appendix.

Theorem 1 *Assume that complete factorial data with a numerical data vector* $\mathbf{x}_1$ *of a response and categorical data vectors* $\mathbf{x}_2, \ldots, \mathbf{x}_p$ *of* $p-1$ *factors is given. Then the solution* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \ldots, \hat{\boldsymbol{\beta}}_p^T)^T$ *which minimises* (3) *under the constraint* (4) *is given by* $\hat{\boldsymbol{\beta}}_1 = c$ *and* $\hat{\boldsymbol{\beta}}_j = c(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T\tilde{\mathbf{x}}_1/R$, $j = 2, \ldots, p$, *where* $c = \sqrt{np/(2\|\tilde{\mathbf{x}}_1\|)}$ *and* R *is the multiple correlation coefficient derived from the linear model* (6).

Theorem 1 indicates that, for complete factorial design data, the textile plot provides a plot of

$$(\mathbf{y}_1, \ldots, \mathbf{y}_p) = c\left(\tilde{\mathbf{x}}_1, \frac{1}{R}\mathbf{P}_2\tilde{\mathbf{x}}_1, \ldots, \frac{1}{R}\mathbf{P}_p\tilde{\mathbf{x}}_1\right)$$

for $\alpha_0 = 0$, since $\hat{\boldsymbol{\beta}}_j$ $j = 2, \ldots, p$ are the same as in (7) except for a constant multiplications, and the location $\hat{\alpha}_1$ for the response $\mathbf{x}_1$ also equals to $\hat{\alpha}$ in (7). Therefore, the textile plot is showing the mean value of the response at each level of each factor, as same as in the mean plot (Chambers et al. (1992)).

4.2 Interpretation of axis ordering

The proof of Theorem 1 implies that the minimum of (3) is $\hat{S}^2 = (p-1-R)n$ for complete factorial design data. This indicates that $\hat{S}^2 \geq np/2$ for more than three factor data. Therefore, from (5), the axis ordering in which the leftmost axes have bigger variations is confusing right to left so that the rightmost axis is closer to the ideal coordinate vector.

However, the axis ordering is considered to be that, for complete factorial design data, the further left factors contribute to the total variation of the response variable since $\|\mathbf{y}_j - \bar{y}_{\cdot j}\mathbf{1}\|^2 = \|\mathbf{P}_j\tilde{\mathbf{x}}_1\|^2$. In Figure 2, it is readily seen that the axes from *Opening* to *PadType* have bigger variations and are important factors, whereas axis *Panel* has smaller variation and is less important. Note here that the response variable always located on the leftmost axis since $\|\tilde{\mathbf{x}}_1\|^2 \geq \|\mathbf{P}_j\tilde{\mathbf{x}}_1\|^2/R^2$ for $j = 2, \ldots, p$.

5 Concluding remarks

This article has presented an interpretation of the textile plot for complete factorial design data with one response and multiple factors. The interpretations of the axis ordering, locations and scales of the textile plot has been provided in the context of ANOVA with a practical example, the soldering data. The initial result have encouraged the potential usefulness of the textile plot as an aid to the interpretation of the result of ANOVA.

It is important to reveal what can be seen on the textile plot of incomplete factorial design data or multiple response data. The introduction of interaction variables on the textile plot would also be important improvements. Such developments are left for further investigation.

Appendix

The following lemma is used to prove the Theorem 1.

Lemma 1 *Let* $\mathbf{C}$ *be a* $p \times p$ *symmetric matrix such that*

$$\mathbf{C} = \begin{pmatrix} 1 & \mathbf{c} \\ \mathbf{c}^T & \mathbf{I} \end{pmatrix},$$

with $(p-1)$*-dimensional vector* $\mathbf{c} \neq \mathbf{0}$*. Then the eigenvector* $\boldsymbol{\gamma}$ *of* $\mathbf{C}$ *with respect to the largest eigenvalue* $1 + \|\mathbf{c}\|$ *such that* $\|\boldsymbol{\gamma}\| = 1$ *is given by* $\boldsymbol{\gamma} = \sqrt{1/2}(1, \mathbf{c}^{*T})^T$*, where* $\mathbf{c}^* = \mathbf{c}/\|\mathbf{c}\|$.

Proof Let $\mathbf{c}^*, \mathbf{c}_2, \ldots, \mathbf{c}_{p-1}$ be orthogonal bases of $\mathbb{R}^{p-1}$, then a set of orthogonal eigenvectors of $\mathbf{C}$ is

$$\left\{ \begin{pmatrix} 1 \\ \mathbf{c}^* \end{pmatrix}, \begin{pmatrix} 0 \\ \mathbf{c}_2 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ \mathbf{c}_{p-1} \end{pmatrix}, \begin{pmatrix} 1 \\ -\mathbf{c}^* \end{pmatrix} \right\}$$

with corresponding eigenvalues $\{1 + \|\mathbf{c}\|, 1, \ldots, 1, 1 - \|\mathbf{c}\|\}$. □

Then we have the following proof of Theorem 1.

Proof (Theorem 1) Let $\mathbf{Q}_j \mathbf{R}_j = \tilde{\mathbf{X}}_j$, $j = 1, \ldots, p$, be the results of QR decomposition of $\tilde{\mathbf{X}}_j$. It is easily seen that the eigenvector $\hat{\boldsymbol{\beta}}$ of $\mathbf{A}$ with respect to $\mathbf{B}$ with the largest eigenvalue, such that $\hat{\boldsymbol{\beta}}^T \mathbf{B} \hat{\boldsymbol{\beta}} = np$ in Proposition 1 follows from an eigenvector $\boldsymbol{\gamma} = \mathbf{R}\hat{\boldsymbol{\beta}}$ of

$$\mathbf{C} = (\mathbf{R}^{-1})^T \mathbf{A} (\mathbf{R}^{-1}) = \frac{1}{p} \begin{pmatrix} 1 & \mathbf{Q}_1^T \mathbf{Q} \\ \mathbf{Q}^T \mathbf{Q}_1 & \mathbf{I} \end{pmatrix},$$

where $\mathbf{Q} = (\mathbf{Q}_2, \ldots, \mathbf{Q}_p)$ and $\mathbf{R} = \mathrm{diag}(\mathbf{R}_1, \ldots, \mathbf{R}_p)$ since $\mathbf{B} = \mathbf{R}^T \mathbf{R}$. By using the Lemma 1, we have $\boldsymbol{\gamma} = \sqrt{np/2}(1, \mathbf{Q}_1^T \mathbf{Q} / \|\mathbf{Q}^T \mathbf{Q}_1\|)^T$, which satisfies $\|\boldsymbol{\gamma}\|^2 = \hat{\boldsymbol{\beta}}^T \mathbf{B} \hat{\boldsymbol{\beta}} = np$. Note here that $\mathbf{R}^{-1} = \mathrm{diag}(1/\|\tilde{\mathbf{x}}_1\|, \mathbf{R}_2^{-1}, \ldots, \mathbf{R}_p^{-1})$ and $\mathbf{Q}_1 = \tilde{\mathbf{x}}_1 / \|\tilde{\mathbf{x}}_1\|$, we have $\mathbf{R}^{-1}\boldsymbol{\gamma} = \hat{\boldsymbol{\beta}}$ such that

$$\hat{\beta}_1 = \sqrt{\frac{np}{2\|\tilde{\mathbf{x}}_1\|}} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_j = \frac{\hat{\beta}_1 (\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_j)^{-1} \tilde{\mathbf{X}}_j^T \tilde{\mathbf{x}}_1}{R}; \ j = 2, \ldots, p,$$

where

$$\|\mathbf{Q}^T \mathbf{Q}_1\| = \frac{\|\mathbf{Q}^T \tilde{\mathbf{x}}_1\|}{\|\tilde{\mathbf{x}}_1\|} = \frac{\|\mathbf{Q}\mathbf{Q}^T \tilde{\mathbf{x}}_1\|}{\|\tilde{\mathbf{x}}_1\|} = \frac{\|\sum_{j=2}^p \mathbf{P}_j \tilde{\mathbf{x}}_1\|}{\|\tilde{\mathbf{x}}_1\|} = R$$

is the multiple correlation coefficient, since $\mathbf{Q}$ is a column orthogonal matrix so that $\mathbf{Q}^T \mathbf{Q} = \mathbf{I}$. □

References

TUKEY, J.W. (1977): *Exploratory Data Analysis.* Addison-Wesley, Reading, MA.

CHAMBERS, J.M. and HASTIE, T.J. (1992): *Statistical Models in S.* Wadsworth & Brooks, California.

FRIENDLY, M. and KWAN, E. (2003): Effect ordering for data displays. *Computational Statistics & Data Analysis 43 (4), 509-539.*

INSELBERG, A. (1985): The plane with parallel coordinates. *The Visual Computer, 1, 69-91.*

KUMASAKA, N. and SHIBATA, R. (2008): High dimensional data visualisation: the textile plot. *Computational Statistics & Data Analysis, 52 (7), 3616-3644.*

WEGMAN, E. (1990): Hyperdimensional data analysis using parallel coordinates. *Journal of The American Statistical Association , 85, 664-675.*

Lifting Between the Sets of Three-Way Contingency Tables and R-Neighbourhood Property

Toshio Sakata[1] and Toshio Sumi[2]

[1] Faculty of Design, Kyushu University
4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan,
sakata@design.kyushu-u.ac.jp
[2] Faculty of Design, Kyushu University
4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8540, Japan,
sumi@design.kyushu-u.ac.jp

Abstract. In this paper we consider the sequential conditional test for three-way contingency tables. Conditional tests of no interaction for three-way contingency tables use as the frame of inference the set of all contingency tables with three fixed two-way marginal tables. Lifting between three-way contingency tables means a method of calculating the frame Ω_t of the t-stage from Ω_{t-1} of the $(t-1)$-stage. This makes it easy to perform the sequential conditional test. We will show the meaning of a r-neighbourhood property for this problem. Then, we first show a 1-neighbourhood property for $I \times J \times 2$ tables with arbitrary I and J, and next, a 2-neighbourhood property for $3 \times 3 \times 3$ tables.

Keywords: three-way contingency table, sequential conditional test, Markov basis, lifting

1 Introduction

For the conditional test of contingency tables, the Fisher's exact test is commonly used. When testing the independence of a two dimensional contingency table, the conditional test treats the distribution on the set of all contingency tables with fixed row sums and column sums. For the problem of testing no three factor interaction of a three-way table, the conditional test treats a distribution on the set of all contingency tables with three fixed two-way tables. These sets of contingency tables are called the frame of conditional inference, or in short, the frame. In this paper, we examine the construction of the frames of conditional inference in a sequential conditional test. Let Ω_t be the frame at the t-stage of the sequential conditional test. Then, we have a sequence of frames in an experiment,

$$\Omega_1 \to \Omega_2 \to \cdots \to \Omega_t \to \Omega_{t+1} \to \cdots$$

and it may be unnecessary to calculate each Ω_t from scratch each time. Thus we have arrived to the following question: how can we calculate Ω_t from Ω_{t-1}

successively, where Ω_t is the set of marginal fixed contingency tables. In the case of two-way tables this problem was dealt with in Sakata and Sawae (2003), which was motivated by Saito et al. (1997) that treated the integer programming problem. In this case fortunately, we have Sasaki's operator, and by using this we can lift up the frame of the $(t-1)$-stage Ω_{t-1} into the frame of the t-stage Ω_t. That is, for a fixed row sum vector r and column sum vector c, the generating function of $\Omega(r,c)$, the set of all contingency tables with row sum vector r and column sum vector c, $\Phi(u)$ is defined by

$$\Phi(u|r,c) = \sum_{x \in \Omega(r,c)} \frac{u^x}{x!} = \sum_{x \in \Omega(r,c)} \frac{\prod_{i,j} u_{ij}^{x_{ij}}}{\prod_{i,j} x_{ij}!}.$$

Then, Sasaki's operator for $\Phi(r,c)$ is defined by

$$C_{ij} : u_{ij} + \sum_{p=1}^{s} \sum_{q=1}^{t} u_{pj} u_{iq} \frac{\partial}{\partial u_{pq}}.$$

Then it holds that

$$C_{ij}\Phi(u|r,c) = (1+r_i)(1+c_j)\Phi(u|r+e_i, c+e_j), \tag{1}$$

where e_k is the unit vector with 1 in the k-th coordinate and 0 in other coordinates.

The equation (1) means that C_{ij} is an onto mapping from $\Omega(r,c)$ to $\Omega(r+e_i, c+e_j)$. From this we have the following algorithm.

Algorithm:
Let $\Omega_{t-1} = \Omega(r,c)$ be given at the $(t-1)$-stage and an observation falls in the (i,j)-cell at the t-stage. Then applying C_{ij} to $\Phi(u|r,c)$ we get $(1+r_i)(1+c_j)\Phi(u|r+e_i, c+e_j)$, and hence obtain the set $\Omega_t = \Omega(r+e_i, c+e_j)$, if duplications are deleted.

For the case of three-way tables, as far as the authors know, such a compactly describable creation operators as Sasaki's operator have not appeared in the literature. The purpose of this paper is to construct an algorithm lifting up Ω_{t-1} into Ω_t for three-way tables. In place of pursuing an operator like Sasaki's operator for three-way tables, we establish r-neighborhood theorems. Assume that at the t-stage we had a sample in a (i,j,k)-cell. Then, the r-neighbourhood theorem asserts that for any three-way table H in Ω_t with $H_{ijk} = 0$ there is a table H' with $H'_{ijk} > 0$ which is transmittable at most r steps of moves by using moves in the minimal Markov basis. Since H' with $H'_{ijk} > 0$ can be obtained by some H_0 in Ω_{t-1} by simply adding 1 in the (i,j,k)-cell, this means that H can be obtained by one operation of adding 1 and at most r moves by using the elements in a minimal Markov basis. Fortunately, we can use the unique minimal Markov basis for $3 \times 3 \times 3$ tables which has been obtained by Aoki and Takemura (2003, 2004).

For the case of three-way tables an appropriate lifting operator has not been obtained, as far as we know. So, in place of seeking a compactly describable operator like Sasaki's operator, here we trace a different root. Our starting point is the following theorem which is easily proven.

Theorem 1. *Let us assume that at the t-stage we had a data in an (i,j,k)-cell and that the frame changed from Ω_{t-1} to Ω_t. Let H be any table of Ω_t with $H_{ijk} > 0$. Then H is obtainable by some $H' \in \Omega_{t-1}$ by simply adding 1 in the (i,j,k)-cell of H'.*

From Theorem 1, we need to consider how we can generate $H \in \Omega_t$ with $H_{111} = 0$. Before stating main theorems we give a definition.

Definition 1. Let us assume that the unique minimal Markov basis $\mathcal{B}$ exists. Then, for H and $H' \in \Omega_t$, H' is said to be in the *r-neighbourhood* of H if H' is reachable from H by at most r moves of $\mathcal{B}$. Ω_t has a *r-neighbourhood property* if for each $H \in \Omega_t$ there is $H' \in \Omega_t$ with $H'_{ijk} > 0$ in the r-neighbourhood of H and there is $H \in \Omega_t$ such that the $(r-1)$-neighbourhood of H has no H' with $H'_{ijk} > 0$.

Now we state the following main theorems, whose proofs are given in the later sections.

Theorem 2. *The set of $I \times J \times 2$ contingency tables has a 1-neighbourhood property.*

Theorem 3. *The set of $3 \times 3 \times 3$ contingency tables has a 2-neighbourhood property.*

It is clear that Theorem 1 and Theorem 3 means Theorem 4.

Theorem 4. *For the case of $3 \times 3 \times 3$ tables, if a new data was obtained in the (i,j,k)-cell, then the set of operations of adding 1 in the (i,j,k)-cell and operations by at most 2 consequent moves of the minimal Markov basis forms a lifting operator from Ω_{t-1} to Ω_t.*

2 A proof of Theorem 2

Let $S = \{(i,j) \mid 1 \le i \le I,\ 1 \le j \le J\}$. We consider an unoriented cycle

$$\sigma = \sigma_1, \sigma_2, \ldots, \sigma_{2n}, \sigma_1$$

of length $2n \ge 4$ with labels, where σ_t's are points in S and labels are in $\{P, N\}$ and adjacent verteces have no identical labels. A cycle σ is called *simple* if $s \ne t$ implies $\sigma_s \ne \sigma_t$. We call a cycle σ a *Lawrence cycle* if it is a simple cycle with labels and $\sigma_{2s-1} = (i_s, j_s)$, $\sigma_{2s} = (i_s, j_{s+1})$ for each

$1 \leq s \leq n$ such that $i_s \neq i_{s+1}$ and $j_s \neq j_{s+1}$ for $1 \leq s < n$ and $i_n \neq i_1$ and $j_{n+1} = j_1$. For convenience we put $\sigma_t = \sigma_{t+2an}$ for an integer a.

For two simple cycles σ of length $2n$ and τ of length $2m$, as follows we define a operator, called a joint operator, which is how to create a new simple cycle.

For each segment

$$\sigma_{a+x} = \tau_{b\pm x},\ 0 \leq x \leq s,\ s \geq 0, \quad \sigma_{a-1} \neq \tau_{b\mp 1},\ \sigma_{a+s+1} \neq \tau_{b\pm(s+1)}$$

of the intersection $\sigma \cap \tau$, if the label at σ_a is the same as one at τ_b, we do nothing, otherwise we change a route as follows. If three points σ_{a-1}, σ_a and τ_{b-1} lie on a line, then we exchange the route at the neighbourhood of $\sigma_a = \tau_b$ from

$$\cdots \sigma_{a-2}, \sigma_{a-1}, \sigma_a, \sigma_{a+1} \cdots$$

to

$$\cdots \sigma_{a-2}, \sigma_{a-1}, \tau_{b\mp 1}, \tau_{b\mp 2} \cdots$$

and exchange the route at the neighbourhood of $\sigma_{a+s} = \tau_{b\pm s}$ from

$$\cdots \tau_{b\pm(s+2)}, \tau_{b\pm(s+1)}, \tau_{b\pm s}, \tau_{b\pm(s-1)} \cdots$$

to

$$\cdots \tau_{b\pm(s+2)}, \tau_{b\pm(s+1)}, \sigma_{a+s+1}, \sigma_{a+s+2} \cdots$$

since three points $\tau_{b\pm s}$, $\tau_{b\pm(s+1)}$ and σ_{a+s+1} lie on a line. If $s > 0$ then three points σ_{a-1}, σ_a and τ_{b-1} lie on a line and so we exchange the route as above. If three points $\tau_{b\pm s}$, $\tau_{\pm(s+1)}$ and σ_{a+s+1} do not lie on a line, then $s = 0$ and we exchange from

$$\cdots \sigma_{a-2}, \sigma_{a-1}, \sigma_a, \sigma_{a+1} \cdots$$

to

$$\cdots \sigma_{a-2}, \sigma_{a-1}, \tau_{b\pm 1}, \tau_{b\pm 2} \cdots$$

at the neighbourhood of $\sigma_a = \tau_b$. The labels are not changed. In general after applying the joint operator, simple cycles are made. We select one appreciate cycle among them. It is easy to see that $\sigma \# \tau$ is a Lawrence cycle if σ and τ are Lawrence cycles, where we denoted by $\sigma \# \tau$ the simple cycle obtained from σ and τ by applying the joint operator.

A Lawrence cycle σ of length $2n$ is called a *minimal* Lawrence cycle if $1 \leq a < b \leq n$ implies that $i_a \neq i_b$ and $j_a \neq j_b$, where $\sigma_{2s-1} = (i_s, j_s)$, $\sigma_{2s} = (i_s, j_{s+1})$ for each $1 \leq s \leq n$. For a Lawrence cycle σ as above, we define a move $m(\sigma)$ by

$$m(\sigma)_{i_s j_s 1} = -1, \quad m(\sigma)_{i_s j_{s+1} 1} = 1$$

for each s and all other $(i, j, 1)$-cells are zero and $m(\sigma)_{ij2} = -m(\sigma)_{ij1}$ when the label of σ_1 is P. The values ± 1 of cells are determined by their labels. If the label is P then the value is -1 and otherwise 1.

Recall that the Markov basis for the set of $I \times J \times 2$ tables are well-known (see Sturmfels (1996)).

Theorem 5. *The set of moves $m(\sigma)$ for all minimal Lawrence cycles σ of even length ≥ 4 becomes a minimal Markov basis. Furthermore, a minimal Markov basis is unique.*

Proof. First we show that $m(\sigma)$ for a minimal Lawrence cycle is indispensable for a Markov basis. Let σ be a minimal Lawrence cycle. We put two-way tables M_+ and M_- with smallest L^1-norm so that $m(\sigma)_{\cdot\cdot 1} = M_+ - M_-$. That is, $(M_+)_{ij} = 1$ only if $m(\sigma)_{ij1} = 1$, and otherwise it is 0, and $(M_-)_{ij} = 1$ only if $m(\sigma)_{ij1} = -1$, and otherwise it is 0. Let H and H' be contingency tables defined by

$$H_{\cdot\cdot 1} = H'_{\cdot\cdot 2} = M_+ \text{ and } H_{\cdot\cdot 2} = H'_{\cdot\cdot 1} = M_-.$$

Then H and H' have the same marginals which are either 1 or 0.

Let $\mathcal{B}$ be a Markov basis. Consider the moves

$$H, H'', \cdots, H'$$

from H to H' in $\mathcal{B}$. Put $M = H'' - H \in \mathcal{B}$. We show that $M = m(\sigma)$. The two-way table $M_{\cdot\cdot 1}$ determines $M_{\cdot\cdot 2}$. We take first $(i_1, j_1, 1)$ so that $M_{i_1 j_1 1} < 0$ and next choose j_2 so that $M_{i_1 j_2 1} > 0$. Note that $j_2 \neq j_1$. Next, we choose i_2 so that $M_{i_2 j_2 1} < 0$. Note that $i_2 \neq i_1$. Similarly, we continue to choose $j_3, i_3, j_4, i_4, \ldots$ so that $M_{i_a j_a 1} < 0$ and $M_{i_a j_{a+1} 1} > 0$ for each a. Then, we have a cycle

$$\tau := (i_1, j_1), (i_1, j_2), (i_2, j_2), \ldots, (i_{2m}, j_1), (i_1, j_1).$$

We define the label of τ_1 is P. Since $H_{i_a j_a 1} \leq 1$, it holds that $H_{i_a j_a 1} = 1$ and $M_{i_a j_a 1} = -1$. Similarly, $H_{i_a j_{a+1} 2} \leq 1$ implies that $M_{i_a j_{a+1} 2} = -1$ and so $M_{i_a j_{a+1} 1} = 1$. Recall that $H_{ij1} = H_{ij2} = 0$ for (i, j) not belonging to σ. Thus, we obtain that $M = m(\tau)$ and that τ coincides with σ. Therefore, $m(\sigma)$ is an indispensable transition from H to $H' = H + m(\sigma)$. So it suffices to show that the set of moves $m(\sigma)$ for all minimal Lawrence cycles σ is a Markov basis. Let H, K be distinct tables with same marginals. Set X as a move which comes from $H - K$ by applying moves $m(\sigma)$ for minimal Lawrence cycles σ and minimizes $|X|$, L^1-norm of X. Suppose that $X \neq 0$, and we lead to a contradiction. Take (i_1, j_1) with $X_{i_1 j_1 1} < 0$. Then we can take i_2 so that $X_{i_2 j_1 1} > 0$, equivalent to that $X_{i_2 j_1 2} < 0$, and take j_2 so that $X_{i_2 j_2 1} < 0$. If $X_{i_1 j_2 1} > 0$, then by applying the move

$$m((i_1, j_1), (i_2, j_1), (i_2, j_2), (i_1, j_2), (i_1, j_1))$$

where the label of (i_1, j_1) is N, $|X|$ becomes smaller, which is a contradiction. Thus $X_{i_1 j_2 1} \leq 0$. We can take $i_3 \neq i_1, i_2$ so that $X_{i_3 j_2 1} > 0$. Next, we want to choose j_3 with $X_{i_3 j_3 1} < 0$. If $j_3 = j_1, j_2$, by applying some move, the norm $|X|$ becomes smaller, which is a contradiction. Succeeding this process, we can find (i_r, j_r) with $X_{i_r j_r 1} < 0$ and $X_{i_1 j_r 1} > 0$, and then we obtain a minimal Lawrence cycle, which is also a contradiction. Thus $X = 0$. Therefore, the set of moves $m(\sigma)$ for all minimal Lawrence cycles σ of even length ≥ 4 becomes a unique minimal Markov basis. □

Let $\mathcal{B}$ be a unique minimal Markov basis. The assignment from a minimal Lawrence cycles σ to $m(\sigma) \in \mathcal{B}$ is bijective.

Lemma 1. *Let H be an $I \times J \times 2$ table and let σ and τ be Lawrence cycles. If $H + m(\sigma)$ and $H - m(\tau)$ has no negative values, then $(H - m(\tau)) + m(\sigma \# \tau)$ has no negative values.*

Proof. If (i, j) is not a point of $\sigma \# \tau$, then $(H - m(\tau)) + m(\sigma \# \tau) = H - m(\tau)$ which has no negative values. Let (i, j) be a point of $\sigma \# \tau$ and $k = 1, 2$. By the straightforward computation, we confirm that $(H - m(\tau) + m(\sigma \# \tau))_{ijk}$ is $(H + m(\sigma))_{ijk}$, H_{ijk}, or $(H - m(\tau))_{ijk}$, since $m(\sigma \# \tau)_{ijk}$ is $m(\sigma)_{ijk}$, $m(\tau)_{ijk}$ or 0. □

We may assume $(i, j, k) = (1, 1, 1)$ without loss of generality. Theorem 2 easily follows from the following lemma.

Lemma 2. *Let H be an $I \times J \times 2$ table with $H_{111} = 0$. Assume that there is a three-way table F with $F_{111} > 0$ which has the same marginals as H. Then, there is a minimal Lawrence cycle σ containing $(1, 1)$ with label N such that $H + m(\sigma)$ has no negative values.*

Proof. First we show that there is a Lawrence cycle containing $(1, 1)$ with label N. To do this, we consider a transition from F to H by moves in $\mathcal{B}$. As a subsequence of this transition, we have a transition

$$F^{(r+1)}, F^{(r)}, \ldots, F^{(2)}, F^{(1)}$$

from $F^{(r+1)}$ to $H = F^{(1)}$ such that $F^{(r+1)}_{111} = 1$ and $F^{(s)}_{111} = 0$ for $1 \leq s \leq r$. Let $\sigma = m^{-1}(F^{(r+1)} - F^{(r)})$ be a minimal Lawrence cycle determined by $F^{(r+1)} - F^{(r)} \in \mathcal{B}$. Then $F^{(r)} + m(\sigma_r) = F^{(r+1)}$ and in particular has no negative values. Let $\tau_s = m^{-1}(F^{(s+1)} - F^{(s)})$ be a minimal Lawrence cycle determined by $F^{(s+1)} - F^{(s)} \in \mathcal{B}$ for each $1 \leq s \leq r - 1$. Noting that $m(\sigma)_{111} = 1$ and $m(\tau_s)_{111} = 0$ for $1 \leq s \leq r - 1$, the label of $(1, 1)$ of σ is N and the cycle τ_s for $1 \leq s \leq r - 1$ does not contain $(1, 1)$. In a joint operator, we choose a cycle containing $(1, 1)$. Then we obtain a Lawrence cycle

$$\rho = (\cdots(\sigma \# \tau_{r-1}) \# \cdots \# \tau_2) \# \tau_1$$

such that the label of $(1, 1)$ in ρ is N. By Lemma 1, $F^{(1)} + m(\rho)$ has no negative values. It is clear that $(F^{(1)} + m(\rho))_{111} = 1$. By using a shortcut of this cycle, we can find a minimal Lawrence cycle ρ' containing $(1, 1)$ with label N so that $F^{(1)} + m(\rho')$ has no negative values. □

3 A proof of Theorem 3

In this section, we introduce an algorithm to get r so that the set of $I \times J \times K$ tables with same marginals has a r-neighbourhood property and show the computational results.

Let $\Omega(\alpha,\beta,\gamma)$ mean the set of all three-way contingency tables with marginals α,β,γ and $\Omega^u(\alpha,\beta,\gamma)$ the subset of $\Omega(\alpha,\beta,\gamma)$ consisting H with $H_{i_1j_1k_1}=u$. Similarly let $\mathcal{B}^u$ be the subset consisting M with $M_{i_1j_1k_1}=u$ for a Markov basis $\mathcal{B}$. We write $\Omega(\alpha,\beta,\gamma)$ and $\Omega^s(\alpha,\beta,\gamma)$ by Ω and Ω^s respectively for short.

Theorem 6. *Suppose that neither Ω^1 nor Ω^0 is empty. Fix a positive integer r. Suppose that for any $H\in\Omega^1$, $M_1^0,\ldots,M_r^0\in\mathcal{B}^0$ and $M_{r+1}^1\in\mathcal{B}^1$, if*

$$(\cdots(H+M_1^0)+M_2^0)+\cdots+M_r^0)+M_{r+1}^1$$

lies in Ω, then there are $M_1'^0,\ldots M_{r-1}'^0\in\mathcal{B}^0$, $M_r'^1\in\mathcal{B}^1$ such that

$$(\cdots(H+M_1'^0)+M_2'^0)+\cdots+M_{r-1}'^0)+M_r'^1$$

lies in Ω. Then Ω^1 intersects with the r-neighbourhood of H for any $H\in\Omega^0$.

Proof. Let H be an element of Ω^0. Take a transition from H to some

$$(\cdots(H+M_1)+M_2)+\cdots+M_{s-1})+M_s\in\Omega^1$$

with minimal length s. The minimality implies that $M_j\in\mathcal{B}^0$ $(j=1,\ldots,s-1)$ and $M_s\in\mathcal{B}^1$. Suppose that $s>r$. For $\widehat{H}=(\cdots(H+M_1)+\cdots+M_{s-r-2})+M_{s-r-1}$, by the assumption, we obtain that there exist $M_1'^0,\ldots M_{r-1}'^0\in\mathcal{B}^0$ and $M_r'^1\in\mathcal{B}^1$ such that $(\cdots(\widehat{H}+M_1'^0)+M_2'^0)+\cdots+M_{r-1}^0)+M_r'^1\in\Omega$. However the transition from H has length $s-1$, which is a contradiction. Therefore, it holds $s\le r$. □

For $M_1^0,\ldots,M_r^0\in\mathcal{B}^0$ and $M_{r+1}^1\in\mathcal{B}^1$, we define a three-way table

$$N=N(M_1^0,M_2^0,\ldots,M_r^0,M_{r+1}^1)$$

by

$$\begin{aligned}N_{ijk}=-\min(0,&(M_1^0)_{ijk},(M_1^0+M_2^0)_{ijk},\ldots,\\&(M_1^0+\ldots+M_r^0)_{ijk},(M_1^0+\ldots M_r^0+M_{r+1}^1)_{ijk}).\end{aligned}$$

Note that $N_{i_1j_1k_1}=0$.

The following lemma is one of keys:

Lemma 3. *The following two claims are equivalent.*

1. *Ω has a r-neighbourhood property.*
2. *Ω^1 intersects with the r-neighbourhood of $N(M_1^0,\ldots,M_r^0,M_{r+1}^1)$ for any $M_1^0,\ldots,M_r^0$ of $\mathcal{B}^0$ and any M_{r+1}^1 of $\mathcal{B}^1$, but does not intersect with the $(r-1)$-neighbourhood of $N(M_1'^0,\ldots,M_{r-1}'^0,M_r'^1)$ for some $M_1'^0,\ldots,M_{r-1}'^0$ of $\mathcal{B}^0$ and some $M_r'^1$ of $\mathcal{B}^1$.*

Now we assume $I = J = K = 3$. Then the assumption of Theorem 6 with $r = 2$ is confirmed by using a computer program, which proves Theorem 3. Further, by another computer program, we can confirm the following result, where we use their notations about the minimal Markov basis for $3 \times 3 \times 3$ contingency tables in Aoki and Takemura (2003).

Theorem 7. *Suppose that Ω^1 is not empty. A table N of Ω^0 is transmitted to some table of Ω^1 by at least one of the following Markov moves.*

$$m_4(i_1i_2, j_1i_2, k_1i_2),\ m_4(i_1i_3, j_3j_1, k_2k_3) + m_4(i_1i_2, j_1j_2, k_1k_3),$$
$$m_6^I(i_1i_2, j_1j_3j_2, k_1k_2k_3),\ m_4(i_1i_3, j_3j_2, k_1k_3) + m_4(i_1i_2, j_1j_3, k_1k_2),$$
$$m_6^J(i_1i_3i_2, j_1j_2, k_1k_2k_3),\ m_4(i_2i_3, j_3j_1, k_1k_3) + m_4(i_1i_3, j_1j_2, k_1k_2),$$
$$m_6^K(i_1i_3i_2, j_1j_2j_3, k_1k_2),\ m_4(i_2i_3, j_2j_3, k_2k_3) + m_4(i_1i_2, j_1j_2, k_1k_2),$$
$$m_4(i_2i_3, j_3j_1, k_1k_3) + m_6^I(i_1i_3, j_1j_3j_2, k_1k_2k_3)$$

Conclusion and future work

In this paper we proved the 1-neighbourhood property for $I \times J \times 2$ tables and the 2-neighbourhood property for $3 \times 3 \times 3$ tables. When considering the sequential conditional inference for these types of contingency tables, these properties enable us to make a program for generating frames of the exact conditional inference after obtaining data from the frame of the previous stage. Note that for testing hypothesis we need to combine a process of excluding duplication with one of generating members of the frame. In the future work we will strive for larger tables.

References

AGRESTI, A. (1996): An Introduction to Categorical Data Analysis. *John Wiley & Sons, Inc.*

AOKI, S. (2004): Exact methods and Markov chain Monte Carlo methods of conditional inference for contingency tables. *Doctor Thesis, Tokyo University.*

AOKI, S. and TAKEMURA, A. (2003): Minimal basis for connected Markov chain over $3 \times 3 \times K$ contingency tables with fixed two-dimensional marginals. *Australian and New Zealand Journal of Statistics 45, 229–249.*

SAITO M., STRUMFELS B. and TAKAYAMA, N. (1997): Hypergeometric polynomials and integer programming. *Compositio Mathematica 115, 185–204.*

SAKATA, T. and SAWAE, R. (2003): A study of the sequential conditional test for contingency tables. *Journal of the Japanese Society of Computational Statistics, 15(2), 169–174.*

SAKATA, T. and SAWAE, R. (2005): Creation operators of three way contingency tables. *International Conference of Theoretical Effectiveness and Practical Effectivity of Gröbner Bases at Rikkyo University.*

SASAKI, T. (1991): Contiguity relations of Aomoto-Gelfand's hypergeometric functions and applications to Apell's system F_3 and Goursat's system ${}_3F_2$. *SIAM Journal of Mathematical Analysis 22, 821–846.*

STURMFELS, B. (1996): Gröbner bases and convex polytopes. *American Mathematical Society, University Lecture Series 8.*

Matrix Visualization and Rasch Models

Anatol Sargin, Ali Ünlü, and Antony Unwin

Department of Computer Oriented Statistics and Data Analysis, University of Augsburg, D-86159 Augsburg, Germany
anatol.sargin@math.uni-augsburg.de, ali.uenlue@math.uni-augsburg.de, unwin@math.uni-augsburg.de

Abstract. Visualization is essential in modern applied statistics. In psychometrics, however, graphics have been neglected so far.

This paper tries to fill the gap between psychometric statistical modeling on the one hand, and visualization as a practical tool on the other. Matrix visualizations are proposed to display the data and results of a Rasch analysis using the one-parameter logistic model in item response theory.

Matrix visualizations can conveniently address and summarize the data based on specifics of the Rasch model in a single graphic. Matrix visualization is further enhanced for outlier regions to identify and study improbable response behavior. Extensions of matrix visualization are described for dealing with large datasets. The usefulness of matrix visualization in psychometrics is illustrated with data from the Programme for International Student Assessment (PISA).

Keywords: psychometrics, Rasch model, graphics, matrix visualization

1 Introduction

Graphics are widely used tools in modern applied statistics, because they are easy to create, convenient to use, and they can present information effectively (Cook and Swayne (2007) or Unwin et al. (2006)). Real datasets are often large, complex, and difficult to analyze. Even if that is possible, the results of an analysis, or the analysis itself, are too bulky or complicated to communicate easily. This applies particularly in psychometrics, where graphics have seldom been used so far.

Consider, for instance, dichotomous data and the classical but fundamental one-parameter logistic (Rasch) model, with difficulty parameter σ_i for item i and ability parameter θ_v for person v (Fischer and Molenaar (1995), Rasch (1960)):

$$\mathbb{P}(x_{vi} = 1) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)}.$$

The Rasch model parameterizes the probability of a correct answer for an item-person combination as the logistic cumulative distribution function applied to the difference of the corresponding ability and difficulty parameters. The larger θ_v or smaller σ_i is, the higher the probability of a correct answer.

The Rasch model is based on the dichotomy of the data and the central role of the (manifest) total score as a sufficient statistic for (latent) person ability. These properties suggest informative graphics, especially suited for the Rasch model.

In this paper, the recently proposed idea of matrix visualization (Chen (2002) and Chen et al. (2007)) is modified to apply to the Rasch model. In the original publications on matrix visualization the main focus was on sorting cases and variables. Here sorting is determined by the model (as detailed below), and the main focus of the visualization is on investigating such issues as total score groups and improbable response behavior.

A matrix visualization can display the data in just one graphic, and can be enhanced to study outliers, in the sense of improbable responses. Since datasets in psychometrics can be large, it is essential for a good graphic to be applicable for such datasets. Therefore modifications of matrix visualization are discussed for dealing with large datasets.

In the following, two datasets on mathematical literacy are used. One dataset consists of 317 people and 12 items, the other one of 340 people and 32 items. Both are part of the 2003 PISA data (http://www.pisa.oecd.org). To apply the Rasch model to the data and to create the graphics, the statistical computing environment `R` (http://www.r-project.org) was used.

2 Matrix visualization

The main purpose of a matrix visualization is to show the structure of the data, another focus is on outlier detection. First, we define matrix visualization, and then discuss its advantages. Later, approaches are provided for extending it for large datasets.

2.1 Definition of matrix visualization

Matrix visualization draws all item-person combinations from the data matrix in such a way that a dot represents a correct answer of a person (x-axis) to an item (y-axis). An incorrect answer is simply left blank in the graphic. To gain as much information as possible from the graphic a proper sorting of the matrix, suited for the Rasch model, is necessary:

- From left to right people are sorted according to their total score, where the people with the highest scores are to the left.
- From bottom to top items are sorted according to their difficulty, where the easiest item is at the bottom.
- Within every total score group, people solving more difficult items are placed more to the left.

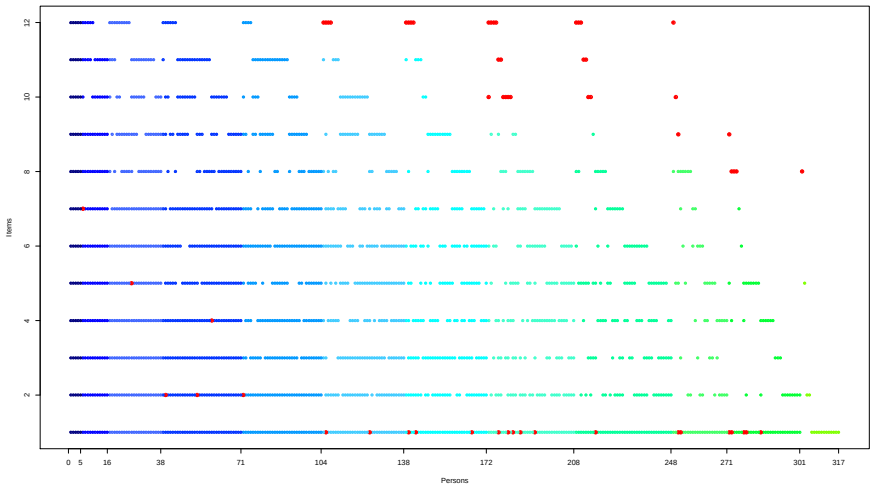


Fig. 1. Matrix visualization for 12 items and 317 people (first PISA dataset). Correct and incorrect responses having a probability of less than 0.15 are marked red, in the upper and lower triangular regions, respectively.

The dots are colored according to a person's total score, and the boundaries of the total score are marked on the x-axis. This takes into account, that the total score is a sufficient statistic under the Rasch model. Figure 1 shows a matrix visualization of the first PISA dataset.[1] A more detailed labeling of the x-axis could show the values of the total score and the numbers of people in every group.

2.2 Advantages of matrix visualization

Matrix visualization helps to distinguish between total score groups, which are a main interest of a Rasch analysis. Under the Rasch model, the total score is a measurement statistic for estimating the ordering of subjects on the latent trait (Fischer and Molenaar (1995)). Furthermore, it is possible to examine whether the items are too hard for the participants, in which case the matrix is sparsely filled. One can also visually check whether correct answers to an item are spread over all individuals, and whether they are concentrated on special total score groups.

Matrix visualization can also be utilized for detecting outliers. Improbable correct and incorrect responses are located in the upper and lower triangular regions of a matrix visualization, respectively. The probabilities of individual responses are computed using the Rasch model formula, and for a fixed outlier

[1] The graphics in color and the R-code can be found on http://stats.math.uni-augsburg.de/mitarbeiter/sargin/.

threshold value, the responses are marked in red according to whether they are unexpectedly correct (a person with a low total score solving a difficult item) or incorrect (a person with a high total score failing an easy item). Figure 1 displays outliers (marked red) of the first PISA dataset, with a threshold value of 0.15.

The investigation of outliers is of practical relevance in psychometrics, for the efficient identification of critical items and people. This is useful in conjunction with fit statistics. A fit statistic is used to determine whether an individual item or person has a large discrepancy from the model, and outliers generally have a great impact on the value of a fit statistic. With matrix visualization it is possible to see in which items and people discrepancies occur and why they occur. In Figure 1, for example, the most difficult item has a critical value of 3.27, for a standard normally distributed fit statistic. This is explained by the fact that many people with lower abilities (16 out of 45) were able to solve this item and were therefore marked as outliers. For details about fit statistics, see Baker (1992).

Matrix visualization can be extended using interaction. Tools such as querying, linking, and zooming enhance our ability to explore and analyze data. These methods allow the user to interact with one or more graphics directly to gain more information about specific data characteristics. With matrix visualization it is possible to highlight or link subgroups by variables, such as sex or age, or to select interesting total score or item groups. More details on interactive methods are provided by Hofmann (1998), Theus and Urbanek (2008) and Unwin et al. (2006).

2.3 Matrix visualization of large datasets

Though graphics are a powerful instrument for exploring data, some problems can arise when datasets are too large. One has to consider that a large dataset in psychometrics is different from common ideas of large datasets in statistics. Psychological experiments often imply concentrated datasets, though related datasets from, for instance, the PISA study can be substantially larger.

When the number of items (and hence total score groups) increases, it is generally difficult to find a good coloring of the graphic. This results in such a smooth color gradient that distinguishing the individual color components becomes very difficult. A slight improvement is the use of a gray scale, or if this fails, discriminable colors for the total score groups. Such a modified graphic can still be used for analyzing the raw data and detecting outliers. One should note, that in the case of a gray coloring, differentiating between two total score groups can be hard. That is why the labels on the abscissa are indispensable for visualizing large datasets. Nevertheless gray coloring is advisable, as it improves the overall impression and serves as a supporting guide. In this situation only approximate, not precise disrimination is needed. Figure 2 shows a matrix visualization of the second PISA dataset using a gray

coloring. One can see that the graphic is very dense, which indicates a good performance of the participants on the items.

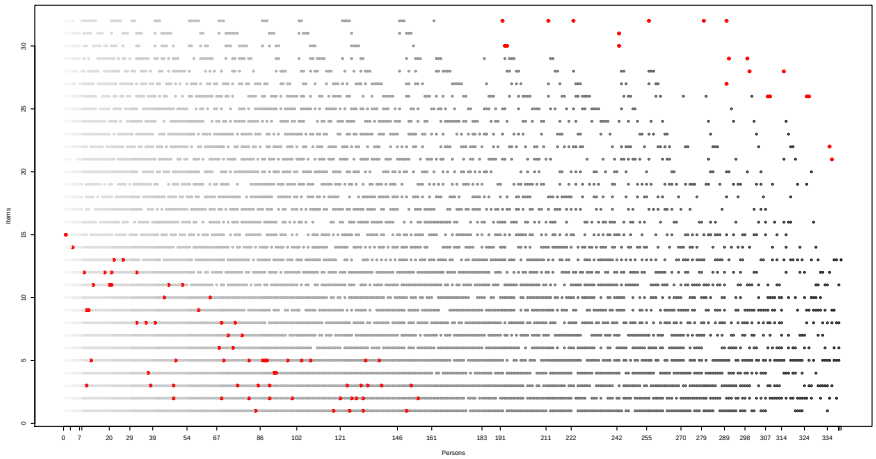


Fig. 2. Matrix visualization for 32 items and 340 people (second PISA dataset). A gray coloring of the dots is used. Correct and incorrect responses having a probability of less than 0.10 are marked red, in the upper and lower triangular regions, respectively. Due to the number of total score groups, not all numbers can be shown on the x-axis, but the more important total score boundaries are.

The dataset (usually the number of people) can be so large that it is not useful to display all items and people in one graphic, because the graphic would be overcrowded. In this case a partitioning of the matrix is a possible alternative. Only parts of the dataset, for specific subsets of items and/or people, are then drawn in a matrix visualization. This is of practical importance, for example, when investigating the performances of the participants with highest total scores on the most difficult items. Another possibility is the idea of grouping items or people. Interactive logical zooming could then be used to drill down into the groups if it is desired to explore local differences.

3 Conclusion

Matrix visualization is a powerful and useful tool for data and model visualization in Rasch analysis. This is especially worth noting as graphics have been little used in psychometrics. Not only does a matrix visualization compactly display the dichotomous data, it also supports detecting and analyzing outliers.

To extend the use of matrix visualization, one possibility is to apply this graphic to polytomous data, for example with different symbols for different

response categories. The development of variants of matrix visualization for other psychometric models (Boomsma et al. (2001) or Van der Linden and Hambleton (1997)) is also an important direction for future research.

It may also be interesting to enhance matrix visualization to more than two dimensions. No printed graphic can display more than two dimensions fully at once and interactive visualization methods can be used to gain insights into multivariate datasets (Theus and Urbanek (2008) and Unwin et al. (2006)). Multiple linked simple displays of the same dataset can be easier to interpret than single complex multivariate plots.

Much can be gained in psychometrics when both modeling and visualization are combined to better understand the data and support model building, and much progress is still to be made in this direction.

References

BAKER, F. (1992): *Item Response Theory: Parameter Estimation Techniques*. Dekker, New York.

BOOMSMA, A., VAN DUIJN, M.A.J. and SNIJDERS, T.A.B. (Eds.) (2001): *Essays on Item Response Theory*. Springer, New York.

CHEN, C.H. (2002): Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica 12, 7-29*.

CHEN, C.H., HAERDLE, W. and UNWIN, A. (Eds.) (2007): *Handbook of Data Visualization*. Springer, New York.

COOK, D. and SWAYNE, D. (2007): *Interactive and Dynamic Graphics for Data Analysis*. Springer, New York.

FISCHER, G.H. and MOLENAAR, I.W. (Eds.) (1995): *Rasch models: Foundations, Recent Developments, and Applications*. Springer, New York.

HOFMANN, H. (1998): Simpson on board the Titanic? Interactive methods for dealing with multivariate categorical data. *Statistical Computing & Statistical Graphics Newsletter 9, 16-19*.

RASCH, G. (1960): *Probabilistic Models for some Intelligence and Attainment Tests*. Nielsen & Lydiche, Copenhagen.

THEUS, M. and URBANEK, S. (2008): *Interactive Graphics for Data Analysis*. CRC Press, London.

UNWIN, A., THEUS, M. and HOFMANN, H. (2006): *Graphics of Large Datasets*. Springer, New York.

VAN DER LINDEN, W.J. and HAMBLETON, R.K. (Eds.) (1997): *Handbook of Modern Item Response Theory*. Springer, New York.

Monte Carlo Evaluation of Model Search in Graphical Models for Ordinal Data

Volkert Siersma[1] and Svend Kreiner[2]

[1] Research Unit for General Practice in Copenhagen
Øster Farimagsgade 5, P.O. Box 2099
1014 Copenhagen K, Denmark, *V.Siersma@gpract.ku.dk*
[2] Department of Biostatistics, University of Copenhagen
Øster Farimagsgade 5, P.O. Box 2099
1014 Copenhagen K, Denmark, *S.Kreiner@biostat.ku.dk*

Abstract. We describe a method to simulate ordinal data from a user specified graphical model so that model search heuristics may be evaluated by Monte Carlo methods. Moreover, inference on relations of conditional independence in graphical models also depends on the model search that precedes it. Monte Carlo samples are used to investigate the distribution of the partial gamma coefficient - the relevant test statistic when data is ordinal - taking the outcome of model search into account.

Keywords: graphical model, model search, Monte Carlo, Goodman and Kruskal's gamma, bootstrap

1 Introduction

A graphical model, an encoding of conditional independence relations in a mathematical graph, is a useful framework for the analysis of multivariate data. Specifically, graphical models give strengthened inference on relationships of interest as, because of the graph's decomposition, we may control only for a subset of all possible confounders, and the graph structure itself can be used to investigate causal mechanisms.

1.1 Estimation of association in graphical models

Central in the analysis of graphical models is the assessment of conditional independence by means of significance tests pertaining the association between two variables conditional on others. In the true graphical model, we may know the correct inference for various statistics that assess conditional independence. However, in practice we do not know the true model and have to derive this from the data as well. Hence, the distribution of the corresponding test statistic is dependent on the result of a model search heuristic, unlikely to be the true graphical model. Since the model search result is a random variable itself with an in general unknown distribution, the distribution of the statistic is also unknown.

In this paper we investigate the power of a specific model search method to infer the true model from the data. Moreover, we investigate the distribution of a certain test statistic when taking into account that the model in which the test is performed is the result of a model search.

1.2 Graphical models and the partial γ coefficient

Often the multivariate data consist of ordinal categorical variables, and a graphical model is identified with a log-linear model (Darroch et al. 1980). This suggests straightforward LR inference for conditional independence. However, the log-linear model does not straightforwardly account for the ordinal nature of the variables. A non-parametric rank correlation coefficient offers a more convenient way to test independence, and moreover describes the strength and sign of the dependence when conditional independence is rejected; a measure for conditional dependence is then given by the corresponding partial coefficient.

A conceptually simple rank correlation coefficient is Goodman and Kruskal (1954)'s γ coefficient; a partial γ coefficient is well-established (Davis 1967; Agresti 1984). Empirical investigation of conditional independence, which corresponds to $\gamma = 0$, is then based on the asymptotic Gaussian distribution of the partial γ (Agresti 1984) or on exact tests (Kreiner 1987). These tests are implemented in the DIGRAM software (Kreiner 2003). In this paper we consider graphical models where the edges correspond to non-zero values for the partial γ coefficient.

2 Simulating ordinal data from a graphical model

To evaluate model search strategies we need to be able to generate data from specific graphical models, i.e. multivariate data where the conditional association between each pair of variables is of a given strength $\gamma = \gamma_0$.

2.1 Probability tables with a given γ coefficient

Let X and Y be two ordinal variables with n_X and n_Y categories respectively, and category labels x_i and y_j. We want to determine a joint distribution of X and Y so that their association has a specific value $\gamma = \gamma_0$. Assume the marginal distributions of X and Y to be given. Define a vector β, with elements β_j, $j = 1, \ldots, n_Y$, as

$$\beta_j := \frac{2}{1 + \exp\left(-s_0(-\frac{(n_Y-1)}{2} + j - 1)\right)} \tag{1}$$

for some s_0. Notably, β is a monotone array with unit mean and median, and all $\beta_j > 0$ for all values of s_0. We assume that the distribution of X is

influenced by Y through β as

$$P(X \leq x_i|Y = y_j) := P(X \leq x_i)^{\beta_j} \tag{2}$$

and the entries of the probability table are found as

$$P(X = x_i, Y = y_j) = P(X = x_i|Y = y_j)P(Y = y_j) \tag{3}$$

where

$$P(X = x_i|Y = y_j) = P(X \leq x_i|Y = y_j) - P(X \leq x_{i-1}|Y = y_j) \tag{4}$$

We then determine s_0 by numerical optimization so that, for a specific constellation of β in (1) and for uniform marginal distributions, the relation has the value γ_0.

2.2 Simulation from a graphical model

Consider a graphical model G on a set of ordinal random variables V where each edge corresponds to a monotone association with a specific partial γ coefficient. The distribution of a variable X conditional on all others $V \setminus \{X\}$ is then characterized by the distribution of X conditional only on $\mathrm{Bd}(X)$, i.e. those variables with which it is directly connected in G (Lauritzen 1996).

$$P(X|V \setminus \{X\}) = P(X|\mathrm{Bd}(X)) \propto \prod_{Y \in \mathrm{Bd}(X)} P(X|Y) \tag{5}$$

The rightmost characterization furthermore assumes that there are no third and higher order interactions. For any realization of third variables the association between two variables X and Y has γ equal to the partial γ coefficient corresponding to the XY edge in the graph. When we specify the marginal distributions for X and Y as uniform, we can determine the joint distribution $P(X, Y)$ with the wanted value for the γ coefficient. The terms in the product on the right-hand-side of (5) are then easily derived from this probability table for given values of Y by selecting the appropriate row or column and reweighting the entries to sum to one.

We use a Gibbs sampler as described in Lauritzen and Richardson (2002) to construct the data sample. Given a data line for the variables in V we use (5) to sample a new value for a specific variable. We cycle through the variables so that a new data line is sampled after each cycle. To avoid dependence on the starting values, we discard early cycles as burn-in. Thereafter, to avoid serial correlation, we keep only every 10th data line.

2.3 A note on the simulation procedure

The simulated data is of a very regular nature: we assume uniform marginal distributions, a kind of linear monotonicity assumption in the form of the vector β in (1), and we assume that we have at most second-order interactions.

Notably the latter is strong, but it is not clear how higher order interactions are to be specified. The assumptions are needed to identify the joint distribution; other assumptions may identify another joint distribution that also conforms to the collection of partial γ coefficients. We could inject irregularity in the procedure by modifying the assumptions in a random way while still adhering to the wanted values of the partial γ coefficients. However, we feel at present that it may be better to realize that the data are generated in a specific regular way, and to draw conclusions in accordance with this, rather than to put much faith in designed irregularities.

3 A model search strategy

To estimate the graphical model from data we employ a model search strategy based on two steps, general to many suggested model search methods.

1. An initial screening to obtain a base model.
2. Stepwise model search starting from and improving the base model

The initial screening is an analysis of two- and three-way tables in the data, previously described in Kreiner (1986,1987,2003), a procedure with similarities to the PC algorithm (Sprites et al. 2000). Notably, the screening is not a proper model search strategy in itself, but merely a means to construct a parsimonious model to make the second step more efficient.

In this paper we specify the stepwise model search by one backwards sequence where edges with the highest p-value are removed until all edges have $p < 0.01$, and subsequently one forward sequence where absent edges with the lowest p-value are added until all remaining absent edges have $p > 0.01$.

Both the initial screening and the stepwise procedures are implemented in DIGRAM where either the complete above procedure may be performed automatically, or manually with the analyst's control at each iteration in the stepwise model search. Moreover, this is the approach that we, and presumably many others, use in statistical practice, see e.g. Klein et al. (1995); Kreiner (2003).

4 Monte Carlo evaluation of the model search strategy

We evaluate the model search strategy from Sec. 3 by investigating the model search outcomes from Monte Carlo (MC) samples from a known graphical model. Investigation of the empirical distribution of selected partial γ coefficients then gives an indication of its distribution.

We simulate 100 data sets with $n = 500$ observations each from the graphical model shown in Fig. 1A. When we assess the edges in the true model structure for all $n = 50000$ observations and for only the first MC sample of $n = 500$, we obtain the estimates in Fig. 1B and 1C respectively. As these

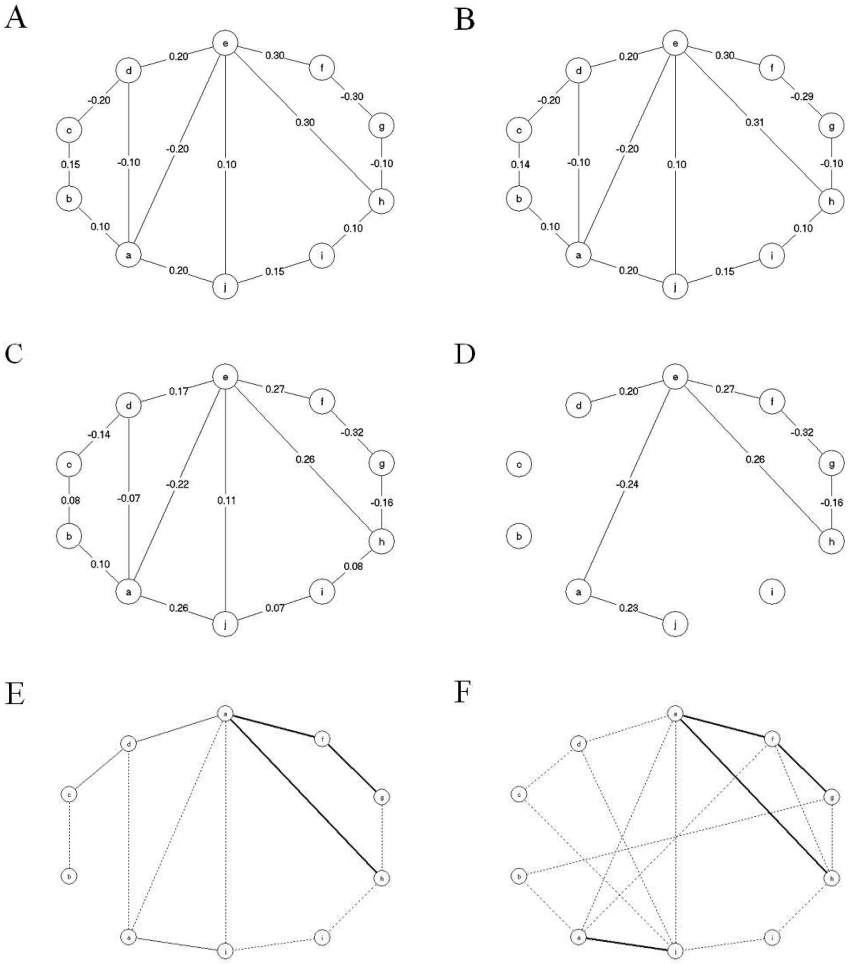


Fig. 1. Graphical models. **A** True model. **B** Partial γ estimates from all data ($n = 50000$). **C** Partial γ estimates from the first of the 100 MC samples ($n = 500$). **D** Model estimate from the first of the 100 MC samples. **E** Power graph from the 100 MC samples. **F** Power graph for 100 bootstrapped data sets from the first of the 100 MC samples. Power graph legend: solid line 95-100% power, normal line: 80-95% power, dotted line: 20-80% power.

resemble the true values with a precision expected from the respective sample sizes, we may be confident that the simulation procedure in Sec. 2 provides adequate data.

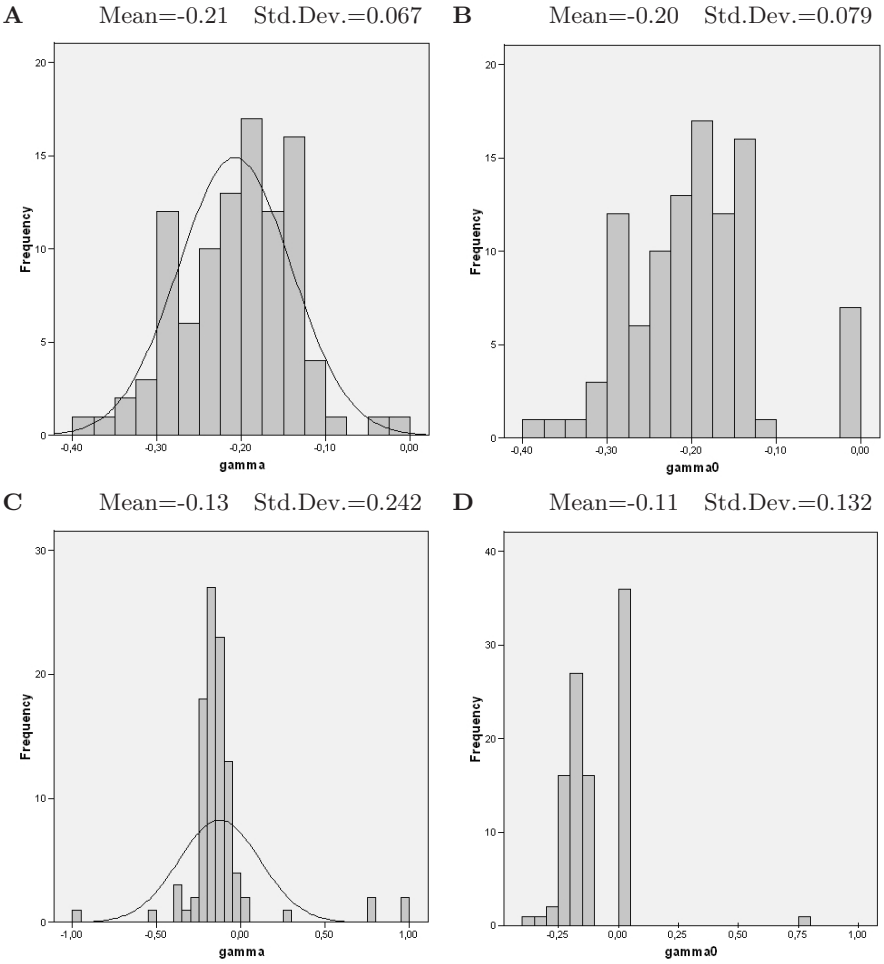


Fig. 2. Distribution of the partial γ coefficient of the CD edge. **A** Distribution estimated from the 100 MC samples. **B** Distribution as in Fig. 2A where insignificant values are set to zero. **C** Distribution estimated from 100 bootstrap samples from the first of the 100 Monte Carlo samples. **D** Distribution as in Fig. 2C where insignificant values are set to zero.

4.1 The model search strategy

We use the model search strategy on the first of the MC samples and estimate the partial γ coefficients of the edges in the resulting model. This gives Fig. 1D; a model slightly different from the true model in Fig. 1C. Notably, the estimates for the edges in the left-hand-side of the graph are different in the two models because of the different decomposition properties.

The outcomes of the model search for the 100 MC samples are summarized in an empirical power graph in Fig. 1E, i.e. a graph on the variables in the data where the thickness of an edge represents the percentage of times this edge is included in a model search result; this percentage can be interpreted as the power of the model search strategy to detect the edge in the graphical model.

The distribution of the partial γ of the CD edge after model search is estimated from the 100 corresponding values from the MC samples in Fig. 2A. We observe that the distribution resembles the asymptotic theory Gaussian distribution, and that the estimate is unbiased. Model search is relatively efficient for the present model and sample size as the corresponding distribution in the true model has standard deviation of 0.064 which is only a little smaller than the 0.067 from in Fig. 2A. This result is especially remarkable as the correct set of control variables defined by the decomposition of the graphical model Fig. 1A - AB - only occurs once for the 100 models; efficiency is lost with larger sets, and the partial γ is biased for sets that do not include AB.

In practice we set $\gamma = 0$ when an edge is not in the result of the model search. This results for the CD edge in distribution Fig. 2B which shows that estimates of partial γ after model search have distributions with a problematic concentration at zero.

4.2 Non-parametric bootstrap

The power graph of Fig. 1E and the distribution in Fig. 2A are the product of Monte Carlo sampling and cannot be produced for real-life data for which we do not know a true graphical model. However, we may mimic the repeated sampling with a non-parametric bootstrap. In this approach a new data set is obtained by sampling with replacement from the original data. The results from model search on 100 bootstrapped data sets from the first of the MC samples is summarized in the power graph in Fig. 1F. This result is poorer than the true performance of the model search method represented by Fig. 1E. It tentatively captures more of the true model structure than the single model search in Fig. 1D, although the number of errors - edges in the graph that should not be there and vice versa - is only 1 higher in the latter. Notably, while the power graph in itself merely gives an indication of the variation in the model search, it can be used as model search itself when we include all edges with an estimated power above a certain threshold.

The distribution of the partial γ for the CD edge after model search may also be approximated by the bootstrap; this is shown in Fig. 2C. We see that the distribution has long tails caused by some extreme estimates. While Gaussian inference does not seem appropriate, inference based on robust moment estimators may give acceptable efficiency, e.g. the median is -0.16 which has less bias than the mean. Likewise, robust estimators may appropriately be used for inference in Fig. 2D where insignificant γ are set to zero, even though the distribution is far from Gaussian.

5 Discussion

In this paper we presented a simulation procedure as a practical an useful tool for the Monte Carlo evaluation of model search strategies in graphical models for ordinal data. However, we investigated the performance of only one example of search strategy, graphical model and sample size. Future efforts are to broaden the scope of the simulation studies both by a generalization of the simulation procedure, and by examining many more instances of graphical models.

The investigated model search strategy is simple and straightforward, and does not seem to capture weak associations in the model for data of the present size. However, model search results are much poorer if LR tests are used instead of γ coefficients. Future efforts are to evaluate and compare further model search strategies.

A non-parametric bootstrap approach is promising but overrates variance in the distribution of the partial γ coefficient. However, it gives a practicable method to incorporate the model search into the inference on the edges of the graph. The power graph summarizing the bootstrap results may be used as a model search strategy in itself. The optimal parameter settings for such model search is to be determined in future research.

References

AGRESTI, A. 1984: *Analysis of ordinal categorical data*. John Wiley, New York.

DARROCH, J.N., LAURITZEN, S.L. and SPEED, T.P. (1980): Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics 8 522-539*.

DAVIS, J.A. (1967): A partial coefficient for Goodman and Kruskal's gamma. *Journal of the American Statistical Association 62 189-193*.

GOODMAN, L.A. and KRUSKAL, W.H. (1954): Measures of association for cross classifications. *Journal of the American Statistical Association 49 732-764*.

KLEIN, P.J., KEIDING, N. and KREINER, S. (1995): Graphical models for panel studies, illustrated on data from the Framingham heart study. *Statistics in Medicine 14 1265-1290*.

KREINER, S. (1986): Computerized exploratory screening of large-dimensional contingency tables. In: F. De Antoni, N. Lauro and A. Rizzi (Eds.): *Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg, 43-48.

KREINER, S. (1987): Analysis of multidimensional contingency tables by exact methods. *Scandinavian Journal of Statistics 14 97-112*.

KREINER, S. (2003): Introduction to digram. *Department of Biostatistics Research Report 03/10, University of Copenhagen*.

LAURITZEN, S.L. (1996): *Graphical Models*. Clarendon Press, Oxford.

LAURITZEN, S.L. and RICHARDSON, T.S. (2002): Chain graph models and their causal interpretation (with discussion). *Journal of the Royal Statistical Society, Series B 64 321-361*.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000): *Causation, Prediction and Search, 2nd edition*. MIT Press, Cambridge.

Clustering with Finite Mixture Models and Categorical Variables

Cláudia Silvestre[1], Mário Figueiredo[2], and Margarida Cardoso[3]

[1] Escola Superior de Comunicação Social, Instituto Politécnico de Lisboa
Campus de Benfica do IPL, 1549-014 Lisboa, Portugal, *csilvestre@escs.ipl.pt*
[2] Instituto de Telecomunicações, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal, *mtf@lx.it.pt*
[3] ISCTE, Business School, Department of Quantitative Methods
Av. das Forças Armadas 1649-026 Lisboa, Portugal, *margarida.cardoso@iscte.pt*

Abstract. Finite mixture models are widely used for clustering. Maximum likelihood (ML) estimation of finite mixture models via the EM algorithm has some limitations, one being the determination of the number of clusters. The EM-type approach proposed by Figueiredo and Jain (2002) overcomes several of these limitations. In the present work we implement a new version of the referred algorithm for the estimation of a finite mixture of multinomials. An application to clustering TV viewers illustrates the proposed approach.

Keywords: cluster analysis, marketing, finite mixture model, TV audiences

1 Introduction

Finite mixture models are widely used for cluster analysis. Statistical tools based on mixtures models are used in several areas, such as social sciences, medicine, biology, engineering, computer science, and marketing. Finite mixture models with a fixed number of components are usually estimated through likelihood maximization using the *expectation-maximization* (EM) algorithm or variants thereof. The EM algorithm has well known limitations, namely sensitivity to initialization (it may converge to local maxima) and numerical instabilities at the boundary of parameter space.

In finite mixture models, the number of components is, in general, unknown, thus it must be inferred from the data. Estimating the number of components (known as the model selection problem) is often done via information criteria: the Bayesian information criterion (BIC) (Schwarz, 1978), Akaike's information criterion (AIC) (Akaike, 1973), and the minimum message length (MML) criterion (Wallace, 1968).

In the present work, an EM-type algorithm (Figueiredo and Jain, 2002), based on the MML criterion, is used. The novelty of the approach is that it does not rely on selecting among a set of preestimated candidate models, but rather integrates estimation and model selection in a single algorithm.

A new implementation of the Figueiredo and Jain (2002) algorithm is proposed to deal with categorical variables by estimating a finite mixture of multinomials. In the present work, we want to identify clusters of Portuguese TV viewers based on time spent watching TV, and profile the obtained clusters according to their social and demographic characteristics.

The paper is organized as follows: in Section 2, we review finite mixture models and the EM algorithm. In Section 3, we review the criterion developed by Figueiredo and Jain (2002) and the EM-type algorithm which considers this criterion as its objective function. Further, we develop a new approach for the estimation of a mixture of multinomials. In section 4, we use the new algorithm to cluster Portuguese TV viewers and report the obtained results. Finally, in Section 5, conclusions and future research directions are presented.

2 Finite mixture models and the EM algorithm

The basic idea of finite mixture models is that the observations in a sample are assumed to arise from k ($k \geq 2$) clusters that are mixed in unknown proportions. These proportions, or mixing probabilities α_m, are subject to the usual constraints $\sum_{m=1}^{k} \alpha_m = 1$ and $\alpha_m \geq 0$, $m = 1, \ldots, k$. Finite mixture models assume a specific intra-cluster probability function for each cluster in base variables, which may belong to the same family but differ in the parameter values. The purpose of model estimation is to identify the clusters and estimate the parameters of the distribution underlying the observed data within each cluster.

Let $y = \{y_1, \ldots, y_n\}$ be a sample of n independent and identically distributed (i.i.d.) random variables, $Y = \{Y_1, \ldots, Y_n\}$, where each Y_i is a d-dimensional random variable, $Y_i = \{Y_{i1}, \ldots, Y_{id}\}$. Using the total probability theorem, the probability (density) of Y_i is

$$p(y_i|\Theta) = \sum_{m=1}^{k} \alpha_m p(y_i|\theta_m),$$

where $\Theta = (\theta_1, \ldots, \theta_k, \alpha_1, \ldots, \alpha_k)$ is the set of all the parameters of the model and θ_m are the parameters defining the m-th cluster. The log-likelihood of the whole sample, given the independence assumption, is

$$\log p(y|\Theta) = \sum_{i=1}^{n} \log \sum_{m=1}^{k} \alpha_m p(y_i|\theta_m).$$

The ML estimators cannot be found analytically, and the EM algorithm (Dempster et al., 1977) has been often used as an effective method for approximating the corresponding estimates. The basic idea behind the EM algorithm is regarding the observations Y as incomplete data (clusters allocation being unknown). In finite mixture models each variable Y_i for $i = 1, .., n$, (the

incomplete data) is augmented by a component-label variable Z_i which is a set of k binary indicator latent variables, that is, $Z_i = (Z_{i1}, \ldots, Z_{ik})$, with $Z_{im} \in \{0, 1\}$. One and only one of the elements of Z_i equals one, indicating which of the component density describes Y_i; that is, $Z_{im} = 1$ if and only if the density of Y_i is $p(y_i|\theta_m)$. Assuming that the Z_i are i.i.d., following a multinomial distribution of k categories, with probabilities $\alpha_1, \ldots, \alpha_k$, the log-likelihood of a complete data sample (y, z), where $z = \{z_1, \ldots, z_n\}$ is a sample of $Z = \{Z_1, \ldots, Z_n\}$, is given by

$$\log p(y, z|\Theta) = \sum_{i=1}^{n} \sum_{m=1}^{k} z_{im} \log(\alpha_m p(y_i|\Theta_m)). \tag{1}$$

The EM algorithm produces a sequence of estimates $\hat{\Theta}(t)$, $t = 1, 2, \ldots$ using two alternating steps, until some convergence criterion is met.

- **E-step**: calculates the expectation of the complete log-likelihood, with respect to the missing variables, given y and the current parameter estimate: $E[\log p(y, Z|\Theta)|y, \hat{\Theta}^{(t)}]$. Linearity of the complete log-likelihood with respect to Z (see (1)) implies that this expectation equals $\log p(y, w^{(t)}|\Theta)$, where $w^{(t)} \equiv E[Z|y, \hat{\Theta}^{(t)}]$. Since, for some binary variable B, $E[B] = P[B = 1]$, and each Z_{im} is binary, we have

$$w_{im}^{(t)} = P[Z_{im} = 1|y_i, \hat{\theta}^{(t)}] = \frac{\hat{\alpha}_m^{(t)} p(y_i|\hat{\theta}_m^{(t)})}{\sum_{j=1}^{k} \hat{\alpha}_j^{(t)} p(y_i|\hat{\theta}_j^{(t)})}. \tag{2}$$

- **M-step**: updates the estimate: $\hat{\Theta}^{(t+1)} = \arg\max_{\Theta} \log p(y, w^{(t)}|\Theta)$.

3 Methological approach

Figueiredo and Jain (2002) developed a new EM variant for the estimation of a Gaussian mixture model which is based on the MML (minimum message length) criterion of Wallace and Boulton (1968). Their approach seamlessly merges estimation and model selection in a single algorithm. In this work, we implement a new version of that algorithm, for the purpose of clustering categorical data via the estimation of a mixture of multinomials.

3.1 The minimum message length criterion

MML-type criteria choose the model providing the shortest description (in an information theory sense) of the observations (Wallace and Boulton, 1968). According to Shannon's information theory, if Y is some random variable with probability distribution $p(y|\Theta)$, the optimal code-length for an outcome y is $l(y|\Theta) = -\log_2 p(y|\Theta)$, measured in bits (from the base-2 logarithm) and ignoring that $l(y)$ should be integer (Cover and Thomas, 1991). Since

the parameters Θ also need to be encoded, the total message length is in fact $l(y, \Theta) = l(y|\Theta) + l(\Theta)$, where the first part encodes the observation y, and the second the parameters of the model. The several variants of MML and the related MDL (minimum description length) criteria differ essentially in the way they compute $l(\Theta)$ (Lanterman, 2001).

Under an MML criterion, the estimate of Θ is the one minimizing $l(y, \Theta)$. Figueiredo and Jain (2002) proposed a new EM variant which considers the following description length function:

$$l(y, \Theta) = -\log p(\Theta) - \log p(y|\Theta) + \frac{1}{2}\log|I(\Theta)| + \frac{c}{2}\left(1 + \log\frac{1}{12}\right) \qquad (3)$$

where $p(\Theta)$ is a prior on Θ, $p(y|\Theta)$ is the likelihood function, c is the dimension of Θ, $I(\Theta) \equiv -E\left[\frac{\partial^2}{\partial\Theta^2}\log p(Y|\Theta)\right]$ is the expected Fisher information matrix, and $|I(\Theta)|$ its determinant. Since $I(\Theta)$ cannot be obtained analytically for mixtures, the authors replace $I(\Theta)$ by its the complete-data counterpart $I_c(\Theta) \equiv -E\left[\frac{\partial^2}{\partial\theta^2}\log p(Y, Z|\Theta)\right]$. Also they adopt independent Jeffreys' priors for the mixture parameters. The message length function becomes

$$l(y, \Theta) = \frac{N}{2}\sum_{m:\alpha_m>0}\log\left(\frac{n\alpha_m}{12}\right) + \frac{k_{nz}}{2}\log\frac{n}{12} + \frac{k_{nz}(N+1)}{2} - \log p(y|\Theta)$$

where N is the number of parameters specifying each component (the dimension of each θ_m), and k_{nz} the number of components with non zero probability (for details, see Figueiredo and Jain, 2002).

3.2 An EM-type algorithm

In order to estimate a finite mixture model for the purpose of clustering, an EM-type algorithm is used which enables not only the determination of the mixture parameters, but also the estimation of the number of clusters (Figueiredo and Jain, 2002). Furthermore, the proposed approach enables to avoid some estimation problems: it is not as sensitive to the initialization as traditional EM and it avoids parameters near the boundary (where the likelihood is unbounded). The idea is simply to use the EM algorithm to minimize $l(y, \Theta)$.

It can be observed that, w.r.t. the α_m parameters, $l(y, \Theta)$ is formally equivalent to a posterior density under a conjugate Dirichlet-type prior, with negative parameters, $-N/2$. The maximization of $l(y, \Theta)$ w.r.t. these parameters (under the constraints $\sum_{m=1}^{k}\alpha_m = 1$ and $\alpha_m \geq 0$) is thus simply

$$\hat{\alpha}_m^{(t+1)} = \frac{\max\left\{0, \sum_{i=1}^{n} w_{im}^{(t+1)} - \frac{N}{2}\right\}}{\sum_{j=1}^{k}\max\left\{0, \sum_{i=1}^{n} w_{ij}^{(t+1)} - \frac{N}{2}\right\}}, \quad \text{for } m = 1, 2, \ldots, k. \qquad (4)$$

Notice that (4) may return zero values for some component probability estimates, thus "pruning" the mixture model. Since the corresponding θ_m values become irrelevant for the log-likelihood value, their calculation is simply omitted thereafter. The algorithm is initialized with a maximum number of clusters, k_{max}, and automatically removes unnecessary ones.

Since (apart from $p(y|\Theta)$) no other terms depend on the θ_m parameters, the corresponding maximization is as in the standard EM algorithm, and depends on the particular probabilistic model of each mixture component. In the present work, we consider multinomial component models.

The use of the CEM^2 (component-wise EM) algorithm (Celeux et al., 1999) overcomes problems concerning boundary solutions. In CEM^2, after the update of the parameter estimate of each cluster, a full E-step is computed. When arriving at a null α_m, this allows the observations that were allocated to the corresponding component to be redistributed by the remaining clusters. After convergence, a series of further CEM^2 trials are run, with decreasing values of k, trying to find lower values of $l(y, \Theta)$ (Figueiredo and Jain, 2002).

3.3 Estimating a mixture of multinomials

In finite mixture models, the normal distribution is frequently used for continuous variables, whereas the multinomial distribution is commonly used for categorical data. In this work, we adapt the above described algorithm to perform estimation of a mixture of multinomials.

Consider that each Y_i is a d-variate categorical variable $Y_i = \{Y_{i1}, \ldots, Y_{id}\}$, where each component has c_l categories, for $l = 1, \ldots, d$. Conditionally on belonging to the m-th component of the mixture, each Y_{il} is modeled by a multinomial distribution with n_l trials, c_l categories, and non-negative parameters $\theta_{ml} = \{\theta_{mlc},\ c = 1, \ldots, c_l\}$, with $\sum_{c=1}^{c_l} \theta_{mlc} = 1$. Thus, with $\theta_m = \{\theta_{m1}, \ldots, \theta_{md}\}$ and $\Theta = \{\theta_1, \ldots, \theta_k, \alpha_1, \ldots, \alpha_k\}$, the finite multinomial mixture can be written as

$$p(y_i|\Theta) = \sum_{m=1}^{k} \alpha_m p(y_i|\theta_m) = \sum_{m=1}^{k} \alpha_m \prod_{l=1}^{d} \left[n_l! \prod_{c=1}^{c_l} \frac{(\theta_{mlc})^{y_{ilc}}}{y_{ilc}!} \right] \tag{5}$$

where y_{ilc} is the number of observations of y_{il} in category c. Finally, assuming that the set of variables $Y = \{Y_1, \ldots, Y_n\}$ are i.i.d. we obtain the log-likelihood function,

$$\log p(y|\Theta) = \sum_{i=1}^{n} \log \sum_{m=1}^{k} \alpha_m \prod_{l=1}^{d} \left[n_l! \prod_{c=1}^{c_l} \frac{(\theta_{mlc})^{y_{ilc}}}{y_{ilc}!} \right]$$

Applying the EM-type algorithm described in the previous section to estimate the parameters of this mixture models, we obtain the following steps. The E-step is given by (2), where

$$p(y_i|\hat{\theta}_m^{(t)}) = \prod_{l=1}^{d} \left[n_l! \prod_{c=1}^{c_l} \frac{(\hat{\theta}_{mlc}^{(t)})^{y_{ilc}}}{y_{ilc}!} \right]$$

The M-step has two parts: the α_m parameter estimates are updated according to (4), where

$$N = \sum_{l=1}^{d} (c_l - 1),$$

is the number of parameters of each component; the multinomial parameters of each component are updated by

$$\hat{\theta}_{mlc}^{(t+1)} = \frac{\sum_{i=1}^{n} w_{im}^{(t)} y_{ilc}}{n_l \sum_{i=1}^{n} w_{im}^{(t)}}, \text{ for } m = 1, \ldots, k, \; l = 1, \ldots, d, \; c = 1, \ldots, c_l, \quad (6)$$

which are weighted maximum likelihood estimates.

4 Application: clustering Portuguese TV viewers

Audience analysis is a key input to effective marketing strategies. It is both an indicator of TV viewers' behavior, and the basis for advertising pricing. In the present application, the goal is to cluster Portuguese TV viewers based on TV watching patterns. Clusters obtained are then profiled according to socio-demographic characteristics of their members.

4.1 The data

The Portuguese audience analysis panel consists of 1000 representative households, with 2500 viewers. Panel members are carefully selected in order to represent the target population. The households are selected based on geographic location and socio-demographic characteristics, such as gender, age, occupation and social class.

In the present work, we deal with the watch/non watch TV indicators, grouped into intervals of 30 minutes. We have a sample of $n = 464$ viewers during prime-time (20h00-23h00) and want to determine TV watching patterns, but not patterns referring to programs choice. Every viewer Y_i is modeled by a mixture of multinomial distributions with 7 categories, where the first six categories $c = 1, \ldots, 6$ (6 half-hours during prime-time) represent the events of watching TV during the interval c, and the seventh is the event of not watching TV during the prime-time.

The estimates of the multinomial mixture parameters obtained by the method above described are shown in Table 1, where the α parameters indicate the clusters relative dimensions. It is clear that the main differences between clusters regard the time spent not watching TV. For example, a viewer in cluster 1 has a 0,896 probability of not watching TV while 0,357 is the correspondent probability in cluster 3.

In order to obtain a more complete clusters profile, Chi-Square tests of independence between clusters and some additional available variables are

performed. Some significant differences between the clusters emerge of this analysis: cluster 1 has more female viewers and more viewers between 15 and 24 years than clusters 2 and 3; children under 6 years represent the majority of viewers in cluster 3; cluster 2 includes more viewers from the higher social classes.

Table 1. Estimative of model parameters by algorithm described.

Multinomial Categories	Cluster 1 $\hat{\alpha}_1 = 0.293$	Cluster 2 $\hat{\alpha}_2 = 0.442$	Cluster 3 $\hat{\alpha}_3 = 0.265$
20:00 - 20:30	0.018	0.045	0.101
20:30 - 21:00	0.018	0.050	0.108
21:00 - 21:30	0.017	0.053	0.112
21:30 - 22:00	0.017	0.056	0.112
22:00 - 22:30	0.017	0.058	0.108
22:30 - 23:00	0.017	0.057	0.102
not watching	0.896	0.681	0.357

Using the standard EM and the BIC criterion, an alternative solution is obtained with 5 clusters (see Table 2). This solution has many clusters and one of them has a low probability ($\hat{\alpha} = 0.089$), so this clustering is not easily interpretable, thus arguably there was overfitting of the data.

Table 2. Estimative of model parameters by the standard EM.

Multinomial Categories	Cluster 1 $\hat{\alpha}_1 = 0.172$	Cluster 2 $\hat{\alpha}_2 = 0.223$	Cluster 3 $\hat{\alpha}_3 = 0.167$	Cluster 4 $\hat{\alpha}_4 = 0.349$	Cluster 5 $\hat{\alpha}_5 = 0.089$
20:00 - 20:30	0.010	0.054	0.092	0.036	0.122
20:30 - 21:00	0.010	0.060	0.100	0.037	0.129
21:00 - 21:30	0.011	0.066	0.103	0.038	0.135
21:30 - 22:00	0.011	0.071	0.100	0.039	0.137
22:00 - 22:30	0.011	0.073	0.095	0.039	0.136
22:30 - 23:00	0.011	0.071	0.088	0.038	0.130
not watching	0.936	0.605	0.422	0.773	0.211

5 Conclusions

In the present work we deal with the estimation of a finite mixture of multinomial variables for the purpose of clustering. An EM-type approach is implemented which is based on the proposal of Figueiredo and Jain (2002). The

main advantage of this approach is related with the model selection procedure, specifically with the determination of the number of clusters. An application to clustering Portuguese TV viewers illustrates the proposed approach. Results obtained indicate that a more parsimonious solution is derived which has fewer clusters than the solution derived by a standard EM procedure. In fact, although the BIC criterion seems to have a general good performance (Fonseca and Cardoso, 2007) it may sometimes overestimate the adequate number of clusters (e.g Chiu et al., 2001). Therefore the present results seem promising, in particular regarding the issue of finite mixture model selection. Naturally, further research is required to validate these empirically drawn conclusions.

References

AKAIKE, H.(1973): Maximum Likelihood Identification of Gaussian Autorregressive Moving Average Models. *Biometrika 60, 255-265.*

CELEUX, G., CHRÉTIEN, S., FORBES, F. AND MKHADRI, A. (1999): A Component-Wise Algorithm for Mixtures. *Technical Report 3746,* INRIA, France.

CHIU, T., FANG, D., CHEN, J., WANG, Y. and JERIS C. (2001): A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* San Francisco, California, 263-268

COVER, T., THOMAS, J. (1991): *Elements of Information Theory.* Wiley.

DATIA, N., MOURA-PIRES, J., CARDOSO, M. and PITA, H. (2005): Temporal Patterns of TV watching for Portuguese Viewers. *EPIA 2005*, 151-158.

DEMPSTER, A. LAIRD, N., RUBIN, D.(1977): Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm. *Journal of Royal Statistical Society 39 (B), 1-38.*

ELMORE, R. and WANG, S. (2003): Identifiability and Estimation in Finite Mixture Models with Multinomial Components. *Technical Report.*

FIGUEIREDO, M.A.T. and JAIN, A.K. (2002): Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 381-396.*

FONSECA, J. and CARDOSO, M. (2007): Mixture-model Cluster Analysis Using Information Theoretical Criteria. *Intelligent Data Analysis 11, 155-173.*

LANTERMAN, A. (2001): Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Order Estimation. *International Statistical Review 69, 185-212.*

SCHWARZ, G.(1978): Estimating the Dimension of a Model. *The Annals of Statistics 6, 461-464.*

WALLACE, C. and BOULTON, D. (1968): An Information Measure for Classification. *The Computer Journal 11, 195-209.*

Part III

Classification and Discrimination

The Unimodal Supervised Classification Model in a New Look at Parameter Set Estimation

Hugo Alonso, Joaquim F. Pinto da Costa, and Teresa Mendonça

Faculdade de Ciências da Universidade do Porto
Rua do Campo Alegre, 687, 4169-007 Porto, Portugal
{*hugo.alonso,jpcosta,tmendo*}*@fc.up.pt*

Abstract. The goal of this paper is to introduce a new look at the problem of parameter set estimation and propose solving it using the unimodal model (Pinto da Costa et al. (2008)). This is done in the context of system identification. Hence, given a model structure and measurements from a system, we are interested in estimating a set rather than a point for the model parameterization. Our interest reflects the uncertainty associated with the measurement process. The new look consists in dealing with the problem of parameter set estimation as an ordinal supervised classification problem. Therefore, a supervised classification technique which takes into account the existence of an order relation between the classes should be used to solve it. Here, we propose a new one: the unimodal model. Our approach is illustrated using a mechanical system and the performance of the unimodal model is compared with the performance of other methods in a simulation study.

Keywords: parameter set estimation, ordinal supervised classification, unimodal model, neural networks

1 Introduction

Parameter set estimation plays an important role in system identification and control in areas such as mechanics (Madi et al. (2004)), physics (Hurtig and Yurkovich (2002)) and chemistry (Braems et al. (2005)). The objective is to estimate a region in the parameter space where a parameterization of interest is, given some data from a system. We will consider regions associated with classes of systems and introduce in this work the use of a supervised classification method to predict the class of a system, and hence the associated region, given some data from it.

The system data used for estimation purposes are invariably corrupted by random noise during the measurement process. In spite of this, the usual methods of parameter set estimation, like the ones described in the previous references, approach the problem in a deterministic fashion. They do so in a bounding error context and assuming a value for the error bound. Their goal is to over-bound the set of all parameterizations which are consistent with the measured data, the system model and the value taken for the error bound. To that end, hyper-rectangles and ellipsoids are usually considered.

The over-bound is determined from an iterative process with high computational demands and its size depends in particular on the value assumed for the error bound. This presents at least two drawbacks. Firstly, issues of divergence and time to convergence may arise, making these iterative methods unsuited for practical purposes. Secondly, the over-bound may be very large in size, and thus little informative about the location of a parameterization of interest. This may be due, for instance, to a bad choice of its geometry, or to a high value taken for the error bound, particularly when the true error bound is not known.

Here, we propose a different approach, a stochastic one. The error is also supposed to be bounded, but no value for the error bound is assumed. The space of parameters is *a priori* partitioned into several regions. This is done in such a way that different regions are associated with different classes of systems, *i.e.*, sets of systems sharing a similar behavior or similar characteristics, which differ from those of systems in other classes. Approaching the problem from this perspective, any supervised classification method could now be applied in order to predict the right class (region). However, there is a natural order relation between these different classes (regions) and so a classification method specific for these type of classes will be used; namely the one introduced in (Pinto da Costa et al. (2008)).

We will start by defining our classes (of systems). These classes will correspond to regions in a partition of the space of parameters. Our goal is to estimate a region (predict the class), from those in the partition considered, where a parameterization of interest is, given some data from a system. The geometry of the regions may vary from problem to problem and in the same problem according to the definition of the classes. Hence, it is not restricted to hyper-rectangles and ellipsoids, but rather to the nature of the problem in hands, which in our opinion makes more sense. Moreover, the size of the regions depends in particular on their number, *i.e.*, on the number of classes.

Given a new system, our method will find the probability of each class for that particular system and predict the most likely class. We will need to use available information regarding data collected and the corresponding class of other systems to build the decision function. The problem of parameter set estimation can and will therefore be viewed by us as an ordinal supervised classification problem. Finally, note that a suitable classifier is able to non-iteratively suggest a class for a system, thus avoiding issues of convergence.

The rest of this paper is organized as follows. In Section 2, we formalize our view of a parameter set estimation problem as an ordinal supervised classification problem. The unimodal model is presented in Section 3. In Section 4, we illustrate the application of our approach in a simulation study involving a mechanical system. The conclusions are given in Section 7.

2 Parameter set estimation as a supervised classification problem with ordered classes

The systems Σ considered throughout this paper are assumed to follow the fixed-regressor model

$$y_i = f(\mathbf{x}_i, \theta) + \varepsilon_i,$$

where y_i is the output measured at time t_i with a bounded error ε_i, $\mathbf{x}_i$ is a vector of r regressors measured up to t_i, and f is a function representing a theoretical relationship between y_i and $\mathbf{x}_i$ such that

$$\mathcal{E}[y_i|\mathbf{x}_i] = f(\mathbf{x}_i, \theta),$$

where $\mathcal{E}$ denotes mathematical expectation and θ is a vector of p parameters. Now, let Θ represent the space of parameters, a compact (bounded and closed) set of dimension p. Define a partition P of Θ into K regions $R_1, \ldots, R_K$ and a total order $\leq_P$ on P such that

$$R_1 \leq_P \ldots \leq_P R_K.$$

In the definition of P, different regions R_k and R_ℓ of values for the parameters should be associated with different classes C_k and C_ℓ of systems. A class of systems is understood as a set of systems sharing a similar behavior or similar characteristics. Two classes differ in the behavior or characteristics of their systems. Formally, a system Σ belongs to class C_k if the data $(y_i, \mathbf{x}_i)_{i=1,\ldots,n}$ collected from Σ, with $n \geq p$, is such that $y_1, \ldots, y_n$ is a random sample from a distribution where for some $\theta \in R_k$ one has $\mathcal{E}[y_i|\mathbf{x}_i] = f(\mathbf{x}_i, \theta)$ and $\mathcal{E}[(y_i - f(\mathbf{x}_i, \theta))^2|\mathbf{x}_i] = \mathcal{E}[\varepsilon_i^2]$ for $i = 1, \ldots, n$. Finally, the order $\leq_P$ between $R_1, \ldots, R_K$ should be related with a natural ordering existing between $C_1, \ldots, C_K$, namely

$$R_k \leq_P R_\ell \Leftrightarrow C_k \leq_C C_\ell. \tag{1}$$

Example 1. A system of considerable interest in vibration analysis is the mass-spring-damper system in Fig. 1 (Meirovitch (1986)), where y is a variable representing the displacement of the system from the equilibrium position, which coincides with the position in which the spring is unstretched, and m, s and d are parameters corresponding respectively to the mass, the spring constant and the damper constant.

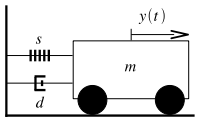
Under non-zero initial conditions, *i.e.*, a displacement $y(0) \neq 0$ or a velocity $\dot{y}(0) \neq 0$, and for a fixed mass m, each system is either under-damped (oscillatory) or over-damped (non-oscillatory) depending on the values of the parameters s and d. An example of the displacement characteristic of these two types of systems is depicted in the left side of Fig. 2. A way to quantify the amount of damping in a system is by using the so-called viscous damping factor, given by $\zeta = \frac{d}{2\sqrt{ms}}$. In fact, the greater the value of ζ, the greater the amount of damping. Furthermore, a system is under-damped if $0 < \zeta \leq 1$

and over-damped if $\zeta > 1$, *i.e.*, if s and d are such that $0 < \frac{d}{\sqrt{s}} \leq 2\sqrt{m}$ and $\frac{d}{\sqrt{s}} > 2\sqrt{m}$, respectively. Hence, there is an association between the two classes of systems, C_1: under-damped and C_2: over-damped, and two regions in the space Θ of the parameters s and d, given by

$$R_k = \left\{ (s,d) \in \Theta : b_{k-1} \leq \frac{d}{\sqrt{s}} \leq b_k \right\} \tag{2}$$

for $k = 1, 2$ with $(b_0, b_1, b_2) = (0, 2\sqrt{m}, +\infty)$. This is illustrated in Fig. 2. Moreover, there is a natural ordering between the classes, namely $C_1 \leq_C C_2$, which is associated with the order $\leq_P$ in the partition $P = \{R_1, R_2\}$ of Θ defined as

$$R_k \leq_P R_\ell \Leftrightarrow b_k \leq b_\ell.$$



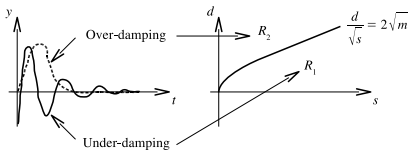


Fig. 1. Mass-spring-damper system. **Fig. 2.** Relation between classes and regions.

Back from the example, it is assumed that a set $T = \{(D_{\Sigma_j}, R_{\Sigma_j})_{j=1,\ldots,N}\}$ is known *a priori*, where $D_{\Sigma_j} = \{(y_i, \mathbf{x}_i)_{i=1,\ldots,n_j}\}$ is a data set obtained from the j-th system Σ_j, with $n_j \geq p$, and R_{Σ_j} is the region of Θ in P where the model of Σ_j is supposed to have its parameterization. Note that this is equivalent to know the set $\{(D_{\Sigma_j}, C_{\Sigma_j})_{j=1,\ldots,N}\}$, and in particular the classes assigned to N systems. Based on this prior knowledge, our problem is to find for a general system Σ the region R_Σ where the parameterization θ is, given a data set D_Σ. Given the association between R_Σ and the class C_Σ of Σ, this is the same as finding C_Σ, given D_Σ. Hence, this problem can be viewed as a supervised classification problem. The goal is then to define a classifier represented by a map $g_T : D \to \{C_1, \ldots, C_K\}$ which minimises some cost functional with respect to T. Bayes decision theory suggests classifying Σ into the class C_k maximising the *a posteriori* probability $P(C_k|D_\Sigma)$, *i.e.*, g_T should be such that $g_T(D_\Sigma) = \arg\max_{C_k}\{P(C_k|D_\Sigma)\}$. To that end, g_T must estimate the *a posteriori* probabilities. Furthermore, this should be done taking into account the fact that there is an order relation between the classes. The reason is simple and can be illustrated using the example above as a motivation. Assume that there are four instead of two classes, namely C_1: very under-damped, C_2: under-damped, C_3: over-damped and

C_4: very over-damped. Given a new query D_Σ, if the highest *a posteriori* probability is, for instance, $P(C_3|D_\Sigma)$, then we should have $g_T(D_\Sigma) = C_3$. Now, if we use a classifier which does not take into account the order relation between the classes, the second highest *a posteriori* probability can be, for instance, $P(C_1|D_\Sigma)$. This does not make any sense to have the most likely an over-damped system and the second most likely a very under-damped system. Given that there is an order relation between the classes, C_2 and C_4 are closer to C_3 and therefore the second highest *a posteriori* probability should be attained in one of these classes. This means that if the most likely is an over-damped system, then the second most likely should be either a under-damped or a very over-damped system. More generally, the probabilities should decrease monotonically to the left and to the right of the class where the maximum probability is attained. This is the main idea behind the method described next.

3 The unimodal model for parameter set estimation

The unimodal model (Pinto da Costa et al. (2008)) is a new supervised classification technique which takes into account the existence of an order relation between the classes. As suggested by the name and according to the motivation presented at the end of the previous section, it does so by assuming that the random variable class follows a unimodal discrete distribution. In this context, the output of a classifier where the *a posteriori* class probabilities are estimated is obliged to be unimodal, *i.e.*, to have only one local maximum. There are different ways to impose unimodality. In (Pinto da Costa et al. (2008)), we suggested assuming either a particular unimodal discrete distribution, like the binomial and Poisson's, or no distribution at all, and simply train the classifier in such a way that its output becomes unimodal. These two approaches were named parametric and non-parametric, respectively. In the first case, all there is to do by the classifier is to estimate some parameters of the assumed distribution; in the second case, the distribution itself has to be estimated non-parametrically. In all practical experiments, the parametric approach led to the best results, in particular when the binomial distribution was considered. The superior performance when using this distribution was also justified in theoretical terms. For these reasons, our focus here will be only on the binomial model. Furthermore, since the classifier chosen by us is a neural network (Hastie et al. (2001)), although others could have been considered, we refer hereafter to a binomial network. Its description applied to the problem of parameter set estimation is given next.

In the binomial network, the output values follow the binomial distribution $B(K-1, p_s)$. As this distribution takes integer values in the set $\{0, 1, \ldots, K-1\}$, we will take value 0 to represent class C_1, 1 to C_2 and so on until value $K-1$ to represent class C_K. This distribution is unimodal in most cases and when it has two modes these are for contiguous values,

which in the context of our problem could be interpreted as a suggestion that the parameterization is not identifiable. Now, since K is known, the only unknown parameter is the probability of success p_s. Hence, we consider a network architecture as in Fig. 3 and train it to adjust all connection weights from layer 0 to layer L. Note that the connection from layer L to layer $L+1$ has a fixed weight equal to one and serves only to forward the value of p_s to the output layer of the network where the probabilities from the binomial distribution are calculated. For a given data set D_Σ collected from a system Σ, the output of layer L will be a single numerical value in the range $[0, 1]$, which we denote by p_Σ. Then, the probabilities $P(C_k|D_\Sigma)$ in layer $L+1$ are calculated from the binomial distribution:

$$P(C_k|D_\Sigma) \;=\; \frac{(K-1)!p_\Sigma^{k-1}(1-p_\Sigma)^{K-k}}{(k-1)!(K-k)!},\; k = 1, 2, \ldots, K.$$

In fact, they can be calculated recursively to save computing time (see (Pinto da Costa et al. (2008))). When the training instance is presented, the error is here defined as

$$\sum_{k=1}^{K} \left(P(C_k|D_\Sigma) - \delta(k - h(C_\Sigma))\right)^2,$$

where $\delta(k) = 1$ if $k = 0$, 0 otherwise, and $h(C_\Sigma)$ is the number corresponding to the class C_Σ of Σ. The binomial network is trained to minimise the average value over all training cases of this error. Finally, in the testing phase, we choose the class C_k which maximises the probability $P(C_k|D_\Sigma)$.

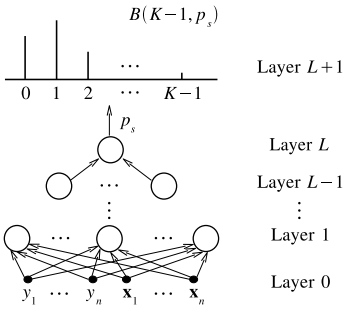


Fig. 3. Binomial network.

Before ending this section, it should be noted that regardless of the number K of classes, *i.e.*, of regions in which the parameter space is partitioned, there is only one "true" output in the binomial network, which corresponds to the estimate of the probability of success. A practical consequence is that this is a simple model to train and to apply in the testing phase.

4 Simulation study

The simulation study here presented is related with the example given in Section 2. A set of 1000 mass-spring-damper systems was generated by taking $m = \frac{1}{4}$ and 1000 pairs (s, d) uniformly distributed in $\Theta = [0, 1]^2$. Under the initial conditions $y(0) = 0$ and $\dot{y}(0) = 1$, the displacement of each of those systems was simulated from 0 to 10 units of time, which is sufficiently long for most of them to reach the equilibrium position, *i.e.*, $y = 0$. Then, the 1000 displacements were all sampled in 100 time instants equally spaced in $[0, 10]$ and the resulting samples were contaminated with random noise $\varepsilon \sim N(0, 0.005^2)$. Our goal is to estimate a region in the parameter space $\Theta = [0, 1]^2$ where a system has its parameterization (s, d), given the associated time series. Since this is done in a context where we look at parameter set estimation as a supervised classification problem, we decided to compare the unimodal model implemented by the binomial network with two supervised classification techniques. The first one is the neural network implementation of the algorithm by Frank and Hall (2001), which also takes into account the existence of an order relation between the classes. The second one is the nearest neighbors method with the Dynamic Time Warping distance (NN-DTW) (see for example (Ratanamahatana and Keogh (2005)) and references therein), which does not take into account the existence of such an order relation, but which is specifically intended for time series classification contrary to the other two techniques. In order to study how the performance of the three models varies with the number of regions considered *a priori* in the parameter space, Θ was partitioned into 5 and 10 regions, defined as in (2) with $(b_0, \ldots, b_5) = (0, \frac{1}{3}, \frac{2}{3}, 1, \sqrt{2}, +\infty)$ and $(b_0, \ldots, b_{10}) = (0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1, \sqrt{5/4}, \sqrt{5/3}, \sqrt{5/2}, \sqrt{5}, +\infty)$, respectively. In both cases, a class would be assigned to each system according to the region in Θ where the associated parameterization (s, d) is, *i.e.*, according to the value of $\frac{d}{\sqrt{s}}$. However, with the aim of simulating the existence of some uncertainty in the knowledge about each system, the assignment was done using instead the value of $\frac{d}{\sqrt{s}} + \eta$, where $\eta \sim N(0, \sigma_\eta^2)$ with $\sigma_\eta = 0.075$ and $\sigma_\eta = 0.0375$ for 5 and 10 regions, respectively. As a result, 14.4% and 14.3% of the systems were correspondingly misclassified. Having the 1000 time series labeled, they were randomly split into training and testing patterns. Two training sets with different sizes, one with 50 patterns, the other with 100, were defined. The testing set gathered the remaining 900 patterns. Our goal in doing this was to enable an investigation on how the amount of prior knowledge about the problem in hands influences the performance of the three models. The neural network models were trained using the Levenberg-Marquardt algorithm with back-propagation and weight decay. Furthermore, their inputs were either the original time series or their most important principal components. We used 5-fold cross-validation to select the regularization parameter and the number of hidden neurons in the networks, and to select the number of neighbors in

the NN-DTW model.

Table 1 shows the actual and predicted accuracies, quantified by the misclassification error rate (MER). It can be seen that the unimodal model outperforms the other two methods, being NN-DTW the worst of all. This suggests in particular that the respect for the order relation between the classes is more important than taking into account the fact that the data are time series. The better performance of the unimodal model becomes more marked with an increase in the number of regions considered *a priori* in the parameter space. Furthermore, there is an improvement in the value of MER with an increase in the number of training patterns. In the 5 regions case, 100 patterns were enough to achieve a MER (17.4%) very close to the Bayes error (14.4%). In the 10 regions case, there is nevertheless the suggestion that more patterns are needed to further approximate the value of MER (24.6% was the best) to the Bayes error (14.3%). Moreover, it seems that the more the training patterns, the less the difference between using the original times series or their most important principal components, which are just a few contrary to the large length of the series.

Model	Data	5 regions		10 regions	
		Training patterns		Training patterns	
		50	100	50	100
Unimodal	Orig.	19.4; 8.0±8.4	17.4; 15.0±5.0	31.9; 22.0±4.5	24.6; 22.0±7.6
	PCs	28.2; 18.0±13.0	18.8; 16.0±8.2	39.0; 26.0±15.2	25.7; 22.0±9.8
Frank and Hall's	Orig.	22.1; 12.0±8.4	20.0; 15.0±8.7	40.6; 30.0±18.7	34.6; 28.0±4.5
	PCs	29.7; 20.0±12.3	20.2; 17.0±6.7	46.3; 32.0±4.5	37.1; 29.0±2.2
NN-DTW	Orig.	24.8; 26.0±18.2	28.0; 29.0±9.6	49.2; 42.0±13.0	44.3; 41.0±10.8

Table 1. Misclassification error rate (%): testing set; 5-fold cross-validation (mean±standard-deviation). Orig. stands for original, PCs for principal components.

5 Conclusions

This paper introduced a different look at parameter set estimation, where each problem is treated as an ordinal supervised classification task. In this context, we suggested the use of a new supervised classification method which takes into account the existence of an order relation between the classes, called the unimodal model. This method outperformed two others in a simulation study involving a mechanical system. In the future, we plan to consider other kinds of systems, namely biomedical ones, and compare the performance of the unimodal model with other methods, like some of the usual deterministic approaches to parameter set estimation. Other applications of the unimodal model can be found in (Pinto da Costa et al. (2008)).

6 Acknowledgments

The first author was supported by Unidade de Investigação Matemática e Aplicações, Universidade de Aveiro, Portugal, through FCT's program "Ciência e Tecnologia e Inovação", co-financed by the EU fund FEDER.

References

BRAEMS, I., BERTHIER, F., and FRANGER, S. (2005): Set-membership techniques for reliable electrochemical parameter estimation. In: *Proc. WMSCI 2005*, 6, 41-46.

FRANK, E., and HALL, M. (2001): A simple approach to ordinal classification. In: *Proc. ECML 2001*, 1, 145-156.

HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. (2001): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

HURTIG, J., and YURKOVICH, S. (2002): Parameter set estimation for non-linear systems. *Int. J. Control*, 75(2), 111-122.

MADI, M.S., KHAYATI, K., and BRIGAS, P. (2004): Parameter estimation for the LuGre friction model using interval analysis and set inversion. In: *Proc. IEEE SMC 2004*, 1, 428-433.

MEIROVITCH, L. (1986): *Elements of vibration analysis*. McGraw-Hill, 2nd ed..

PINTO DA COSTA, J.F., ALONSO, H., and CARDOSO, J.S. (2008): The unimodal model for the classification of ordinal data. *Neural Networks*, 21, 78-91.

RATANAMAHATANA, C.A., and KEOGH, E. (2005): Three myths about Dynamic Time Warping. In: *Proc. SDM 2005*, 506-510.

Robust Supervised Classification with Gaussian Mixtures: Learning from Data with Uncertain Labels

Charles Bouveyron[1] and Stéphane Girard[2]

[1] Samos-Matisse, CES, Université Paris 1 Panthéon-Sorbonne
90 rue de Tolbiac, 75634 Paris Cedex 13, France.
Email: *charles.bouveyron@univ-paris1.fr*
[2] Mistis, INRIA Rhône-Alpes – LJK
Inovallée, 655 avenue de l'Europe, Montbonnot, 38334 Saint-Ismier, France.
Email: *stephane.girard@inrialpes.fr*

Abstract. In the supervised classification framework, the human supervision is required for labeling a set of learning data which are then used for building the classifier. However, in many applications, the human supervision is either imprecise, difficult or expensive. In this paper, the problem of learning a supervised classifier from data with uncertain labels is considered and a model-based classification method, called Robust Mixture Discrimant Analysis (RMDA), is proposed to solve this problem. The idea of the proposed method is to confront an unsupervised modeling of the data with the supervised information carried by the labels of the learning data in order to detect inconsistencies. The method is able afterward to build a robust classifier taking into account the detected inconsistencies into the labels.

Keywords: supervised classification, data with uncertain labels, model-based classification, robustness, label noise, high-breakdown methods

1 Introduction

In the supervised classification framework, the human supervision is required to associate labels with a set of learning observations in order to construct a classifier. However, in many applications, this kind of supervision is either imprecise, difficult or expensive. For instance, in bio-medical applications, domain experts are asked to manually label a sample of learning data (MRI images, DNA micro-array, ...) which are then used for building a supervised classifier. The cost of the supervision phase is usually high due to the difficulty of labeling complex data. Furthermore, an human error is always possible in such a difficult task and an error in the supervision phase could have big effects on the decision phase, particularly if the size of the learning sample is small. It is therefore very important to be able to provide supervised classifier flexible enough to deal with data with uncertain labels.

1.1 The label noise problem

Since the main assumption of supervised classification is that the labels of learning samples are true, existing methods giving a full confidence to the labels of the learning data naturally provide disappointing classification results when the learning dataset contains some wrong labels. Particularly, model-based discriminant analysis methods such as Linear Discriminant Analysis (LDA) or Mixture Discriminant Analysis (MDA, see Hastie *et al.* (1996)) are sensitive to label noise. This sensitivity is mainly due to the fact that these methods estimate the model parameters from the learning data and label noise naturally perturbs these estimates.

1.2 Related works

For years, researchers tried to solve the problem of learning a supervised classifier from data with uncertain labels using different strategies. Early approaches tried to clean the data by removing the misclassified instances using some form of nearest neighbor algorithm (*e.g.* Gates (1972)) and further checking by human experts (*e.g.* Guyon *et al.* (1996)). However, approaches cleaning the data could introduce a bias in the learning procedure by removing correctly labeled instances. Therefore, other researchers proposed not to remove any learning instance and to build instead supervised classifiers robust to label noise. Hawkins *et al.* (1997) and Bashir *et al.* (2005) focused on robust estimation of the model parameters in the mixture model context but only observed a slight reduction of the average probability of misclassification. Similarly, Mingers (1989) and Sakakibara (1993) proposed noise-tolerant approaches to make decision tree classifiers robust to label noise. Among all these solutions, the model proposed by Lawrence *et al.* (2001) and extended in Li *et al.* (2007) has the advantage of explicitely including the label noise in the model with a sound theoretical foundation.

1.3 The proposed approach

In this paper, we propose a supervised classification method, called Robust Mixture Discriminant Analysis (RMDA), made for dealing with label noised data. The main idea of our approach is to compare the supervised information given by the learning data with an unsupervised modeling of the data based on the Gaussian mixture model. With such an approach, if some learning data have wrong labels, the comparison of the supervised information with an unsupervised modeling of the data will allow to detect the inconsistent labels. It will be possible afterward to build a supervised classifier by giving a low confidence to the learning observations with inconsistent labels.

The remainder of this paper is organized as follows. The model of the proposed method RMDA is presented in Section 2. Section 3 is devoted to the inference aspects. Experimental studies on simulated and real datasets are reported in Section 4. Finally, some extensions are discussed in Section 5.

2 Our approach

In order to compare the supervised information given by the learning data with an unsupervised modeling of the data, we propose to use an unsupervised mixture model in which the supervised information is introduced.

2.1 Model

We consider a mixture model in which two different structures coexist: an unsupervised structure of K clusters (represented by the random variable S) and a supervised structure, given by the learning data, of k classes (represented by the random variable C). As in the standard mixture model, we assume that the data $(x_1, ..., x_n)$ are independent realizations of a random vector $X \in \mathbb{R}^p$ with density function:

$$p(x) = \sum_{j=1}^{K} P(S = j)p(x|S = j), \tag{1}$$

where $P(S = j)$ is the prior probability of the jth cluster and $p(x|S = j)$ is the conditional density of the jth cluster. Let us now introduce the supervised information carried by the learning data. Since $\sum_{i=1}^{k} P(C = i|S = j) = 1$ for all $j = 1, ..., K$, we can introduce this quantity in (1) to obtain:

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} P(C = i|S = j)P(S = j)p(x|S = j), \tag{2}$$

where $P(C = i|S = j)$ can be interpreted as the probability that the jth cluster belongs to the ith class. It measures the consistency between classes and clusters. The conditional density $p(x|S = j)$ is modelled by a Gaussian distribution with mean μ_j and covariance Σ_j. Under this assumption and adopting the classical notations of Gaussian mixtures, (2) can therefore be rewritten as:

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} r_{ij}\pi_j\phi(x; \mu_j, \Sigma_j), \tag{3}$$

where $r_{ij} = P(C = i|S = j)$, $\pi_j = P(S = j)$ and ϕ is the Gaussian density. Therefore, equation (3) exhibits both the "modeling" part of our approach, based on the Gaussian mixture model, and the "supervision" part through the parameters r_{ij}.

2.2 Link with Mixture Discriminant Analysis

It is possible to establish a link between model (3) and the fully supervised method MDA in which each class is modeled by a mixture of K_i Gaussian

densities. Denoting by $K = \sum_{i=1}^{k} K_i$ the total number of Gaussian components, and keeping in mind the notations of the previous paragraph, MDA assumes that the class conditional density of the ith class, $i = 1, ..., k$, is

$$p(x|C = i) = \sum_{j=1}^{K} \pi_{ij} \phi(x; \mu_j, \Sigma_j), \tag{4}$$

where $\pi_{ij} = P(C = i, S = j)$ is the prior probability of the jth mixture component of the ith class. Note that $\pi_{ij} = 0$ if the jth mixture component is not included in the ith class. Moreover, remarking that, $\pi_{ij} = r_{ij}\pi_j$, we obtain

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} r_{ij} \pi_j \phi(x; \mu_j, \Sigma_j), \tag{5}$$

which formally corresponds to model (3). The main difference is that, in the MDA case, the labels are certain (fully supervised context). Thus $r_{ij} = P(C = i|S = j)$ is known and reduces to $r_{ij} = 1$ if the jth mixture component belongs to the ith class and $r_{ij} = 0$ otherwise. Consequently, in the case where the labels of learning data are all consistent with the modeling of these data, RMDA should provide the same classifier as MDA.

2.3 Classification step

In model-based discriminant analysis, new observations are usually assigned to a class using the maximum a posteriori (MAP) rule. The MAP rule assigns a new observation x to the class for which x has the highest posterior probability. Therefore, the classification step mainly consists in calculating the posterior probability $P(C = i|X = x)$ for each class $i = 1, ..., k$. It can be expressed using the Bayes rule as:

$$P(C = i|X = x) = \sum_{j=1}^{K} r_{ij} P(S = j|X = x). \tag{6}$$

Therefore, the classification step of RMDA relies on (6) and requires the estimation of the consistency probabilities r_{ij} as well as the unsupervised classification probabilities $P(S = j|X = x)$.

3 Estimation procedure

Due to the nature of the model proposed in Section 2, the estimation procedure is made of two steps corresponding respectively to the unsupervised and to the supervised part of the comparison. The first step consists in estimating the parameters of the mixture model in an unsupervised way leading to the clustering probabilities $P(S = j|X = x)$. In the second step, the parameters r_{ij} linking the mixture model with the information carried by the labels of the learning data are estimated by maximization of the likelihood.

3.1 Estimation of the parameters π_j, μ_j and Σ_j

In this first step of the estimation procedure, we do not use the labels of the data in order to form K homogeneous groups. Therefore, this step consists in estimating the parameters of the Gaussian mixture: the proportions π_j, the means μ_j and the variance matrices Σ_j, for $j = 1, ..., K$. The classical procedure for estimating the parameters of a Gaussian mixture model is the Maximum Likelihood (ML) method. Unfortunately, it is not possible to find directly a solution of the maximum likelihood problem. In such a case, the Expectation-Maximization (EM) algorithm proposed by Dempster *et al.* (1977) allows to obtain the ML estimates of the parameters using an iterative procedure.

3.2 Estimation of the parameters r_{ij}

In this second step of the procedure, we introduce the labels of the data to estimate the parameters r_{ij} and we use the parameters learned in the previous step as the mixture parameters. Since we consider a supervised problem, the labels $c_1, \ldots, c_n$ of the learning data $x_1, \ldots, x_n$ are known, and we can therefore introduce $\mathcal{C}_i = \{x_\ell, \ell = 1, \ldots, n \text{ such that } c_\ell = i\}$. From (6), the log-likelihood associated to our model can be expressed as:

$$\begin{aligned}\ell(R) &= \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \log P(X = x, C = i), \\ &= \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \log \left(\sum_{j=1}^{K} r_{ij} P(S = j | X = x) \right) + C^{ste},\end{aligned}$$

where C^{ste} does not depend on R. This relation can be matricially rewritten as:

$$\ell(R) = \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \log < R_i, \Psi(x) > + C^{ste},$$

with the $\mathbb{R}^K$-vectors $\Psi(x) = (P(S = 1|X = x), \ldots, P(S = K|X = x))^t$ and $R_i = (r_{i1}, \ldots, r_{iK})^t$. Consequently, we end up with a constrained optimization problem:

$$\begin{cases} \text{minimize} & \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \log < R_i, \Psi(x) >, \\ \text{with respect to} & R_i \in [0,1]^K, \text{ for all } i = 1, ..., k, \\ \text{and} & \sum_{i=1}^{k} R_i = \mathbb{I}, \end{cases}$$

where $\mathbb{I}$ denotes the $\mathbb{R}^K$ vector $(1, \ldots, 1)^t$. Since it is not possible to find an explicit solution to this optimization problem, an iterative optimization procedure has to be used to compute the maximum likelihood estimators of the parameters R_i.

4 Experimental results

In this section, we present experimental results on artificial and real datasets in situations illustrating the problem of supervised classification under uncertainty. For the sake of simplicity, the following experiments consider only the problem of discriminating two classes.

4.1 Experimental setup

In the following studies, we consider the general problem of label switching between the classes. In this case, complex models are very sensible but parsimonious models can also be affected if the contamination rate is high. In order to simulate a label noise, the observation labels have been switched ($1 \rightarrow 2$ if the true label is 1 and $2 \rightarrow 1$ otherwise) following a Bernoulli distribution with parameter η ranging from 0 to 1 and representing the contamination rate. In all studies, the performance of the methods, measured by the correct classification rate, was computed on a test dataset and the experiments have been repeated 25 times in order to average the classification results.

4.2 Simulated data study

For this first experiment, we simulated the data following the mixture model of MDA and RMDA. The simulated dataset is made of 2 classes and each class was modeled with a Gaussian mixture of 2 components in a 25-dimensional space. We used for the mixture components of each class a diagonal Gaussian model with a covariance matrix $\Sigma_j = \sigma_j I_p$ where $\sigma_j \in \mathbb{R}$. The means of the different mixture components were chosen in order to obtain two separated enough classes. The left panel of Figure 1 shows the performance of LDA, MDA, RLDA (proposed by Lawrence *et al.* (2001)) and RMDA (introduced in this paper) on the simulated dataset for different contamination rates. On the one hand, LDA and MDA appear to be sensitive to contamination. Particularly, MDA becomes very instable for contamination rates higher than 0.2. The behavior of these two supervised methods is not surprising since they both have full confidence in the labels of the data. On the other hand, RLDA turns out to be more robust than LDA but its performance decreases quickly for contamination rates higher than 0.325. Finally, RMDA appears to be particularly robust for a large panel of contamination rates (up to 0.4). Furthermore, this experiment shows as well that RMDA is as efficient as MDA when there is no label noise and thus demonstrates the equivalence between both methods in this special case.

4.3 Real dataset study

We consider now a dataset from the real world, called USPS-24, extracted from the well-known USPS dataset[1]. The learning dataset is made of 1383

[1] The USPS dataset is available for download at *www.kernel-machines.org*.

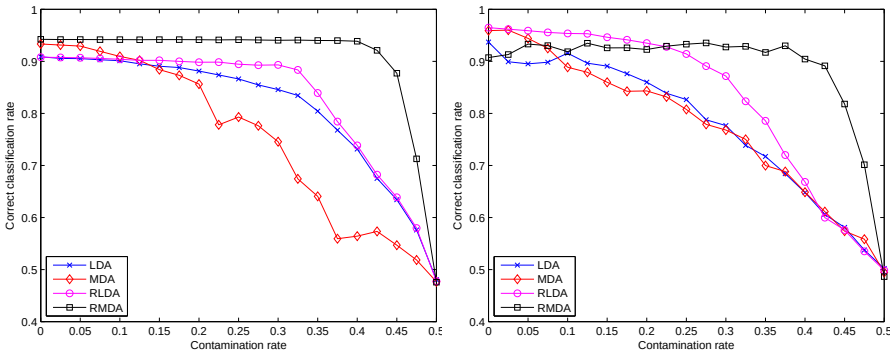


Fig. 1. Performance of LDA, MDA, RLDA and RMDA for different contamination rates on the simulated dataset (left) and on the USPS-24 dataset (right).

observations: 731 observations belonging to the class of the digit 2 and of 652 observations belonging to the class of the digit 4. Similarly, the test dataset contains 298 elements: 198 and 200 observations respectively from the classes of the digits 2 and 4. These two classes have been chosen since they are the classes with the highest misclassification rate in the original USPS dataset. Each observation of the USPS-24 dataset corresponds to a 16×16 grey level image of a digit and is represented as a 256-dimensional vector. Figure 2 shows some examples of the dataset. For both MDA and RMDA, each class was modeled by a mixture of 5 Gaussians and, due to the high dimension of the data, we used for each mixture component a diagonal Gaussian model with a covariance matrix $\Sigma_j = \sigma_j I_p$ where $\sigma_j \in \mathbb{R}$. The right panel of Figure 1 shows the performance of LDA, MDA, RLDA and RMDA on the USPS-24 dataset for different contamination rates. As in the previous experiment, LDA and MDA appear to be very sensitive to contamination. RLDA is again more robust than LDA and MDA but its performance decreases quickly for contamination rates higher than 0.2. Finally, RMDA appears to be robust for contamination rates up to 0.375 and to be almost as efficient as the other methods when the label noise is low. In this experiment, RMDA has therefore demonstrated its ability to deal with label noise in real and complex situations.

5 On the way to weakly-supervised classification

We have proposed in this paper a supervised classification method, called Robust Mixture Discriminant Analysis (RMDA), for performing classification in the presence of label noise. The experimental studies have shown that RMDA is as efficient as fully supervised techniques when the label noise is low and that RMDA is particularly robust to label noise, even in very noisy situations. In addition, we believe that this work opens the way to a new

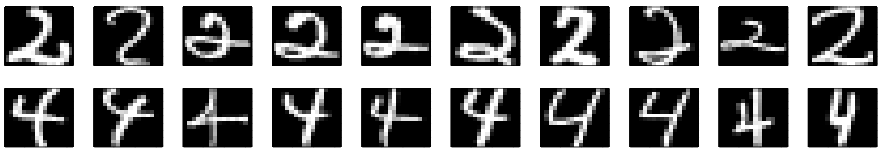


Fig. 2. Some examples of the USPS-24 dataset.

kind of learning in which a complete human supervision is not possible and replaced by a less expensive supervision. For example, in computer vision, the problem of object recognition requires that human experts segment a very large number of images for each object category. It is clear that this work is impossible given the infinite number of existing object categories. However, it is easy to obtain images containing a given object (using Google Image for instance) and to assume that all pixels of these images are representative of the studied object even if we know that it is wrong. By doing that, we consciously introduce a label noise between the class "object" and the class "background" but, using the approach proposed in this paper, it will be possible to identify all pixels which actually belong to the class "object" and, finally, localize the studied object in the images. To summarize, the classification method proposed in this paper could be the first step to solve an important problem of learning theory: how to learn under weak supervision?

References

BASHIR, S. and CARTER, E. (2005): High breakdown mixture discriminant analysis. *Journal of Multivariate Analysis, 93(1), pp. 102-111.*

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, 39(1), pp. 1-38.*

GATES, G. (1972): The reduced nearest neighbor rule. *IEEE Transactions on Information Theory, 18(3), pp. 431-433.*

GUYON, I., MATIC, N. and VAPNIK, V. (1996): Discovering informative patterns and data cleaning. *Knowledge Discovery and Data Mining, pp. 181-203.*

HASTIE, T. and TIBSHIRANI, R. (1996): Discriminant analysis by Gaussian mixtures. *J. of the Royal Statistical Society, 58, pp. 155-176.*

HAWKINS, D. and MCLACHLAN, G. (1997): High-Breakdown Linear Discriminant Analysis. *J. of the American Statistical Association, 92(437), pp. 136-143.*

LAWRENCE, N. and SCHOLKOPF, B. (2001): Estimating a Kernel Fisher Discriminant in the Presence of Label Noise. *In Proc. of 18th International Conference on Machine Learning, pp. 306-313.*

LI, Y., WESSELS, L., DE RIDDER, D. and REINDERS, M. (2007): Classification in the presence of class noise using a probabilistic Kernel Fisher method. *Pattern Recognition, 40, pp. 3349-3357.*

MINGERS, J. (1989): An empirical comparison of pruning methods for decision tree induction. *J. of Machine Learning, 4(2), pp. 227-243.*

SAKAKIBARA, Y. (1993): Noise-tolerant Occam algorithms and their applications to learning decision trees. *J. of Machine Learning, 11(1), pp. 37-62.*

Optimal Screening Methods in Gene Expression Profiles Classification

Sandra Ramos[1], Antónia Amaral Turkman[2] and Marília Antunes[2]

[1] Instituto Superior de Engenharia do Porto - Instituto Politécnico do Porto
Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
sfr@isep.ipp.pt
[2] Faculdade de Ciências, Universidade de Lisboa
DEIO, Bloco C6, Campo Grande 1749-016 Liboa, Portugal
antonia.turkman@fc.ul.pt, marilia.antunes@fc.ul.pt

Abstract. We propose the application of a Bayesian Optimal Screening Method to classify an individual in one of two groups (presence/absence of disease) based on the observation of pairs of covariates, namely the expression level of pairs of genes. The method is general and can be applied to any correlated pair of covariates with bivariate normal distribution or that can be tranformed in a bivariate normal. In this case, the boundaries of the optimal screening region are approximated by a quadratic function of the screening variables. The classifier was evaluated on data from three gene expression studies - Leukemia, Prostate and Breast cancers - found in the literature. The classification error rates were calculated using the leave-one-out cross-validation approach.

Keywords: screening methods, DNA microarrys, classification

1 Introduction

Microarray technology is a powerful tool for genomic research, which allows the monitoring of expression profiles for tens of thousands of genes in parallel and is already producing huge amounts of data (Duggan et al. (1999)). However, the number of profile measurements per experimental study remains quite small, usually fewer than one hundred. The small-sample dilemma in the statistical methods for classification in micrroarray data is well documented in the literature (Dudoit et al. (2003)), with some simplifying assumptions appearing as necessary (such as the reduction of the dimensionality of the data). Geman et al. (2004) and Bo et al. (2002) propose the use of marker gene pairs for classification. In this paper, we propose the use of optimal screening methods applied to pairs of gene expression levels for classification purposes.

The screening method consists in the identification of successful individuals of the population, based on the observation, $\mathbf{x}$, of a feature vector $\mathbf{X}$ for each individual.

The purpose of screening is to find a region $C_{\mathbf{x}}$ such that if $\mathbf{x} \in C_{\mathbf{x}}$ the probability that the individual is considered a success is maximized (Turkman and Amaral Turkman (1989)). In section 2 we describe the fundamental concepts of the screening methodology, applied to classification based on the observation of expression levels of pairs of genes.

We demonstrate the usefulness of this methodology using several public data sets involving leukemia, breast and prostate cancers. The performance of the procedure will be evaluated using leave-one-out cross-validation and will be displayed for each data set. Results are presented in section 3 and conclusions and final remarks in section 4.

2 Method

We suggest an application of the screening methodology in supervised classification based on observation of pairs of genes. In this section we explain the main theoretical tools that are necessary to understand the methodology, and its application in classification problems.

2.1 Optimal screening methods in classification of gene pairs

Consider two genes whose expression levels $\mathbf{X} = (X_1, X_2)$ (measured using DNA microarrays) are regarded as random variables, each profile $\boldsymbol{X}$ having a true class label in $\{0, 1\}$. Let Y be a binary random variable that assumes value 1 (success) if the profile $\boldsymbol{X}$ has class 1 and assumes the value 0 otherwise. Suppose that we have a random sample of n individuals, $\mathcal{D} = \{(y_1, x_{11}, x_{21}), \cdots, (y_n, x_{1n}, x_{2n})\}$, for which the true classification label is known. The optimal screening problem has been stated by Turkman and Amaral Turkman (1989) and in this case the optimal region is

$$C_{\mathbf{x}} = \left\{\mathbf{x} \in \mathbb{R}^2 : P\left(Y = 1 | \mathbf{x}, \mathcal{D}\right) \geq k\right\} \tag{1}$$

or equivalently

$$C_{\mathbf{x}} = \left\{\mathbf{x} \in \mathbb{R}^2 : \frac{P\left(Y = 1|\mathcal{D}\right) p\left(\mathbf{x}|Y = 1, \mathcal{D}\right)}{\sum_{i=0,1} P\left(Y = i|\mathcal{D}\right) p\left(\mathbf{x}|Y = i, \mathcal{D}\right)} \geq k\right\} \tag{2}$$

where k is such that

$$P\left(\mathbf{X} \in C_{\mathbf{x}}|\mathcal{D}\right) = \alpha. \tag{3}$$

We consider the case where Y has a Bernoulli distribution with parameter θ ($Y \sim \text{Ber}(\theta)$, $\theta \in (0, 1)$), and for $i = 0, 1$, $\log \mathbf{X}|Y = i$ has a bivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and precision matrix $\boldsymbol{\Lambda}_i$ $\left(\log \mathbf{X}|Y = i \sim N_2\left(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i = \boldsymbol{\Sigma}_i^{-1}\right)\right)$. The model parameters are $(\theta, \boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1)$, where $\boldsymbol{\Theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$, $i = 0, 1$. We assume that *a priori* the parameters θ, $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ are independent.

If we assume a Beta prior distribution for θ ($\theta \sim$Be(a,b), $a>0$, $b>0$) and a conjugate prior for $(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)$ of the form $p(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i)\, p(\boldsymbol{\Lambda}_i)$, where $p(\boldsymbol{\mu}_i|\boldsymbol{\Lambda}_i)$ is $N_2(\boldsymbol{\mu}_{0i}, c_i\boldsymbol{\Lambda}_i)$ and $p(\boldsymbol{\Lambda}_i)$ is Wishart$_2(\alpha_i, \boldsymbol{\beta}_i)$, the predictive distribution of a future observation in class $Y=i$ is a non-centred, scaled, bivariate Student with $2\alpha_{ni}$ degrees of freedom, (Aitchison and Dunsmore, 1975)

$$\log \mathbf{X}|(Y=i, \boldsymbol{D}) \sim \mathrm{St}_2\left(\boldsymbol{\mu}_{ni}, (c_i+n_i+1)^{-1}(c_i+n_i)\,\alpha_{ni}\boldsymbol{\beta}_{ni}^{-1}, 2\alpha_{ni}\right) \quad (4)$$

where

$$\alpha_{ni} = \alpha_i + \frac{1}{2}(n_i - 1),$$

$$\boldsymbol{\mu}_{ni}(c_i+n_i)^{-1}(c_i\boldsymbol{\mu}_{0i} + n_i\bar{\mathbf{x}}_i),$$

and

$$\boldsymbol{\beta}_{ni} = \boldsymbol{\beta}_i + \frac{1}{2}S_i + \frac{1}{2}(n_i+c_i)^{-1}(\boldsymbol{\mu}_{0i} - \bar{\mathbf{x}}_i)(\boldsymbol{\mu}_{0i} - \bar{\mathbf{x}}_i)^t.$$

The predictive probability of a future individual to be a success, $(Y=1)$, is

$$\gamma = P(Y=1|\mathcal{D}) = \frac{n_1 + a}{n + a + b}, \quad (5)$$

with $n = n_0 + n_1$, where n_i is the number of individuals in the sample for which $Y=i$.

The following predictive probabilities are called operating characteristics (OC) of the screening region,

1. $\alpha = P(\mathbf{X} \in C_{\mathbf{x}}|\mathcal{D})$
2. $\gamma = P(Y=1|\mathcal{D})$
3. $\delta = P(Y=1|\mathbf{X} \in C_{\mathbf{x}}, \mathcal{D})$
4. $\varepsilon = P(Y=1|\mathbf{X} \notin C_{\mathbf{x}}, \mathcal{D})$

2.2 Classification

Prior to the classification procedure, the selection of differentially expressed genes, results in a family $\mathcal{P} = \{\mathbf{X}_j = (X_{j1}, X_{j2}), j = 1, \ldots, m\}$ of m distinct pairs. Usually m is very small, that is, there are only a few pairs of genes good for discrimination purposes; for example, in two of the three experiments presented here there is only one such pair, and in the Leukemia study there are three pairs (Geman et al. (2004), Bo and Jonassen (2002)). We use $\mathcal{P}$ as input of our method. For each pair in $\mathcal{P}$, the classification rule and the operating characteristics are obtained for several values of k, defined in (1). The optimal k is the one which renders the best collection of operating characteristics and gives the smallest number of profiles incorrectly classified.

Consider a new individual, with family of profiles $\mathcal{P}$ with m pairs. Based on the $j-th$ pair, he is classified in $C_j = 1$ if the observed profile $\mathbf{x}_j \in C_{\mathbf{x}_j}$.

Otherwise he is classified in $C_j = 0$. Let δ_j be the corresponding predictive probability of success given that his $j-th$ profile belongs to $C_{\mathbf{x}_j}$. Then, the final classification rule is given by

$$C = \text{Round}\left(\frac{C_1\delta_1 + C_2\delta_2 + \cdots + C_m\delta_m}{\delta_1 + \delta_2 + \cdots + \delta_m}\right). \tag{6}$$

where $C_j \in \{0, 1\}$

2.3 Error estimation

The classification performance, for all data sets, is assessed using leave-one-out cross validation procedure. The Leukemia study (Golub et al. (1999)) has one training and test data set, but in order to use the same method of error estimation on all studies, we combined these two data sets into one.

3 Application

3.1 Data sets

Prostate study- The data is drawn from the study of prostate cancer reported in Singh et al. (2002). This study assigns profiles to either tumor or normal tissues classes based on expression values for 12600 genes. There are $n_1 = 52$ prostate tumor samples and $n_0 = 50$ non-tumor samples, selected from among several hundred radical prostatectomy patients. The top scoring gene pair used as input for the screening classifier is M84226 and M55914. The joint behaviour of this pair of genes, as we will see, is highly discriminative of prostate tumor versus non-tumor samples, yielding an error of 5.43%.

Leukemia study- This study (Golub et al, (1999)) compares two different types of leukemia (Acute Myeloid and Acute Lymphoplastic, ALL *vs* AML) with 7129 probes (6187 human genes) from 27 samples of ALL and 11 samples of AML. There is also a test set consisting of 34 samples (20 ALL and 14 AML). In order to use the same method of error estimation on all studies, we combined the two data sets into one of size $n = 72$ (47 ALL and 25 AML). Negative values due to normalization and/or background correction were eliminated in order to apply the logarithmic transformation and hence the final data set has size $n = 63$ with $n_1 = 38$ ALL samples and $n_0 = 25$ AML samples. The screening classifier uses three gene pairs (five genes) and classifies 60 samples correctly out of 63.

Breast Study- The data set (Huang et al. (2003)) consists of gene expression profiles measured in 52 women with breast cancer. $n_0 = 34$ women did not experience recurrence of the tumor during a 3 years time period and $n_1 = 18$ experienced the recurrence of the tumor. The screening classifier uses only one pair of genes ($38895 - i - at$ and $32625 - at$). The estimated error rate is 11.54%.

3.2 Classification results

For each study, and for each gene pair in the corresponding $\mathcal{P}$ family, the approximated optimal screening region was computed together with the operating characteristics. All the procedure is automatically implemented in R.

For each study, we present the scatterplot of the log expression levels for two genes - the unique pair for Prostate (Fig. 1) and Breast data (Fig. 3) and one of the three pairs for the Leukemia data (Fig. 2).

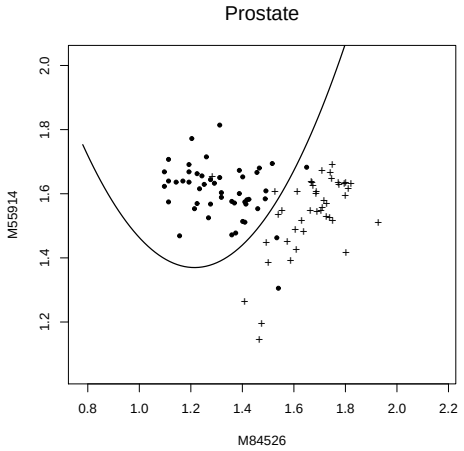


Fig. 1. Scatterplot for a pair of genes for Prostate study. Classes are represented using dots (C_1) and crosses (C_0). The axes represent the logarithm of the expression levels of the two genes. The curve, $x_2 = 4.3726 - 4.9457x_1 + 2.0364x_1^2$, represents the decision boundary.

Table 1 shows the operating characteristics of the optimal screening region for the represented gene pairs. The estimated prediction error rate of the classifier for each study is displayed in Table 2.

Problem	k	$P(Y=1\|\mathcal{D})$	$P(\mathbf{X}\in C_{\mathbf{x}}\|\mathcal{D})$	$P(Y=1\|\mathbf{X}\in C_{\mathbf{x}},\mathcal{D})$	$P(Y=1\|\mathbf{X}\notin C_{\mathbf{x}},\mathcal{D})$
Prostate	0.63	0.5319	0.5153	0.9399	0.0981
Leukemia	0.70	0.6000	0.5456	0.9814	0.1421
Breast	0.42	0.3519	0.3128	0.8458	0.1269

Table 1. Operating characteristics for the best value of k.

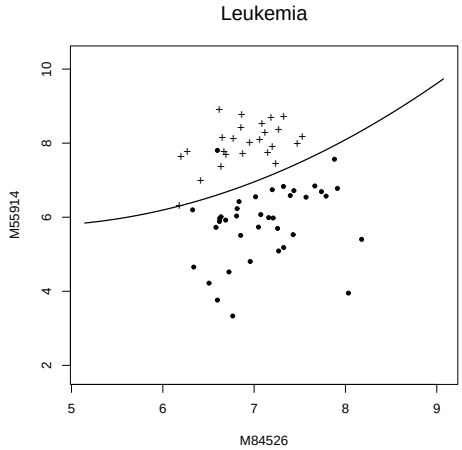


Fig. 2. Scatterplot for a pair of genes for Leukemia study. Classes are represented using dots (C_1) and crosses (C_0). The axes represent the logarithm of the expression levels of the two genes. The curve, $x_2 = 9.5632 - 1.6949x_1 + 0.1889x_1^2$, represents the decision boundary.

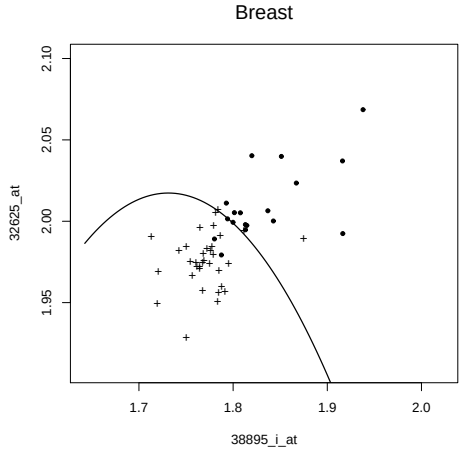


Fig. 3. Scatterplot for a pair of genes for Breast study. Classes are represented using dots (C_1) and crosses (C_0). The axes represent the logarithm of the expression levels of the two genes. The curve, $x_2 = -9.7506 + 13.5948x_1 - 3.9263x_1^2$, represents the decision boundary.

4 Conclusions and further work

We have introduced a new classification methodology for microarray data based entirely on expression levels of pairs of genes. In bivariate normal case,

Problem	Sample Size	# genes	Error (%)
Prostate	102	2	5.43%
Leukemia	63	5	4.76%
Breast	52	2	11.54%

Table 2. Classification error rate for a pair of genes for each study. The results are based on leave-one-out cross-validation.

the optimal screening region is approximated by a quadratic function of the screening variables. This method is more general than the one used by Geman et al. (2004) since they advocate the use of $x_1 = x_2$ as a decision boundary. We have chosen leave-one-out cross-validation to estimate the error rate of the classifier. For the three data sets presented here the estimated prediction rate is very satisfactory.

The computer code used to obtain the optimal screening regions, compute the operating characteristics and perform the final classification has been written in R.

It is our aim to make the programs fully automatic so that it can be generally used and made available to the R community.

5 Acknowledgments

The authors acknowledge the support given by CEAUL (Centro de Estatística e Aplicações da Universidade de Lisboa) and FCT (Fundação da Ciência e Tecnologia) - projects FCT/POCI/2010 and FCT/PTDC/MAT/64353/2006.

References

AITCHISON,J. and DUNSMORE, I.R. (1975): *Statistical Prediction Analysis. Cambdridge University Press.*

BO, T.H. and JONASSEN, I. (2002): New feature subset selection procedures for classification of expression profiles. *Genome Biology, 3(4): research0017.1-0017.11.*

DUDOIT and FRIDLYAND, J. (2003): *Classification in microarrays experiments. In T. Speed, editor, Statistical Analysis of Gene Expression Microarray Data. Chapman and Hall.*

DUGGAN, D.J., BITTNER, M., CHEN, Y., MELTZER, P. and TRENT, J.M. (1999): Expression profiling using cDNA microarrays. *Nature Genetics Supplement, 21:10-14.*

GEMAN, D., d'AVIGNON, C., NAIMAN, D. and WINSLOW, R. (2004): Classification Gene Expression Profiles from Pairwise mRNA Comparisons. *Statistical Applications in Genetics and Molecular Biology, 3 (1).*

GOLUB, T.R., SLOMIN, D.K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J.P., COLLIER, H., LOH, M.L. et al. (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science, 286, 531-537*.

HUANG, E., CHENG, S. H., DRESSMAN, H., PITTMAN, J., TSOU, M., HORNG, C., BILD, A., INVERSEN, E., LIAO, M. and CHEN, C. (2003): Gene expression predictors of breast cancer outcomes. *The Lancet, 361 (9369), 1590-1596*.

SINGH, D., FEBBO, P.G., JACKSON, D.G., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A.A., D'AMICO, A.V., RICHIE, J.P., LANDER, E. S., LODA, M., KANTOFF, P.W., GOLUB, T.R. and SELLERS, W.R. (2002): Gene expression correlates of clinical prostate cancer behaviour. *Cancer Cell, 1(2):203-209*.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2003): Class prediction by nearest shrunken centroids, with applications to DNA microarray. *Statistical Science, 18, 104-117*.

TURKMAN, K.F. and AMARAL TURKMAN, M.A. (1989): Optimal Screening Methods. *Journal of the Royal Statistical Society, B 51, 287-295*.

Part IV

Clustering

A Method for Outlier Detection in Grouped Data

Daniela G. Calò

Department of Statistics, University of Bologna
via delle Belle Arti, 41, 40126 Bologna, Italy, *danielagiovanna.calo@unibo.it*

Abstract. The method proposed by Hadi (1994) for multiple outlier detection in a single group of multivariate data is adapted to the multiple cluster setting. The idea is to replace, in Hadi's algorithm, the Gaussian distribution and the Mahalanobis distance with the K-component normal mixture model (with $K > 1$) and a coherent measure of discrepancy from a mixture distribution, respectively. The performance of the proposed procedure is illustrated on a real data set and compared, through a simulation study, with the method proposed by Caroni and Billor (2007) for detecting multiple outliers in grouped multivariate data.

Keywords: robust clustering, forward search, normal mixture models, outlier detection

1 Introduction

The problem of detecting multiple outliers has been deeply studied for the case of a single multivariate sample. In the last few years some methods have been proposed for identifying outlying points in grouped multivariate data. The interest in this topic is due to the fact that many clustering algorithms are derailed by the presence of observations which do not belong to any cluster, and that robust methods do not usually identify such particular points.

In this context, Caroni and Billor (2007) have recently proposed an adaption of the BACON algorithm (Billor et al. (2000)) to the multiple cluster setting. BACON is a computationally efficient version of the multiple outlier detection method proposed by Hadi (1994) for a single sample of n independent d-dimensional observations ($d > 1$). Hadi's method starts from a basic subset B_m of $m = m_0$ observations ($d < m_0 < n$) that can be safely presumed to be free of outliers, and then performs a "forward search" (FS): the subset is allowed to grow one observation at a time as long as the new subset remains clean of outliers. More precisely, at each step of the search a Gaussian model is fitted to the current subset B_m, and the subset is updated as the $m + 1$ observations with the smallest values of Mahalanobis distance. Denoted by $D^2_{[m+1]}$ the maximum squared Mahalanobis distance in the new basic subset, the search stops if $D^2_{[m+1]} \geq c_{n,d}\chi^2_{d,\alpha/n}$, where $c_{n,d}$ is a correction factor. In such a case, the tested observation and all observations not included in the subset, if any, are nominated as outliers.

In this paper, we extend Hadi's procedure to the multiple cluster setting and compare our proposal with Caroni and Billor's one. We adapt Hadi's algorithm by replacing the Gaussian distribution with the K-component normal mixture model ($K > 1$). When the number of components, K is treated as fixed – as we do indeed – finite mixtures represent a natural way of extending FS-based outlier detection, because maximum likelihood estimation is not robust against outlying data. For this extension to be possible, a coherent criterion for ordering observations by closeness to the fitted model is needed, as well as a testing procedure to be used as a stopping rule.

2 A measure of typicality in mixture models

The issue of assessing how much an observation is typical of a normal mixture model has been considered by McLachlan and Basford (1988). Let X be a d-dimensional random vector distributed according to a mixture with K normal components:

$$p(x) = \sum_{k=1}^{K} w_k \phi(x|\mu_k, \Sigma_k), \tag{1}$$

where each Gaussian density $\phi(\cdot)$ is parameterized by its mean vector $\mu_k \in \mathbb{R}^d$ and covariance matrix Σ_k, belonging to the set of positive definite $d \times d$ matrices, and w_k $(k = 1, \ldots, K)$ are mixing proportions.

McLachlan and Basford suppose that a sample of n observations $\{x_{hk}; h = 1, \ldots, n_k, k = 1, \ldots, K\}$ stemming from model (1) is available, where x_{hk} are known to come from the k-th normal population, P_k. In this situation, the following two distributional results can be proved (for $k = 1, \ldots, K$):

for a generic observation x_{hk} coming from P_k, the quantity

$$\frac{(\frac{\nu_k n_k}{d}) D^2(x_{hk}; \bar{x}_k, S_k)}{(\nu_k + d)(n_k - 1) - n_k D^2(x_{hk}; \bar{x}_k, S_k)} \tag{2}$$

has the F_{d,ν_k} distribution, where $D^2(.; \bar{x}_k, S_k)$ denotes the Mahalanobis squared distance from group k (with $\bar{x}_k$ and S_k denoting the mean vector and covariance matrix of group k), and $\nu_k = n_k - d - 1$ $(k = 1, \ldots, K)$;

for a new unclassified observation $y \in \mathbb{R}^d$, the quantity

$$\frac{n_k(\nu_k + 1)}{(n_k + 1)d(\nu_k + d)} D^2(y; \bar{x}_k, S_k) \tag{3}$$

has the F_{d,ν_k+1} distribution, where $D^2(.; \bar{x}_k, S_k)$ and ν_k are defined as before.

These results are (approximately) valid also in case of unclassified data $\{x_j; j = 1, \ldots, n\}$ – like those considered in the present paper – provided that observations are first clustered by fitting a K-component heteroscedastic

normal mixture model and the resulting K clusters are taken as a "true classification" of the data: that is, in (2) and (3), $\bar{x}_k$ and S_k are replaced by $\hat{\mu}_k$ and $\hat{\Sigma}_k$, and n_k represents the number of observations put in cluster k.

In light of the results mentioned above, McLachlan and Basford conclude that an assessment of how typical an observation $z \in \mathbb{R}^d$ is of the k-th component of the mixture is given by the tail area to the right of the observed value of (2) or (3) under the F distribution with the appropriate degrees of freedom, depending on whether z belongs to the available set of observations ($z = x_{hk}$) or not ($z = y$). Denoted this tail area by $a_k(z)$, they propose to assess how typical an observation z is of the mixture by the following measure:

$$a(z) = \begin{cases} a_k(z) & \text{if } z = x_{hk} \\ \max_k a_k(z) & \text{if } z = y \end{cases} \tag{4}$$

and to assess z as being atypical of the mixture if:

$$a(z) \leq \alpha, \tag{5}$$

where α is some specified threshold. Thus, an observation will be labelled as outlying of the mixture if it is outlying of all the mixture components.

3 Mixture-based forward search

Given a sample of n independent d-dimensional observations $B = \{x_j; j = 1, \ldots, n\}$ stemming from a mixture of $K > 1$ normal populations, and supposing that some contamination is present in the data, we want to identify clusters and outliers at the same time. We propose to modify Hadi's FS algorithm by replacing the Gaussian distribution and the Mahalanobis distance with the normal mixture model in (1) and the typicality measure in (4), respectively. Thus, at each step of the search, the mixture is fitted to the current basic subset B_m and the subset is updated as the $m+1$ most typical observations; the search stops when rule (5) signals that a potential outlier is is going to be included.

More precisely, given a specified significance level α, the proposed mixture-based forward search (MFS) algorithm proceeds as follows:

- *Phase* 0: (Initialization).
  For $k = 1, \ldots, K$, we find a subset of m_{0k} observations that are presumably located inside cluster k. Let B_{m_0} be the union of these K subsets, consisting of $m = m_0$ observations. It is taken as the starting basic subset.

- *Phase* 1: (Mixture-based clustering).
  Model (1) is fitted to the current basic subset B_m via the EM algorithm: let $\{\hat{w}_{k,m}, \hat{\mu}_{k,m}, \hat{\Sigma}_{k,m}; k = 1, \ldots, K\}$ denote the resulting set of parameter estimates. Observations in B_m are clustered accordingly; let m_k be the current size of cluster k, with $\sum_{k=1}^{K} m_k = m$.

- *Phase* 2: (Order by typicality).
  For $j = 1, \ldots, n$, the typicality of x_j is assessed according to the measure in (4) (depending on whether $x_j \in B_m$ or $x_j \in B \setminus B_m$, respectively); we denote the quantities involved in the rhs of (4) by by $a_{k,m}(x_j)$, where the subscript m helps to remind that computation involves the estimates and cluster sizes obtained on subset B_m: $\{\hat{\mu}_{k,m}, \hat{\Sigma}_{k,m}, m_k; k = 1, \ldots, K\}$. Then, the n observations are sorted in decreasing order of typicality; let $x_{[i],m}$ be the observation with the i-th ordered typicality value (computed on subset B_m.

- *Phase* 3: (Test).
  Observation $x_{[m+1],m}$ is tested to be an outlier with respect to the mixture fitted to B_m. If the following inequality holds:

$$a_{\hat{k},m}(x_{[m+1],m}) > (\alpha/K)/(m_{\hat{k}} + 1), \tag{6}$$

  where $\hat{k} = argmax\{[a_{k,m}(x_{[m+1],m})](m_k+1))\}$, then the basic subset B_m is updated as the set of the $m + 1$ most typical observations:

$$B_{m+1} = \{x_{[i],m} : i = 1, \ldots, m + 1\},$$

  and we return to Phase 1, with m increased by 1. On the contrary, if inequality (6) is not true, then the algorithm stops and all observations not included in B_m, if any, are nominated as outliers.

After this summary of the proposed procedure, some remarks are needed.

The starting basic subset B_{m_0} is selected as the union of K subsets, each located inside a distinct cluster. Caroni and Billor's procedure starts form a step analogous to *Phase 0*, and selects the starting subset by the following algorithm: the K-medoids method is used to find K representative observations that are centrally located, each in the respective cluster; then, for $k = 1, \ldots, K$, the m_{0k} points with the smallest Euclidean distances from the k-th medoid are selected. In the simulation reported in Section 4, Caroni and Billor's initialization strategy was used in our algorithm too, for the sake of comparison. More generally, the K starting subsets could also be specified by the researcher, after some inspection of the data: in this context, the exploratory tools proposed in Chapter 7 of Atkinson et al. (2004) or in Atkinson and Riani (2007) can be useful.

In Phase 1, mixture fitting is obtained after multiple runs of EM algorithm from different random starts, in order to achieve stable results (avoid spurious estimators). The use of mixture-based clustering allows for partially overlapping clusters. This makes the proposed algorithm more flexible than Caroni and Billor's one, where clustering is carried out by the nearest centroid rule (with Mahalanobis distance as a metric).

The ordering criterion used in Phase 2 has an appealing feature: for a given mixture component, it depends on the Mahalanobis distance from that

component. As a result, the proposed strategy amounts to perform K single-sample forward searches simultaneously, each one referred to a distinct mixture component.

A special attention must be deserved to the testing phase. Let $x_{[m+1],m}$ be the observation that is being tested to be atypical of the mixture fitted on B_m. There are two "orderings" we have to account for:

1. Observation $x_{[m+1],m}$ is currently the least typical observation with respect to any cluster. We allow for this by Bonferroni correction. It is applied by adjusting the tail area to the right of $x_{[m+1],m}$ for each $k = 1, \ldots, K$. Since the test is for an outlier in a sample of size $m_k + 1$, the right tail area is adjusted as follows:

$$[a_{k,m}(x_{[m+1],m})](m_k + 1).$$

2. Rule (5) amounts to testing the outlyingnes of $x_{[m+1],m}$ with respect to the component, $\hat{k}$, which $x_{[m+1],m}$ is most typical of. Following McLachlan and Peel (2000), the ordering of the K tipicality values is accounted for by Bonferronization. Therefore, in the rhs of inequality (6), the specified nominal significance level α is replaced with α/K.

Finally, Caroni and Billor (2007) use a $\chi^2_{d,\alpha/n}$ cutoff for Mahalanobis squared distance. On the contrary, our strategy involves F-bounds, which do not require asymptotic arguments, with cluster-specific Bonferronization.

4 Some empirical results

MFS procedure has been implemented in `R`, resorting to package `mvtnorm` for random number generation and package `MCLUST` (Fraley and Raftery, 2006) for mixture modelling.

As an illustrative example, we report MFS results on a real data set, containing the values of $d = 3$ variables on $n = 103$ investment funds operating in Italy. In light of the results of some preliminary explorative analysis (see Atkinson et al. (2004)), we decided to take $K = 2$ and identified a starting subset of $m_0 = 30$ observations. For $\alpha = 0.05$, observations 52 and 77 were nominated as outliers by MSF; with $\alpha = 0.10$ the following larger set of observations was identified: $21, 50, 52, 54, 77$ (see Figure 1).

MFS has been carried out on the simulation study presented in Caroni and Billor (2007), in order to compare the performance of MFS with that of Caroni and Billor's algorithm for various configurations of data, with and without outliers. In both settings, 1000 independent samples of $n = 100$ observations were generated, with $d = 2$ or 5 (due to the too small sample size to dimensionality ratio, we decided to skip the case $d = 8$).

In the uncontaminated setting, samples of $n = 100$ observations are simulated from a d-dimensional uniform mixture of K normal components, with

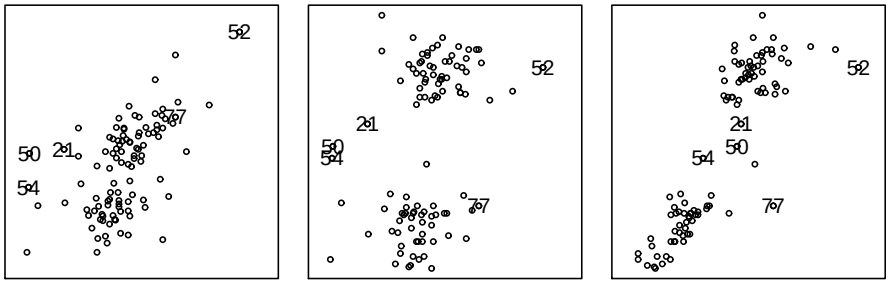


Fig. 1. Funds data. Scatterplot of X_2 against X_1 (left panel), X_3 against X_1 (middle panel) and X_3 against X_2 (right panel).

$K = 2$ or 3. In the case $K = 2$, the components are centered in $\mu_1 = (0, 0)$ and $\mu_2 = (0, 0.866)$ (with zero in the remaining dimensions if $d > 2$); the covariance matrix is set to I_d for both groups, but the variance in the first dimension is increased to 2 for the second group. In the case $K = 3$, the components are centered in $\mu_1 = (0, 0)$, $\mu_2 = (0, 0.866)$ and $\mu_3 = (7.50, 4.33)$ (with zero in the remaining dimensions if $d > 2$); the covariance matrix is set to I_d for all the groups, but the variance in the first dimension is increased to 2 for the second group. Table 1 shows the results obtained by CB and MFS at the nominal significance level $\alpha = 0.05$ for various choices of the size $m_{0,k}$ of the starting subsets ($m_{0,k}$ is fixed for $k = 1, \ldots, K$). The entries in column CB are taken from Caroni and Billor (2007). The performance of the algorithms in the absence of outliers are quite similar: in some situations they show some conservatism, which anyway appears to be tolerable.

In the contaminated setting the same configurations described above are considered, but 6 outliers are generated by shifting as many simulated observations: the first 3 points in each group if $K = 2$, and the first 2 points in each group if $K = 3$. For $K = 2$, the slippages are $\{(-6, 0), (0, 6), (0, -6)\}$ in group

Dimension d	Groups K	Starting subset size m_{0k}	Percentage of sets CB	MFS
2	2	10	3.4	3.8
2	2	30	3.8	3.2
2	3	10	5.9	5.0
2	3	30	5.2	4.0
5	2	15	3.6	4.8
5	2	30	4.3	3.2
5	3	15	3.0	5.0
5	3	30	3.8	4.2

Table 1. Percentage of 1000 sets in which any observations were declared as outliers by Caroni and Billor's algorithm (CB) and the proposed one (MFS); $\alpha = 0.05$.

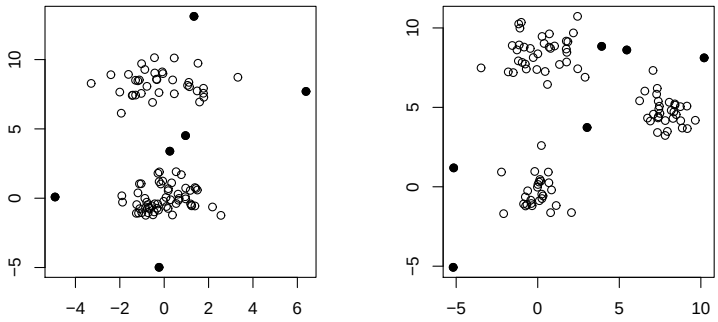


Fig. 2. Simulated instances of two contaminated situations considered in the study: $d = 2$, $K = 2$ (left); $d = 2$, $K = 3$ (right). Black dots represent planted outliers.

1 and $\{(6, 0), (0, 6), (0, -6)\}$ in group 2; for $K = 3$, the pairs of slippages are $\{(-x, 0), (-y, -y)\}$, $\{(x, 0), (y, -y)\}$ and $\{(y, y), (-y, y)\}$ respectively, where $x = 6$ and $y = 4.24$ (and zero in the remaining dimensions if $d > 2$). A graphical display of two simulated instances is given in Figure 2. The performance of the compared algorithms has been evaluated by the percentage of planted outliers detected, averaged over the 1000 simulated sets. The results are shown in Table 2: the entries of column CB are taken from Caroni and Billor's paper. They were obtained using starting subsets of $m_{0k} = 30$ observations for both CB and MFS, for a nominal significance level $\alpha = 0.05$

The results show that MFS outperforms CB in all the examined situations. Moreover, CB seems to be relatively more sensitive to increasing dimensionality: for both the values of K, the decrement in its performance is about 60% when passing from $d = 2$ to $d = 5$, whereas the decrement of MFS performance is about 30%. This is probably due to the fact that the latter situation differs from the former only in the last three variables, which are totally uninformative with respect to both clustering and outlier identification. The presence of such noisy variables blurs the data structure. CB algorithm is severely negatively affected, because its clustering phase relies entirely on the use of Mahalanobis distance. MFS probabilistic clustering is affected to

Dimension d	Groups K	Percentage of outliers detected CB	MFS
2	2	35.43	55.67
2	3	27.70	44.30
5	2	14.35	42.07
5	3	10.03	29.33

Table 2. Mean percentage of planted outliers detected by Caroni and Billor's algorithm (CB) and the proposed one (MFS) in 1000 replicates; $\alpha = 0.05$.

a lesser extent, mostly because the number of free parameters in model (1) grows quadratically with the dimensionality and causes over-fitting. In this respect, the methods proposed in the literature for reducing the number of free parameters in normal mixture models could be employed for improving MFS performance. The price to be paid for MFS better results is in terms of computational cost. MFS requires that an heteroscedastic normal mixture is fitted $n - m_0$ times, usually starting EM algorithm from different initial values in order to achieve stable results.

5 Concluding remarks and open issues

A mixture-based procedure for extending Hadi's outlier detection method to the multiple cluster setting has been presented, as an alternative to the strategy recently proposed by Caroni and Billor (2007). It assumes that the number of clusters is known in advance and that the "core" has been somehow identified for each cluster; we are currently working on an algorithm for automatic identification of these "core" subsets.

By using some results on order statistics and on estimation in truncated samples, Riani et al. (2008) have shown that Hadi's method has a low detection rate for moderate outliers, since Bonferroni bounds are too large. This unpleasant feature is inherited by our algorithm. The possibility of devising a more powerful procedure by exploiting the same results in the context of normal mixture models represents a direction for further research.

References

ATKINSON, A.C. and RIANI, M. (2007): Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis 52, 272-285.*

ATKINSON, A.C., RIANI, M. and CERIOLI A. (2004): *Exploring Multivariate Data with the Forward Search.* Springer, New York.

BILLOR, N., HADI, A.S., VELLEMAN, P.F. (2000): BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis, 34, 279-298.*

CARONI, C. and BILLOR, N. (2007): Robust detection of multiple outliers in Grouped multivariate data. *Journal of Applied Statistics 34:10, 1241-1250.*

FRALEY, C. and RAFTERY A.E. (2006): MCLUST Version 3 for R : Normal Mixture Modeling and Model-based Clustering. *Technical Report No. 504, Department of Statistics, University of Washington.*

HADI, A.S. (1994): A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society (B) 56, 393-396.*

MCLACHLAN, G.J. and BASFORD K.E. (1988): *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

MCLACHLAN, G.J. and PEEL, D. (2000): *Finite Mixture Models.* Wiley, New York.

RIANI, M. ATKINSON, A.C. and CERIOLI, A. (2008): Finding an unknown number of multivariate outliers in large data sets. (Submitted)

Patterns of Functional Dependency Discovery in Schizophrenia by Using Clustering Based on Rules

K. Gibert[1], L. Salvador-Carulla[2], J. C. Martín[1], S. Ochoa[2], V. Vilalta[2], and M. Nadal[2]

[1] Dpt. Statistics and Operations Research, Universitat Politècnica de Catalunya C. Jordi Girona 1-3, Barcelona 08034, Spain, *karina.gibert@upc.edu*
[2] PSICOST Research Association, Spain

Abstract. Functional impairment (FI) in schizophrenia and other severe mental disorders show a different pattern than FI in physical disability or in ageing population. It is important to identify, describe and to operationalize FI in schizophrenia in order to develop elegibility criteria and services for functional dependency in this particular population. This is specially relevant for decision-making related to the implantation of the Spanish Dependency Low, acting from 2007. This work aims to develop an operational classification for schizophrenia based on functional dependency. This *Knowledge Discovery* has been faced by using *clustering based on rules*, an hybrid AI and Statistics technique, which combines some Inductive Learning (from AI) with clustering (from Statistics) to extract knowledge from certain complex domains in form of typical profiles. In this paper, the results of applying this technique to a sample of patients with mental disorders are presented and their advantages with regards to other more classical analysis approaches are discussed. Advantages of proper pre and post treatment of data are also stressed.

Keywords: Data mining and Knowledge Discovery, clustering based on rules, decision support and Knowledge management, class panel graph, prior expert knowledge, schizophrenia, clinical test, dependency

1 Introduction

In 1998, the European Council recommended the member states to develop services for people with dependency. *Dependency* was defined as a condition where, due to the lack or loss of physical, psychological or intellectual functions, the person needs assistance and/or significant aids to perform daily living activities related to self-help and autonomy. Thus *functional dependency* defines a population characterized by high special needs, including the aged, and disabled (either physical or psychological). However, when the global concept of dependency was applied to specific eligibility criteria and related services in different European countries, it became clear that severe mental disorders did not fit into the model developed for physical

and age-related dependency. Among mental disorders, schizophrenia is a major cause of functional impairment (Prince et al. (2007)). In Ustun et al. (1999), relationship between disability and physical and mental conditions was studied, and positive symptoms of schizophrenia (active psychosis) were ranked the third most disabling condition, higher than paraplegia and blindness, by the general population. In the Global Burden of the Disease study (WHO (2001)), schizophrenia accounted for 1.1% of total disability-adjusted life years (DALYs) and 2.8% of years lived with disability (YLDs). The use of services and the economic cost of schizophrenia to society are also high (Haro et al. (2006)). However the functional impairment related to schizophrenia an other mental disorders widely differs from functional impairment in physical disabilities or ageing. Daily living activities such as grooming or moving around are impaired in the later groups. However, schizophrenia concerns social isolation, difficulty in medication compliance and behavioral problems which need monitoring from carers among other distinct impairments. Most of them produce dependency, even if patients are physically able to perform daily activities by themselves.

Spain was the first Mediterranean country to adopt a policy on dependency which also included severe mental disorders. The *Law for the Promotion of personal autonomy and care for persons with dependency (LPAD, 39/2006, 14th December)* was approved by the Spanish government in 2006, to be enacted from 2007 on by regional dependency agencies. The dependency agency of Catalonia (PRODEP) funded a specific project to adapt the dependency concept and elegibility criteria for accessing dependency services and benefits derived by the Law by persons with severe mental disorders (schizophrenia) (Salvador-Carulla et al. (2006)). This work helped to update the know-what and know-how about dependencies in schizophrenia and was a relevant support for the improvement of the official instrument to assess dependency in persons with mental disabilities. In this study, patterns of dependency in schizophrenia were identified by KDD techniques from a collected database regarding patient' clinical, socio-demographic characteristics, as well as psychometric batteries of tests about functional impairment, about the use of private or public Health services and the amount of support required from their carers. Main idea is to induce from data homogeneous groups in schizophrenic population as well as their distinctive characteristics, contributing to an operational definition of functioning in schizophrenia. This is useful for better understanding dependency patterns in our immediate environment and also to support proper decisions about the planning allocation of resources derived from the application of LPAD to the psychic disabled persons.

Nowadays it is well known that *Knowledge Discovery (KDD)* provides a good framework to analyze complex phenomena, as the one presented here, for getting novel and valid knowledge that can improve the background *doctrine corpus* (Fayyad et al. (1996)). We are, in fact, facing a clustering prob-

lem. It has been seen in Gibert and Sonicki (1999) that classical clustering techniques cannot well recognize certain domain structures, so producing some non-sense classes, which cannot be interpreted by the experts. In fact, this arises when dealing with ill-structured domains (ISD) (Gibert and Cortés (1998), Gibert and Cortés (1994)), where numerical and qualitative information coexists (see Gibert et al. (1997), Gibert et al. (2005)), and there exists some relevant semantic additional (but partial) knowledge to be regarded. *Clustering based on rules (ClBR)* (Gibert and Cortés (1994)) is a technique described bellow especially introduced by Gibert to improve clustering results on ISD. In fact, a main advantage is to guarantee the semantic meaning of the resulting classes. In previous works (Gibert and Sonicki (1999), Gibert et al. (2003), Annichiarico et al. (2004)) the improvements in results related with ClBR instead of using other classical clustering techniques have been discussed.

However, *KDD* is, as proposed by Fayyad (Fayyad et al. (1996)), the high level process combining *DM* methods with different tools for extracting *knowledge* from data. In fact, Fayyad's proposal, pointed to a new paradigm in KDD research: *"Most previous work on KDD has focussed on [...] DM step. However, the other steps are of considerable importance for the successful application of KDD in practice."* From this point of view, KDD includes prior and posterior analysis tasks as well as the application of DM algorithms. This work fits in this integral approach and some tools have been used to support prior analysis (data cleaning, transformation, acquiring prior expert knowledge) as well as to assist interpretation of results and reporting (Class Panel Graph, Gibert et al. (2005)).

2 Methods

The analysis is performed on the database PSICOST-II, a naturalistic study of assisted prevalence pre-post with a follow up of two years and three waves of data collection (beginning of study, first year, second year) of a representative sample of six-month's prevalence. There were 306 patients between 18 and 65 years with diagnoses of schizophrenia DSM-IV (APA (2000)). Patients were in contact with mental health services in 4 small healthcare areas in Spain which represented different socio-economical contexts regarding familiar rent, construction levels and mental health services provided. For 205 patients it was also possible to interview the main care giver. Four persons were specially trained to interview the patients for evaluating a battery of assessment scales over the patient. Independent interviews with psychiatrist, and main care giver as well as a revision of the clinical history of the patient were performed, provided the informed consent of the patient. The assessment scales used for the evaluation concerned disease (PANSS, Kay et al. (1986), Prudo and Blum (1987)), quality of life (EuroQol, Brooks (1996)), functioning (through the scales GAF, Endicot et al. (1976), DAS, Janca et al. (1996)), familiar

help requirements (ECFOS, Vilaplana et al. (2007)) about daily activities performances, behaviour, economic management, etc), health services use (CECE, Vazquez-Polo et al. (2005)).

First, descriptive statistics was done. Very simple statistical techniques (Tukey (1977)) were used to describe data and to get preliminary information about. Next, data cleaning, including missing data treatment or outlier detection was performed. It is a very important phase, since the quality of final results directly depends on it. Decisions were taken on the basis of descriptive statistics and background knowledge of the experts. A selection of relevant variables among the whole battery of scales was also done together with the experts. Redundant items through different scales were eliminated.

Data was analyzed using two methods: i) A hierarchical clustering was performed, using chained reciprocal neighbors method, with Ward criterion and the Gibert's mixed metrics (Gibert et al. (1997)), since both numerical and categorial variables were considered. ii) A Clustering based on rules (ClBR), described bellow, was used on the same data set. In this paper, just an intuitive idea is given (see details in Gibert and Cortés (1998) and Gibert and Cortés (1994)). It is a hybrid AI and Statistics technique which combines inductive learning (AI) and clustering (Statistics) especially designed for KDD in ISD. A Knowledge Base (KB) expressing the existent prior domain knowledge is considered to properly bias the clustering on the database. It is implemented in the software KLASS and it has been successfully used in several real applications. Our experience there, see Annichiarico et al. (2004), or Gibert and Sonicki (1999), Comas et al. (2001), also Gibert et al. (2003) or see Gibert et al. (2008), is that using ClBR use to be better than using any statistical clustering method by itself, since an important property of the method is that semantic constraints implied by the KB are hold in final clusters; what guarantees interpretability of the resulting classes. Also, it uses to be better than pure inductive learning methods, since it reduces the effects of missing some implicit knowledge in the KB:

1. Build a (KB) with additional prior knowledge provided by the expert, which can even be a partial description of the domain
2. Evaluate the KB on data. Induce an initial partition over data from it; build a residual class (RC) with the data not included in this partition.
3. Independent hierarchical clustering for every rules-induced class (RIC).
4. Generate prototypes of each rules-induced class.
5. Build the extended residual class as the union of RC with the set of prototypes of RIC, conveniently weighted by the number of objects they represent.
6. Perform a weighted hierarchical clustering of the extended residual class.
7. In the resulting dendrogram, substitute every rules-induced prototype by its hierarchical structure, obtained in 3, integrating a single hierarchy.

For both methods, clustering results can be graphically represented in a dendrogram. Final number of classes was determined on best horizontal cut (maximizing the ratio of between-classes inertia versus within-classes inertia). This identifies a partition of the data. Interpretation of the classes

use to be difficult and time consuming and requires much human guidance. Here, it was supported by Class panel graph (CPG), where conditional distributions of the variables through the classes, displayed through multiple histograms, is shown in a compact way (Gibert et al. (2005)); relevance of differences between classes is assessed using ANOVA, Kruskall-Wallis or χ^2 independence test, depending on the required assumptions hold by the variables. Experts use the CPG to get a meaningful description of classes by identifying which variables indicate particularities of every class regarding the others and making a conceptualization process which leads to a class-labeling proposal regarding the semantic entity represented by each class.

3 Results

Clustering of the 306 patients was made. The dominant profile in the sample is a man (68%), single (77%), with primary school (49%), getting a pension (62%) and mainly leaving with parents (67%). Disease started about 24 years and stay for more than 14 years, (sample of long duration schizophrenia).

i) With classical hierarchical clustering, 5 classes emerged. However, most of the variables shown no significant differences *vs* classes and their interpretation was confusing and psychiatrists could not learn too much from the results. Patients with different levels of dependency were mixed in the different clusters and it was no possible to understand the underlying clustering criteria. ii) Several iterations of the ClBR process conformed the prior knowledge acquisition phase, and experts could elicit their implicit knowledge. The knowledge provided by the experts concerned some clear situations of dependency or autonomy (as an example, they stated that patients with bad levels of functioning (GAF), high familiar support requirements in daily activities (ECFOS section A) and behavioral problems (ECFOS section B) are patients in ill conditions; in another rule they stated that patients able to work and with high levels of functioning (GAF) are in good condition). Finally 5, classes with different patterns of dependency were found. A clearer conceptual interpretation of the classes is possible from the experts' point of view, looking at significant variables (Salvador-Carulla et al. (2006)). Fig. 1 shows the CPG with some variables supporting this interpretation:

Autonomous (c299) : 93 persons with the better conditions: They are autonomous and can do tasks by themselves; they can work; they require little support from their carers (less than 4 h. a week) and they do not make intensive use of healthcare services. This group has the greater educational level: 28 started secondary school and 18 could finish it; other 18 started higher education and 11 could finish it. Younger disease than other groups (13 years in average)

Singles (C300) : 87 persons whose main characteristic is to live alone. They tend to complete primary school. They have been ill for 15 years in average. They have intermediate scores in the assessment scales. But their condition is not good; probably they would require higher supervision, but they have not. Treatment adherence and contacts with doctors are very low. They show a healthcare

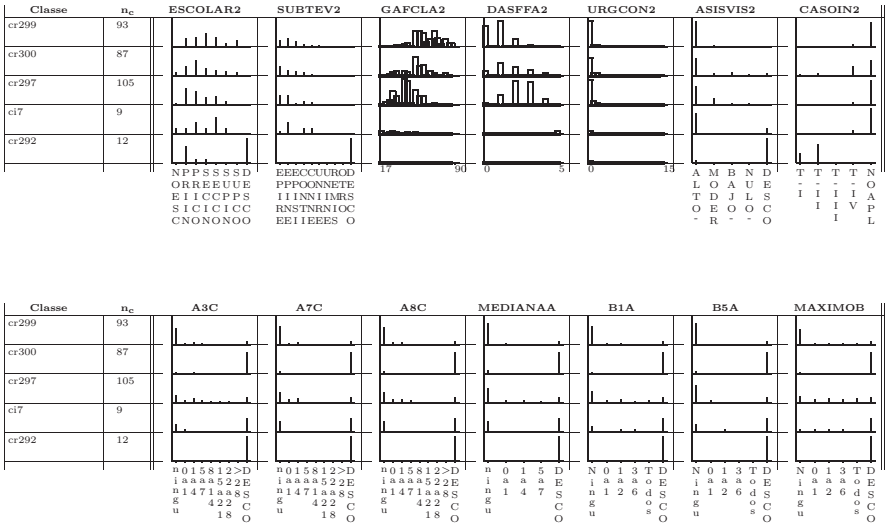


Fig. 1. Panel graph of most relevant variables.

pattern very different from other groups, in fact they use the services in an inappropriate way: they may fail to the scheduled visits with professionals, whereas they may overuse emergency services (up to 15 times), probably because they feel bad and they are alone. When ECFOS could be evaluated, they show low familiar support requirements (less than 7 h. a week), mainly focused on domestic tasks. They do not generate family burden due to behavioral problems.

Institutionalized (ci7) This group includes all the persons (a total of 9) which along the first year of the study ended-up to long-term residential care. They completed secondary school, they tend to be slightly older than the other groups, but not significantly, and they have a longer course of disease (23 years in average), they have non-contributive pension, they have an evolutive subtype of episodes with residual inter-episodic symptoms and severe negative symptoms. They show worse functioning levels that other groups and higher levels of severity as well as more self-harm attempts. They provoke family burden due to behavioral problems and they use to require support in daily activities.

Dependents (c297) It is a class of 105 patients with high levels of dependency. They couldn't finish primary school. This are the patients in worse conditions; thus they make a high use of health care. This is the group requiring higher support from carers, up to 28 hours a week.

Uncomplete (c292) They are 12 patients which dropped out the study by different reasons. Only socio-demographic and clinical data is available.

4 Discussion

Clustering techniques allow detecting different groups of patients of different dependency profiles. The analysis of the data under method i) only provided

a confusing partition of patients difficult to understand. Facing such a complicated phenomenon as dependency, concerned with a lack of clear patterns and difficulties for establishing relationships between patient characteristics and patient needs of support, requires to take into account as much prior expert knowledge as possible, even if it is a partial description of the phenomenon. Actually, mental disorders can be considered an ISD, as stated in Gibert and Cortés (1998) and clustering use to be unable to capture the complex structure of ISD by itself. The additional knowledge provided by experts is expressed by means of logical rules; it use to be a partial description of the domain (as usual for ISD, it is very difficult to make explicit a complete KB for the domain, and this is a great handicap for using pure AI methods). Here, the KB expressed 5 rules with antecedents involving no more than 4 variables of the whole set of 75 measurements available. None of the classical statistical methods support expert knowledge influencing the analysis. ClBR is a hybrid technique which sensibly improved results by integrating clinical knowledge inside the analysis, which produces classes with proper interpretation (Annichiarico et al. (2004)). Finally, a set of 5 classes was recommended by the system. Several tools were used to assist the interpretation of final classes. Among them, CPG appeared as a successful support for the conceptualization process. From the medical point of view, ClBR provided a set of classes which fit well with different patterns of increasing degrees of dependency. All the patients that dropped out the study appear in a single group. Patients with dependency are subdivided in three different profiles: those who ended in long-term residential care, those who leave alone and those in such an ill condition that they cannot leave alone, but stay at home. Particularly interesting to the experts, elicitation of the special situation of the **Singles** group, which do not have extremely high dependency regarding daily leaving activities or functioning, but they show behavioral problems and as they leave alone and they are not properly supervised, whereas they should, they finally loss treatment adherence and make an irrational use of services, *i. e.* missing scheduled visits and using emergency service as their main care resource.

The use of ClBR produces meaningful classes and sensibly improves, from a semantics point of view, the results of classical clustering, according to our opinion that hybrid techniques combining AI and Statistics are more powerful for KDD than pure ones. This work allowed the experts to formulate further hypothesis to be confirmed in future works and contributed to increase the knowledge about the dependency situations in the population with severe mental disorders. A clearer knowledge about how dependency behaves in schizophrenic population was achieved and this should help to a better resource allocation and planning of dependency services derived from the LPAD. Indeed, from this results assigning specific *packages of care/support* according to the dependency profile of the patient becomes possible. An independent group of experts was asked to manually elaborate a second profiles

proposal, to validate the methodology. Later, predictors of the different profiles will be identified to properly assign resources and benefits to LPAD applicants with severe mental health problems.

Acknowledgements: This research was supported and leaded by PRODEP, the specific program of the Generalitat de Catalunya for encouraging and structuring the promotion to the personal autonomy and the attention to persons with dependencies. Thanks also to APPS, for partially financing the research.

References

ANNICHIARICO, R. *et. al.* (2004): Qualitative profiles of disability. *JRRD* 41(6A), 835–845.

APA (2000): *Diagnostic and Statistical Manual of Mental Disorders*. DSM-IV-TR. Washington: American Psychiatric Association.

BROOKS (1996): EuroQol: The current state of play *Health Policy* 37, 53-72.

COMAS, J. *et al.* (2001): Knowledge discovery by means of inductive methods in wastewater treatment plant data. *AI Communications*, 14(1), 45-62.

EENDICOTT, J. , SPITZER, R. L., FLEISS, J. L. and COHEN, J. (1976): The global assessment scale. A procedure for measuring overall severity of psychiatric disturbance *Archives of General Psychiatry* 33, 766-771.

FAYYAD, U. *et alt.* (1996): *Advances in Knowledge Discovery and Data Mining*, chapter From Data Mining to KDD: An overview. AAAI/MIT Press.

GUTIERREZ-RECACHA, P., CHISHOLM, D., HARO, J.M., SALVADOR-CARULLA, L., AYUSO-MATEOS, J.L. (2006): Cost-effectiveness of different clinical interventions for reducing the burden of schizophrenia in Spain. *Acta Psychiatrica Scandinavica*, (111(Suppl. 432)), 29-38.

GIBERT, K., GARCÍA-RUDOLPH, A. *et al.* (2008): Response to Traumatic Brain Injury Neurorehabilitation through an Arteficial Intelligence and Statistics Hybrid Knowledge Discovery from Databases Methodology *Medical Archives* 62(3), 132-135.

GIBERT, K., CORTÉS, U. (1994): Combining a knowledge-based system and a clustering method for a construction of models in ill-structured domains. In: P. Cheeseman and R. W Oldford (Eds.): Selecting Models from Data. Artificial Intelligence and Statistics IV, Lecture Notes in Statistics, Springer-Verlag, 351-360.

GIBERT, K. *et al.* (1997): Weighing quantitative and qualitative variables in clustering methods. *Mathware and Soft Computing* 4(3), 251-266.

GIBERT, K., CORTÉS, U. (1998): Clustering based on rules and knowledge discovery in ill-structured domains. *Computación y Sistemas.* 1(4), 213-227.

GIBERT, K., NONELL, R. *et al.* (2005): Knowledge Discovery with Clustering: Impact of metrics and reporting phase by using KLASS. *Neural Network World*, 4/05, 319–326.

GIBERT, K., RODAS, J. *et al.* (2003): Using kdsm for knowledge discovery in electroconvulsive therapy. *Medicinska Informatica* 6, 15-21.

GIBERT, K., SONICKI, Z. (1999): Classification based on rules and thyroids dysfunctions. *A*MSDA, 15(4), 319–324.

JANCA, A., KASTRUP, M. *et al.* (1996): The World Health Organization Short Disability Assessment Schedule (WHO DAS-S): a tool for the assessment of difficulties in selected areas of functioning of patients with mental disorders. *Soc Psychiatry Psychiatr Epidemiol* 31, 349-54.

KAY, S.R., OPLER L.A. *et al.* (1986): The positive and negative symptom scale (PANSS). Rating manual. *Social Behav Sci Doc* 17, 28-29.

PRINCE, M., PATEL, V. *et al.* (2007): No health without mental health. Lancet, 8,370(9590),859-77.

PRUDO, R., BLUM, H.M. (1987): Five-year outcome and prognosis in schizophrenia. *British Journal of Psychiatry* 150, 345-54.

SALVADOR-CARULLA, L., GIBERT, K. *et alt* (2006): Estudio DEFDEP: Definición operativa de dependencia en personas con discapacidad psíquica, vols. 1 y 2. PRODEP, Barcelona, Spain.

TUKEY, J.W. (1977): *Exploratory Data Analysis*. Addison-Wesley.

USTUN, T.B, REHM, J. *et al.* (1999): Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH Joint Project CAR Study Group. *LANCET*, 354 (9173), 111–115.

VAZQUEZ-POLO, F.J., NEGRIN, M. *e*t al (2005): An analysis of the costs of treating schizophrenia in Spain: a hierarchical Bayesian approach. *J Ment Health Policy Econ* 8(3), 153-65

VILAPLANA, M., OCHOA, S. *et al*(2007): Validacin en poblacin espaola de la entrevista de carga familiar objetiva y subjetiva (ECFOS-II). Validacin en poblacin espaola del ECFOS-II. *Actas Esp Psiquiatr* 35(6), 372-81.

WHO (2001): *The World Health Report 2001 – Mental Health: New Understanding, New Hope*. Geneva: WHO.

Fitting Finite Mixtures of Linear Mixed Models with the EM Algorithm

Bettina Grün

Department für Statistik und Mathematik, Wirtschaftsuniversität Wien
Augasse 2-6, 1090 Wien, Austria, *Bettina.Gruen@wu-wien.ac.at*

Abstract. Finite mixtures of linear mixed models are increasily applied in different areas of application. They conveniently allow to account for correlations between observations from the same individual and to model unobserved heterogeneity between individuals at the same time. Different variants of the EM algorithm are possible for maximum likelihood (ML) estimation. In this paper two different versions for fitting this model class are presented. One variant of the EM algorithm requires weighted ML estimation. As this fitting method might not be readily available in standard software sufficient conditions which allow to transform a weighted into an unweighted ML estimation problem are derived.

Keywords: EM algorithm, finite mixture, linear mixed model, unobserved heterogeneity

1 Introduction

Finite mixture models are a popular method for modelling unobserved heterogeneity. In the last decades the original model of finite mixtures of distributions has been extended in several ways and nearly arbitrary component specific models are nowadays used in applications. This development has been faciliated by estimation techniques which constitute a common framework for fitting arbitrary mixture models and which require only to modify the component specific model estimation for different mixture models. This holds for the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin (1977)) for maximum likelihood (ML) estimation.

Finite mixtures of mixed effects models allow to account for different kinds of heterogeneity between individuals (Frühwirth-Schnatter 2006). The components of the mixture represent different groups with distinct parameterizations while the random effects allow for individual differences which cluster around a common mean value. These models are applied in several different areas such as marketing (Lenk and DeSarbo (2000)), medicine (Xu and Hedeker (2001)) and bioinformatics (Luan and Li (2004)).

This paper is organized as follows: Section 2 introduces the model. Section 3 outlines two variants of the EM algorithm for ML estimation of this model class and derives sufficient conditions for allowing the use of implementations of fitting algorithms for unweighted mixed-effects models. A short sketch of a possible implementation in R is provided.

2 Model specification

In the following finite mixtures of mixed effects models are considered where the mixed effects are needed to account for correlations between observations from the same individual and the finite mixture models the unobserved heterogeneity between the individuals. This implies that the component memberships of the individuals are fixed.

Assume observations from N individuals are given and for each individual i the data (Y_i, X_i, Z_i, w_i) is given which consists of n_i observations on the dependent variables $Y_i = (y_{ij})_{j=1,\ldots,n_i}$, the covariates for the fixed effects $X_i = (x_{ij})_{j=1,\ldots,n_i}$ and the covariates for the random effects $Z_i = (z_{ij})_{j=1,\ldots,n_i}$. w_i denote the individual specific concomitant variables.

The finite mixture density of mixed effects models with K components is given for the observations of individual i by

$$\begin{aligned} h(Y_i|X_i, Z_i, w_i, \Theta) &= \sum_{k=1}^{K} \pi_k(w_i) \int \prod_{j=1}^{n_i} \phi_1(y_{ij}; x_{ij}\beta_k + z_{ij}b_i^k, \sigma_k^2)\phi_q(b_i^k; 0, \Psi_k)db_i^k \\ &= \sum_{k=1}^{K} \pi_k(w_i)\phi_{n_i}(Y_i; X_i\beta_k, Z_i\Psi_k Z_i^T + \sigma_k^2 I_{n_i}). \end{aligned}$$

$\phi_d(\cdot; \mu, \Sigma)$ denotes the d-dimensional normal distribution with mean μ and variance-covariance matrix Σ. The fixed effects are given by β_k and the random effects by b_i^k. The random effects are assumed to have mean zero which implies that any constant influence is already captured by the fixed effects. This can be ensured by constraining that the covariates Z_i span a subset of the space spanned by X_i over all individuals $i = 1, \ldots, N$.

The variance-covariance matrix of Y_i for component k is given by

$$\Sigma_i^k = \sigma_k^2 \Sigma_{i0}^k = Z_i \Psi_k Z_i^T + \sigma_k^2 I_{n_i}.$$

It is assumed that $\Psi_k = \sigma_k^2 \theta_k$ and $\Sigma_i^k = \sigma_k^2 (Z_i \theta_k Z_i^T + I_{n_i})$.

The component weights $\pi_k(w_i)$ are assumed to fulfill the following conditions for all i:

$$\pi_k(w_i) > 0 \quad \forall k \qquad \text{and} \qquad \sum_{k=1}^{K} \pi_k(w_i) = 1.$$

The most common concomitant variable model for w_i is the multinomial logit model (Dayton and Macready (1988)).

This model specification implies that there exist no common parameters which are constant over the components and hence, each of the components can be separately estimated given the component memberships of the individuals. As the component memberships are fixed for all observations $j = 1, \ldots, n_i$ of individual i, it is also assumed that the concomitant variables

w_i are constant for each individual. A different model specification where the concomitant variables for individual i are given by $W_i = (w_{ij})_{j=1,\ldots,n_i}$ and the component membership π_k is not fixed for each individual is for example given in Yau et al. (2003) and Hall and Wang (2005).

3 Estimation with the EM algorithm

The EM algorithm is in general applied in a missing data context. It is an iterative procedure which alternates between an E(xpectation)-step and a M(aximization)-step. The EM algorithm works on the complete likelihood derived by also including the missing data and exploits the fact that the complete likelihood is in general easier to maximize than the original likelihood. The missing data is integrated out in the E-step by determining the expectation of the complete likelihood given the available data and the current parameter estimates. The expected complete likelihood is then maximized in the M-step.

The EM algorithm has been shown to increase the likelihood in each step and hence to converge for bounded likelihoods. The implementation of the EM algorithm can often be simplified by introducing more variables as missing data. However, the disadvantage is that the convergence of the EM algorithm depends on the amount of missing data and hence, more iterations are needed if the amount of missing data is increased.

For finite mixtures of linear mixed effects models different variants for ML estimation with the EM algorithm have been proposed. In the following two different versions are discussed in detail which differ with respect to the variables they use as missing data.

3.1 Random effects and component memberships as missing data

The most popular variant of the EM algorithm for fitting finite mixtures of linear mixed effects models is where the component memberships as well as the random effects are treated as missing data and imputed in the E-step (see for example Xu and Hedeker (2001) or Celeux et al. (2005)).

For this variant the E-step consists of determining

1. the a posteriori probabilities that an individual i is from component k:

$$\tau_{ik} = \frac{\pi_k(w_i)\phi_{n_i}(X_i\beta_k, Z_i\Psi_k Z_i^T + \sigma_k^2 I_{n_i})}{\sum_{l=1}^{K} \pi_l(w_i)\phi_{n_i}(X_i\beta_l, Z_i\Psi_l Z_i^T + \sigma_l^2 I_{n_i})}$$

   and
2. the mean and the variance of the random effects b_i conditional on the current parameter estimates Θ, the observations Y_i, the covariates X_i and Z_i and the component k. These are calculated using that b_i and Y_i

follow a joint multivariate normal distribution conditional on Θ, X_i, Z_i and k:

$$\begin{aligned}\mu_{b_i,k} &= \mathbb{E}[b_i|Y_i, X_i, Z_i, \Theta, k] \\ &= [\frac{1}{\sigma_k^2} Z_i^T Z_i + \Psi_k^{-1}]^{-1} \frac{1}{\sigma_k^2} Z_i^T (Y_i - X_i\beta_k) \\ \Sigma_{b_i,k} &= \mathbb{V}[b_i|Y_i, X_i, Z_i, \Theta, k] = \left[\frac{1}{\sigma_k^2} Z_i^T Z_i + \Psi_k^{-1}\right]^{-1}.\end{aligned}$$

The expected complete likelihood is given by

$$\sum_{k=1}^{K}\sum_{i=1}^{N} \tau_{ik} \Bigg[\log \pi_k(w_i) - \frac{1}{2}\Bigg((n_i + q)\log(2\pi) + n_i \log \sigma_k^2 + \log|\Psi_k| + \sum_{j=1}^{n_i} \frac{(y_{ij} - z_{ij}\mu_{b_i,k} - x_{ij}\beta_k)^2 + z_{ij}\Sigma_{b_i,k} z_{ij}^T}{\sigma_k^2} + + \mathrm{tr}(\Psi_k^{-1}(\Sigma_{b_i,k} + \mu_{b_i,k}\mu_{b_i,k}^T))\Bigg)\Bigg].$$

$\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

For the M-step the parameters of the concomitant variable model and the component specific model can be separately determined. For the concomitant variable model a weighted multinomial logit model has to be estimated if the component weights are determined through a multinomial logit model. This estimation method is often already available in standard statistical software. For the component specific model the parameters can be determined in closed form by solving the equations derived by determining the derivatives of the expected complete likelihood and setting them to zero:

$$\begin{aligned}\hat{\beta}_k &= \frac{1}{\sum_{i=1}^N \tau_{ik}} \Big(\sum_{i=1}^{N} \tau_{ik} \sum_{j=1}^{n_i} x_{ij}^T x_{ij}\Big)^{-1} \left[\sum_{i=1}^{N} \tau_{ik} \sum_{j=1}^{n_i} x_{ij}^T (y_{ij} - z_{ij}\mu_{b_i,k})\right] \\ \hat{\sigma}_k^2 &= \frac{1}{\sum_{i=1}^N \tau_{ik} n_i} \sum_{i=1}^{N} \tau_{ik} \sum_{j=1}^{n_i} (y_{ij} - z_{ij}\mu_{b_i,k} - x_{ij}\hat{\beta}_k)^2 + z_{ij}\Sigma_{b_i,k} z_{ij}^T \\ \hat{\Psi}_k &= \frac{1}{\sum_{i=1}^N \tau_{ik}} \sum_{i=1}^{N} \tau_{ik}(\Sigma_{b_i,k} + \mu_{b_i,k}\mu_{b_i,k}^T).\end{aligned}$$

3.2 Component memberships as missing data

An alternative implementation would be the straightforward application of the EM algorithm as in general used for finite mixtures, i.e., only the component membership is treated as missing data. This implementation requires

the weighted ML estimation of the linear mixed model for the M-step and the determination of the posterior probabilities in the E-step. Standard software for fitting linear mixed effects models often does not allow for weighted ML estimation or does only account for different variance-covariance matrices for the error term. Under certain conditions an unweighted ML estimation can be used for weighted ML estimation where the observations are suitably transformed. The following corollary gives sufficient conditions.

Corollary 1 (Weighted ML estimation). *The weigthed ML estimate of θ of a linear mixed model with observations (Y_i, X_i, Z_i) and weights τ_i for $i = 1, \ldots, N$ is equivalent to the ML estimate of θ of a linear mixed model with transformed variables $\tilde{X}_i = \sqrt{\tau_i} X_i$ and $\tilde{Y}_i = \sqrt{\tau_i} Y_i$ and the same Z_i if*

$$Z_i \equiv Z \quad \forall i = 1, \ldots, N.$$

Proof. The weighted deviance which is equivalent to -2 log-likelihood is given by

$$\operatorname{dev}(\beta, \theta, \sigma^2) = \sum_{i=1}^{N} \tau_i n_i \log(2\pi\sigma^2) + \tau_i \log|\Sigma_{i0}| + \\ + \frac{\tau_i}{\sigma^2}(Y_i - X_i\beta)^T \Sigma_{i0}^{-1}(Y_i - X_i\beta).$$

The ML estimates of the coefficients $\hat{\beta}$ and the variance $\hat{\sigma}^2$ depend on the weighted residual sum of squares $r_{\tau i}^2$ and for determining the profile deviance they are all functions of θ:

$$r_{\tau i}^2(\theta) = \tau_i (Y_i - X_i \hat{\beta}(\theta))^T \Sigma_{i0}^{-1} (Y_i - X_i \hat{\beta}(\theta)) \tag{1}$$

$$\hat{\sigma}^2(\theta) = \frac{\sum_{i=1}^{N} r_{\tau i}^2(\theta)}{\sum_{i=1}^{N} \tau_i n_i} \tag{2}$$

Given Equation (1) $\hat{\beta}(\theta)$ is given by the generalized least squares estimate for the variance-covariance matrix Σ_{i0}.

The profile deviance is then given by

$$\operatorname{dev}(\theta) = \sum_{i=1}^{N} \tau_i n_i \log\left(2\pi \frac{\sum_{i=1}^{N} r_{\tau i}^2}{\sum_{i=1}^{N} \tau_i n_i}\right) + \tau_i \log|\Sigma_{i0}| + \tau_i n_i. \tag{3}$$

If $Z_i \equiv Z$ for all i and hence also $n_i \equiv n$, this gives

$$\operatorname{dev}(\theta) = \tilde{\tau}\left[n\left(1 + \log(\frac{2\pi}{\tilde{\tau} n}) + \log(\sum_{i=1}^{N} r_{\tau i}^2)\right) + \log|Z\theta Z^T + I_n|\right]$$

where $\tilde{\tau} = \sum_{i=1}^{N} \tau_i$.

The profile deviance for $\tilde{X}_i = \sqrt{w_i}X_i$ and $\tilde{Y}_i = \sqrt{w_i}Y_i$ is given by

$$\text{dev}(\theta) = N\left[n\left(1 + \log(\frac{2\pi}{Nn}) + \log(\sum_{i=1}^{N} r_{\tau i}^2)\right) + \log|Z\theta Z^T + I_n|\right].$$

As the profile deviances are equivalent up to an additive constant and a constant factor they are maximized for the same θ.

The estimates for β and σ^2 are then determined using Equations (1) and (2). The estimate of β is the same for the weighted and the unweighted but transformed fitting problem, because the residual sum of squares term is identical up to a constant factor. Only the estimate of σ^2 has to be modified if the estimates of the unweighted but transformed fitting problem are used. As can be seen in Equation (2) the denominator of the weighted estimation problem is $\sum_{i=1}^{N} \tau_i n_i$ while it is $\sum_{i=1}^{N} n_i$ for the unweighted but transformed problem.

From the weighted profile deviance (Equation 3) it can be seen which changes are necessary to allow for weighted ML estimation. It is not sufficient to only change the residual sum of squares but the weights also influence the sum over the logarithm of the determinant of the individual variance-covariance matrices. Accounting for the weights in the estimation might then not be easily possible if for example the following simplification is used by the software for the determining the determinant

$$|\tilde{Z}\tilde{Z}^T + I_{\sum_{i=1}^{N} n_i}| = |\tilde{Z}^T\tilde{Z} + I_q|.$$

The sufficient conditions indicate that standard software can easily be used in the case where a balanced design is given, i.e., the same observations are available for each individual. Without missing data this occurs for example in bioinformatics where gene expression data is observed over time at a priori specified time points. The conditions might also be more likely applicable in the case where only a random intercept is fitted.

If the entire data set does not fulfill the sufficient conditions, only the subsample fulfilling the conditions might be used in a first step to pre-analyse the data. The entire data set can then be fitted using the EM algorithm where also the random effects are used as missing information but which is initialized in the previously found solution.

If the transformation of the weighted into an unweighted ML estimation problem is not possible, the Classification EM algorithm (CEM; Celeux and Govaert (1992)) can be used instead of the classical EM algorithm. The CEM algorithm allows to use unweigthed ML estimation methods. However, it does not maximize the likelihood but the classification likelihood. The advantage of the CEM algorithm is that it converges in general faster than the EM algorithm, i.e., it needs less iterations. It has been therefore proposed to use the CEM algorithm with different random initializations to find a good

starting point for the ordinary EM algorithm which in the case of finite mixtures of mixed effects models might be the variant where the random effects are also included in the missing data.

3.3 Implementation in R

Both variants of the EM algorithm can easily be implemented in R, an environment for statistical computing and graphics (R Development Core Team (2007)). Package **flexmix** (Leisch (2004)) for example implements the EM algorithm for ML estimation of finite mixture models. It provids the E-step and all data handling and arbitrary mixture models can be fitted by modifying the M-step. The implementation of the package aims at easy extensibility and tries to enable rapid prototyping. In general only a model driver for the component specific model needs to be written which specifies the fitting function. In addition the package also allows fitting of finite mixture models with the CEM algorithm.

The recommended package in R for fitting linear mixed effects models is **nlme** (Pinheiro and Bates (2000)). Function `lme()` allows to specify a weights argument, which can be used to describe the within-group heteroscedasticity structure. An alternative implementation is provided by the package **lme4** (Bates (2007)). The weights argument of function `lmer()` specifies that a weighted residual sum of squares is minimized. Hence, the recommended functions in R do not allow for weighted ML estimation of linear mixed models. The sufficient conditions can be used to determine when it is possible to estimate the transformed problem using this functionality in combination with package **flexmix**.

4 Conclusion and future work

The most common way of fitting finite mixtures of mixed effects models with the EM algorithm is by introducing the component memberships and the random effects as missing data. However, this signifies that this EM algorithm is different from the general application of the EM algorithm for finite mixture models where only the component memberships are used as missing data and the M-step consists of weighted ML estimation of the component specific models.

As the reason for the preference of this variant might be that weighted ML estimation of linear mixed models is not readily available in standard statistical software, this paper investigates which conditions need to be fulfilled that the weighted ML problem is equivalent to an unweighted but transformed ML problem. The results indicate that this is possible in applications where a balanced design is used to collect the data. In addition it is likely to be at least applicable for a subset of the data in random intercept models.

In the future the performance of the two EM algorithms should be compared. The variant where the component memberhips as well as the random effects are used as missing data can be expected to need more iterations while each iteration will take less time as the M-step is given in closed form. The advantage of the other variant is that if the fitting function of the linear mixed model is improved this can be exploited in the M-step. In addition it might be useful to investigate how the fitting function of the linear mixed models has to be modified to allow for weighted ML estimation.

Acknowledgments

This piece of research was supported by the Austrian Science Foundation (FWF) under grant T351.

References

BATES, D. (2007): *lme4: Linear mixed-effects models using S4 classes*. R package version 0.99875-8.

CELEUX, G. and GOVAERT, G. (1992): A Classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis, 14(3), 315–332*.

CELEUX, G., MARTIN, O. and LAVERGNE, C. (2005): Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling, 5, 243–267*.

DAYTON, C.M. and MACREADY, G.B. (1988): Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83(401), 173–178*.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D. B. (1977): Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B, 39, 1–38*.

FRÜHWIRTH-SCHNATTER, S. (2006): *Finite Mixture and Markov Switching Models*. Springer.

HALL, D.B. and WANG, L. (2005): Two-component mixtures of generalized linear mixed effects models for cluster correlated data. *Statistical Modelling, 5, 21–37*.

LEISCH, F. (2004): FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software, 11(8)*, 1–18.

LENK, P.J. and DESARBO, W.S. (2000): Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika, 65(1), 93–119*.

LUAN, Y. and LI, H. (2004): Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics, 20(3), 332–339*.

PINHEIRO, J.C. and BATES, D.M. (2000): *Mixed-Effects Models in S and S-Plus*. Springer.

R DEVELOPMENT CORE TEAM (2007): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

XU, W. and HEDEKER, D. (2001): A random-effects mixture model for classifying treatment response in longitudinal clinical trials. *Journal of Biopharmaceutical Statistics, 11(4), 253–273*.

YAU, K.K., LEE, A.H. and NG, A.S. (2003): Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics & Data Analysis, 41, 359–366.*

Clustering Rows and/or Columns of a Two-Way Contingency Table and a Related Distribution Theory

Chihiro Hirotsu

Faculty of Science and Technology, Meisei University
2-1-1 Hodokubo, Hino-City, 191-8506 Tokyo, Japan, *hirotsu@ge.meisei-u.ac.jp*

Abstract. The row-wise multiple comparison procedure proposed in Hirotsu (1977, 1983) has been verified to be useful for clustering rows and/or columns of a contingency table in several applications. Although the method improved the preceding work there is still a gap between the squared distance between the two clusters of rows and the largest root of a Wishart matrix as a reference statistic for evaluating the significance of the clustering. In this paper we extend the squared distance to a generalized squared distance among any number of rows or clusters of rows and dissolves the loss of power in the process of the clustering procedure. If there is a natural ordering in columns we define an order sensitive squared distance and then the reference distribution becomes that of the largest root of a non-orthogonal Wishart matrix, which is very difficult to handle. We therefore propose a very nice χ^2-approximation which improves the usual normal approximation in Anderson (2003) and also the first χ^2-approximation introduced in Hirotsu (1991). A two-way table reported by Guttman (1971) and analyzed by Greenacre (1988) is reanalyzed and a very nice interpretation of the data has been obtained.

Keywords: χ^2-approximation, correspondence analysis, generalized squared distance, non-orthogonal Wishart matrix, row-wise multiple comparisons

1 Introduction

An overall goodness of fit chi-square test for independence is a well known approach to a contingency table. It cannot, however, give any detailed information on the association between the rows and columns. Therefore some multiple comparison approaches have been proposed, among which the row- and/or column-wise multiple comparisons proposed in Hirotsu (1977, 1983) have been verified to be useful in several occasions as compared with other multiple comparison approaches, see Greenacre (1988), Hirotsu (1991, 1993) and Hirotsu et al. (2003). The row-wise multiple comparisons are essential for the one-way layout with categorical responses instead of usual normal variables. The multiple comparison procedure proposed first in Hirotsu (1977, 1983) is essentially the Scheffé type but the actual procedure is based on the squared distances between every two rows or two clusters of rows for which

those distances are uniquely defined by the normalization and orthogonality conditions. Then there is an inevitable loss of power in the procedure. Therefore in Section 3 of this paper the method is extended to defining the generalized squared distance among any number of rows or the clusters of rows and the loss of power is dissolved in the process of Scheffé type multiple comparison procedure.

An interesting extension of the method is to the one-way layout with the ordered categorical responses. In this case the procedure is essentially unchanged excepting the definition of the squared distance and the related asymptotic distribution. Then the reference distribution becomes that of the largest root of a non-orthogonal Wishart matrix which is very difficult to handle. The usual normal approximation given in Anderson (2003) is quite unsatisfactory especially when the first and the second largest roots are close to each other. In Section 4 of this paper we therefore propose a χ^2-approximation as a more reasonable one. It nicely improves the normal approximation and also the χ^2-approximation introduced in Hirotsu (1991). In Section 5 a two-way table reported by Guttman (1971) and analyzed in Greenacre (1988) is reanalyzed by the method of the generalized squared distance and a very nice interpretation of the data has been obtained. Finally in Section 6 a conclusion is mentioned.

2 Row-wise multiple comparisons in a two-way contingency table

Let a two-way contingency table be denoted by $\{y_{ij}\}_{a\times b}$ and the row, column and the grand totals by $R_i = y_{i\cdot}$ for all i (i th row total), $C_j = y_{\cdot j}$ for all j (j th column total) and $N = y_{\cdot\cdot}$ (grand total), respectively, where we use the usual dot notation to express the summation with respect to the subscript replaced by dot. We assume a multinomial distribution with the cell probabilities $\{p_{ij}|p_{\cdot\cdot} = 1\}$. The null hypothesis of interest is then

$$H : p_{ij} = p_{i\cdot}p_{\cdot j} \quad \text{for all } i,\ j$$

and the statistical inference is based on the conditional distribution given $\{R_i\}$ and $\{C_j\}$. For the row-wise multiple comparisons define

$$\boldsymbol{r} = N^{-1/2}\Big(\sqrt{R_1}, \cdots, \sqrt{R_a}\Big)', \quad \boldsymbol{c} = N^{-1/2}\Big(\sqrt{C_1}, \cdots, \sqrt{C_b}\Big)'$$

and then define $R'_{a-1\times a}$ and $C'_{b-1\times b}$ so that $\begin{pmatrix} \boldsymbol{r}' \\ R' \end{pmatrix}$ and $\begin{pmatrix} \boldsymbol{c}' \\ C' \end{pmatrix}$ are the a- and b-dimensional orthogonal matrices, where the prime denotes a transpose of a matrix. Define a column vector $\boldsymbol{z}$ with the elements $z_{ij} = y_{ij}/\sqrt{R_iC_j/N}$ arranged in the dictionary order. Then under the null hypothesis H the conditional expectation and variance of $(R' \otimes C')\boldsymbol{z}$ given R_i and C_j are

$$E\{(R' \otimes C')\boldsymbol{z}\} = \boldsymbol{O}_{(a-1)(b-1)},$$
$$V\{(R' \otimes C')\boldsymbol{z}\} = (N/(N-1))\boldsymbol{I}_{(a-1)(b-1)}$$

and become very easy to handle, where $\boldsymbol{O}_n$ and $\boldsymbol{I}_n$ are n-dimensional zero vector and the identity matrix, respectively and $\otimes$ denotes the Kronecker's product. In the following we ignore the coefficient $(N/(N-1))$ in the variance since our example of the contingency table is usually large. Then

$$\chi^2 = \|(R' \otimes C')\, \boldsymbol{z}\|^2 \tag{1}$$

is nothing but the goodness of fit χ^2 for H and every row of $(R' \otimes C')\, \boldsymbol{z}$ gives the partition of χ^2 into one degree of freedom, where $\|\cdot\|^2$ denotes the squared norm of a vector. However, the multiple comparison approach based on one degree of freedom statistic like this cannot have a reasonable power if the two-way table is moderately large. Then the row-wise multiple comparison procedure proposed in Hirotsu (1977, 1983) is based on

$$\chi^2(i;i') = \|(\boldsymbol{r}'(i;i') \otimes C')\, \boldsymbol{z}\|^2 \tag{2}$$

$$\boldsymbol{r}'(i;i') = \left(\tfrac{1}{R_i} + \tfrac{1}{R_{i'}}\right)^{-1/2} \left(0 \cdots 0 R_i^{-1/2} 0 \cdots 0 - R_{i'}^{-1/2} 0 \cdots 0\right), \quad i,\ i' = 1, \cdots, a.$$

This has been called the squared distance between the two rows i and i'. It is naturally extended to the squared distance between the two clusters of rows. Without any loss of generality we assume the two clusters to be $G_1 = \{1, \cdots, q_1\}$, $G_2 = \{q_1 + 1, \cdots, q_1 + q_2\}$ and then the squared distance between G_1 and G_2 is defined by

$$\chi^2(G_1;G_2) = \|(\boldsymbol{r}'(G_1;G_2) \otimes C')\, \boldsymbol{z}\|^2, \tag{3}$$

$$\boldsymbol{r}'(G_1;G_2) = \left(\tfrac{1}{T_1} + \tfrac{1}{T_2}\right)^{-1/2} \left(\tfrac{\sqrt{R_1}}{T_1} \cdots \tfrac{\sqrt{R_{q_1}}}{T_1} - \tfrac{\sqrt{R_{q_1+1}}}{T_2} \cdots - \tfrac{\sqrt{R_{q_1+q_2}}}{T_2} 0 \cdots 0\right),$$
$$T_1 = \sum_{i \in G_1} R_i, \quad T_2 = \sum_{i \in G_2} R_i.$$

Those squared distances are obviously a part of χ^2 (3) and bounded above by

$$\max_{\boldsymbol{\gamma}'\boldsymbol{r}=0,\ \|\boldsymbol{\gamma}\|=1} \|(\boldsymbol{\gamma}' \otimes C')\, \boldsymbol{z}\|^2 \tag{4}$$

whose asymptotic distribution is shown to be that of the largest root of the Wishart matrix $W(I_{\min(a-1,b-1)}, \max(a-1, b-1))$. This reference distribution has been introduced in Hirotsu (1983) and employed by other authors including Greenacre (1988). Although it improved Gilula (1986) who proposed similar clustering procedure there was still some gap between the squared distances (5), (1) and the maximal reference statistic (5). It is because some optimization procedure is required for defining the squared distance among more than two clusters of rows whereas those squared distances (5) and (1) are uniquely defined by the normalization and orthogonalization. The problem has, however, been solved in Hirotsu (1991) for ANOVA model by defining the

generalized squared distance among any number of clusters of rows and the loss of power in evaluating the significance of clustering has been dissolved. This is actually the Scheffé type multiple comparisons of all the contrasts in rows. In the following Section it will be extended to the contingency table.

3 The generalized squared distance among any number of clusters of rows

Without any loss of generality we assume a partition of rows into m clusters: $G_1 = \{1, \cdots, q_1\}, G_2 = \{q_1 + 1, \cdots, q_1 + q_2\}, \cdots, G_m = \{q_1 + \cdots + q_{m-1} + 1, \cdots, q_1 + \cdots + q_m\}$. Then the generalized squared distance among them is defined by

$$\chi^2 (G_1; \cdots; G_m) = \max \|(\boldsymbol{\gamma}' \otimes C')\, \boldsymbol{z}\|^2, \tag{5}$$

where the maximization is taken with respect to $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_a)'$ under the condition

$$\begin{aligned} &\boldsymbol{\gamma}'\boldsymbol{r} = 0, \qquad \|\boldsymbol{\gamma}\| = 1, \\ &\gamma_i \equiv \lambda_k (R_i/T_k)^{1/2}, \quad i \in G_k, \\ &T_k = \sum_{i \in G_k} R_i, \quad k = 1, \cdots, m. \end{aligned}$$

It is actually the maximization by $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_m)'$ under the condition

$$\sum_{k=1}^{m} \sqrt{T_k}\lambda_k = 0, \quad \sum_{k=1}^{m} \lambda_k^2 = 1. \tag{6}$$

The basic idea is to give a constant coefficient for the rows within a cluster so that it cannot contribute to the maximization.
Let $Y_{kj} = \sum_{i \in G_k} y_{ij}$, $k = 1, \cdots, m$, denote the frequency of the k th cluster at the j th column so that $\{Y_{kj}\}$ gives the $m \times b$ table with the row total T_k collapsing those pooled rows. Then the equation (7) becomes

$$\chi^2(G_1; \cdots; G_m) = \max_{\lambda} \lambda' \begin{pmatrix} \boldsymbol{w}_1' \\ \vdots \\ \boldsymbol{w}_m' \end{pmatrix} (\boldsymbol{w}_1 \cdots \boldsymbol{w}_m)\lambda \tag{7}$$

with $\boldsymbol{w}_k = (T_k/N)^{-1/2}\, C'\left(C_1^{-1/2} Y_{k1}, \cdots, C_b^{-1/2} Y_{kb}\right)'$.
In particular we have

$$(\boldsymbol{w}_1 \cdots \boldsymbol{w}_m)\left(\sqrt{T_1}, \cdots, \sqrt{T_m}\right)' = \sum_k \sqrt{T_k}\boldsymbol{w}_k = NC'\boldsymbol{c} = \boldsymbol{0}$$

suggesting $\left(\sqrt{T_1}, \cdots, \sqrt{T_m}\right)'$ to be the latent vector of $(\boldsymbol{w}_1 \cdots \boldsymbol{w}_m)'(\boldsymbol{w}_1 \cdots \boldsymbol{w}_m)$ corresponding to a zero root. Then the maximization reduces to the problem of the largest root and the condition (6) is automatically satisfied. The

statistic (6) is the same type statistic for the pooled $m \times b$ table as (5) from the original $a \times b$ table.

There may be a case where one cluster is found to be definitely different from the other rows so that the clustering of the rest of rows is interested excluding that cluster, see the example given in Section 5. Then a significant clustering can be searched by applying the maximization with respect to $\boldsymbol{\lambda}$ of (6) with an additional constraint $\lambda_m = 0$, say. This problem is not the problem of the largest root of a matrix but can be solved by a standard optimization procedure. Including this case all the squared distances are bounded above by the statistic (5). In considering the case of one row in each cluster it is found that there is no loss of power in applying this reference distribution in the clustering procedures.

4 Clustering rows of the contingency table with natural ordering in columns

When there is a natural ordering in columns such as the severity of disease several procedures have been proposed taking it into consideration among which the max cumulative chi-squared (max χ^{*2}) is a very natural extension of the previous Section just replacing the C' by the matrix $C^{*'}$ of the cumulative contrasts. It has been verified to have higher power, as compared with the linear rank statistic, over the wide range of the ordered alternatives, see Hirotsu (1983, 1991, 1993) for details as well as the definition of $C^{*'}$. In case of $a \geq b$ the reference distribution is that of the largest root $w_{(1)}$ of the non-orthogonal Wishart matrix $W\left(C^{*'}C^{*}, a-1\right)$. Although the distribution is very difficult to handle, a very simple χ^2-approximation

$$\rho_{(1)}\chi^2_{a-1} \tag{8}$$

has been introduced in Hirotsu (1991) for the ANOVA model and shown nicely to improve the usual normal approximation (Anderson, 2003), where $\rho_{(1)}$ is the largest root of $C^{*'}C^{*}$. Now we can calculate the asymptotic cumulants of the largest root from the asymptotic expansion of the latent roots of the non-orthogonal Wishart matrix in terms of the i.i.d. standardized normal variables obtained in Sugiura (1977). The dominating term of the first two cumulants of $w_{(1)}$ are $\kappa_1 = (a-1)\rho_{(1)}$ and $\kappa_2 = 2(a-1)\rho^2_{(1)}$, respectively, which explains the very nice property of the χ^2-approximation. The precision of the approximation depends, however, heavily on the difference between the largest and the second largest roots of $C^{*'}C^{*}$ and the ANOVA model corresponds to the balanced case of the contingency table with $C_1 = \cdots = C_b$, where the difference of the second root $(b/6)$ and the largest root $(b/2)$ is large enough. In the unbalanced case the approximation is even better if the first and the second roots are more different each other than the balanced case but the approximation becomes poor if they become closer and a is not

so large. Then we recommend the χ^2-approximation adjusting the first two cumulants with higher order terms obtained after a considerable calculation based on Sugiura's expansion. We do not present the detailed equation here because of the limited space and give only Table 1 where we compare the upper 5 percentile by the proposed method with the previous approximations. In Table 1 the first line gives the approximation based on the Zonal polynomial expansion by Aida and Hirotsu (1983). The 0-approximation denotes the usual normal approximation, the 1-approximation the χ^2-approximation of (4) and the 2-approximation the proposed method using the higher order terms of cumulants. We give here only the case of $b = 3$ where $\rho_{(1)} + \rho_{(2)} = 2$ and $\rho_{(1)} = 1.5$ is the balanced case. It is seen that the proposed method using the higher order terms of cumulants works well for $\rho_{(1)}$ as small as 1.2 and $a > 5$. In case $b > 4$ the situation does not seem to become worse since the difference between the first and the second largest roots becomes usually large. We also performed a MCMC(Marcov Chain Monte Carlo) simulation for the balanced case and verified it to be very close to the Zonal polynomial expansion.

		$\rho_{(1)}$					
$a-1$	method	1.2	1.3	1.4	1.5	1.6	1.7
5	Zonal	15.01	15.62	16.40	17.27	18.21	19.17
	0-approx.	12.24	13.26	14.28	15.30	16.32	17.34
	1-approx.	13.28	14.39	15.50	16.61	17.71	18.82
	2-approx.	15.23	15.45	16.17	17.06	18.02	19.02
10	Zonal	23.78	25.05	26.54	28.14	29.79	31.47
	0-approx.	20.83	22.56	24.30	26.03	27.77	29.51
	1-approx.	21.97	23.80	25.63	27.46	29.29	31.12
	2-approx.	21.99	24.30	26.13	27.89	29.62	31.36
20	Zonal	39.58	42.12	44.91	47.81	50.76	53.75
	0-approx.	36.48	39.52	42.56	45.60	48.64	51.69
	1-approx.	37.69	40.83	43.97	47.12	50.26	53.40
	2-approx.	38.32	41.62	44.65	47.65	50.66	53.69

Table 1. Comparing the 5 percentile of the approximation methods.

5 Example

Greenacre (1988) applied the method of Hirotsu (1983) for evaluating the significance of clustering of rows of the 8×5 contingency table reported by Guttman (1971). It cross tabulates 1554 Israeli adults according to the row categories of principal worries and the column categories depending on their place of residence and that of their respective fathers. The row categories are as follows:

1. OTH-other worries
2. POL-political situation
3. MIL-military situation
4. ECO-economic situation
5. ENR-enlisted relative
6. SAB-sabotage
7. MTO-more than one worry
8. PER-personal economics.

He could isolate PER from the other rows since the squared distance $\chi^2(1,2,3,4,5,6,7;8) = 77.90$ exceeds the critical value 23.55, the upper 0.05 point of the largest root of $W(I_4, 7)$. However, any partition of the larger cluster into two clusters cannot give a significant distance since the largest is $\chi^2(1;2,3,4,5,6,7) = 20.77$. Therefore he stopped here suggesting there might be some heterogeneity in the larger cluster. Now we can try the generalized squared distance among any number of clusters by the method of Section 3 obtaining $\chi^2(1;2,4;3,5,6,7) = 21.76$ and $\chi^2(1;2;3,4,7;5,6) = 24.47$ as the maximal generalized distance among three and four clusters, respectively. Thus finally we have five significant clusters (1), (2), (3,4,7), (5,6), and (8). This nicely explains Fig. 2 of Greenacre (1988) which shows the chi-square components along first two principal axes in the correspondence analysis. The homogeneity within the two clusters (3, 4, 7) and (5, 6) is very acceptable from the original squared distances between two rows.

6 Conclusion

The generalized squared distance has been introduced which dissolves the loss of power in the process of clustering rows of a contingency table by the row-wise multiple comparison approach. When there is a natural ordering in columns an order sensitive squared distance is introduced. In this case a very nice χ^2-approximation has been obtained for the largest root of a non-orthogonal Wishart matrix as a reference distribution. A two-way table reported by Guttman (1971) and analyzed by Greenacre (1988) is reanalyzed and a very nice interpretation of the data has been obtained.

Acknowledgment

The author thanks Dr. S. Tsukada who verified the author's lengthy calculation of the asymptotic cumulants to be correct by a computer software MATHEMATICA.

References

AIDA, M. and HIROTSU, C. (1983): A method for comparing the multinomial distributions under order constraints and the table of percentiles. *Applied Statistics 12, 101-110.* (In Japanese)

ANDERSON, T.W. (2003): *An introduction to multivariate statistical analysis.* (3rd Ed.) Wiley Intersciences, New York.

GILULA, Z. (1986): Grouping and association in contingency tables: An exploratory canonical correlation approach. *J. American Statistical Association 81, 773-779.*

GREENACRE, M.J. (1988): Clustering the rows and columns of a contingency Table. *J.Classification 5, 39-51.*

GUTTMAN, L. (1971): Measurement as Structural Theory. *Psychometrika 36, 329-347.*

HIROTSU, C. (1977): Multiple comparisons and clustering rows in a contingency table. *Quality 7, 27-33.* (In Japanese)

HIROTSU, C. (1983): Defining the pattern of association in two-way contingency tables. *Biometrika 70, 579-589.*

HIROTSU, C. (1991): An approach to comparing treatments based on repeated measures. *Biometrika 75, 583-594.*

HIROTSU, C. (1993): Beyond Analysis of Variance Techniques: Some Applications in Clinical Trials. *International Statistical Review 61, 183-201.*

HIROTSU, C., OHTA, E., HIROSE, N. and SHIMIZU, K. (2003): Profile analysis of 24-hours measurements of blood pressure. *Biometrics 59, 907-915.*

SUGIURA, N. (1973): Derivatives of the characteristic root of a symmetric or a Hermitian matrix with two applications in multivariate analysis. *Communications in Statistics 1, 393-417.*

Projection-Based Clustering for High-Dimensional Datasets

Iulian Ilieş[1] and Adalbert Wilhelm[2]

School of Humanities and Social Sciences, Jacobs University Bremen
Campus Ring 1, 28759 Bremen, Germany, *a.wilhelm@jacobs-university.de*

Abstract. Methods for finding groups of similar objects in large data sets with the purpose of facilitating data interpretation play an important role in exploratory data analysis. However, classical cluster analysis methods do not scale well with an increased number of objects and/or dimensions. Recent work in the field has focused on designing algorithms that can overcome these difficulties (e.g. Goil, Nagesh, & Choudhary, 1999; Aggarwal and Yu 2000) while providing meaningful solutions. We propose an extension of the OptiGrid algorithm (Hinneburg and Keim 1999), with influences from hierarchical divisive methods. Given a group of objects, the present algorithm selects potentially favorable one-dimensional projections using Principal Component Analysis, searches for low-density points in these projections, and then partitions the data by a hyperplane passing through the best point found, if any.A number of measures were taken to maintain the memory load and the execution time as low as possible: non-recursive implementation of the algorithm, sampling of objects and dimensions for quick finding of near-optimal projections, and simplified local minima search and scoring using histograms. Tests on synthetic datasets indicate our method can detect and recover efficiently groups of objects that are distinguishable from the remaining data on at least one direction, provided that all extant clusters are linearly separable; interlocked clusters are generally subdivided. The running time is sublinear with respect to the number of objects and of found clusters, and subquadratic in the number of dimensions.

Keywords: cluster analysis, principal component analysis, projections, low-density points, sampling

1 Introduction

Cluster analysis is one of the traditional statistical routines with the aim of classifying objects based on their features in such a way that similar objects constitute a homogeneous subgroup while distinct subgroups or clusters are as separate as possible. The principal task in cluster analysis can be transformed in various related questions such as density estimation problems (Scott, 1992) or data compression problems (Gersho and Gray, 1992). Hence, a variety of different approaches for cluster analysis have been investigated in the fields of statistics, pattern recognition and machine learning, see Jain et al. (1999) and Berkhin (2002).

During the last decades there has been a growing emphasis on data mining (exploratory analysis) in very large data sets (Gordon, 1999). This imposes severe additional computational constraints on cluster analysis methods (Berkhin, 2002). Traditional algorithms normally do not address the problem of processing large data sets with a limited amount of resources (i.e. system memory and processor time). These challenges led to the development of a variety of new clustering methods, and also constitute the focus of this paper.

2 High dimensionality

Clustering high dimensional data presents a two-fold problem. Firstly, the higher the dimensionality, the more likely is to have a large number of irrelevant attributes, and the clusters become very hard to find (Berkhin, 2002). Secondly, the"curse of dimensionality" is manifesting: the data becomes sparse, and the concept of proximity looses meaning in more than 15 dimensions (Aggarwal, Hinneburg, & Keim, 2001). The distance to the nearest objects becomes of the same order as the distance to any other object, and the proportion of populated grid cells decays rapidly (Hinneburg & Keim, 1999). Direct use of feature transformation techniques (e.g. factor analysis) cannot help, since the relative distances between objects are preserved. If the noise level is very high, the effectiveness of such methods is significantly decreased (Parsons, Haque, & Liu, 2004). The alternative - dimensionality reduction via feature selection algorithms (Becher, Berkhin, & Freeman, 2000) is also prone to problems. If clusters reside in different subspaces, it is difficult to restrict the set of dimensions without pruning attributes that are relevant to only some of the clusters. This type of data motivated the development of subspace clustering methods, the most successful ones in high dimensional settings. Notable examples are the algorithms MAFIA (Goil et al., 1999), OptiGrid (Hinneburg & Keim, 1999), and ORCLUS (Aggarwal and Yu 2000).

3 Proposed method

We propose a projection-based partitioning method based on OptiGrid (Hinneburg and Keim, 1999). The algorithm selects potentially favorable one-dimensional projections using correlations and principal component analysis, and then searches for low-density points in these projections. The analyzed set of objects is divided by an orthogonal hyper- plane passing through the best split point found, if any. Unless special stopping conditions are imposed, the process will terminate when no more splits are found for any of the extant subsets. The algorithm returns a partitioning of the data set in non-dividable groups (i.e. all intermediate clusters are discarded). Measures such as iterative implementation, objects and dimensions sampling, and simplified search

for projections and local minima, ensure the computational efficiency of the algorithm. These topics are presented in more detail below.

3.1 Implementation

A recursive algorithm would have memory requirements proportional to the size of the data and the logarithm of the number of clusters, since it would involve creating successive copies of parts of the data that are stored in the memory until the recursion stops. Since we would want the algorithm to be able to work with large data sets (e.g. of size comparable to the available physical memory), it is necessary to adopt an iterative implementation, which allows for having only one copy of the data (the entire set) loaded in the working memory. To do that, we store clusters as sets of indices of the included objects (thus, the values on the different attributes are easily accessible via reading operations). These sets are organized as a first-in-first-out list; clusters are processed in order, and, if partitioned, the resulting sub-clusters are appended at the end of the list, while the current cluster is removed.

3.2 Algorithm

In this section, we give a detailed description of the analysis process for one group of objects (see also Figure 1). The procedure can be roughly separated into two parts: in the first one (Steps 1 to 3 below), the method seeks potentially favorable projections, while in the second part (Steps 4 to 6) it searches for the best separating points along these projections. If a satisfactory point is found, the group is partitioned (Step 7). For each step, we also specify the computational complexity (in brackets). We use the following notations: n is the number of objects in the cluster, d is the number of dimensions, and σ is a bounded sub-linear function.

Step 1 - Correlation - estimation ($O(\sigma(n)d^2)$)

A random subset of objects is selected with the purpose of estimating the correlations between all attributes. Standard Pearson correlations are then calculated using this reduced sample of objects. The selection is index based: a random vector of length equal to the cluster size is generated by sampling from the standard uniform distribution (with values in $[0, 1]$). An object is selected if the corresponding random number is smaller than the ratio between the desired sample size and the number of objects in the cluster. The size of this sample is between 100% of the total number of objects and 10000, depending on the cluster size (following an empirically determined function).

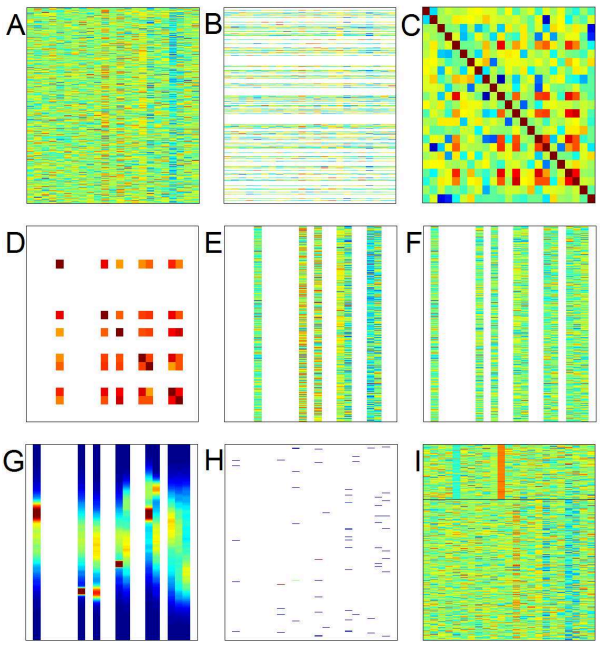


Fig. 1. A. Initial data. B. A random subsample selected for estimating correlations. C. Correlations matrix (in absolute values). D, E. Correlations matrix and data restricted to the dimensions involved in the highest correlations. F. Projections of the data on the directions calculated via PCA. G. Smoothed histograms (ASHs) of the projected data. H. Position of local minima points, color coded by score. I. Resulting partition of the initial data.

Step 2 - Principal component analysis ($O(\sigma(d)^2 \log(\sigma(d))))$

The data from one group of objects constitutes a cloud in its high-dimensional space. Typically, the "natural" axes of this cloud are not parallel to the coordinates. In particular, this is true if there are (large enough) correlations between some of the attributes. Having a group of dimensions pairwise correlated indicates that in the corresponding subspace there is a strong preference for one direction. Along this line, the distribution of objects would have a higher variance (the objects are more widely distributed). If we would project the data along this direction, it would be more likely to find good local minima, which separate well the objects into two or more groups. In order to find such directions of highest variances, we rely on Principal Component Analysis (PCA). To avoid the introduction of noise from irrelevant attributes, the analysis is restricted to the dimensions involved in the largest correlations. We sort all attributes by their largest correlations (in absolute value) in descending order, and select the first few. The correlation matrix and the working

data (i.e. the data of all objects in the current cluster) are reduced correspondingly. The directions of highest variance within the selected subspace are then calculated via eigenvalue decomposition of the reduced correlation matrix (following Rencher 2002).

Step 3 - Data projection ($O(n\sigma(d)^2)$)

The set of selected projections includes all coordinate projections in the restricted subspace, as well as the principal components with eigenvalues above the 0.95 threshold (i.e. having variance larger than or equal to one attribute, with a $\pm 5\%$ error margin). These projections are encoded as a loadings (weights) matrix. To project the objects, we first normalize the working data, and then multiply it with the loadings matrix.

3.3 Step 4 - Construction of histograms ($O(n\sigma(n)\sigma(d))$)

On each projection, we approximate the probability distribution function of the current group of objects by average shifted histograms (ASH) (Scott, 1992). These are a smoother version of histograms, obtained by averaging several histograms of the same bin size and different offsets, or, equivalently, by filtering a histogram constructed on a correspondingly finer grid. The number of bins depends on the group size (we tried to have, on average, at least 10 objects per bin, such that very sharp differences between neighboring bins have a low probability of occurrence), while the size of the averaging filter depends on the number of bins. The histograms are subsequently normalized to an average bin count of one (by dividing to the total number of objects in the cluster and then multiplying with the number of bins) in order to facilitate further processing.

Step 5 - Local minima search ($O(\sigma(n)\sigma(d))$)

For each histogram, all points where the first-order divided difference changes sign from $+$ to $-$ (i.e. for which both the left and right neighboring bins have higher values) are marked as local minima. Their positions are saved in a list for quick access.

Step 6 - Local minima scoring ($O(\sigma(n)\sigma(d))$)

In their paper presenting the OptiGrid method, Hinneburg and Keim (1999) suggested that the quality of the possible splitting points (defined as minimal densities separating between dense areas that are above a certain noise level) should be inverse proportional to their densities. We found their approach rather cumbersome (especially since the baseline noise level is hard to specify), and therefore decided to construct a simpler scoring system. In

doing this, we aimed on one hand to define very intuitive rules, and on the other hand to have low computational requirements. We started from the basic definition of a local minimum - a trough in the graph of the distribution function, which separates two regions of higher density. The appropriateness of a local minimum as a partitioning point depends firstly on how well it separates the two high-density regions, and secondly on how dense these regions are actually. Applied to our estimation of the density function by a histogram, these conditions translate into comparing the maximal and average bin counts of the two dense regions to the local minimum bin. This yields two pairs of numbers, which need to be averaged in some way; we chose to employ a geometric mean (i.e. taking the square root of the product), since it penalizes asymmetric cases.

Step 7 - Data partitioning ($O(n)$)

If the best score is larger than a minimal threshold (our test suggest using values of about 0.15; non-meaningful fluctuations in the ASHs have scores up to 0.1), the data is divided by a hyper-plane passing through that point, orthogonal to the corresponding projection. All objects lying to the left of the local minimum (i.e. having a value on the corresponding projection less than the value at the center of the split bin) are distributed to the first resulting group, while the remaining objects (lying to the right) are assigned to the second group.

4 Experimental results

The Fundamental Clustering Problems Suite (FCPS; Ultsch, 2005) is a collection of ten data sets with known classification. All sets are rather simple, having low dimensionality (2-3 attributes) and relatively few objects (in the range of 200 to 4000). They address several clustering problems that cannot be handled properly by the more traditional methods (such as k-means or single linkage; see Ultsch 2005 for details). We ran our method on all these data sets, and compared the obtained solutions to the true classifications (see Figure 2 for some examples). These results indicate that our method recovers correctly clusters that are linearly separable and non-overlapping. If the first condition does not hold, any interlocked clusters are subdivided into smaller groups; the algorithm effectively cuts away pieces from such clusters until the remaining data is linearly separable. This results in several smaller groups which are subsets of the actual clusters. If the second condition fails, and the data contains clusters that are overlapping or close enough to each other that the objects in the contact region cannot be assigned unambiguously to any of the clusters, then misclassifications may occur in that region. However, the main characteristics of the clusters (e.g. high-density regions or centroids) are correctly retrieved. Additional tests on large synthetic data

sets (around 50000 objects in 20-50 dimensions, grouped in several clusters) confirmed these observations.

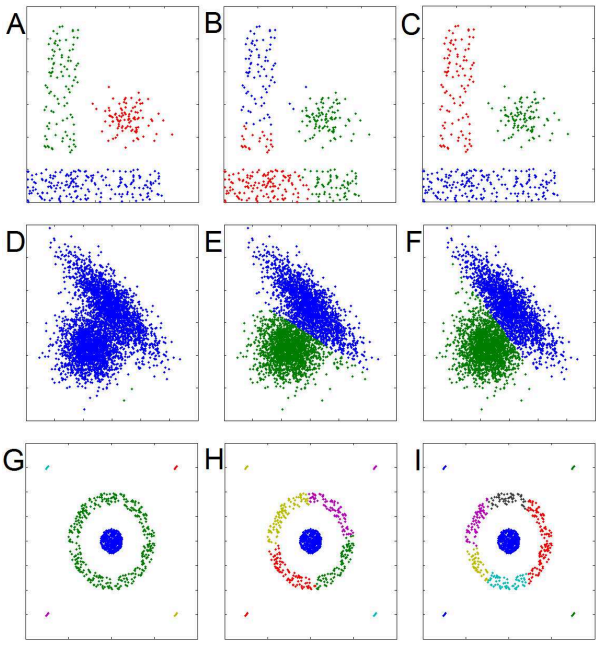


Fig. 2. Comparison between single linkage (A, D, G), k-means (B, E, H), and the proposed method (C, F, I) on three example data sets. For single linkage and k-means, the correct number of clusters was used as input argument. Example 1 (A, B, C): k-means does not perform well if clusters have different variances. Example 2 (D, E, F): single linkage cannot find the correct solution if clusters are touching or overlapping. Example 3 (G, H, I): if the clusters are not linearly separated, our method subdivides them.

5 Discussion

In this paper, we present a new method for finding clusters in data of high dimensionality. What distinguishes our algorithm from other subspace based methods is the low reliance on user specified parameters. Most importantly, the number of clusters does not need to be specified a priori: running in "free mode" (with no parameters), the algorithm will produce a cut through the clusters tree at a level defined by the split score threshold. Based on this preliminary result, a better solution could be obtained at a second run, using

the inferred parameters (e.g. number of clusters, minimal cluster size).Since our algorithm relies on hyper-planes for partitioning the data, it cannot split groups that are linearly non-separable, e.g. dense clusters surrounded by noise on all dimensions, or interlocked clusters. The algorithm will however divide such clusters into smaller subgroups that do not cross cluster boundaries, and hence could allow for better cluster recovery via post-processing. However, the additional investment of computational resources may not be justified: even if such clusters of irregular shapes would exist, they would be more difficult to interpret in a high dimensional context.

References

AGGARWAL, C. and YU, P. (2000): Finding generalized projected clusters in high dimensional spaces. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. Dallas, TX, 70–81.

AGGARWAL, C. and HINNEBURG, A. and KEIM, D. (2001): On the surprising behavior of distance metrics in high dimensional space. In: *Proceedings of the 8th International Conference on Database Theory*, 420-434.

BECHER, J. and BERKHIN, P. and FREEMAN, E. (2000): Automating exploratory data analysis for efficient data mining. In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, MA, 424-429.

BERKHIN, P. (2002): Survey of clustering data mining techniques. *Technical Report*. Accrue Software, San Jose, CA.

GERSHO, A., and GRAY, R.M. (1992): *Vector Quantization and Signal Processing Communications and Information Theory*. Kluwer Academic Publishers, Norwell, MA.

GOIL, S., and NAGESH, H., and CHOUDHARY, A. (1999): MAFIA: Efficient and scalable subspace clustering for very large data sets. *Technical Report*. Northwestern University, Evanston, IL.

GORDON, A.D. (1999): *Classification* (2nd ed.). Chapman & Hall / CRC, Boca Raton, FL.

HINNEBURG, A., and KEIM, D. (1999): Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. In: *Proceedings of the 25th International Conference on Very Large Data Bases*. Edinburgh, Scotland, pp. 506-517.

JAIN, A.K, and MURTY, M.N, and FLYNN, P.J. (1999): Data Clustering: A Review. *ACM Computing Surveys, 31 (3), 264 - 323*.

MILENOVA, B.L. and CAMPOS, M.M. (2002): O-cluster: scalable clustering of large high dimensional data sets. In: *Proceedings of the 2nd IEEE International Conference on Data Mining*, 290-297.

PARSONS, L. and HAQUE, E. and LIU, H. (2004): Subspace clustering for high dimensional data: A review. *ACM SIGKDD Explorations Newsletter, 6 (1), 90- 105*.

RENCHER, A.C. (2002): *Methods of multivariate analysis* (2nd ed.). Wiley, New York, NY.

SCOTT, D.W. (1992): *Multivariate density estimation*. Wiley, New York, NY.

ULTSCH, A. (2005): Clustering with SOM: U*C. In: *Proceedings of the Workshop on Self-Organizing Maps*. Paris, France, pp. 75-82.

A New Approach to Spatial Clustering Based on Hierarchical Structure

Fumio Ishioka[1] and Koji Kurihara[2]

[1] Graduate School of Environmental Science, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan,
fishioka@ems.okayama-u.ac.jp
[2] Graduate School of Environmental Science, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan,
kurihara@ems.okayama-u.ac.jp

Abstract. The importance of statistical analyses for spatial data has grown in a variety of scientific fields. Spatial lattice data is comprised of measurements or observations taken at specific locations or within specific regions. However, there are few approaches for cluster analysis of spatial lattice data. In this paper, we explore cluster analysis for spatial lattice data using echelon analysis. Echelon analysis is a useful technique for investigating the phase-structure of spatial lattice data systematically and objectively. We propose a new zone classification based on the peak of an echelon. The zone classification is demonstrated by some examples.

Keywords: echelon, spatial clustering, lattice data

1 Introduction

The importance of statistical analyses for spatial data has grown in various scientific fields. Spatial lattice data is comprised of measurements or observations taken at specific locations or within specific regions. Cluster analysis based on some types of measures such as a similarity or defined distance has been performed in many fields. However, there are few approaches for cluster analysis of spatial lattice data. GIS (geographic information system) provides powerful tools to study the spatial lattice data, but it is very difficult to describe the spatial clustering based on its spatial structure. For example, a statistical map with shading is used to show how quantitative information varies geographically, but we can only find contiguous clusters in this map with the poor accuracy of visual decoding.

Echelon analysis (Myers et al., 1997) is a useful technique for investigating the phase-structure of spatial lattice data systematically and objectively. The echelons are derived from changes in topological connectivity. Various approaches were performed to analyze the spatial lattice data such as remote sensing data and multi-dimensional image data (Kurihara et al., 2000; Ishioka et al., 2007). In this paper, we define the neighbors and families of spatial lattice data in order to enable the clustering procedure. In addition, we propose a new zone classification based on the peak of an echelon.

2 Classification of spatial data based on echelons

2.1 One-dimensional spatial data

One-dimensional spatial data has the position x and the value $h(x)$ on the horizontal and vertical axes, respectively. For k divided lattice (interval) data, data is taken at the following intervals $l_1(i) = (i-1, i]$, $i = 1, ..., k$. Table 1 shows the 23 intervals named from A to W in order and their values. In order

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
ID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
$h(i)$	1	2	3	4	3	4	5	4	3	2	3	4	5	6	5	6	7	6	5	4	3	2	1

Table 1. One-dimensional spatial interval data.

to utilize the information of spatial positions, we use a cross sectional view of a topographical map as shown in Figure 1.

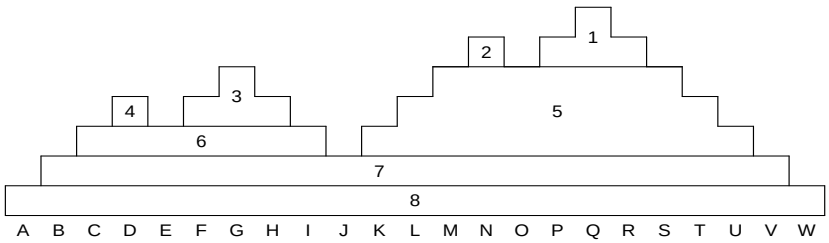


Fig. 1. The hypothetical set of hillforms in one-dimensional spatial data.

There are eight numbered parts with the same topological structures in these hillforms. These parts are called echelons. These echelons consist of peaks, foundations of peaks, and foundation of foundations. Thus we have eight clusters $G(i)$, $i = 1, ..., 8$ for specified intervals based on these echelons. These are the following clusters of the four peaks:

$$G(1) = \{\text{Q,P,R}\},\ G(2) = \{\text{N}\},\ G(3) = \{\text{G,F,H}\},\ G(4) = \{\text{D}\}$$

These are the clusters of foundations:

$$G(5) = \{\text{M,O,S,L,T,K,U}\},\ G(6) = \{\text{C,E,I}\},\ G(7) = \{\text{B,J,V}\},\ G(8) = \{\text{A,W}\}$$

The cluster $G(6)$ is a parent of $G(3)$ and $G(4)$, subsequently $G(6)$ has two children of $G(3)$ and $G(4)$. Therefore we can define the family $FM(G(6))$ for the cluster $G(6)$ by:

$$FM(G(6)) = G(6) \cup G(3) \cup G(4)$$

2.2 Procedure for making echelon classification

In this section we describe the procedure for clustering spatial lattice data. At first, we define the neighbors of spatial data $l_1(i)$, say $NB(i)$. For $l_1(i) = (i-1, i]$, $i = 1, ..., k$, the neighbor is given by $NB(i)$.

$$NB(i) = \begin{cases} \{i+1\}, & i = 1 \\ \{i-1, i+1\}, & 1 < i < k \\ \{i-1\}, & i = k \end{cases} \tag{1}$$

The neighbor of the jth cluster $G(j)$, $NB(G(j))$, is also given by

$$NB(G(j)) = \bigcup_{i \in FM(G(j))} NB(i) - \bigcup_{i \in FM(G(j))} \{i\} \tag{2}$$

where $A - B = A \cap \{B^C\}$ for the sets of A and B.

In order to make the clusters $G(i)$, we find the peaks and the foundations using the following steps in Algorithm 1 and 2. To simplify the procedure, we assume there are no ties.

Algorithm 1. Find the peaks for k cells.
Step1 Initial setting: $i = 0$ and $RN = \{i | 1 \le i \le k\}$.
Step2 Find the i-th peak $G(i)$: $i = i + 1$ and $G(i) = \phi$.
Step3 Find the cell of candidate $M(i)$ for the i-th peak $G(i)$:
If $G(i) = \phi$ then $h(M(i)) = \max_{j \in RN} h(j)$.
Else $h(M(i)) = \max_{j \in NB(G(i))} h(j)$.
Step4 Check the cell $M(i)$ which belongs to $G(i)$ or not:
Set $RN = RN - M(i)$. If $RN = \phi$ then END.
If $h(M(i)) > \max_{j \in NB(M(i)) - G(i)} h(j)$ then $G(i) = G(i) \cup M(i)$ and go to Step 3.
Else If $G(i) = \phi$ then $i = i - 1$ and go to Step2.
Else Go to Step2.

Algorithm 2. Find the foundations. Let l be the number of peaks.
Step1 Initial setting: $i = l$ and $RN = \{i | 1 \le i \le k\} - \bigcup_{j=1}^{l} G(j)$.
If $RN = \phi$ then END.
Else $GN = \{i | 1 \le i \le l\}$ and $FM(G(j)) = G(j)$, $j = 1, ..., l$.
Step2 Find the i-th foundation $G(i)$: Set $i = i + 1$ and $G(i) = \phi$.
Step3 Find the cell of candidate $M(i)$ for the i-th foundation $G(i)$:
If $G(i) = \phi$ then
$h(M(i)) = \max_{j \in RN} h(j)$ and
$CN = \{j | NB(M(i)) \cup FM(G(i)) \neq \phi,\ j \in GN\}$ and
$FM(G(i)) = \bigcup_{j \in CN} FM(G(j))$ and
$GN = GN \cup \{i\} - CN$.
Else $h(M(i)) = \max_{j \in NB(FM(G(i))) \cap RN} h(j)$.
Step4 Check the cell $M(i)$ which belongs to $G(i)$ or not:

Set $RN = RN - M(i)$. If $RN = \phi$ then $G(i) = G(i) \cup M(i)$ and END.
If $h(M(i)) > \max_{j \in NB(FM(G(i)) \cup M(i))} h(j)$ then
$G(i) = G(i) \cup M(i)$ and
$FM(G(i)) = FM(G(i)) \cup M(i)$ and go to Step3.
Else $RN = RN \cup M(i)$ and go to Step2.

2.3 Echelon classification of two-dimensional spatial data

Two-dimensional spatial data has the value $h(x, y)$ for response variable at the position (x, y). In applications such as remote sensing the data are given as pixels of digital values over the $M \times N$ lattice area $l_2(i, j) = \{(x, y) : x_{i-1} < x < x_i, y_{j-1} < y < y_j\}$, $i = 1, ..., N$, $j = 1, ..., M$. The neighbors of cell $l_2(i, j)$ are given by

$$NB(l_2(i,j)) = \{(k,l)|i-1 \le k \le i+1, j-1 \le l \le j+1\} \cap \{(k,l)|1 \le k \le N, 1 \le l \le M\} - \{(i,j)\}. \quad (3)$$

	A	B	C	D	E
1	10	24	10	15	10
2	10	10	14	22	10
3	10	13	19	23	25
4	20	21	12	11	17
5	16	10	10	18	10

Fig. 2. The digital values over a 5×5 array.

By analyzing the two-dimensional spatial data, we can obtain its echelon structure as well as one-dimensional data by the foregoing Algorithms. To illustrate this, we will apply the digital values in the 5×5 array shown in Figure 2. By using the Algorithms, $G(i)$ are calculated in Table 2. The graphical representation based on Table 2 for these array data is shown as the dendrogram in Figure 3.

3 New classification method based on the peaks

In the previous section, we showed that spatial lattice data with neighbor information is classable by the echelons. In this section, we will define a new zone classification based on the peak of echelon. Zone $GE(i)$ can be calculated by using following steps in Algorithm 3.

Algorithm	i	$G(i)$	$FM(G(i))$
1	1	E3, D3, D2	$G(1)$
1	2	B1	$G(2)$
1	3	B4, A4	$G(3)$
1	4	D5	$G(4)$
2	5	C3	$G(j)$ $j = 1, 3, 5$
2	6	E4, A5, D1	$G(j)$ $j = 1, 3, 4, 5, 6$
2	7	C2, B3, C4, D4 and others	$G(j)$ $j = 1, 2, 3, 4, 5, 6, 7$

Table 2. The echelon clusters of the 5 × 5 array.

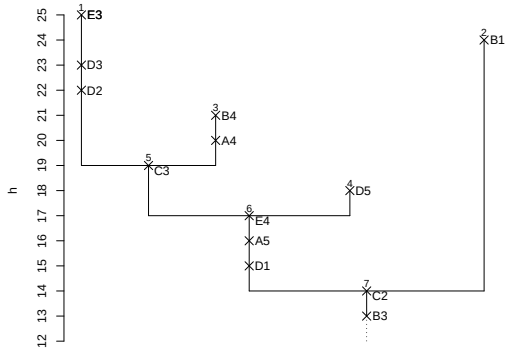


Fig. 3. The relation of clusters for 5 × 5 array.

Algorithm 3. Find the zone. Let l be the number of peaks.
$GE(i) = G(i)$, $i = 1, ..., l$. (Zones $GE(i)$ include i-th peak.)
Step1 Initial setting: $RN = \{i | 1 \leq i \leq k\} - \bigcup_{j=1}^{l} G(j)$.
If $RN = \phi$ then END.
Else $M = \phi$.
Step2 Find the cell of candidate M for the i-th zone $GE(i)$:
$RN = RN - M$. If $RN = \phi$ then END.
Else $h(M) = \max_{i \in RN} h(i)$.
Step3-1 Check the cell M which belongs to $GE(i)$ or not (1):
$ZN = \{i | M \cup NB(G(i)) \neq \phi, i = 1, ..., l\}$.
If $ZN = \phi$ then go to Step3-2.
Else $GE(i) = GE(i) \cup M$, $i \in ZN$ and go to Step2.
Step3-2: Check the cell M which belongs to $GE(i)$ or not (2):
$ZN = \{i | NB(M) \cup GE(i) \neq \phi, i = 1, ..., l,$
(here, except i satisfied at the same time.)$\}$.
$GE(i) = GE(i) \cup M$, $i \in ZN$ and go to Step2.

By using the Algorithm 3, we can classify the Table 1 into each zone based on four peaks; $G(1)$ ={Q,P,R}, $G(2)$ ={N}, $G(3)$ ={G,F,H} and $G(4)$ ={D}.

In the cluster $G(5)$, {S} and {M} are classable in zone $GE(1)$ and $GE(2)$, respectively. As a more interesting result, there are two classes, $GE(1)$ and $GE(2)$, where {O} belongs. These are the following clusters of the four zones:

$$GE(1) = \{Q,P,R,O,S,T,U,V,W\},\ GE(2) = \{N,O,M,L,K,J\},$$
$$GE(3) = \{G,F,H,J,E,I\},\ GE(4) = \{D,E,C,B,A\}$$

These results can be summarized as shown in Table 3 and drawn as an image as shown in Figure 4.

	lattice	$GE(1)$	$GE(2)$	$GE(3)$	$GE(4)$
$G(1)$	P, Q, R	○			
$G(2)$	N		○		
$G(3)$	G, F, H			○	
$G(4)$	D				○
	O	○	○		
	S, T, U, V, W	○			
	M, L, K		○		
	J		○	○	
	I			○	
	E			○	○
	C, B, A				○

Table 3. The zone classification of the one-dimensional data.

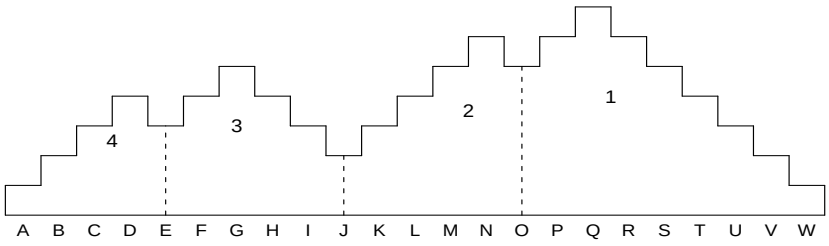


Fig. 4. The zone classification of the one-dimensional data.

Next, we perform the zone classification according to the 5×5 array shown in Figure 2. We consider four peaks, $G(1)$ ={E3, D3, D2}, $G(2)$ ={B1}, $G(3)$ ={B4, A4} and $G(4)$ ={D5}, as the central part of classification, and apply the Algorithm 3. As a result, Figure 2 can be classified as four zones as shown in Table 4 and Figure 5.

4 Classification of geospatial lattice data

Geospatial lattice data are areal-referenced values $h(D_i)$ within spatial regions D_i, $i = 1, ..., n$. The regional features are mainly investigated over

	lattice	$GE(1)$	$GE(2)$	$GE(3)$	$GE(4)$
$G(1)$	E3, D3, D2	◯			
$G(2)$	B1		◯		
$G(3)$	B4, A4			◯	
$G(4)$	D5				◯
	C3	◯		◯	
	E4	◯			◯
	A5			◯	
	D1	◯			
	C2	◯	◯		
	B3			◯	
	C4	◯		◯	◯
	D4	◯			◯
	others				

Table 4. The zone classification of the the 5×5 array.

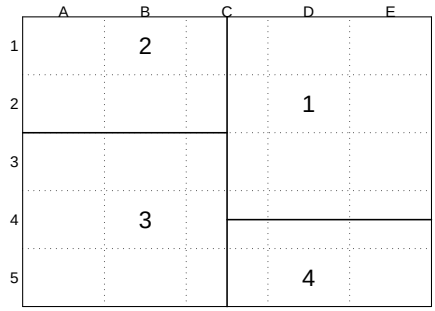


Fig. 5. The zone classification of the 5×5 array.

lattice regions like the watersheds in the state, the counties in the state, the states in the USA and so on. If we have neighbors $NB(D_i)$ for each spatial region, we can also classify them using geospatial lattice data based on the Algorithms.

Table 5 shows the rates of unemployment in 1997 corresponding to 50 states in the USA. This data is quoted from the Statistical Abstract of the United States 1998. The USA is divided into irregular states, so the lattice is irregular. The neighbor information is also shown in this table; neighbors are defined by shared borders. Although the states of Alaska and Hawaii are isolated states, we assume that the states of Alaska and Hawaii are connected to the states of Washington and California, respectively. We show the results of classifications based on Algorithms in Table 5. There are nine peaks of unemployment in the USA. Accordingly, we can classify these into nine zones. The result of the zone classification can be drawn as Figure 6. By using these new spatial clustering methods, we gain a better understanding of regional unemployment patterns.

State name		neighbor of states	rates	G	GE
Alabama	AL	FL, GA, MS, TN	51	13	6
Alaska	AK	(WA)	79	1	1
Arizona	AZ	CA, CO, NM, NV, UT	46	15	4, 5
Arkansas	AR	LA, MS, MO, OK, TN, TX	53	10	6
California	CA	AZ, NV, OR, (HI)	63	4	4
⋮		⋮	⋮	⋮	⋮

Table 5. States in USA, their neighbors, unemployment tates, G and GE based on Algorithms. For reasons of space, this table has been shortened.

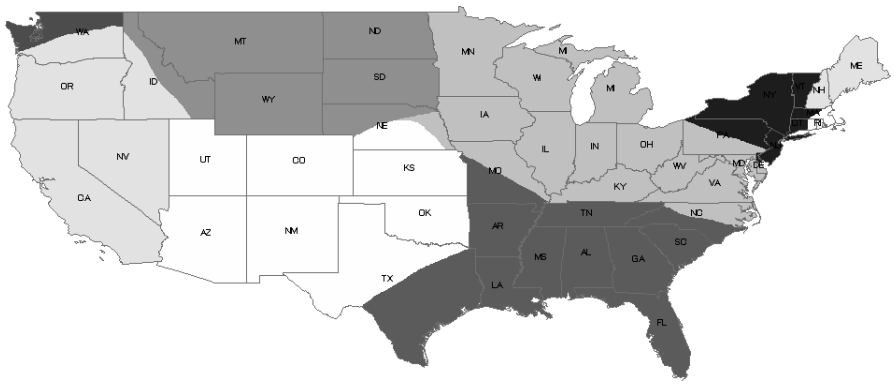


Fig. 6. The zone classification of the rates of unemployment for the states in 1997.

5 Conclusion

We proposed a new cluster analysis for spatial lattice data. We established Algorithms to perform the spatial clustering based on an echelon approach. In particular, zone classification can easily provide us with an objective expression of areal segmentation for several types of geographical data.

References

ISHIOKA, F., KURIHARA, K., SUITO, H., HORIKAWA, Y. and ONO, Y. (2007): Detection of Hotspots for 3-dimensional Spatial Data and Its Application to Environmental Pollution Data. *Journal of Environmental Science for Sustainable Society, 1, 15-24.*

KURIHARA, K., MYERS, W.L. and PATIL, G.P. (2000): Echelon analysis of the relationship between population and land cover patterns based on remote sensing data. *Community Ecology, 1, 103-122.*

KURIHARA, K. and ISHIOKA, F. (2007): Classification of Spatial Data based on the Pattern of Hierarchical Structure and its Applications. *Journal of the Japan Statistical Society, 37(1), Series J, 113-132.*

MYERS, W.L., PATIL, G.P. and JOLY, K. (1997): Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics, 4, 131-152.*

A Toolbox for Bicluster Analysis in R

Sebastian Kaiser and Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstrasse 33, 80539 München, Germany,
firstname.lastname@stat.uni-muenchen.de

Abstract. Over the last decade, bicluster methods have become more and more popular in different fields of two way data analysis and a wide variety of algorithms and analysis methods have been published. In this paper we introduce the R package `biclust`, which contains a collection of bicluster algorithms, preprocessing methods for two way data, and validation and visualization techniques for bicluster results. For the first time, such a package is provided on a platform like R, where data analysts can easily add new bicluster algorithms and adapt them to their special needs.

Keywords: biclustering, two-way-clustering, software, R

1 Introduction

Biclustering is an important new technique in two way data analysis. After Cheng and Church (2000) followed the initial bicluster idea of Hartigan (1972) and started to calculate bicluster on microarray data, a wide range of different articles were published dealing with different kinds of algorithms and methods to preprocess and analyze the results of such methods. Comparisons of several bicluster algorithms can be found, e.g., in Madeira and Oliveira (2004) or Prelic et al. (2006).

Consider a two-way data set of form

$$\begin{array}{c|cccccc} & c_1 & \dots & c_i & \dots & c_m \\ \hline r_1 & a_{11} & \dots & a_{i1} & \dots & a_{m1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_j & a_{1j} & \dots & a_{ij} & \dots & a_{mj} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ r_n & a_{1n} & \dots & a_{in} & \dots & a_{mn} \end{array}$$

with rows r_i and columns c_j and entries a_{ij}. The goal of biclustering is to find subgroups of rows and columns which are as similar as possible to each other and as different as possible to the rest.

As noted above, the recent boom in biclustering has originated in the analysis of genetic data, where rows r_i correspond to genes and columns c_j to conditions, and a_{ij} is the expression level of gene r_i under condition

c_j. The task is to find groups of genes which are co-regulated under some conditions. However, two way data appear also in other research fields. E.g., in marketing biclusters can be used to group consumers into market segments which have several preferences in common. Traditional market segmentation methods like k-means cluster analysis or mixture models use the same set of variables for all clusters, while bicluster methods are able to select different sets of variables for different segments (Goveart and Nadif (2003)).

This article is organized as follows: In Section 2 we give an introduction to the structure of R package `biclust` including a brief description of the theory and algorithms of the five bicluster methods that have been implemented yet. We also focus on preprocessing of the data, validation and visualization methods, and show the advantage of storing the resulting bicluster output in consistent classes for all methods. In Section 3, we demonstrate usage of the package using the popular yeast data (Barkow et al., 2006). Finally, we give a short summary and point out future plans for extending `biclust`.

2 R package `biclust`

Most bicluster methods have been developed for a particular data analysis problem, some authors provide standalone software for their algorithms, while others only describe the algorithms in papers. Barkow et al. (2006) provide a first toolbox with several algorithms within a single graphical user interface. While the GUI has the advantage that it is easy to use, it has also the disadvantage that it is again monolithic software, which cannot be changed easily by users, and results cannot be directly used as input for other statistical methods.

We have therefore started to implement a comprehensive bicluster toolbox in R (R Development Core Team, 2007). It provides a growing list of bicluster methods, together with pre-processing and visualization techniques, using S4 classes and methods (Chambers, 1998). The software is open source and freely available from R-Forge at `http://R-Forge.R-project.org`.

One of the main design principles of the package is to provide the results as an entity of `Biclust-Class`, an S4-class containing all information needed for postprocessing of results. It consists of the four slots `Parameters`, `RowxNumber`, `NumberxCol` and `Number`. Slot `Parameters` contains parameters and algorithm used, `Number` the number of biclusters found. The `RowxNumber` and `NumberxCol` slots represents the biclusters that have been found. They are both logical matrices of dimension (rows of data $\times$ number of biclusters found) with a TRUE-value in `RowxNumber[i,j]` if row i is in bicluster j. `NumberxCol` is the same for the columns, but due to computational reasons, here the rows of the matrix represent the number of biclusters and the columns represent the columns of the data. So by simply calling

```
data[ Biclust@RowxNumber[,a] * Biclust@NumberxCol[a,] ]
```

the values of the bicluster `a` can be extracted.

Objects of class `Biclust-class` are created using a uniform interface for all bicluster methods by calls of form `biclust(x,method=BiclustMethod,...)`. This generic function takes as inputs the preprocessed data matrix `x`, a bicluster algorithm represented as a `Biclustmethod-Class` and additional arguments (`...`) for the latter.

In the following we give a brief description of the five algorithms already implemented in the package, subsection headings correspond to the name of the respective `Biclustmethod-Class`. The naming scheme is `BCxxx` where `xxx` is an abbreviation for the name of the algorithm. Some methods have been chosen because open source code from the original authors is available, others have been newly implemented to make the overall toolbox as comprehensive as possible. Of course, there is always room for improvement, and more methods will be added to the package in the future. See also van Mechelen and Schepers (2006) for a discussion on main directions of bicluster calculation. Algorithms are described in alphabetic order and, if not stated otherwise, functions were implemented in interpreted S code.

2.1 `BCBimax()`

The Bimax algorithm presented by Prelic et al. (2006) finds subgroups in a binary matrix where all entries are one. The algorithm iterates the following two steps:

1. Rearrange the rows and columns to concentrate ones in the upper right of the matrix.
2. Divide the matrix into two submatrices.

Whenever in one of the submatrices only ones are found, this submatrix is returned. In order to get satisfying results the method has to be restarted several times with different starting points.

Although the algorithm was originally designed to deliver ideas for bicluster validation, it can also be used as a bicluster method itself. In our implementation we used the original and fast C Code of Prelic et al. (2006).

2.2 `BCCC()`

The CC method implements the algorithm by Cheng and Church (2000). Starting from an adjusted matrix, where normalization or simple standardization preprocessing is suggested, they define a score

$$H(I,J) = \frac{1}{\|I\|\|J\|} \sum_{i\in I, j\in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2, \tag{1}$$

where a_{iJ} is the mean of row i,a_{Ij} is the mean of column j and a_{IJ} is the overall mean. They call a subgroup a bicluster if the score is below a level

α and above a δ-fraction of the whole data. The algorithm itself has three major steps:

1. Deleting rows and columns with a score larger than *alpha* times the matrix score.
2. Deleting rows and columns with largest scores.
3. Adding Rows or Columns until *alpha* level is reached.

These steps are repeated until a maximum number of biclusters is reached or no bicluster is found. The result are constant bicluster where all a_{ij} are nearly on the same level. Choosing an appropriate preprocessing methods is essential for good solutions.

2.3 `BCPlaid()`

This algorithm is an improvement of the plaid model of Lazzeroni and Owen (2002) by Turner et al. (2005). The original algorithm was fitting layers k to the model

$$Y_{ij} = (\mu_0 + \alpha_{i0} + \beta_{j0}) + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk} + \varepsilon_{ij} \tag{2}$$

using ordinary least squares (OLS), where μ, α, β represent mean, row and column effects and ρ and κ identify if a row or column is member of the layer, respectively. After the computation of the residuals of the obtained data, the calculation has the following steps:

1. Update all parameters one after another S times.
2. Calculate the sum of squares of the layer (LSS) using the resulting parameters.
3. Compare Result with random permutation and return bicluster if LSS is higher.

The algorithm terminates when no new layer (bicluster) is found. In the new faster algorithm of Turner et al. (2005), OLS is replaced with a binary least square algorithm. In our implementation we used the original code from Turner et al. (2005).

2.4 `BCSpectral()`

The bicluster algorithm described by Kluger et al. (2003) includes several preprocessing steps, like normalization, independent scaling, bistochastization and log interactions. The goal is to find a checkerboard structure of the data matrix and in order to identify it, the following steps are performed:

1. Reorder the data matrix and choose a normalization method.

2. Compute a singular value decomposition to get eigenvalues and eigenvectors.
3. Depending on the chosen normalization methods, construct biclusters beginning from the largest or second largest eigenvalue.

The quantity of bicluster depends on the number and value of the eigenvalues. The biclusters found have higher or lower values than the rows and columns around them and are arranged in a checkerboard structure.

2.5 `BCXmotifs()`

The Xmotifs algorithm of Murali and Kasif (2003) searches for rows with constant values over a set of columns. For gene expression data, they call the biclusters "conserved genes expression motifs", short "Xmotifs". Again, finding a good preprocessing method is crucial, because the main aspect of their algorithm is to define a gene state where a gene (row) is called conserved, if it has the same state in all samples (columns). One way to deal with gene states is to simply discretize the data (for example with function `discretize()`). Once the data matrix represents the states, the algorithm works by choosing a random number of columns n times and performs the following steps:

1. Choose a subset from these columns and collect all rows with equal state in this subset.
2. Collect all columns where these rows have the same state.
3. Return the bicluster if it has the most rows from all found and is also larger than a *alpha* fraction of the data.

To collect more than 1 bicluster the calculation can be reran without the rows and columns already found or just return the smaller combinations found.

This algorithm finds submatrices where all rows have the same value structure over the columns. So here it is possible to find groups with a large variance in their values in the row direction.

2.6 Other functions of package `biclust`

In addition to the cluster algorithms described above, our package provides several methods for data pre-processing, some of which have already been described above. Other functions provide utilities for cluster validation, like an adaptation of the well known Jaccard index for comparison of two cluster results in `jaccardind()`. Function `constantVariance()` implemements the variation index of Madeira and Oliveira (2004).

Another important focus of the package is visualization of bicluster results. As the results are stored in consistent classes, it is easy to implement visualization techniques which work for results of different algorithms. Parallel

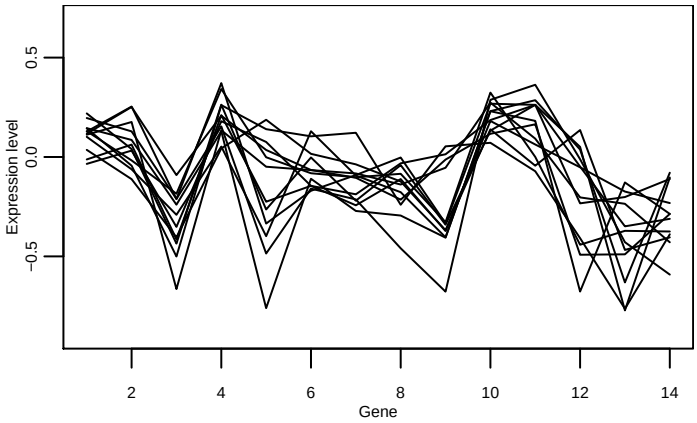


Fig. 1. Example for parallel coordinates plot: Expression levels of conditions across their genes in the 4th bicluster in the result of the Xmotifs algorithm.

coordinates (function `parallelCoordinates()`) can be used to visualize similarity of rows over columns within a bicluster. Heatmaps (`drawHeatmap()`) highlight the difference between the bicluster and the surrounding rows and columns. The bubbleplot (`Bubbleplot()`) of Santamaria et al. (2007) represents biclusters in two dimensions, its position in the graph is a two dimensional representation of the row and column combination in the cluster, the size of the bubble corresponds to the size of the bicluster. Hence, a bicluster containing another bicluster is drawn as a big bubble around a smaller one.

3 Example with yeast data

After introducing the main functions of the package we now want to show how the package works. As a standard example we ran all the algorithms on the BicatYeast data from Barkow et al. (2006). To do so the data has to be preprocessed and committed to the biclust function together with the chosen algorithm (here Xmotifs) and parameters:

```
> data(BiclustYeast)
> x<-discretize(Bicatyeast)
> res<-biclust(x, method=BCXmotifs(), alpha=0.05, number=50)
```

To visualize the result you can simply call any visualization function on the result, for example:

```
> parallelCoordinates( x=BicatYeast, result=res, bicluster=4)
```

The output of this code can be seen in Figure 1.

Table 1 shows the pairwise Jaccard indices of all bicluster algorithms. The Jaccard index is a measure of similarity between two cluster results, zero

	BCPlaid	BCXmotifs	BCCC	BCSpect.	BCBimax
BCPlaid	1.0000	0.0007	0.0116	0.0000	0.0000
BCXmotifs	0.0007	1.0000	0.1789	0.0935	0.0000
BCCC	0.0116	0.1789	1.0000	0.0898	0.0036
BCSpectral	0.0000	0.0935	0.0898	1.0000	0.0000
BCBimax	0.0000	0.0000	0.0036	0.0000	1.0000

Table 1. Bicluster results similarity measure with an adaptation of Jaccard index.

means no concordance, one means that the results are identical. It can be seen that all algorithms find very different sets of biclusters. This can be partly explained by different pre-processing steps which were necessary such that the data conform to the respective assumptions of the algorithms. Another important aspect is that we selected the first algorithms to implement to get a collection of algorithms which differ from each other as much as possible. It is now very easy for practitioners to try various bicluster methods in R and choose the one which works best for given data set.

4 Summary and future work

In this article, we gave a short introduction to R package `biclust` with special emphasis on the algorithms that have already been implemented. We explained some of the design principles and object structures of the packages, and demonstrated usage of the software on a real word data set. All methods implemented share common infrastructure for data pre-processing, storing results and visualization. It is now very easy to add new bicluster methods, because the modular design allows for re-use of existing building blocks.

As a next step, we will do benchmark experiments comparing all the algorithms. As demonstrated in the example above, bicluster results do strongly depend on the clustering algorithm. It will be interesting to investigate which algorithm works best for which type of data, and how sensitive the algorithms are with respect to their parameters and data pre-processing steps.

5 Acknowledgments

Package `biclust` is joint work with Rodrigo Santamaria.

References

BARKOW, S., BLEULER, S., PRELIC, A., ZIMMERMANN, P., and ZITZLER, E. (2006): Bicat: a biclustering analysis toolbox. *Bioinformatics, 22,1282–1283.*

CHAMBERS, J.M. (1998): *Programming with data: A guide to the S Language.* Chapman & Hall, London.

CHENG, Y. and CHURCH, G.M. (2000): Biclustering of expression data. In: *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, 1,93–103.

GOVEART, G. and NADIF, M. (2003): Clustering with block mixture models. *Pattern Recognition*, 36, 463–473.

HARTIGAN, J.A. (1972): Direct Clustering of a data matrix. *Journal of The American Statistical Association, 67,12079–12084.*

KLUGER, Y., BASRI, R., CHANG, J.T., and GERSTEIN, M. (2003): Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research, 13,703–716.*

LAZZERONI, L. and OWEN, A. (2002): Plaid models for gene expression data. *Statistica Sinica, 12,61–86.*

MADEIRA, S.C. and OLIVEIRA, A.L. (2004): Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1(1),24–45.*

VAN MECHELEN, I. and SCHEPERS, J. (2006): A unifying model for biclustering. In: *Compstat 2006 - Proceedings in Computational Statistics*, 81–88.

MURALI, T. and KASIF, S. (2003): Extracting conserved gene expression motifs from gene expression. In: *Pacific Symposium on Biocomputing*, 8,77–88.

PRELIC, A., BLEULER, S., ZIMMERMANN, P., WIL, A., BÜHLMANN, P., GRUISSEM, W., HENNING, L., THIELE, L., and ZITZLER, E. (2006): A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics, 22(9),1122–1129.*

SANTAMARIA, R., THERON, R., and QUINTALES, L. (2007): A framework to analyze biclustering results on microarray experiments. In: *8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07)* ,Springer, Berlin, 770–779.

TURNER, H., BAILEY, T., and KRZANOWSKI, W. (2005): Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis, 48,235–254.*

Robust Classification and Clustering Based on the Projection Depth Function

Daniel Kosiorowski[1]

Department of Statistics, Cracow University of Economics
ul. Rakowicka 27, 31 - 510 Cracow, Poland, *daniel.kosiorowski@uek.krakow.pl*

Abstract. In this paper we propose classification and clustering procedures based on the projection depth function. We studied the performance of the propositions on various multivariate data sets simulated from skewed, fat tailed and including outliers distributions.

Keywords: depth function, robust classification, robust clustering

1 Introduction

We assume we have measured p variables of n observations that are sampled from k different populations $\mathcal{C}_1$,...,$\mathcal{C}_k$. We also assume we know the membership of each observation with respect to the population. We set such measurements into a $n \times p$ data matrix $\mathbf{Z}$ whose rows are portioned into k groups corresponding to k considered populations. We call the matrix $\mathbf{Z}$ a training sample.

One of the objects of classification is to rationally allocate a new observation $\mathbf{x}$ to one of these populations on the base of his measurement and the information included in the training sample. An index $i \in \{1, 2, ..., k\} = \mathbb{Y}$ corresponding to the population $\mathcal{C}_i$ is entitled as a label. A classification rule is a function $L : \mathbb{R}^p \ni \mathbf{x} \longrightarrow i \in \mathbb{Y}$. The function assigns to the vector $\mathbf{x}$ the prediction of the label $L(\mathbf{x}) \in \mathbb{Y}$.

In a clustering issue our aim is to group n objects considered with respect to p variables $C_0 = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ into k homogeneous classes, where k is also unknown. In other words our aim is to find an optimal partition of C_0 into k homogeneous disjoint nonempty subsets C_1,...,C_k, $k \geq 2$, $C_i \cap C_j = \oslash$, $i \neq j$, $\bigcup C_i = C_0$.

Classical classification or clustering methods like linear or quadratic discriminant functions, k means algorithm often assume a multivariate normality of the components of the mixture generating observations (see Hand (1981)). They are highly influenced by outliers because they are based on an empirical mean vector and covariance matrix or a pair-wise Euclidean distance matrix of the data. They are inappropriate at contaminated data sets or in the case of skewed populations and became useless in the case of the nonexistence of moments of the population.

These facts motivate us to propose nonparametric and robust classification and clustering procedures referring to a data depth concept (for details, see Dyckerhoff (2004) or Zuo (2003)).

Let $\mathcal{P}_0$ denote a set of all probability measures on $(\mathbb{R}^p, \mathcal{B}^p)$ and let $\mathcal{P}$ be a subset of $\mathcal{P}_0$. A statistical depth assigns to each probability measure $P \in \mathcal{P}$ a real function $D(\cdot|P) : \mathbb{R}^p \longrightarrow [0,1]$, the so called depth function w.r.t. P. The depth function measures the centrality of a point $\mathbf{x} \in \mathbb{R}^p$ w.r.t. P enabling for extending in a unified way to the multivariate setting the univariate methods of signs and ranks, order statistics, quantiles and outlyingness measures. The set of all points that have a depth of at least α is called the $\alpha-$ central region (for theoretical details see Dyckerhoff (2004) and references therein).

A symmetric projection depth of a point $\mathbf{x} \in \mathbb{R}^p$ being a realization of some p dimensional random vector $\mathbf{X}$ with probability distribution F, $PD(\mathbf{x}, F)$ is defined as

$$PD(\mathbf{x}, F) = \left[1 + sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^t\mathbf{x} - Med(\mathbf{u}^t\mathbf{X})|}{MAD(\mathbf{u}^t\mathbf{X})}\right]^{-1}, \tag{1}$$

where Med denotes the univariate median, $MAD(Z) = Med(|Z - Med(Z)|)$.

The projection depth function possesses among others an affine invariance property, induced location and scatter estimators have high finite sample replacement breakdown points and good properties in terms of Hampel's influence function and Huber's maximum bias (for details see Zuo (2003)).

It is hoped that robustness measures introduced in Section 2 will enable us to make further proposition validation. In Section 3 our classification rule proposition is introduced. Section 4 presents our clustering procedure propositions. Section 5 describes results of our propositions simulation studies. In Section 6 some conclusions are presented.

Furthermore the sample projection depth $PD(\mathbf{y}|C)$, where $C = \{\mathbf{x}_1, ..., \mathbf{x}_j\}$ denotes the sample drawn from a population $\mathcal{C}$, has been calculated using an approximation algorithm proposed by R. Dyckerhoff with an application of thousand one dimensional random projections (see Dyckerhoff (2004)).

2 Robustness quantification

In order to measure the quality of our classification and clustering procedure propositions in a context of robustness, we focus our attention on a notion of a finite sample breakdown point. However, an acceptable definition of the breakdown point of the classification and especially, the clustering procedure seems to be a challenge. This is partly due to the fact that these procedures jointly use location, scale and regression parameters estimators (for a general framework of the issue see Davies and Gather (2005)). Below we propose a simple and intuitive proposition of the breakdown point of the classification rule referring to the training sample. We say the classification procedure

breaks down if probability of misclassification is greater than 50% (Some authors believe rather that this probability should depend on the number of k clusters. For a simplicity of the considerations we move a discussion of the issue to further studies).

DEFINITION 1 : Consider k populations $\mathcal{C}_1$,...,$\mathcal{C}_k$, $k \geq 2$ and a fixed training sample $\mathbf{z}$ representing the populations. **An actual prediction error** of a classification rule L is equal

$$Err\left(L, \mathbf{z}\right) = P\left\{L\left(\mathbf{X}\right) \neq i : \ \mathcal{C}_i \ generates \ \mathbf{X}, i = 1, ..., k\right\}, \tag{2}$$

where $\mathbf{X}$ denotes an observation independent from the training sample.

DEFINITION 2 : Consider k populations $\mathcal{C}_1$,...,$\mathcal{C}_k$, $k \geq 2$ in p dimensions and a training sample $\mathbf{Z}$ representing these populations. A breakdown point of the training sample $\mathbf{Z}$ of a classification rule L in the j class $\mathcal{C}_j$ is defined as

$$BP_j\left(L\right) = inf_m\left\{m/n_j : P\left\{L\left(\mathbf{X}\right) \neq j : \ \mathcal{C}_j \ generates \ \mathbf{X}\right\} > 1/2\right\}, \tag{3}$$

where m rows of the $n_j \times p$ sub matrix $\mathbf{Z}_j$ of the training sample $\mathbf{Z}$ corresponding to a sample of n_j observations drawn from population $\mathcal{C}_j$ are replaced by arbitrary rows (outliers), $\mathbf{X}$ denotes an observation independent from the training sample $\mathbf{Z}$.

An overall breakdown point of the training sample $\mathbf{Z}$ of the classification rule L is defined as

$$BP\left(L, \mathcal{C}_1, ..., \mathcal{C}_k\right) = min_j BP_j\left(L\right). \tag{4}$$

In order to examine a clustering procedure proposition we used a method of silhouettes proposed by Rousseeuw (1987). Silhouettes offer the advantage that they depend on the actual partition of the objects and not on the clustering algorithm. We studied how the silhouettes changed for a data set consisted of n observations generated by a known mixture of k, $k \geq 2$ populations after a replacement of m points of the data set by arbitrary points. We studied a nature of changes (i.e. a width and/or a height of the silhouettes) for mixtures of various numbers and types of the distributions. However the method we used was heuristic.

In order to study local robustness of the clustering proposition we used the method of the empirical Hampel's influence curve. We discriminated between point contamination inside and outside convex hulls of the clusters and between influence on probability of misclassification and on a quality of clustering (i.e. on the silhouettes features).

3 Projection depth classification rule

Let $C_j = \left\{\mathbf{x}_1, ..., \mathbf{x}_{n_j}\right\}$, $j = 1, ..., k$, denote $n_j \times p$ sub matrice $\mathbf{Z}_j$ of the training sample $\mathbf{Z}$ corresponding to the sample of n_j observations drawn

from population $\mathcal{C}_j$ and let $\mathbf{y}$ denote a new observation independent from $\mathbf{Z}$. Our aim is to assign $\mathbf{y}$ to one of the considered populations $\mathcal{C}_j$, $j = 1, ..., k$.

It is well known that any location depth function D provides a depth classification rule

$$L(\mathbf{y}|\mathbf{z}) = argmax_j D(\mathbf{y}|C_j), \tag{5}$$

which assigns $\mathbf{y}$ to that population $\mathcal{C}_j$ in which $\mathbf{y}$ is deepest.

The depth induced classification rules studied among others Mosler and Hoberg (2006). They introduced a classification rule based on a modified zonoid depth. Jörnsten (2004) introduced a classification rule based on L_1 depth. Both propositions however depend on an existence of moments of the population generating data. Christmann (2006) introduced a classifier related to support vector machines and based on a regression depth.

It is worth noticing that most of the location depth functions vanish outside the convex hull of the data set. This implies that a point $\mathbf{y}$ lying outside the convex hulls of all classes cannot be classified by the depth. In order to cope with that problem Mosler and Hoberg (2006) proposed to combine the zonoid depth with the Mahalahanobis depth, Jörnsten (2004) used L_1 depth which does not vanish outside the convex hull of the data set.

We propose a strategy involving using one of the best depth functions, namely projection depth, and joining a classified point into a training sample during the depth ranking of the considered populations calculation.

PROPOSITION 1: Let C_j, $j = 1, ..., k$, denote a sample from a population $\mathcal{C}_j$ used as a part of a training sample $\mathbf{z}$. In order to classify a new observation $\mathbf{y} \in \mathbb{R}^p$ consider the following projection depth classification rule

$$L(\mathbf{y}|\mathbf{z}) = argmax_j PD(\mathbf{y}|C_j \cup \{\mathbf{y}\}), \tag{6}$$

where $PD(\mathbf{y}|C)$ denote the sample $C = \{\mathbf{x}_1, ..., \mathbf{x}_s\}$ projection depth of $\mathbf{y}$.

Note: In order to identify outliers we can prefix a certain threshold $\beta \in (0, 0.5)$ say $\beta = 0.2$. If $PD(\mathbf{y}|C_j \cup \{\mathbf{y}\}) < 0.2$ for $j = 1, ..., k$, then we flag the observation $\mathbf{y}$ as being a potential outlier and suspend the classification decision to a careful content-related analysis.

4 Projection depth clustering procedure

There are many fields of interest for which well known algorithms partitioning a set of objects into k clusters such as the $k-$ means or $k-$ nearest neighborhoods, have not appropriate statistical properties. These methods are not robust to outliers. In order to find a robust alternative to these techniques we focus our attention to the projection depth function. There are not many propositions in the literature referring to the data depth concept. We only found Jörnsten (2004) who introduced a clustering procedure based on L_1 data depth. In our opinion it is due to a computational complexity of many well known combinatorial depth functions. We propose therefore using the approximation algorithm described by Dyckerhoff (2004).

PROPOSITION 2: Suppose we are interested in a partition of a data set $C_0 = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ into k homogonous clusters C_1,...,C_k, $2 \leq k < n$. Consider the following procedure:

Step 1: Start with an initial partition C_1,..., C_k, e.g. with the partition obtained by an application of some known clustering technique like a $k-$ multivariate medians partitioning algorithm.

Step 2: Compute $PD(C_1)$,...,$PD(C_n)$ i.e. sample projection depths of the points belonging to the clusters separately for each of them.

Step 3: For the cluster C_i , $i = 1, ..., k$, identify the set of observations $S_i = \{\mathbf{x}_i \in C_i : PD(\mathbf{x}_i|C_i) \leq \beta\}$, where β is a prefixed threshold.

Step 4: For a random subset $D \subset S := \bigcup S_i$ define a new partition $\widetilde{C_1}$,..., $\widetilde{C_k}$ with D randomly relocated to the initial partition.

Step 5: Denote by $vol\,(PD_\alpha(C_i))$ a volume of the $\alpha-$ sample central projection region in the cluster C_i, where α is a prefixed threshold, $PD_\alpha(C_i) = \{\mathbf{x} \in C_i : PD(\mathbf{x}|C_i) \geq \alpha\}$. If following condition holds

$$\sum_{i=1}^{k} vol\,(PD_\alpha(C_i)) > \sum_{i=1}^{k} vol\left(PD_\alpha(\widetilde{C_i})\right), \tag{7}$$

then C_1,..., $C_k \leftarrow \widetilde{C_1}$,..., $\widetilde{C_k}$ otherwise keep C_1,..., C_k.

Step 6: Iterate 2 – 5 for a prefixed number of times say 30 times.

Step 7: If no moves were accepted for the last m iterations terminate the algorithm.

Notes: We propose prefixing $\beta = 0.2$, $\alpha = 0.3$ and $m = 10$ for the parameters in the above clustering procedure. In order to obtain the initial partition of the data we propose using a modified $k-$ medians algorithm where projection medians are chosen as clusters representatives or partitioning around medoids (relatively fast technique). Depending on the subject matter and the task at hand we also propose to identify observations in each obtained cluster with relatively small value of depth to further content related analysis.

Example: Let $\mathbf{x}_1 = (-3.8, 1.7)$, $\mathbf{x}_2 = (-2.4, 1.5)$, $\mathbf{x}_3 = (1.9, -0.19)$, $\mathbf{x}_4 = (-0.99, 0.57)$, $\mathbf{x}_5 = (-2.7, 2.9)$, $\mathbf{x}_6 = (20.8, 20.7)$, $\mathbf{x}_7 = (21.6, 18.9)$, $\mathbf{x}_8 = (20.7, 20.15)$, $\mathbf{x}_9 = (18.9, 21.8)$, $\mathbf{x}_{10} = (19.6, 22.8)$. Let $C_0 = \{\mathbf{x}_1, ..., \mathbf{x}_{10}\}$ be a data set we are interested in a partition into clusters. Let $C_1 = \{\mathbf{x}_1, ..., \mathbf{x}_5\}$, $C_2 = \{\mathbf{x}_6, ..., \mathbf{x}_{10}\}$, $\tilde{C}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8\}$, $\tilde{C}_2 = \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_9, \mathbf{x}_{10}\}$ denote the clusters.
We calculate $PD(C_0 = \{0.31, 0.37, 0.17, 0.28, 0.32, 0.46, 0.3, 0.47, 0.26, 0.24\}$, $PD(C_1) = \{0.2, 0.44, 0.13, 0.36, 0.1\}$, $PD(C_2) = \{0.32, 0.26, 0.28, 0.07, 0.16\}$, $PD(\tilde{C}_1) = \{0.01, 0.01, 0.35, 0.1, 0.4\}$,
$PD(\tilde{C}_2) = \{0.03, 0.02, 0.009, 0.035, 0.034\}$. We have $vol\,(PD_{0.01}(C_0)) = 226.4$, $vol\,(PD_{0.01}(C_1)) = 11.2$, $vol\,(PD_{0.01}(C_2)) = 5.7$, $vol\left(PD_{0.01}(\tilde{C}_1)\right) = 75.6$,

$vol\left(PD_{0.01}(\tilde{C}_2)\right) = 198.8$. Hence for $\alpha = 0.01$ the partition C_1, C_2 is better than the trivial C_0 and $\tilde{C}_1$, $\tilde{C}_2$ partitions.

5 Results

Statistical properties of the proposed projection depth classification rule (PDR) in comparison to linear (LDF) and quadratic (QDF) discriminant functions were investigated using simulations and well known empirical data sets.

A. Table 1 and Table 2 show a performance of the proposed classification rule on the Fisher's well known data set consisting of 150 measurements on three species of iris considered with respect to sepal length, sepal width, petal length and petal width. We considered $25 \times 25 \times 25$ and $40 \times 40 \times 40$ training sample sizes, on base of them all observations belonging to data set was classified. The results show that in this case proposed procedure exhibits comparable properties to the classical methods.

Classification rule	LDF	QDF	PDR
Actual prediction error	2.6%	4%	3.3%

Table 1. Fisher's data on three species of Iris. A comparison of linear discriminant function (LDF), quadratic discriminant function (QDF) and the proposed classification rule (PDR) in the case of the 3×25 training sample.

Classification rule	LDF	QDF	PDR
Actual prediction error	3.8%	2.6%	2.6%

Table 2. Fisher's data on three species of Iris. A comparison of linear discriminant function (LDF), quadratic discriminant function (QDF) and the proposed classification rule (PDR) in the case of the 3×40 training sample.

B. We simulated 100 two dimensional data sets of sizes 3000 from the equal size contamination of Marshall - Olkin distribution (1000) , isotropic normal distributions (1000) and skewed Student T with two degrees of freedom distribution (1000). We drew training samples of sizes $100 \times 100 \times 100$ from the data sets. On the basis of the training samples, data sets were classified using the proposed procedure. Table 3 shows performance of the proposed classification rule. The results shows that in this case proposed procedure exhibits much better properties in comparision to the classical methods.

Classification rule	LDF	QDF	PDR
Actual prediction error	12.6%	12.6%	0.3%

Table 3. A comparison of linear discriminant function (LDF), quadratic discriminant function (QDF) and the proposed classification rule (PDR) in the case of the equal contribution mixture of Skewed T, Marshall-Olkin and isotropic normal distributions in the case of 3×100 the training sample.

C. In order to estimate the overall breakdown point of the training sample we simulated data sets consisting of 3000 observations generated by an equal contribution contamination of three skewed T Student distribution with different location and shape parameters. Next we replaced the 0%,1%,...,10% observations in the 3×100 training sample drawn from simulated earlier data sets by outlying observations. We calculated the actual error of classification after that replacement. Table 4 shows the results for the fraction of outliers in the training sample varying from 0% to 10%. The results show very good properties of the proposition in terms of robustness to the outliers.

Fraction of outliers	0%	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
Err(L)	0.004	0.008	0.017	0.023	0.03	0.043	0.04	0.05	0.05	0.07	0.07

Table 4. Results of the estimation of the overall BP of the projection depth classification rule in case of an equal contribution mixture of three skewed bivariate Cauchy distributions. A fraction of outliers in the training sample vs. an actual prediction error.

D. In order to examine our clustering procedure proposition we simulated data sets from various mixtures of skewed and heavy tailed distributions and populations including a moderate number of outliers. We validated the resulted partitions of the simulated data sets using silhouettes. The results show that our proposition properly reflected a clustering structure actually present in the simulated data sets. In the case of the clearly separated convex clusters our proposition produce clusters with relatively small "within" distances in comparison to "between" clusters distances. The silhouettes we obtained were relatively wide. In the case of an existence of outliers our proposition is much more robust than $k-$ means algorithm, in the case of the proposition we obtained a higher overall average silhouette width. The proposition is also useful in the case of nonexistence of moments of the population generating data (i.e. a mixture of multivariate Cauchy distributions). We did not observe significant changes of the silhouettes properties and probability of misclassification of the observation estimates in cases of point contaminations (i.e. our proposition seems to be locally robust).

6 Conclusion

We presented two projection depth based methods for classification and clustering. The simulation studies showed that classification rule proposition have good statistical properties in a context of the robustness. The proposition seems to be a competitive classifier to well known classifiers and others depth induced classification rules. The proposed clustering procedure performed well on the simulated data sets. The clustering results analysis showed that our proposition was robust to a moderate fraction of outliers and generated clusters that could be used to predict sample labels better than $k-$ means algorithm.

We are currently working on a further development of the proposed methods i.e. among others for a simplification of the computational aspects of the procedures (we focus our attention on the properties of a projection pursuit approach proposed by Dyckerhoff (2004)) and obtaining an idea about the number of natural clusters that are really present in the data sets (we study the possibilities of replacing a Schwarz information criterion by maximum depth criterion in a mixture based clustering modeling). For an application of the presented issues in a statistical theory of shape see Kosiorowski (2007).

References

CHRISTMANN, A. (2006): Regression depth and support vector machine, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science 72, 71–85*.

DAVIES, P.L. and GATHER, U. (2005): Breakdown and groups(with discussion and rejoinder), *The Annals of Statistics 33, 977–1035*.

DYCKERHOFF, R. (2004): Data Depths Satisfying the Projection Property. *Allgemeines Statistisches Archiv 88, 163–190*.

HAND, D. (1981): *Discrimination and Classification*. Wiley, Chichester.

MOSLER, K. and HOBERG, R. (2006): Data analysis and classification with the zonoid depth. In: Liu, R. Serfling, D. Souvaine (Eds.): *Data Depth: Robust Multivariate Analysis,Computational Geometry and Applications*. American Mathematical Society,49–59.

HUBERT, M. and DRIESSEN, K.V. (2004): Fast and Robust Discriminant Analysis, *Computational statistics and data analysis 45 (2), 301–320*.

JÖRNSTEN, R. (2004): Clustering and Classification based on the L_1 Data Depth, *Journal of Multivariate Analysis 90 (1) ,67–89* .

KOSIOROWSKI, D. (2007): Nonparametric Equity of Two Shapes Test Based on Multivariate Quantile Functional, *Bulletin of the ISI 56th Session*.

ROUSSEEUW, P.J. (1986): Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics 20, 53–65*.

ZUO, Y. (2003): Projection Based Depth Functions and Associated Medians. *The Annals of Statistics 31 (5), 1460 – 1490*.

Random Generation of Pyramids: A New Method Proposed

Vasco Machado[1] and Fernanda Sousa[2]

[1] Departamento de Saúde Pública, ARSN, I.P.
Rua Anselmo Braancamp, 144 4000-078 Porto,
Faculdade de Engenharia, Universidade do Porto - CEC
Rua Dr. Roberto Frias, s.n., 4200-465 Porto, Portugal, *vmpmachado@clix.pt*
[2] Faculdade de Engenharia, Universidade do Porto - DEC and CEC
Rua Dr. Roberto Frias, s.n., 4200-465 Porto, Portugal, *fcsousa@fe.up.pt*

Abstract. In this paper we propose a new method for random generation of pyramids. A pyramid is the most common result of an Ascending Pyramidal Clustering method, and represents a classificatory structure, with particular properties, over the elements of the set to classify. A pyramid is a dendrogram generalization, and then has a more complex structure associated. The method we propose now, called *QuikRAP*, presents some improvements face to a previous one, the *RAP* method. The performance of the two methods is discussed, based on theoretical and simulation studies, as well as a comparison between the two algorithms.

Keywords: cluster analysis, pyramidal clustering, pyramid, pyramids random generation, simulation

1 Introduction

In the last decades many works were developed in the context of random generation of trees. In a general form we may say that this interest has been motivated by the applications of Monte Carlo studies in classification. As important works, for the presented paper and for a more complete list of references on the subject, we refer to Furnas (1984), Lapointe and Legendre (1991), Podani (2000) and Sousa (2000). The classificatory structures comparison is an incontestable requirement in classification and many measures have been constructed to compare different kinds of those structures. However, statistical distributions of these measures, in this context, are unknown. A solution to this problem is given by the random generation of classificatory structures, which allows assessing the significance of structures comparisons.

A dendrogram, or a classification tree, is the most common graphical result drawn out an Ascending Hierarchical Clustering (A.H.C.) method. In what follows, we will use the term dendrogram for the graphic representation as well as for the hierarchical structure associated. The random generation of dendrograms has gained an increasing attention and several algorithms

have been proposed to generate weighted dendrograms. In a weighted dendrogram a numerical value of a continuous variable, called fusion level index, is attributed to each internal node. The distribution of this variable is, in general, unknown and depends on the two choices required in clustering process: the comparison function between pairs of elements of the set to classify and the comparison function between clusters associated to the aggregation criterion. The use of the fusion level index numerical values is frequently questioned, then to retain the corresponding ordinal values, of the values of the fusion level index, is an alternative approach, which corresponds to work with fully ranked or global-order invariant dendrograms (GOI). The *Double Permutation* method, Lapointe and Legendre (1991), the *Uniform Generation* method, Sousa (2000), and the *RA - Random Agglomeration* method, Podani (2000), allow it to generate uniform weighted dendrograms, (*sensu*) Furnas (1984). Analytical and simulation studies, Sousa (2000) and Tendeiro (2005) showed that these three methods have similar behaviors and produce identical results. Sousa (2000) proposed also a method of random generation with a form parameter, which allows producing dendrograms with propensity to an advance fixed type.

The work here presented aims at random generation of a more complex type of classificatory structures, the pyramids. The application of an Ascending Pyramidal Clustering (A.P.C.) method, to a set of multivariate data, produces a pyramid as principal result. In a natural sequence of investigation it was proposed, Machado (2007), a method of random generation of pyramids, called *RAP - Random Generation Algorithm of Pyramids*, which generates a pyramid for a fixed number of terminal nodes (any natural number great than 1). In essential the *RAP* generation algorithm acts in accordance with the method of A.P.C., proposed by Bertrand (1986).

To improve some particular aspects, alternative A.P.C. methods were proposed subsequently. Among these we refer the *QuikCAP* method, developed by Mfoumoune (1998). In this paper a new method for pyramids generation is proposed, the *QuikRAP*, which introduces some improvements to the *RAP* method, and which takes as a base the Mfoumoune's A.P.C. method.

In the next section, Section 2, some definitions of Pyramidal Ascending Clustering are presented, necessary to the next sections understanding, and the characteristics of the principal methods are referred.

In Section 3 the motivation for the random generation of pyramids subject is introduced, the algorithm *RAP* is detailed as well as his performance and limitations.

The new method for pyramids generation, proposed in this work, is presented in Section 4. An explanation of his acting and the most important differences from the *RAP* method are including.

Finally, in Section 5, some experimental results are presented and the algorithms performance is discussed.

2 Pyramidal classification

Pyramids are a generalization of hierarchies and were introduced by Diday (1984) and Bertrand (1986). A pyramid is a collection of subsets of the set of objects to be clustered, with specific properties. In particular two not disjointed clusters are not necessarily nested. Another attractive property of a pyramidal representation is its ability to produce a few numbers of orders on the set of objects, which are compatible with the proximities between objects.

A hierarchy is a nested sequence of partitions of the data set. Closer to the initial data and giving a more accurate order on the objects, a pyramidal classification produces overlapping classes instead of partitions. An A.P.C. begins with the most refined partition (with singleton classes) and in each step (corresponding to a pyramid level) the most resembled classes are merged.

Let E be a set of n elements to be clustered. Formally we said that P, $P \in \mathcal{P}(E)$[1], is a *pyramid* on E if:

(i) $E \in P$;
(ii) $\{a\} \in P,\ \forall\, a \in E$;
(iii) $\forall\, p, p' \in P,\ p \cap p' = \emptyset$ or $p \cap p' \in P$;
(iv) an order θ, compatible with P, exists.

Given a pyramid P, a class $p \in P$ is a *successor* of a class $p' \in P$ (and p' *predecessor* of p) if $p \subseteq p'$ and doesn't exist $p'' \in P :\ p \subseteq p'' \subseteq p'$. Any class has no more than two predecessors.

Let f be an application on a pyramid P that assumes nonnegative real values. (P, f) is called a *indexed pyramid* if, $\forall\, p, p' \in P$, we have

$$f(p) = 0 \iff \exists\, a \in E : p = \{a\} \quad \text{and} \quad p \subset p' \Longrightarrow f(p) \leq f(p').$$

The quantity $f(p)$ is called the heigth of cluster p. An order θ defined on E is compatible with a *pyramidal index* d if

$$\forall\, x, y, z \in E, \quad x <_\theta y <_\theta z \quad \Longrightarrow \quad d(x,z) \geq max\{d(x,y), d(y,z)\}.$$

Given a class $p \in P$, we designate by C_p the connected component of p, by $min(p)$ the minimum element of p and by $max(p)$ the maximum element of p , according to the order θ (for mathematical definitions see Bertrand (1986) and Mfoumoune (1998)). A class $p \in P$ is defined as *maximal* if it doesn't have predecessors and as *extreme class* if $\big(min(p) = min(C_p)$ or $max(p) = max(C_p)\big)$. $p \in P$ is an *internal class* if it exists a class q such that $\big(min(q) < min(p)$ and $max(p) < max(q)\big)$.

The introduction of other definitions would make the understanding of the next sections easier, nevertheless his extension makes his inclusion inadvisable. For this subject we propose as references Bertrand (1986), Mfoumoune

[1] set of the parts of E.

(1998) and Machado (2007). In spite of some notions aren't defined formally, they can be explained in the intuitive form. For example, we understood for *active class* a class that can be joined in following agregations and for *free class* a class that, with an inversion in the order of the elements, becomes an extreme class.

3 Random generation of pyramids

The random generation algorithms for different kinds of clustering structures are very useful tools in clustering methods performance analysis and in validation studies.

In several aspects the previously proposed method for random generation of pyramids, called ***R**andom Generation **A**lgorithm of **P**yramids* (***RAP***), works similar to the algorithm of A.P.C. proposed by Bertrand (1986). In the RAP method the pair of classes to merge is random generated. It may be understood as an extension, for pyramids, of the RA method (Podani (2000)) for dendrograms. The RAP algorithm was developed in Matlab and it generates pyramids for any number of terminal nodes.

Before explaining the RAP algorithm some used notation must be introduced.

- c^0_{act} - vector of classes with the n terminal nodes, numbered of 1 up to n;
- c^i_{act} - vector of active classes in the iteration i;
- c^i_j $(j = 1, 2)$ - class j generated in the iteration i;
- c^i_{agr} - vector of possible classes to merge with c^i_1;
- c^i - new class formed in the iteration i;
- θ - vector of dimension n associated to the compatible order on the elements[2];
- CC - matrix associated to the connected components;
- P - matrix associated to the pyramid. When the class k, with $k \geq n+1$, is formed a new line is added to the matrix P.
- M - matrix $n \times n$, at the begin all the elements are equal to zero, built step by step and, at the end, when the order θ is defined, it will be a Robinson matrix.

Next we present, in a simplified form, the principal steps of the RAP algorithm.

[2] on an abusive form we are using the same letter for the order and for the associated vector.

The *RAP* algorithm

Iteration 0: the number of terminal nodes, n, is introduced and the variables c_{act}^0, CC, θ, P and M are begun.

For i from 1 to the algorithm stopping[3].

Iteration i: .

- random generation of the couple $\left(c_1^i, c_2^i\right)$ of classes to merge:
  - random and uniform generation of an element of the vector c_{act}^{i-1}, the class c_1^i;
  - definition of the vector of posible classes to merge, c_{agr}^i: a vector is built with the classes that are possible to join with c_1^i, satisfying the conditions of aggregation (see Bertrand (1986) and Machado (2007));
  - random and uniform generation of an element of the vector c_{agr}^i, the classe c_2^i.
- updating of: order θ, matrix CC, associated matrix of the pyramid P, matrix M and vector of the active classes c_{act}^i.

Output: order θ, matrix P associated to the pyramid and Robinson matrix $MR = M(\theta)$.

The *RAP* algorithm, in his iterative process (from the base to the top), presents fundamentally two limitations that were identified in Machado (2007): the way the couple of classes to merge is produced and the unjust elimination of internal classes of the set of active classes.

Fig. 1. Example to illustrate some limitations of the *RAP* algorithm.

In the *RAP* algorithm the couples of classes to merge is not produced in an uniform way, i.e., the possible generated couple of classes to merge don't have equal probability. For example, in Figure 1, the pairs of possible classes to merge are $(1,3)$, $(1,4)$, $(2,3)$, $(2,4)$, $(3,4)$, $(3,5)$ and $(4,5)$. Therefore, in a uniform distribution, the couple $(3,5)$ has probability $\frac{1}{7}$ to be generated. With the *RAP* algorithm the probability of this couple being obtained is $\frac{3}{20}$, that results from the first produced class to be 3 and the second one 5, or the first one to be 5 and the second one 3. Then the probability of generating the couple $(3,5)$, using the *RAP* algorithm, is $\frac{1}{5} \times \frac{1}{4} + \frac{1}{5} \times \frac{1}{2} = \frac{3}{20}$.

[3] the algorithm stops when the class that contains all the elements is formed.

The random generation of a pyramid implicates the resolution of numerous algorithimcs problems, in particular, the algorithms of partial or total inversion of the connected components. The aggregation of two classes attributes normally an arbitrary order between the elements, that can be revised in a second aggregation. The partial inversion in the order of the elements is not effectuated in the *RAP* algorithm.

For example, in Figure 1, the aggregation of the classes 3 and 5 forms the class 6. With the *RAP* algorithm, after this moment, the class 2 is excluded because it is an internal class. However, a partial inversion of the new formed connected component (constituted by the elements 1, 2 and 3), i.e., an inversion in the order of the elements of the class 5, allows class 2 as a possible class to merge in next iterations. So, in the *QuikRAP* algorithm are considered active classes, moreover the extreme classes and the maximal classes, the free classes that are transform into extreme classes with a partial inversion of the connected component to which they belong.

4 *QuikRAP* algorithm

The new random generation of pyramids algorithm, *QuikRAP*, appears like an adaptation of the *RAP* algorithm presented in the previous section. Thus, *QuikRAP* algorithm aims at removing some limitations of the *RAP* algorithm. Alternative methods of A.P.C. were developed with progressive improvements, namely the *QuikCAP* method proposed by Mfoumoune (1998). The improvements introduced on the *QuikRAP* algorithm are based on this method.

Two main aspects characterize the new random generation of pyramids algorithm: (i) a total identification of all the couples with possible aggregation and (ii) an efficient management of the matrix that contains all these couples. The aim of point (ii) is to make quicker the search and elimination proceeding of the couples of classes to merge, in each stage of the pyramid construction.

On the *QuikRAP* algorithm we will use the same notation that was introduced in the *RAP* algorithm. The matrix M_p^i is defined with all the couples of classes potentially to be merge in the iteration i.

The main steps of the *QuikRAP* algorithm are described as follows:

Iteration 0: the number of terminal nodes, n, is introduced and are defined the intervening variables: M_p^0, CC, θ, P and M.

For i from 1 to the algorithm stopping, i.e., until $M_p^i = \emptyset$.
Iteration i: .

- random and uniform generation of the couple of classes to merge from M_p^i, $\left(c_1^i, c_2^i\right)$;
- aggregation of the couple of classes produced, $c^i = c_1^i \cup c_2^i$, and updating of: order θ (respecting the pyramidal structure), matrix CC, P and M;

- elimination in M_p^i of the couple $\left(c_1^i, c_2^i\right)$ and all the couples that aren't possible to be merged[4];
- deduction of the new possible classes to merge with c^i and matrix M_p^i update.

Output: order θ, matrix P associated to the pyramid and Robinson matrix $MR = M(\theta)$.

The *QuikRAP* algorithm already contemplates the free classes as being active to future aggregations and, consequently, these are not excluded in the construction of a pyramid. Therefore, the algorithm has in account not only the total inversion, but also the partial inversion of a connected component. Moreover, the algorithm allows a more effective administration of the pairs of classes joined, accelerating the searching process of the joining pairs of classes in each iteration, and the elimination process of those pairs not possible to be merged in future iterations.

5 Simulation and discussion

With the purpose of evaluating and comparing the performance of developed algorithms, theoretical and simulation studies were made. For a fixed terminal nodes number (n), the theoretical study gives: (i) the number of different topologic types of pyramids, (ii) the number of non isomorphic pyramids, (iii) the distribution of the non isomorphic pyramids for the different topologic types and (iv) the probability of obtaining a pyramid for each topologic type. In the particular case of dendrograms, for any number of terminal nodes, mathematical formulas providing these quantities are available. For pyramids, unfortunately, analogous formulas don't exist. In this work the theoretical study was accomplished, only for small values of n, using combinatory calculus tools. For $n = 4$ ($n = 3$ is a trivial case) 666 non isomorphic pyramids were counted and eight different topologic types were identified. Some of the obtained information is displayed in Table 1.

Topologies	I	II	III	IV	V	VI	VII	VIII
Probability	0,016	0,057	0,023	0,292	0,057	0,169	0,217	0,169

Table 1. Probability of obtaining a pyramid for each topologic type, $n = 4$.

The algorithms *RAP* and *QuikRAP* were used to simulate random pyramids and the correspondent topologic type was identified for each of them.

[4] The extension of definitions and necessary conditions for the understanding of all the couples of classes to exclude does not allow to us presented them here. For more details see Mfoumoune (1998) and Machado (2007).

Table 2 presents the observed frequencies for the eight topologies, with simulation of 100000 pyramids with 4 terminal nodes.

Topologies	I	II	III	IV	V	VI	VII	VIII
RAP algorithm	0,011	0,076	0,024	0,245	0,077	0,195	0,175	0,197
QuikRAP algorithm	0,016	0,057	0,023	0,291	0,059	0,168	0,218	0,168

Table 2. Observed frequencies, with the algorithms *RAP* and *QuikRAP*, for the eight topologies of pyramids, $n = 4$.

Analyzing the theoretical and simulation results, we conclude that, for the 4 terminal nodes pyramids, the *QuikRAP* algorithm achieves the correct proportions. Therefore, some of the *RAP* algorithm limitations were overcome and the new proposed algorithm - *QuikRAP* - seems to produce results in agreement with the expected.

References

BERTRAND, P. (1986): *Etude de la représentation pyramidale*. 3th Cycle Thesis, Université Paris IX-Dauphine.

DIDAY, E. (1984): Une représentation visuelle des classes empiétantes: les pyramides. Rapport de recherche I.N.R.I.A. n. 291, Rocquencourt, France.

FURNAS, G.W. (1984): The generation of random, binary unordered trees. *Journal of Classification 1, 187-233*.

LAPOINTE, F. and LEGENDRE, P. (1991): The generation of random ultrametric matrices representing dendrograms. *Journal of Classification 8, 177-200*.

MACHADO, V. (2007): *Geração aleatória de estruturas classificatórias*. Master Thesis, Faculdade de Engenharia, Universidade do Porto.

MACHADO, V. and SOUSA, F. (2007): A methodology to simulate random pyramids: some comments. *Proc. International Conference Probability and Statistics in Science and Technology, Bernoulli Society satellite meeting of ISI 2007*, Porto (Portugal), *17-18*.

MFOUMOUNE, E. (1998): *Les aspects algorithmiques de la Classification Ascendante Pyramidale et Incrémentale*. PhD Thesis, Université Paris IX-Dauphine.

PODANI, J. (2000): Simulation of random dendrograms and comparison tests: some comments. *Journal of Classification 17, 123-142*.

SOUSA, F. (2000): *Novas metodologias e validação em classificação hierárquica ascendente*. PhD Thesis, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

TENDEIRO, J. (2005) *Comparação de dendrogramas: obtenção de distribuições empíricas de alguns coeficientes*. Master Thesis, Faculdade de Engenharia, Universidade do Porto.

Optimized Clusters for Disaggregated Electricity Load Forecasting

Michel Misiti[1,2], Yves Misiti[1],
Georges Oppenheim[1,3], and Jean-Michel Poggi[1,4]

[1] Université Paris-Sud, Mathématique, Bât. 425, 91405 Orsay, France,
{*yves.misiti, georges.oppenheim, Jean-Michel.Poggi*}*@math.u-psud.fr*
[2] Ecole Centrale de Lyon, France, *michel.misiti@ec-lyon.fr*
[3] Université de Marne-la-Vallée, France
[4] Université Paris Descartes, France

Abstract. In order to take into account the variation of the EDF (the French electrical company) portfolio due to the liberalization of the electrical market, it is essential to conveniently disaggregate the global signal. The idea is to disaggregate the global load curve in such a way that the sum of disaggregated predictions improve significantly the prediction of the global signal considered as a whole. The strategy is to optimize with respect to a predictability index, a preliminary clustering of individual load curves. The optimized clustering scheme is directed by forecasting performance via a cross-prediction dissimilarity index and proceeds as a discrete gradient type algorithm.

Keywords: clustering, disaggregation, forecasting, optimization, wavelets

1 Introduction

The goal is to improve the accuracy of electrical load forecast and to allow to take into account the variation of the portfolio of EDF (the French electrical company) linked to the liberalization of the electrical market. One way to deal with this problem is to conveniently disaggregate the basic signal in order to improve the prediction performance. The problem is to find clusters so that the sum of disaggregated predictions improve significantly the prediction of the global signal considered as a whole. We propose an optimized clustering scheme directed by a cross-prediction dissimilarity index and based on a discrete gradient type algorithm.

This paper is organized as follows. After this short introduction, Section 2 is devoted to the problem and the data. Section 3 recalls briefly a wavelet-based procedure for clustering load curves. Then, in Section 4, we propose the optimized clustering for forecasting by disaggregation. Section 5 contains experimental results. Finally, Section 6 sketches some perspectives.

2 The data and the problem

2.1 The data

Individual power electricity demand curves along the time are available for 2309 industrial customers, during two years: 2000 and 2001. The sampling period is one hour, leading to 17520 samples.

To get an idea of the differences between the individual curves let us examine load curves for four different customers during the two years (see Figure 1).

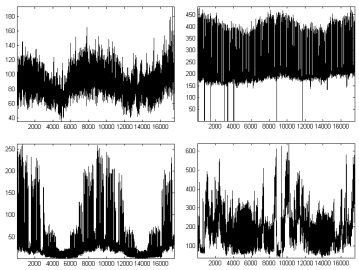


Fig. 1. Raw data: 4 customers load curves during years 2000 and 2001.

The long term shape can differ a lot from a customer to another: climate-free for customer at the bottom right and three different climatic sensitivity for the others. The load curves during one week of 2000 for the same 4 customers are displayed in Figure 2.

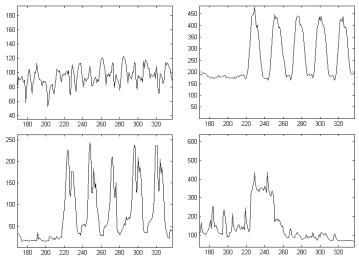


Fig. 2. Raw data: 4 customers load curves during one week of 2000.

The customers located on the main diagonal of the array of plots of the figure are truly different and the social rhythm is not at all visible at this time scale. On the contrary for the two other customers, the global shape is similar and the week-end is easy to detect while the difference is localized at the middle of the day: bimodal shape instead of a single bump.

2.2 Aggregated versus disaggregated

The problem of forecasting based on disaggregation is the following. Let us denote by $X_i(t)$ the value at time t of the load curve for the ith customer and consider the aggregated electricity consumption signal $S(t) = \sum X_i(t)$. Then the aggregated forecast is obtained by modeling and forecasting it leading to:

$$\widehat{S_{aggr}(t)} = \widehat{S(t)}. \tag{1}$$

Now, associated with any partition of individuals in clusters, we can define the consumption of each cluster g: $S_g(t) = \sum_{i \in g} X_i(t)$. Then, the disaggregated forecast is obtained by modeling and forecasting the signal within each cluster calculating $\widehat{S_g(t)}$ and then summing over the clusters leading to:

$$\widehat{S_{dis}(t)} = \sum_g \widehat{S_g(t)}. \tag{2}$$

The challenge is to find a partition of the individuals which is as accurate as possible from the forecasting perspective and whose performance is significantly better than the aggregated forecast. Of course a natural question is why such a result can be expected? We sketch two simple elements valid in narrow contexts which are useful suggestions.

A first indication is provided by the estimation of the mean μ of a variable Y on a given population using the random sample mean $\overline{Y}$. It is an unbiased estimator of variance $\sigma^2(Y)/n$ where n is the sample size. Using a stratified representative sampling with respect to a given partition, the variance of the associated stratified estimator (which is the disaggregated one) reduces to the within variance over n: $\sigma^2_{within}(Y)/n$, which is always smaller.

A second indication is a simple result stated for two clusters but true for more. Let $X_1(t)$ and $X_2(t)$ be two sequences of stationary square integrable random variables and define $S(t) = X_1(t) + X_2(t)$. Then denoting by $\widehat{Z(t)} = E(Z(t) \mid (Z(t-1), Z(t-2), \ldots))$ the conditional mean of $Z(t)$ by its own past, let us define the two error indices $Err_{aggr} = E(S(t) - \widehat{S(t)})^2$ and $Err_{dis} = E(S(t) - \widehat{X_1(t)} - \widehat{X_2(t)})^2$. So, if X_1 and X_2 are independent then $Err_{dis} \leq Err_{aggr}$. In other words, at least if the signals of different clusters are independent and if the conditional mean (or some corresponding accurate estimation) is used to predict, disaggregated forecast is of better quality.

So the two indications say that it could be useful to disaggregate the global signal in order to improve forecasting, and the idea is to find a tradeoff between the homogeneity within the clusters and the quality of the models estimation: the first increases with the number of clusters while the second decreases. So we propose a three steps strategy: 1) a preprocessing using wavelets; 2) a first clustering of the customers in numerous very homogeneous clusters and 3) an aggregation step using a stepwise optimization algorithm based on a dissimilarity index linked to a cross-prediction error and a discrete gradient type algorithm.

Before entering in more details, let us precise the basic forecasting model.

2.3 Eventail-like forecasting model

Handling recent statistical time series tools for load forecasting (see for example Hippert et al. (2001)) is out of the scope of this paper: we prefer for this work to restrict our attention on a single "black-box" method to design the prediction model starting from a given time series. More precisely, we use a fully automatic version of the EDF operational model called Eventail (see Bruhns et al. (2005)) which is designed to predict the aggregated electricity consumption. Daily, weekly and annual components for the endogenous variable are considered, together with exogenous variables: temperature, cloud cover, calendar events and a long-term trend. The middle-term model is a highly parameterized climate-free SARIMA model corrected additively by a weather dependent term, which delivers an accurate prediction.

As a reference, for the considered sample of 2309 customers, the forecasting performance, measured by the long-term MAPE (for Mean Absolute Percentage Error) is about 4.06% for the global aggregated signal. On the other hand, the performance of the completely disaggregated forecasting strategy given by the sum of the 2309 individual forecasts reaches 2.94%. So, we have to obtain less than these two reference values corresponding to the two extreme situations.

We emphasize that the error reduction provided by the new scheme is only due to clustering optimization since we do not perform any ad-hoc adaptation of the model design strategy to clusters.

3 Clustering using wavelets

The key idea of the preliminary step is to make profit of the hierarchical multiresolution structure of wavelet decomposition (see Misiti et al. (2007a)) for clustering signals. The procedure described in Misiti et al. (2007b) is an hybrid scheme mixing (following the terminology of James, Sugar (2003)) regularization and filtering approaches: individual denoising using a signal-adapted wavelet basis, projection on a single common wavelet basis to get a huge dimensionality reduction effect (see Biau et al. (2007)) and finally clustering the coefficients using the Ward method with squared Euclidean distance, in order to preserve distances between signals through wavelet coefficients encoding.

We generate hierarchies of partitions corresponding to various wavelet representations (typically approximations of decreasing resolution level) and different numbers of clusters. Then we choose the best one according to the normalized variance ratio index, similar to the statistic of Calinski, Harabasz (1974) (see also Tibshirani et al. (2001)):

$$I_Z^N(P) = \frac{Var_{between}(Z,P)}{C(P).Var_{within}(Z,P)}. \tag{3}$$

where $C(P)$ is the number of clusters of partition P. It allows to select a convenient number of clusters as well as a critical level of wavelet decomposition.

In our electrical context, we have in a first study obtained various partitions using this clustering scheme which does not take into account the forecasting objective. The most interesting partitions have about 15 to 19 clusters and highlight wavelet approximation coefficients at level 6 (about 2 coefficients per week) as well as detail coefficients at level 2 (about 5 coefficients per day). Let us remark that the forecasting performance reached by these partitions is about a 2.75% long-term MAPE. Even if it is better than the fully aggregated or the fully disaggregated forecasts, nevertheless, these partitions are essentially not improvable by the optimization process described below.

So, in the sequel we propose to retain this initial pre-processing step by selecting wavelet approximation coefficients at level 6 in order to capture the global shape of load curves but we propose to relax the unsupervised clusters constraints by starting from a larger number of clusters (we find, by using variance ratio again, that more than 90 guarantee strong homogeneity) and then aggregate these numerous clusters using an optimization trick supervised by predictability.

4 Optimized clustering directed by forecasting

4.1 A multistage procedure

The proposed optimized clustering scheme is the following:

(i) *Wavelet preprocessing.*
Wavelet representation of each signal after standardization: approximation coefficients at level 6 are used to characterize a customer;

(ii) *First clustering* around numerous centroids.
At least 90 clusters of very homogeneous customers. Each cluster is then represented by the corresponding aggregated signal;

(iii) *Iterative optimization.*
Starting from this initial partition, we perform an optimization process supervised by a cross-prediction dissimilarity index. A discrete gradient type procedure based on D matrix (defined in the next section) explores the set of partitions.

4.2 Cross-prediction dissimilarity

To quantify the interest of an aggregation we introduce a cross-prediction dissimilarity between two elements which can be individual or aggregated signals. This dissimilarity index between X_k and X_j is based on the following idea: to use the model fitted on the past observations of $X_j(t)$ to predict the

future of $X_k(t)$ and vice-versa. More precisely in our electrical context, let us denote by

$$forec_{k|j}^{2001} = forecast(X_j^{2000}, X_k^{2001}), \tag{4}$$

the forecasts of X_k on the year 2001 (the test period) obtained from the model fitted (using the previously mentioned Eventail-like design tool) on X_j on the year 2000 (the learning period). Then, the associated error is defined by:

$$E_{k|j} = error(X_k^{2001}, forec_{k|j}^{2001}), \tag{5}$$

and then a natural symmetric measure of dissimilarity is:

$$D = (D_{j,k}) = ((E_{k|j} + E_{k|j})/2). \tag{6}$$

4.3 Zooming in on the optimization step

The iterative optimization of the initial partition is supervised by the cross-prediction dissimilarity and can be adapted to the prediction horizon and to the error criterion. The principle is a discrete gradient via a neighborhood definition through a dissimilarity between an element and a cluster induced by the matrix D. The basic step is an iterative exploration of elements, which are always candidates for cluster change, using nearest D-neighbors. It should be noted that the partition evolves and that the basic step consists in changing an element from a cluster to another one. So, this process generates a non monotonic sequence of partitions (this is not a hierarchical approach), evolving by modification of element assignments. The number of clusters decreases slowly along the iterations, and a cluster disappears only if it becomes empty. The optimization scheme is as follows:

(i) *Compute matrix* D of dissimilarities between elements;
(ii) *Compute dissimilarities* between each element and the current clusters using D and a linkage function (the minimum for example);
(iii) *Select* a neighbor: a couple (E,C), an element E candidate to move to a cluster C;
(iv) *Test* the gain of the affectation for the disaggregated prediction associated with the resulting partition
- *if* the error does not decrease *then*
  - *if* there are candidates *then* select the next one and *go to* step 4
  - *else end* (no improvement by moving an element from a cluster to another)
- *if* the error decreases *then* modify partition and *go to* step 2

5 Experimental results

Starting from 90 clusters, the optimized partition reaches the performances measured by long-term and short-term MAPE, given by Table 1.

	Aggregated	Disaggregated	Gain
MAPE long-term (LT)	4.06%	2.39% with 19 clusters	41.13%
MAPE short-term (ST)	2.47%	1.51% with 28 clusters	38.86%

Table 1. Performances of optimized partition starting from 90 clusters.

The proposed procedure is anytime: it can be stopped at any step of the optimization process, delivering an admissible solution which improves the previous one. The process of 195 steps leading to an error rate gain of 41%, is illustrated by Figure 3, starting from 90 clusters and ending with 19. The error reduction of the optimization stage is huge and the benefits are obvious.

The first step of the global procedure (wavelet preprocessing and initial clustering using wavelets) is useful. Indeed if one performs directly a hierarchical clustering of the original 2309 customers using the dissimilarity matrix D and then optimizes the associated 90 clusters partition, the MAPE-LT error criterion stabilizes around 2.7% instead of 2.4%.

The optimization step is useful. Indeed, starting from the 90 clusters partition, if one constructs the hierarchy of partitions (by hierarchical clustering using D), it is difficult to select a critical number of clusters and the MAPE-LT error criterion remains about 2.6% instead of 2.4%.

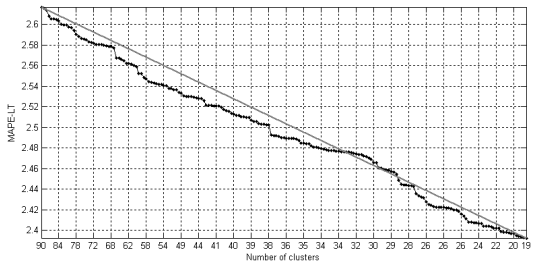


Fig. 3. Optimization process: from 90 to 19 clusters leading to a gain of 41%.

Finally, let us mention that the number of initial clusters (taken here to be equal to 90) is a parameter which could be important, especially when the method will be used for a significantly larger number of customers. Indeed, on the actual data set the performance is slightly improved by increasing the number of clusters: the initial 2.39% performance for 90 becomes 2.31% for 200 and even 2.26% for 500 increasing the reduction rate from 41.1 to 44.3%.

6 Future work

Let us sketch some future applied developments. The first direction is to evaluate the method using a more representative set of customers including

more information. The second direction proposes to experiment on these kind of electrical data, forecasting methods using wavelets, see Antoniadis et al. (2006), Amin Ghafari, Poggi (2007) and to adapt the models to the clusters mimicking the approach of Hathaway, Bezdek (1993). In addition to this last topic, a useful idea is to make profit of external information (meteorological and economical) for interpretation and performance improvement. More generally, the global procedure should integrate data-driven choices of several parameters: the wavelet and the representation basis, the obtained partition and the adaptation of the model to cluster specificities.

Moreover, an interesting topic will be to study theoretically the conditions to maximize the benefits of disaggregation in more general contexts.

Acknowledgements

This work is part of a scientific collaboration between EDF Clamart R&D, OSIRIS department and Orsay University and the authors thank EDF for the problem and the data, and Alain Dessertaine for helpful discussions.

References

ANTONIADIS, A., PAPARODITIS, E., SAPATINAS, T. (2006): A functional wavelet-kernel approach for time series prediction. *J. of the Royal Stat. Soc., Series B, 68, 837–857.*

AMIN GHAFARI, M., POGGI, J.M. (2007): Forecasting time series using wavelets, *Int. Journ. of Wavelets, Multiresolution and Inf. Proc., 5(5), 709–724.*

BIAU, G., DEVROYE, L. and LUGOSI, G. (2007): On the performance of clustering in Hilbert spaces, *IEEE Trans. on Inf. Theory, in press.*

BRUHNS, A., DEURVEILHER, G., ROY, J.S. (2005): A non linear regression model for mid-term load forecasting and improvements in seasonality, *Proceedings of the 15th Power Systems Computation Conference 2005, Liege Belgium.*

CALINSKI, R.B. and HARABASZ, J. (1974): A dendrite method for cluster analysis. *Comm. Stat., 3, 1-27.*

HATHAWAY, R.J., BEZDEK J.C. (1993): Switching regression models and fuzzy clustering, *IEEE Trans. Fuzzy Systems 1 (3), 195–203.*

HIPPERT, H.S. PEDREIRA, C.E., SOUZA, R.C. (2001): Neural networks for short-term load forecasting: A review and evaluation, *IEEE Transactions on Power Systems, 16(1), 44–55.*

MISITI, M., MISITI, Y., OPPENHEIM, G., POGGI, J.-M. (2007a): *Wavelets and their applications.* Hermes Lavoisier, ISTE Publishing Knowledge.

MISITI, M., MISITI, Y., OPPENHEIM, G., POGGI, J.-M. (2007b): Clustering signals using wavelets, F. Sandoval et al. (Eds.): *IWANN 2007*, Lecture Notes in Computer Science, 4507, 514–521, Springer.

JAMES, G., SUGAR, C. (2003): Clustering for sparsely sampled functional data. *JASA, 98, 397–408.*

TIBSHIRANI, R., WALTHER, G., HASTIE, T. (2001): Estimating the number of clusters in a data set via the gap statistic. *J. of the Royal Stat. Soc., Series B 63 (2), 411-423.*

A Spectral Analysis Approach for Univariate Gaussian Mixture Estimation

Nicolas Paul, Michel Terre, and Luc Fety

CNAM, Laboratoire Électronique et Communication
292 rue Saint-Martin, 75003 Paris, France, *nicolas.paul@cnam.fr*

Abstract. This paper deals with the estimation of one-dimensional Gaussian mixture. Given a set of observations of a K-component Gaussian mixture, we focus on the estimation of the component expectations. The number of components is supposed to be known. Our method is based on a spectral analysis of the estimated first characteristic function. We construct a Toeplitz matrix $\mathbf{R}_M$ with M ($M > K$) estimated samples of the first characteristic function and show that the mixture component expectations can be derived from the eigenvector decomposition of $\mathbf{R}_M$. Simulations illustrate the performance of our algorithm on several configurations of a six-component Gaussian mixture. In the investigated scenarios the proposed method outperforms the Expectation-Maximization algorithm.

Keywords: Gaussian mixture estimation, spectral analysis

1 Introduction

In this paper we deal with Gaussian mixture estimation. Given a set of one-dimensional observations originating from K possible Gaussian components, we focus on the estimation of the component expectations. The number of components is supposed to be known, and the component expectations are supposed to be all different.

One method consists in estimating a sampling of the observations probability density function (pdf), a mixture of K pdf, by associating a kernel to each observation and adding the contribution of all the kernels (Parzen 1962). If the mixture components do not strongly overlap, a search of the pdf modes then leads to the component expectations. The drawback of such a method is that it requires the selection of extra-parameters (kernel design, sampling intervals). Furthermore, the final mode search algorithm might fail because of spurious local maxima in the estimated pdf.

An alternative method consists in using the Expectation-Maximization (EM) algorithm (Dempster 1977). Each EM iteration consists of two steps. The Expectation step estimates the probability for each observation to come from each mixture component. Then, during the Maximization step, these estimated probabilities are used to update the estimation of the mixture parameters. This procedure converges to any stationary point of the log-likelihood. The convergence to the global maximum of the log-likelihood is not

guaranteed. Some solutions consist, for instance, in using smart initializations or stochastic optimization (McLachlan 2000).

In this contribution we propose a new approach based on a spectral analysis of the first characteristic function (CF). We define a Toeplitz matrix $\mathbf{R}_M$ with M ($M > K$) estimated samples of the CF and show that the mixture component expectations can be estimated from an eigenvector decomposition of $\mathbf{R}_M$. The proposed method is strongly inspired from the *mu*ltiple *si*gnal *c*lassification algorithm MUSIC (Schmidt 1981) which aims at estimating the frequencies in a sum of sinusoids. The paper is organized as follow: In section 2 the observation model is presented and an analytical expression of the CF of a Gaussian mixture is given. In section 3 the matrix $\mathbf{R}_M$ is defined and some properties of $\mathbf{R}_M$ are described. Section 4 then presents the complete estimation algorithm. Section 5 illustrates the estimation performances on a six-component Gaussian mixture with different configurations. Conclusions are finally given in section 6, as well as perspectives for using the proposed method to estimate the number of components in a mixture.

2 Gaussian mixture

2.1 Probability density function (pdf)

Let $\{p_k\}_{k\in\{1\cdots K\}}$ be a set of K positive mixing weights that sum up to one. The multimodal pdf of the random observable variable Z is a finite mixture given by:

$$f_Z(z) := \sum_{k=1}^{K} p_k g(z, a_k, \sigma_k), \tag{1}$$

where $g(z, a_k, \sigma_k)$ is the Gaussian pdf given by:

$$g(z, a_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(z-a_k)^2}{2\sigma_k^2}\right)$$

and a_k and σ_k are respectively the expectation and the standard deviation of component k. Given a set of N observed realizations $\{z_n\}_{n\in\{1,\cdots,N\}}$ of Z we focus on the estimation of the K component expectations $\{a_k\}_{k\in\{1,\cdots,K\}}$. Our proposal is mainly based on the estimated first characteristic function (CF) of the mixture.

2.2 First characteristic function (CF)

In general, the CF of a random variable X is defined by:

$$\phi_X(t) := E_X\{e^{itX}\}, \qquad t \in \mathbb{R},$$

where $E_X\{\cdot\}$ is the mathematical expectation with respect to the pdf of X. For instance, the CF of a Gaussian random variable X with pdf $g(x, a, \sigma)$ is given by:

$$\begin{aligned}\phi_X(t) &= \int_{x=-\infty}^{\infty} e^{itx} g(x, a, \sigma) \mathrm{d}x \\ &= e^{-\frac{\sigma^2 t^2}{2}} e^{iat}.\end{aligned}$$

Consequently the CF of the random variable Z with the pdf described in (1) is given by:

$$\begin{aligned}\phi_Z(t) &= \sum_{k=1}^{K} p_k \int_{z=-\infty}^{\infty} e^{itz} g(z, a_k, \sigma_k) \mathrm{d}z \\ &= \sum_{k=1}^{K} p_k e^{\frac{-\sigma_k^2 t^2}{2}} e^{ita_k}. \end{aligned} \tag{2}$$

Now let ϕ_m be the sampled version of $\phi_Z(t)$ with a sampling period T_e. According to (2), we have:

$$\begin{aligned}\phi_m &:= \phi_Z(mT_e), \qquad m \in \mathbb{Z} \\ &= \sum_{k=1}^{K} p_k \alpha_{k,m} w_k^m, \end{aligned} \tag{3}$$

where w_k and $\alpha_{k,m}$ are defined by:

$$w_k := e^{ia_k T_e} \tag{4}$$

$$\alpha_{k,m} := e^{-\frac{\sigma_k^2 (mT_e)^2}{2}}. \tag{5}$$

In practical situation, ϕ_m can be estimated from a set of N observations $\{z_n\}_{n=1,\cdots,N}$ using:

$$\hat{\phi}_m = \frac{1}{N} \sum_{n=1}^{N} e^{iz_n (mTe)} \tag{6}$$

In section 3 we will show how the $\{w_k\}_{k=1,\cdots,K}$ defined in (4) can be estimated from the sampled CF. Once the $\{w_k\}_{k=1,\cdots,K}$ are estimated, the $\{a_k\}_{k=1,\cdots,K}$ can be obtained without ambiguity if the sampling period T_e is less than $\frac{2\pi}{\max\{z_n\} - \min\{z_n\}}$: If we for instance choose:

$$T_e = \frac{2\pi}{2(\max\{z_n\} - \min\{z_n\})}, \tag{7}$$

then $\frac{2\pi}{T_e} = 2(\max\{z_n\} - \min\{z_n\})$. Since $\min\{z_n\} \leq a_k \leq \max\{z_n\}$ there is exactly one integer l_{w_k} such as:

$$\frac{\mathrm{angle}(w_k)}{T_e} + l_{w_k} \frac{2\pi}{T_e} \in [\min\{z_n\} \quad \max\{z_n\}],$$

and we have:

$$a_k = \frac{\text{angle}(w_k)}{T_e} + l_{w_k}\frac{2\pi}{T_e}, \qquad k = 1, \cdots, K. \tag{8}$$

Furthermore, if T_e verifies (7), and since the $\{a_k\}_{k=1,\cdots,K}$ are supposed to be all different, then the $\{w_k\}_{k=1,\cdots,K}$ are also all different. This will be used in section 3.

3 Definition and properties of $\mathbf{R}_M$

Let $\mathbf{R}_M := (r_{jl})_{j,l=1,\cdots,M} \in \mathbb{C}^{M\times M}$ be the Toeplitz matrix with the following elements:

$$r_{jl} := \phi_{l-j}, \qquad j, l = 1, \cdots, M \tag{9}$$

where ϕ_m has been defined in (3). Note that $\phi_{-m} = \phi_m^*$ so $\mathbf{R}_M$ is a Hermitian matrix and one only has to compute M samples of the CF to build $\mathbf{R}_M$. Including (3) into (9):

$$\begin{aligned} r_{jl} &= \sum_{k=1}^{K} p_k \alpha_{k,l-j} w_k^{l-j} \\ &= \sum_{k=1}^{K} p_k w_k^{l-j} + \sum_{k=1}^{K} p_k(\alpha_{k,l-j} - 1) w_k^{l-j} \\ &= \sum_{k=1}^{K} w_k^{*j-1} p_k w_k^{l-1} + \sum_{k=1}^{K} p_k(\alpha_{k,l-j} - 1) w_k^{l-j}, \end{aligned} \tag{10}$$

where in (10) we used that $w_k^{l-j} = w_k^{*j} w_k^l = w_k^{*j-1} w_k^{l-1}$ since $w_k^{*-1} w_k^{-1} = 1$. A consequence of (10) is that $\mathbf{R}_M$ can be expressed as the sum of a "signal" matrix $\mathbf{S}_M$ and a "perturbation" matrix $\mathbf{P}_M$:

$$\mathbf{R}_M = \mathbf{S}_M + \mathbf{P}_M, \tag{11}$$

where the "signal" matrix $\mathbf{S}_M$ is given by:

$$\mathbf{S}_M := (\mathbf{w}_1, \cdots, \mathbf{w}_K)\mathbf{D}\begin{pmatrix} \mathbf{w}_1^H \\ \vdots \\ \mathbf{w}_K^H \end{pmatrix} \in \mathbb{C}^{M\times M}, \tag{12}$$

$$\mathbf{w}_k := (1, w_k^1, \cdots, w_k^{M-1})^H \in \mathbb{C}^M, \tag{13}$$

$$\mathbf{D} := \text{diag}(p_1, \cdots, p_K) \in \mathbb{R}^{K\times K}, \tag{14}$$

and the "perturbation" matrix $\mathbf{P}_M$ is given by:

$$\mathbf{P}_M := \sum_{k=1}^{K} p_k \left((\alpha_{k,l-j} - 1) w_k^{l-j}\right)_{l,j=1,\cdots,M} \in \mathbb{C}^{M\times M}. \tag{15}$$

The "signal" matrix $\mathbf{S}_M$ is a well-known matrix in the spectral analysis community. It is the auto-correlation matrix of a received sum of K sinusoids with angular frequencies a_k and power p_k. High resolution algorithm such as MUSIC (Schmidt 1981) estimate $\mathbf{S}_M$ from some (potentially corrupted) signal samples then estimate the sinusoid frequencies from its eigenvector decomposition. Indeed, since the w_k are all different (section 2.2), one can show that the rank of $\mathbf{S}_M$ is equal to K and that the signal vectors $\mathbf{w}_k$ defined in (13) are orthogonal to any vector of the kernel of $\mathbf{S}_M$ (Schmidt 1981). Consequently, if $\mathbf{V} := (\mathbf{v}_{K+1}, \cdots, \mathbf{v}_M) \in \mathbb{C}^{M \times M-K}$ contains $M-K$ orthogonal eigenvectors belonging to $\mathrm{Ker}\{\mathbf{S}_M\}$ we have:

$$\mathbf{w}_k^H \mathbf{V}\mathbf{V}^H \mathbf{w}_k = 0, \qquad k = 1, \cdots, K. \tag{16}$$

A consequence of (16) is that if t_j denotes the sum of the jth diagonal of $\mathbf{V}\mathbf{V}^H$ ($j \in \{-M+1, \cdots, M-1\}$ and $t_0 = \mathrm{trace}\{\mathbf{V}\mathbf{V}^H\}$) and if $q(y)$ is the polynomial defined by:

$$q(y) := \sum_{j=-M+1}^{M-1} t_{-j} y^j, \qquad y \in \mathbb{C}, \tag{17}$$

then the zeros of $q(y)$ exhibit inverse symmetry with respect to the unit circle, and $q(y)$ exactly has K zeros on the unit circle, equal to $\{w_k\}_{k=1,\cdots,K}$ (see for instance Haykin (1991) for a detailed description of the "root-MUSIC" algorithm).

In our Gaussian mixture estimation case, the "signal" matrix $\mathbf{S}_M$ (12) is corrupted with the "perturbation" matrix $\mathbf{P}_M$ (15). When all the component variances tend to zero (ideal case) the perturbation matrix $\mathbf{P}_M$ tends to a null matrix: using (5) and (15) we have:

$$\lim_{\sigma_k \to 0} \alpha_{k,l-j} = 1, \qquad k = 1, \cdots, K$$
$$\lim_{(\sigma_1, \cdots, \sigma_K) \to (0, \cdots, 0)} \mathbf{P}_M = \mathbf{0}_{M \times M}.$$

Yet, in the general case, $\mathbf{P}_M$ is not null and unfortunately depends on w_k.

4 Estimation algorithm

The proposed algorithm for estimating the set of component expectations is based on the eigenvector decomposition of the estimation of $\mathbf{R}_M$ (9), thus neglecting the effect of the perturbation matrix $\mathbf{P}_M$ (15). Given a set of N observations $\{z_n\}_{n \in \{1, \cdots N\}}$ the algorithm steps are the following:

(i) define a sampling period T_e using (7)
(ii) estimate the M[1] first samples ($M > K$) of the CF using (6)

[1] $M = 2K$ seems to be a good choice from our simulations but more investigations are needed to optimize the value of M. Note that the optimal size of $\mathbf{R}_M$ depends on the number of components but does not depend on the number of observations.

(iii) build the Hermitian Toeplitz matrix $\mathbf{R}_M$ using (9)
(iv) perform a eigenvector decomposition of $\mathbf{R}_M$
(v) construct the matrix $\mathbf{V} = (\mathbf{v}_{K+1}, \cdots, \mathbf{v}_M)$ with the $M-K$ eigenvectors associated to the $M-K$ smallest eigenvalues of $\mathbf{R}_M$
(vi) calculate the coefficient of $q(y)$ defined in (17)
(vii) calculate the roots of $q(y)$. These roots exhibit inverse symmetry with respect to the unit circle. Keep the roots inside the unit circle then identify the K roots that are closest to the unit circle, call them $\{\hat{w}_k\}_{k=1,\cdots,K}$
(viii) derive $\{\hat{a}_k\}_{k=1,\cdots,K}$ from $\{\hat{w}_k\}_{k=1,\cdots,K}$ using (8)

5 Simulation

In our simulations several types of a six-component Gaussian mixture have first been considered. The set of expectations is equal to $(0, 1, 2, 4, 5, 6)$, with a difference of one or two between two successive component expectations. Four cases have been studied: common variance and common weight (scenario 1), different variances and common weight (scenario 2), common variance and different weights (scenario 3) and different variances and different weights (scenario 4). A summary of the scenarios is given in Table 1. The parameter σ in Table 1 enables to simulate different overlapping situation. The number N of observations per simulation run is 200. The $\mathbf{R}_M$-based algorithm has been run as described in section 4 with $M = 2K$(see footnote 1). This algorithm has been compared to the EM algorithm (Dempster (1977)) with a uniform random start and a maximal number of 100 iterations. See (McLachlan (2000)) for a detailed description of the Gaussian mixture estimation with EM. A constrained version of the EM (EM_c) which imposes a common variance and a common mixing weight has been used to prevent the convergence to components with an almost null variance. In all the scenario, EM_c provides better estimates than the standard EM, even in the scenario where the component variances or the component weights are different. Therefore only the performances of EM_c are presented here. To get rid of the permutation

	scenario 1		scenario 2		scenario 3		scenario 4	
mean	var.	weight	var.	weight	var.	weight	var.	weight
0	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.2	σ^2	0.2
1	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.2	$\frac{\sigma^2}{2}$	0.2
2	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.1	σ^2	0.1
4	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.2	$\frac{\sigma^2}{2}$	0.2
5	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.2	σ^2	0.2
6	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.1	$\frac{\sigma^2}{2}$	0.1

Table 1. Means, variances, weights of the simulated mixture.

ambiguity, the estimation performance is evaluated as follows: If $\mathbf{a} \in \mathbb{R}^K$ is the vector of the true component expectations and $\hat{\mathbf{a}}_r \in \mathbb{R}^K$ is the vector of the estimated component expectations at simulation run r, the performance criterion e_r is defined as the maximal absolute distance between the true and estimated ordered vector of component expectations:

$$e_r := \left\| \text{sort}(\mathbf{a}) - \text{sort}(\hat{\mathbf{a}}_r) \right\|_\infty ,$$

where $\text{sort}(\mathbf{x})$ is the ordered permutation of $\mathbf{x}$ and $\|\cdot\|_\infty$ is the infinity norm in $\mathbb{R}^K$.

The simulation results are presented in Figure 1 for different values of σ. When σ is greater than 0, there is a risk that the constrained EM converges to a wrong set of estimated component expectations. Typically one estimated component expectation is located in the middle of two true component expectations. For instance, for $\sigma = 0.1$, EM_c provides a wrong set of estimates ($e_r > 0.5$) for 60% of the run and a good set of estimates ($e_r < 0.1$) for 40% of the run. Concerning the $\mathbf{R}_M$-based algorithm, the higher σ is, the higher the influence of the pertubation matrix $\mathbf{P}_M$ (15) is. Yet, in all the investigated scenario and for all the values of σ, the proposed method outperforms the EM_c algorithm. For instance the $\mathbf{R}_M$-based algorithm always provide a good set of estimates ($e_r < 0.1$) when $\sigma = 0.1$ and e_r is never greater than 0.5 when $\sigma \leq 0.3$.

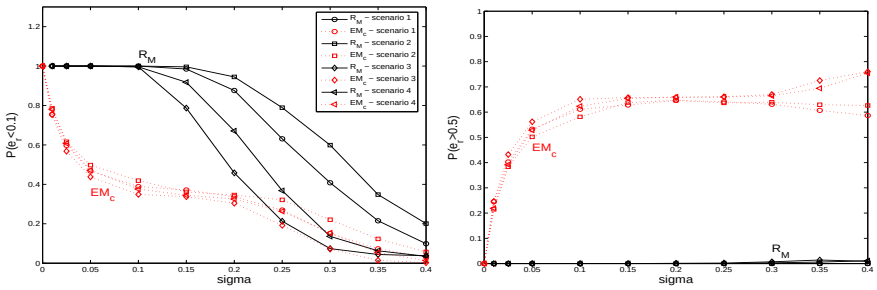


Fig. 1. Performances of the constrained EM (EM_c, dotted lines) and $\mathbf{R}_M$ based algorithm with $M = 2K = 12$ (full lines) for different values of σ. For each value of σ and for each scenario 10000 simulation runs have been performed. The performance criteria are the probabilities for e_r to be smaller than 0.1 (left) and to be greater than 0.5 (right).

6 Conclusion

Given a set of observations originating from a K-component univariate mixture, we focused on the estimation of the component expectations when the number K of components is known. We proposed a method based on the eigenvector decomposition of a Toeplitz matrix $\mathbf{R}_M$ built from some estimated samples of the first characteristic function. Simulations illustrated

the superiority of the proposed method compared with the Expectation-Maximization algorithm on various configurations of a six-component Gaussian mixture. More theoretical investigations are now needed to study the influence of the perturbation matrix (15) on the performances. Our current research also deals with the case of an unknown number of components. In figure 2 we plot the eigenvalues of $\mathbf{R}_M$ with $M = 10$ obtained in scenario 4 of Table 1 (where the mixture components have different weights and variances) with $N = 200$ observations and $\sigma = 0.15$. One can see that $K = 6$ eigenvalues are clearly greater than 0 while the $M - K = 4$ other eigenvalues are almost null. In general, one can therefore expect the eigenvalue decomposition of the matrix $\mathbf{R}_M$ to provide relevant information on the number of components in an observed mixture.

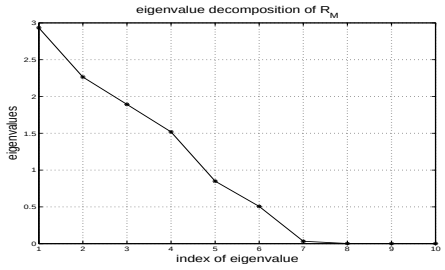


Fig. 2. eigenvalue decomposition of the matrix $\mathbf{R}_M$ with $M = 10$ in scenario 4 (6 mixture components) for $\sigma = 0.15$.

Acknowledgements

This work is supported by France-Telecom under External Research Contract number 16-132-234.

References

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B 39:1-38.*

HAYKIN, S. (1991): *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, New Jersey.

MACLACHLAN, G. and PEEL, D. (2000): *Finite Mixture Models*, John Wiley and Sons, New-York.

PARZEN, E. (1962): On estimation of a probability density function and mode. *Annals of Mathematical Statistics Vol. 33, pp. 1065-1076*

SCHMIDT, R.O. (1981): *A signal subspace approach to multiple emitter location and spectral estimation* Ph.D. thesis, Stanford University, Standford, CA.

A Simple Algorithm to Recognize Robinsonian Dissimilarities

Morgan Seston

Laboratoire d'Informatique Fondamentale
Faculté des Sciences de Luminy, Université de la Méditerranée
13288 Marseille cedex 9, France, *morgan.seston@lif.univ-mrs.fr*

Abstract. In this paper, we present a $O(n^3)$-time algorithm to recognize if a dissimilarity defined on n objects is Robinsonian. A dissimilarity d is Robinsonian if there exists a total order $\preceq$ such that $x \preceq y \preceq z$ implies $d(x,z) \geq \max\{d(x,y), d(y,z)\}$. Moreover, our algorithm provides all such orders with a polynomial representation (PQ-trees).

Keywords: Algorithm, Recognition, Robinson, PQ-tree

1 Introduction

The classical seriation problem consists in finding a simultaneous ordering (or permutation) of the rows and the columns of the dissimilarity matrix d on a finite set X with the objective of revealing an underlying one-dimensional structure (d is a dissimilarity if $d(x,y) = d(y,x) \geq 0$ and $d(x,y) = 0$ iff $x = y$). The basic idea is that small values should be concentrated around the main diagonal as closely as possible, whereas large values should fall as far from it as possible. This goal is best achieved by considering the so-called *Robinson property* (Robinson (1956)): a dissimilarity matrix d on a finite set X is Robinsonian if there exists a total order $\preceq$ on X such that $x \preceq y \preceq z$ implies $d(x,z) \geq \max\{d(x,y), d(y,z)\}$. Such an order is said *compatible*. Robinsonian dissimilarities are of importance in various domains as DNA analysis, overlapping clustering or archaeology. Additionally, there exists a one-to-one correspondence between Robinsonian dissimilarities and weakly indexed pseudo-hierarchies (Batbedat (1990); Bertrand (1995); Critchley, Fichet (1996)). For a dissimilarity d, deciding if it admits a compatible order is an interesting problem of seriation. Mirkin and Rodin (1984) presented an $O(n^4)$-time algorithm for this problem using the linear algorithm of Booth and Lueker (1976) for recognition of interval graph. Further, Chepoi and Fichet (1997) gave an $O(n^3)$-time algorithm based on a "divide and conquer" strategy. In this paper, we present another algorithm with same complexity but in a simpler way, using known tools of combinatorics like connected components and PQ-trees. Moreover, our algorithm finds all compatible orders with a polynomial representation (PQ-trees).

2 Preliminaries

Let d be a dissimilarity defined on the set X of objects. Recall that a dissimilarity d is a symmetric function from X^2 to the nonnegative real numbers and vanishing on the main diagonal, i.e. $d(x, y) = d(y, x) \geq 0$ and $d(x, y) = 0$ iff $x = y$.

For a subset $A \subseteq X$, we denote by $\delta(A) = \max\{d(u, v) : u, v \in A\}$ the diameter of A. Let $\preceq$ be a quasi-order on X, i.e. a reflexive, transitive and linear binary relation (any binary relation, considered in this paper, will be total). We note $x \prec y$ if $x \preceq y$ and not $y \preceq x$. Note that any quasi-order can be represented as an ordered partition $(B_1, B_2, \ldots, B_m)$, where $x \preceq y$ if and only if $x \in B_i, y \in B_j$ and $i \leq j$. Thus, $B_1, B_2, \ldots, B_m$ are called *blocks*. A quasi-order $\preceq_2$ refines a quasi-order $\preceq_1$, if $x \preceq_2 y$ implies $x \preceq_1 y$.

A quasi-order $\preceq$ on X is said *compatible* with a dissimilarity d if for all $x, y, z \in X$, $x \preceq y \prec z$ implies $d(x, z) \geq d(x, y)$ and $x \prec y \preceq z$ implies $d(x, z) \geq d(y, z)$. A dissimilarity d is *Robinsonian* if there is an order compatible with d.

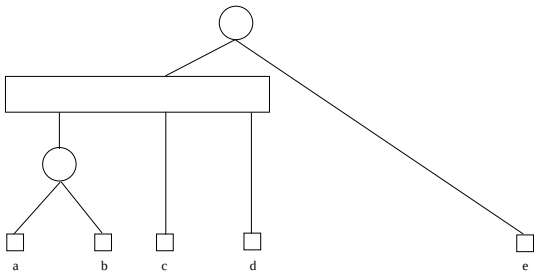


Fig. 1. An example of PQ-tree. As usually, P-nodes and Q-nodes are labeled by a circle and a rectangle, respectively.

In this paper, we also use *PQ-trees*, a structure introduced by Booth and Lueker to recognize interval graphs. This structure permits to represent an exponential set of orders with a polynomial storage. A PQ-tree T defined on a set X is a rooted tree which has the elements of X as leaves. For a node a, let $L(a)$ be the leaves (the elements of X) which are descendant of a. Additionally, a PQ-tree has two kinds of internal nodes: P-nodes, Q-nodes. The children of a Q-node are ordered. Thus a Q-node induces a quasi-order on its descendants whose ordered blocks are the sets induced by its children. We denote by $Q(n_1, n_2, \ldots, n_m)$ and $P\{n_1, n_2, \ldots, n_m\}$, a Q-node and a P-node, respectively, where $n_1, \ldots, n_m$ are the children of those nodes. Using this notation, a Q-node $Q(n_1, n_2, \ldots, n_m)$ induces the quasi-order $L(n_1) \preceq L(n_2) \preceq \ldots \preceq L(n_m)$. Now, we can define *the orders of a PQ-tree T*. An order $\leq$ on X is an order of T if it fulfills the following two properties: (i) For every node a, $L(a)$ is an interval of $\leq$; (ii) For every Q-node q, the restriction

of $\leq$ on $L(q)$ refines the quasi-order induced by q or its converse order. For example, the orders of the PQ-tree $P\{Q(P\{a,b\},c,d),e\}$ (represented by the Figure 1) are: $abcde, bacde, dcbae, dcabe$, up to the converse order.

3 The algorithm

Our algorithm is based on a "divide and conquer" strategy. We want to divide the set X in subsets (ordered or not) $B_1, B_2, \ldots, B_m$ having the following property: if d is Robinsonian then each B_i is an interval of any compatible order, and if $(B_1, B_2, \ldots, B_m)$ are ordered then any compatible order refines this quasi-order or its converse. We also requires that the sets B_i are *independent*: two distinct sets U, V are independent, if for all $u, u' \in U$ and for all $v, v' \in V$, we have $d(u,v) = d(u,v') = d(u',v') = d(u',v)$. Now, let $\leq_i$ be an order compatible with d restricted on B_i and let $\preceq$ be a compatible quasi-order whose blocks are $B_1, B_2, \ldots, B_m$. Then, the order refining $\preceq$ and each $\leq_i$, is compatible with d. Thus, we recursively apply the algorithm on each B_i to obtain a PQ-tree T_i whose orders are the orders compatible with d restricted on B_i. Finally, we return a PQ-tree whose root is a P-node or Q-node (depending on whether the blocks $B_1, B_2, \ldots, B_m$ are ordered or not) with the roots of $T_1, T_2, \ldots, T_m$ as children. If a block B_i does not admit a compatible order then d is not Robinsonian.

Let d be a dissimilarity defined on X. Now, we describe how we partition X in blocks $B_1, B_2, \ldots, B_m$. To this end, we construct the graph $G = (X, E)$, where $E = \{xy : d(x,y) < \delta(X)\}$. We will distinguish two cases: G is not connected and G is connected.

First, suppose that G is not connected and let $B_1, B_2, \ldots, B_m$ be the connected components of G. By definition of G, its connected components are independent. So, we can recursively call our algorithm on each connected component B_i and we denote by $T(B_i)$ the obtained PQ-tree. Finally, we returned the PQ-tree which has the P-node $P\{T(B_1), T(B_2), \ldots, T(B_m)\}$ as root.

Lemma 4. *Suppose that the restrictions of d on each B_i are Robinsonian. For every B_i, let $T'(B_i)$ be a PQ-tree such that the orders of $T'(B_i)$ are exactly the orders compatible with d restricted on B_i. Let T be the PQ-tree which has the P-node $P\{T'(B_1), T'(B_2), \ldots, T'(B_m)\}$ as root. Then the orders of T are exactly the orders compatible with d.*

Proof. First, we have to show that every order $\leq$ compatible with d is an order of T. We assert that each B_i is an interval of $\leq$. Let x, y be two elements of B_i, and z be an element of B_j with $i \neq j$. By contradiction, suppose that $x < z < y$. Since x and y belong to the same connected component, there exists $x', y' \in B_i$ such that $x'y'$ belongs to E and $x' < z < y'$. As z does not belong to B_i, $x'z$ and $y'z$ do not belong to E. That implies $d(x', y') < \delta(X) = d(x', z) = d(y', z)$ and contradicts that the order $\leq$ is

compatible with d. Moreover, the restriction of $\leq$ on each B_i is an order of $T'(B_i)$, and any permutation of the intervals $B_1, B_2, \ldots, B_m$ yields a quasi-order which is refined by an order of T. Thus $\leq$ is an order of T.

Now, we have to show that every order of T is compatible with d. Let $\leq$ be an order of T and x, y and z be three elements of X such that $x < y < z$. Note that if x and y belong to different connected components then x and z belong to different connected components. If x, y and z belong to a common component B_i, we have $\max\{d(x,y), d(y,z)\} \leq d(x,z)$ by definition of $T'(B_i)$. If x, y belong to the same component B_i and z does not belong to B_i (the case $x \in B_i$ and $y, z \in B_j$ with $i \neq j$ is similar), we have $d(x,y) \leq \delta(X) = d(y,z) = d(x,z)$ by definition of G. If x, y and z belong to three different components, we have $\delta(X) = d(x,y) = d(y,z) = d(x,z)$. Thus, we can conclude that $\leq$ is compatible with d. □

Now, suppose that G is a connected graph. In this case, we pick two elements x, y of X such that xy does not belong to E. We denote by N_x and N_y, the neighborhood of x and y respectively in G, i.e. $N_x = \{x' : xx' \in E\}$ and $N_y = \{y' : yy' \in E\}$. We define the graph $\Gamma_{xy} = (X \setminus \{x, y\}, E')$ where $E' = E \setminus \{uu' : u, u' \in N_x \text{ ou } u, u' \in N_y\}$.

Let C_{xy} be the union of all the connected components of Γ_{xy} containing at least one element of N_x and one of N_y. We define the set $\mathcal{P}_{xy}$ of paths as follows: A path $P = \{p_1, p_2, \ldots, p_m\}$ of G belongs to $\mathcal{P}_{xy}$, if it fulfills the following two properties: (i) for any i different from 1 and m, p_i is different from x and y; (ii) there does not exists p_i, p_{i+1} such that both p_i and p_{i+1} belong to N_x or belong to N_y. Note that, for any element z of C_{xy}, there exists a path of $\mathcal{P}_{xy}$ between x and y passing through z. Pick a path P' in Γ_{xy} between an element x' of N_x and an element y' of N_y passing through z. By definition of Γ_{xy}, P' belongs to $\mathcal{P}_{xy}$. If we add x and y as the predecessor of x' and as the successor of y', respectively, then we obtain a path of $\mathcal{P}_{xy}$ between x and y passing through z.

We denote by C_x and C_y the connected components of $G \setminus C_{xy}$ containing x and y respectively.

Lemma 5. *The sets C_x, C_y, C_{xy} are not empty and define a partition of X.*

Proof. The sets C_x and C_y are not empty, because they contain x and y respectively. Concerning C_{xy}, take a minimal path P between x and y : a path is minimal if there does not exist an edge between two non-consecutive elements of P. Such a path necessarily exists because G is connected. Let x' and y' be the element of N_x and N_y respectively, which belongs to P. Note that it may happen $x' = y'$. By definition of Γ_{xy}, the path between x' and y' obtained by removing x and y in P belongs to Γ_{xy}. Thus, x' and y' belong to C_{xy} and C_{xy} is non empty. In passing, we also have proved that there is no minimal path between x and y in $G \setminus C_{xy}$. Therefore, C_x and C_y are disjoint. Moreover, C_x and C_y clearly do not intersect C_{xy}.

Now, it remains to show that any element z of X belongs to $C_x \cup C_y \cup C_{xy}$. As G is connected, we necessarily have either (i) a minimal path P between x and z which does not contain y or (ii) a minimal path P between y and z which does not contain x. Suppose that we are in the case (i), the proof for the case (ii) is similar. Let x' be the neighbor of x in P. By the definition of Γ_{xy} and the minimality of P, the path $P' = P \setminus \{x\}$ is a path between x' and z in Γ_{xy}. If x' belongs to C_{xy}, by the existence of P', z belongs to the same connected component as x'. Thus, z belongs to C_{xy}. Now, if x' does not belong to C_{xy}, then no element of P' belongs to C_{xy}. Therefore, the graph $G \setminus C_{xy}$ contains P, and z belongs to C_x. □

Lemma 6. *If d is Robinsonian, then every order compatible with d refines the quasi-order $(\preceq) = (C_x, C_{xy}, C_y)$ or its converse.*

Proof. Let $\leq$ be an order compatible with d. Without loss of generality, we suppose that $x < y$. Let z be an element of C_{xy}.

First, we assert that $x < z < y$. By way of contradiction, we suppose that $z < x < y$ (the case $x < y < z$ is similar). Let P be a path of $\mathcal{P}_{xy}$ between x and y passing through z. There exists two consecutive elements u, v of P such that $u < x < v$. Since P belongs to $\mathcal{P}_{xy}$, we have $\max\{d(x,u), d(x,v)\} = \delta(X) > d(u,v)$. That contradicts $\leq$ is compatible. Thus $x < C_{xy} < y$.

Now, we show that $C_x < C_{xy}$ (the proof that $C_{xy} < C_y$ is similar). By way of contradiction, we suppose that there exists $x' \in C_x \setminus \{x\}$ and $z \in C_{xy}$ such that $x < z < x'$. We distinguish two cases: $x < z < x' < y$ and $x < z < y < x'$. Suppose that $x < z < x' < y$. Let P be a path of $\mathcal{P}_{xy}$ between y and z. Let u, v be two consecutive elements of P such that $u < x' < v$. Since P belongs to $\mathcal{P}_{xy}$, u and v are different from x and one of them, say v, is not an element of N_x. Thus, if there exists an edge in G between v and x', vx' is also an edge of Γ_{xy} (except if $y = v$, but in this case x' does not belong to C_x). Moreover, note that $P \setminus \{y\}$ is a path of Γ_{xy}. Therefore, we have a path in Γ_{xy} between x' and z. We obtain a contradiction with $x' \in C_x$ and $z \in C_{xy}$. So, vx' is not an edge of G. We deduce that $d(v, x') = \delta(X) > d(u, v)$. Thus, if $x < z < x' < y$ then $\leq$ is not compatible. It remains to consider the case $x < z < y < x'$. Pick a minimal path P of $G \setminus C_{xy}$ between x and x'. This path belongs to $\mathcal{P}_{xy}$. Let u, v be two consecutive elements of P, such that $u < y < v$. As previously, if yu is an edge of G, x' does not belong to C_x (except if $u = x$, but in this case xy is not an edge of G). Thus, we have $d(y, u) > d(u, v)$ and we obtain a contradiction with $\leq$ is a compatible order. □

Some elements of N_x and N_y may belong to C_{xy}. Therefore, it is possible that the sets C_x, C_y and C_{xy} are not independent. So, we need to refine the quasi-order $(\preceq') = (C_x, C_{xy}, C_y)$ to obtain a new quasi-order $\preceq$ such that the blocks of $\preceq$ are independent and every compatible order refines $\preceq$ or its converse. To construct such a quasi-order, we will use the procedure refine presented by Chepoi, Fichet (1997). The strategy used in this paper is quite

standard and also used in different works (Mirkin, Rodin (1984); Durand (1989); Batdedat (1990)). This procedure is based on the following property: Let $(\preceq') = (B_1, B_2, \ldots, B_m)$ be a quasi-order such that every order compatible with d refines $\preceq'$ or its converse. Let $\leq$ be such an order. Then, for any $x \in B_i$, and for any $y, z \in B_j$ such that $d(x, y) < d(x, z)$, we have $y < z$ if $i < j$ and $z < y$ if $j < i$.

We now shortly describe the procedure Refine on a a quasi-order. For an element $x \in B_i$, we remove the relation $y \preceq' z$ for all $y, z \in B_j$ such that $d(x, y) < d(x, z)$ and $j < i$ or such that $d(x, y) > d(x, z)$ and $j > i$. It remains to recursively call the procedure on the blocks which have been divided. In (Chepoi, Fichet (1997)), the following property has been proved:

Lemma 7. *Let $\preceq'$ be a quasi-order such that every order compatible with d refines $\preceq'$ or its converse. Let $(\preceq) = (B_1, B_2, \ldots, B_m)$ be the quasi-order returned by the procedure* refine *on $\preceq'$. Then, every order compatible with d refines $\preceq$ or its converse. Thus if $\preceq$ is not compatible with d, then d is not Robinsonian. Moreover, for all $x \in B_i$, and for all $y, z \in B_j$ with $i \neq j$, we have $d(x, y) = d(x, z)$.*

Let $(\preceq) = (B_1, B_2, \ldots, B_m)$ be the quasi-order returned by the procedure Refine on the quasi-order $(\preceq') = (C_x, C_{xy}, C_y)$. The lemma 6 implies that any compatible order refines $\preceq'$. By lemma 7, if $\preceq$ is not compatible then d is not Robinsonian. Thus, we test if $\preceq$ is not compatible with d. In this case, we returned that d is not Robinsonian. Otherwise, the lemma 7 implies that the blocks of $\preceq$ are independent. Therefore, we recursively call our algorithm on each B_i and denote by $T(B_i)$ the returned PQ-tree. Then, we construct the PQ-tree which has the Q-node $Q(T(B_1), T(B_2), \ldots, T(B_m))$ as root.

Lemma 8. *Let $(\preceq) = (B_1, B_2, \ldots, B_m)$ be the quasi-order returned by the procedure* Refine *on (C_x, C_{xy}, C_z). Suppose that the restriction of d on each B_i is Robinsonian. For every B_i, let $T'(B_1), T'(B_2), \ldots, T'(B_m)$ be PQ-trees such that the orders of $T'(B_i)$ are exactly the orders compatible with d restricted on B_i. If T is the PQ-tree which has $Q(T'(B_1), T'(B_2), \ldots, T'(B_m))$ as root, then the orders of T are exactly the orders compatible with d.*

Proof. First, let $\leq$ be a order compatible with d. We will prove that $\leq$ is an order of T. By lemma 6 and lemma 7, $\leq$ refines $(\preceq) = (B_1, B_2, \ldots, B_m)$ or its converse. Moreover, the restriction of $\leq$ on B_i is an order of $T'(B_i)$. Thus, $\leq$ is an order of T.

Now, we have to show that every order of T is compatible with d. Let $\leq$ be an order of T and x, y and z be three elements of X such that $x < y < z$. If x, y and z belong to a common component B_i, we have $\max\{d(x, y), d(y, z)\} \leq d(x, z)$ by definition of $T(B_i)$. Otherwise, since we test if $\preceq$ is a compatible quasi-order and two distinct blocks of $\preceq$ are independent, we have $\max\{d(x, y), d(y, z)\} \leq d(x, z)$. We can conclude that $\leq$ is compatible with d. □

The lemma 4, 7 and 8 directly implies the following theorem:

Theorem 1. *If the algorithm returns that d is not Robinsonian then d is not Robinsonian. Otherwise, let T be the PQ-tree returned by our algorithm. The orders of T are exactly the orders compatible with d.*

4 Complexity

In this section, we justify that our algorithm has a complexity in $\mathcal{O}(n^3)$ in time and in $\mathcal{O}(n^2)$ in space.

For the procedure Refine, we need a matrix $D_{\leq}$ of size n^2 described in (Chepoi and Fichet (1997)). Note that this matrix can be constructed in time $\mathcal{O}(n^2)$. For each recursive call, we also need the graph G of size n^2, but we only need one graph at the same time. Therefore, the complexity in space of this algorithm is $\mathcal{O}(n^2)$.

Now, concerning the complexity in time, we first prove that the complexity of one recursive call, without considering the procedure refine, is $\mathcal{O}(n^2)$. It is possible to construct G and its connected components in $\mathcal{O}(n^2)$. Since C_{xy} is the union of some connected components of Γ_{xy}, C_{xy} can be constructed in $\mathcal{O}(n^2)$. In the same way, C_x and C_y are the connected components of $G \setminus C_{xy}$ containing x and y respectively. Thus, they also can be constructed in $\mathcal{O}(n^2)$. We can conclude that a recursive call of our algorithm without the procedure refine has a complexity in $\mathcal{O}(n^2)$. Moreover, it has been proved that the general complexity of the procedure refine is $\mathcal{O}(n^3)$ (Chepoi, Fichet (1997)). As the sets C_x, C_y and C_{xy} cannot be empty, there is at most n recursive calls and the general complexity of our algorithm is $\mathcal{O}(n^3)$.

Refine
Input: A quasi-order $(\preceq') = (B_1, B_2, \ldots, B_m)$ and a dissimilarity d defined on X
Output: A quasi-order $\preceq$ whose blocks are independent

1. $L := \emptyset$;
2. **if** $m = 1$
3. **return** $(\preceq) := (\preceq')$
4. **else**
5. $i := 1$
6. **until** $i = m$
7. **for any** $b \in B_i$
8. $(\preceq) := refine(X, \preceq', b)$
9. Let $(\preceq) = (B_{(1,1)}, \ldots, B_{(1,k_1)}, \ldots, B_{(i-1,k_{i-1})}, B_i, B_{(i+1,1)}, \ldots, B_{(m,k_m)})$ be the obtained quasi-order
10. $m := \sum_{j=1}^{i-1} k_j + \sum_{j=i+1}^{m} k_j + 1$
11. $i := \sum_{j=1}^{i-1} k_j + 1$
12. **return** $(\preceq) := (refine((B_{(1,1)}, \ldots, B_{(1,k_1)}), R), \ldots, refine((B_{(m,1)}, \ldots, B_{(m,k_m)}), R))$

Robinson
Input: A finite set X and a dissimilarity defined on X
Output: a PQ whose orders are all the compatible orders with d
1. **if** $|X| = 1$
2. **return** $\{X\}$
3. **else**
4. Construct $G = (X, E)$
5. **if** G is not connected
6. Let $B_1, B_2, \ldots, B_p$ be the connected components of G
7. **return** $T = P\{Robinson(B_1), Robinson(B_2), \ldots, Robinson(B_p)\}$
8. **else**
9. Let x, y be two elements such that $xy \notin E$
10. Construct C_{xy}, C_x, C_y
12. $(\preceq) :=$ refine$(X, (C_x, C_{xy}, C_y))$
14. Let $(\preceq) = (B_1, B_2, \ldots, B_m)$ the quasi-order obtained
15. **if** $\preceq$ is not compatible with d
16. **return** d is not Robinsonian
17. **else**
18. **return** $T = Q(Robinson(B_1), Robinson(B_2), \ldots, Robinson(B_m))$

5 Acknowledgements

I am grateful to Bernard Fichet for his help during the work on this paper. This research was partly supported by the ANR grant BLAN06-1-138894 (projet OPTICOMB).

References

BATBEDAT, A. (1990): Les approches pyramidales dans la classification arborée. *Paris: Masson.*

BERTRAND, P. (1995): Structural properties of pyramidal clustering, *DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 19, 35-53.*

BOOTH, K. and LUEKER, G. (1976): Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. Sys. Sci. 13, 355-379.*

CHEPOI, V. and FICHET, B. (1997): Recognition of Robinsonian dissimilarities. *Journal of Classification 14, 311–325.*

CRITCHLEY, F. and FICHET, B. (1994): The partial order by inclusion of the principal classes of dissimilarity on a finite set, and some of their basic properties. *In B. van Cutsem (Ed.) Classification and Dissimilarity Analysis. Lecture Notes in Statistics, 5–65.*

DURAND, C. (1989): Ordres et graphes pseudo-hiérarchiques: théorie et optimisation algorithmique. *Thèse de l'Université de provence, Marseille.*

MIRKIN B. and RODIN, S. (1984): Graphs and Genes. *Springer-Verlag, Berlin.*

ROBINSON, W.S. (1951): A method for chronologically ordering archaeological deposits. *American Antiquity, **16**, 293–301.*

Part V

Computational Methods for Industry

AOQL Plans by Variables when the Remainder of Rejected Lots is Inspected

Jindřich Klufa

University of Economics
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, *klufa@vse.cz*

Abstract. In this paper we shall deal with the AOQL single sampling plans for inspection by variables when the remainder of rejected lots is inspected. We shall report on an algorithm allowing calculation[1] of these plans. For the calculation we shall derive a new theorem (see Theorem 1) and we shall use an original method.

Keywords: acceptance sampling, AOQL plans, software Mathematica

1 Introduction

In a book written by Dodge and Romig (1998) sampling plans (n, c) are considered which minimize the mean number of items inspected per lot of process average quality, I_s in (1), assuming that the remainder of rejected lots is inspected

$$I_s = N - (N - n)\cdot L(\bar{p}; n, c) \tag{1}$$

under the condition

$$\max_{0<p<1} AOQ(p) = p_L \tag{2}$$

(AOQL single sampling plans), where N is the number of items in the lot (the given parameter), $\bar{p}$ is the process average fraction defective (the given parameter), p_L is the average outgoing quality limit (the given parameter, denoted AOQL), n is the number of items in the sample ($n < N$), c is the acceptance number (the lot is rejected when the number of defective items in the sample is greater than c), $L(p)$ is the operating characteristic (the probability of accepting a submitted lot with fraction defective p), $AOQ(p)$ is average outgoing quality (the mean fraction defective after inspection when the fraction defective before inspection was p). The average outgoing quality (all defective items found are replaced by good ones) is approximately

$$AOQ(p) \approx \left(1 - \frac{n}{N}\right)\cdot p \cdot L(p; n, c). \tag{3}$$

Condition (2) protects the consumer against the acceptance of a bad lot. The AOQL plans are extensively tabulated - see Dodge and Romig (1998).

[1] The calculation is considerably difficult, we must use sequentialy three numerical methods.

2 AOQL plans by variables and attributes

The Dodge-Romig AOQL plans can be used under the assumption that each inspected item is classified as either good or defective *(inspection by attributes).* The problem to find AOQL plans for *inspection by variables* has been solved by Klufa (1997) under the following assumptions: measurements of a single quality characteristic X are independent, identically distributed normal random variables with unknown parameters μ and σ^2. For the quality characteristic X is given either an upper specification limit U (the item is defective if its measurement exceeds U), or a lower specification limit L (the item is defective if its measurement is smaller than L). It is further assumed that the unknown parameter σ is estimated from the sample standard deviation s (unknown standard deviation plans).

The inspection procedure is as follows: draw a random sample of n items and compute $\bar{x}$ and s; accept the lot if

$$\frac{U-\bar{x}}{s} \geq k, \quad \text{or} \quad \frac{\bar{x}-L}{s} \geq k. \tag{4}$$

The problem is to determine the sample size n and the critical value k. There are different solutions of this problem. In the present paper we shall look for the acceptance plan (n, k) minimizing the mean inspection cost per lot of process average quality, C_{ms}, under the condition

$$\max_{0<p<1} \left(1-\frac{n}{N}\right) \cdot p \cdot L(p; n, k) = p_L. \tag{5}$$

Assuming that the sample is inspected by variables and the remainder of rejected lots is inspected by attributes *(the inspection by variables and attributes),* the inspection cost per lot is $n\, c_m^*$ with probability $L(p; n, k)$, and $[n\, c_m^* + (N-n)\, c_s^*]$ with probability $[1-L(p; n, k)]$, where c_s^* is the cost of inspection of one item by attributes, and c_m^* is the cost of inspection of one item by variables. The mean inspection cost per lot of process average quality is therefore

$$C_{ms} = n \cdot c_m^* + (N-n) \cdot c_s^* \cdot [1 - L(\bar{p}; n, k)]. \tag{6}$$

Let us denote

$$c_m = \frac{c_m^*}{c_s^*}. \tag{7}$$

Instead of C_{ms} we shall look for the acceptance plan (n, k) minimizing

$$I_{ms} = n \cdot c_m + (N-n) \cdot [1 - L(\bar{p}; n, k)], \tag{8}$$

noting that both functions C_{ms} and I_{ms} have a minimum for the same acceptance plan ($C_{ms} = I_{ms} \cdot c_s^*$), under the condition (5). For these AOQL plans, the parameter c_m in (7) must be statistically estimated in each real situation. Usually is $c_m > 1$. Putting formally $c_m = 1$ into (8) (I_{ms} in this case is denoted I_m) we obtain $I_m = N - (N-n) \cdot L(\bar{p}; n, k)$, i.e. the mean

number of items inspected per lot of process average quality, assuming that both the sample and the remainder of rejected lots is inspected by variables. Consequently we shall study *the AOQL plans for inspection by variables* as a special case of *the AOQL plans by variables and attributes* for $c_m = 1$. From I_m is evident that for the determination AOQL plans by variables it is not necessary to estimate c_m ($c_m = 1$ is not a real value of this parameter).

Summary: For the given parameters N, $\bar{p}$, p_L and c_m we must determine the acceptance plan (n, k) for inspection by variables and attributes, minimizing I_{ms} in (8) under the condition (5).

In the first place we shall deal with the solution of the equation (5). The operating characteristic is (e.g., Klufa (1999))

$$L(p; n, k) = \int_{k\sqrt{n}}^{\infty} g(t; n-1, u_{1-p}\sqrt{n})\, dt, \tag{9}$$

where $g(t; n-1, u_{1-p}\sqrt{n})$ is probability density function of non-central t distribution with $(n-1)$ degrees of freedom and noncentrality parameter $u_{1-p}\sqrt{n}$.

Instead of (9), using the normal distribution as an approximation of the non-central t distribution (Johnson and Welch (1940)), we have

$$L(p; n, k) \approx \Phi\left(\frac{u_{1-p} - k}{A}\right), \text{ where } A = \sqrt{\frac{1}{n} + \frac{k^2}{2(n-1)}}\,. \tag{10}$$

The function Φ in (10) is a standard normal distribution function and u_{1-p} is a quantile of order $1-p$. The equation (5), using (10), has an equivalent form

$$\max_{0<p<1}\ p \cdot \Phi\left(\frac{u_{1-p} - k}{A}\right) = \frac{p_L}{1 - \frac{n}{N}}. \tag{11}$$

Let us denote

$$G(p; n, k) = p \cdot \Phi\left(\frac{u_{1-p} - k}{A}\right), \quad M(n, k) = \max_{0<p<1} G(p; n, k). \tag{12}$$

Let n, N, p_L be given parameters (for given n we shall write $M_n(k)$ instead of $M(n, k)$). In the first place we shall look for the critical value k for which (11) holds, i.e.

$$M_n(k) = p_L/(1 - \tfrac{n}{N}). \tag{13}$$

Solution of the equation (13) is unique (see proof in Klufa (1997)). Since an explicit formula for k does not exist, we have to solve (13) numerically. We use Newton's method, therefore we must determine $M_n(k)$ and $M_n'(k)$. According to (12) one has

$$M_n(k) = p_M \cdot \Phi\left(\frac{u_{1-p_M} - k}{A}\right), \tag{14}$$

where $p_M \in (0, 1)$ is the value of p, for which the function $G(p)$ has a maximum. Evidently, it holds that $G(0) = G(1) = 0$ and $G(p) > 0$ for $p \in (0, 1)$.

Since the function $G(p)$ is continuous for $p \in \langle 0,1\rangle$, the value p_M exists. We determine the value p_M as a solution of the equation $G'(p) = 0$, i.e.

$$\Phi\left(\frac{u_{1-p}-k}{A}\right) - \frac{p}{A}\,\exp\left[-\frac{1}{2A^2}\left[(1-A^2)u_{1-p}^2 - 2k\,u_{1-p} + k^2\right]\right] = 0. \quad (15)$$

Theorem 1. *Let n be given parameter, $n \in \langle 7, (1{-}4p_L)N\rangle$, $k_r = (n-1)\sqrt{\frac{2}{n}}$. If $k \in \langle 0,\infty) - \{k_r\}$, then solution p_M of the equation (15) is between p_a and p_r, where*

$$p_a = \Phi\left(\frac{-k - A\sqrt{k^2 - 2(1-A^2)\ln A}}{1-A^2}\right), \quad p_r = \Phi\left(-\frac{k}{1+A}\right). \quad (16)$$

The proof of this theorem is presented in Klufa (2008).

Instead of p_M we shall look for $x_M = u_{1-p_M}$ $(p_M = \Phi(-x_M))$ as a solution of the equation $G'(x) = 0$, i.e.

$$\Phi\left(\frac{x-k}{A}\right) - \frac{\Phi(-x)}{A}\cdot\exp\left[-\frac{1}{2A^2}\left[(1-A^2)x^2 - 2kx + k^2\right]\right] = 0. \quad (17)$$

The equation (17) must be solved once more numerically. Numerical solution of the equation $G'(x) = 0$ depends on good first approximation x_0. Under assumptions of Theorem 1, solution x_M of the equation (17) is between x_r and x_a, where

$$x_r = \frac{k}{1+A}, \quad x_a = \frac{k + A\sqrt{k^2 - 2(1-A^2)\ln A}}{1-A^2}. \quad (18)$$

Using (18) we choose for x_0 following point (numerical investigations show that this point is good start value for solution of the equation (17))

$$x_0 = \frac{(100+n)x_r + nx_a}{2n+100}. \quad (19)$$

If we find x_M for which (17) holds, then we determine $M_n(k)$ from the formula

$$M_n(k) = \Phi(-x_M)\cdot\Phi\left(\frac{x_M - k}{A}\right) \quad (20)$$

and the derivative $M_n'(k)$ from the formula

$$M_n'(k) = -\frac{\Phi(-x_M)}{A^3\sqrt{2\pi}}\cdot\left[\frac{1}{n} + \frac{k\ x_M}{2(n-1)}\right]\cdot\exp\left[-\frac{1}{2A^2}(x_M - k)^2\right]. \quad (21)$$

Determination of the acceptance plans (n,k) for which (13) holds is considerably difficult. From these plans we must choose the acceptance plan (n,k) minimizing $I_{ms} = n\cdot c_m + (N-n)\cdot\alpha$, where

$$\alpha = 1 - L(\bar{p}; n, k) = \Phi\left(\frac{k - u_{1-\bar{p}}}{A}\right) \quad (22)$$

is producer's risk (the probability of rejecting a lot of process average quality). This problem we shall solve once more numerically.

3 Calculation of the AOQL plans for inspection by variables and attributes

For calculation of the AOQL plans by variables and attributes we shall use software Mathematica - see Wolfram (1991).

Example. Let $N = 1000$, $p_L = 0.0025$, $\bar{p} = 0.001$ and $c_m = 1.8$ (the cost of inspection of one item by variables is higher by 80% than the cost of inspection of one item by attributes). We shall look for the AOQL plan for inspection by variables and attributes. Furthermore we shall compare this plan and the corresponding Dodge-Romig AOQL plan for inspection by attributes.

Solution. In the first step we shall determine x_M as a solution of equation $G'(x)=0$ (see (17)). According to (18) and (19) we have ($\bar{p}$=pbar, N=nbig)

```
In[1]:= <<Statistics‘ContinuousDistribution‘
In[2]:= ndist = NormalDistribution[0,1]
In[3]:= cm = 1.8
In[4]:= AOQL = 0.0025
In[5]:= pL = 0.0025
In[6]:= pbar = 0.001
In[7]:= nbig = 1000
In[8]:= A[n_,k_] := Sqrt[1/n + k^2/(2n-2)];
G’[x_,n_,k_] := CDF[ndist, (x - k)/A[n,k]] - CDF[ndist, -x]*
Exp[-((1 - A[n,k]^2) x^2 - 2k x + k^2)/ (2A[n,k]^2)]/A[n,k];
xr[n_,k_] := k/(1 + A[n,k]);
xa[n_,k_] := (k + A[n,k]*Sqrt[k^2 - 2(1 - A[n,k]^2)*
Log[A[n,k]]])/(1 - A[n,k]^2);
x0[n_,k_] := ((100 + n)*xr[n,k] + n*xa[n,k])/(2n + 100);
FR[n_,k_] := FindRoot[G’[x,n,k] == 0, {x, x0[n,k]}];
xM[n_,k_] := x / . FR[n,k];
```

Now we shall solve equation $M_n(k) = p_L/(1 - \frac{n}{N})$ (for given n we shall look for critical value k for which (13) holds). Using Newton's method (see (20) and (21)) with start point o=1.6 we have

```
c[n_,k_] := -(CDF[ndist,-xM[n,k]]*CDF[ndist,(xM[n,k] - k)/
A[n,k]] - pL/(1 - n/nbig))/ (-CDF[ndist, -xM[n,k]]*
(1/n + k xM[n,k]/ (2n - 2))*Exp[-(xM[n,k] - k)^2/(2A[n,k]^2)]/
(A[n,k]^3*Sqrt[2Pi]));
o = 1.6;
fRecAux[n_,i_] := fRecAux[n,i]=fRecAux[n,i-1] +
c[n,fRecAux[n,i-1]]; fRecAux[n_,0]=o;
k[n_] := fRecAux[n,7];
```

Finally in the third step we shall determine the acceptance plan (n, k) minimizing $I_{ms} = n{\cdot}c_m + (N - n){\cdot}\alpha$, where α is producer's risk (see (22)), under condition (13). Solution of this problem is as follows (half-intervals method):

```
   a[n_] := CDF[ndist, (k[n] - Quantile[ndist, 1- pbar])/
Sqrt[1/n + k[n]^2/(2n - 2)]];
   Ims[n_] := n cm + (nbig - n)*a[n];
   FMinSearch[nl_,nu_] := nl /; nl == nu;
FMinSearch[nl_,nu_] := FMinSearch[nl, nl+Floor[(nu-nl)/2]] /;
Ims[nl+Floor[(nu-nl)/2]] <= Ims[nl+Floor[(nu-nl)/2]+1];
FMinSearch[nl_,nu_]:= FMinSearch[nl+Floor[(nu-nl)/2]+1,nu] /;
n=FMinSearch[7, nbig/2];
```

Correction for non-central t distribution (see (9)):

```
   In[25]:= lambda[p_]=:Quantile[ndist,1-p]*Sqrt[n]
   In[26]:= nonctdist[p_]=:
NoncentralStudentDistribution[n-1, lambda[p]]
   In[27]:= L1[p_]:= 1-CDF[nonctdist[p], k[n]*Sqrt[n]]
   In[28]:= AOQ[p_]:= (1-n/nbig)*p*L1[p]
   In[29]:= d=0.00001
   In[30]:= fMSmodq[pl_, pu_] := pl /; pl == pu
fMSmodq[pl_, pu_]:= fMSmodq[pl, pl+Floor[(pu-pl)/(2d)]*d] /;
-AOQ[pl + Floor[(pu - pl)/(2d)]*d] <= -AOQ[pl + Floor[(pu -
pl)/(2d)]*d + d]
fMSmodq[pl_, pu_] := fMSmodq[pl+Floor[(pu-pl)/(2d)]*d+d, pu]
   In[33]:= pLtrue := AOQ[fMSmodq[0.00001, 0.01] ]
   In[34]:= pcpL = 0.00000001
   In[35]:= samplan := n, k[n] /; (konst = pLtrue;
Abs[konst - AOQL] < pcpL );
samplan := (pL = pL + AOQL - konst; Clear[fRecAux];
fRecAux[n_, i_] := fRecAux[n, i] = fRecAux[n, i - 1] +
c[n, fRecAux[n, i - 1]];
fRecAux[n_, 0] = o; samplan)
   In[37]:= samplan
   Out[37]= {49, 2.57617}[2]
```

The AOQL plan for inspection by variables and attributes is $n = 49$, $k = 2.57617$. The corresponding AOQL plan for inspection by attributes is $n_2 = 130$, $c = 0$ (see Dodge and Romig (1998)). For the comparison of these two plans from an economical point of view we use parameter $e = (I_{ms}/I_s) \cdot 100$ (see (8) and (1), the operating characteristic $L(\bar{p}; n_2, c)$ - see e.g. Hald (1981)). The Mathematica gives

```
   In[38]:= n = 49
   In[39]:= k = 2.57617
   In[40]:= n2=130
   In[41]:= c=0
```

[2] This algorithm with correction for non-central t distribution takes about six minutes (algorithm for non-central t distribution takes several hours).

```
   In[42]:= L1[p_] := 1 - CDF[nonctdist[p], k*Sqrt[n]]
   In[43]:= L2[p_] := Sum[ Binomial[nbig*p, i]*
Binomial[nbig - nbig*p, n2 - i]/Binomial[nbig, n2], i, 0, c]
   In[44]:= e = 100*(n*cm + (nbig - n)*(1 - L1[pbar]))/
(nbig - (nbig - n2)*L2[pbar])
   Out[44]:= 52.2008
```

Since $e = 52.2008\%$, using the AOQL plan for inspection by variables and attributes $(49, 2.57617)$ it can be expected approximately **48% saving of the inspection cost**[3] in comparison with the corresponding Dodge-Romig plan. Further we shall compare the operating characteristics of these plans:

```
   In[45]:= Table[{p, N[L1[p], 6], N[L2[p], 6]},
{p, 0.001, 0.025, 0.002}]
   In[46]:= TableForm[%]
   Out[46]//TableForm=
      0.001     0.959306     0.87
      0.003     0.734422     0.658207
      0.005     0.522047     0.497674
      0.007     0.365862     0.376067
      0.009     0.256813     0.284003
      0.011     0.181453     0.214346
      0.013     0.129244     0.161675
      0.015     0.0928235    0.121872
      0.017     0.0672038    0.0918112
      0.019     0.049026     0.0691225
      0.021     0.0360195    0.0520083
      0.023     0.0266387    0.039107
      0.025     0.019822     0.0293876
```

For example we get $L_1(\bar{p}) = L_1(0.001) = 0.959306$, i.e. the producer's risk for the AOQL plan for inspection by variables and attributes is therefore approximately $\alpha = 1 - L_1(\bar{p}) = 0.04$. The producer's risk for the corresponding Dodge-Romig plan is $\alpha = 1 - L_2(\bar{p}) = 1 - 0.87 = 0.13$. Finally graphic comparison of the operating characteristics of these two plans:

```
   In[47]:= oc1 =  Plot[L1[p], {p, 0, 0.025}, AspectRatio -> 0.9,
AxesLabel -> {"p", "L(p)"}, PlotStyle -> Thickness[0.0045]]
   In[48]:= oc2=ListPlot[Table[{p, L2[p]}, {p,0,0.025,0.0003}]]
   In[49]:= Show[oc1, oc2]
   Out[49] - see Figure 1
```

Conclusion. From these results it follows that the AOQL plan for inspection by variables and attributes is more economical than the corresponding

[3] Under the same protection of consumer the AOQL plans for inspection by variables and attributes are in many situations **more economical** than the corresponding Dodge-Romig AOQL attribute sampling plans - see Klufa (1997).

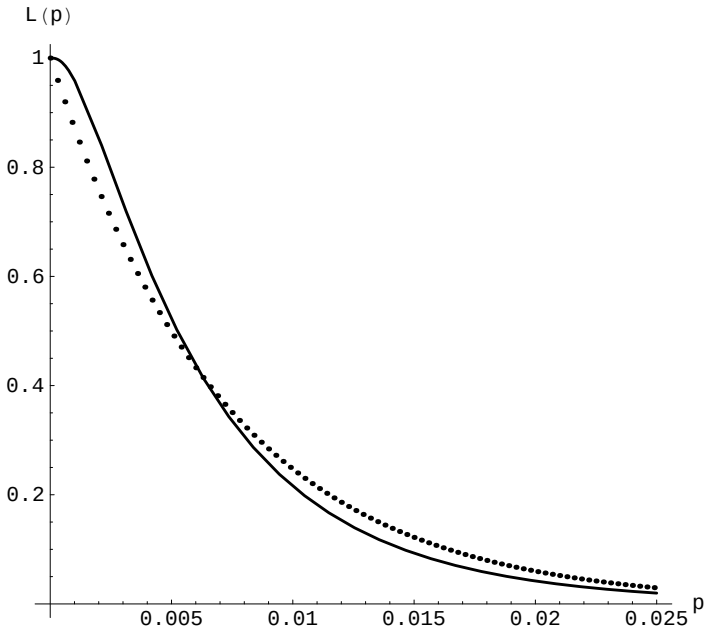


Fig. 1. OC curves for the AOQL sampling plans for inspection by variables and attributes $(49, 2.57617)$ ——— for inspection by attributes $(130, 0)$ .....

Dodge-Romig AOQL attribute sampling plan (48% saving of the inspection cost). Furthermore the OC curve for the AOQL plan by variables and attributes is better than corresponding OC curve for the AOQL plan by attributes - see Figure 1 (for example the producer's risk for the AOQL plan by variables and attributes $\alpha = 0.04$ is less than for the corresponding Dodge-Romig plan $\alpha = 0.13$).

Acknowledgements

This work was prepared with support from the Grant Agency of the Czech Republic, and under contract number 402/06/0115.

References

DODGE, H.F. and ROMIG, H.G. (1998): *Sampling Inspection Tables: Single and Double Sampling*. John Wiley, New York.

HALD, A. (1981): *Statistical Theory of Sampling Inspection by Attributes*. Academic Press, London.

JOHNSON, N.L. and WELCH, B.L. (1940): Applications of the Non-central t-distribution. *Biometrika 31, 362 - 389*.

KLUFA, J. (1997): Dodge-Romig AOQL single sampling plans for inspection by variables. *Statistical Papers 38, 111 - 119*.

KLUFA, J. (1999): *Economical Aspects of Acceptance Sampling*. Ekopress, Prague.
KLUFA, J. (2008): Dodge-Romig AOQL plans for inspection by variables from numerical point of view. *Statistical Papers 49, 1 - 13*.
WOLFRAM, S. (1991): *Mathematica*. Addison-Wesley

Improved Local Sensitivity Measure for Regression Models with Correlated Parameters

Hana Sulieman

Department of Mathematics and Statistics, American University of Sharjah
Sharjah, United Arab Emirates, *hsulieman@aus.edu*

Abstract. In parameter estimation local sensitivity assessment; conventionally measured by the first-order derivative of the predicted response with respect to a parameter of interest fails to provide a representative picture of the prediction sensitivity in the presence of significant parameter co-dependencies. In this article we propose a profile-based sensitivity measure defined by the total derivative of the model predicted response with respect to a parameter. Although inherently local, the profile-based measure is shown to account for parameter co-dependencies and the underlying model nonlinearity and so it has broader range of validity than the conventional local sensitivity measure. The comparison between the two measures is illustrated by two examples.

Keywords: local sensitivity analysis, nonlinear parameter estimation, profile-based sensitivity coefficient

1 Introduction

Parametric sensitivity analysis in general describes the impact of perturbations in the values of model input parameters on the model outputs. The objective of the sensitivity assessment is to improve the quality of an existing model representing a physical system perhaps by reducing complexity or by guiding further experiments to reduce uncertainty. Through better understanding of the interplay between the model input parameters and the relative importance they have on the model outputs, we can target specific parameters for more detailed study to reduce model uncertainty.

This article investigates the sensitivity of the predicted responses from nonlinear regression models to variations in parameter values. The numerical values of these parameters are usually estimated using available experimental data. The resulting uncertainties associated with these parameter estimates propagate into the model predictions via sensitivities. Results of the sensitivity analysis applied to an existing model of a physical system can be used to strengthen the knowledge base by guiding subsequent research in order to increase the confidence in the model and its predictions.

Parametric sensitivity can be local in the sense that the information generated from the analysis is valid over small ranges of parameter uncertainties.

Local sensitivity is usually carried out by computing derivatives of the model function with respect to the parameters. The conventional measure of local parametric sensitivity is defined by the first-order partial derivatives of the model response function with respect to the parameters resulting from the linear approximation of the model function in the parameter space.The resulting sensitivity coefficients measure the *marginal* impact of the parameter of interest on model predictions since only the value of this parameter is perturbed while all other parameters are held fixed at their nominal values. Therefore, co-dependencies among parameter estimate are ignored by these measures, Sulieman *et. al.* (2001) called them *marginal sensitivity coefficients*. To surmount the drawbacks of the local sensitivity assessment, Sulieman *et al.* (2001) proposed an alternative local assessment procedure in which simultaneous perturbations in the values of all model parameters are achieved using the profiling scheme introduced by Bates and Watts (1988) for nonlinearity assessment of regression models. The profile-based sensitivity measure; defined by the total derivative of the model function with respect to parameter of interest, was shown to account for both nonlinearity within the parameter estimation problem and parameter estimate co-dependencies. Like any derivative measure, profile-based sensitivity is inherently local, it provides, however, a more representative picture of the prediction sensitivity in the presence of parameter co-dependencies and model nonlinearity. Detailed discussion of the profile-based sensitivity measures is presented in section 2.

This article explores a comparison between the conventional local sensitivity and the profile-based sensitivity measures. The comparison is demonstrated by the implementation of the two measures to two parameter estimation problems. The first one presents a single response model while the other presents a multi-response model. The reliability of the sensitivity results generated by the two measures is tested using a simulation exercise in which the use of sensitivity information to guide future experimentation is demonstrated.

2 Profile-based sensitivity

We consider the general multi-response regression model given by :

$$y_{nj} = f_j(\mathbf{x}_n, \Theta) + z_{nj}, \quad n = 1, \ldots, N \quad j = 1, \ldots, J \tag{1}$$

where y_{nj} is the random variable associated with the measured value of the j-th response for the n-th experimental setting, f_j is a known model function for the j-th response depending on some or all of the experimental settings $\mathbf{x}_n$ and on some or all of the parameters in Θ. z_{nj} is the disturbance term and Θ is a p-element vector of unknown parameters. When $J = 1$ the model becomes a uni-response model.

For fixed set of experimental conditions, $\mathbf{x}_n$, $n = 1, \ldots, N$, the expected responses $f_j(\mathbf{x}_n, \Theta)$ depend only on Θ and therefore can be defined as:

$$\mathbf{f}(\mathbf{x}_n, \Theta) = \mathbf{H}(\Theta) \tag{2}$$

where $\mathbf{f}$ and $\mathbf{H}$ are $N \times J$ matrices of expected responses $f_j(\mathbf{x}_n, \Theta)$.
Motivated by the profiling algorithm by Bates and Watts (1988), Sulieman *et al.* (2001, 2004) developed a profile-based sensitivity measure to assess sensitivity of the predicted responses from the model in equation (1) to the parameters Θ. The developed measure quantifies the net change in a predicted response due to perturbations in a parameter of interest, after the remaining parameters are updated to their estimates conditioned on the perturbed value of the parameter. The p-element vector of parameters Θ is partitioned as $\Theta = (\theta_i, \Theta_{-i})$ where θ_i is the parameter of interest and Θ_{-i} is $(p-1)$-element vector of the remaining parameters. The value of θ_i is varied across a specified range of uncertainty, and for each fixed value of θ_i, conditional estimates of the remaining parameters Θ_{-i}, denoted by $\tilde{\Theta}_{-i}$, are obtained by minimizing a parameter estimation criterion over the $(p-1)$-dimensional parameter space. For a selected design point $\mathbf{x}_0$, the profile-based sensitivity measure is defined by the total derivative of the predicted response with respect to θ_i:

$$\mathbf{PSC}_i(\mathbf{x}_0) = \frac{D\mathbf{H}_0'(\theta_i, \tilde{\Theta}_{-i}(\theta_i))}{D\theta_i} \tag{3}$$

where $\mathbf{H}_0'$ is J-element vector of response variables evaluated at $\mathbf{x}_0$.
For single response regression model $J = 1$, Sulieman *et al.* (2001) used the least squares estimation criterion for which the profile-based sensitivity coefficient in equation (3) is shown to equal:

$$PSC_i(\mathbf{x}_0) = \frac{\partial H_0}{\partial \theta_i} - \frac{\partial H_0}{\partial \Theta_{-i}} \left(\frac{\partial^2 S}{\partial \Theta_{-i} \partial \Theta'_{-i}}\right)^{-1} \frac{\partial^2 S}{\partial \theta_i \partial \Theta_{-i}}\Big|_{\tilde{\Theta}_{-i}} \tag{4}$$

where S is the sum of squares function, $S(\Theta) = \sum_{i=1}^{n} (y_i - H_i(\Theta))^2$, $\frac{\partial H_0}{\partial \theta_i}$ is the first-order derivative of the predicted response at $\mathbf{x}_0$ with respect to θ_i which Sulieman *et al.* (2001) called *Marginal Sensitivity Coefficient, MSC.* Expressing equation (4) in terms of the first and second derivatives of H with respect to Θ yields

$$PSC_i(\mathbf{x}_0) = v_{0_i} - \mathbf{v}'_{0_{-i}} (V'_{-i} V_{-i} - [z'][V_{-i-i}])^{-1} (V'_{-i} \mathbf{v}_i - \mathcal{D}'_{\cdot\cdot} z) \tag{5}$$

where v_{0_i} is the i-th component of the first derivative vector $\mathbf{v}_0$ evaluated at $\mathbf{x}_0$; V_{-i} is an $n \times (p-1)$ matrix consisting of first derivative vectors of $H(\Theta)$ with respect to Θ_{-i}; $\mathbf{v}_{0_{-i}}$ is a $(p-1)$ dimensional vector consisting of the elements in the row of V_{-i} which corresponds to $\mathbf{x}_0$;V_{-i-i} is the $n \times (p-1) \times (p-1)$ array of the second derivatives of $H(\Theta)$ with respect to Θ_{-i}; $\mathcal{D}_{\cdot\cdot}$ is the $n \times (p-1)$ matrix of the second derivatives of $H(\Theta)$ with respect to Θ_{-i} and θ_i, and z is the n-element residuals vector.

For the multi-response regression models, $J > 1$, Sulieman *et al.* (2004) used Box-Draper estimation criterion and obtained the following expression for the vector-valued profile-based sensitivity coefficient:

$$\mathbf{PSC}_i(\mathbf{x}_0) = \frac{\partial \mathbf{H}_0'}{\partial \theta_i} - \frac{\partial \mathbf{H}_0'}{\partial \Theta_{-i}} \left(\frac{\partial^2 d}{\partial \Theta_{-i} \partial \Theta_{-i}'}\right)^{-1} \frac{\partial^2 d}{\partial \theta_i \partial \Theta_{-i}}\Big|_{\tilde{\Theta}_{-i}} \tag{6}$$

where $d(\Theta)$ is the determinant given by $d(\Theta) = |\mathbf{Z}'\mathbf{Z}|$, $\mathbf{Z}$ is the $N \times J$ matrix of residuals defined by $\mathbf{Z}(\Theta) = \mathbf{Y} - \mathbf{H}(\Theta)$, the vector $\frac{\partial \mathbf{H}_0'}{\partial \theta_i}$ has J elements of the marginal sensitivity coefficients. $\frac{\partial \mathbf{H}_0'}{\partial \Theta_{-i}}$ is an $J \times (p-1)$ matrix consisting of the marginal sensitivity coefficients of the J model functions with respect to Θ_{-i} evaluated at $\mathbf{x}_0$. The matrix $\frac{\partial^2 d}{\partial \Theta_{-i} \partial \Theta_{-i}'}$ is the $(p-1) \times (p-1)$ sub-matrix of the Hessian matrix of $d(\Theta)$, and $\frac{\partial^2 d}{\partial \theta_i \partial \Theta_{-i}}$ is a $(p-1)$-element vector of the Hessian terms corresponding to θ_i and Θ_{-i}. The (r,s)th component of the Hessian matrix is given by:

$$\frac{\partial^2 d}{\partial \theta_s \partial \theta_r} = |\mathbf{Z}'\mathbf{Z}| \begin{pmatrix} tr(\mathbf{U}_s) tr(\mathbf{U}_r) - tr[\mathbf{U}_s \mathbf{U}_r] + tr\left[(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}_s'\mathbf{Z}_r + \mathbf{Z}_r'\mathbf{Z}_s)\right] \\ + tr\left[(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z}_{rs} + \mathbf{Z}_{sr}'\mathbf{Z})\right] \end{pmatrix} \tag{7}$$

where

$$\mathbf{U}_r = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Z}_r + \mathbf{Z}_r'\mathbf{Z})$$

$$\mathbf{Z}_r = \frac{\partial \mathbf{Z}}{\partial \theta_r} = -\frac{\partial \mathbf{H}(\Theta)}{\partial \theta_r}$$

$$\mathbf{Z}_{sr} = \frac{\partial^2 \mathbf{Z}}{\partial \theta_s \partial \theta_r} = -\frac{\partial^2 \mathbf{H}(\Theta)}{\partial \theta_s \partial \theta_r}, \quad r, s = 1, \ldots, i-1, i+1, \ldots, p.$$

The $(p-1)$ elements of the vector $\frac{\partial^2 d}{\partial \theta_i \partial \Theta_{-i}}$ are obtained by taking $s = i$ and $r = 1, \ldots, i-1, i+1, \ldots, p$ in equation (7). The quantities in the above equations are evaluated at the estimated values $(\hat{\theta}_i, \tilde{\Theta}_{-i}(\hat{\theta}_i))$.

Close examination of equations (4) and (6) shows that the profile-based sensitivity measure is a sum of two terms. The first term gives the marginal sensitivity coefficient and the second term gives an adjustment factor containing the marginal effects of Θ_{-i} on the predicted response at $\mathbf{x}_0$ and the vector of slopes for $\tilde{\Theta}_{-i}$ with respect to θ_i. These slopes measure the changes in the values of the parameters Θ_{-i} caused by the changes in the values of θ_i. These slopes consist of two components: the pairwise co-dependencies among the elements of $\tilde{\Theta}_{-i}$ and the co-dependencies between $\hat{\theta}_i$ and $\tilde{\Theta}_{-i}$. As equations (5) and (7) show, the two sets of co-dependencies are computed using the second-order derivatives of the model function so as to account for the non-linearity in the estimation problem making profile-based sensitivity a local

measure that has broader ranges of validity than the corresponding marginal sensitivity coefficient. The reliability of the marginal sensitivity information depends on the magnitude of the adjustment term which in turn depends on the extent of the nonlinear co-dependencies between the parameter estimates induced by the model formulation, parameterization and experimental design.

To produce dimensionless sensitivity coefficients, Sulieman *et al.* (2001) scaled the resulting profile-based and marginal sensitivity coefficients by the factor:

$$\frac{se(\hat{\theta}_i)}{se(\hat{H}_0)}) \tag{8}$$

where $se(\hat{H}_0)$ is the estimated standard error of the predicted response at $\mathbf{x}_0$ and $se(\hat{\theta}_i)$ is the estimated standard error of the i-th parameter estimate.

3 Illustrative examples

Example 1: Osborne (1972) fitted the model :

$$E(y) = \theta_1 + \theta_2 exp(-\theta_3 x) + \theta_4 exp(-\theta_5 x) \tag{9}$$

to data supplied by A.M. Sargeson of the Research School of Chemistry in the Australian National University. The results of model fitting using least squares estimation is given in Table 1. A common feature of fitting

Table 1. Summary of parameter estimates for the model fitted to the Osborne data.

Parameter	Estimate	St.Error	
θ_1	0.375	0.002	
θ_2	1.936	0.220	$s^2 = 1.95 \times 10^{-6}$ with 28 *df*
θ_3	0.013	0.0004	
θ_4	-1.465	0.221	
θ_5	0.022	0.0009	

linear combination of exponentials is the pronounced parameter estimate co-dependencies that are induced by model formulation. In this example, parameter estimates correlations involving θ_2, θ_3, θ_4 and θ_5 approach 1. Using equation (5) profile-based and marginal sensitivity coefficients for the five parameters were calculated for each of the 33 cases. The results are displayed in Figure 1. The solid and dotted lines in each plot represent,respectively, the values of $\widehat{MSC}_i$ and $\widehat{PSC}_i$. Due to model formulation the values of of $\widehat{MSC}_3$ and $\widehat{MSC}_5$ at minimum $\mathbf{x}_0 = 0$ are zeros. These values are corrected

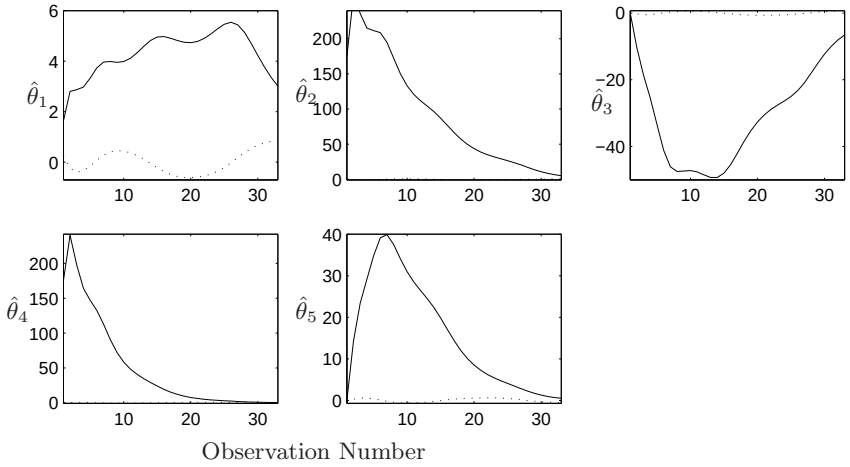


Fig. 1. Parameter sensitivities for the exponential model. The solid lines join the $\widehat{MSC}_i(\mathbf{x}_0)$ values and the dotted lines join the $\widehat{PSC}_i(\mathbf{x}_0)$ values.

by the corresponding non-zero values $\widehat{PSC}_3 = 0.34$ and $\widehat{PSC}_5 = -0.6$. The $\widehat{PSC}$ values show no distinguishing strong effects of any of the parameters on the predicted response at any $\mathbf{x}_0$. In other words, when the parameters are adjusted for their strong co-dependencies, their influences on the predicted responses are relatively small, implying that further experimentation at any of the design points could have negligible impact on the estimated values of the five parameters and the model predictions. Adding more experiments to the design will not improve the precision of the predicted response. The relationships between the estimated parameters and the predicted responses become clearer when these correlations are reduced, which can be achieved by using appropriate parameter transformations. In contrast, the $\widehat{MSC}$ values suggest some very strong sensitivity relationships.

Example 2: The model here is a multi-response model that was taken from Bates and Watts (1987). The three expected responses are linear functions in two parameters,

$$f_1(x_{n1}, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{n1}, \quad f_2(x_{n2}, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{n2}, \quad f_3(x_{n3}, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{n3},$$

where x_{nj} is the value of the jth independent variable at the nth experimental settings, $n = 1, 2, \ldots, 8$ and $j = 1, 2, 3$. The interesting feature of this simple linear multi-response model is that there is no nonlinearity associated with the parameters, however, the use of the determinant criterion produces a nonlinear estimation problem. Using the data given by Bates and Watts (1987), the parameters were estimated by minimizing $d(\Theta)$. The resulting

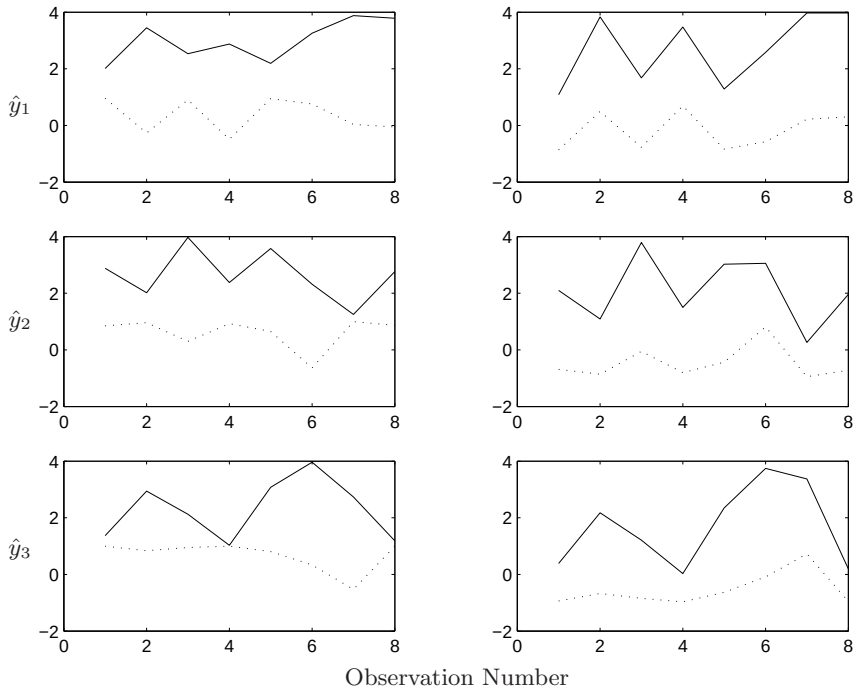


Fig. 2. Parametric sensitivities for the multi-response linear model. The solid lines join the $\widehat{MSC}_i^j(\mathbf{x}_0)$ values and the dotted lines join the $\widehat{PSC}_i^j(\mathbf{x}_0)$ values. Plots in the first column correspond to the sensitivities to $\hat{\beta}_0$ and in the second column to the sensitivities to $\hat{\beta}_1$.

values of the parameter estimates are $(\hat{\beta}_0, \hat{\beta}_1) = (0.408, 2.555)$ with a determinant of 569 and $corr(\hat{\beta}_0, \hat{\beta}_1) = -0.968$. The estimated variance-covariance matrix is

$$\hat{\Sigma} = \frac{\hat{\mathbf{Z}}'\hat{\mathbf{Z}}}{8} = \begin{bmatrix} 0.962 & 1.130 & 1.231 \\ & 2.464 & 1.166 \\ & & 2.660 \end{bmatrix} \tag{10}$$

Using equations (6) and (7) the vector-valued profile-based and marginal sensitivity coefficients for the two parameters were calculated for each of the 8 cases. The results are displayed in Figure 2. Due to the strong co-dependency between the two parameter estimates, the two sensitivity curves in each plot differ in magnitude and behavior.

According to the marginal sensitivity coefficients, the predicted response $\hat{y}_1$, shown in the plots of the top row,is most sensitive to both parameter estimates at cases number 7 and 8 while the smallest sensitivities to both parameter estimates appear at case # 1. Comparing the $\widehat{PSC}_i^1$ curves indicates

that the predicted value of y_1 at case # 1 exhibits the greatest magnitude of sensitivities to both parameter estimates, while at case # 7, $\hat{y}_1$ manifests nearly zero sensitivities to both parameter estimates. Thus, when the two parameters are estimated jointly, their individual impacts on the predicted value of y_1 at case # 7 decline to almost zero.

Similar conclusions can be drawn for the sensitivities of $\hat{y}_2$ and $\hat{y}_3$. According to the profile-based sensitivities, $\hat{y}_2$ is most sensitive to $\hat{\beta}_0$ and $\hat{\beta}_1$ at case # 7, the location of least marginal sensitivities, while $\hat{y}_3$ is most sensitive to both parameter estimates at the location of case # 4, again the location of least marginal sensitivities. $\widehat{MSC}_i^2$ and $\widehat{MSC}_i^3$ curves show that the greatest sensitivities to both parameter estimates occur at case # 3 for $\hat{y}_2$ and at case # 6 for $\hat{y}_3$.

3.1 Simulation exercise

To assess the above scenarios suggested by the marginal and profile-based sensitivities of the three responses, we carried out the following simulation exercise. An additional observation to the data set was simulated at the conditions of case number 1 using the fitted model equations and adding a normally distributed noise triplet with 0 means and variance-covariance matrix given in equation (10). The parameter estimates were then obtained from the expanded data set consisting of the original 8 cases and the newly generated case. This simulation exercise was repeated at the conditions of case numbers 3, 4, 6 and 7. In each instance, the improvement in the precisions of the predicted responses was measured by the reduction in the standard errors obtained using the original data set. A sufficient number of simulations were carried out for each case to establish a consistent pattern in the improvement of the parameter estimate precisions. Table 2 shows the average values of the improvement in the precisions of the three predicted responses at the five cases (1, 3, 4, 6 and 7) over 500 simulations. Examination of the values in Table 2 reveal that the greatest improvement in the precision of the predicted values occurred for $\hat{y}_3$ at case # 4. This outcome was anticipated by the

Case Number	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
1	0.199	0.101	0.314
3	0.147	0.013	0.162
4	0.140	0.136	0.827
6	0.071	0.077	0.021
7	0.017	0.402	0.095

Table 2. Average improvement in the precision of the three predicted responses from the multi-response linear model over 500 simulations.

profile-based sensitivity results discussed above. According to the marginal sensitivity results, $\hat{y}_3$ would be expected to experience the largest improvement in its precision at case # 6; however, Table 2 shows the least average improvement in the precision of $\hat{y}_3$ at this case. Additionally, $\hat{y}_1$ experienced the greatest improvement in its precision at case # 1 as was indicated by the values of $\widehat{PSC}_i^{\,1}$, while $\hat{y}_2$ achieved the greatest improvement in its precision at case # 7, the location indicated by the values of $\widehat{PSC}_i^{\,2}$ values. These results could not be anticipated by the values of the marginal sensitivities for $\hat{y}_1$ and $\hat{y}_2$. The locations where the least improvement in the precisions of the three predicted responses occurred are the locations that were suggested by the profile-based sensitivity results and not by the marginal sensitivities.

The inconsistency between the indications provided by the marginal sensitivities and the improvements in the precision of the predicted responses can be attributed to the strong co-dependencies between the parameter estimates. The $\widehat{\mathbf{MSC}}_i(\mathbf{x}_0)$ measures do not account for simultaneous changes in the values of the parameter estimates when they are estimated jointly using the larger data set. However, $\widehat{\mathbf{PSC}}_i(\mathbf{x}_0)$ measures do account for the systematic co-dependencies among the parameter estimates and correctly identify the locations in the design space where the predicted responses are most sensitive to changes in the values of both parameter estimates.

4 Acknowledgments

The authors gratefully acknowledge the financial support of the American University of Sharjah, United Arab Emirates.

References

BATES, D.M. and WATTS, D.G. (1988): *Nonlinear Regression Analysis and Its Applications*, Wiley: New York.

BATES, D.M. and WATTS, D.G.(1987): A generalized Gauss-Newton procedure for multi-response parameter estimation. *SIAM Journal of Scientific and Statistical Computing 7(1), 49-55.*

BOX, G.E.P. and DRAPER, N. (1965): The Bayesian estimation of common parameters from several responses. *Biometrika 52, 355-365.*

OSBORNE, M.R. (1972): Some Aspects of Non-linear Least Squares Calculations. *Numerical Methods For Nonlinear Optimization*, Academic Press: London.

SULIEMAN, H., MCLELLAN, P.J. and BACON, D.W. (2004):A Profile-based approach to parametric sensitivity in multiresponse regression models. *Computational Statistics & Data Analysis 45, 721-740.*

SULIEMAN, H., MCLELLAN, P.J. and BACON, D.W. (2001): A Profile-Based Approach to Parametric Sensitivity Analysis of Nonlinear Regression Models. *Technometrics 43(4), 425-33.*

Part VI

Computational Methods in Official Statistics

The Birth Rate and the Marriage Rate in the Czech Republic in Years 1960-2006*

Josef Arlt[1], Markéta Arltová[1], and Jitka Langhamrová[2]

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *arlt@vse.cz*, *arltova@vse.cz*

[2] Department of Demography, Faculty of Informatics and Statistics, University of Economics Prague, W. Churchill sq. 4, 130 67 Prague, Czech Republic, *langhamj@vse.cz*

Abstract. The Czech Republic has undergone a profound political, economical and social transformation in the past few decades. The changes that have occurred in society showed strongly in the change in the demographic behaviour of society and the change in the general population climate. The transition to the market economy brought with it new opportunities for self-realisation, young people give preference to other values than the starting of a family at an early age. From the analysis of the time series it emerges that there is a long-run relationship between the birth rate and the marriage rate in the Czech Republic.

Keywords: birth rate, marriage rate, co-integration, VAR model

1 Analysis of the birth rate and the marriage rate

In the age composition of the population of the Czech Republic there are increasing signs of a long-term drop in the level of the birth-rate. Specific to the Czech Republic is the increase in the level of the birth rate already from the beginning of the forties during the period of Nazi occupation, then its decline in the late fifties, which is linked with the legalisation of artificial termination of pregnancy, the slight increase in the mid-sixties as a result of the passing of pro-birth measures and the subsequent decline as the result of the social-economic crisis. The numerically strongest generations in the population are those of 1974 and 1975. These are the children born in the first half of the seventies as a result of the passing of a series of measures promoting births. Since the nineties the Czech Republic is characterised by a significant decline in the number of births, when young people reacted very sensitively and strongly to the new changed political and social-economic situation in the country. At the present time the number of births is slightly rising.

* This article came into being within the framework of the long-term research project 2D06026, "Reproduction of Human Capital", financed by the Ministry of Education, Youth and Sport within the framework of National Research Program II.

In our paper we are concentrating on the analysis of the birth rate and the marriage rate connected with it in the Czech Republic from 1960 up to the present and trying to clarify the relations that apply among these indicators. For this purpose it is necessary to identify suitable model of these indicators. This model is than possible to use also for forecasting.

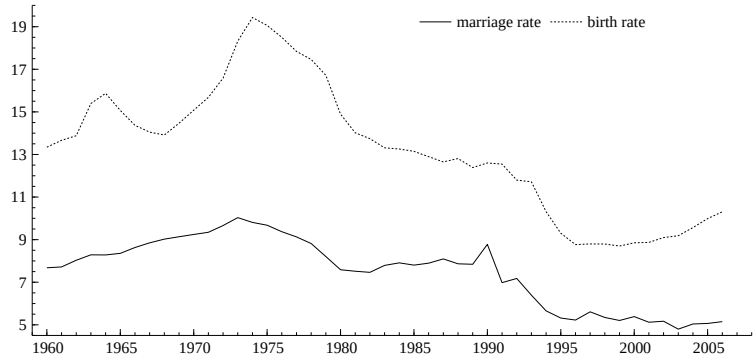


Fig. 1. The marriage and birth rate in the Czech Republic.

The relationship between the birth rate and the marriage rate in the period 1960-2007 is analyzed on the base of the yearly time series ('Pohyb obyvatelstva v Českých zemích 1785-2006'). We are aware of the fact that for modeling purposes these time series are short. This problem can be solved by analysis of monthly data. But from the preliminary investigation it follows that these time series are very volatile and extremely sensitive to exogenous interventions (above all the birth rate) which complicates the analysis considerably. In addition, some of time series we need for modeling (mid-year population, birth rate according to births order) are officially available only since 1991.

For the analysis we are dividing the time sequences into two parts; the first part will document the development of the time series in the years 1960-1989, i.e. in the period of socialist economy directed by the state, and the second part in the period of 1990-2006, i.e. after the transition to a market economy.

1.1 The period of 1960-1989

The time series of the marriage rate and especially of the birth rate have a rising tendency at the beginning, caused to a considerable extent by the pro-population economic measures of the socialist government. The reason for the marked growth of the birth rate at the beginning of the sixties might be the introduction of the differentiation in the retirement age for women and

reductions in rent depending on the number of children or the prolonging of maternity leave.

The effect of the pro-population measures was short-lived. It was only further measures at the end of the sixties: the increasing of child benefit (1968), the maternity allowance (1970), the increase in benefits for second and further children (1973) and the introduction of loans for young couples that brought about a further rise in the birth rate. All this, together with the arrival of the strong post-war generations of women at the age of highest fertility, brought with it an increase in the birth rate. The climax of this period is 1974. In further years both time series showed a long-term decline.

From the graph and from the ADF test of unit root it follows that both the birth rate and the marriage rate are of I(1) type. It is also evident that both time series develop similarly. This fact is also confirmed by the multi-dimensional analysis of the time series (Arlt and Arltová, 2007). We identified the model VAR(3)[1] (Table 1.).

Table 1.

VAR(3) model Sample: 1960-1989		
Variable	Birth rate	Marriage rate
Birth rate(-1)	1.436938	-0.175496
t-value	[10.3641]	[-1.89149]
Birth rate(-2)	-0.660253	0.169834
t-value	[-3.19574]	[1.22838]
Birth rate(-3)	-0.019153	-0.082592
t-value	[-0.16419]	[-1.05802]
Marriage rate(-1)	1.253393	1.494732
t-value	[3.85003]	[6.86097]
Marriage rate(-2)	-2.179356	-0.372328
t-value	[-3.51562]	[-0.89752]
Marriage rate(-3)	1.347481	0.031411
t-value	[2.91297]	[0.10147]
D1	1.034794	0.200854
t-value	[4.03692]	[1.17090]
Resid. Corr. Matrix		
	Birth rate	Marriage rate
Birth rate	1.000000	0.029291
Marriage rate	0.029291	1.000000

From this model[2] it emerges that the birth rate depends on the birth rate in the preceding two years and on the marriage rate in the preceding three years. The marriage rate depends only on the marriage rate in the preceding year.

From the residual correlation matrix of VAR(3) model follows that the linear dependence of the birth rate and marriage rate in the same year is very week ($r = 0.02929$). From this we may deduce that a wedding in a given year is not accompanied by the birth of a child in the same year.

[1] EViews6 was used for all computations

[2] D1 is dummy variable, 1963 (1), 1973 (1)

Johansen´s test (Johansen, 1991) indicates the co-integration between birth rate and marriage rate and the long-run relation (Engle and Granger, 1987) takes the form

$$c = Birth\,rate_t - 1.6909 Marriage\,rate_t.$$

The long-run development of the birth rate is in direct relationship to the long-run development of the marriage rate.

1.2 The period of 1990-2006

At the beginning of this period both time series continue to decline. We consider the most important cause to be the reaction to the new social-economic situation, which is accompanied by the deep drop of fertility, the deferring of the birth of a child until a later age and an increase in work engagement (Koschin, Fiala, Langhamrová, and Roubíček, 2001). The economic reforms introduced are bringing a rise in unemployment and rent and home ownership is becoming inaccessible to a large proportion of young couples. The government is also cancelling a number of benefits for families with children. For instance the stopping of the advantageous loans for young couples (as of 1.1.1991) is also reflected very strongly in the time series of the marriage rate with a marked increase in 1990, compensated by a strong decline in the following year ('Populační vývoj České republiky 1990-2002').

From 2000 the birth rate has shown a rising tendency that is not, however, accompanied by an increase in the marriage rate. Strong years of women of fertile age are gradually reaching the age of greatest fertility. Children are beginning to be born to women who have left the birth of the first child to a later age. According to the opposite development of the both time series it may be deduced that to a considerable extent the phenomenon is beginning to appear that is well known from advanced Western economies, this being the significant increase in the number of children born outside wedlock (in 1990 this proportion was 8.58 % and in 2006 already 33.34 %). Young people are living increasingly frequently in factual marriages.

From the graphic expression and from the result of ADF test it follows that both time series are of I(1) type. In the first period, roughly up to 1996, the two time series develop similarly. After 1996 there is a turning point when, in spite of the declining number of weddings, the birth rate is beginning to rise. Through multi-dimensional analysis of the time series we identified the VAR(2) model (Table 2.).

From this model[3] it follows that the birth rate depends on the development of the birth rate in the preceding year and on the marriage rate in the preceding two years. The marriage rate depends only on the marriage rate of the preceding year. The residual correlation matrix of VAR(2) model indicates the strong linear relationship of the birth rate and marriage rate in the

[3] D2 is dummy variable, 1990 (1)

Table 2.

VAR(2) model Sample: 1990-2006		
Variable	Birth rate	Marriage rate
Birth rate(-1)	1.383480	0.007026
t-value	[6.20673]	[0.02623]
Birth rate(-2)	-0.403440	-0.173065
t-value	[-1.91113]	[-0.68216]
Marriage rate(-1)	0.620609	1.116360
t-value	[2.30367]	[3.34875]
Marriage rate(-2)	-0.572507	0.161248
t-value	[-2.63585]	[0.61804]
D2	-0.842965	-2.025730
t-value	[-1.45139]	[-2.90261]
Resid. Corr. Matrix		
	Birth rate	Marriage rate
Birth rate	1.000000	0.72311
Marriage rate	0.72311	1.000000

same year ($r = 0.72311$). From this we may deduce that a wedding in a given year is accompanied by the birth of a child in the same year. By Johansen´s test of co-integration no long-run relationship was identified in this period.

2 Analysis of the birth rate according to births order and the marriage rate

In spite of the fact that the birth rate in the Czech Republic has recorded in many respects very marked changes, the distribution of the children born according to birth order has remained relatively stable in recent years. In 2006 the percentage of first-born babies was 48.96 % and that of second-born infants was 36.86 %. The development of the numbers of the first and the second children born copies the development of the total number of babies born. The drop in the birth rate in years 1993-1996 concerned children of all birth orders, but it affected the first-born children to a greater degree. This is linked to the establishment of a family as such. The marriage rate is at a relatively low level in the Czech Republic. Weddings are postponed to a higher age and part of the population rejects marriage completely. Almost a third of all children are born out of marriage and in the case of the first-born children it was as high as 41.6 %. The strongest growth in the number of babies born was at the time of the baby-boom in the first half of the seventies, when the numbers of not only first-born, but also second-born children increased.

The time series of the marriage rate and birth rate in the first order (*birth rate1*) develop in a similar manner to begin with; up to the first half of the seventies they show a tendency for slow growth as opposed to the time series of the birth rate in the second order, which first shows similar development, drops rapidly around 1965 and around 1970 begins a rapid growth. In 1974 the birth rate in the second order (*birth rate2*) is almost at the same level as the birth rate in the first order. Up to the beginning of the eighties this

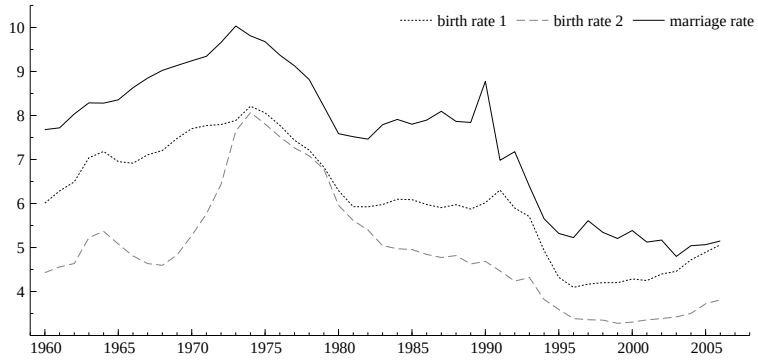


Fig. 2. The marriage and birth rate according to births order in the Czech Republic.

growth, evoked by the pro-population measures, is compensated by a rapid decline. The whole of the eighties are characterised by roughly constant development of the birth rate in the first order, a continuing declining tendency in the second order and a slight increase in the marriage rate. After 1989 the marriage rate and the birth rate in the first and the second order drop sharply and from the second half of the nineties their development begins to deviate slightly.

2.1 The period of 1960-1989

All the time series analysed are of I(1) type. Their relationship is catched by VAR(2) model (Table 3.).

Table 3.

VAR(2) model Sample: 1960-1989			
Variable	Birth rate1	Birth rate2	Marriage rate
Birth rate1(-1)	1.051860	-0.514326	0.162014
t-value	[6.11902]	[-1.47076]	[0.58786]
Birth rate1 (-2)	-0.128496	0.678774	0.042746
t-value	[-0.76078]	[1.97605]	[0.15791]
Birth rate2(-1)	-0.021582	1.535390	-0.205616
t-value	[-0,26352]	[9.21049]	[-1.56481]
Birth rate2(-2)	-0.026254	-0.711759	0.051196
t-value	[-0.28079]	[-3.74216]	[0.34153]
Marriage rate(-1)	0.654554	0.513054	1.317090
t-value	[4.83066]	[1.86159]	[6.06639]
Marriage rate(-2)	-0.563051	-0.529356	-0.379636
t-value	[-3.60007]	[-1.66412]	[-1.51430]
Resid. Corr. Matrix			
	Birth rate1	Birth rate2	Marriage rate
Birth rate1	1.000000	0.28526	0.25490
Birth rate2	0.28526	1.000000	0.29397
Marriage rate	0.25490	0.29397	1.00000

From this model it follows that the birth rate in the first order depends on the birth rate in the first order in the preceding year and on the marriage rate in the preceding two years. The birth rate of the second order depends on the birth rate of the second order in the preceding two years. The marriage rate depends only on the marriage rate in the preceding year.

The residual correlation matrix indicates a weaker linear relationships a) the birth rate in the first order and in the second order ($r = 0.28526$), b) the birth rate in the first order and the marriage rate ($r = 0.25490$), c) the birth rate in the second order and the marriage rate ($r = 0.29397$). By co-integration analysis no long-run relationship between the time series was indicated.

2.2 The period of 1990-2006

The time series investigated in this period are also of I(1) type. The relations between the time series are catched by VAR(2) model (Table 4.).

Table 4.

VAR(2) model Sample: 1990-2006			
Variable	Birth rate1	Birth rate2	Marriage rate
Birth rate1(-1)	1.511673	0.698800	0.567018
t-value	[4.57725]	[3.51781]	[0.69005]
Birth rate1(-2)	-0.909006	-0.482513	-1.953842
t-value	[-2.89109]	[-2.55139]	[-2.49760]
Birth rate2(-1)	-0.047366	0.078199	0.234423
t-value	[-0.07541]	[0.20697]	[0.14999]
Birth rate2(-2)	0.827128	0.888130	3.251660
t-value	[1.44051]	[2.57155]	[2.27609]
Marriage rate(-1)	0.219001	0.105553	0.151070
t-value	[1.81992]	[1.45831]	[0.50457]
Marriage rate(-2)	-0.379381	-0.317764	-0.082782
t-value	[-2.64399]	[-3.68181]	[-0.23188]
C	-0.029773	0.301500	-1.080333
t-value	[-0.05406]	[0.91018]	[-0.78843]
Resid. Corr. Matrix			
	Birth rate1	Birth rate2	Marriage rate
Birth rate1	1.000000	0.764965	0.460084
Birth rate2	0.764965	1.000000	0.885446
Marriage rate	0.460084	0.885446	1.000000

From this model it follows that the birth rate in the first order depends on the birth rate of the first order in the preceding two years and on the marriage rate two years before. The birth rate in the second order depends on the birth rate of the first order in the preceding two years and on the birth rate in the second order and the marriage rate two years before. The marriage rate depends on the birth rate of the first and the second order two years before.

The residual correlation matrix indicates the following linear relationships a) very strong for the birth rate of the first order and the birth rate of the

second order ($r = 0.76497$), b) medium strong for the birth rate of the first order and the marriage rate ($r = 0.46008$), c) very strong for the birth rate of the second order and the marriage rate ($r = 0.88545$). The Johansen´s test identified one co-integration relationship between the first order birth rate, the second order birth rate and the marriage rate, it has form

$$c = Birth\,rate1_t - 1.4904 Birth\,rate2_t - 0.02223 Marriage\,rate_t + 0.87879.$$

3 Conclusion

The Czech Republic has undergone a profound political, economical and social transformation in the past few decades. The changes that have occurred in society showed strongly in the change in the demographic behaviour of society and the change in the general population climate.

From the analysis of time series it emerges that in years 1960-1989 there was a long-run relationship between the birth rate and the marriage rate in the Czech Republic. We do not reach the same conclusions, however, if we include in the analysis the birth rate according to the order of the children born. It may be assumed that this is the consequence of the influence of the pro-population measures in the seventies, especially on the numbers of second children born. Children were born mainly in the years following the wedding.

The transition to the market economy brought with it new opportunities for self-realisation, young people give preference to other values than the starting of a family at an early age. The state is gradually restricting the support for families with children and social securities are being reduced. No long-run relationship was identified among the time series in this period. If we included the birth rate according to the order of the children born, then a long-run relationship was emerged.

References

ARLT, J. and ARLTOVÁ, M. (2007): *Ekonomické časové rady.* Grada Publishing, Prague.

ENGLE, R.F. and GRANGER, C. W. J. (1987): Cointegration and Error Correction: Representation, Estimation and Testing. *Econometrica 55, 251-276.*

JOHANSEN, S. (1991): Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica 59, 1551-80.*

KOSCHIN, F., FIALA, T., LANGHAMROVÁ, J., ROUBÍČEK, V. (2001): Fertility in the Czech Republic in the Nineties: University of Economics Prague.

Pohyb obyvatelstva v Českých zemích 1785-2006. ČSÚ, Praha. (www.czso.cz).

Populační vývoj České republiky 1990-2002. Katedra demografie a geodemografie: Přírodovědecká fakulta, Univerzita Karlova v Praze, 2002.

Improved Efficient Mean Estimation in Incomplete Data Using Auxiliary Information

Waqas Ahmed Malik and Ali Ünlü

Department of Computer Oriented Statistics and Data Analysis, University of Augsburg, Germany
malikwaqasahmed@gmail.com, ali.uenlue@math.uni-augsburg.de

Abstract. Incomplete data are a common feature of medical studies, agricultural experiments, and socio-economic investigations. Missing data may lead to substantial biases in analyses if they are not taken into consideration.

In this paper, we consider the problem of estimating the population mean of a study characteristic when some observations in the sample are missing at random and the population mean of an auxiliary characteristic is unknown. Two estimators are proposed and compared to the usual mean estimator and estimators proposed by Toutenburg and Srivastava. It is shown theoretically that our estimators are unbiased and they are more efficient than alternative ones. This is also confirmed by simulation study.

Keywords: mean estimation, missing data, auxiliary information, improved efficiency, unbiasedness

1 Introduction

Missing data are a common problem in almost all surveys, for example in medical, agricultural and socio-economic data, and in opinion polls. Due to variety of reasons, for a fraction of the subjects, either no data at all are available or information on one or more variables is missing. Missing data can contribute to biases in estimation and make results unreliable.

In a perfect world, a survey has no missing data; all selected units participate and provide all requested information. However, reality is very different. Missing data due to some accidental loss of data are a normal although undesirable feature of any survey. The most frequently used method to compensate for missings is imputation (e.g., Little and Rubin (1987)). But this method manipulates the original information and it has a large effect on the results; especially in medical and agriculture data. In many empirical studies, they simply discard the subjects with missing information in any attribute and employ the standard inference procedure. As a result, much useful information is lost.

While assuming that the deleted observations may contain valuable information, an alternative approach is to try to improve the precision of the

estimators by including all cases available for their calculation rather than deleting the incomplete cases. The precision of estimators can be improved significantly by using correlated auxiliary information. Indirect estimation methods are comprehensible techniques for the estimation of population mean when an auxiliary characteristic correlated with the study characteristic is available (e.g., Sukhatme et al. (1984) or Singh (2003)). These methods of estimation assume that sample data contain no missing observation.

Some authors have defined indirect estimators when a sample is drawn with simple random sampling without replacement, when some observations are missing and the population mean of the auxiliary characteristic is available (e.g., Tracy and Osahan (1994), Toutenburg and Srivastava (1998), and Rueda and Gonzalez (2004)).

In this paper, we will consider the situation when both the assumptions are violated simultaneously, that is, some observations are missing in study or auxiliary characteristics, and the population mean of the auxiliary variable is unknown. We propose two new estimators for the mean of the study variable, using all available information of study and auxiliary characteristics.

2 Mean estimation using auxiliary information in missing data

Consider a finite population $U = (U_1, U_2, \ldots, U_N)$ of size N from which a random sample s of size n is drawn according to simple random sampling without replacement. Let (x_i, y_i) be the values of unit U_i of the auxiliary characteristic x and the study characteristic y for the population U with population means $\bar{X}$ and $\bar{Y}$ respectively. We want to estimate $\bar{Y}$ on the basis of the units in random sample s from U.

When all observations are available and the population mean $\bar{X}$ of the auxiliary characteristic is known, a well-known regression method of estimation provides the following estimator of $\bar{Y}$.

$$\hat{\bar{Y}}_{reg} = \bar{y} + b(\bar{X} - \bar{x}) \tag{1}$$

where $\bar{y}$ and $\bar{x}$ are the means of the sample observations on the study and auxiliary characteristic respectively, and b is the sample regression coefficient.

The estimation procedure (1) does not work when some of the observations are missing. Assuming that some observations are missing on either the study characteristic or the auxiliary characteristics, or both, the only possibility of using (1) is to discard the incomplete pairs of observations and consider only the complete pairs. This means that the actual sample size is less than the planned one, biases in estimation arise and sampling variances become larger (Kalton and Kasprzyk (1986)).

It is assumed that a set of $n - p - q - k$ complete observations (x_1, y_1), $(x_2, y_2), \ldots, (x_{n-p-q-k}, y_{n-p-q-k})$ on the selected units in the sample is avail-

able. Apart from these, a set of p observations $x_1^*, x_2^*, \ldots, x_p^*$ on the x characteristic in the sample are available but the corresponding y characteristic are missing. Similarly, q observations $y_1^{**}, y_2^{**}, \ldots, y_q^{**}$ of the y characteristic in the sample are available but the associated values on the x characteristic are missing. Furthermore, there are k sampling units on which observations on both the study and auxiliary characteristics are missing. The numbers p, q and k are assumed to be random.

Thus, the information available from the sample s has the following structure (M stands for 'Missing'):

$$\begin{array}{cccccc}
 & & \overbrace{\qquad\qquad\qquad}^{p} & & & \overbrace{\quad}^{k} \\
y_1 \ldots y_{n-p-q-k} & & M \quad \ldots \quad M & & y_{n-q-k+1} \ldots y_{n-k} & M \ldots M \\
\underbrace{x_1 \ldots x_{n-p-q-k}}_{n-p-q-k} & & x_{n-p-q-k+1} \ldots x_{n-q-k} & & \underbrace{M \quad \ldots \quad M}_{q} & M \ldots M
\end{array}$$

With this structure consider four disjoint sets of units from the sample s:

$$\begin{aligned}
s_1 &= \{i \in s : x_i, y_i \text{ are available}\},\\
s_2 &= \{i \in s : x_i \text{ are available, but } y_i \text{ are not available}\},\\
s_3 &= \{i \in s : y_i \text{ are available, but } x_i \text{ are not available}\},\\
s_4 &= \{i \in s : x_i,\ y_i \text{ are both not available}\}.
\end{aligned}$$

When the population mean of the auxiliary characteristic is not available, it is customary to make the use of a large preliminary sample for finding an estimate of it. But in many scientific and medical experiments taking a large preliminary sample is not possible. When some observations are missing and the mean of the auxiliary characteristic is not available, Toutenburg and Srivastava (2003) suggested the four ratio-type estimators.

Motivated by Toutenburg and Srivastava (2003), we propose two unbiased estimators when some observations are missing and also the mean of the auxiliary characteristic is not available:

$$\hat{\bar{Y}}_{w1} = \frac{(n-p-q-k)\bar{y} + q\bar{y}^{**}}{n-p-k} + b\Big[\frac{(n-p-q-k)\bar{x} + p\bar{x}^{*}}{n-q-k} - \bar{x}\Big], \quad (2)$$

$$\hat{\bar{Y}}_{w2} = \bar{y} + b\Big[\frac{(n-p-q-k)\bar{x} + p\bar{x}^{*}}{n-q-k} - \bar{x}\Big], \quad (3)$$

where

$$\bar{x} = \frac{\sum x_i}{n-p-q-k},\ \bar{y} = \frac{\sum y_i}{n-p-q-k},\ \bar{x}^* = \frac{\sum {x_i}^*}{p},\ \bar{y}^{**} = \frac{\sum {y_i}^{**}}{q},$$

and b is the sample regression coefficient in (1). It may be observed that the estimator $\hat{\bar{Y}}_{w1}$ utilizes all information available from incomplete observations, while $\hat{\bar{Y}}_{w2}$ ignores all cases in which the auxiliary characteristic x is missing.

On the lines of Toutenburg and Srivastava (2003), we can solve equations (2) and (3) up to first order approximation and get biases and variances of $\hat{\bar{Y}}_{w1}$ and $\hat{\bar{Y}}_{w2}$ respectively as

$$Bias\big(\hat{\bar{Y}}_{w1}\big) = 0, \tag{4}$$

$$Var\big(\hat{\bar{Y}}_{w1}\big) = S_y^2\big[f_{p+k} + \rho^2\big(f_{p+q+k} - f_{q+k}\big)\big], \tag{5}$$

and

$$Bias\big(\hat{\bar{Y}}_{w2}\big) = 0, \tag{6}$$

$$Var\big(\hat{\bar{Y}}_{w2}\big) = S_y^2\big[f_{p+q+k} - \rho^2\big(f_{p+q+k} - f_{q+k}\big)\big]. \tag{7}$$

Here, S_y^2 is the variance of y, ρ the population correlation between x and y, and

$$f_s = E\Big(\frac{1}{n-s}\Big) - \frac{1}{N},$$

where the expectation operator E refers to all possible values of the non-negative integer-valued random variable (standard deviation) s, and

$$f_{p+q+k} \geq f_{p+k},$$
$$f_{p+q+k} \geq f_{q+k},$$
$$f_p \begin{cases} > f_q & : \quad p > q \\ < f_q & : \quad p < q. \end{cases}$$

3 Theoretical comparison with alternative estimators

In order to compare the performance of the proposed estimators under the criteria of the bias and mean squared error or variance to the first order of approximation, we consider the usual mean estimator $\hat{\bar{Y}}_0 = \bar{y}$ and four estimators $\hat{\bar{Y}}_i\,(i = 1, \ldots, 4)$ of Toutenburg and Srivastava (2003).

The biases of above estimators are given by

$$Bias\left(\hat{\bar{Y}}_0\right) = 0, \tag{8}$$

$$Bias\left(\hat{\bar{Y}}_1\right) = S_y^2 \frac{C_x}{C_y}\left[\frac{C_x}{C_y} - \rho\right]\big(f_{p+q+k} - f_{q+k}\big), \tag{9}$$

$$Bias\left(\hat{\bar{Y}}_2\right) = S_y^2 \frac{C_x}{C_y}\rho\big(f_{p+q+k} - f_{q+k}\big), \tag{10}$$

$$Bias\left(\hat{\bar{Y}}_3\right) = S_y^2 \frac{C_x^2}{C_y^2}\big(f_{p+q+k} - f_{q+k}\big), \tag{11}$$

$$Bias\left(\hat{\bar{Y}}_4\right) = 0, \tag{12}$$

where C_x and C_y are the coefficients of variation of the x and y variables, respectively.

It can be seen from the above that the proposed estimators are unbiased but $\hat{\bar{Y}}_1$, $\hat{\bar{Y}}_2$ and $\hat{\bar{Y}}_3$ are biased estimators.

The Variance or MSE of the alternative estimators are given by

$$Var\left(\hat{\bar{Y}}_0\right) = S_y^2 f_{p+q+k}, \tag{13}$$

$$MSE\left(\hat{\bar{Y}}_1\right) = S_y^2\left[f_{p+q+k} + \frac{C_x}{C_y}\left(\frac{C_x}{C_y} - 2\rho\right)\left(f_{p+q+k} - f_{q+k}\right)\right], \tag{14}$$

$$MSE\left(\hat{\bar{Y}}_2\right) = S_y^2\left[f_{p+q+k} + \frac{C_x}{C_y}\left(\frac{C_x}{C_y} + 2\rho\right)\left(f_{p+q+k} - f_{q+k}\right)\right], \tag{15}$$

$$MSE\left(\hat{\bar{Y}}_3\right) = S_y^2\left[f_{p+k} + \frac{C_x^2}{C_y^2}\left(f_{p+q+k} - f_{q+k}\right)\right], \tag{16}$$

$$MSE\left(\hat{\bar{Y}}_4\right) = S_y^2\left[f_{p+k} + \frac{C_x^2}{C_y^2}\left(f_{p+q+k} - f_{q+k}\right)\right]. \tag{17}$$

Now looking at the expressions for Var and MSE, it can be observed from Equations (7), (13), (14) and (15) that the proposed estimator $\hat{\bar{Y}}_{w2}$ will always be more efficient than $\hat{\bar{Y}}_0$, $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_2$, while $\hat{\bar{Y}}_{w1}$ will be more efficient than $\hat{\bar{Y}}_{w2}$ when

$$\rho^2 < \frac{\left(f_{p+q+k} - f_{p+k}\right)}{2\left(f_{p+q+k} - f_{q+k}\right)}. \tag{18}$$

From Equations (5) and (13), we find that $\hat{\bar{Y}}_{w1}$ is a more efficient estimator than $\hat{\bar{Y}}_0$ when

$$\rho^2 < \frac{f_{p+q+k} - f_{p+k}}{f_{p+q+k} - f_{q+k}}, \tag{19}$$

and $\hat{\bar{Y}}_{w1}$ is more efficient than $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ when

$$\rho^2 - \frac{C_x^2}{C_y^2} + 2\rho\frac{C_x}{C_y} < \frac{f_{p+q+k} - f_{p+k}}{f_{p+q+k} - f_{q+k}}. \tag{20}$$

It is seen from (5), (16) and (17) that $\hat{\bar{Y}}_{w1}$ is better than $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ when

$$\rho^2 < \frac{C_x^2}{C_y^2}. \tag{21}$$

It is interesting to observe from (21) that it does not depend upon the missing rate.

In a similar manner, comparing $\hat{\bar{Y}}_{w2}$ with $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$, we can see from (7), (16) and (17) that $\hat{\bar{Y}}_{w2}$ is a more efficient estimator than $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ when

$$\rho^2 + \frac{C_x^2}{C_y^2} < \frac{f_{p+q+k} - f_{p+k}}{f_{p+q+k} - f_{q+k}}. \tag{22}$$

4 Simulation study

A simulation study was carried out to investigate the behaviour and relative efficiency of the proposed estimators compared to alternative ones. The simulations were carried out using `Microsoft VBA`[1] on the following three empirical populations:

Accidents (F. B. S. (2005)): $N = 125$ cities; y: number of accidents in the city in 2004; x: number of vehicles in the city in 2004.

Labour Force (Valliant et al. (2003)): $N = 474$ persons; y: usual amount of weekly wages; x: usual number of hours worked per week.

Electricity Production (United Nations (2003)): $N = 98$ countries; y: production of electricity of the country in year 2001; x: population of a country in year 2001.

We realized the simulation using the following algorithm.

Step 1: Draw a sample of size n according to the procedure of simple random sampling without replacement.

Step 2: Set the missingness rates p, q and k. (The missingness rate for x, for instance, is the proportion of missing data in x.)

Step 3: Eliminate from the sample, p elements of the study characteristic, q elements of the auxiliary characteristic, and k elements of both the characteristics randomly.

Step 4: Define the subsamples s_1, s_2, s_3 and s_4.

Step 5: Generate t =10,000 independent random sample of size n.

Step 6: Calculate: $\hat{\bar{Y}}_0$, $\hat{\bar{Y}}_1$, $\hat{\bar{Y}}_2$, $\hat{\bar{Y}}_3$, $\hat{\bar{Y}}_4$, $\hat{\bar{Y}}_{w1}$, $\hat{\bar{Y}}_{w2}$ for each sample t times.

Step 7: The simulated bias and MSE of $\hat{\bar{Y}}_r$ were calculated as

$$Bias = \frac{1}{t}\sum_{b=1}^{t}\left(\hat{\bar{Y}}_r^{(b)} - \bar{Y}\right),$$

$$MSE = \frac{1}{t}\sum_{b=1}^{t}\left(\hat{\bar{Y}}_r^{(b)} - \bar{Y}\right)^2,$$

where $\hat{\bar{Y}}_r^{(b)}$ is the value of $\hat{\bar{Y}}_r$ for the b^{th} simulation run, and where $r = 0, 1, 2, 3, 4, w1, w2$.

[1] Source files for all the computations in this section are freely available from the authors.

For all three populations, 10,000 samples of sizes almost 10, 20 and 30 percent of N and missing rate of approximately 20, 30 and 40 percent of n were taken, for instance in Table 1 with $n = 15$, the values of missing were $9, 11$ and 16 were taken.

One can see from Tables 1–3 that both of our proposed estimators performed better than all of the alternative ones. Consistently, over different sample sizes and missing rates, the efficiences of $\hat{\bar{Y}}_{w1}$ and $\hat{\bar{Y}}_{w2}$ were higher than the usual mean estimator and estimators proposed by Toutenburg and Srivastava (2003).

n	p	q	k	$\hat{\bar{Y}}_0$	$\hat{\bar{Y}}_1$	$\hat{\bar{Y}}_2$	$\hat{\bar{Y}}_3$	$\hat{\bar{Y}}_4$	$\hat{\bar{Y}}_{w1}$	$\hat{\bar{Y}}_{w2}$
15	2	2	1	100.00	111.34	71.31	97.42	97.42	114.77	112.71
15	2	3	1	100.00	108.79	75.45	99.34	99.34	118.14	109.98
15	3	4	2	100.00	117.41	62.65	105.32	105.32	122.33	119.12
28	2	3	1	100.00	109.12	77.11	99.01	99.01	112.01	110.21
28	3	4	2	100.00	113.22	68.23	101.98	101.98	117.91	114.94
28	4	6	3	100.00	120.39	61.35	103.71	103.71	125.45	122.85
36	3	3	1	100.00	109.13	75.31	99.62	99.62	114.77	112.69
36	3	4	2	100.00	115.89	68.01	102.39	102.39	117.94	116.72
36	5	6	3	100.00	124.42	57.28	103.12	103.12	127.18	125.88

Table 1. Relative efficiency of estimators for population *Accidents.*

n	p	q	k	$\hat{\bar{Y}}_0$	$\hat{\bar{Y}}_1$	$\hat{\bar{Y}}_2$	$\hat{\bar{Y}}_3$	$\hat{\bar{Y}}_4$	$\hat{\bar{Y}}_{w1}$	$\hat{\bar{Y}}_{w2}$
48	4	4	2	100.00	102.29	93.74	108.85	108.85	110.12	109.49
48	5	7	3	100.00	103.04	91.74	118.73	118.73	121.49	119.37
48	7	9	4	100.00	104.73	87.90	126.92	126.92	127.69	105.98
95	7	7	4	100.00	102.11	94.08	108.49	108.49	110.80	109.19
95	10	13	5	100.00	103.47	91.01	118.53	118.53	121.93	119.63
95	14	17	7	100.00	104.92	87.56	127.10	127.10	129.33	128.20
140	8	10	6	100.00	101.81	94.84	109.47	109.47	111.46	110.85
140	15	17	10	100.00	103.58	90.45	117.71	117.71	119.07	118.69
140	18	23	15	100.00	104.97	87.56	127.13	127.13	129.13	128.15

Table 2. Relative efficiency of estimators for population *Labour Force.*

n	p	q	k	$\hat{\bar{Y}}_0$	$\hat{\bar{Y}}_1$	$\hat{\bar{Y}}_2$	$\hat{\bar{Y}}_3$	$\hat{\bar{Y}}_4$	$\hat{\bar{Y}}_{w1}$	$\hat{\bar{Y}}_{w2}$
20	2	2	1	100.00	107.08	79.44	104.44	104.44	113.97	109.82
20	2	3	1	100.00	107.42	79.27	112.32	112.32	118.25	114.83
20	3	4	2	100.00	113.61	69.09	115.81	115.81	121.75	117.40
30	2	3	1	100.00	104.76	84.57	108.53	108.53	114.44	110.56
30	3	4	2	100.00	107.95	77.75	110.79	110.79	116.42	112.33
30	4	6	3	100.00	111.85	70.95	119.60	119.60	125.54	121.30
40	3	3	2	100.00	106.07	81.80	109.68	109.68	115.59	111.34
40	4	6	2	100.00	108.39	76.81	115.59	115.59	119.93	117.64
40	6	7	3	100.00	113.92	68.25	117.48	117.48	121.83	119.91

Table 3. Relative efficiency of estimators for population *Electricity Production.*

5 Conclusion

Estimating population means in the presence of missing values is difficult. Information from auxiliary variable can be used to construct efficient estimators. In the case where the mean of the auxiliary variable is not known, several estimators have been proposed. In this article, we have presented two new estimators and shown that they have better bias and mean square error properties. Simulations from three real datasets support these results and show how much better the new estimators can be in practice. In any given situation, the choice of which of the two new estimators to use is governed by a condition we have found, involving the level of correlation between the variable of interest and the auxiliary variable.

References

FEDERAL BUREAU OF STATISTICS (2005): *Statistical Yearbook*. Pakistan.

KALTON, G. and KASPRZYK, D. (1986): The treatment of missing data. *Survey Methodology 12, 1-16*.

LITTLE, R.J.A. and RUBIN, D.B. (1987): *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

RUEDA, M. and GONZALEZ, S. (2004): Missing data and auxiliary information in surveys. *Computational Statistics 19, 551-568*.

SINGH, S. (2003): *Advanced Sampling Theory with Applications. How Michael Selected Amy*. Kluwer Academic Press, London.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S., and ASOK, C. (1984): *Sampling Theory of Surveys with Applications*. Iowa State University Press and Indian Society of Agricultural Statistics.

TOUTENBURG, H. and SRIVASTAVA, V.K. (1998): Estimation of ratio of population means in survey sampling when some observations are missing. *Metrika 48, 177-178*.

TOUTENBURG, H. and SRIVASTAVA, V.K. (2003): Efficient estimation of population mean using incomplete survey data on study and auxiliary characteristics. *Statistica 63, 223-236*.

TRACY, D.S. and OSAHAN, S.S. (1994): Random non-response on study variable versus on study as well as auxiliary variables. *Statistica 54, 163-168*.

UNITED NATIONS (2003): *Statistical Yearbook*. New York.

VALLIANT, R., DORFMAN, H. and ROYALL, M. (2000): *Finite Population Sampling and Inference*. John Wiley & Sons, New York.

Part VII

Data Mining and Machine Learning

ORF Length in Yeast is Negative Binomial – Why?

Anna Bartkowiak[1,2]

[1] Institute of Computer Science, University of Wrocław
Joliot-Curie 15, 50–383 Wrocław, Poland, *aba@ii.uni.wroc.pl*
[2] Wroclaw High School of Applied Informatics
Wejherowska 28, 54–239 Wrocław, Poland,

Abstract. The genetic code is inscribed into pieces of chromosomes called 'Open Reading Frames'. The code found in an ORF appears in triplets constituted by 20 amino-acids. We show that the ORF length is described by the negative binomial (NBIN) distribution, which was barely recognized so far. The NBIN model is a typical model for describing contagious or heterogeneous events. We show that in the case of yeast length this may be due to heterogeneity of frequency distribution of amino-acids constituting the ORFs.

Keywords: Open Reading Frame, ORF or gene length, heterogeneity, negative binomial distribution

1 Introduction

The genetic code is something great and mysterious developed by the Nature. Discovering of the way how the code is organized took quite a time, see Harrison et al. 2002, Wolfe et Li (2003), Bartkowiak et al. (2001), Luo et al. (2003), Hayes (2004). The code is inscribed into the cell's part called genome, which contains a number of chromosomes being carriers of the code (a yeast genome contains 16 chromosomes). A chromosome is composed of two strands. These are twisted together and code essentially the same information using four basic nucleotides (bases): A,C,T,G appearing in the two strands in complementary positions (A versus T, C versus G). The essential genetic information appears in some pieces of chromosomes; these pieces are called *Open Reading Frames*, short: ORFs. The code found in an ORF appears in triplets of the basic nucleotides. The triplets, in turn, designate 23 amino-acids. It happens that different triplets of bases may denote the same amino-acid. The begin of the coded information is indicated by the START codon (ATG, Methionine); the end of the coded information is indicated by one of three amino-acids (TAG, TAA, TGA, named: Opal, Ochre, Amber) playing the role of STOP codons. The true genetic information appearing between the START and STOP codons is coded by twenty amino-acids, which are listed in Table 1. The START codon ATG may appear also inside of the ORFs.

An obvious question in mathematical biology is: What is the probability distribution of ORF length counted in codons? Quite surprisingly, the exact answer to this natural question in scarcely reflected in the literature. There are hundreds of papers with the key word 'ORF length', but they consider this topic in terms of reporting some empirical distributions or their count statistics, which are derived under specific random walk or evolutionary models, or are concerned with changes of the ORF length under mutational pressure – as e.g., Luo et al. (2003), Mackiewicz et al. (1999), (2002), Polak et al. (2004), Cebrat et al. (2006), Bouaynaya and Schonfeld (2007). We found only reference indicating outright the NBIN p.d.f. of ORF length: Larsen and Krogh (2003). This paper considers the ORF length distribution in bacteria. The authors derive the ORF length distribution from a hidden Markov Model (HMM) architecture using looping codons submodels. It is not shown exactly, how to arrive to the NBIN distribution. Their illustration (Figure 5 in their paper) of the fit of the NBIN is based on a sparse histogram fit, without no formal test of the quality of fit; one may see by 'eye' that the fit is very rough.

In the following we will consider ORFs found in four yeast chromosomes, containing 855, 593, 348 and 513 ORFs appropriately. We will show, that for each of these chromosomes the distribution of ORF length (counted in number of amino-acids constituting the ORF) follows the negative binomial (NBIN) distribution, which will be confirmed both by graphs and by Kolmogorov-Smirnov test. The fit is surprisingly good. Why should ORF length be distributed according to the NBIN p.d.f? It is known that the NBIN distribution describes the probabilities of events which are contagious or form a heterogenous mixture (Lundberg, 1940, Hilbe, 2007). What kind of heterogeneity might lead to the NBIN distribution observed in the ORF length?

We know that the ORF length is obtained as the sum of counts from appearance of the 20 amino-acids, which may appear with various probabilities in the given ORF. For each ORF, apart from its length (the variable *len*), we have recorded also the vector $\mathbf{k} = (k_1, \ldots, k_{20})$ of *frequency counts* for the 20 amino-acids encountered in the given ORF. We will show that the set of the vectors $\mathbf{k}$ (gathered in a chromosome), may be subdivided into several more homogeneous subgroups with statistically different means of the variable *len*. This may be a hint for explaining the heterogeneity of the events leading to the NBIN model stated for ORF length.

This introduction constitutes Section 1 of our paper. In next section we describe the ORF data used in our elaboration. Section 3 recalls the properties of the negative binomial (NBIN) distribution and shows its perfect fit to the ORF length. The following Section 4 considers the distribution of frequency counts of amino-acids appearing in the ORF. The entire data set of the amino-acids frequency counts, after rescaling, is subdivided into several topologically more close clusters. where the variable *len* appears statistically differentiated.

A discussion of the results and some closing remarks follow in Section 5.

2 The data and their primary statistical characterization

The data were downloaded from: www.yeastgenome.org the 3rd April 2008. The downloaded files were in txt format and contained only the ORF data (without the inter-orf sequences). We downloaded data of four chromosomes: no. 4, no. 7, no. 11 and no. 113; thus we got for analysis four sets of data. The amino-acids in the downloaded files appeared in the one-letter code shown in Table 1.

1	2	3	4	5	6	7	8	9	10
A	R	N	D	C	Q	E	G	H	I
'Ala'	'Arg'	'Asn'	'Asp'	'Cys'	'Gln'	'Glu'	'Gly'	'His'	'Ile'
11	12	13	14	15	16	17	18	19	20
L	K	M	F	P	S	T	W	Y	V
'Leu'	'Lys'	'Met'	'Phe'	'Pro'	'Ser'	'Thr'	'Trp'	'Tyr'	'Val'

Table 1. Labels of the 20 amino-acids coding an ORFs content. The letters A, R, ... , Y, V constitute one-letter labels; the strings 'Ala', 'Arg', ... , 'Tyr', 'Val' constitute the 3-letter labels.

From the data we calculated for each ORF the frequencies of appearing of the 20 amino-acids (see – Table 1 for the labels of the 20 amino-acids). The sum of the frequency counts yielded the ORFs total length (*len*), which will be the main subject of our analysis. Next, we calculated some primary statistics of the variable *len*, like: n, the number of ORFs found in the analyzed set; *min* and *max* – the minimum and maximum value of *len* found in the given set; *median*, *mean*, and *variance* of *len*. The calculations were done using the statistical toolbox of Matlab.

Let n denote the number of ORFs found in the given chromosome, let $x_1, \ldots, x_n$ denote the observed ORF lengths, and let $\overline{x}$ denote the arithmetic mean of all the observations $x_1, \ldots, x_n$. Then:

$$skewness = \sum_{i=1}^{n}(x_i - \overline{x})^3 / (\sum_{i=1}^{n}(x_i - \overline{x})^2)^{3/2},$$
$$kurtosis = \sum_{i=1}^{n}(x_i - \overline{x})^4 / (\sum_{i=1}^{n}(x_i - \overline{x})^2)^2.$$

The obtained values are shown in Table 2.

One may see that for all the four investigated chromosomes the median is always greater then the mean. The positive skewness points to the right hand asymmetry of the distribution. The elevated kurtosis indicates for a heavy-tailed distribution (kurtosis for the normal distribution equals 3.0).

Chromosome	n	min	max	median	mean	variance	skewness	kurtosis
no. 4	855	24	3268	372.00	453.87	139759.26	2.11	10.33
no. 7	593	28	2672	368.00	453.91	135166.46	1.87	7.78
no. 11	348	29	4092	368.50	471.08	159452.16	3.27	24.08
no. 13	513	26	2123	383.00	461.72	127987.45	1.57	5.89

Table 2. Statistical characterization of the variable *len* (length of ORFs) as found in the four investigated chromosomes.

The variance is much larger as the mean – this points to an 'overdispersed' Poisson distribution.

Many empirical distributions, which appear as *overdispersed* when compared to the Poisson distribution, have been modelled using the negative binomial (NBIN) distribution. A closer look at the NBIN distribution is shown in next Section.

3 The negative binomial distribution and its fit to ORF length

The negative binomial (NBIN) probability distribution (probability mass function, p.m.f) is modelling contagious or nonhomogeneous events, appearing with a diversified probability. It has a long history – see e.g., wikipedia (2008), Hilbe (2007), Bartkowiakowa (1968), Lundberg (1940). It may be derived from several probabilistic models and may be parameterized in several ways. Below we will use the parametrization shown in wikipedia (2008). It describes the NBIN p.d.f. by the equation

$$f(k;r,p) = \binom{k+r-1}{k} p^r (1-p)^k, \quad \text{for } k = 0,1,2,\ldots. \tag{1}$$

Its parameters are: $r > 0$ (real) and $0 < p < 1$ (real). With $r \to \infty$, it converges to the $Poisson(k;\lambda)$ p.m.f. with parameter $\lambda = r\,(p^{-1}-1)$.

The mean, variance, skewness and kurtosis of the distribution are explicit functions of r and p (see wikipedia, 2008):

$$\mu = E\{k\} = r\,\frac{1-p}{p}, \quad \sigma^2 = E\{(k-\mu)^2\} = r\,\frac{1-p}{p^2}, \tag{2}$$

$$skewness = E\{\frac{(k-\mu)^3}{\sigma^3}\} = \frac{2-p}{\sqrt{r\,(1-p)}}, \tag{3}$$

$$kurtosis = E\{\frac{(k-\mu)^4}{\sigma^4}\} = 3.0 + \frac{6}{r} + \frac{p^2}{r(1-p)}. \tag{4}$$

The NBIN distribution may be also obtained as a compound distribution composed as Gamma-Poisson mixture defined for $k = 0, 1, 2, \ldots$ (the proof is shown in wikipedia 2008, see also Lundberg 1940) defined for $k = 0, 1, 2, \ldots$:

$$f(k; r, p) = \int_0^\infty Poisson(k|\lambda)\, Gamma(\lambda \,|\, r,\ \frac{(1-p)}{p}\ )\, d\lambda. \tag{5}$$

We have investigated the fit of the NBIN p.d.f. (1) to the variable *len* (ORF length) in the investigated four chromosomes. The parameters r and p were estimated by the Matlab function nbinfit using the ML (Maximum Likelihood principle). For all the four distribution the fit is perfect. The Kolmogorov-Smirnov test (Matlab function KSTEST) run at the $\alpha = 0.0005$ rejection level cannot reject the hypothesis that the investigated sample comes from the NBIN distribution. In parallel, we have also tested the hypothesis that the underlying sample comes from the geometrical distribution. The KSTEST test rejected this hypothesis for all four cases. The graphical fit of both distributions – for the ORFs from the 7th chromosome – is shown in Figure 1. One may notice there the perfect fit of the NBIN distribution and the inadequacy of the geometrical distribution.

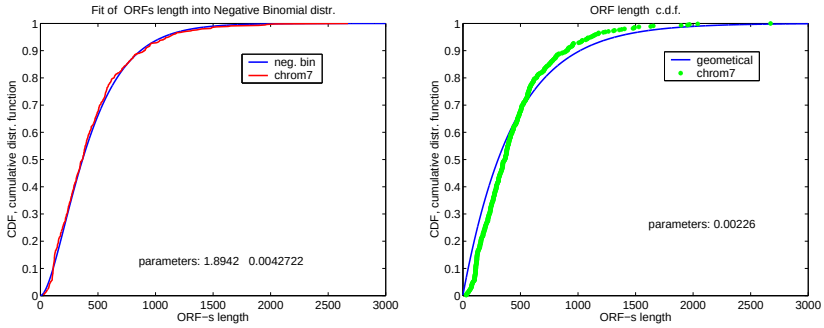


Fig. 1. ORF length. Cumulative probability distribution – observed and expected as: *Left:* Negative binomial distribution. *Right:* Geometrical distribution.

The estimates of the parameters r and p for the four distributions are:

	r	p	conf.interval for r	conf.interval for p
chr 4:	1.7376	0.0038	1.5740 – 1.9012	0.0035 – 0.0042
chr 7:	1.8100	0.0040	1.5948 – 2.0253	0.0035 – 0.0044
chr 11:	1.8529	0.0039	1.6040 – 2.1018	0.0034 – 0.0044
chr 13:	1.7974	0.0039	1.5739 – 2.0208	0.0034 – 0.0044

What might be the reason for the perfect fit of the NBIN distribution to the ORF length data? We will consider this question in next Section.

4 Seeking for homogenous subgroups of the amino-acids occurrences

We start from the fact that ORF length was obtained as the sum of occurrences of 20 amino-acids, which appeared in the ORF with probabilities $\pi_1, ..., \pi_{20}$. We ask now: Are these probabilities the same for all ORFs?

For each chromosome we have the records $(k_1, k_2, \ldots, k_{20})$ of the amino-acid occurrences for each ORF. By dividing each occurrence by the ORF length we obtain the proportions $(k_1/k_., \ldots, k_{20}/k_.)$, with $k_. = \sum_{i=1}^{n} k_i$. Converting these proportions to percentage (for our convenience) and putting them together into a two-dimensional data array A we obtain: $A = (a_{ij})$, $(i = 1, \ldots, n,\ j = 1, \ldots 20)$, with n denoting the number of ORFs in the chromosome. Formally, the data array A contains a set of multivariate data, whose rows may be viewed as data points located in R^{20}. The data may be subjected to multivariate analysis.

We will subdivide these records into several homogenous clusters containing similar subjects (a data row in A is now considered as a data point in R^{20}). We will apply for this purpose the *k-means method* combined with the *Davies-Boldin principle*, which will tells us, which is the reasonable number of clusters suitable for our data. Next we investigate the mean length of ORFs belonging to each cluster. A statement that ORF length differs statistically in the found clusters will authorize us to the belief that ORF length is a heterogeneos mixture.

The calculations were performed using the function kmeans_clusters from the SOM Toolbox for Matlab – see Vesanto et al. (2000). An exemplary output from the run is shown in Figure 2. On may see there the changes of the 'error' of clustering (designated as the within class squared deviations of the data points from the centroid of the cluster to which they belong) and the 'Davies-Bouldin' index, described in Jain and Dubes (1988). The 'error',

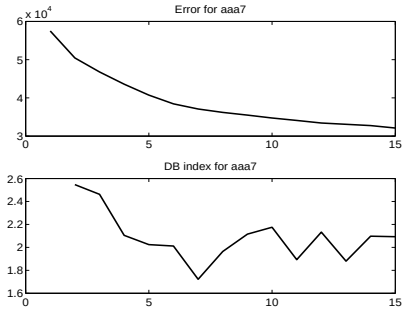


Fig. 2. Search for the right number of subgroups for ORFs from the 7th chromosome. *TOP:* Decay of the error with increasing number of clusters. *Bottom:* Changes of the Davies-Bouldin statistics. Seven clusters seems to be the best solution.

which in fact is the quantization error, exhibits with increasing number of clusters a systematic decay, which may be seen in the upper plot of Figure 2. On the opposite, the 'Davies-Bouldin' index should show – for data containing essential clusters – a minimum of the index for the number of essential clusters. Such a situation is depicted in the bottom plot of Figure 2.

Applying the above procedure we have subdivided the amino-acid data contained in the array A into k clusters (we assumed $k = 7, 10, 11, 12, 13$). For each cluster we found the corresponding values of the variable *len* denoting the ORF length corresponding to the subjects belonging to the found clusters. The values of *len* was transformed to yield another variables: *X1=sqrt(len)* and *X2=log(len)*. Both variable were subjected to the ANOVA1 test taking the obtained clusters as grouping variables. The high resulting F statistics rejects at a high significance level the hypothesis that there is no cluster effect.

Thus the ORF length can be viewed as a mixture of several clusters which have significantly different ORF lengths. Specificity of locations of short (≤ 150) and long (≥ 1200) ORFs is shown in Figure 3.

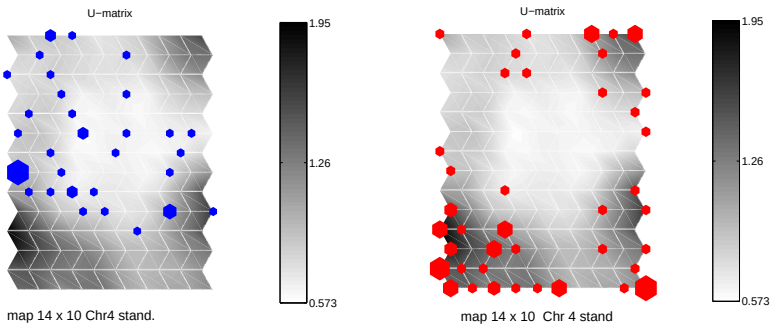


Fig. 3. Location of short (left exhibit) and long (right exhibit) ORFs in Kohonen self-organizing map. Notice that short ORFs (≤ 150) are located in dark regions at the extremes of the map, while longer ORFs (≥ 1200) are located in the center where distances between ORFs are small.

5 Closing remarks

We have shown that ORF length can be described the negative binomial distribution. This distribution may be derived from Poisson distributions with varying parameter λ.

An ORF is composed from 20 amino-acids. The analysis of their frequency distribution, as presented in Section 4, indicates that their distribution may be viewed as a mixture of several more homogeneous parts. ORF length in

these parts is significantly differentiated. Thus, we have shown that also ORF length may be modelled as a mixture of more homogeneous parts.

References

BARTKOWIAKOWA, A. (1968): Stochastic processes in biology and medicine (in Polish). *Listy Biometryczne/Biometrical Letters, No. 19–22, 1–62.*

BARTKOWIAK, A. (2001): An Attempt of Recognizing Genes in DNA sequences. *Book of Short Papers, CLADAG2001, Palermo, Italy, 205-208.*

BOUAYNAYA, N. and SCHONFELD, D. (2007): Protein communication system: Evolution and genomic structure. *Algorithmica 48, 375–397.*

CEBRAT; S., DUDEK, M.S. and MACKIEWICZ, P. (2006): *Modeling gene's length distribution in genomes. arXiv:q-bio/0607029v1 [q-bio.GN] 19 Jul, 1–7.*

HARRISON, P.M. et al. (2002): A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Research. 30 (1), 1083–1090.*

HAYES, B. (2004): Genome biology: on the genetic code. *American Scientist 92, 494.* http://www.americanscientist.org

HILBE, J. (2007): *The Negatibe Binomial Distribution.* Cambridge University Press

JAIN, A.K., and DUBES, R.C. (1988): *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, New Jersey

LARSEN, T.Sch. and KROGH, A. (2003): EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics. 1–15. Open access at Biomedcentral http://www.biomedcentral.com/info/publishing_adv.asp

LUNDBERG, O. (1940): *On Random Processes and Their Application To Sickness and Accident Statistics.* Almquist and Wicksells Boktryckeria, Uppsala.

LUO, L., LI, H. and ZHANG, L. (2003): ORF organization and gene recognition in the yeast genome. *Comparative and Functional Genomics 4, 318–328.*

MACKIEWICZ, P. et al. (1999): Mechanisms generating long-range correlation in nucleotide composition of the Borelia burgdorferi genome. *Physica A 273, 103–115.*

MACKIEWICZ, P. et al. (2002): Replication associated mutational pressure generating long-range correlation in DNA. *Physica A 314 646–654.*

POLAK, N. et al. (2004): How gene survival depends on their length. In: M. Bubak et al. (Eds): *ICCS, LNCS 3039, 694–699*, Springer.

VESANTO, J. et al. (2000) *SOM Toolbox for Matlab 5.* Som Toolbox Team, Helsinki University of Technology, Finland, Libella Oy, Espoo , 1–54. Downloadable from: http://www.cis.hut.fi/projects/somtoolbox.

WOLFE, K.H. and LI, W-H. (2003): Molecular evolution meets the genomic revolution. Review. *Nature Genetics supplement 33, March, 255–265.* Nature Publishing Group http://www-nature.com/naturegenetics.

WIKIPEDIA. http://en.wikipedia.org/wiki/Negative_binomial_distribution

Bench Plot and Mixed Effects Models: First Steps Toward a Comprehensive Benchmark Analysis Toolbox

Manuel J. A. Eugster and Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstrasse 33, 80539 München, Germany,
firstname.lastname@stat.uni-muenchen.de

Abstract. Benchmark experiments produce data in a very specific format. The observations are drawn from the performance distributions of the candidate algorithms on resampled data sets. In this paper we introduce new visualisation techniques and show how formal test procedures can be used to evaluate the results. This is the first step towards a comprehensive toolbox of exploratory and inferential analysis methods for benchmark experiments.

Keywords: benchmark experiments, visualisation, hypothesis tests

1 Introduction

In statistical learning, benchmark experiments are empirical experiments with the aim of comparing and ranking algorithms with respect to a certain performance measure. New benchmark experiments are published on almost a daily basis. Especially in the machine learning community benchmarking is the primary method of choice to evaluate new learning algorithms. However, there are surprisingly few publications on *how* to evaluate benchmark experiments. Some newer exceptions are Hothorn et al. (2005), Demsar (2006), Yildiz and Alpaydin (2006) and Hornik and Meyer (2007).

Hothorn et al. (2005) use the bootstrap method as a sampling scheme such that the resulting performance observations are iid and can be analyzed using standard statistical methods. However, their paper describes a general framework, not precise instructions for a concrete benchmark experiment. To use a metaphor, it describes how to cook in general, but contains no recipes for a nice dinner. Using the foundations laid out by the general framework, our goal is now to implement a toolbox of exploratory and inferential methods for the analysis of benchmark experiments.

Due to space restrictions, we cannot give a comprehensive overview of all our work in this direction in this paper. Hence, we chose to describe one new visualization technique (the benchplot), and how benchmark data can be seen as coming from a blocked design and analyzed as such (using mixed effects models) as examples. All computations are done using R (R

Development Core Team, 2007), the corresponding R functions are part of an R package for the analysis of benchmark experiments which is currently under development and will be released on CRAN later this year.

Following Hothorn et al. (2005), we set up a regression benchmark experiment with the mean squared error as loss function. Given a data set $\mathfrak{L} = \{z_1, \ldots, z_n\}$, we draw B learning samples using sampling with replacement

$$\mathfrak{L}^i = \{z_1^i, \ldots, z_n^i\}$$

for $i = 1, \ldots, B$ (bootstrap). Furthermore we assume that there are $K > 1$ candidate algorithms a_k $(k = 1, \ldots, K)$ available for the solution of the underlying problem. For each algorithm a_k the function $a_k(\cdot \mid \mathfrak{L}^b)$ is the fitted model based on the sample $\mathfrak{L}^b$. This function itself has a distribution $\mathcal{A}_k$ as it is a random variable depending on $\mathfrak{L}^b$:

$$a_k(\cdot \mid \mathfrak{L}^b) \sim \mathcal{A}_k(\mathfrak{L}),\ k = 1, \ldots, K$$

The performance of the candidate algorithm a_k when provided with the training data $\mathcal{L}^b$ is measured with the mean squared error function p (a scalar function):

$$p_{kb} = p(a_k, \mathfrak{L}^b) \sim \mathcal{P}_k = \mathcal{P}_k(\mathfrak{L})$$

The p_{kb} are samples drawn from the distribution $\mathcal{P}_k(\mathfrak{L})$ of the mean squared error of the algorithm k on the data set $\mathfrak{L}$. As we are not able to calculate p_{kb} analytically, we have to use the empirical analogue $\hat{p}_{kb}$ based on a test sample $\mathfrak{T}$. A common choice to define $\mathfrak{T}$ is in terms of out-of-bootstrap observations: $\mathfrak{T} = \mathfrak{L} \setminus \mathfrak{L}^b$. This leads to non-independent observations of the performance measure, but their correlation vanishes as n tends to infinity.

The first step is to analyse the benchmark experiment in an exploratory way. Based on findings in this step, the second step tests hypothesis of interest and yields an ordered ranking of the candidate algorithms.

To demonstrate the methods, we use an examplar benchmark study using the motorcycle data set, see Figure 1. The candidate algorithms used (with corresponding R functions in parenthesis) are linear regression (`lm`), nonlinear least-squares regression (`nls`), neural networks (`nnet`), regression trees (`rpart`), generalized additive models (`gam`), loess regression (`loess`) (all, e.g., in Venables and Ripley, 2002), and boosted generalized additive models (`gamboost`, Hothorn & Bühlmann, 2006) as candidate algorithms. In order to present an experiment with a wide variety of algorithm performances, we included linear regression, although the data are clearly nonlinear. The number of bootstrap samples B is 250.

2 Exploratory analysis

Common analyses of benchmark experiments consist of the comparison of the empirical performance measure distributions based on some summary

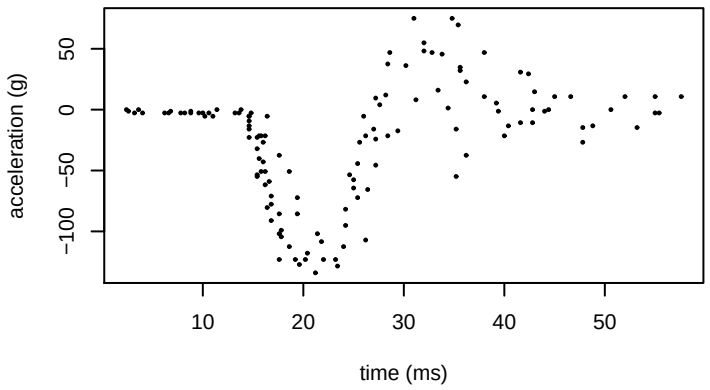


Fig. 1. The motorcycle data set (Silverman (1985)): time and head acceleration of a PTMO (post mortem human test object) after a simulated impact with motorcycles. The number of observations is $n = 133$.

	Mean	SD	95% CI	Median	IQR
`nnet`	1438.1	868.4	[−263.9, 3140.1]	977.2	1697.1
`lm`	2209.2	294.1	[1632.8, 2785.5]	2209.8	384.1
`rpart`	812.4	181.2	[457.2, 1167.6]	809.2	248.8
`gamboost`	583.7	116.5	[355.2, 812.1]	582.1	151.4
`gam`	565.2	122.6	[324.9, 805.6]	563.6	138.1
`nls`	1818.1	242.3	[1343.2, 2292.9]	1808.5	307.6
`loess`	604.3	134.6	[340.4, 868.1]	596.6	169.2

Table 1. Common summary statistics of the example experiment: based on the 250 benchmark experiment runs, the mean, standard deviance (SD), 95% confidence interval, median and interquartiles range (IQR) of the empirical mean squared error distributions are calculated.

statistics. Table 1 shows the most established ones. In many cases, these heavily compacted numbers are the only analysis and basis for a ranking of the algorithms. But in doing so, one loses a lot of interesting and primarily important information about the experiment.

Based on the mean performance values, the order of the candidate algorithms is `gam` < `gamboost` < `loess` < `rpart` < `nnet` < `nls` < `lm`. As indication for the significance of differences, one can use the corresponding 95% confidence intervals. For our data the mean is approximately equal to the median for all except `nnet`, i.e. the performance of the latter seems to be skewed. Figure 2 shows a dot plot with the algorithms on the abscissa (sorted after their mean performance) and their performances on the ordinate, represented with a dot for each benchmark run (i.e., bootstrap sample). It can bee seen that the distribution for `nnet` is not only skewed, but also bimodal. Manual replications of fitting neural networks to the data show that training often gets stuck in local minima. Figure 3 shows a box plot with a box for

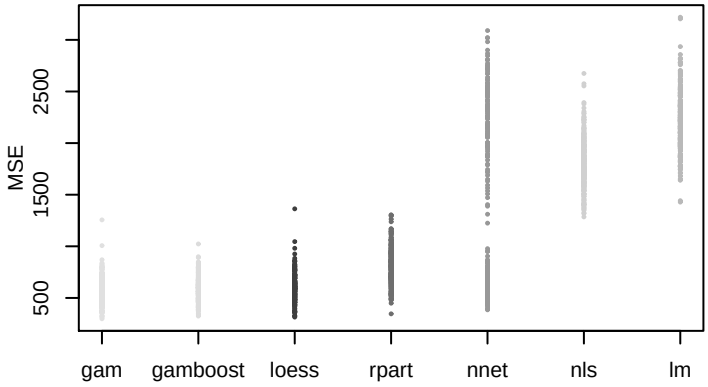


Fig. 2. Dot plot of the example experiment: the performance of each algorithm on each benchmark run is shown as a dot.

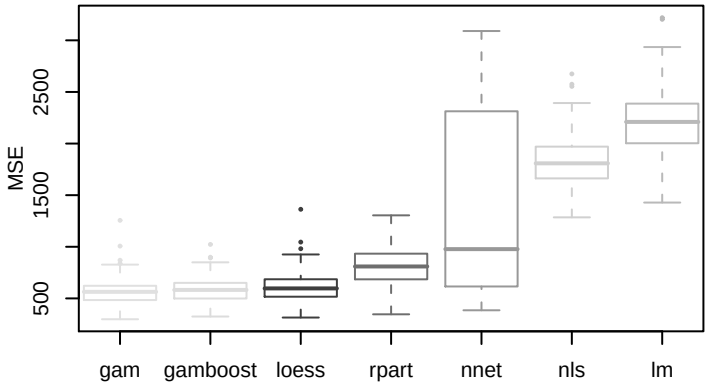


Fig. 3. Box plot of the example experiment: outliers are identified. In comparison to the dot plot, information about local minima is lost.

each algorithm. This plot allows the indication of outlier performances, but information about local minima is lost.

Both, Figure 2 and 3, also give an idea about the overall order of the algorithms. `gam`, `gamboost` and `loess` have basically the same performance, the small differences in mean performance being caused mostly by a few outliers. `rpart` has slightly worse performance, but with an isolated point close to the best value of the other three algorithms. `nls` and `lm` are in the upper MSE range, whereas `nls` has a lower minimal value as `lm`, but similar variance. `lm` also has some outliers near to the minimal value from `nls`. As said above, `nnet` ranges in two areas, whereas the lower MSE range (which corresponds to good performance) is similar to `gam`, `gamboost` and `loess`.

One massive problem of the dot plot is the overdrawing of dots. We do not know how many "lower" outliers of `rpart` there are, but the number of them really influences the impression of an order. This could be partly

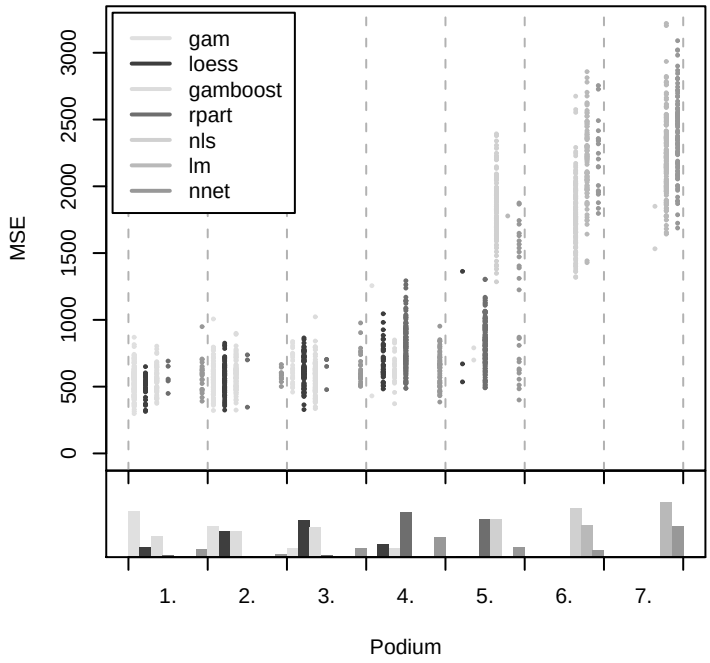


Fig. 4. Benchmark experiment plot of the example: the abscissa is a podium 7 places. For each benchmark run, the algorithms are sorted according to their performance values and a dot is drawn on the corresponding place. To visualise the count of an algorithm on a specific position, a bar plot is shown for each of podium places.

solved by jittering the plots, i.e., adding some random noise to the data. Additionally, the standard dot plot suggests the independence of the bootstrap samples. Indeed we know that, for example, `gam`, `gamboost` and `loess` perform similarly over all benchmark runs, but we do not know their ranking per benchmark run, which algorithm is on which rank and how often. The benchmark experiment plot was developed to overcome these limitations and to get a better understanding of benchmark experiments.

Instead of random jittering, we use the ranks of the algorithms on each bootstrap sample to horizontally "stretch out" the dots. For each benchmark run, the algorithms are ordered according to their performance value, and we draw seperate dot plots for each rank, ties are broken at random. This can be seen as creating a "podium" with K places, and having separate dot plots for each podium place, see Figure 4. Note that the plot is much easier to read when in color.

While the mean performances of `gam`, `gamboost` and `loess` (as shown in Table 1), and the performace distributions of these three (as shown in Figure 2 and Figure 3) all look very similar, we see in Figure 4 that `gam` is by

far most often the best algorithm for single bootstrap samples. Another aspect that is impossible to infer from the marginal distributions of the performance measures alone is that there are a few bootstrap samples where `rpart` works best.

The dots in Figures 2 and 4 are not independent from each other, because all algorithms were evaluated on each bootstrap sample. This dependency can be displayed by connecting the dots corresponding to one bootstrap sample with a line, resulting in a modified version of a parallel coordinates plot. In our implementation, the line segment between two podium places is drawn with the color of the algorithm in the lower position. To overcome the problem of overdrawing lines we use transparency (alpha shading). In this "full benchmark experiment plot" one can also see correlations between algorithm performances (parallel vs. crossing lines). In greyscale the plot looks like a big mess of grey lines and dots, and hence had to be excluded from this manuscript (a color version is available from `http://www.statistik.lmu.de/~eugster/`).

3 Inference

To make a statistically correct order and ranking we need more formal tools: statistical inference and primarily the testing of hypothesis provides them. The design of a benchmark experiment is a random block design. This type of experiment has two classification factors: the experimental one, for which we want to determine systematic differences, and the blocking one, which represents a known source of variability. In terms of benchmark experiments, the experimental factor is the set of algorithms and the blocking factor is that all algorithms perform on the same bootstrap samples.

We use a mixed effects model (e.g. Pinheiro and Bates, 2000) to analyze the output of a benchmark experiment. The variable of primary interest, i.e., the set of algorithms, is modelled as fixed effect β_j. The blocking factor, i.e. the sampling, is modelled as random effect b_j:

$$p_{ij} = \beta_0 + \beta_j + b_i + \epsilon_{ij}$$

with $b_i \sim N(0, \sigma_b^2)$, $\epsilon_{ij} \sim N(0, \sigma^2)$ and $i = 1, \ldots, B$, $j = 1, \ldots K-1$. Hence, we estimate only one parameter σ_b^2 for the effect of the data set. A modelling, by contrast, with the effect of the data set as main effect, would have lead to B parameters. Since we are able to draw as many random samples B from the performance distributions as required, we can rely on asymptotic normal theory. In case of our example the estimates for the parameters have been calculated as

$$\hat{\sigma}_b = 121.31,\ \hat{\sigma} = 353.73,$$

and

Intercept	Δ`gamboost`	Δ`lm`	Δ`loess`	Δ`nls`	Δ`nnet`	Δ`rpart`
$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
565.24	18.41	1643.91	39.03	1252.85	872.86	247.16,

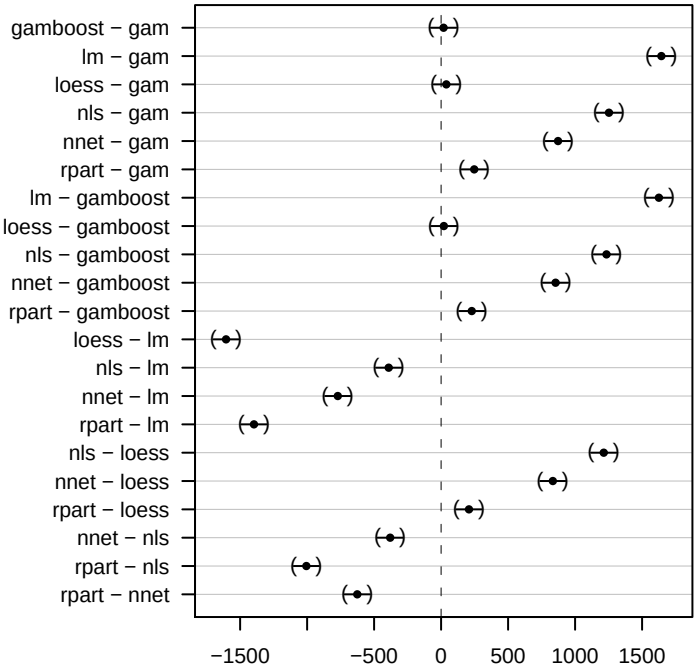


Fig. 5. Simultaneous 95% confidence intervals for multiple comparisons of means using Tukey constrast based on the mixed effects model of the example experiment.

with Δ denotes the difference between the Intercept and the corresponding algorithm.

The global test, whether there are any differences between the algorithms which do not come from the sampling, can be performed with ANOVA and the F-test. For our model this test rejects the null hypothesis that all algorithms have the same performance. We then use Tukey contrasts to test pairwise differences. Figure 5 shows the corresponding 95% family-wise confidence intervals. The differences between `gam`, `gamboost` and `loess` are not significant, the corresponding confidence intervals intersect 0 and overlap each other. As we can not establish a strict total order $<$ or a total order $\leq$, we define a reflexive and symmetric order relation $\approx$: two algorithms are $\approx$-related if their difference is not significant. The differences between all other algorithms are significant and we can establish a strict total order $<$ for each pair. Based on this set of ordered pairs ($\approx$- and $<$- ordered) we can use use a topological sort to define an overall order of the algorithms. In case of our benchmark experiment example, the final order of the candidate algorithms is `gam` $\approx$ `loess` $\approx$ `gamboost` $<$ `rpart` $<$ `nnet` $<$ `nls` $<$ `lm`.

4 Summary and future work

In this paper we gave a short introduction to our current work on formal statistical analysis of benchmark experiments and introduced the benchmark experiment plot as a new visualisation method. The random block design of a benchmark experiment has been modelled using mixed effects. This allows to test various hypothesis of interest, amongst others, the pairwise differences. We introduced an order relation for algorithms with non-significant differences, and inferred a statistically correct order of the candidate algorithms.

This paper is the first step towards a comprehensive toolbox for exploratory and inferential analysis of benchmark experiments. There are lots of things to do. Two examples are (1) sequential testing to reduce computation time and (2) alternative order mechanisms like the minimax principle. Besides the analysis of a set of candidate algorithms on one data set, the extension to a set of data sets is obvious.

Acknowledgments

The authors want to thank Torsten Hothorn for discussions and ideas.

References

DEMSAR, J. (2006): Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7, 1–30.*

HORNIK, K. and MEYER D. (2007): Deriving consensus rankings from benchmarking experiments. In: R. Decker and H.-J. Lenz (Eds.): *Advances in Data Analysis.* Springer-Verlag, 163–170.

HOTHORN, T. and BÜHLMANN, P. (2006): Model-based boosting in high dimensions. *Bioinformatics 22 (22), 2828–2829.*

HOTHORN, T., LEISCH, F., ZEILEIS, A. and HORNIK, K. (2005): The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics 14 (3), 675–699.*

PINHEIRO, J.C. and BATES, D.M. (2000): *Mixed-Effects Models in S and S-PLUS.* Springer-Verlag.

R DEVELOPMENT CORE TEAM (2007): *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

SILVERMAN, B.W. (1985): Some aspects of the spline smoothing approach to non-parametric regression curve (with discussion). *Journal of the Royal Statistical Society Series B (47), 1–52.*

VENABLES, W. and RIPLEY B. (2002): *Modern Applied Statistics with S.* Springer-Verlag.

YILDIZ, O.T. and ALPAYDIN, E. (2006): Ordering and finding the best of k>2 supervised learning algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (3), 392–402.*

Mining Temporal Associations Between Air Pollution and Effects on the Human Health

Corrado Loglisci and Donato Malerba

Dipartimento di Informatica, Università degli Studi di Bari, Italy
{*loglisci,malerba*}*@di.uniba.it*

Abstract. The task of monitoring and improving the urban air quality has attracted a great deal of interest both from national governments and scientific communities. In order to implement policies for the environmental protection, the recent urban planning decisions are often based on the results produced by several research fields. An important research direction aims at understanding the pollution phenomenon by means of data mining approaches, which support decision makers with information extracted directly from data. In this work we investigate the effect of air pollution on human health by taking into account the temporal variability of environmental data. Since the repercussions of air pollution on humans are perceived only after a certain lapse of time, we propose to discover temporal associations which relate a change at time t_j of the population health conditions with a change at time t_i (t_j>t_i) of the polluting emissions. Information conveyed by discovered temporal associations could be exploited both to support policies for environmental protection and to adopt strategies for the reduction of human health risks.

Keywords: data mining, temporal associations, air pollution effects

1 Introduction

In the last decades, we have observed the deterioration of air quality of highly populated urban zones, specially because of motor vehicle emissions (HEI (2003)). For this reason, policies of urban planning have been addressed to contrast the eco-system degradation. One of the measures adopted is that of making information systems able to manage urban air data coming from monitoring stations and provide advanced capabilities for the analysis of environmental data as well. This approach follows the mainstream of integrating knowledge-based technologies into existing information systems to provide more effective solutions to urban problems (Han et al. (1989)). In particular, two main lines of research can be identified: first, the formulation of physics models from dynamic simulations; second, the induction of models from observations. The latter is based on Knowledge Discovery (Fayyad et al. (1996)), which can be applied to the large amounts of collected data and which can support the decision makers with knowledge directly unearthed from environmental data (Read (2000)).

An important scientific issue is the investigation of the effects of pollutants on the population health, since most of the strategies of urban planning are aimed to reduce social deprivation cases and epidemiologic risks caused by the industrialization process (King et al. (2000)). In this work we investigate the effect of air pollution on human health by taking into account the temporal variability of environmental data. This study is based on two assumptions: 1) the actual consequences of the air pollution on the human health becomes evident and perceptible after some temporal delay, and 2) an objective estimation of the air quality is reliable only after a certain lapse of time. The proposed method of temporal data mining discovers temporal associations which relate an event occurring at time t_i in the polluting emissions with an event occurring at time t_j (t_i<t_j) in the population health conditions. Discovered patterns can be interpreted as follows: an event observed when monitoring the population health conditions might have been triggered by an event related to polluting emissions.

The paper is organized as follows. Section 2 illustrates the background of this work on the adverse effects of air pollution on human health. In Section 3 we define the problem of mining temporal associations from air pollution and human health data and report a method to solve it. In Section 4, the application to a real dataset and the experimental results are discussed.

2 Background and related works

Potential associations existing between adverse effects of air pollution and human health have been traditionally investigated by biostatisticians and by environmental protection agencies. A well-documented result (Schwartz (1994), Dockery et al. (1994), Wordley et al. (1997)) is that the emissions of "particulate matter" are associated with decreased respiratory function, aggravation of existing respiratory and cardiovascular conditions, altered defense mechanisms and even premature death (Bascom (1996)). However, different studies conducted on several urban zones lead a slightly contrasting results: the discrepancies may be due to the fact that the chemical composition of the pollutants differs with the geographic areas.

The widely used approach in cited works is that of time-series statistical analysis by means of *Generalized Additive Models*. This technique offers three relevant advantages: i) conducting nonlinear regression analysis on the short-terms effects (Dominici et al. (2002)), ii) adjusting confounding effects (e.g., trends and seasonality) with non-parametric functions, and iii) modeling the effects in terms of time lags. However, the results can be heavily dependent on the chosen smoothing functions and the estimation of regression coefficients can be strongly biased (Baccini et al. (2007)). Data Mining approaches mainly tackle the issue of forecasting future levels of polluting emissions (Efraimidou et al. (2006)). For instance, Read (2000) faces the problem of short-term forecasting of pollutant categorical values by proposing a methodology based

on rules induction from daily maximum data. Osrodka et al. (2005) resort to a two-stepped procedure which initially identifies groups of similar meteorological situations through Kohonen's self organizing networks, and then, for each group, learns SOM networks used for the prediction of emissions. Although these works allow to model the relationships between past and future values of pollutants, two important aspects are not carefully investigated: i) the temporal variability of data of air pollution (no one temporal information is associated to the predicted data), and ii) the relationships between the polluting emissions and the human health.

3 Mining temporal associations

As clarified before, we are interested in mining data observed over time for both air pollution and human health. Data consist of time-stamped measurements which can be collected on a calendaric basis (e.g., hourly or daily) (Li et al. (2004)) and are represented by multi-dimensional (or multi-variate) time-series. The scientific problem of interest in this work can thus be formulated as follows:

Given:
the m^{AP}-dimensional time-series AP of air pollution data and the m^{HH}-dimensional time-series HH of human health data;

Find:
1) a finite set of temporal states S^{HH}:$\{S_1^{HH}, S_2^{HH}, \ldots, S_s^{HH}\}$ induced from HH, and
2) for each pair (S_j^{HH}, S_{j+1}^{HH}), j=1, ..., s-1, the **temporal association** $L_{j,j+1}$: $\langle e_1, e_2, \ldots, e_h, \ldots, e_p \rangle$, where each e_h is a **event**.

In the following subsections we report the notions of temporal state and event, and then we describe the method proposed to find them.

3.1 Discovering the temporal states

Informally speaking, a *temporal state* S_j consists of the set C_j of all facts characteristic of population heath conditions which are true over a certain lapse of time. Typically a state is associated to a period of time $[ts^j ... te^j]$ ($ts^j \leq te^j$)[1] during which there are no changes in the facts C_j: a state can thus be seen as a snapshot that lasts over the time $[ts^j ... te^j]$. Therefore, finding out the temporal states S^{HH} means detecting the several snapshots occurring in *HH*. At this aim we resort to the approach proposed by Loglisci et al. (2006) which allows to extract the temporal states S^{HH} in the following way. First, a *time-series segmentation* step is performed on *HH* to identify

[1] It is assumed that the partial order relation '$\leq$' holds for the time-points ts^j and te^j.

the time-periods [ts^j...te^j] for each state: it generates the time-periods by detecting non overlapped segments of HH, which are characterized by low data variability with respect to (w.r.t.) a threshold ω of statistical coefficient variation, and by highly correlated attributes w.r.t. a threshold ρ of a statistical linear correlation. Next, for each of resulting segment, an *inductive learning process* generates the facts C_j such that the following conditions are met: given two states S_j:⟨ts^j,te^j,C_j⟩, S_{j+1}:⟨ts^{j+1},te^{j+1},C_{j+1}⟩, the facts C_j do hold in [ts^j...te^j] but do not in [ts^{j+1}... te^{j+1}], while C_{j+1} do hold in [ts^{j+1}... te^{j+1}] but do not in [ts^j...te^j], te^j<ts^{j+1}.

3.2 Finding out the events

The meaning of event corresponds to the usual notion of a whatever action that occurs within a certain time-period and that leads the examined phenomenon to evolve. An event **e** is represented in terms of:

- ⟨ed$_1$,...,ed$_r$,...,ed$_{m'}$⟩, which corresponds to a subset of the m^{AP} attributes;
- w$_k$:[ts^k,te^k] and w$_{k+1}$:[ts^{k+1},te^{k+1}], which are time-windows on AP. We mean the event **e** lasts over the period [ts^k...te^k] ([ts^k...te^k]⊆[ts^j...te^j] of the state S_j);
- ⟨cv$_1$, ..., cv$_r$,..., cv$_{m'}$⟩, which is a set of categorical values for each ed$_r$ in w$_k$. Roughly speaking, each cv$_r$ represents qualitatively ed$_r$ in w$_k$;
- ⟨[inf$_{ed_1}$...sup$_{ed_1}$], ..., [inf$_{ed_r}$...sup$_{ed_r}$],..., [inf$_{ed_{m'}}$,sup$_{ed_{m'}}$]⟩, which is a set of numerical ranges for each ed$_r$ in w$_k$. Roughly speaking, each interval [inf$_{ed_r}$...sup$_{ed_r}$] represents quantitatively ed$_r$ in w$_k$.

Moreover, **e** is described by a statistical parameter λ, $\lambda > 0$, which accounts the fraction of the found events $\mathbf{e}_v$ such that:
⟨ed$_1$,...,ed$_r$,...,ed$_{m'}$⟩^{e_v} ⊇ ⟨ed$_1$,...,ed$_r$,...,ed$_{m'}$⟩^{e} and ts^{k_e}>ts$^{k_{e_v}}$ ∧ te^{k_e}<te$^{k_{e_v}}$ and for each ed$_r$: cv$_r^e$=cv$_r^{e_v}$ and [inf$_{ed_r}$...sup$_{ed_r}$]^e ⊆ [inf$_{ed_r}$...sup$_{ed_r}$]^{e_v}.
In other terms, **e** consists of the most frequent events.

A temporal association is a sequence of events $L_{j,j+1}$: ⟨e$_1$,e$_2$,...,e$_h$,...,e$_p$⟩ where ts^j≤ ts$^{k_{e_1}}$, ts$^{k_{e_{h+1}}}$=te$^{k_{e_h}}$+1, te$^{k_{e_p}}$≤te^j. We are interested in temporal associations whose events have high values of λ.

The procedure to generate these sequences resorts to the work by Loglisci and Malerba (2008), which originates in the fact that the events can be seen as *significative variations* occurring in AP. The idea is that whatever variation in the data is reflected also in the underlying models. Therefore, to detect the events that can trigger the evolution from S_j^{HH} to S_{j+1}^{HH} we detect differences between the models induced respectively from w$_j$:[ts^j,te^j] and w$_{j+1}$:[ts^{j+1},te^{j+1}] on AP. In particular, linear models are considered:

$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$,

where the dependent and independent variables x_i correspond to the attributes ed_1,ed_2,... describing an air pollution phenomenon. For instance, given the coefficients β_1^k,β_2^k,... and β_1^{k+1},β_2^{k+1},... of the regression models induced from w_k and w_{k+1}, an event is generated if $|\beta_r^k - \beta_r^{k+1}| \geq \epsilon$, for some ed_r, where ϵ is a user-specified threshold. Parameters β_i are estimated by least squares.

4 Application

We evaluated the mining of temporal associations between air pollution and human health on some of datasets resulting from *Morbidity, Mortality, and Air Pollution Study* (NMMAPS[2]) funded by Health Effects Institute, Boston, MA. Several datasets are available: they concern daily data of pollution emissions, meteorological conditions and hospitalization for the ninety largest cities in the US. The interest on the weather parameters is justified by the fact that high levels of air pollution are associated with particular meteorological conditions. Because of space limitations, we report only the details of the experiments on the dataset of the time-period Jan 1, 1987 - Dec 31, 2000 in the city of Phoenix, Arizona. Since Phoenix is an arid south-western city, it has a large proportion of population susceptible to particular health conditions: in this sense it is an interesting location for our analysis. The attributes that describe the human health conditions are: *Chronic Obstructive Pulmonary Diseases* (COPD),*Cardiovascular Deaths* (CVD), *Influenza Cases* (INF) and *Respiratory Deaths* (RESP). We consider as weather attributes *Maximum Temperature* (TMAX, $\circ$F) and *Maximum Relative Humidity* (MXRH, %). The air pollution attributes, *Particulate Matter* (PM, $\mu g/m^3$) with aerodynamic diameter ≤ 10 and 25 μm), *Carbon Monoxide* (CO, $\mu g/m^3$), *Nitrogen Dioxide* (NO_2, $\mu g/m^3$), *Sulfur Dioxide* (SO_2, $\mu g/m^3$) and *Ozone Concentration* (O_3, $\mu g/m^3$)) were originally averaged across monitors and the missing values were replaced adding the trimmed mean value with the daily median value of 1-year trend.

By varying the input parameters specified in Section 3.1, we obtain two different state sets, the former consisting of 38 states, each of which spans at least 3 months (90 time-points), and the latter consisting of 75 states spanning at least 1 month (30 time-points). This allows us to discover relationships between AP and HH with respect to monthly, weekly and daily temporal axis. A first interesting result is the association discovered over the states S_{33}: $\langle t_{4330}$, t_{4428}, $C_{33}\rangle$ and S_{34}:$\langle t_{4429}$, t_{4525}, $C_{34}\rangle$[3]: the sequence of events reported in Table 1 occurs in S_{33} and might trigger the increase of the number of cases of COPD and RESP in S_{34}. More precisely, it can be interpreted as follows

[2] http://www.ihapss.jhsph.edu/data/data.htm

[3] C_{33}:{... $\wedge$ RESP in [1..5]$\wedge$ COPD in [1..6]...}, C_{34}:{... $\wedge$ RESP in [1..9]... COPD in [2..10]...}

(see Figure 1): *a strong increase of SO2 and of NO2 in [t_{4366}...t_{4386}] (a 20-days period in winter), an increase of O3 in [t_{4408}...t_{4428}] (a 20-days period in winter) and an increase of TMAX in [t_{4408}...t_{4428}] (a 20-days period in winter) occur before than the increase of COPD in [t_{4429}...t_{4525}] (a 3-months period in spring).* Table 1 reports also the fraction (e.g. 6/6) of all of found events that support the selected event.

Table 1. Temporal association mined over S_{33} and S_{34}.

$\langle \ldots, ed_r, \ldots, \rangle$	$\langle \ldots, [inf_{ed_r} \ldots sup_{ed_r}], \ldots \rangle$	w_k	$\langle \ldots, cv_r, \ldots \rangle$	λ
TMAX	[40.34 ... 85.29]	[$t_{4408} \cdots t_{4428}$]	INCREASE	6/6
O3	[9.06 ... 25.29]	[$t_{4408} \cdots t_{4428}$]	INCREASE	6/6
PM10	[40.34 ... 85.29]	[$t_{4387} \cdots t_{4407}$]	STEADY	7/7
NO2	[22.37 ... 54.32]	[$t_{4366} \cdots t_{4386}$]	INCREASE	7/7
SO2	[1.84 ... 7.41]	[$t_{4366} \cdots t_{4386}$]	VERY_INCREASE	7/7

This means that high concentrations of chemical components (SO2,NO2) of PM can influence the respiratory function not immediately, but have repercussions in few months. These results are in agreement with the findings of Mar et al. (2000), who observed that high values of chemical components, although in colder months, can aggravate critical respiratory conditions.

A result of the experiments on states with at least 30 time-points (30 days) shows that the human with cardiovascular diseases are more susceptible and, thus, immediately affected by high levels of polluting emissions. In Table 2 we report the sequence of most frequent events that occur during the state S_{56}:$\langle t_{4286}, t_{4336}, C_{56} \rangle$ but before than S_{57}: $\langle t_{4337}, t_{4389}, C_{57} \rangle$[4].

Table 2. Temporal association mined over S_{56} and S_{57}.

$\langle \ldots, ed_r, \ldots, \rangle$	$\langle \ldots, [inf_{ed_r} \ldots sup_{ed_r}], \ldots \rangle$	w_k	$\langle \ldots, cv_r, \ldots \rangle$	λ
CO	[775 .. 1700]	[$t_{4286} \cdots t_{4315}$]	VERY_INCREASE	7/7
TMAX	[78.5 .. 85.29]	[$t_{4286} \cdots t_{4315}$]	STEADY	7/7
NO2	[16.62 .. 40.82]	[$t_{4315} \cdots t_{4336}$]	INCREASE	6/6
SO2	[0.7 .. 5.5]	[$t_{4315} \cdots t_{4336}$]	VERY_INCREASE	6/6

It means that the increasing of cardiovascular deaths detected during [t_{4337}... t_{4389}] might be explained in terms of: *strong increase of CO in [t_{4286}...t_{4315}] (that is, 20 days before), then increase of NO2 and SO2 in [t_{4315}...t_{4336}] (that is immediately before).* Our results find consensus again in the conclusions of Mar et al. (2000) according to which the cardiovascular mortality is positively correlated with the chemical components of PM_{25} and PM_{10} and, moreover, is associated to them with several degrees of time-lags. Our temporal associations provide additionally the temporal interval information during which these relationships hold.

Finally, we evaluated the result accuracy by introducing the notion of *True Positive* event (Heagerty et al. (2005)). An event **e** is True Positive iff

i) there exists $\{ed'_1,\ldots,ed'_k,\ldots,ed'_{m''}\} \subseteq \{ed_1,\ldots,ed_r,\ldots,ed_{m'}\}$ such that each ed'_k= TMAX $\vee$ MXRH $\vee$ PM_{10} $\vee$ PM_{25} $\vee$ CO $\vee$ NO_2 $\vee$ SO_2 $\vee$ O_3, and $|\{ed'_1,\ldots,ed'_k,\ldots,ed'_{m''}\}| > |\{ed_1,\ldots,ed_r,\ldots,ed_{m'}\}|/2$,

ii) **e** is the most frequent event.

[4] C_{56}:{... $\wedge$ CVD in [7..17]$\wedge$ CVD in [6..12]...}, C_{57}:{... $\wedge$ CVD in [9..21]... CVD in [7..29]...}

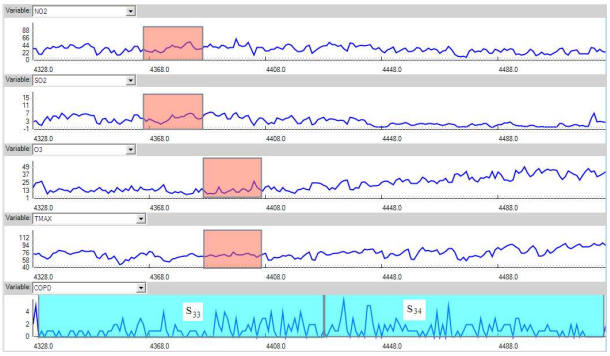


Fig. 1. Representation (from top to bottom) of the attributes NO2, SO2, PM10, O3, TMAX and COPD from the state S_{33} to S_{34}. The time-windows of the states are drawn only on the axis of COPD, while the time-windows of the events are drawn on the axis of the other attributes.

The accuracy of the events detected in S_{56} w.r.t. S_{57} is reported below. We observe that the influence of the time-period width on the accuracy: this confirms the importance of considering the temporal variability in the analysis of the adverse effects of air pollution.

accuracy(%)	88.8	100	50	62.5	83.3	83.3
minimum time-period of events (days)	5	10	15	20	25	30

5 Discussion

We investigated the influence of air pollution on population health by discovering associations that temporally relate changes in health conditions with changes in polluting emissions: these provide quantitative and qualitative information about 'how human health is affected', 'which are the changes and when they occur' and 'how the influence is statistically supported'. These associations are described in terms of sequentially ordered events: the occurrence of particular events in the polluting emissions might trigger changes in the health conditions. The events are mined by means of a procedure that we reported in the paper and which makes use of computational statistical techniques (linear regression and temporal segmentation). Information conveyed by these associations is operational and makes the evaluation and optimization of classical physical modelling possible (Osrodka et al. (2003)).

Acknowledgment

This work is partial fulfillment of the research objective of ATENEO-2008 project "Scoperta di conoscenza in domini relazionali".

References

BACCINI, M., BIGGERIA, A., LAGAZIO, C., LERTXUNDIA ,A. and SAEZD, M. (2007): Parametric and semi-parametric approaches in the analysis of short-term effects of air pollution on health. *Computational Statistics and Data Analysis* 51(9), 4324-4336.

BASCOM, R. (1996): BASCOM, R., BROMBERG, P.A., COSTA, D.A., DEVLIN, R., DOCKERY, D.W., FRAMPTON, M.W., LAMBERT, W., SAMET, J.M., SPEIZER, F.E., UTELL, M. Health effects of outdoor air pollution. Part I. *Am. J. Respir. Crit. Care Med.* 153(1), 3-50.

DOCKERY, D.W. and POPE, C.A. III (1994): Acute respirator effects of particulate air pollution. *Annu. Rev. Public Health* 15, 107-132.

DOMINICI, F., McDERMOTT, A., ZEGER, S.L. and SAMET, J. (2002): On the use of generalized additive models in time-series studies of air pollution and health. *Am. J. Epidemiol.* 156(3), 193-203.

EFRAIMIDOU, M., KANAKI, M., ATHANASIADIS, I., MITKAS, P. and KARATZAS, K. (2006): Data mining air quality data for Athens, Greece. *Proc. of the 20th EnviroINFO*, Tochermann Eds, 505–508.

FAYYAD, U.M., PIATESKY-SHAPIRO, G. and SMYTH, P. (1996): The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39(11), 27-34.

HEAGERTY, P.J. and ZHENG, Y.(2005): Survival model predictive accuracy and ROC curves. *Biometrics* 61(1), 92–105.

HAN, S.Y. and KIM, T.J.(1989): Can expert systems help with planning? *Journal of the American Planning Association* 55(3), 296–308.

HEI (2003): Special Report, Revised Analyses of Time-Series Studies of Air Pollution and Health, Health Effects Institute, Boston.

KING, K. and STEDMAN, J. (2000): Analysis of Air Pollution and Social Deprivation. *Report AEAT/R/ENV/0241 AEA*, Technology Environment.

LI, S.T. and SHUE, L.Y. (2004): Data mining to aid policy making in air pollution management. *Expert Systems with Applications* 27(3), 331-340

LOGLISCI, C. and BERARDI, M. (2006): Segmentation of Evolving Complex Data and Generation of Models. *Proc. of the 6-th ICDM - Workshops*, IEEE Computer Society, 269–273.

LOGLISCI, C. and MALERBA, D. (2008): Discovering Explanations from Longitudinal Data. In: A. An, S. Matwin, Z.W. Ras and D. Slezak (Eds.): *Foundations of Intelligent Systems, 17th International Symposium On Methodologies For Intelligent Systems*, LNAI 4994, Springer-Verlag, Berlin, 196–202.

MAR, T.F., NORRIS, G.A, KOENIG, J.Q. and LARSSON, T.V. (2000): Association between air pollution and mortality in Phoenix. *Environ. Health. Perspect*, 108(4), 347–353.

OSRODKA, L., WOJTYLAK, M., KRAJNY, E., DUNAL, R. and KLEJNOWSKI, K. (2005): Application Data Mining for forecasting of high-level air pollution in urban-industrial area in southern Poland. *Proc. of 10th Conference HARMO for Regulatory Purposes*.

READ, B.J. (2000): Data mining and science? Knowledge Discovery in Science as opposed to business, *Proc. of 12-th ERCIM Workshop on Database Research*.

SCHWARTZ, J. (1994): What are people dying of on high air pollution days? *Environ. Res.* 64(1), 26-35.

Mining Information from Plastic Card Transaction Streams

Dimitris K. Tasoulis[1], Niall M. Adams[2], David J. Weston[1], and David J. Hand[1,2]

[1] Institute for Mathematical Sciences, Imperial College London, SW7 2PG, UK
[2] Department of Mathematics, Imperial College London, SW7 2AZ, UK,
{*d.tasoulis,d.weston,n.adams,d.j.hand*}*@imperial.ac.uk*

Abstract. Detecting fraudulent plastic card transactions is an important problem in retail financial services. The problem is challenging because the data are streaming, heterogeneous, dynamic, and there is a large imbalance between fraudulent and legitimate transaction classes. Due to the complexity of the problem, a variety of knowledge discovery approaches can be fruitfully deployed. However, existing information systems' infrastructure constrains the information available to analyse each transaction. To provide a richer representation we therefore consider extending the practitioner's toolkit to extract feature information using recently proposed stream clustering algorithms. Such algorithms have to account for the temporal structure of the data. In this paper, we explore the utility of on-line density-based stream clustering methods for plastic card transaction fraud detection. Experiments with real data suggest that such methods have merit for both fraud detection and for and for revealing aspects of temporal structure of the transaction stream. This suggests that new features can be discovered to enhance existing detection algorithms.

Keywords: fraud detection, data clustering, data streams

1 Introduction

Plastic cards have been a major advance in the banking and credit services offered by lenders. However, accompanying such advances, plastic card providers are challenged with the serious problem of fraudulent card transactions. To illustrate the magnitude of the problem, it is estimated that losses attributed to such fraud in the UK in the first six months of 2006 amounted to £209 million (APACS (2008)). Note that during this period, UK lenders introduced a scheme requiring PIN authentication for the majority of transactions. Since then, fraudsters have adopted new tactics to circumvent the scheme. Fraud detection methods must therefore have the potential to adapt to shifting fraudulent behaviour. In this arms-race the data available to lenders is constrained by existing information processing infrastructure. To enhance the data available to lenders, in this paper we explore the utility of extracting new features from the stream of transactions. Large and sophisticated infrastructure exists to rapidly process plastic card transactions.

Loosely, fraud implies unauthorized and illegal use of the facilities of a legitimate account. Tackling fraud in the context of plastic card finance is a daunting problem. A number of complicating factors are involved including the sheer volume of transactions to process, the asynchronous and heterogeneous nature of transactions, and the adaptive behaviour of fraudsters. The effort to handle fraud can be broadly divided into *fraud prevention*, that attempts to block fraudulent transactions as they occur, and *fraud detection* where successful fraud transactions are subsequently implicated. For fraud prevention purposes, lenders typically challenge all transactions with rule based and other filters, often based on third party software such as, for example, VISA's VISOR fraud detection tool (Visa (2003)). Fraud detection should find fraudulent transactions as rapidly as possible after they occur. In the case of both fraud prevention and detection, the problem is magnified by specific characteristics of plastic card finance. First, to avoid customer irritation, the number of incorrectly implicated transactions needs to be kept to a minimum, Second, most lenders routinely process vast numbers of transactions, more than 20000 per day is not uncommon, of which only a small fraction is fraudulent, often less than 0.1%.

Many approaches to fraud problems have been considered (for example Kou et al. (2004), provide general discussion). Statistical views are explored by Bolton and Hand (2002). In the context of plastic card fraud, various authors Brause *et al.* (1999), Maes *et al.* (2002), have approached fraud detection as a classification problem. There are a number of difficulties with this approach, including the extensive processing requirement associated with irregularly timed transaction sequences, and the conversion of the data into a representation suitable for classification algorithms. Moreover, the approaches may ignore important temporal aspects of fraud, particularly that fraudsters change tactics – classification approaches can only find existing tactics. We propose tools that complement standard detection methods by providing new features.

There are many other ways to approach this problem, such as peer group analysis Weston *et al.* (2008) or outlier detection Juszczak *et al.* (2008). It is unlikely that just one approach will successfully detect all types of fraud. Practical systems adopt more than one approach. We are attempting to provide new features that should enhance these hybrid systems. To this end, we consider the application of streaming clustering algorithms Cao *et al.* (2006); Tasoulis *et al.* (2006), that to the best of our knowledge, have not yet been applied to this problem.

Streaming data, consisting of multiple indefinitely long and time-evolving sequences, is becoming ubiquitous. Such data presents new challenges to data mining algorithms. These challenges arise primarily from the dynamically changing nature of the streams, thus clustering algorithms must have the capacity to adapt rapidly to changing dynamics of the sequences. Additionally, timely results and scalability in the number of sequences is becoming

increasingly desirable, as data collection technology develops. Plastic card transaction data, has precisely these characteristics. In this contribution we utilize density-based clustering methods, that are an important category of clustering algorithms. These methods partition data into clusters of high density, surrounded by regions of low density. To address the issues of stream dynamics, timeliness and scalability, recent developments (Cao *et al.* (2006); Tasoulis *et al.* (2006)) have extended the most successful density clustering algorithms to the streaming data model. An analysis of these methods is presented in Section 2.

Among the collection of methods used in plastic card fraud detection, stream clustering algorithms are potentially advantageous for two particular reasons. Such algorithms can operate asynchronously at the transaction level in real time. Also as exploratory tools, these algorithms have the potential to identify different types of fraud, and characterize temporal components of fraud transactions.

In the next section we briefly review the literature on data stream clustering, and describe the algorithms used in this work. In Section 3 we describe the plastic card transaction data stream. Next, in Section 4, we present experimental results using streaming clustering. We conclude with a discussion in Section 5.

2 Data stream clustering

Traditional clustering methods are not able to accommodate the needs of the streaming data model, since they rely on the assumption that the data are available in a permanent memory structure, from which global information can be obtained at any time. Recently however new algorithms (Cao *et al.* (2006); Tasoulis *et al.* (2006)), have been developed that embrace the need of clustering in streaming applications. Here we focus on density based clustering methods and in particular on the DenStream and WStream algorithms, for three reasons. First, both of them are based on successful static clustering algorithms. Second, they have the desirable characteristic of providing approximations to the cluster number without prior knowledge, and at the same time they are able to detect non-convex clusters without any particular data transformation. Finally, they are both computationally efficient, which is a fundamental requirement for the application to credit card transactions.
The WStream algorithm The WStream algorithm (Tasoulis *et al.* (2006)) uses containers (*windows*) in the form of hyper-rectangles that are adjusted though time to discover and track the evolution of the underlying clusters. This is achieved using two procedures, "*movement*" and "*enlargement-contraction*". The "*movement*" of windows incrementally recenters windows every time a new streaming data point arrives. Windows are recentered to the mean of the points they include at each time point in a manner that also depends on each point's timestamp. A fading function, that decreases

with time, associates a weight with each timestamp. This function depends on a parameter called the *forgetting factor*, that provides a compromise between the ability to track changes and the need to suppress the uninformative stochastic behaviour of the data. The larger the value of this parameter the faster the algorithm forgets previous information. On the other hand when it attains small values the algorithm is strongly influenced by historic information. The "*enlargement-contraction*" procedure aims to iteratively adapt the window widths, that relate to the scales of the different variables. The width of each co-ordinate of a window is enlarged or contracted depending on rules that also depend on user defined parameters.

The algorithm maintains a list of windows that are iteratively adjusted each time a new streaming data point arrives. New windows are created when the new data are not included in any of the windows. For a detailed description of WStream and a sensitivity analysis of its parameters see Tasoulis *et al.* (2006).

The DenStream algorithm DenStream (Cao *et al.* (2006)) was developed from its static counterpart (DBSCAN Sander *et al.* (1998)), which dictates that in a neighbourhood of a given radius, for each point in a cluster at least a minimum number of points should be contained. DenStream utilizes micro-clusters to extend this concept to the spatio-temporal setting. Micro-clusters are defined as quantities that capture the weight of points that reside in an area of a specific size, called the *diameter* of the micro-cluster. The *weight* is computed by again utilizing a fading function applied on the timestamps of the points inside each such area. Similar to WStream, the fading function is based on a parameter called forgetting factor.

Two types of micro-clusters are considered based on user defined parameters. We have a *core-micro-cluster*, if the cluster weight is large enough, and its diameter is small enough. These account for a "dense" region of the data. Otherwise the micro-cluster is called an *outlier-micro-cluster*. DenStream maintains two lists; one for the core-micro-clusters, and the other for the outlier-micro-clusters. These lists are updated each time new data arrives. Initially, an attempt is made to merge the new data into its nearest core-micro-clusters. If the resultant micro-cluster has a significantly large diameter (based on user defined rules) the merge is omitted. In this case, a new attempt is made to merge the new data into its nearest outlier-micro-cluster, using a similar rule. If this merge also fails a new outlier-micro-cluster is created, centered at the new data. If however the merge changed the outlier-micro-cluster such that it can be considered a core-micro-cluster it is moved to the appropriate list.

In order to derive the clustering result a variant of the DBSCAN algorithm is applied on the list of core-micro-clusters. Each core-micro-cluster is regarded as a virtual point located at its center, having its respective weight. The concept of density-connectivity (Sander *et al.* (1998)), is used to derive the final clustering result.

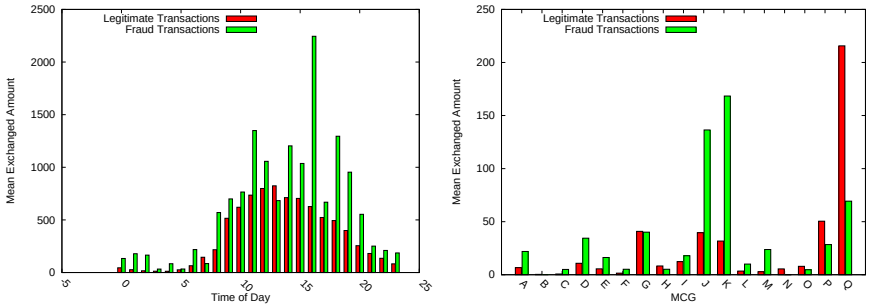


Fig. 1. Time of Day – Amount Plot(left), MCG – Amount Plot (right).

3 The transaction stream

A credit card transaction record is a complex entity. A fundamental identifier is the particular *account* associated with the transaction. The focus of this paper is the sequence of transaction as they are processed by the lender. We are less concerned with account level structure, as with exploring the spatio-temporal structure of all transactions.

One of our commercial collaborators provided transactions records with 77 fields. These refer to a diverse set of information, including process-oriented fields like card reader response status codes. Such fields can be used to precisely identify important details of the transaction. A fundamental distinction is provided by the service ID, that indicates transaction type, determining whether the transaction was conducted at an automatic teller machine (ATM) or at a point-of-sale (POS) terminal. Note that it is possible to complete a variety of transaction types at an ATM in addition to cash withdrawals. For example, transferring money between accounts.

We extract a variety of data from transaction records. All POS transactions have a merchant category code marker. The merchant category code (MCC) is used to identify in which market segment the transaction was performed. For example "Book Stores". The MCC is a four digit number, though there are far fewer than 10000 codes currently in use. A merchant category group (MCG) is a grouping of MCCs into a broader market segment using business criteria. More details of MCCs and MCGs are given in Weston et al. (2008). We use MCGs to reduce the dimensionality of the merchant categories to 17. We introduce one further merchant category group to label ATM transactions. It is clear that streams consisting of transactions from heterogeneous accounts, require extra processing to manage this extra structure. We handle this by comparing each transaction with the profile of the respective account (Piotr *et al.* (2008)) as explained later.

The data we extract from each transaction is the extended MCG and two other features, the time of day the transaction occurred and the amount of money exchanged. For illustration the left plot of Fig 1 shows the relation

between time of day and mean amount exchanged, for 20000 transactions, by fraud status. Similarly the right plot of Fig. 1 illustrates the relation between the mean amount and the MCG (in arbitrary order). We include these figures to illustrate the complex structure of the data. However easy it may seem to identify fraudulent behaviour, we must note first that the figure refers to a single day's data and second that some structure in the graph may be an artefact of imbalanced classes. For such reasons we are interested in methods that view the data differently.

To handle heterogeneity induced by accounts, for each account holder we construct a simple updating profile and measure a transaction's deviation from its profile. This profile consists of the mean amount exchanged and the number of times each MCG was used in a transaction, as a proportion of the total number of transactions the account holder performed. Thus each transaction in the stream is composed of 20 features, the first of which is the time it arrives. The second is the deviation from the profile mean for the amount associated with the transaction. The remaining 18 are the MCG counts as described above.

4 Streaming clustering

In this section we explore the behaviour of streaming clustering on fraud transaction data. Our objective is not to provide a comparative analysis of this method for fraud detection, but rather to characterise the structures that the method reveals.

To evaluate the stream clustering performance over the complete dataset we used the following methodology. We deployed WStream to continuously adapt to the complete data stream. Each time a fraud transaction occurred, we examined the correspondence of the most recent 2000 transactions, on the clustering adapted so far. Each transaction that did not belong to a window was regarded an outlier. We define as the False Positive (FP) ratio the proportion of legitimate transactions flagged as outliers. Respectively, the proportion of fraudulent transactions appearing in those 2000 transactions that are flagged as outliers is defined as the True Positive (TP) ratio. The results exhibited in Fig. 2 report the False Positive ratio (solid line), and the True Positive ratio (dotted line), for two runs of the algorithm with different values for the forgetting factor. In the top graph of Fig. 2, the forgetting factor was set to emphasize recent transaction activity while in the bottom graph of Fig. 2, the forgetting factor was set to emphasize historical data.

Regardless of the choice of forgetting factor, certain fraudulent transactions can be detected. These transactions are dissimilar from other transactions regardless of the time they appear. The choice of forgetting factor affects which fraudulent transactions we detect. Choices of forgetting factor that rapidly adapt to immediate changes in the stream lead to a lower mean FP rate. Rapid adaptation occasionally produces very high FP rate, which

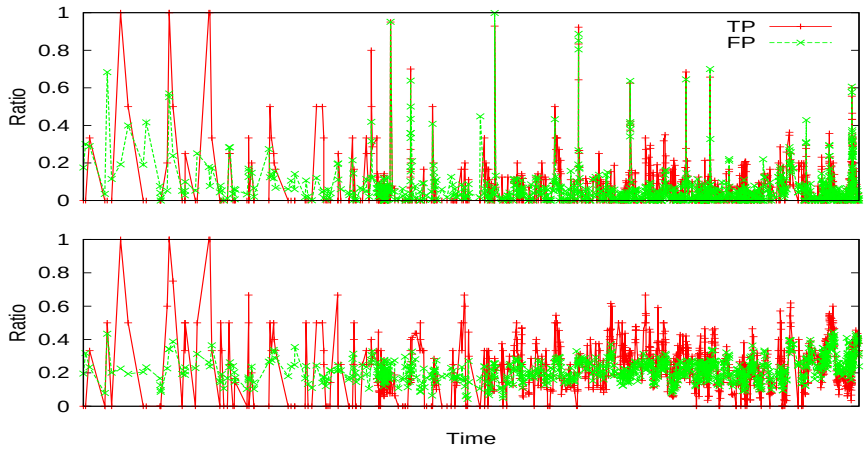


Fig. 2. Top: Fast forgetting, Bottom: Slow forgetting.

we interpret as evidence for temporally local transactions far apart in the feature space. Slower adaptation induces a more steady FP rate, indicative of a slowly changing structure of the streams. Slower adaptation also appears to enhance fraud detection in the sense that TP rate is higher than the FP rate.

5 Discussion

In this paper we propose stream clustering as an innovative feature extraction method to enhance fraud detection methodology. Such innovation is essential in battling the adaptive behaviour of fraudsters. The streaming tools we propose attempt to adapt to both the measurement and temporal structure of the data. In this way, we hope to reveal different aspects of the character of fraud.

Our experimental results, based on real data, suggest that it is possible to deploy these streaming clustering methods for knowledge discovery. Strikingly, different rates of forgetting appear to reveal different types of fraud structure. This mined information can be utilized to produce novel features for fraud detection. For example, determining if a transaction is an outlier based on the clustering structure for the rest of the stream, provides a binary variable which enriches the description of the transaction, as as the experimental results show. This enriched representation could be passed to standard fraud detection technology.

6 Acknowledgements

The work of Dimitris Tasoulis work was undertaken as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Systems) project and is jointly funded by a BAE Systems and the EPSRC (Engineering and Physical Research Council) strategic partnership, under EPSRC grant EP/C548 051/1. The work of David Weston was supported by grant number EP/C5 32589/1 from the UK Engineering and Physical Sciences Research Council. The work of David Hand was partially supported by a Royal Society Wolfson Research Merit Award. We would like to express appreciation to the bank that provided the fraud data.

References

APACS (2008): *Card fraud facts and figures* http://www.apacs.org.uk/resources_publications/card_fraud_facts_and_figures.html.

VISA (2003): *VISA EU launches new advanced fraud detection tool, http://www.visaeurope.com/pressandmedia/newsreleases/press178_pressreleases.jsp.*

BOLTON, R.J., and HAND, D.J. (2002): Statistical fraud detection: A review. *Statistical Science, 17(3), 235–255.*

BRAUSE, R., LANGSDORF, T., and HEPP, M. (1999): Neural data mining for credit card fraud detection. *Int. Conference on Tools With Artificial Intelligence, 103.*

Cao, F., Ester, M., Qian, W., and ZHOU, A. (2006): Density-based clustering over an evolving data stream with noise. *SIAM Conference on Data Mining, 326–337.*

JUSZCZAK, P., ADAMS, N.M., HAND, D.J., WHITROW, C., and WESTON, D.J. (2008): Off-the-peg or bespoke classifiers for fraud detection? *Computational Statistics and Data Analysis, in press.*

KOU, Y., LU, C.-T., SIRWONGWATTANA, S., and HUANG, Y.-P. (2004): Survey of fraud detection techniques. *IEEE Int. Conference on Networking, Sensing and Control, 2, 749–754.*

MAES, S., TUYLS, K., VANSCHOENWINKEL, B., and MANDERICK, B. (2002): Credit card fraud detection using Bayesian and neural networks. *International NAISO Congress on Neuro Fuzzy Technologies.*

SANDER, J., ESTER, M., KRIEGEL, H.-P., and XU, X. (1998): Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery, 2(2), 169–194.*

TASOULIS, D.K., ADAMS, N.M., and HAND, D.J. (2006): Unsupervised clustering in streaming data. *6th IEEE Int. Conference on Data Mining 638–642.*

WESTON, D.J., HAND, D.J., ADAMS, N.M., WHITROW, C., and JUSZCZAK, P. (2008): Plastic card fraud detection using peer group analysis. *Advances in Data Analysis and Classification, 2(1), 45–62.*

Part VIII

Econometrics

Modeling of the Household Incomes in the Czech Republic in 1996–2005

Jitka Bartošová and Vladislav Bína

University of Economics in Prague, Faculty of Management, Jarošovská 1117/II, 37701 Jindřichův Hradec, Czech Republic, {*bartosov, bina*}*@fm.vse.cz*

Abstract. This paper addresses the problem of Czech household income modeling in the years 1996–2005. The transformation and globalization processes experienced in the Czech Republic have led to, among other things, fundamental changes in the income structure of the inhabitants. Privatization of the economy allowed emergence of new income sources and thus a change in the social structure of households and differentiation of incomes took place. The presented contribution discusses the question of construction and usability of a simple parametric model under altered conditions.

Keywords: income distribution, probability model, iterative procedure

1 Introduction

The subject of this contribution ensued from actual questions related to the development of the Czech economy after the year 1989; the period of transition from a centrally planned economy to the free market, in which there was a change of many economic indicators. The transformation process and consequent globalization of the Czech economy within the framework of the EU also affected the social level of the inhabitants. Monitoring and analysis of social status of individuals and households is performed by means of an exploration of net incomes, which are often regarded as a better indicator of social status than net consumption. The reason for this preference is that net incomes are not affected by consumer preferences (see e.g. Longford and Pittau (2006)).

Contemporary research focuses mainly on analyzing the income development dynamics, its stability and detection of important factors affecting incomes levels. Some works concentrated on the description of income development dynamics are, e.g.: Paap and van Dijk (1998), Di Prete and McManus (2000). Stability of these dynamics in EU countries is considered in Longford and Pittau (2006). The articles of Kneip and Utikal (2001), Pittau (2004) are focused on the search for the most influencing factors and regional diversities in the EU.

Present income distributions are considerably complicated and thus many authors use more general classes for their modeling (see e.g. Longford and Pittau (2006)). Theoretical foundations of the modern mathematical and

Social class	1996		2002		2005	
	Size	%	**Size**	%	**Size**	%
Employees	15966	56.72	4094	51.35	2148	49.37
Self-employed	1879	6.68	757	9.49	391	8.99
Retired with EA members	1156	4.11	278	3.49	178	4.09
Retired without EA members	8651	30.73	2533	31.77	1425	32.75
Unemployed	260	0.92	172	2.16	131	3.01
Others	236	0.84	139	1.74	78	1.79
All	28148	100.00	7973	100.00	4351	100.00

Table 1. Structure of data sets in the years 1996 – 2005.

statistical classification methods (modeling with the use of mixtures, hidden Markov models, generalized linear and additive models) can be found in the works of McCullagh and Nelder (1994) and McLachlan and Peel (2000).

Many modern authors give priority to modeling that uses kernel density estimates (Kneip and Utikal (2001)), which allows the choice of a suitable degree of curve smoothing, thus constructing detailed models of density function. The progressive method of modeling based upon the use of properties of generalized lambda distribution quantile function (RS GLD) (see Ramberg and Schmeiser (1974), Pacáková and Sodomová (2003)), or generalized Pareto distribution (see Luceno (2006)) is also widely applied.

2 Data sets

Samples of household incomes used for the modeling were collected in two different studies – Microcensus and SILC. The sample survey Microcensus is a periodical sample survey performed every 3 to 5 years since 1957. After the Czech Republic accession to the EU, the former Microcensus was replaced by the SILC survey. In Table 1 we can see surveyed social groups as well as the decreasing size of the sample.

3 Modeling of income distributions

The planned form of the economy used prior to the revolution is characterized by its high homogeneity of population income in all social classes. After 1989, however, significant changes in income distribution among the population have been occurring due to the transformation from a planned to a market economy. The impact of regional and demographic factors is rising; incomes of households of pensioners and unemployed deserve particular attention. Significant changes that the transformation process brings about lead to impairment of the current statistical model and discontinuation of time

lines representing the distribution characteristics. This situation requires certain precautions with respect to previously valid research methods as well as conclusions that were based on those methods.

For modeling of empirical income distribution in the Czech Republic prior to 1989, two- or three-parametric lognormal models were often used. Nowadays, many authors suggest using more general classes of models. In the article Longford and Pittau (2006), the authors say that "The logarithm is the natural transformation for income data in most populations because it reduces the skew and asymmetry of its distribution. Comparisons of income are practical on the multiplicative scale, by changes expressed in percentages. In the ECHP data, the distribution of log-transformed income is close to normality, but deviations from it are perceptible. A single normal distribution may be inadequate for describing log-income. Mixtures of normal distribution form a much more general class." Thus the question arises whether the two- or three-parametric lognormal model allows us to obtain a good approximation of the income distribution of Czech households even after the year 1989.

In order to objectively decide whether the chosen model is suitable, it is necessary to estimate the model's parameters and test it with maximum accuracy. The basic aim for construction of the theoretic model is its maximum correspondence to the empirical distribution (Bartošová (2006)). Consequently, sufficient flexibility and elasticity must be included as conditions for proper choice of the model. Because of the fact that the sample sets of household incomes in years 1996 – 2005 are sufficiently large for the construction of logarithmic-normal models with two parameters μ and σ^2 or three parameters μ, σ^2 and γ (where γ is the theoretical minimum), the maximum likelihood method was applied. The maximum likelihood estimate of parameter γ of a three-parametric logarithmic-normal distribution could be calculated only numerically. Toward this aim several methods could be used:

- searching maximum of the modified log-likelihood function. In the case of sample size n

$$\tilde{\ell}(\gamma) = -n[\hat{\mu}(\gamma) + \frac{1}{2}\log\hat{\sigma}^2(\gamma)],$$

  where $\hat{\mu}(\gamma)$ and $\hat{\sigma}^2$ are maximum likelihood estimates of parameters and γ is a chosen value of theoretical minimum in the model,
- searching minimum of the likelihood ratio

$$LR(\mu, \sigma^2, \gamma|n) = 2[\ell(\boldsymbol{p}|n) - \ell(\boldsymbol{\pi}(\mu, \sigma^2, \gamma)|n)],$$

  where $\boldsymbol{p}$ is the vector of income empirical probability, $\boldsymbol{\pi}(\mu, \sigma^2, \gamma)$ are the probabilities of particular class occupation and $\ell(\boldsymbol{p}|n)$, $\ell(\boldsymbol{\pi}(\mu, \sigma^2, \gamma)|n)$ are corresponding log-likelihood functions.

Considering the character of a particular feature, we obtained the corresponding estimate by searching the maximum value of function $\tilde{\ell}(\gamma)$ (minimum of the function $LR(\gamma|n)$) in the interval $(-x_{\max}, x_{\min})$, where $x_{\min}$ and $x_{\max}$ are the minimum and maximum income values. The task was solved by iteration method – on a grid which is refined in each iteration step. This iteration procedure was implemented by the R language script.

4 Proposal of the iterative procedure

The proposed iterative procedure is composed of two cycles (outer and inner) and is realized in R language. At the beginning of the ith **outer cycle** of the iterative procedure ($i \in N$), the maximal scope of γ parameter estimate is given by the following interval

$$\left\langle \gamma_{\min}^{(i)(0)}, \gamma_{\max}^{(i)(0)} \right\rangle = \left\langle -x_{\max}^{(i)}, x_{\min}^{(i)} \right\rangle,$$

which is split by m points (m is a constant) into $m+1$ intervals of length

$$\Delta^{(i)(j)} = \frac{\gamma_{\max}^{(i)(j)} - \gamma_{\min}^{(i)(j)}}{m+1}.$$

In the first phase of jth inner cycle of the iterative procedure, the values $\hat{\mu}(\gamma_k^{(i)(j)})$, $\hat{\sigma}^2(\gamma_k^{(i)(j)})$ and $\tilde{\ell}(\gamma_k^{(i)(j)})$ are given at each grid point $\Delta^{(i)(j)}$ for $j = 0, 1, \ldots J$. Then such a point $\hat{\gamma}^{(i)(j)}$ is chosen where the reduced logarithmical likelihood function achieves its maximum so that the following equality holds

$$\tilde{\ell}(\hat{\gamma}^{(i)(j)}) = \max_{\gamma_k^{(i)(j)}} \tilde{\ell}\left(\gamma_k^{(i)(j)}\right).$$

The second phase of the iterative cycle consists of increasing accuracy for the maximum likelihood estimates concerned. Within each $(j+1)$th step, the improvement is given by the reduction and shift of the searching interval $\left\langle \gamma_{\min}^{(i)(j)}, \gamma_{\max}^{(i)(j)} \right\rangle$ to the neighborhood of the retrieved grid maximum $\hat{\gamma}^{(i)}$ so that the new bounds fulfill the conditions

$$\gamma_{\min}^{(i)(j+1)} = \hat{\gamma}^{(i)(j)} - \Delta^{(i)(j)} \qquad \text{and} \qquad \gamma_{\max}^{(i)(j+1)} = \hat{\gamma}^{(i)(j)} + \Delta^{(i)(j)}.$$

Then the whole process is iterated. The inner cycle ends when the maximum of likelihood function $\tilde{\ell}(\hat{\gamma}^{(i)(J)})$ is achieved with the prescribed precision. The accuracy of maximum achievement is given by the choice of exponent α which gives the total decreasing of the search scope

$$\frac{\Delta^{(i)(J)}}{\Delta^{(i)(0)}} \leq 10^{-\alpha}.$$

The **outer cycle** continues by deciding whether the retrieved value $\tilde{\ell}(\hat{\gamma}^{(i)(J)})$ is the maximum on the whole interval $(-\infty, x_{\min})$.

Social class	LR 1996		LR 2002		LR 2005	
	$\gamma = 0$	γ_{LR}	$\gamma = 0$	γ_{LR}	$\gamma = 0$	γ_{LR}
Employees	477.3548	349.9550	76.1012	73.5210	59.7239	51.3693
Self-employed	99.9104	99.8761	52.3678	52.3581	27.3061	27.2604
Retired with EA	32.6141	21.3661	25.8631	16.4592	21.8096	21.6421
Ret. without EA	3874.4552	3698.5442	1174.2735	1113.5243	565.2051	535.0171
Unemployed	29.2938	23.0240	15.0465	15.0458	7.3785	7.2391
Others	26.8715	25.3154	11.0360	10.5905	13.7848	12.3391
All	3239.8009	3224.4095	790.2746	756.7107	398.9364	397.6695

Table 2. Comparison of conformity of empirical distribution with two- and three-parametric logarithmic-normal models. (Incomes per household, 1996 – 2005.)

- If $\hat{\gamma}^{(i)(J)} \leq \gamma_{\min}^{(i)(0)}$ then $\left\langle \gamma_{\min}^{(i+1)(0)}, \gamma_{\max}^{(i)(0)} \right\rangle = \left\langle \gamma_{\min}^{(i)(0)} - \Delta^{(i)(0)}, \gamma_{\min}^{(i)(0)} \right\rangle$ and the procedure repeats once more.
- If $\hat{\gamma}^{(i)(J)} > \gamma_{\min}^{(i)(0)}$, then the procedure stops.

Insertion of the outer decision cycle into the iterative procedure ensures convergence to maximum likelihood estimates of parameter vector components $\boldsymbol{\theta} = \left(\mu, \sigma^2, \gamma\right)$ regardless of the initial choice of the search scope $\left\langle \gamma^{(1)(0)}, \gamma_{\max}^{(1)(0)} \right\rangle$.

5 Validity of the logarithmic-normal models

Validity of the logarithmic-normal models was quantified by the statistic LR. The results are also influenced by the number of classes in which data are gathered during calculation. The problem of optimizing the number of classes m is a subject of many papers. In this case we chose $m = 15\sqrt[5]{(n/100)^2}$, which is suitable for a sufficiently large sample, i.e. for $n > 80$ (see Williams (2001)).

In Table 2 the values of likelihood ratio LR are written for the constructed logarithmic-normal models with two and three parameters. In all social groups, greater agreement of empirical distribution with the model was achieved for three-parametric logarithmic-normal models. The use of three-parametric logarithmic-normal models with the above-described iterative estimation of parameter γ led to an improvement of the models' validity. In the iterative procedure, the method of minimizing the likelihood ratio was used in order to estimate the value of parameter γ.

Estimates of the three parameters of logarithmic-normal models of incomes in Czech households in the years 1996–2005 are listed in Tables 3–5. For consideration of the models' validity, the estimates of μ, σ^2 and γ are supplemented by the values of LR statistics and 95% quantiles of $\chi^2_{0.95}(m-4)$.

From Tables 3–5 we could infer that in most cases the values of LR and $\chi^2_{0.95}(m-4)$ are comparable. The strong discrepancy between the empirical

Social class	Parameter			Statistic	Quantile	Number
	μ	σ^2	γ	LR	$\chi^2_{0.95}(m-4)$	of classes
Employees	12.2385	0.1543	-30133	349.9550	135.4802	114
Self-employed	12.2292	0.3470	1062	99.8761	60.4809	48
Retired with EA	11.7822	0.1904	36141	21.3662	50.9985	40
Retired without EA	10.9931	0.2222	13331	3698.5442	107.5217	89
Unemployed	11.5093	0.2359	-21716	23.0240	28.8692	22
Others	11.0788	0.5114	4239	25.3154	27.5871	21
All	11.7562	0.4042	4343	3224.4095	167.5143	143

Table 3. Estimates of the parameters for the three-parametric logarithmic-normal models. (Incomes per household, 1996.)

Social class	Parameter			Statistic	Quantile	Number
	μ	σ^2	γ	LR	$\chi^2_{0.95}(m-4)$	of classes
Employees	12.4654	0.2095	-11159	73.5210	81.3810	66
Self-employed	12.5610	0.3296	1316	52.3581	43.7730	34
Retired with EA	11.9953	0.2783	87257	16.4592	30.1435	23
Retired without EA	11.3734	0.2482	28477	1113.5243	68.6693	55
Unemployed	11.5123	0.3912	297	15.0458	24.9958	19
Others	11.4531	0.4073	8999	10.5905	22.3620	17
All	12.0783	0.4181	12933	756.7107	104.1387	86

Table 4. Estimates of the parameters for the three-parametric logarithmic-normal models. (Incomes per household, 2002.)

Social class	Parameter			Statistic	Quantile	Number
	μ	σ^2	γ	LR	$\chi^2_{0.95}(m-4)$	of classes
Employees	12.6150	0.1756	-34033	51.3693	64.0011	51
Self-employed	12.7075	0.3404	-4562	27.2604	33.9244	26
Retired with EA	12.4765	0.1022	29636	21.6421	24.9958	19
Retired without EA	11.4014	0.2919	36890	535.0171	54.5722	43
Unemployed	11.4730	0.4600	6215	7.2391	22.3620	17
Others	11.8529	0.2687	-27916	12.3391	18.3070	14
All	12.2010	0.3803	4818	397.6695	83.6753	68

Table 5. Estimates of the parameters for the three-parametric logarithmic-normal models. (Incomes per household, 2005.)

distribution and the model appears only in the case of **retired without economically active members** and **all households**. Those income sets have bimodal distribution (see Figure 1) and could not be modeled using simple parametric models. Modeling of such mixtures is the topic of a previous paper (Bartošová and Bína (2007)).

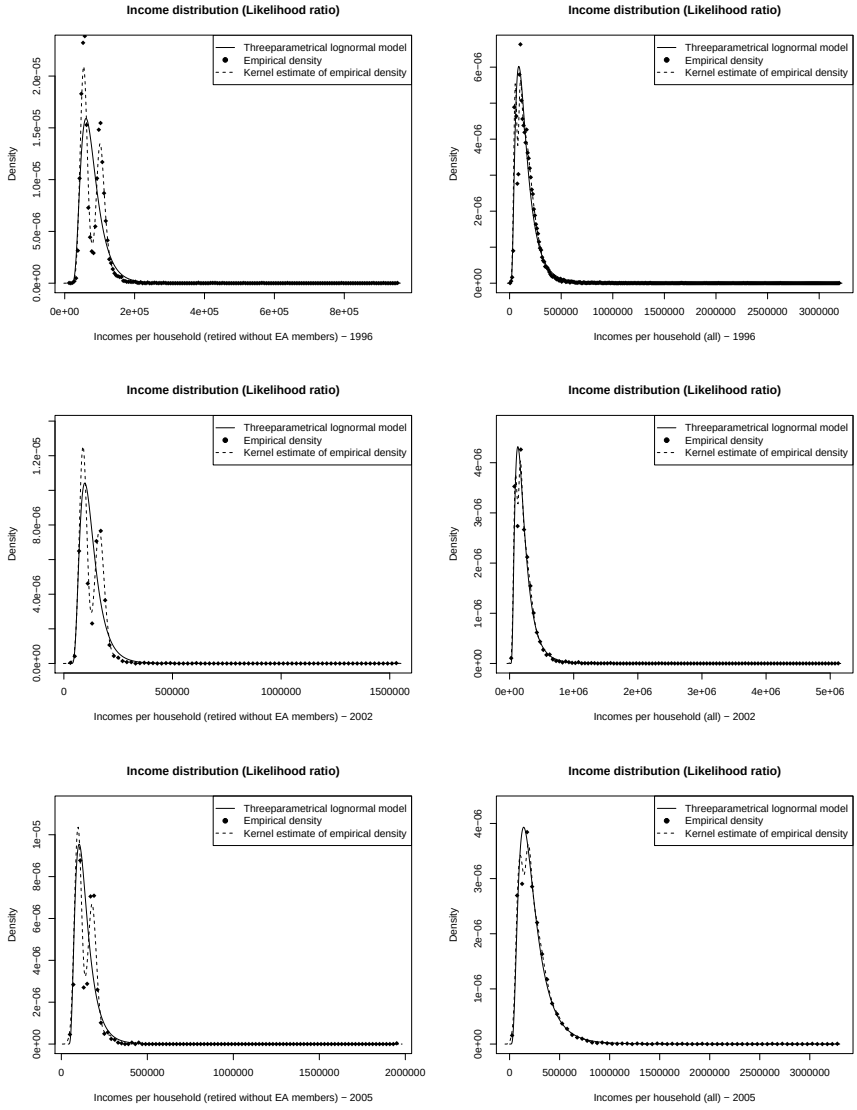


Fig. 1. Three-parametric model, empirical density and kernel estimate of empirical density. (Retired without EA members and all households in 1996–2005.)

6 Conclusions

In most cases the presented results of modeling using a three-parametric model shows good agreement with the empirical distribution. Only in the

case of **retired without economically active members** and in the case of **all households** the distribution is bimodal and significant discrepancy is shown. In all surveyed cases, the three-parametric model was better than the two-parametric one. The use of the three-parametric model always led to an improvement in the resulting model's validity. In the iterative procedure the minimization of the likelihood ratio was used for estimating γ. Similar results could be achieved using maximization of the log-likelihood function.

Acknowledgement

The work was partly supported by Czech Ministry of Education, Youth and Sports under grant no. 2C06019 "ZIMOLEZ".

References

BARTOŠOVÁ, J. (2006): Logarithmic-normal model of household income distribution in the Czech Republic after 1990. *Forum Statisticum Slovacum 3*, Slovak Statistical and Demographical Society, Bratislava, 3–10.

BARTOŠOVÁ, J., BÍNA, V. (2007): Mixture models of household income distribution in the Czech Republic. In: M. Kováčová (Ed.): *6th International Conference APLIMAT 2007, Part I.* Slovak Univ. of Technology, Bratislava, 307–316.

DI PRETE, T.A., MCMANUS, A. (2000): Family change, employment transition, and welfare state: household income dynamics in the United States and Germany. *American Sociological Revue 65, 343–370.*

KNEIP, A., UTIKAL, K.J. (2001): Inference for density families using functional principal component analysis. *Journal of the American Statistical Association 96 (454), Theory and Methods, 519–533.*

LONGFORD, N.T., PITTAU, M.G. (2006): Stability of household income in European countries in the 1990s. *Computational Statistics & Data Analysis 51, 1364–1383.*

LUCENO, A. (2006): Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. *Computational Statistics & Data Analysis 51, 904–917.*

MCCULLAGH, P., NELDER, J.A. (1994): *Generalized Linear Models.* Chapman and Hall, London.

MCLACHLAN, G., PEEL, D. (2000): *Finite mixture models.* John Wiley & Sons, New York.

PAAP, R, van DIJK, H.K. (1998): Distribution and mobillity of wealth of nation. *European Economic Revue 42, 1269–1293.*

PACÁKOVÁ, V., SODOMOVÁ, E. (2003): Modelling with quantille distribution functions. *Ekonomika a informatika 1, 30–44.*

PITTAU, M.G. (2004): Regional income distributions in the European Union. *Oxford Bulletin Economic Statistics 67, 135–161.*

RAMBERG, J., SCHMEISER, B. (1974): An approximate method for generating asymmetric random variables. *Communications of the ACM 17 (2), 78–82.*

WILLIAMS, D. (2001): *Weighing the Odds, A Course in Probability and Statistics.* Cambridge Univ. Press, Cambridge.

Detecting Social Interactions in Bivariate Probit Models: Some Simulation Results

Johannes Jaenicke

Faculty of Economics, Law and Social Sciences, University of Erfurt
Nordhäuser Straße 63, D-99089 Erfurt, Germany,
Johannes.Jaenicke@uni-erfurt.de

Abstract. This paper analyzes the possibility of detecting observable and non-observable social interactions in a bivariate probit model with an endogenous dummy regressor via Monte Carlo simulation. The main result is that in small samples, we only find low probability of detecting observable and non-observable social interactions. In large samples, however, we find the z-parameter test to be very powerful.

Keywords: parameter tests, bivariate probit model, Monte Carlo study
JEL classification: C35, C15

1 Introduction

For the researcher, interactions between two persons, e.g., spouses or brothers and sisters may only partly be observable, due to psychological reasons. In order to detect the neglected or non-observable interactions between the respective decision processes a bivariate probit model is recommendable (Jaenicke, 2004). In small samples with 76 or 132 observations (Dean, 1995, and Greene, 1998), parameter tests in a bivariate model may have bad size and power properties.

Our intention is to find out whether in the presence of social interactions in the data, it is possible to detect these interactions in a bivariate probit model with an endogenous dummy regressor. Hence we analyze the power of the usual z-coefficient test concerning the parameters of the observable and non-observable interactions, i.e. endogenous dummy variable and the residual covariance between both equations of this bivariate probit model.

2 A bivariate probit model of social interactions

The maximum likelihood estimation of a bivariate probit model involves the numerical problem of the evaluation of double integrals over the normal distribution. This estimation procedure is implemented in several statistic software packages and widely used in practice. We use a two equation binary choice

model with an endogenous dummy regressor, first proposed by Maddala and Lee (1976). The regression equations of the individual I and the peer P are

$$\begin{aligned} Y_I^* &= X_I\beta_1 + Y_P\beta_2 + u_I, \quad & Y_I = 1 \text{ if } Y_I^* > 0, 0 \text{ otherwise} \\ Y_P^* &= X_P\gamma_1 + u_P, & Y_P = 1 \text{ if } Y_P^* > 0, 0 \text{ otherwise} \\ [u_I, u_P] &\sim \Phi_2\left(0, 0, 1, 1, \rho\right), \end{aligned}$$

with the observable discrete choice behavior Y, latent variables Y_I^*, and exogenous variables X. The residual vector $[u_I, u_P]$ is bivariate normal distributed with $E(u_i) = 0$, $var(u_i) = 1$, $i = I, P$, and $cov(u_I, u_P) = \rho$. As a condition of identification, we only need exclusion restrictions if there is no variation of the exogenous regressors (Wilde, 2000).

In our model, the observable part of the social interactions, the influence of the decision of the peer P on the behavior of the individual I is tested by the hypothesis $H_0 : \beta_2 = 0$. The non-observable part of the social interactions may be revealed through the residual covariance structure. A residual covariance $cov\left(u_I, u_P\right)$, i.e. ρ, significantly different from zero, may serve as an indicator of unobserved social interactions between the two decisions or as an indicator of simultaneously neglected third-party effects. Restricting residual correlation of the bivariate probit model to zero may result in biased and inconsistent estimations (Murphy, 1995). Fitting separate probit models for the first- and the second decision equation can involve significant endogeneity biases in the estimation (Lollivier, 2001). The joint estimation of the two equations provides substantial efficiency gains compared to separate estimation based on two-stage technique. Such estimation accounts for potential correlation between the two decisions (Hoffnar and Greene, 1995).

3 Monte Carlo results for the bivariate probit model

In a small Monte Carlo study, we analyze the size and the power of the usual z-coefficient tests concerning the parameters of the observable and non-observable social interactions, β_2 and ρ. The test statistics are $z\left(\beta_2\right) = \hat{\beta}_2/se(\hat{\beta}_2) \xrightarrow{d} N(0,1)$ and $z\left(\rho\right) = \hat{\rho}/se(\hat{\rho}) \xrightarrow{d} N(0,1)$ and its square gives the Wald test (Greene 2008, p. 820).

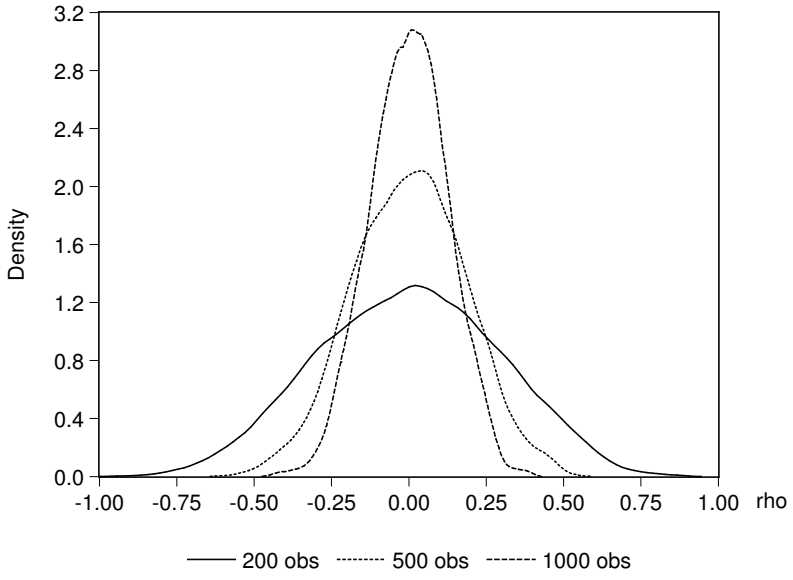
The non-observable influences stem from missing variables. These may be uncorrelated, weakly or strongly correlated or identically for the two persons. To create the non-observable interactions, we use an omitted variable vector $[v_I, v_P] \sim \Phi_2\left(0, 0, 1, 1, r\right)$ with $r\epsilon[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ in one set of experiments. In this case, the residuals u_i, $i = I, P$, are the sum $u_i = v_i + \varepsilon_i$ with $[\varepsilon_I, \varepsilon_P] \sim \Phi_2\left(0, 0, 1, 1, 0\right)$, therefore $[u_I, u_P] \sim \Phi_2\left(0, 0, 2, 2, \frac{r}{2}\right)$. Because the assumption of the unit residual variance $var\left(u_i\right)$ is not met, we expect some problems resulting from the misspecification of the model.

In the experiments with the extended model, we include v_i as additional explanatory variables. In this case, the residuals are $u_i = \varepsilon_i$ and are independent normal distributed. Because of the independence of the resiuduals,

$\rho = 0$, the model is overparametrized. Two single equation models would be more efficient. Anyway, since we do not know the true parameter set in the empirical research situation, in the simulation experiments we remain in the bivariate probit model class.

The variables $X_i, i = I, P$ are standard normal distributed, $X_i \sim N(0, 1)$, $i = I, P$. All parameters β and γ in the omitted variable model and in the extended model are set equal to one. We use the econometric software package Limdep 7.0. It performs well in nonlinear estimation benchmark tests (McCullough, 1999). We estimate the bivariate models with the default settings of the procedure (algorithm: BFGS, maximum iterations: 100). The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is rather time consuming, but it shows a convergence rate of between 99.5 percent (in small data sets with 100 observations) and 100 percent (in data sets with 10,000 observations) in our Monte Carlo study. The number of replications in the Monte Carlo experiment is N=1000.

Figure 1: Kernel estimation, rho=0

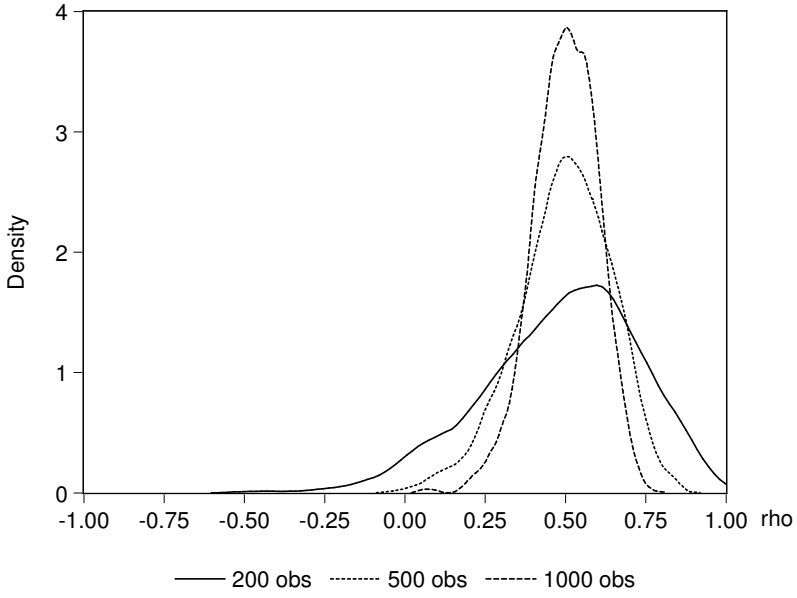


The estimated parameters $\widehat{\rho}$ show no severe bias. Figure 1 presents the density estimation with the Epanechnikov kernel function in the case that the true parameter $\rho = 0$. With increasing sample size from $T = 200$ to $T = 1000$, the dispersion of the estimated parameters becomes smaller. The parameters are distributed more or less symmetrically (with skewness S_T between -0.083

and -0.040) and means (with $\overline{\widehat{\rho_T}}$ between -0.001 and 0.003) very close to the theoretical value zero.

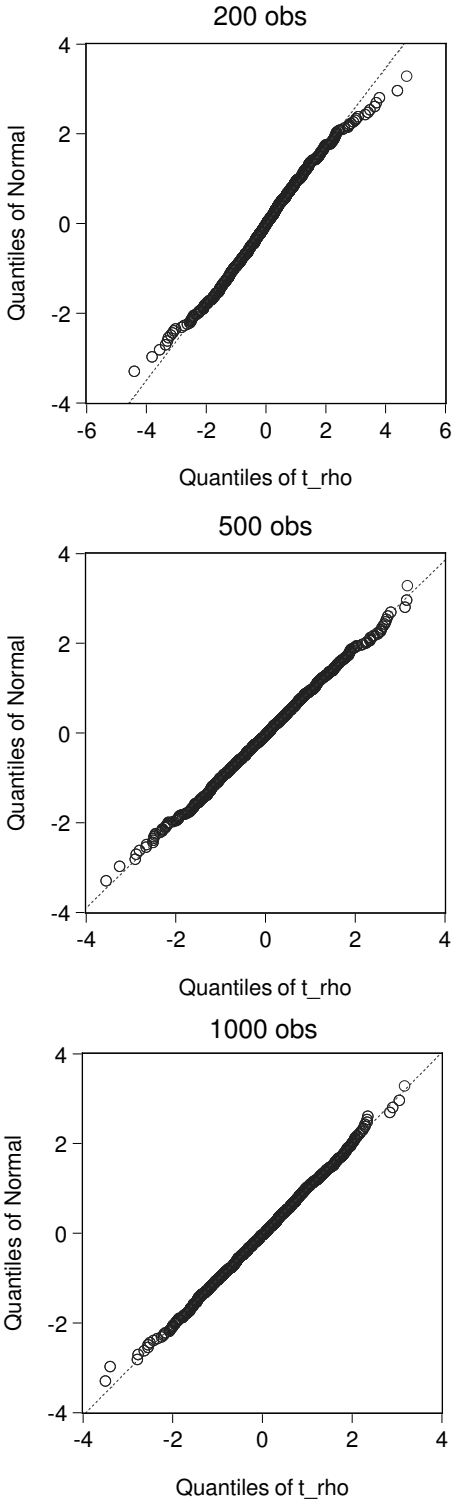
The picture changes if we assume with $\rho = 0.5$ strong residual correlation between both decision equations in small data sets. In the case of $T = 200$, we find with $\overline{\widehat{\rho_{200}}} = 0.485$ some deviations from the theoretical parameter value. In all three cases, the distribution of the estimated parameters $\widehat{\rho_{200}}$ is left skewed with a skewness S_T between -0.568 and -0.403. In the $T = 200$ case, some $\widehat{\rho_{200}}$ are very close the theoretical limite of 1. In Figure 2, we present a Kernel density estimation for $\rho = 0.5$ and $T = 200$, $T = 500$ and $T = 1000$.

Figure 2: Kernel estimation, rho=0.5



Looking at the z-statistics in the case $\rho = 0$, we find some indication that the $z-$statistics may not be normally distributed in small samples. To analyze this graphically, we compare the quantiles of these statistics with the standard normal distribution in figure 3. Especially in the case of $\rho = 0, T = 200$ observations, we find some deviation from normality, not only at the tails of the distribution. In this case skewness $S_{200} = 0.092$ and the kurtosis $K_{200} = 3.945$.

Figure 3: QQ-Plots for t_rho
rho=0, and T=200,500,1000



The finite sample behavior of the $z-$parameter test for ρ and β_2 is summarized in the table 1. We set the nominal significance level of the tests equal to 5 percent. The first two blocks of this table presents the size and power of $z(\rho)$ and the power of $z(\beta_2)$ in in the omitted variable model. The power results are not size adjusted. The size distortions of the test $z(\rho)$ are relatively small in medium and large data sets, but more pronounced in the 200 observations case.

Table 1: Size and power of z-parameter tests concerning the social interactions β and ρ in the bivariate probit model (in percent).

correlation	model with omitted variables v_i						extended model		
$\text{corr}(u_I, u_P)$	size and power of $z(\rho)$			power of $z(\beta_2)$			power of $z(\beta_{2ext})$		
$= \rho$	T=200	T=500	T=1,000	T=200	T=500	T=1,000	T=200	T=500	T=1,000
0.50	52.2	86.1	99.6	34.2	74.9	97.0	55.9	93.5	100.0
0.45	50.2	76.5	96.0	33.2	72.8	96.2	62.4	97.2	100.0
0.40	38.4	67.0	91.6	34.8	74.9	97.6	66.3	96.9	100.0
0.35	32.7	56.4	82.5	33.0	72.9	96.5	69.8	99.0	100.0
0.30	28.3	47.8	72.0	30.0	68.8	95.4	72.7	98.8	100.0
0.25	19.4	34.6	61.2	31.6	70.1	95.8	73.7	98.8	99.8
0.20	16.8	25.5	39.0	34.5	72.0	94.6	74.6	99.4	100.0
0.15	13.2	16.9	25.1	35.9	69.6	95.0	76.6	99.4	100.0
0.10	11.0	10.5	13.4	36.1	72.2	94.8	77.3	99.5	100.0
0.05	9.6	6.9	8.2	37.9	72.7	95.8	78.3	99.6	100.0
0.00	8.1	5.7	4.6	37.3	72.4	95.8	76.8	99.3	100.0

We find that the power of these tests varies with the correlation ρ and is very low in a sample size of $T = 200$ observations. For example, in the case of $\rho = 0.4$, only 38.4 percent of the true ρ-coefficients and 34.8 percent of the true β_2 -coefficients are significantly different from zero at the 5-percent significance level. In the case of $\rho = 0.1$, only 11.0 percent of the ρ-coefficients and 36.1 percent of the β_2 -coefficients are significant.

The power can be dramatically increased (e.g., to 77.3 percent in the case of $T = 200$ and $\rho = 0.1$) if it is possible to find the neglected variables that cause high residual correlation (see table 1, third block). This is true although the extended model is overparametrized because the true correlation-coefficient is zero in this model. In the data set of $T = 1,000$ observations, the power of the test $z(\beta_2)$ concerning the obervable social interactions is very high, i.e. between 94.6 and 97.6 percent in the omitted variable model and nearly always 100 percent in the extendet model.

To study the dependence of the test from the sample size in detail, we run several experiments for the case of weak correlation ($\rho = 0.1$) between the two equations. The sample size ranges from 100 to 10,000 observations. From the Monte Carlo experiments presented in table 2, we see that only in the case of 10,000 observations we have a good chance (72.2 percent) to detect the correlation between the two equations with the $z(\rho)$-test. In all other

cases, the probability is less than a half. In cases of 100 to 500 observations, the probability is about ten percent.

Table 2: The dependence of the size and power from the sample size T: The case of $\rho = 0.1$ (in percent).

T	100	150	200	250	500	1,000	1,500	2,000	2,500	5,000	10,000
$z(\rho)$	13.6	10.5	11.0	11.5	10.5	13.7	17.0	22.2	26.7	44.6	72.2
$z(\beta_2)$	22.6	28.3	36.1	42.7	72.2	94.5	100.0	100.0	99.9	100.0	100.0
$z(\beta_{2ext})$	43.9	63.1	77.3	88.1	99.5	100.0	100.0	100.0	100.0	100.0	100.0
$z(\beta_2 - \beta_{2ext})$	12.4	12.2	12.4	17.6	27.5	47.2	61.0	75.5	80.5	98.3	100.0

The power of the test $z(\beta_2)$ in the omitted variables model is less than one third in the case of 100 or 150 observations. With 1,000 or more observations, however, the power of the test is quite good. In the extended model, the power of $z(\beta_{2ext})$ is far better and lies between 43.9 and 100 percent.

In the extended model, we additionally raise the question whether β_2 differs significantly from β_{2ext}. The results are presented in the fourth row of table 2. It becomes obvious that omitted variables cause significant deviations from the estimated β_{2ext} in the extended model. In large data sets, the hypothesis of equal β_2-coefficients is rejected in about 100 percent of all cases. This underlines the importance of searching for possibly neglected variables.

Our results are only partly in line with a recent Monte Carlo study by Monfardini and Radice (2006) using STATA. To obtain maximum likelihood estimates of the bivariate probit model with an endogenous dummy regressor, they replaced the Newton-Raphson algorithm of a STATA procedure and used a Hessian-based estimator of the asymtotic covariance matrix. They find severe size distortions of some Lagrange Multiplier tests with, e.g., deviations from the 5 percent level by factor 8 up to 100, depending on the data generation mechanism for $T = 500, 1000, 2,000$. In the case of the Wald-test, the nominal $p-$level differs from the empirical one by factor up to 3. Our rejection rates for the Wald-test, using the BFGS algorithm with LIMDEP to estimate a smaller bivariate probit model are obviously better. With the same number of observations and looking at the same nominal level, we only find deviations up to 14 percent.

Our power results are, however, comparable with the one of Monfardini and Radice (2006), looking at their first (easiest) data generating mechanism. They analyze $\pm\rho = 0.25, 0.5, 0.75$ and present with $\pm\rho = 0.5, 0.75$ only the strong-correlation cases. Looking at the size adjusted power results in the study, the Wald test was often the best or second best performer.

4 Conclusions

In our paper, we find that the power of z-parameter tests concerning the residual correlation between the two decision equations in the bivariate probit

model is very low in small samples. This is is especially true if this correlation is only weak. The power of the parameter test concerning the endogenous dummy variable is around one third in small samples. If it is possible to find omitted variables, the power of this test can be increased notedly.

From an empirical point of view, we may often fail to find significant social interactions in the data sets although they exist. An extensive search for omitted variables may therefore be essential to prove social interactions in empirical models.

Acknowledgements

The author wishes to thank two anonymous referees of this paper, the participants of the discussion at the 3rd IASC world conference on Computational Statistics and Data Analysis, Cyprus 2005, and Pfingsttagung der Deutschen Statistischen Gesellschaft, Hamburg 2006, and Rose-Gerd Koboltschnig, IHS Carinthia, for valuable comments. I am especially grateful to Margot Petersen-Jaenicke for her valuable help.

References

DEAN, J.M. (1995): Market disruption and the incidence of VERs under the MFA, *Review of Economics and Statistics* 72, 383-88.

MONFARDINI, C. and RADICE, R. (2006): Testing exogeneity in the bivariate probit model: A Monte Carlo study, Working paper, University of Bologna.

GREENE, W. (1998): Gender economics courses in liberal art colleges: Further results, *Journal of Economic Education* 29, 291-300.

GREENE, W.H. (2008): Econometric Analysis, 6th ed., Pearson, Prentice Hall.

HOFFNAR, E. and GREENE, M. (1995): The effect of relative group size on the employment prospects of African-American and white males, *Review of Regional Studies* 25, 207-218.

MCCULLOUGH, B.D. (1999): Econometric software reliability: EViews, LIMDEP, SHAZAM and TSP, *Journal of Applied Econometrics* 14, 191-202.

JAENICKE, J. (2004): *Observable and non-observable social interactions in labor supply*, Discussion paper No. 2003/05, Rev. version, May 2004, University of Osnabrück.

LOLLIVIER, S. (2001): Endogénéité d'une variable explicative dichotomique dans le cadre d'un modèle probit bivarié, *Annales d'Économie et de Statistique* 62, 251-269.

MADDALA, G.S. and LEE, L.-F. (1976): Recursive models with qualitative endogenous variables, *Annals of Economic and Social Measurement* 5, 525-545.

MURPHY, A. (1995): Female labour force participation and unemployment in Northen Ireland: Religion and family effects, *Economic and Social Review* 27, 67-84.

WILDE, J. (2000): Identification of multiple equation probit models with endogenous dummy regressors, *Economics Letters* 69, 309-312.

Distributional Least Squares Based on the Generalized Lambda Distribution

Pier Francesco Perri and Agostino Tarsitano

Department of Economics and Statistics, University of Calabria
Via P. Bucci, Cubo 0C, 87036 Arcavacata di Rende (CS), Italy,
pierfrancesco.perri@unical.it, *agotar@unical.it*

Abstract. Ordinary least squares is an optimal procedure in many senses when the stochastic component has a Gaussian distribution or when linear estimates are required (Gauss-Markov Theorem). Nevertheless, departures from normality are quite plausible in many situations. In this paper, we propose an iterative procedure for estimating the regression coefficients modelling the residual term by a five-parameter version of the generalized lambda distribution. Distributional and regression parameters are estimated in a unique procedure and the effectiveness of the technique is analyzed on real and simulated data.

Keywords: quantile function, controlled random search, error distribution

1 Introduction

The validity of linear regression models in finite samples is built on the utopian hypothesis that the disturbances are normally distributed. This assumption is evoked in statistical inference procedures concerning the regression parameters, the interval prediction of the response variable as well the model fitting. Moreover, the normality of the disturbances is a convenient assumption when minimum variance unbiased estimates are required (Gauss-Markov Theorem).

Recent findings suggest that the most commonly used methods of estimation exhibit varying degrees of nonrobustness to certain violations of the normality assumption. Despite the large and expanding literature on robust estimation, many researchers continue to employ traditional methods of analysis. A good deal of this lack of acceptance can be attributed to absence of familiarity and computational convenience, but there also appears to be a suspicion that the error distributions necessitating the use of robust methods are only rarely encountered in practice.

Usually, the estimates of the regression parameters are obtained with no specific reference to the distribution of the disturbances. In this sense, the regression models could be referred to as semiparametric methods because only the deterministic part is parametrically defined. If a distribution is needed for the stochastic term, it is usually the Normal with zero mean and standard

deviation estimated from the data. Gilchrist (2000) noted that the semiparametric approach focuses on the deterministic component of the model and the estimated residuals are used to provide information on the error distribution in a totally separate exercise.

In this paper, according to Gilchrist's (2000) seminal idea, we intend to implement a least squares regression procedure based on a distributional approach. The stochastic component is parametrically defined by a quantile function without any requirement of symmetry, finiteness of the variance or support to $(-\infty, +\infty)$. This approach is known as distributional least squares (DLS) regression. According to it, we model the unknown distribution of the errors by a five-parameter version of the generalized lambda distribution (FPLD) discussed, for instance, in Gilchrist (2000). The FPLD is of practical relevance in many fields and applications since it has been found to adapt to a wide variety of theoretical and practical distributions. Hence, it is useful for the representation of data when the underlying model is unknown because it avoids the need to make an *a priori* choice among the embedded cases.

Our method involves many repeated fits using different parametric specification for the stochastic term of the regression model, each fit providing an estimate of the regression parameters. These, in turn, generate a sample of "observed" errors which are used to estimate the parameters of the distribution underlying the stochastic component. We look across the different fits for that one yielding the minimum sum of squared residuals.

The content of the present paper is organized as follows. In Section 2 we describe the estimation algorithm for the DLS regression. In Section 3 we compare the performance of the proposed method with the ordinary least squares (OLS). The effectiveness of the methods is analyzed on simulated data. Section 4 concludes the paper with some final considerations.

2 Distributional regression

The conceptual model *observation=deterministic component+stochastic component* underlies most uses of regression analysis. The observable data are assumed to be generated by the model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where $\{\mathbf{x}_i\}_{i=1}^n$ is a sequence of n design vectors of dimension m and $n > m+1$. The symbol $\boldsymbol{\beta}$ denotes a conformable vector of unknown parameters and $\{e_i\}_{i=1}^n$ is a set of unobservable independent and identically distributed random variables with cumulative distribution function F and quantile function Q. We suppose that the disturbances are with zero mean and that are uncorrelated with the individual predictors.

Usually, the estimates of $\boldsymbol{\beta}$ are obtained with no specific reference to the distribution of the residuals and only the deterministic component is

parametrically defined. Gilchrist (2000) suggested using a quantile function to model (1)

$$y = \mathbf{x}'\boldsymbol{\beta} + Q\left(p, \boldsymbol{\lambda}\right)$$

where $p = F\left(y - \mathbf{x}'\boldsymbol{\beta}\right)$ and $\boldsymbol{\lambda}$ is a vector of unknown parameters ($\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ are assumed to be independent). The merit of this formulation over standard regression schemes is that all the equational and distributional parameters, arising in the estimation procedure, are explicit in the one equation that defines the model. In such a way, both the regression and distributional parameters may be simultaneously estimated.

In this work, we adopt a five-parameter version of the generalized lambda distribution (FPLD) defined by its quantile function

$$Q\left(p, \boldsymbol{\lambda}\right) = \lambda_1 + \frac{\lambda_2}{2}\left\{\left(1 - \lambda_3\right)\left(\frac{p^{\lambda_4} - 1}{\lambda_4}\right) - \left(1 + \lambda_3\right)\left[\frac{\left(1 - p\right)^{\lambda_5} - 1}{\lambda_5}\right]\right\} \quad (2)$$

which is a flexible and manageable tool for modeling a broad class of empirical and theoretical distributions (Gilchrist, 2000). If $\lambda_2 \geq 0$ and $\lambda_3 \in [-1, 1]$ then (2) is a continuous and increasing function of p. Here, λ_1 controls, albeit not exclusively, the location of $Q\left(p, \boldsymbol{\lambda}\right)$; λ_2 is a scale parameter, while $\lambda_3, \lambda_4, \lambda_5$ influence the shape of $Q\left(p, \boldsymbol{\lambda}\right)$. We observe that parameters $\lambda_1, \lambda_2, \lambda_3$ appear in the linear form whereas λ_4 and λ_5 are in the nonlinear form.

2.1 Distributional least squares

To investigate the relationship between y and $\mathbf{x}$ in (1), we resort to the well-known least squares criterion

$$\min_{\boldsymbol{\beta},\boldsymbol{\lambda}} S\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = \min_{\boldsymbol{\beta},\boldsymbol{\lambda}} \sum_{i=1}^{n} \left[\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) - Q\left(p_{r_j}, \boldsymbol{\lambda}\right)\right]^2 \quad (3)$$

where r_j denotes the anti-rank of $e_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, that is $r_j = i$ if, and only if, e_i is the j-th smallest of $e_1, ..., e_n$ in the set of the n residuals. According to Gilchrist (2000), we define this approach distributional least squares (DLS). The stochastic component in (3) contains two exponential parameters (λ_4, λ_5) and there exists a complex interaction between $\{p_{r_j}\}_{j=1}^{n}$ and $\{-\mathbf{x}_i'\boldsymbol{\beta}\}_{i=1}^{n}$. Consequently, $S\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right)$ is difficult to minimize because the available packages of nonlinear quantile regression cannot be applied.

The criterion (3) can be considered in a parametric form with respect to the constant term. In practice, we do not directly estimate the parameter λ_1 since it is determined by the constraint

$$E\left(e_i\right) = E\left[Q\left(p_i, \boldsymbol{\lambda}\right)\right] = \lambda_1 - 0.5\lambda_2\left(\frac{1 - \lambda_3}{1 + \lambda_5} - \frac{1 + \lambda_3}{1 + \lambda_4}\right) = 0, \quad i = 1, 2, ..., n.$$

To initialize the iterative estimation procedure, a starting value of the quantile function is requested. In absence of any information on the error

distribution, we assume $e_i^{(0)} = Q(p_{r_j}, \boldsymbol{\lambda}) = 0$ and we set $y_i^{(0)} = y_i - e_i^{(0)}$, $i = 1, 2, ..., n$. Obviously, if a preliminary reliable guess on the behavior of the residuals is available, one can exploit this information to provide a more realistic representation of the initial approximation of the quantile function. Let be $r^{(0)} = (1, 2, \cdots, n)$. The vector $\mathbf{y}$ is regressed on $\mathbf{X}$ to obtain an initial OLS estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{(0)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(0)}$. Given the current parameter estimates $\hat{\boldsymbol{\beta}}^{(0)}$, and the predicted vector $\hat{\mathbf{y}}^{(0)}$, we obtain the estimated residuals $\mathbf{e}^{(1)} = \mathbf{y}^{(0)} - \hat{\mathbf{y}}^{(0)}$ which act as a sample of independent observations from $Q(p, \boldsymbol{\lambda})$ and are used to estimate the parameters $\boldsymbol{\lambda}$. In so doing, it is useful to order the residuals $\mathbf{e}^{(1)}$ obtaining $\mathbf{r}^{(\mathbf{1})}$ as the vector of anti-ranks determined by the sorting of $\mathbf{e}^{(1)}$. The i-th ordered estimated residual may be expressed as

$$e_{(i)}^{(1)} = E\left[e_{(i)}^{(1)}\right] + \xi_i \tag{4}$$

where ξ_i is a measure of the discrepancy between the observed and modelled i-th value. Here, the errors are such that $E(\xi_i) = 0$ and $\sigma^2(\xi_i) = \sigma_i^2$. The expected value of the i-th order statistic from a FPLD is available in a closed form, so that the deterministic component of (4) may be written as

$$E\left[e_{(i)}^{(1)}\right] = \gamma_1 U_{1,i} + \gamma_2 U_{2,i} \tag{5}$$

where

$$\gamma_1 = (1 - \lambda_3)\lambda_2, \quad \gamma_2 = (1 + \lambda_3)\lambda_2 \tag{6}$$

$$U_{1,i} = (2\lambda_4)^{-1}\left[\frac{\Gamma(n+1)\Gamma(i+\lambda_4)}{\Gamma(i)\Gamma(n+1+\lambda_4)} - 1\right]$$

$$U_{2,i} = (2\lambda_5)^{-1}\left[1 - \frac{\Gamma(n+1)\Gamma(n+1-i+\lambda_5)}{\Gamma(n+1-i)\Gamma(n+1+\lambda_5)}\right].$$

In order to avoid numerical problems caused by the repeated use of the gamma function and to reduce computational time, Öztürk and Dale (1985) approximated $E\left[e_{(i)}^{(1)}\right]$ with $e_{(i)}^{*(1)} = Q(p_i, \boldsymbol{\lambda})$

$$e_{(i)}^{*(1)} = \gamma_1 V_{1,i} + \gamma_2 V_{2,i} \tag{7}$$

where γ_1 and γ_2 are as in (6) while

$$V_{1,i} = (2\lambda_4)^{-1}\left[(p_i)^{\lambda_4} - 1\right]$$

$$V_{2,i} = (2\lambda_5)^{-1}\left[(p_{n+1-i})^{\lambda_5} - 1\right]$$

with plotting positions $p_i = i/(n+1)$, $i = 1, 2, \cdots, n$.

Focusing on this approximation, although the description of the procedure is the same as for the exact version of $E\left[e_{(i)}\right]$, the DLS approach calls for choosing γ_4 and γ_5 to minimize

$$C(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left[e_{(i)}^{(1)} - e_{(i)}^{*(1)} \right]^2 .$$

The solution can be achieved by first making an initial guess at the value of (λ_4, λ_5) and then applying the OLS to solve the linear problem for γ_1, γ_2.

Once γ_1 and γ_2 have been estimated, we obtain the fitted residual $\hat{e}^{(1)}_{r_i^{(1)}}$ which should be a better approximation to $Q(p, \boldsymbol{\lambda})$ than the one provided by the initial $e_i^{(0)}$ (notice that $\mathbf{r^{(1)}}$ is the sorting determined by $\mathbf{e}^{(1)}$). The new residuals are now used to refine $y_i^{(0)}$ from the estimated errors by defining $y_i^{(1)} = y_i^{(0)} - \hat{e}^{(1)}_{r_i^{(1)}}$, $i = 1, 2, ..., n$. A better estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}^{(1)}$, can now be determined by solving

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[y_i^{(1)} - \mathbf{x}_i' \boldsymbol{\beta} \right]^2 .$$

We note that the design matrix of both the β-estimation and the γ-estimation are fixed, so that the inversion of the coefficient matrices must be executed just one time. Given the new estimate, we reset the values of $\hat{\beta}^{(0)}$ with $\hat{\beta}^{(1)}$ and go through exactly the same procedure described above restarting from (4) after that the new regression residuals $\mathbf{e}^{(2)} = \mathbf{y}^{(1)} - \hat{\mathbf{y}}^{(1)}$ have been calculated and reordered obtaining $\mathbf{r^{(2)}}$. The process is repeated until the correction for $C(\boldsymbol{\gamma})$ becomes negligible. The optimum $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ which meet the stopping rule $C(\boldsymbol{\gamma}) \leq 10^{-5}$ are then used to evaluate the criterion in (3). According to Gilchrist (2000, p.258), we expect that $C(\boldsymbol{\lambda})$ decreases since it takes into consideration the shape of the estimated errors. However, we have observed an oscillatory behavior in the criterion $C(\boldsymbol{\lambda})$, that is, it showed small increases followed by sustained overall decreases. At the end of the procedure the quantity $\sum_k \hat{\mathbf{e}}^{(k)}_{\mathbf{r}^{(k)}}$ approximates the true $Q(p, \boldsymbol{\lambda})$.

Finally, the conversion of γ_1, γ_2 into λ_2, λ_3 is straightforward

$$\hat{\lambda}_2 = \frac{\hat{\gamma}_1}{1 - \hat{\lambda}_3} = \frac{\hat{\gamma}_2}{1 + \hat{\lambda}_3}, \quad \hat{\lambda}_3 = \frac{1 - \hat{\gamma}_1/\hat{\gamma}_2}{1 + \hat{\gamma}_1/\hat{\gamma}_2}$$

while $\hat{\lambda}_1 = \hat{\beta}_0$.

The estimation procedure we have described is based on the assumption that λ_4 and λ_5 are *a priori* fixed. Indeed, the procedure needs to be repeated for different values of (λ_4, λ_5) until there is no further improvement in (3) or until further refinements become negligible.

The optimization of the criterion (3) over a wide spectrum of (λ_4, λ_5) has been performed using a direct search optimization technique, called controlled

random search, and proposed by Price (1977). This is a systematic search technique over possible values of (λ_4, λ_5) falling within a proper restricted region of the parametric space. The most promising region that we have found in our applications is the rectangle

$$D = \{(\lambda_4, \lambda_5) : \ -0.999 \leq \lambda_4 \leq 3.5; \quad -0.999 \leq \lambda_5 \leq 3.5\}.$$

Briefly, the search requires repeated evaluations of the objective function $S(\boldsymbol{\beta}, \boldsymbol{\lambda})$ at points randomly chosen from the set D which is progressively contracted by substituting the worst point with a better one. The search continues until an iteration limit is reached, or a desired tolerance between minimum and maximum values in the $S(\boldsymbol{\beta}, \boldsymbol{\lambda})$ value storage is achieved. Obviously, for each point $(\lambda_4, \lambda_5) \in D$, the estimation procedure previously described for regression and distributional parameters has to be implemented.

3 Comparison of efficiency

In this section we present results of some numerical experiments to study the behavior of the DLS method vs OLS. Moreover, we perform a comparison between estimated and theoretical values to test the accuracy of the estimates discussed in the previous sections and to investigate their properties.

3.1 Application

To illustrate the efficacy of the DLS approach, we applied the method to the *Martin Marietta Data* already analyzed in a very similar context by Taylor (2004). This example consists of $n = 60$ measurements from January 1982 to December 1986. The excess return on equity for the firms and the excess of return on the market portfolio were considered. The data set includes an evident outlier and two more points are suspected outliers. Results for OLS estimates are shown in Table 1 together with our findings related both to the exact DLS and its approximate version (DLSA) based on (5) and (7), respectively. The example highlights that the DLS approach is not really different from the OLS. In particular, the regression parameter estimates are very close to one another while the root mean squared error, $RMSE = \left[n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]^{0.5}$, does not reveal any significant change in the fitting data performance. Nevertheless, the benefits which can derive from the DLS approach mainly rely on the possibility of estimating, in a unique procedure, the coefficients of the regression model and the parameters of the error distribution.

Residual pictures concerning the DLS estimates are given in Figure 1. The histogram of the estimated residuals shows a marked departure from normality. Skewness and kurtosis statistics (2.3 and 12.6, respectively) suggest that a model for error distribution should take into account both leptokurtosis and positive asymmetry. The estimated FPLD model seems to offer a satisfactory answer to the problem.

Table 1. Results for Martin Marietta Data.

Method	β_0	β_1	λ_1	λ_2	λ_3	λ_4	λ_5	RMSE
OLS	0.0011	1.8026						0.0944
DLS	0.0684	1.5592	0.0104	0.0145	-0.7199	2.7461	-0.2965	0.0929
DLSA	0.0016	1.8071	0.0476	0.0072	-0.6619	3.0743	-0.2190	0.0933

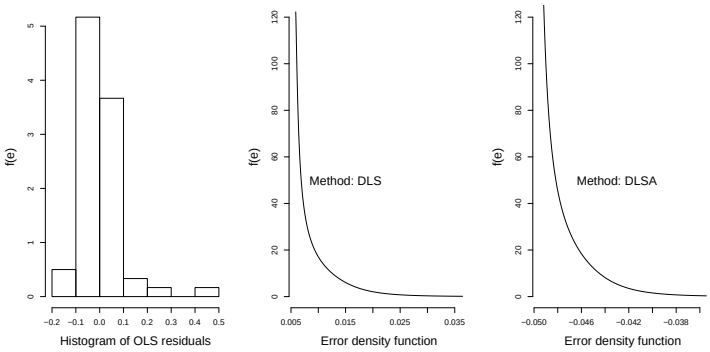


Fig. 1. Histogram and estimated densities of the residuals.

3.2 Simulation

We carry out a simulation study to evaluate the performance of the DLS approach with respect to bias and mean squared error for different sample sizes. Data were simulated from $y = \beta_0+\beta_1 x_1+\beta_2 x_2+e$ with parameters $\beta_0 = 0, \beta_1 = \beta_2 = 1$. The values of (x_1, x_2, e) were generated from independent $[0, 1]$ uniform distributions (these values are kept fixed across experiments). The regression parameter estimates are obtained by generating $N = 1000$ different random samples of size $n = 30, 60, 120$ from a $[-1.5, 1.5]$ uniform distribution. The results are summarized in Table 2.

Table 2. Results for simulated data.

		OLS			**DLSA**			**DLS**		
n		β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
30	Bias	-0.0015	1.0108	0.9885	0.0004	1.0088	0.9862	0.0010	1.0084	0.9853
	RMSE	0.4225	0.6012	0.6382	0.4184	0.5987	0.6355	0.4189	0.5998	0.6363
60	Bias	-0.0018	0.9928	1.0023	-0.0020	0.9940	1.0016	-0.0027	0.9948	1.0021
	RMSE	0.2711	0.4469	0.3794	0.2666	0.4380	0.3698	0.2667	0.4374	0.3710
120	Bias	-0.0027	1.0085	0.9869	-0.0020	1.0076	0.9867	-0.0024	1.0074	0.9871
	RMSE	0.2116	0.2791	0.2848	0.2043	0.2696	0.2757	0.2040	0.2695	0.2757

For all methods, the RMSE decreases as the sample size increases, an indication that all the estimates tend to be consistent. Moreover, DLSA and

DLS approaches provide results which are competitive with the OLS, even for the smallest sample size (here, the former has a better performance than the latter). When n increases, the methods are virtually indistinguishable.

4 Conclusion

The aim of this paper is to make a contribution toward the use of the quantile statistical methods in regression analysis. The approach we present assumes that the class of distributions describing the stochastic component of the linear regression model can be characterized parametrically by the five-parameter version of the generalized lambda distribution (FPLD).

A computational iterative algorithm is developed to implement the distributional least squares (DLS) approach which allows to estimate, in a unique procedure, both the regression coefficients and the parameters of the FPLD which are more compatible with the unknown distribution of the regression disturbances. The estimation of the error density may be particularly useful to check departures from the Normal hypothesis usually assumed for OLS estimates. For small sample size, the violation of this assumption may deteriorate the estimates efficiency as well as invalidate the inference on the model.

At the moment, the properties of DLS estimates are under study for a wide range of non-normalities. Moreover, it must be noted that the final estimate of the error density is strictly linked to the optimization criterion used to determine the regression parameters. Therefore, the distributional approach, applied in this paper only to the canonical OLS, is under investigation for alternative regression methods, such as the least absolute deviations.

References

GILCHRIST, W. (2000): *Statistical Modelling with Quantile Functions*. Chapman & Hall, CRC, Boca Raton, USA.

ÖZTÜRK, A., DALE, R.F. (1985): Least squares estimation of the parameters of the generalized lambda distributions. *Technometrics 19 (1), 37-45.*

PRICE, W.L. (1977): A controlled random search procedure for global optimisation. *The Computer Journal 20 (4), 367-370.*

TAYLOR, J. (2004): Joint modelling of location and scale parameters of the t distribution. *Statistical Modelling 4 (2), 91-112.*

Maximum Likelihood Estimation for Brownian-Laplace Motion and the Generalized Normal-Laplace (GNL) Distribution

William J. Reed

Department of Mathematics and Statistics, University of Victoria
P.O. Box 3045, Victoria, B. C, Canada, V8W 3P4, *reed@math.uvic.ca*

Abstract. The generalized normal-Laplace (GNL) distribution arises as the distribution of the increments of the Lévy process known as Brownian-Laplace motion. Similar to the generalized hyperbolic distribution, the GNL distribution can exhibit excess kurtosis and skewness, making Brownian-Laplace motion potentially a good model for the evolution of financial returns. However unlike the generalized hyperbolic distribution, the GNL does not possess a known closed-form for its density, being instead defined in terms of its characteristic function. This causes difficulties for parameter estimation via maximum likelihood. In this paper three methods of numerically computing the GNL density, and hence the liklelihood function are considered. One involves inverting the characteristic function, while the remaining two rely on representations of the GNL distribution – one as a convoltion of normal and generalized Laplace components and the other as a normal variance-mean mixture.

Keywords: financial modelling, normal variance-mean mixture, Brownian-Laplace motion

1 Introduction

Brownian-Laplace motion was introduced by Reed (2007) as a possible model for the evolution of logarithmic returns of financial assets. It is a Lévy process whose increments can exhibit excess kurtosis and skewness, over short intervals, but tending to normality as the time interval increases. Such behaviour is often seen in time series of logarithmic returns of stocks and other financial assets. Reed (2007) derived a method for computing the value of a European call option on a stock whose logarithm follows Brownian-Laplace motion, and presented examples.

One of the main difficulties with this model involves paarmeter estimation. While in principle method-of-moments estimation is straightforward, difficulties can often arise because of contstraints on the parameters - four of the five model parameters are constrained to be non-negative. Estimation by maximum likelihood is rendered difficult by the fact that there is no known closed-form expression for the density of the increments of Brownian-Laplace motion.

In this article three methods of numerically computing the density of the increments of Brownian-Laplace motion, and hence the likelihood function and maximum likelihood estimates of model parameters, are presented and compared.

2 The generalized normal-Laplace distribution

Brownian-Laplace motion is defined as a Lévy process $\{X_t\}_{t\geq 0}$, for which the increments $X_{t+\tau} - X_\tau$ have characteristic function $(\phi(s))^t$ where

$$\phi(s) = \left[\frac{\alpha\beta\exp(\mu i s - \sigma^2 s^2/2)}{(\alpha - is)(\beta + is)}\right]^\rho \tag{1}$$

with α, β, ρ and σ positive parameters and $-\infty < \mu < \infty$ (Reed, 2007).

A distribution with characteristic function (1) is known as a *generalized normal-Laplace* (GNL) distribution and the notation

$$Y \sim \mathrm{GNL}(\mu, \sigma^2, \alpha, \beta, \rho)$$

is used to indicate that the random variable Y follows such a distribution. Properties of the GNL distribution are given in Reed (2007). In particular the mean and variance of the $\mathrm{GNL}(\mu, \sigma^2, \alpha, \beta, \rho)$ distribution are

$$\mathrm{E}(Y) = \rho\left(\mu + \frac{1}{\alpha} - \frac{1}{\beta}\right); \qquad \mathrm{var}(Y) = \rho\left(\sigma^2 + \frac{1}{\alpha^2} + \frac{1}{\beta^2}\right)$$

while the higher order cumulants are (for $r > 2$)

$$\kappa_r = \rho(r-1)!\left(\frac{1}{\alpha^r} + (-1)^r\frac{1}{\beta^r}\right).$$

The parameters μ and σ^2 influence the central location and spread of the distribution, while α and β affect the symmetry. If $\alpha > \beta$ the distribution is skewed to the left, and vice versa. The parameter ρ affects the lengths of the tails. The nature of the tails can be determined from the order of the poles of its characteristic function. Precisely $f(y) \sim c_1 y^{\rho-1}e^{-\alpha y}$ $(y \to \infty)$ and $f(y) \sim c_2(-y)^{\rho-1}e^{\beta y}$ $(y \to -\infty)$, (where c_1 and c_2 are constants). Thus for $\rho < 1$, both tails are fatter than exponential; for $\rho = 1$ they are exactly exponential and for $\rho > 1$ they are less fat than exponential. This exactly mimics the tail behaviour of the generalized Laplace distribution. Thus in the tails the generalized Laplace component of the GNL distribution dominates over the normal component.

The parameter ρ affects all moments. However the coefficients of skewness ($\gamma_1 = \kappa_3/\kappa_2^{3/2}$) and of excess kurtosis ($\gamma_2 = \kappa_4/\kappa_2^2$) both decrease with increasing ρ (and converge to zero as $\rho \to \infty$) with the shape of the distribution becoming more normal with increasing ρ, (exemplifying the central

limit effect since the sum of n iid GNL$(\mu, \sigma^2, \alpha, \beta, \rho)$ random variables has a GNL$(\mu, \sigma^2, \alpha, \beta, n\rho)$ distribution). When $\alpha = \beta$ the distribution is symmetric. In the limiting case $\alpha = \beta = \infty$ the GNL reduces to a normal distribution.

A GNL random variable $Y \sim \text{GNL}(\mu, \sigma^2, \alpha, \beta, \rho)$ can be represented as

$$Y \stackrel{d}{=} \rho\mu + \sigma\sqrt{\rho}Z + \frac{1}{\alpha}G_1 - \frac{1}{\beta}G_2 \tag{2}$$

where Z, G_1 and G_2 are independent with $Z \sim$ N(0,1) and G_1, G_2 gamma random variables with scale parameter 1 and shape parameter ρ, *i.e.* with probability density function (pdf)

$$g(x) = \frac{1}{\Gamma(\rho)} x^{\rho-1} e^{-x}. \tag{3}$$

This representation provides a straightforwd way to generate pseudo-random deviates following a GNL distribution.

Since the difference between two gamma random variables, with the same shape parameter follows a *generalized Laplace* (also known as *variance-gamma* or *Bessel function*) distribution (Kotz *et al.*, 2001) it follows that a random variable $Y \sim \text{GNL}(\mu, \sigma^2, \alpha, \beta, \rho)$ can be represented as

$$Y \stackrel{d}{=} W + V \tag{4}$$

where $W \sim \text{N}(\mu, \sigma^2)$ and V follows a generalized Laplace distribution with probability density function

$$f_{GL}(v) = \frac{(\alpha\beta)^\rho}{\sqrt{\pi}\Gamma(\rho)} \left(\frac{|v|}{\alpha+\beta}\right)^{\rho-1/2} \exp\left(\frac{\beta-\alpha}{2}\, v\right) K_{\rho-1/2}\left(\frac{\alpha+\beta}{2}\, |v|\right) \tag{5}$$

where K_λ is a modified Bessel function of the third kind with index λ (Kotz *et al.*, 2001). Thus the pdf of a GNL$(\mu, \sigma^2, \alpha, \beta, \rho)$ distribution can be obtained as the convolution of the pdf (5) and the pdf of an N(μ, σ^2) distribution.

Another characterization of the GNL distribution which will prove useful is that it arises as the state of a Brownian motion, with initial state normally distributed, after a time which follows a gamma distribution. This leads to a normal variance-mean mixture representation of the GNL distribution. Precisely $Y \sim \text{GNL}(\mu, \sigma^2, \alpha, \beta, \rho)$ can be represented as

$$Y \stackrel{d}{=} \rho\mu + \left(\frac{1}{\alpha} - \frac{1}{\beta}\right) T + Z\sqrt{\rho\sigma^2 + \frac{2}{\alpha\beta}\, T} \tag{6}$$

where $Z \sim \text{N}(0, 1)$ and T follows a gamma distribution with pdf (3). The proof of this follows from evaluating the characteristic function of the right-hand side by conditioning on T. Precisely as

$$E(E(\exp(isY|T)) = \exp[i\rho\mu s - \rho\sigma^2 s^2/2]\ \phi_T\left(\left(\frac{1}{\alpha} - \frac{1}{\beta}\right) s + \frac{is^2}{\alpha\beta}\right)$$

where $\phi_T(s) = (1 - is)^{-\rho}$ is the characteristic function of T. Using this in the above leads to the characteristic function (1) of the GNL$(\mu, \sigma^2, \alpha, \beta, \rho)$ distribution.

3 Numerical computation of the GNL pdf and log-likelihood maximization

For iid observations $y_1, \ldots, y_n$ from the GNL $(\mu, \sigma^2, \alpha, \beta, \rho)$ distribution the log likelihood is

$$\sum_{i=1}^{n} \log f(y_i; \mu, \sigma^2, \alpha, \beta, \rho)$$

where $f(y_i; \mu, \sigma^2, \alpha, \beta, \rho)$ is the pdf of the GNL $(\mu, \sigma^2, \alpha, \beta, \rho)$. Since there is no closed-form expression for the density f we turn to numerical methods for evaluating the log likelihood.

Three methods of numerically computing the density f can be used.

• *Inversion of the characteristic function.* Since μ is a location parameter, the density $f(y; \mu, \sigma^2, \alpha, \beta, \rho)$ can be computed as $f_0(y - \mu; \sigma^2, \alpha, \beta, \rho)$ where f_0 is the pdf of GNL $(0, \sigma^2, \alpha, \beta, \rho)$. The latter can be obtained by numerically inverting the characteristic function (1) with $\mu = 0$ *i.e*

$$\phi(s) = \left[\frac{\alpha\beta e^{-\sigma^2 s^2/2}}{(\alpha - is)(\beta + is)} \right]^{\rho} \tag{7}$$

This involves (for fixed y and $\sigma^2, \alpha, \beta, \rho$) evaluating the integral

$$\frac{1}{\pi} \int_0^{\infty} r(s) \cos(\theta(s) - sy) ds \tag{8}$$

where $r(s)$ $\theta(s)$ are the modulus and argument (principal value) of the complex number (7).

• *Convolution of normal and generalized Laplace pdfs.* This involves evaluating numerically the convolution integral

$$\int_{-\infty}^{\infty} \frac{1}{\sigma} \phi \left(\frac{y - \mu - t}{\sigma} \right) f_{GL}(t) dt \tag{9}$$

where ϕ is the pdf of a standard normal deviate, and f_{GL} is the generalized Laplace pdf given in (5).

• *Using the normal mean variance mixture representation.* The GNL pdf can be evaluated as the integral of (6) with respect to a gamma density for T, *i.e.* by evaluating

$$\int_0^{\infty} \frac{1}{\sqrt{\rho\sigma^2 + 2t/(\alpha\beta)}} \, \phi \left(\frac{y - \mu - (1/\alpha - 1/\beta)t}{\sqrt{\rho\sigma^2 + 2t/(\alpha\beta)}} \right) g(t) dt \tag{10}$$

where $g(t)$ is the pdf (3) of a gamma distribution with shape parameter ρ and scale parameter 1. The three methods were compared by evaluating f at a thousand values of y using using different parameter values.

The computations were performed in R using the function integrate. Of the three methods, the third appears to be considerably faster than the other two. However it is probably less accurate that the other two. When the pdf $f_{GNL}(y)$ was evaluated for 1000 equally-spaced values of y between -5 and 5, using various parameter values, typically the first two methods would exhibit a maximum absolute difference of the order 10^{-9}, while the third method would a maximum absolute difference from the other two of the order 10^{-5}.

4 Examples

We present two simple examples. For both the computations were performed in R using the Nelder-Mead simplex method in the function optim in the stats package for maximizing the log-likelihood. The first method (inverting characteristic function) for numerically computing the pdf was used. In both examples computing time (on a desktop PC with a 2.66 GHz processor) was several (5 to 10) minutes. This changed little when the second (convolution) method for numerically computing the pdf was used

In the first example the GNL distribution was fitted to 500 simulated (using (2)) GNL$(0, 0.1, 0.2, 0.3, 0.5)$ deviates. The MLEs, with asymptotic standard errors (computed from the inverse of the observed information matrix) in brackets, were

$$\hat{\mu} = -0.123\ (0.104);\quad \hat{\sigma}^2 = 0.110\ (0.129),$$

$$\hat{\alpha} = 0.193\ (0.023);\quad \hat{\beta} = 0.364\ (0.040);\quad \hat{\rho} = 0.559\ (0.078)$$

Fig. 1 (left-hand panel) shows a Q-Q plot of the observed versus fitted distributions.

In the second example Brownian-Laplace motion was fitted to a time series of Qualcomm ordinary share prices (Jan 2006-Nov 2007). This involved fitting the GNL to 474 logarithmic returns ($\log(P_{t+1}/P_t)$). The right-hand panel of Fig. 1 show a Q-Q plot for the fit.

5 Conclusions

The paper has considered maximum likelihood estimation of the parameters of the GNL distribution, needed for fitting the Brownian-Laplace motion model to time-series data. Because there is no closed-form of the density of the GNL distribution, numerical methods are needed in evaluating the likelihood function. For a data set comprising n observations, at each evaluation of the log-likelihood (in each iteration of the maximization routine) n numerical integrations are required (to evaluate the pdf at each point). While the

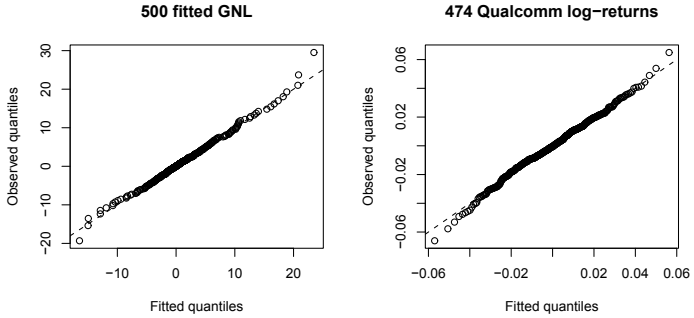


Fig. 1. Q-Q plots (observed *vs.* fitted quantiles) for the GNL distribution fitted to two sets of data. The left-hand panel is for 500 independent GNL$(0, 0.1, 0.2, 0.3, 0.5)$ pseudo-random deviates; the right-hand panel is for logarithmic returns for a time series of 475 Qualcomm share prices.

evalaution method based on the normal varaince-mean mixture representation of the GNL distribution is the fastest of the the three ways discussed for computing the pdf, it appears to be less accurate than the other two methods, at least using standard R functions for quadrature and for computing the normal and gamma pdf, *etc.*

There also appears to be the possibility of multiple local maxima in the likelihood function for the GNL distribution. For some parameter values, using different starting values for the optimization routine can lead to maxima of similar values, but at different locations. This suggests flatness in the likelihood. It is particularly the case when one or more of the parameters ρ, α and β are large. Since the GNL distribution converges to a normal distribution when $\rho \to \infty$, when ρ is large there may be lttle information for estimating it or the tail parameters α and β, and furthermore there will likely be confounding of the estimates. Likewise when $\alpha, \beta \to \infty$, the GNL distribution converges to a normal distribution, and so confounding and difficulties in the estimation of ρ, α and β will likely occur when any of these parameters are large.

References

KOTZ, S., KOZUBOWSKI, T.J. and PODGORSKI, K. (2001): *The Laplace Distribution and Generalizations.* Birkhäuser, Boston.

REED, W.J. (2007): Brownian-Laplace motion and its use in financial modelling. *Communications in Statistics: Theory & Methods 36(3) 473-484.*

White's Estimator of Covariance Matrix for Instrumental Weighted Variables

Jan Ámos Víšek

Faculty of Social Sciences, Charles University
& Inst. of Information Theory and Automation, Academy of Sciences,
Smetanovo nábřeží 6, 110 01 Prague, the Czech Republic, *visek@mbox.fsv.cuni.cz*

Abstract. Under heteroscedasticity of error terms the significances of explanatory variables in a linear regression model have to be established employing the *White's estimator of covariance matrix of regression coefficients*, coefficients estimated by the *(Ordinary) Least Squares*. When the orthogonality condition is broken the *Instrumental Variables* or the *Total Least Squares* are used to preserve unbiasedness of estimation (former in social sciences, latter mostly in natural or technical sciences). If moreover, data are contaminated a robust version of instrumental variables called the *Instrumental Weighted Variables* is to be used to cope both with the break of orthogonality condition as well as with contamination. Significance of explanatory variables (and of instruments) is to be examined by a robust version of White's estimator of covariance matrix of estimates of regression coefficients.

Keywords: robustness, heteroscedasticity, instrumental weighted variables, White's estimator of covariance matrix of estimates of regression coefficients

1 Introduction of basic framework

The set of all positive integers will be denoted by N and p-dimensional Euclidean space by R^p. Let us consider the linear regression model

$$Y_i = X_i'\beta^0 + e_i = \sum_{j=1}^{p} X_{ij}\beta_j^0 + e_i = \beta_0^0 + \sum_{j=1}^{p-1} V_{ij-1}\beta_j^0 + e_i, \quad i = 1, 2, ..., n. \tag{1}$$

We shall assume that:

C1 *The sequence* $\{(V_i', e_i)'\}_{i=1}^{\infty}$ *is sequence of independent* p*-dimensional random variables. There is an absolutely continuous d.f. , say* $F_{V,e}(v, r)$ *(denote density* $f_{V,e}(v, r)$*), so that the d.f.'s* $F_{V,e_i}(v, r) = F_{V,e}(v, \sigma_i \cdot r)$ *and* $\mathbb{E}e_i = 0$ *for all* $i \in N$*. The marginal d.f.'s* $F_V(v)$ *of vectors* V_i*'s are the same for all* $i \in N$ *and have a bounded support, i.e. putting* $M = \sup\{\|v\| \ : \ f_V(v) > 0\}$ *we have* $M < \infty$*. Moreover, the existence of second moments is assumed, the density* $f_{V,e}(v, r)$ *is bounded, say by* B*, and* $\sup_{i\in N} \sigma_i < \infty$*. Finally, consider the sequence* $\{(X_i', e_i)'\}_{i=1}^{\infty}$ *where* $X_{i1} = 1$ *and* $X_{ij} = V_{i,j-1}, \ j = 2, 3, ..., p$ *for all* $i \in N$*.*

Notice please that we assume that the error terms e_i's can be correlated with explanatory variables V_i's. Moreover, error terms are assumed generally

heteroscedastic. Finally, as $f_{V,e_i}(v,r) = \sigma_i \cdot f_{v,e}(v, \sigma_i \cdot r)$, we have $f_{V,e_i}(v,r) < \sup_{i \in N} \sigma_i \cdot B$. For any $\beta \in R^p$ $r_i(\beta) = Y_i - X_i'\beta$ denotes the i-th residual and $r^2_{(h)}(\beta)$ the h-th order statistic among the squared residuals, i.e. we have

$$r^2_{(1)}(\beta) \leq r^2_{(2)}(\beta) \leq \ ... \ \leq r^2_{(n)}(\beta). \tag{2}$$

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 Why instrumental weighted variables?

Let's explain why the classical econometrics employes instrumental variables.

The violation of orthogonality condition $I\!E\,\{e_1|X_1\} = 0$ implies that (remember that e_i's have the same d.f. except for variances)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i e_i \neq 0 \quad \text{in probability} \tag{3}$$

and hence also inconsistency of

$$\hat\beta^{(OLS,n)} = \beta^0 + \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i e_i. \tag{4}$$

The most frequently given examples of failure of the condition of orthogonality are the measurement of explanatory variables with a random error or the (dynamic) regression model with lagged response in the role of explanatory variable (Judge et al. (1985)). Econometricians offer as a remedy the method of the *Instrumental Variables* which defines the estimator as (any) solution of the normal equations

$$\sum_{i=1}^{n} Z_i (Y_i - X_i'\beta) = 0 \tag{5}$$

where the sequence $\{Z_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. instruments for explanatory variables X_i's given as follows: Let $\{U_i\}_{i=1}^{\infty}$ be a sequence of $p-1$-dimensional i.i.d. r.v.'s such that $I\!E U_1 \cdot e_1 = 0$, so that putting

$$Z_{i1} = 1 \qquad \text{and} \qquad Z_{ij} = U_{i,j-1} \quad j = 2, 3, ..., p \tag{6}$$

for all $i \in N$ the orthogonality condition $I\!E Z_1 e_1 = 0$ holds (see **C1**) and hence $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i e_i = 0$ a.s. . If moreover (e. g.) $I\!E Z_1 X_1' = Q$ is regular, the analogy of relation (4), namely

$$\hat\beta^{(IV,n)} = \beta^0 + \left(\frac{1}{n} \sum_{i=1}^{n} Z_i X_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Z_i e_i. \tag{7}$$

hints that the estimator evaluated by means of method of *Instrumental Variables* is consistent.

So much to the explanation how to cope with the break of the orthogonality condition Nevertheless, in the case of contaminated data, we need to use the robustified version of instrumental variables. Let's recall idea of their proposal.

In 1992 Hettmansperger and Sheather showed that the *Least Median of Squares (LMS)* (Rousseeuw (1984)) can be considerably sensitive to some very small changes of data. It appeared later that their result was due to bad algorithm for *LMS* (Víšek (1994)). Nevertheless, evaluating the *Least Trimmed Squares (LTS)* (Hampel (1986)) by total search for data used by Hettmansperger and Sheather (1992) (and hence reaching the exact value of the estimator) revealed that the problem exists for *LTS*. Academic examples in Víšek (1996b) and (2000a) indicated the reason for it (for any robust estimator with high *breakdown point*) and Víšek (1992), (1996a) and (2002b) brought the theoretical justification of the fact that the discontinuous objective functions can cause (extremely) high sensitivity of robust estimators to some changes of data. That was an inspiration for defining the *Least Weighted Squares (LWS)* (Víšek (2000b), see also (2002a))

$$\hat{\beta}^{(LWS,n,w)} = \arg\min_{\beta\in R^p} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r^2_{(i)}(\beta) = \arg\min_{\beta\in R^p} \sum_{i=1}^{n} w\left(F^{(n)}_{\beta}(|r_i(\beta)|)\right) r_i^2(\beta) \quad (8)$$

where

$$F^{(n)}_{\beta}(v) = \frac{1}{n}\sum_{i=1}^{n} I\left\{|r_i(\beta)| < v\right\} = \frac{1}{n}\sum_{i=1}^{n} I\left\{|e_i - X_i'\beta| < v\right\} \quad (9)$$

is the empirical distribution function of the absolute values of residuals and w is a weight function fulfilling:

C2 *Weight function* $w : [0,1] \to [0,1]$ *is nonincreasing, with bounded derivative* $w'(\alpha)$*, i.e.* $|w'(\alpha)| < L < \infty$*, and* $w(0) = 1$.

It is only a technicality to show that $\hat{\beta}^{(LWS,n,w)}$ has to be a solution of

$$\sum_{i=1}^{n} w\left(F^{(n)}_{\beta}(|r_i(\beta)|)\right) X_i\left(Y_i - X_i'\beta\right) = 0. \quad (10)$$

Then again, if

$$w\left(F^{(n)}_{\beta}(|e_1|)\right) X_1 e_1 \neq 0,$$

$\hat{\beta}^{(LWS,n,w)}$ is inconsistent. The remedy is straightforward, given by *normal equations*

$$\sum_{i=1}^{n} w\left(F^{(n)}_{\beta}(|r_i(\beta)|)\right) Z_i\left(Y_i - X_i'\beta\right) = 0 \quad (11)$$

where again the sequence $\{Z_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. instruments for X_i's (see text below the equation (2) and Víšek (2004)).

3 Asymptotics of instrumental weighted variables

For robust version of the method we need some assumptions about the mutual behaviour of X_i's and Z_i's. Let's recall that we assume heteroscedasticity of the error terms (see **C1**) and define a "mean" d.f. for any $n \in N$

$$\overline{F}_{n,\beta}(v) = \frac{1}{n}\sum_{i=1}^{n} P\left(|Y_i - X_i'\beta| < v\right). \tag{12}$$

(a possibility to approximate the empirical distribution $F_\beta^{(n)}(v)$ - see (9) - by $\overline{F}_{n,\beta}(v)$ uniformly in $v \in R$ as well as in $\beta \in R^p$ opened in fact the way for results given below, see Víšek (2008c)). Further define for any $\beta \in R^p$

$$F_{\beta' Z X' \beta}(u) = P(\beta' Z_1 X_1' \beta < u)$$

and put for any $\lambda > 0$ and any $a \in R$

$$\gamma_{\lambda,a} = \sup_{\|\beta\|=\lambda} F_{\beta' Z X' \beta}(a) \quad \text{and} \quad \tau_\lambda = -\inf_{\|\beta\|\le\lambda} \beta' I\!E\left[Z_1 X_1' \cdot I\{\beta' Z_1 X_1' \beta < 0\}\right]\beta.$$

C3 *The $p-1$-dimensional r.v.'s $\{U_i\}_{i=1}^\infty$ are independent and identically distributed with distribution function $F_U(u)$. Moreover, they are independent from the sequence $\{e_i\}_{i=1}^\infty$, the joint distribution function $F_{V,U}(v,u)$ is absolutely continuous. Create Z_i's as in (6). Further, let $I\!E Z_1 Z_1'$ is positive definite and there is $\Delta > 1$ so that $I\!E\left\{\|Z_1\| \cdot \|X_1\|\right\}^\Delta < \infty$. Moreover, there is $n_0 \in N$ so that for all $n > n_0$ $I\!E\left\{\frac{1}{n}\sum_{i=1}^n \left[w(\overline{F}_{n,\beta}(|e_i|))Z_i X_i'\right]\right\}$ is regular. Finally, there is $a > 0$, $b \in (0,1)$ and $\lambda > 0$ so that*

$$a \cdot (b - \gamma_{\lambda,a}) \cdot w(b) > \tau_\lambda \tag{13}$$

For discussion of **C3** see Víšek (2008a).

C4 *There is $n_0 \in N$ so that for all $n > n_0$ the vector equation*

$$\beta' I\!E\left\{\frac{1}{n}\sum_{i=1}^{n}\left[w(\overline{F}_{n,\beta}(|r_i(\beta)|))Z_i\left(e_i - X_i'\beta\right)\right]\right\} = 0 \tag{14}$$

in the variable $\beta \in R^p$ has unique solution $\beta^0 = 0$.

Lemma 9. *Let Conditions* **C1**, **C2**, **C3** *and* **C4** *be fulfilled. Then any sequence $\left\{\hat\beta^{(IWV,n,w)}\right\}_{n=1}^\infty$ of the solutions of normal equations (11) is weakly consistent.*

Proof is given in Víšek (2008a) where also a simulation study demonstrates that the algorithm, firstly presented in Víšek (2006a), works very well. Result in Víšek (2006b) opened way to prove $\sqrt{n}$-consistency and to find an asymptotic representation of $\hat\beta^{(IWV,n,w)}$ under following conditions (denote by $f_{e|V}(r|V_1 = x)$ the conditional density corresponding to the d.f. $F_{V,e}(v,r)$):

NC1 *The density $f_{e|V}(r|V_1 = x)$ is uniformly with respect to x Lipschitz of the first order. Moreover, $f_e'(r)$ exists and is bounded in absolute value by U_e'.*

NC2 *The derivative $w'(\alpha)$ of the weight function is Lipschitz of the first order (with the corresponding constant J_w).*

Lemma 10. *Let the conditions* **C1**, **C2**, **C3**, **C4**, **NC1** *and* **NC2** *be fulfilled. Then any sequence $\left\{\hat\beta^{(IWV,n,w)}\right\}_{n=1}^\infty$ of the solutions of normal equations (11) is $\sqrt{n}$-consistent.*

For the proof see Víšek (2008b). Further, denote by $g_e(r)$ the density of the d.f. $G_e(r) = P(e_1^2 < r)$ (notice that under **C1** density $g_e(r)$ always exists). Moreover, for any $\alpha \in (0,1)$ denote by u_α^2 the upper α-quantile of d.f. G_e, i.e. we have $P(e_1^2 > u_\alpha^2) = \alpha$. Let's recall that $f_e(r)$ is density of d.f. $F_e(r)$ which is marginal d.f. of $F_{V,e}(v,r)$ (see **C1**).

AC1 *For any* $\alpha \in (0,1)$ *there is* $\delta(\alpha) > 0$ *so that*

$$\inf_{r\in(0,u_\alpha^2+\delta(\alpha))} g_e(r) > L_{g,\alpha} > 0 \quad \text{and} \quad \inf_{|r|\in(0,\sqrt{u_\alpha^2+\delta(\alpha)})} f_e(r) > L_{f,\alpha} > 0. \tag{15}$$

Similarly as above (see text under **C1**) the condition **AC1** implies in fact that (15) holds for all densities $g_{e_i}(r)$ and $f_{e_i}(r)$, i.e. for all $i \in N$.

AC2 *There is* $q > 1$ *so that* $\sup_{i\in N} I\!E\,|e_i|^{2q} < \infty$.

Lemma 11. *Let the conditions* **C1**, **C2**, **C3**, **C4**, **NC1**, **NC2**, **AC1** *and* **AC2** *hold. Then*

$$\sqrt{n}\left(\hat\beta^{(IWV,n,w)} - \beta^0\right) =$$

$$\left[\frac{1}{n}\sum_{i=1}^n w\left(\overline{F}_{n,\beta^0}(|e_i|)\right)\cdot Z_i X_i'\right]^{-1}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n w\left(\overline{F}_{n,\beta^0}(|e_i|)\right)\cdot Z_i e_i + o_p(1) \tag{16}$$

as $n \to \infty$.

For the proof see again Víšek (2008b).

4 Robustifying White's estimator of covariance matrix

Having at hand the algorithm for the *IWV* and applying it on data, one needs a test for homoscedasticity of error terms (as disregarding heteroscedasticity my lead to poor identification of regression model, frequently wrongly assuming some insignificant explanatory variables as significant). Such a test was for *IWV* established in Víšek (2007). When the test rejects the homoscedasticity, we need estimators of variances of the estimates of regression coefficient which is consistent under heteroscedasticity. Following Halbert White (1980) and employing (16), we may prove (let's recall that $F_\beta^{(n)}(r)$ is the empirical function of the absolute values of residuals $r_i(\beta)$'s - see (9) and also the first line of the second page):

Lemma 12. *Let the conditions* **C1**, **C2**, **C3**, **C4**, **NC1**, **NC2**, **AC1** *and* **AC2** *hold. Write briefly* r_i^2 *instead of* $r_i^2(\hat\beta^{(IWV,n,w)})$. *Then*

$$\widehat{\mathrm{cov}}(\hat\beta^{(IWV,n,w)}) = n\cdot\left[\sum_{i=1}^n w\left(F_\beta^{(n)}(|r_i^2|)\right)Z_i X_i'\right]^{-1}\times$$

$$\times\sum_{i=1}^n w^2\left(F_\beta^{(n)}(|r_i^2|)\right)r_i^2 Z_i Z_i'\cdot\left[\sum_{i=1}^n w\left(F_\beta^{(n)}(|r_i^2|)\right)Z_i X_i'\right]^{-1} \tag{17}$$

is weakly consistent estimator of covariance matrix of $\hat\beta^{(IWV,n,w)}$.

This is one of two (important?) results of paper.

5 Simulation study of robustified White's estimator

S1 The regression model

$$Y_n = \beta_1 \cdot X_{n1} + \beta_2 \cdot X_{n2} + \beta_3 \cdot X_{n3} + e_n, \quad n = 1, 2, ..., 50, \tag{18}$$

was considered and the experiment - as described below - was 10 times repeated and the results were collected in tables.

S2 Each repetition of experiment contains 100 samples, each sample consists of 50 observations generated as follows: A sequence $\{T_n\}_{n=1}^{52}$ of 3-dimensional random vectors normally distributed with zero mean and unit covariance matrix was generated and the autoregressive sequence $\{U_n\}_{n=1}^{51}$ was defined by $U_n = 0.5 \cdot T_{n+1} + 0.5 \cdot T_n$.

S3 The sequences of explanatory and instrumental variables were constructed

$$X_n = U_{n+1} \qquad \text{and} \qquad Z_n = U_n.$$

Then for $j = 1, 2, 3$ $\mathrm{var}\,(X_{nj}) = \mathrm{var}\,(Z_{nj}) = 0.5$ and

$$\mathrm{cov}\,(X_{nj}, Z_{nj}) = \mathrm{cov}\,(U_{n+1,j}, U_{nj})$$

$$= \mathrm{cov}\,(0.5 \cdot T_{n+2,j} + 0.5 \cdot T_{n+1,j}\,,\, 0.5 \cdot T_{n+1,j} + 0.5 \cdot T_{nj}) = 0.25$$

and hence $\mathrm{corr}\,(X_{nj}, Z_{nj}) = 0.5$. On the other hand for $j \neq k$

$$\mathrm{cov}\,(X_{nj}, Z_{nk}) = \mathrm{cov}\,(0.5 \cdot T_{n+2,j} + 0.5 \cdot T_{n+1,j}\,,\, 0.5 \cdot T_{n+1,k} + 0.5 \cdot T_{nk}) = 0.$$

S4 The error terms $\{e_n\}_{n=1}^{50}$ were created by $e_n = \sum_{k=1}^{3} T_{n+2,k}$. Then again $\mathrm{cov}\,(X_{nj}, e_n) = 0.5$, $j = 1, 2, 3$ and $\mathrm{var}\,(e_n) = 3$ and hence the explanatory variables are correlated with the error terms. On the other hand

$$\mathrm{cov}\,(Z_{nj}, e_n) = 0, \quad j = 1, 2, 3,$$

i. e. the instrumental variables are not correlated with the error terms.

S5 The values of response variables Y_n's were calculated as

$$Y_n = 2.4 \cdot X_{n1} - 3.1 \cdot X_{n2} + 2.8 \cdot X_{n3} + e_n, \quad n = 1, 2, ..., 50.$$

Then contamination was performed as follows: For $n = 1, 2, ..., 5$ we put $Y_n^* = 5 \cdot Y_n$ and $Y_n^* = Y_n$ for $6 \leq n \leq 50$. Moreover, for $n = 46, 47, ..., 50$ we put $X_n^* = 5 \cdot X_n$ and $Z_n^* = 5 \cdot Z_n$. Finally, $X_n^* = X_n$ and $Z_n^* = Z_n$ for $1 \leq n \leq 45$. Then we took into account the data $\left\{(Y_n^*, [X_n^*]', [Z_n^*]')'\right\}_{n=1}^{50}$ and estimates $\hat\beta_{(k)}^{(LS,50)}$, $\hat\beta_{(k)}^{(LWS,50,w)}$ and $\hat\beta_{(k)}^{(IWS,50,w)}$ calculated for 100 repetitions (index of experiment is k) with $w(x)$ continuous, $w(x) = 1$ for $x \in [0, 0.8]$, $w(1) = 0$ and linear on $(0.8, 1)$.

S6 The mean values were calculated $\hat\beta_{(mean)}^{(LS,50)} = \frac{1}{100} \sum_{k=1}^{100} \hat\beta_{(k)}^{(LS,50)}$,

$$\hat\beta_{(mean)}^{(LWS,50,w)} = \frac{1}{100} \sum_{k=1}^{100} \hat\beta_{(k)}^{(LWS,50,w)} \text{ and } \hat\beta_{(mean)}^{(IWV,50,w)} = \frac{1}{100} \sum_{k=1}^{100} \hat\beta_{(k)}^{(IWV,50,w)}.$$

These empirical means are presented in the next triplet of tables, in the columns denoted (at the second row of tables) by 1. In a similar way empirical and estimated variances of $\hat\beta$'s were evaluated.

S7 The whole procedure, starting with **S2** up to **S6**, was 10 times repeated and values collected in the next 5 tables. The first 3 of them show that OLS as well as LWS give bad estimates of regression coefficients while IWS give considerably better.

Ordinary Least Squares										
	1	2	3	4	5	6	7	8	9	10
$\hat{\beta}_1$	3.406	3.393	3.396	3.386	3.395	3.394	3.407	3.412	3.393	3.400
$\hat{\beta}_2$	-2.100	-2.111	-2.088	-2.105	-2.085	-2.104	-2.107	-2.103	-2.095	-2.096
$\hat{\beta}_3$	3.793	3.819	3.788	3.823	3.797	3.809	3.797	3.797	3.801	3.789
Least Weighted Squares										
	1	2	3	4	5	6	7	8	9	10
$\hat{\beta}_1$	3.405	3.393	3.394	3.377	3.396	3.388	3.419	3.403	3.407	3.398
$\hat{\beta}_2$	-2.099	-2.109	-2.085	-2.102	-2.070	-2.101	-2.113	-2.101	-2.096	-2.098
$\hat{\beta}_3$	3.777	3.823	3.784	3.829	3.793	3.805	3.811	3.798	3.796	3.788
Instrumental Weighted Variables										
	1	2	3	4	5	6	7	8	9	10
$\hat{\beta}_1$	2.446	2.296	2.160	2.352	2.221	2.289	2.227	2.311	2.316	2.343
$\hat{\beta}_2$	-3.261	-3.218	-3.222	-3.122	-3.125	-3.172	-3.254	-3.200	-3.102	-2.999
$\hat{\beta}_3$	2.892	2.832	2.748	2.896	2.632	2.603	2.797	2.677	2.688	2.742

The next 2 tables demonstrate that White's-like estimate of variances of the estimates of regression coefficients (see (17)) gives values close to the empirically estimated values (estimates of regression coefficients were evaluated by IWV). The table contains only diagonal elements of covariance matrix (which are interesting).

IWV - empirical variances of the estimates of coefficients										
	1	2	3	4	5	6	7	8	9	10
$\widehat{\mathrm{var}}(\hat{\beta}_1)$	1.038	1.044	1.092	0.994	1.044	1.029	0.966	1.048	1.003	1.026
$\widehat{\mathrm{var}}(\hat{\beta}_2)$	1.007	1.078	1.011	0.964	1.073	0.965	1.032	1.067	1.064	1.000
$\widehat{\mathrm{var}}(\hat{\beta}_3)$	1.026	0.997	0.988	1.100	0.991	1.068	1.079	0.975	1.002	1.034
IWS - variances estimated by (17) (mean values over 100 samples)										
	1	2	3	4	5	6	7	8	9	10
$\widehat{\mathrm{var}}(\hat{\beta}_1)$	1.090	1.077	1.141	1.018	1.084	1.038	0.999	1.062	1.061	1.106
$\widehat{\mathrm{var}}(\hat{\beta}_2)$	1.024	1.106	1.013	0.985	1.123	1.022	1.088	1.097	1.087	1.016
$\widehat{\mathrm{var}}(\hat{\beta}_3)$	1.071	1.013	1.023	1.142	1.028	1.066	1.121	1.012	0.979	1.044

References

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J., STAHEL, W. A. (1986): *Robust Statistics – The Approach Based on Influence Functions.* J.Wiley & Sons, New York.

HETTMANSPERGER, T. P., SHEATHER, S. J. (1992): A Cautionary Note on the Method of Least Median Squares. *The American Statistician 46, 79–83.*

JUDGE, G. G., GRIFFITHS, W. E., HILL, R. C., LUTKEPOHL, H., LEE, T. C. (1985): *The Theory and Practice of Econometrics.* J.Wiley & Sons(second edition), New York.

ROUSSEEUW, P.J. (1984): Least median of square regression. *Journal of Amer. Statist. Association 79, 871-880.*

VÍŠEK, J. Á. (1992): Stability of regression model estimates with respect to subsamples. *Computational Statistics 7 (1992), 183 - 203.*

VÍŠEK, J. Á. (1994): A cautionary note on the method of the Least Median of Squares reconsidered.In: J. Á. VÍšek (Ed.): *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes.* ACADEMIA, Prague, 254 - 259.

VÍŠEK, J. Á. (1996a): Sensitivity analysis of M-estimates. *Annals of the Institute of Statistical Mathematics, 48(1996), 469-495.*

VÍŠEK, J. Á. (1996b): On high breakdown point estimation. *Computational Statistics (1996) 11:137 - 146, Berlin.*

VÍŠEK, J. Á. (2000a): On the diversity of estimates. *Computational Statistics & Data Analysis 34, (2000), 67 - 89.*

VÍŠEK, J. Á. (2000b): Regression with high breakdown point. In: J. Antoch & G. Dohnal (Eds.): *Robust 2000.* MatFyz Press, Prague, 324 - 356.

VÍŠEK, J. Á. (2002a): The least weighted squares I - The asymptotic linearity of normal equations. The least weighted squares II - Consistency and asymptotic normality. *Bulletin of the Czech Econometric Society, no. 15, 31 - 58. & Vol. 9, no. 16, 1 - 28.*

VÍŠEK, J. Á. (2002b): Sensitivity analysis of M-estimates of nonlinear regression model: Influence of data subsets. *Annals of the Institute of Statistical Mathematics, 54 (2002), 261 - 290.*

VÍŠEK, J. Á. (2004): Robustifying instrumental variables. In: J. Antoch, (Ed.): *COMPSTAT'2004.* Physica-Verlag/Springer, Berlin. 1947 - 1954.

VÍŠEK, J. Á. (2006a): Instrumental Weighted Variables - algorithm. In: A. Rizzi & M. Vichi (Eds.): *COMPSTAT 2006.* Physica-Verlag, Heidelberg 2006, 777-786.

VÍŠEK, J. Á. (2006b): Kolmogorov-Smirnov statistics in multiple regression. In: J. Antoch & G. Dohnal (Eds.): *ROBUST 2006.* MatFyz Press, Prague, 367-374.

VÍšek, J. Á. (2007): White's test for the instrumental weighted variables. Submited to the *Bulletin of the Czech Econometric Society*, presented on ICORS 2006.

VÍŠEK, J. Á. (2008a): Consistency of the instrumental weighted variables. To appear in the *Annals of the Institute of Statistical Mathematics.*

VÍŠEK, J. Á. (2008b): $\sqrt{n}$-consistency and asymptotic representation of the instrumental weighted variables. *Preprint*

VÍŠEK, J. Á. (2008c): Empirical distribution function under heteroscedasticity. Submitted to the *Statistics.*

WHITE, H. (1980): A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica, 48, 817 - 838.*

Part IX

Finance and Insurance

Modeling Tick-by-Tick Realized Correlations

Francesco Audrino[1] and Fulvio Corsi[2]

[1] Institute of Mathematics and Statistics, University of St. Gallen
Bodanstrasse 6, 9000 St. Gallen, Switzerland, *francesco.audrino@unisg.ch*
[2] University of Lugano and Swiss Finance Institute, Via Buffi 13, 6904 Lugano, Switzerland, *fulvio.corsi@lu.unisi.ch*

Abstract. We propose a tree-structured Heterogeneous Autoregressive (tree-HAR) process as a simple and parsimonious model for the estimation and prediction of tick-by-tick realized correlations. The model can account for different time and other relevant predictors' dependent regime shifts in the conditional mean dynamics of the realized correlation series. Testing the model on S&P 500 and 30 years Treasury Bond futures realized correlations, we provide empirical evidence that the tree-HAR model reaches a good compromise between simplicity and flexibility, and yields accurate single- and multi-step ahead out-of-sample forecasts, also in comparison with other standard approaches.

Keywords: high frequency data, realized correlation, stock-bond correlation, tree-structured models, HAR, regimes

1 Introduction

Asset returns cross correlation is pivotal to many prominent financial problems such as asset allocation, risk management and option pricing. Recently, the use of high frequency data has been advocated to improve the precision of asset volatility estimation yielding to the so-called Realized Volatility (RV) approach proposed in a series of breakthrough papers by Andersen et al. (2001a), (2001b) and (2003), Barndorff-Nielsen and Shepard (2001), (2002a), (2002b), and Comte and Renault (2001). As for the realized volatility approach, the idea of employing high frequency data in the computation of covariances between two assets leads to the analogous concept of *realized covariance* (or covariation); for more details, see Martens (2004), Hayashi and Yoshida (2005), Griffin and Oomen (2006), and Voev and Lunde (2007). Recently, Corsi and Audrino (2007) proposed a modified tick-by-tick realized covariance estimator in the case where a rounding in the price stamps is needed, typical situation for many practical financial data sets. *Realized correlations* are then constructed as quotients between realized covariances and products of realized standard deviations.

We propose a regime-dependent, tree-structured Heterogeneous Autoregressive (tree-HAR) model for the estimation and prediction of the tick-by-tick realized correlation series. In particular, the conditional mean dynamics

of the realized correlation series follow local linear HAR processes and are subjected to regime shifts in dependence of past values of some relevant predictor variables, like, for example, past returns, past realized volatilities or time. The local HAR processes are standard linear models where the explanatory variables are past realized correlations at three different horizons: daily, weekly and monthly; for more details, see Corsi (2004). This structure allows the model to take into account two important features exhibited by most real data realized correlation series: long memory and structural changes. Moreover, another nice feature of the tree-HAR model is that it belongs to the class of tree-structured threshold regime models and, therefore, can be easily estimated and regimes can be well interpreted in terms of relevant predictor variables; see, for example, Audrino and Bühlmann (2001), or Audrino and Trojani (2006).

We test the accuracy of the tree-HAR model on the series of daily tick-by-tick realized correlations between S&P 500 and 30 years Treasury Bond futures, collecting empirical evidence that tick-by-tick realized correlations show drastic regime shifts, supporting the evidence already found in other studies. We contribute to the literature on US stock-bond correlations by estimating local dynamics and incorporating structural breaks in a threshold-type model. The estimated tree-HAR model for daily US stock-bond realized correlations has three optimal regimes. The first regime is characterized by large losses of the US market index, whereas the second and third regimes react to positive returns or moderate losses of the S&P500 index, with an important structural break in time corresponding to March 1992.

Moreover, we perform a series of out-of-sample tests for superior predictive ability (SPA; see Hansen (2005)) of our model against a number of competitors using different goodness of fit statistics, to verify whether the more flexibility allowed by the tree-HAR model (with a corresponding higher number of parameters to be estimated) is worth value for forecasting. We empirically show that the tree-HAR model systematically outperforms the competitors, in particular when multi-period forecasts are considered.

2 Modeling realized correlations

2.1 The model

Empirical evidence on strong temporal dependence of realized correlations has been already showed, for example, in Andersen et al. (2001a), (2001b). This evidence, together with our empirical results reported in the Section 3, suggests that realized correlation series should be described by long-memory type of models.

Corsi (2004) recently proposed a class of time series models called Heterogeneous Autoregressive (HAR) models that successfully achieves the purpose of modeling the long memory behavior of financial variables in a very simple

and parsimonious way. The basic idea was to explain the long memory observed in the volatility as the superimposition of only few processes operating on different time scales. Hence, Corsi (2004) proposed a stochastic additive cascade of three different realized volatility components corresponding to the three main different time horizons present in the market: daily, weekly and monthly. This stochastic volatility cascade leads to simple AR-type models in the realized volatility with the feature of considering realized volatilities defined over different time horizons (the HAR-RV models). Although the HAR models do not formally belong to the class of long-memory models, they are able to reproduce a memory decay which is almost indistinguishable from that observed in the empirical data. The above mentioned empirical evidence on the high degree of persistence of correlations suggests that the parsimonious HAR models could also be successfully applied to model the time series of realized correlations.

A second important stylized fact that must be taken into account when building up a model for the realized correlations' dynamics is the (possible) presence of structural breaks, as it was shown in a number of recent empirical studies.

We propose a tree-structured local HAR model for the dynamics of tick-by-tick realized correlations able to take into account the above discussed stylized facts of realized correlation series: long-memory and structural breaks. Tree-structured models belong to the class of threshold regimes models, where regimes are characterized by some threshold for the relevant predictor variables. This class of model was introduced by Audrino and Bühlmann (2001) in the financial volatility literature, and has been generalized recently to capture simultaneous regime-shifts in the first and second conditional dynamics of returns series, with good results for different forecasting applications (see, for example, Audrino and Trojani (2006)).

Let $\{\widetilde{RC}\}_{t\geq 1}$ be the daily Fisher-transformed (FT) series of the tick-by-tick realized correlations $\{RC\}_{t\geq 1}$,[1] i.e.

$$\widetilde{RC}_t = \frac{1}{2}\log\Big(\frac{1+RC_t}{1-RC_t}\Big),\ RC_t \in [-1,1].$$

We then model the series $\{\widetilde{RC}\}_{t\geq 1}$ as

$$\widetilde{RC}_{t+1} = \mathbb{E}_t[\widetilde{RC}_{t+1}] + \sigma_{t+1}U_{t+1}, \tag{1}$$

where $\{U_t\}_{t\geq 1}$ is a sequence of i.i.d. innovations following the distribution p_U with expected value 0 and variance 1, and $\mathbb{E}_t[\cdot]$ denotes as usual the

[1] Note that we consider Fisher-transformed correlations not to have to impose any restriction on the parameters in the model to ensure the final estimates and forecasts to lie in the $[-1,1]$ interval. Moreover, by performing the tree-HAR analysis on the original correlations we found qualitatively exactly the same results; for example, the optimal regime structure was found to be the same.

conditional expectation given the information up to time t. The conditional dynamics of the FT-correlations are given by

$$\mathbb{E}_t[\widetilde{RC}_{t+1}] = \sum_{j=1}^{k} \big(a_j + b_j^{(d)}\widetilde{RC}_t + b_j^{(w)}\widetilde{RC}_t^{(w)} + b_j^{(m)}\widetilde{RC}_t^{(m)}\big) I_{[\mathbf{X}_t^{\text{pred}} \in \mathcal{R}_j]}; \quad (2)$$

$$\sigma_{t+1}^2 = \sum_{j=1}^{k} \sigma_j^2 I_{[\mathbf{X}_t^{\text{pred}} \in \mathcal{R}_j]}, \ \sigma_j^2 > 0, j = 1, \ldots, k, \quad (3)$$

where $\theta = (a_j, b_j^{(d)}, b_j^{(w)}, b_j^{(m)}, \sigma_j^2 : j = 1, \ldots, k)$ is a parameter vector that parameterizes the local HAR dynamics in the different regimes, k is the number of regimes (endogenously estimated from the data), and $\widetilde{RC}_t^{(w)}$ and $\widetilde{RC}_t^{(m)}$ are respectively the weekly and monthly FT-realized correlations, obtained as simple averages of 5 resp. 22 daily FT-realized correlations. The regimes are characterized by partition cells $\mathcal{R}_j$ of the relevant predictor space G of $\mathbf{X}_t^{\text{pred}}$:

$$G = \bigcup_{j=1}^{k} \mathcal{R}_j, \ \mathcal{R}_i \cap \mathcal{R}_j = \emptyset \ (i \neq j).$$

In our study, the relevant predictor variables in $\mathbf{X}_t^{\text{pred}}$ are past lagged FT-realized correlations, and past lagged realized volatilities and returns of the two instruments under investigation. All such predictor variables are considered at three different time horizons: daily, weekly and monthly. Moreover, we also consider as an additional predictor variable time.

To completely specify the conditional dynamics given in (2)-(3) of the FT-realized correlations, we decide the shape of the partition cells $\mathcal{R}_j$ which are admissible in the tree-HAR model. Similarly to the standard Classification and Regression Trees (CART) procedure, the only restriction we make is that regimes must be characterized by rectangular partition cells with edges determined by thresholds on the predictor variables. Such partition cells are practically constructed using a binary tree. Introducing this restriction has two big advantages: it allows for a clear interpretation of the regimes in terms of relevant predictor variables, and allows to estimate the model also using large-dimensional predictor spaces G.

2.2 Estimation

The tree-HAR model introduced in (1)-(3) can be estimated using quasi maximum likelihood (QML). Conditionally on some reasonable starting values, the negative quasi log-likelihood for model (1)-(3) is given by

$$-l\big(\theta; (\widetilde{RC}, \mathbf{X}^{\text{pred}})_1^n\big) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\sum_{t=1}^{n} \log\big(\sigma_t^2(\theta)\big) + \frac{1}{2}\sum_{t=1}^{n} \frac{\big(\widetilde{RC}_t - \mathbb{E}_{t;\theta}[\widetilde{RC}_{t-1}]\big)^2}{\sigma_t^2(\theta)} \quad (4)$$

Therefore, for any fixed sequence of partition cells the tree-HAR model can be estimated by QML. The choice of the optimal partition cells (i.e. splitting variables and threshold values) involves a model choice procedure for non-nested hypotheses. Similarly to CART and as already discussed in Audrino and Bühlmann (2001) and Audrino and Trojani (2006) for general tree-structured models, the model selection of the optimal splitting variables and threshold values can be performed via a tree-structured partial search. Within any data-determined tree structure the optimal model is selected using the Bayesian-Schwartz information criterion (BIC). For all details about the flexible procedure used to estimate the model, we refer to Audrino and Trojani (2006), Section 2.3 and Appendix A. Proof of the consistency of the conditional mean and volatility estimates in the tree-HAR model under a possible model mis-specification can be derived from Theorem 1 in Audrino and Bühlmann (2001).

3 Empirical application

3.1 Data and estimation results

We consider a tick-by-tick bivariate returns series of the S&P 500 Index and 30 years US Treasury Bond futures for the period from January 1990 to October 2003, for a total of 3,391 daily observations. The data come from the `Price-data.com` data base with time stamps rounded at the one minute frequency. Combining the First-Last realized covariance estimator introduced by Corsi and Audrino (2007) together with the Multi-Scales Discrete Sine Transform realized volatility estimator (see Curci and Corsi (2003)), we are now able to construct a realized correlation measure where both the volatilities and the covariances are computed from tick-by-tick data. As usual, correlations are computed as quotients between covariances and products of standard deviations.

We start the analysis by estimating the tree-HAR model (1)-(3) on the whole data sample. The estimated tree-HAR model has three optimal regimes (endogenously estimated from the data) that can be clearly interpreted as follows. The first regime is in reaction of US market crashes: in particular, the first regime is characterized by large negative past S&P500 daily returns, conditional mean dynamics of the realized correlations are highly persistent, and the volatility of realized correlations is large. The second and third regimes are both characterized by relatively positive[2] past S&P500 daily returns, but for two different time periods. In fact, we identify a structural break in time corresponding to the period February - March 1992. This structural break may be a consequence of the Western European monetary crisis of 1992-1993. After March 1992 the persistence of the conditional mean dynamics and the

[2] As usual for threshold regimes, positive here means above the optimal threshold value.

volatility of the realized correlations significantly increase. Almost all coefficients in the local dynamics of the conditional mean and variance of the Fisher-transformed (FT) realized correlations (with only one exception) are highly significant.

3.2 Forecasting results

To better validate the goodness of the tree-HAR model for the real data under investigation, we investigate its forecasting ability, always in comparison with different competitors introduced in the literature: the standard AR(1) model, the ARMA(1,1) model, the ARIMA(1,1,1) model introduced for non stationary time series, and the global HAR model. In particular, we perform a series of out-of-sample tests to assess the forecasting power of the tree-HAR model for one-period and multi-periods predictions. As goodness of fit statistics we consider the out of sample MAE and MSE of the residuals. In addition to these performance measures, we also report results for the out-of-sample log-likelihood in (4) in the single-period out-of-sample test, and the R^2 we get when regressing realizations against forecasts at the same time t (Mincer-Zarnowitz regression).

Single-period forecasts To get the daily forecasts we use a rolling strategy. The models are re-estimated every month (22 days) using all past data available in the sample. The initial in-sample period goes from January 1990 to December 1999. As a consequence, we get 926 out-of-sample daily forecasts (until October 2003). Results are summarized in Table 1. Between parentheses we report the p-values of Superior Predictive Ability (SPA) tests introduced by Hansen (2005) for the null-hypothesis that the chosen model is not inferior to any of the alternatives.

Single-period forecasting results.

Model	Loglik.	MAE	MSE	R^2
AR(1)	155.7 (0)	0.2055 (0)	0.0753 (0)	0.3911
ARMA(1,1)	−82.10 (0.036)	0.1615 (0.222)	0.0480 (0.049)	0.4795
ARIMA(1,1,1)	−93.69 (0.088)	0.1601 (0.777)	0.0469 (0.785)	0.4810
HAR	−88.22 (0.042)	0.1602 (0.842)	0.0474 (0.148)	0.4806
Tree-HAR	−109.1 (0.617)	0.1601 (0.407)	0.0471 (0.527)	0.5077

Table 1. Comparative results of 1 day ahead forecasts of S&P - US Bond FT-realized correlations obtained using the classical AR(1), ARMA(1,1), ARIMA(1,1,1) models, the global HAR model, and the tree-HAR model.

The tree-HAR model yields the best results for three out of the four goodness of fit statistics considered: with respect to the MSE statistics the ARIMA(1,1,1)

model is slightly better. However, differences measured by the MAE and MSE statistics are in most cases very small and not statistically significant. Only the simple AR(1) and ARMA(1,1) models are clearly beaten by the competitors. On the contrary, with respect to the out-of-sample likelihood the tree-HAR model yields significant improvements in the accuracy of the FT-realized correlation forecasts over the competitors.

Multi-period forecasts Practical asset allocation applications would typically require correlation forecasts at longer time horizons. Therefore, we perform two out-of-sample tests at weekly (i.e. 5 days) and monthly (i.e. 22 days) horizons to assess the accuracy of the multi-period ahead forecasts obtained using the different approaches. Such multi-period ahead predictions are constructed using Filtered Historical Simulation (FHS): see Barone-Adesi et al. (1999).

Like in the previous out-of-sample experiment, we use the same rolling strategy and initial in-sample period. Results are summarized in Table 2. Once again p-values of SPA tests are reported between parentheses.

Multi-period forecasting results: 1 week horizon.

Model	MAE	MSE	R^2
AR(1)	0.4243 (0)	0.2431 (0)	0.2810
ARMA(1,1)	0.1777 (0.085)	0.0600 (0.002)	0.3853
ARIMA(1,1,1)	0.1763 (0.274)	0.0565 (0.074)	0.3859
HAR	0.1756 (0.346)	0.0579 (0.005)	0.3861
Tree-HAR	0.1742 (0.776)	0.0557 (0.497)	0.4298

Multi-period forecasting results: 1 month horizon.

Model	MAE	MSE	R^2
AR(1)	0.5939 (0)	0.4346 (0)	0.0853
ARMA(1,1)	0.2329 (0)	0.0974 (0)	0.2261
ARIMA(1,1,1)	0.2118 (0.061)	0.0766 (0.051)	0.2234
HAR	0.2135 (0.006)	0.0813 (0)	0.2381
Tree-HAR	0.2066 (0.533)	0.0751 (0.501)	0.2837

Table 2. Comparative results of 1 week and 1 month ahead forecasts of S&P - US Bond FT-realized correlations obtained using the classical AR(1), ARMA(1,1), ARIMA(1,1,1) models, the global HAR model, and the tree-HAR model.

The better forecasting power of the tree-HAR model with respect to all competitors for multi-period predictions is clearly pointed out by the results of the SPA tests. Especially for longer-time ahead forecasts (i.e. 1 month), the predictions obtained using the tree-HAR model outperform those from the alternative approaches. Gains are in most cases statistically significant.

References

ANDERSEN, T.G., BOLLERSLEV, T., DIEBOLD, F.X. and EBENS, H. (2001a): The distribution of realized stock return volatility. *Journal of Financial Economics 61 (1), 43-76.*

ANDERSEN, T.G., BOLLERSLEV, T., DIEBOLD, F.X. and LABYS, P. (2001b): The distribution of exchange rate volatility. *Journal of the American Statistical Association 96, 42-55.*

ANDERSEN, T.G., BOLLERSLEV, T., DIEBOLD, F.X. and LABYS, P. (2003): Modeling and forecasting realized volatility. *Econometrica 71 (2), 579-625.*

AUDRINO, F. and BÜHLMANN, P. (2001): Tree-structured GARCH models. *Journal of the Royal Statistical Society, Series B 63, 727-744.*

AUDRINO, F. and TROJANI, F. (2006): Estimating and predicting multivariate volatility regimes in global stock markets. *Journal of Applied Econometrics 21 (3), 345-369.*

BARNDORFF-NIELSEN, O.E. and SHEPARD, N. (2001): Non-gaussian ornstein-uhlembech-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society, Series B (63), 167-241.*

BARNDORFF-NIELSEN, O.E. and SHEPARD, N. (2002a): Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society, Series B (64), 253-280.*

BARNDORFF-NIELSEN, O.E. and SHEPARD, N. (2002b): Estimating quadratic variation using realized variance. *Journal of Applied Econometrics 17 (5), 457-477.*

BARONE-ADESI, G., GIANNOPOULOS, K. and VOSPER, L. (1999): VaR Without Correlations for Portfolio of Derivative Securities. *Journal of Futures Markets 19 (April), 583-602.*

COMTE, F. and RENAULT, E. (2001): Long memory in continuous time stochastic volatility models. *Mathematical Finance 8, 291-323.*

CORSI, F. (2004): Simple long memory models of realized volatility. Manuscript, University of Lugano.

CORSI, F. and AUDRINO, F. (2007): Realized covariance tick-by-tick in presence of rounded time stamps and general microstructure effects. Working paper, University of Lugano.

CURCI, G. and CORSI, F. (2003): A discrete sine transform approach for realized volatility measurement. NCCR FINRISK Working Paper No. 44.

GRIFFIN, J.E. and OOMEN, R.C.A. (2006): Covariance measurement in the presence of non-synchronous trading and market microstructure noise. Unpublished Manuscript.

HANSEN, P.R. (2005): A test for superior predictive ability. *Journal of Business & Economic Statistics 23, 365-380.*

HAYASHI, T. and YOSHIDA, N. (2005): On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli 11 (2), 359-379.*

MARTENS, M. (2004): Estimating unbiased and precise realized covariances. Social Science Research Network Electronic Library.

VOEV, V. and LUNDE, A. (2007): Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics 5, 68-104.*

Heterogeneous Hidden Markov Models

José G. Dias[1], Jeroen K. Vermunt[2] and Sofia Ramos[3]

[1] Department of Quantitative Methods, ISCTE – Higher Institute of Social Sciences and Business Studies, Edifício ISCTE, Av. das Forças Armadas, 1649–026 Lisboa, Portugal,
jose.dias@iscte.pt

[2] Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands,
J.K.Vermunt@uvt.nl

[3] Department of Finance, ISCTE – Higher Institute of Social Sciences and Business Studies, Edifício ISCTE, Av. das Forças Armadas, 1649–026 Lisboa, Portugal,
sofia.ramos@iscte.pt

Abstract. Heterogeneous hidden Markov models (HHMMs) are models with time-constant and time-varying discrete latent variables that capture unobserved heterogeneity between and within clusters, respectively. We apply HHMMs in modeling financial return indexes from seven markets. The return-risk patterns of the encountered latent states correspond to the well-known bear and bull market states.

Keywords: latent class model, finite mixture model, hidden Markov model, model-based clustering, stock indexes

1 Introduction

Latent class or finite mixture modeling has proven to be a powerful tool for analyzing unobserved heterogeneity in a wide range of social and behavioral science data (see, for example, McLachlan and Peel (2000)). We introduce a specific latent class model for time series analysis that takes into account unobserved heterogeneity by means of time-constant and time-varying discrete latent variables.

Here, this methodology is used to model the dynamics of the returns of seven stock market indexes. As illustrated below, the proposed approach is flexible in the sense that it can deal with the specific features of financial time series data, such as asymmetry, kurtosis and unobserved heterogeneity, aspects that are almost always ignored in finance research. Having selected a heterogeneous sample of countries including both developed and emerging countries from the American region, we expect that heterogeneity in market returns due to country idiosyncrasies will show up in the results. For instance, emerging market return distributions show larger deviations from normality; i.e., are more skewed and have fat tails (Harvey, 1995).

The paper is organized as follows: Section 2 presents the full mixture hidden Markov model; Section 3 describes the seven stock market time series that are used throughout this paper. Section 4 reports HHMM estimates. The paper concludes with a summary of the main findings.

2 The heterogeneous hidden Markov model (HHMM)

We model simultaneously the time series of n stock markets. Let y_{it} represent the response of observation (stock market) i at time point t, where $i \in \{1, \ldots, n\}$, $t \in \{1, \ldots, T\}$, and $y_{it} \in \Re$. In addition to the observed "response" variable y_{it}, the HHMM contains two different latent variables: a time-constant discrete latent variable and a time-varying discrete latent variable. The former, which is denoted by $w \in \{1, ..., S\}$, is used to capture the unobserved heterogeneity across stock markets; that is, stock markets are clustered based on differences in their dynamics. We will refer to a model with S clusters as HHMM-S. The two-state time-varying latent variable is denoted by $z_t \in \{1, 2\}$. Changes between the two states or regimes between adjacent time points are assumed to be in agreement with a first-order Markov or first-order autocorrelation structure.

Let $f(\mathbf{y}_i; \varphi)$ be the (probability) density function associated with the index return rates of stock market i, where φ is the vector of parameters in the model. The HHMM-S defines the following parametric model for this density:[1]

$$f(\mathbf{y}_i; \varphi) = \sum_{w=1}^{S} \sum_{z_1=1}^{2} \cdots \sum_{z_T=1}^{2} f(w) f(z_1|w) \prod_{t=2}^{T} f(z_t|z_{t-1}, w) \prod_{t=1}^{T} f(y_{it}|z_t). \quad (1)$$

As in any mixture model, the observed data density $f(\mathbf{y}_i; \varphi)$ is obtained by marginalizing over the latent variables. Because in our model these are discrete variables, this simply involves the computation of a weighted average of class-specific probability densities where the (prior) class membership probabilities or mixture proportions serve as weights (McLachlan and Peel, 2000). We assume that within cluster w the sequence $\{z_1, \ldots, z_T\}$ is in agreement with a first-order Markov chain. Moreover, we assume that the observed return at a particular time point depends only on the regime at this time point; i.e, conditionally on the latent state z_t, the response y_{it} is independent of returns at other time points, which is often referred to as the local independence assumption. As far as the first-order Markov assumption for the latent regime switching conditional on cluster membership w is concerned, it is important to note that this assumption is not as restrictive as one may initially think. It does clearly not imply a first-order Markov structure for the responses y_{it}. The standard or hidden Markov model (HMM) (Baum et

[1] For details on model specification and estimation, see Dias et al. (2007).

al., 1970) is a special case of the HHMM-S that is obtained by eliminating the time-constant latent variable w from the model, that is, by assuming that there is no unobserved heterogeneity.

The characterization of the HHMM is provided by:

- $f(w)$ is the prior probability of belonging to a particular latent class or cluster w with multinomial parameter $\pi_w = P(W = w)$;
- $f(z_1|w)$ is the initial-regime probability; that is, the probability of having a particular initial regime conditional on belonging to latent class w with Bernoulli parameter $\lambda_{kw} = P(Z_1 = k|W = w)$;
- $f(z_t|z_{t-1}, w)$ is a latent transition probability; that is, the probability of being in a particular regime at time point t conditional on the regime at time point $t-1$ and class membership; assuming a time-homogeneous transition process, we have $p_{jkw} = P(Z_t = k|Z_{t-1} = j, W = w)$ as the relevant Bernoulli parameter. In other words, within cluster w one has the transition probability matrix

$$\mathbf{P}_w = \begin{pmatrix} p_{11w} & p_{12w} \\ p_{21w} & p_{22w} \end{pmatrix},$$

  with $p_{12w} = 1 - p_{11w}$ and $p_{22w} = 1 - p_{21w}$. Note that the HHMM-S allows that each cluster has its specific transition or regime-switching dynamics, whereas in a standard HMM it is assumed that all cases have the same transition probabilities;
- $f(y_{it}|z_t)$, the probability density of having a particular observed stock return in index i at time point t, conditional on the regime occupied at time point t, is assumed to have the form of a univariate normal (or Gaussian) density function. This distribution is characterized by the parameter vector $\theta_k = (\mu_k, \sigma_k^2)$ containing the mean (μ_k) and variance (σ_k^2) for regime k. Note that these parameters are assumed to be equal across clusters, an assumption that may, however, be relaxed.

Since $f(\mathbf{y}_i; \varphi)$, defined by Equation (1), is a mixture of densities across clusters w and regimes, it defines a flexible Gaussian mixture model that can accommodate deviations from normality in terms of skewness and kurtosis. The two-state HHMM-S has $4S+3$ free parameters to be estimated, including $S-1$ class sizes, S initial-regime probabilities, $2S$ transition probabilities, 2 conditional means, and 2 conditional variances.

Maximum likelihood (ML) estimation of the parameters of the HHMM-S involves maximizing the log-likelihood function: $\ell(\varphi; \mathbf{y}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i; \varphi)$, a problem that can be solved by means of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). In the E step, we compute the joint conditional distribution of the $T+1$ latent variables given the data and the current provisional estimates of the model parameters. In the M step, standard complete data ML methods are used to update the unknown model parameters using an expanded data matrix with the estimated densities of the

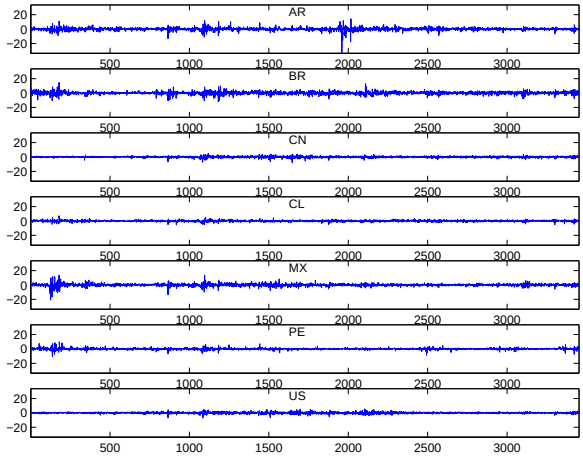


Fig. 1. Time series of index rates for seven American region stock markets.

latent variables as weights. Since the EM algorithm requires us to compute and store the $S{\cdot}2^T$ entries in the E step this makes this algorithm impractical or even impossible to apply with more than a few time points. However, for hidden Markov models, a special variant of the EM algorithm has been proposed that is usually referred to as the forward-backward or Baum-Welch algorithm (Baum et al., 1970). The Baum-Welch algorithm circumvents the computation of this joint posterior distribution making use of the conditional independencies implied by the model.

An important modeling issue is the selection of the value of S, the number of clusters needed to capture the unobserved heterogeneity across stock markets. The selection of S is typically based on information statistics such as the Bayesian Information Criterion (BIC) of Schwarz (Schwarz, 1978). In our application we select S that minimizes the BIC value defined as:

$$BIC_S = -2\ell_S(\hat{\varphi}; \mathbf{y}) + N_S \log n, \tag{2}$$

where N_S is the number of free parameters of the model concerned and n is the sample size.

3 Data set

The data set used in this article are daily closing prices from 4 July 1994 to 27 September 2007 for seven stock market indexes from the American region drawn from Datastream database and listed in Table 1. The series are expressed in US dollars. In total, we have 3454 end-of-the-day observations per country. Let P_{it} be the observed daily closing price of market i on day t, $i = 1, ..., n$ and $t = 0, ..., T$. The daily rates of return are defined as the

Table 1. Summary statistics.

Stock market	Mean	Median	Std. Deviation	Skewness	Kurtosis	Jarque-Bera test	
						statistics	p-value
Argentina (AR)	0.001	0.031	1.940	-1.756	34.648	173761.38	0.000
Brazil (BR)	0.055	0.077	1.961	-0.231	4.944	3527.58	0.000
Canada (CN)	0.054	0.096	0.997	-0.691	4.706	3442.42	0.000
Chile (CL)	0.027	0.000	1.024	-0.150	3.212	1487.90	0.000
Mexico (MX)	0.035	0.081	1.737	-0.807	16.347	38647.91	0.000
Peru (PE)	0.043	0.029	1.144	0.114	12.709	23139.40	0.000
United States (US)	0.038	0.042	1.040	-0.147	4.027	2332.37	0.000

percentage rate of return $y_{it} = 100 \times \log(P_{it}/P_{i,t-1})$, $t = 1, ..., T$, with $T = 3454$.

Table 1 provides descriptive statistics of the time series, while Figure 1 depicts the full time series. The sample period includes periods of market instability as the Mexican crisis of 1994, the 1999 Brazilian crisis, the Argentine crises in 2001-2002, and the global stock market downturn of the 2001 Internet bubble. It can be seen that both the mean and the median return rates are all positive and close to zero. Stock markets show, instead, very diverse patterns of dispersion, where the largest standard deviations are found in Brazil and Argentina and the smallest dispersion in Canada, Chile and the United States. Higher standard deviations are typical for emerging markets, known for their high risk. All but one return rate distributions are negative-skewed and the kurtosis (which equals 0 for normal distributions) shows values above 0, indicating heavier tails and more peakness than the normal distribution. The Jarque-Bera test rejects the null hypothesis of normality for each of the seven stock markets. Overall, market features seem well suited to be modeled using HHMMs.

4 Results

This Section reports the results obtained when applying the HHMM-S described before to the seven stock markets. We estimated models characterized by different number of clusters ($S = 1, ..., 8$), using for the estimation of each of them 300 different starting values for the parameters to avoid local maxima. The model with three latent classes ($S = 3$) yielded the lowest BIC value ($\ell_3(\hat{\varphi}; \mathbf{y}) = -38482.3$, $N_3 = 15$ and $BIC_3 = 76993.8$).[2]

Table 2 summarizes the results related to the distribution of stock market across latent classes which gives the size of each cluster. The estimated prior class membership probability is somewhat larger for Class 1 (0.542). From the posterior class membership probabilities, the probability of belonging to each of the clusters conditional on the observed data (Table 2), we have four

[2] For $n = 7$ BIC and AIC (Akaike Information Criterion) are very similar ($\log n = 1.95 \simeq 2$).

Table 2. Estimated prior probabilities, posterior probabilities, and modal classes for the HHMM-3.

Stock market	Latent class 1	Latent Class 2	Latent Class 3	Modal class
Prior probabilities	0.542	0.292	0.167	
Posterior probabilities				
Argentina (AR)	0.000	1.000	0.000	2
Brazil (BR)	0.000	0.000	1.000	3
Canada (CN)	1.000	0.000	0.000	1
Chile (CL)	1.000	0.000	0.000	1
Mexico (MX)	0.000	1.000	0.000	2
Peru (PE)	1.000	0.000	0.000	1
United States (US)	1.000	0.000	0.000	1

countries assigned to Class 1 (Canada, Chile, Peru and United States), two countries to Class 2 (Argentina and Mexico) and the remaining one – Brazil – to Class 3. Based on this classification, one clearly has to reject the hypothesis that stock markets can clustered regionally. With the exceptions of Canada and the USA, and Peru and Chile, which share the same class, neighbor countries tend to be allocated into different classes. Notice that from the posterior probabilities the modal allocation into classes is precise (the probability of the most likely class is always one). Class 1 contains two developed countries, Canada and the USA, and two emerging markets. By combining the classification information with the descriptive statistics in Table 1, we conclude that Class 1 contains the countries with the lowest volatilities.

Table 3 provides information on the two regimes that were identified; that is, the average proportion of markets in regime k over time and the mean and variance of the returns in regime k. The result is in line with the common dichotomization of financial markets into "bull" and "bear" markets. Consistently, the reported means show that one of the regimes is associated with positive returns (bull market) and the other with negative returns (bear market). The probability of being in the bear and bull regimes is 0.26 and 0.74, respectively. We would also like to emphasize that these results are coherent with the common acknowledgment of volatility asymmetry of financial markets. Volatility is likely to be higher when markets fall than when markets rise.

Table 4 reports the estimated probabilities of being in one of the regimes for each latent class. There is a clear distinction between classes. Class 1

Table 3. Estimated marginal probabilities of the regimes and within Gaussian parameters.

	$P(Z)$		Return (mean)		Risk (variance)	
	Regime 1	Regime 2	Regime 1	Regime 2	Regime 1	Regime 2
Estimate	0.7438	0.2562	0.0884	-0.1251	0.7079	6.5757
Std. error	(0.0603)	(0.0603)	(0.0048)	(0.0238)	(0.0123)	(0.1715)

Table 4. Characterization of the switching regimes.

	Latent class 1		Latent Class 2		Latent Class 3	
	Regime 1	Regime 2	Regime 1	Regime 2	Regime 1	Regime 2
$P(Z\|W)$	0.8903	0.1097	0.6293	0.3707	0.4683	0.5317
	(0.0120)	(0.0120)	(0.0212)	(0.0212)	(0.0299)	(0.0299)
Transitions						
Regime 1	0.988	0.012	0.933	0.067	0.895	0.105
	(0.0014)	(0.0976)	(0.0071)	(0.0071)	(0.0147)	(0.0147)
Regime 2	0.098	0.902	0.113	0.887	0.092	0.908
	(0.0114)	(0.0114)	(0.0125)	(0.0125)	(0.0148)	(0.0148)

has the largest probability of being in bull regime (0.89). For Class 2 this probability becomes 0.63. In case of Brazil (Class 3), it is more likely to be in a bear regime than in a bull regime during the period of analysis. Moreover, Table 4 provides another key result of our analysis. It gives the transition probabilities between the two regimes for each of the three latent classes. First, notice that all classes show regime persistence. Once a stock market jumps to a regime, it is likely to remain within the same regime for a while, which is coherent with stylized facts in financial markets. Second, Class 1 shows the lowest propensity to move from a bull regime to a bear regime. This propensity is higher for Class 2 and even higher for Class 3. Note that Class 2 and 3 were severely affected by crises during the sampled period. Third, Class 2 shows the highest probability of jumping from a bear to a bull regime.

Figure 4 shows the regime-switching dynamics of the countries within each of our three latent classes. It depicts the posterior probability of being in bull regime at period t, where the grey color identifies periods in which this probability is below 0.5 which corresponds to a higher likelihood of being in the bear state. The three clusters of countries have rather different pattern of regime switching. Class 1 is more regime persistent with short duration bear regimes that did not turn out to be endemic during the period of analysis. Both Class 2 and Class 3 are extremely dynamic and tend to move very fast between regimes. However, Class 3 tends to be more persistent in bear regimes.

5 Conclusions

The HHMM allows model-based clustering of time series. In the analysis of a sample of seven stock markets providing observations for a period of 3454 days the best fitting model was the one with three latent classes. The three latent classes clearly distinguished three types of regime switching, which is coherent with many stylized facts in finance.

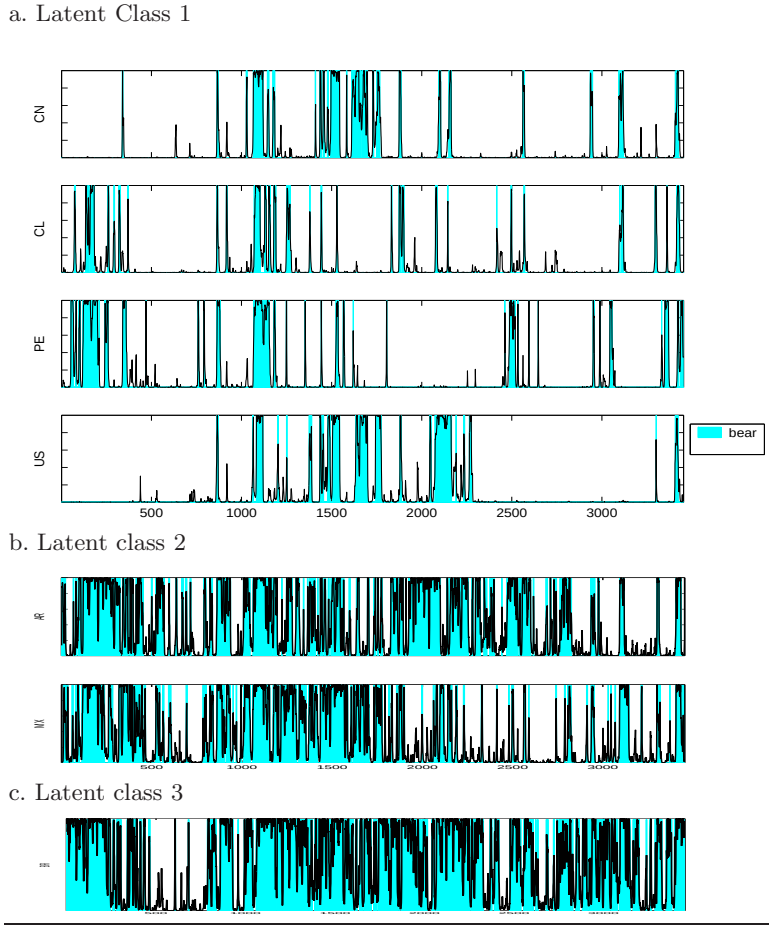


Fig. 2. Estimated posterior bull regime probability and modal regime.

References

BAUM, L.E., PETRIE, T., SOULES, G., WEISS, N. (1970): A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics 41, 164–171.*

DEMPSTER, A.P., LAIRD, N.M., RUBIN, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B 39, 1–38.*

DIAS, J.G., VERMUNT, J.K., RAMOS, S. (2007): Analysis of heterogeneous financial time series using a mixture Gaussian hidden Markov model. *Working Paper, Tilburg University.*

HARVEY, C.R. (1995): Predictable risk and returns in emerging markets. *Review of Financial Studies 8, 773–816.*

McLACHLAN, G.J., PEEL, D. (2000): *Finite Mixture Models*. John Wiley & Sons, New York.
SCHWARZ, G. (1978): Estimating the dimension of a model. *Annals of Statistics 6, 461–464*.

An Insurance Type Model for the Health Cost of Cold Housing: an Application of GAMLSS

Robert Gilchrist, Alim Kamara and Janet Rudge

STORM, London Metropolitan University, UK
Holloway Road, London, N7 8DB, U.K., *r.gilchrist@londonmet.ac.uk*

Abstract. This paper introduces a substantive problem, namely the link between fuel poverty and excess winter morbidity amongst older people, and shows how the GAMLSS suite of programs (www.gamlss.com) can be used to provide a very flexible method of modelling both the number of hospital admissions and the corresponding lengths of stay in hospital. The approach is closely related to the models that have been used to model the number of insurance claims, and their cost (see Heller et al.(2007)). We here fit the Beta Binomial distribution to the number of episodes, and we fit a variety of continuous distributions to the lengths of stay, incorporating random effects to allow for over-dispersion.

Keywords: Beta Binomial, GAMLSS, insurance, morbidity, costs

1 Introduction

Fuel poverty is defined as the inability to afford adequate warmth in the home and is related to poor energy efficiency of homes as well as householders' incomes. The U.K. overnment calculated numbers of fuel poor households in England as 1.2 million in 2004, while those fuel poor households classed as vulnerable numbered one million (DEFRA (2006)). Older households are the group most vulnerable to fuel poverty, and are also particularly susceptible to cold-related health effects. The significant numbers recognised as fuel poor have as yet unrecognised implications for costs to public services.

Conventionally, research has referred to effects of cold homes in terms of excess winter deaths (e.g. Wilkinson et al. (2001)). These deaths are known to be associated with outdoor winter temperatures, but direct evidence of links to low indoor temperatures is limited. Mortality statistics disguise the full extent of potentially long-term chronic conditions exacerbated by cold. Hence we have concentrated on measuring excess winter morbidity (illness) in relation to fuel poverty, rather than mortality, because of the consequent implications for winter pressures on health services.

We have previously demonstrated links between fuel poverty risk and excess winter hospital episodes among older people in Newham, using this excess as a measure of associated morbidity (e.g. Rudge and Gilchrist (2007)). In this paper we refer to that work and describe the means by which this

measure could be developed as a costing element for a health impact assessment tool. The results could contribute to the debate regarding the case for increased energy efficiency investment on public health grounds, in addition to the accepted environmental grounds. Our methodology is closely related to the models that have been used for insurance claims.

2 Substantive background to fuel poverty and poor health

The U.K. Department of Health (2001) recognises that fuel poverty affects health inequalities, particularly among older people. The potential benefits of energy efficiency investment for older fuel poor households involve improvements in comfort, health and well being. Identifying cost savings associated with such benefits is complicated by the many confounding factors involved in showing direct causal links between housing characteristics and health.

There are no current precise methods of calculating the cost to the health services of cold-related disease arising from poor housing. The newly prevailing emphasis on dealing with climate change and carbon emissions may deflect attention from the needs of the fuel poor, who cannot afford to use energy extravagantly. Energy-saving targets tend to skew energy-efficiency investment in favour of fuel-rich households. However, public health implications demand that such investment should also be health-driven.

3 Data and statistical methodology

The main source of data here considered is our existing database for Newham hospital admissions over 1993-96. These data are anonymised with respect to individuals, having been provided at enumeration district (ED) level.

Our previous work examined the excess morbidity for different ages and genders in terms of a range of explanatory variables. We now propose extending this work by analysing daily episodes by length of stay and investigating the associated costs for such episodes. Our proposed methodology is based upon the modelling of the propensity for an individual to be an emergency respiratory hospital admission, together with the duration of stay in hospital for such admissions. This approach is similar to that used for insurance claims (see e.g. Heller, et al.(2007)) in which the probability of a claim and the size of a claim are both modelled. Having modelled the probability of being a hospital respiratory admission and the length of the consequential stays in hospital, we will use data on the average cost of such hospital admissions, adjusted for the duration of stay, to give a model for the cost of the Newham admissions. The effect of FPR will be determined by considering the excess cost in winter over that in summer.

Our methodology utilises the R-based GAMLSS package (see Rigby and Stasinopoulos (2005) and www.gamlss.com). GAMLSS is a suite of programs

written in R (see www.r-project.org). We consider the probability of being admitted as following a Beta Binomial distribution, this being a more flexible extension of the more traditional Binomial distribution. Our 'default' approach is to use logistic regression. The corresponding length of episode is modelled from a selection of continuous distributions.

The GAMLSS package allow us easily to find the the maximum likelihood estimates of the several parameters of a wide range of distributions and to incorporate random effects and smoothing terms. We can make use of the many facilities of R, such as automatic model selection, and we can easily access the wide range of diagnostics available in R. Up to 4-distributional parameters can be modelled in terms of the risk factors. We can assess the expected average stay (and cost) and the variability of the stay (and cost). The potential risk factors are shown in the accompanying **Table 1**. We utilise nominal factors ED, gender and age to allow differing parameters to be fitted for the differing numbers of 'at risk' males and females, of differing ages, in each enumeration district. The definition of the fuel poverty index FPR is discussed further below. Potential confounding factors are considered, using 1991 Census data, including pensioners with limiting long term illness and ethnic composition. Daily weather data were obtained for 1993-1997. The lagged influence of weather is considered, together with maximum, minimum and mean monthly temperature and average monthly rainfall, wind speed, hours of sunshine, and solar radiation levels.

The chance of a repeat admission of an individual appears to be low, although this is not easy to determine precisely as the original data predated inclusion of patient identifier codes. Hence, although an assumption of independence of the observed admissions and of the observed lengths of episode is not too unreasonable, some correlation between occurrences can be expected, as can some correlation between lengths of episode. Thus in modelling both probability of admission and length of stay, we incorporated a random effect in our linear predictor to allow for any over-dispersion caused by the unknown correlation.

4 Definition of the fuel poverty risk index

Our population-based study of the London Borough of Newham involved creating a Fuel Poverty Risk Index (FPR), derived from known risk factors, to compare with a cold-related health indicator, based on excess winter emergency respiratory hospital admissions (see Rudge and Gilchrist (2005)). Our data level was limited to small areas, rather than individuals, for patient anonymity reasons.

Datasets were collated for enumeration districts (EDs), which contain, on average, about 220 households: we collected data on household age, size and tenure from the 1991 Census; Council Tax Benefit (CTB) for 1998; estimated energy ratings for dwellings, based on classification by tenure (census

Variable)	Description
hh1 #	% households with one or more pensioner(s)
hh2	% small households (one or two persons households)
undoc #	% households under-occupied (1 person with 4 rooms; 2 person with 5 rooms)
lowsap #	% dwellings with poor energy efficiency (below SAP35**)
ctb #	% households in receipt of Council Tax Benefit (indicator of low income)
tow	Townsend deprivation score
ch	% households with no central heating
pens	% lone pensioner households with no central heating
pre	% dwellings built before 1945
pensm	total male pensioners as % of total population
pensf	total female pensioners as % of total population
penswh	% of white pensioners in the ED.
FPR	Fuel Poverty Risk Index = (hh1*undoc*lowsap*ctb)* 10^{-3}
pwh	White pensioners (% total pensioners)
mmeant	Monthly mean air temperature, $^{\circ}C$
mmaxt	Monthly maximum air temperature, $^{\circ}C$
mmint	Monthly minimum air temperature, $^{\circ}C$
mrain	Monthly rainfall totals, mm
msun	Monthly sunshine hours
mmwd	Monthly mean wind speed
msol	Monthly solar radiation, W hr m-2
mtdif	Difference from previous month mean temperatures, $^{\circ}K$
dwigs	Total number of dwellings
house	Total number of households
pop	% population 65 years old or more
age	(1) 65-74, (2) 75-84, (3) 85+
nage	Age with 2 levels only: (1) 65-84 (2) 85+
sex	(1) Male, (2) female
q	Season factor with 3 levels: (1) Summer, (2) November, January, February, (3) December
nq	Season factor with 2 levels: (1) Not December (2) December
z	Factor with 48 levels denoting month
E	factor with 450 levels specifying enumeration district (ED)

Table 1. Explanatory variables and factors. $**SAP35$ is energy rating, or measure of energy efficiency, on a scale of 0 - 100, where 0 is poorest. # denotes component of FPR.

data), size and type (from a drive round survey and census) and building age; numbers of emergency episodes for all respiratory diagnoses for patients aged above 64 years for August 93 to July 97 from Hospital Episode Statistics

(HES). (Emergency admissions are more likely to reflect seasonal effects than elective admissions.)

The FPR was calculated for EDs as a product of the following (unweighted) factors, all as percentages of total households or total dwellings:

- households with one or more pensioners
- households in receipt of CTB (indicating low income)
- dwellings with poor energy efficiency (i.e. below the 1991 national average energy rating)
- under occupancy (small households occupying relatively large homes).

5 A statistical model for the expected total duration of emergency respiratory hospital admissions

We here develop a model to explain the observed illness counts in each ED, in each month, in terms of the potential explanatory variables, and notably FPR. We model the counts for males and females, and for the three age categories. We consider data for each of 48 months. Our particular interest is in the difference between the counts observed in summer and winter, and whether we can explain this difference in terms of the explanatory variables. To examine this we model the probability p_{ijkl} of an individual of gender i, in age group j, in ED k, being ill in month l, $i = 1, 2, ; j = 1, .., 3; k = 1, .., 450, l = 1, .., 48$.

Our count data consists of the number of people who are ill in a given month, as a proportion of the total number at risk. Perhaps the most natural model for such data is the Binomial distribution, with the observed counts restricted by a 'Binomial Denominator'. We here use a logistic Beta-Binomial assumption which can allow for potential 'over-dispersion' in our counts.

Thus we assume we have observed numbers Y_{ijkl} of emergency respiratory admissions of gender i, age j, in ED k, in month l, $i = 1, 2; j = 1, 2, 3; k = 1, .., 450; l = 1, .., 48$. The number of people at risk in each 'cell' is $n_{i,j,k,l}$. We assume that $Y_{i,j,k,l}$ is distributed as a Beta Binomial distribution, $\mathbf{BB}(n_{i,j,k,l};p_{i,j,k,l};\sigma_{i,j,k,l})$. Our basic assumption is that we have a logit link, i.e. $p_{ijlk} = 1/(1 + exp(-\eta_{ijkl}))$, where η_{ijkl} is a linear predictor based upon the explanatory variables in **Table 1**.

5.1 Distribution assumption

In defining the probability function, we drop the suffices i, j, k, l for clarity of exposition. The probability function of a random variable, Y which follows the Beta Binomial distribution denoted here as $\mathbf{BB}(n,p,\sigma)$, is given by

$$p_Y(y|p,\sigma) = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(\frac{1}{\sigma})\Gamma(y+\frac{p}{\sigma})\Gamma[n+\frac{(1-p)}{\sigma}-y]}{\Gamma(n+\frac{1}{\sigma})\Gamma(\frac{p}{\sigma})\Gamma(\frac{1-p}{\sigma})}$$

for $y = 0, 1, 2, \ldots, n$, where $0 < p < 1$ and $\sigma > 0$ (and n is a known positive integer). Note that $E(Y) = np$ and $Var(Y) = np(1-p)\left[1 + \frac{\sigma}{1+\sigma}(n-1)\right]$.

For our modelling we have a r.v. Y_{ijkl}, where we model p_{ijlk} and σ_{ijkl} in terms of our explanatory variables and factors. We assume that the duration d_{ijkl} of observed stays of patients for cell i, j, k, l are such that $d_{ijkl} \sim D(\psi_{ijkl}, \lambda_{ijkl}, \gamma_{i,j,k,l}, \tau_{i,j,k,l})$ where D is one of the many 4-parameter distributions available in GAMLSS. (We here restrict ourselves to distributions with a closed form for the mean and variance, as this is more convenient for our derivation of the expectation and variance of the cost to the NHS of fuel poverty). Our default approach is to assume a log link, i.e. $E(d_{ijkl}) = \psi_{i,j,k,l} = exp(\zeta_{ijkl})$, where ζ_{ijkl} is a linear predictor based on the explanatory variates in **Table 1**.

5.2 Model selection strategy

We illustrate our selection strategy for the 2-parameter Beta Binomial. (We extended this naturally for distributions with more parameters). We initially used the step Akaike criterion to select a model for $\mu_{i,j,k,l}$, keeping σ_{ijkl} constant. We then used a step Akaike approach to fit $\sigma_{i,j,k,l}$, for the current 'best' linear predictor for $\mu_{i,j,k,l}$ (with any remaining parameters constant for the more general case). Using the current 'best' linear predictors for σ_{ijkl}, the model for μ_{ijkl} was refitted, and so on. We finally removed the terms whose removal was not significant on a χ^2 scale. We combined levels of factors where this did not result in significant deterioration in scaled deviance.

6 Results

From the 1991 Census, there were about 25,000 people in Newham over 64 years old. The total count of emergency respiratory episodes amongst this age group was 3378 (over 4 years), 16% of which ended in death. Respiratory episodes far outnumber those for other possible cold-related diagnoses.

We fitted Beta Binomial models to explain morbidity counts in terms of the wide range of explanatory variables, removing any that were not statistically significant. We attempted to avoid a so-called ecological fallacy by using a wide range of explanatory variables. Investigation of the monthly data for 450 EDs determined that winter was better defined as November - February, rather than the traditional UK use of December - March.

The accompanying **Table 2** shows our 'best' model, using a logit link, for the probability of being an emergency respiratory hospital admission, and a log link for the σ coefficient. The σ coefficient (a random effect) depends only upon the age of the people and their gender. (The σ coefficient is always negative; it is larger for the over 84 year olds than for the other over 64 year olds, and is larger for men than women). The linear predictor for p_{ijkl} has an interaction between 'season' and FPR, showing that morbidity counts

Variable	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	-7.1510819	2.423e-01	-29.509	1.534e-190
age2	0.6935349	3.977e-02	17.441	5.137e-68
age3	1.7976923	5.006e-02	35.912	1.341e-280
sex2	-0.7128484	3.583e-02	-19.893	7.004e-88
q2	0.1876991	5.790e-02	3.242	1.188e-03
q3	0.4630773	8.975e-02	5.160	2.477e-07
mmaxt[z]	-0.0124803	5.163e-03	-2.417	1.564e-02
hh1[E]	0.0037488	2.375e-03	1.579	1.144e-01
hh2[E]	0.0076334	2.204e-03	3.463	5.336e-04
lowsap[E]	-0.0017245	9.696e-04	-1.779	7.531e-02
ctb[E]	0.0082016	1.166e-03	7.032	2.049e-12
ch[E]	0.0276699	2.963e-03	9.340	9.871e-21
pens[E]	-0.0018379	8.927e-04	-2.059	3.950e-02
fpr[E]	-0.0000794	4.326e-05	-1.835	6.644e-02
penswh[E]	0.0039622	2.133e-03	1.857	6.329e-02
fpr[E]:nq2	0.0001710	7.011e-05	2.439	1.472e-02

Table 2. Best fitting BB model. Logit link for p, using log link for σ. age2 represents age level 2, etc., mmaxtt[z] denotes mmaxtt indexed over months z, hh1[E] denotes hh1 indexed over enumeration districts E, etc.

rise with increasing fuel poverty risk index in 'winter', with a notably large effect in December. This is over and above the underlying effect of winter itself, irrespective of FPR. Effects are evident for age, with higher counts for older people, and sex, with lower counts for women. There was a strong month effect. To understand this further, we considered monthly weather-related factors. Of all these, maximum temperature was most significant, with a higher maximum leading to lower morbidity counts. Having allowed for the maximum temperature effect, other weather related variables were not significant.

Assuming the durations d_{ijkl} follows a 2-parameter Gamma distributions with $E(d_{i,j,k,l}) = \psi_{i,j,k,l}$ and $Var(d_{i,j,k,l}) = \lambda^2_{i,j,k,l}\psi^2_{i,j,k,l}$, our 'best' model for $\psi_{i,j,k,l}$ is shown in **Table 3**. It can be seen that older people stay in hospital longer, especially older women.

7 Conclusion

We model the propensity to be ill by the Beta Binomial distribution. We model length of stay in hospital, to provide a model for the cost of excess winter morbidity attributable to fuel poverty. Our approach is similar to that used in modelling the probability and cost of insurance claims. The GAMLSS software enables us not only to use the Beta Binomial, but also to use a wide range of continuous distributions to model the length of time that a patient stays in hospital.

Variable	Estimate	Std. Error	t value	Pr(> ||t||)
(Intercept)	1.98035	0.04407	44.9382	0.000e+00
age2	0.31509	0.05924	5.3186	1.114e-07
age3	0.51523	0.07568	6.8077	1.169e-11
sex2	-0.06404	0.06718	-0.9533	3.405e-01
age2:sex2	0.25146	0.08865	2.8367	4.585e-03
age3:sex2	0.31298	0.10497	2.9815	2.888e-03

Table 3. Best fitting Gamma model for d, using a log link for ψ and λ, with λ constant. age2 represents age level 2, etc.

References

BARDSLEY M. (2000): Healthier homes: the role of health authorities. In: Rudge J and Nicol F (Eds.): *Cutting the Cost of Cold. Affordable Warmth for Healthier Homes*. E&FN Spon Ltd., London.

DEFRA (2006): The UK Fuel Poverty Strategy 4th Annual Progress Report 2006. At: *http://www.dti.gov.uk/files/file29688.pdf* [Accessed 7/9/06].

DEPARTMENT OF HEALTH (2001): National Service Framework for Older People. At: *http://www.doh.gov.uk/nsf/olderpeople/pdfs/nsfolderpeople.pdf* [Accessed 12/3/03].

HELLER, G., STASINOPOULOS, D.M., RIGBY, R.A. AND DE JONG, P. (2007): Mean and dispersion modelling for policy claim costs. *Scandinavian Actuarial Journal, 1-12*.

RIGBY, R. AND STASINOPOULOS D.M. (2005): Generalized additive models for location, scale and shape (with Discussion). *Applied Statistics, 54, 507-554*.

RUDGE, J. AND GILCHRIST, R. (2005): Excess winter morbidity among older people at risk of cold homes. *Journal of Public Health 27 (4), 353-358*.

RUDGE, J. AND GILCHRIST, R. (2007): Measuring the health impact of temperatures in dwellings: investigating excess winter morbidity and cold homes in the London Borough of Newham. *Energy and Buildings 39, 847-858*.

WILKINSON, P., LANDON, M., ARMSTRONG, B. et al (2001): *Cold Comfort: the social and environmental determinants of excess winter death in England, 1986-96*. B The Policy Press, Bristol.

Part X

Functional Data Analysis

A Functional Data Approach for Discrimination of Times Series

Andrés M. Alonso[1], David Casado[1], Sara López-Pintado[2], and Juan Romo[1]

[1] Universidad Carlos III de Madrid
28903 Getafe (Madrid), Spain
[2] Universidad Pablo de Olavide
41013 Sevilla, Spain

Abstract. A new classification method for time series is proposed. A series is assigned to a class after comparing distances between its integrated periodogram and the mean of the integrated periodograms in each group. The approach can be used with nonstationary time series by computing the periodogram locally. Depth based techniques are used to make the classification robust. The method provides small error rates both with simulated and real data; it also shows good computational behavior.

Keywords: times series, functional data, discrimination, robustness

1 Introduction

Classifying time series is an important task in geology or medicine, among other areas. Previous work has considered both the time and frequency domains to discriminate and classify time series. In particular, methods used by Kakizawa et al. (1998), which depend on spectral density estimates over specific frequency bands, are applicable to stationary processes. Shumway (2003) proposed methods based on time varying spectra that are also applicable to the class of locally stationary processes introduced by Dahlhaus (1997). Sakiyama and Tanigushi (2004) also studied discriminant analysis for locally stationary processes, but their approach required the specification of a parametric model for the time-varying spectral densities. Huang et al. (2004) used methods based on Fourier-type bases using local spectral features of the time series. Recently, Maharaj and Alonso (2007) have proposed classifying locally stationary processes by using discriminant analysis with wavelet variances as input.

We propose a frequency domain technique based on the integrated periodogram. We assign a new time series by considering the distance between its integrated periodogram and the mean of integrated periodograms in each group. Since the integrated periodograms are functional data, we apply depth-based techniques to make the classification robust. The notion of statistical depth has been extended to functional data, and López-Pintado and Romo (2006) have used this concept to classify curves. Since robustness

is an interesting feature of the statistical methods based on depth, we have applied the ideas of López-Pintado and Romo (2006) to add robustness to our time series classification procedure. Their method considers the α-trimmed mean as a reference curve of each group, which is defined as the average of the $1-\alpha$ proportion of the deepest curves of the sample; that is, it leaves $100\alpha\%$ of data out. This trimming is the responsible for adding robustness.

2 Classification method

One of the main points of our classification proposal is that we turn the time series problem into a functional data problem by considering the integrated periodogram of each time series. The *periodogram* $I(\omega)$ is the sample version of the *spectral density* and expresses the contribution of Fourier frequencies to the series total variance. Its cumulative version is the *integrated periodogram* $F_Z(\omega_k) = \sum_{i=1}^{m} I(\omega_i)$. Though the periodogram is properly defined only for stationary series, we shall consider the series are approximately locally stationary in order to classify nonstationary time series. We shall split them into k blocks and compute the integrated periodogram of each block.

When functions—instead of time series—need to be classified, a natural criterion is to assign them to the class minimizing some distance (we have taken L_1) from the new function to the group. As a reference function of each group we take the mean of its elements; since the mean is not robust to the presence of outliers, robustness can be added to the process by considering the α-trimmed mean instead. Both means can be expressed as follows. Let $\Psi_{g_i}(\omega)$; $i = 1, \ldots, N$ be functions of the population g, and let $\Psi_{g_{(i)}}(\omega)$; $i = 1, \ldots, N$ be the same functions ordered by decreasing depth; then the α–trimmed mean is:

$$\bar{\Psi}_g^{\alpha} = \frac{1}{N - [N\alpha]} \sum\nolimits_{i=1}^{N-[N\alpha]} \Psi_{g_{(i)}}(\omega), \tag{1}$$

where $[\cdot]$ is the integer part function. When $\alpha = 0$, the whole sample is taken, while if $\alpha > 0$ the $100\alpha\%$ of the less depth data are left out.

The statistical concept of *depth* is a measurement of the "centrality" of each element inside the sample. Several different definitions of depth for functions can be given. We use the definition of functional generalized band depth proposed by López-Pintado and Romo (2006). Let $G(\Psi) = \{(t, \Psi(t)) : t \in [a, b]\}$ denote the graph in $\mathbb{R}^2$ of a function $\Psi \in C[a, b]$ and let $\Psi_i(t)$; $i = 1, \ldots, m$ be functions in $C[a, b]$; then a subset of this functions, $\Psi_{i_j}(t)$; $j = 1, \ldots, n$, determines a band:

$$V(\Psi_{i_1}, \ldots, \Psi_{i_n}) = \{(t, y) : t \in [a, b], \min_{r=1,\ldots,n} \Psi_{i_r}(t) \leq y \leq \max_{r=1,\ldots,n} \Psi_{i_r}(t)\}.$$

Let $A_j(\Psi) \equiv \{t \in [a, b] : \min_{r=i_1,\ldots,i_j} \Psi_r(t) \leq \Psi(t) \leq \max_{r=i_1,\ldots,i_j} \Psi_r(t)\}$ be the set of points in the interval $[a, b]$ where the function Ψ is inside the band;

if λ is the Lebesgue measure on the interval $[a, b]$, $\lambda(A_j(\Psi))$ is the "proportion of time" that Ψ is inside the band, the following quantity is defined

$$GS_m^{(j}(\Psi) = \binom{m}{j}^{-1} \lambda([a,b])^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_j \le m} \lambda(A(\Psi; \Psi_{i_1}, \ldots, \Psi_{i_j})), \qquad j \ge 2.$$

Finally, the *generalized band depth* of any of the curves Ψ in $\Psi_i(t)$; $i = 1, \ldots,$ m is

$$GS_{m,J}(\Psi) = \sum_{j=2}^{J} GS_m^{(j}(\Psi), \qquad J \ge 2. \tag{2}$$

The algorithm: Let $\{X_1, \ldots, X_{N_x}\}$ be a sample containing time series from the population P_X, and let $\{Y_1, \ldots, Y_{N_y}\}$ be a sample from P_Y. The classification method follows the next steps:

(i) For each time series in the samples, the integrated periodogram of the k blocks is obtained, i.e., we have $\{\Psi_{X_1}, \ldots, \Psi_{X_{N_x}}\}$ and $\{\Psi_{Y_1}, \ldots, \Psi_{Y_{N_y}}\}$, where $\Psi_{X_i} = (F_{X_i}^{(1)}, \ldots, F_{X_i}^{(k)})$, $\Psi_{Y_i} = (F_{Y_i}^{(1)}, \ldots, F_{Y_i}^{(k)})$ and $F_{X_i}^{(j)}$ is the integrated periodogram of the j–th block of the i–th series of the population X; $F_{Y_i}^{(j)}$ is the analogous function for the population Y.
(ii) For both P_X and P_Y samples the α–trimmed class mean is computed: $\bar{\Psi}_X^{\alpha}$ and $\bar{\Psi}_Y^{\alpha}$.
(iii) Let Ψ_Z be the curve associated to a new series Z, $\Psi_Z = (F_Z^{(1)}, \ldots, F_Z^{(k)})$; then Z is classified in the group P_X if $d(\Psi_Z, \bar{\Psi}_X^{\alpha}) < d(\Psi_Z, \bar{\Psi}_Y^{\alpha})$, and in the group P_Y otherwise.

The extension to more than two groups is straightforward. There are two opposite effects as a consequence of splitting the series into blocks: one is that the narrower blocks are, the closer to the locally stationarity assumption we are; the other is that when the length of blocks decreases, also decreases the quality of the integrated periodogram as estimator. Then it can be expected that errors, as functions of k, reach a minimum. This effect can be observed from the simulation exercises. In general, the number of blocks, k, can be selected by cross-validation. To apply the algorithm to stationary series the proper value is $k = 1$. Let us remind that the SLEXbC method splits implicitly the series into blocks.

3 Simulation results

We evaluate our algorithms for $\alpha = 0$ (DbC; letters *–bC* came from *based classification*) and $\alpha = 0.2$ (DbC-α). Also, we take the method proposed in Huang et al. (2004) as a reference (SLEXbC). We have analyzed the three experimental settings considered by these authors. For each comparison between two models we have run 1000 times the training-testing processes. The

three algorithms are called with exactly the same data sets. We present some results of the first setting.

Simulation exercise 1: We compare Gaussian white noise with an autoregressive process of order 1. Each training data set has $N_x = N_y = 8$ series of length $T_x = T_y = 1024$. Six comparisons have been run, with the parameter ϕ of the AR(1) model taking the values -0.5, -0.3, -0.1, $+0.1$, $+0.3$ and $+0.5$. Series are stationary in this exercise.

$$\begin{aligned} X_t^{(i)} &= \phi \cdot X_{t-1}^{(i)} + \epsilon_t^{(i)} \quad t = 1, \ldots, T_x \text{ and } i = 1, \ldots, N_x \\ Y_t^{(j)} &= \epsilon_t^{(j)} \qquad\qquad\qquad t = 1, \ldots, T_y \text{ and } j = 1, \ldots, N_y, \end{aligned} \tag{3}$$

where $\epsilon_t^{(i)}$ y $\epsilon_t^{(j)}$ are i.i.d. $N(0, 1)$.

Simulation exercise 2: We compare two processes composed half by white noise and half by an autoregressive process of order 1. Different combinations of training sample sizes—$N_x = N_y = 8$ and 16—and series lengths—$T_x = T_y = 512$, 1024 and 2048—are considered. In this exercise series are composed of stationary parts, but series themselves are not.

$$\begin{aligned} X_t^{(i)} &= \begin{cases} \epsilon_t^{(i)} & \text{if } t = 1, \ldots, T_x/2 \\ X_t^{(i)} = -0.1 \cdot X_{t-1}^{(i)} + \epsilon_t^{(i)} & \text{if } t = T_x/2 + 1, \ldots, T_x \end{cases} \\ Y_t^{(j)} &= \begin{cases} \epsilon_t^{(j)} & \text{if } t = 1, \ldots, T_y/2 \\ Y_t^{(j)} = +0.1 \cdot Y_{t-1}^{(j)} + \epsilon_t^{(j)} & \text{if } t = T_y/2 + 1, \ldots, T_y \end{cases} \end{aligned} \tag{4}$$

with $i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$.

Simulation exercise 3: In this exercise the stochastic models of both classes are slowly time-varying autoregressive processes of order 2. This means that some coefficient of the autoregressive structure is not fixed but it varies in time. Each training data set has $N_x = N_y = 10$ series of length $T_x = T_y = 1024$. Three comparisons have been done, the first class having always the parameter $\tau = 0.5$, and the second class having respectively the values $\tau = 0.4$, 0.3 and 0.2. In this simulation the processes involved are not stationary. If $a_{t;\tau} = 0.8 \cdot [1 - \tau \cos(\pi t/1024)]$, then

$$\begin{aligned} X_t^{(i)} &= a_{t;0.5} \cdot X_{t-1}^{(i)} - 0.81 \cdot X_{t-2}^{(i)} + \epsilon_t^{(i)} \quad t = 1, \ldots, T_x \\ Y_t^{(j)} &= a_{t;\tau} \cdot Y_{t-1}^{(j)} - 0.81 \cdot Y_{t-2}^{(j)} + \epsilon_t^{(j)} \qquad t = 1, \ldots, T_y \end{aligned} \tag{5}$$

$i = 1, \ldots, N_x$ and $j = 1, \ldots, N_y$ again.

Additionally to Huang et al. (2004) setup, we consider contamination in order to evaluate the robustness of the three considered processes: DbC, DbC-α and SLEXbC.

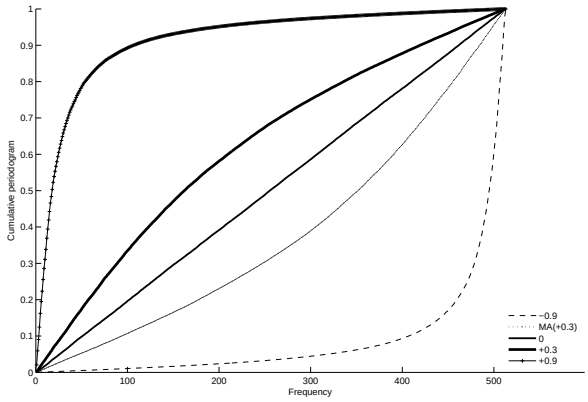


Fig. 1. Integrated periodograms for model $\phi = +0.3$ and its contaminations.

Contamination 1: For exercise 1 it consists in substituting, in one of the series of the AR(1) class, the autoregressive structure for moving average structure; that is, generating a MA(1) model—with the MA parameter equal to the AR parameter—instead of a AR(1) model (see Figure 1). For exercise 2 it consists in doing the same substitution of structures, but only in the autoregressive half of one series of a class (the other half is white noise). For exercise 3 we contaminate the set of slowly time-varying autoregressives of parameter +0.5 with a series of the same model but with parameter value +0.2.

Contamination 2: This contamination consists in using a parameter value of $\phi = -0.9$ in exercises 1 and 2, and $\tau = -0.9$ in exercise 3, instead of the correct value, in one of the series of the training set of the group X. That is, using always the correct model, but mistaking the parameter value in a series.

Contamination 3: Equal to the contamination 2 but using a value +0.9, instead of -0.9.

In Table 1 we present some results for setting 1. Similar results are obtained for the three settings (see Alonso et al. (2008)). When contamination is not present, DbC provides slightly better results than DbC–α (since DbC uses the whole sample), and about half of the errors of SLEXbC. The DbC and SLEXbC errors increase slightly with the weak contamination (1) and substantially with the strong ones (2 and 3), while errors do not change for DbC–α, because its trimming keep contamination out; this shows its robustness. From the whole results, it seems that SLEXbC tends to have higher median, higher errors above this median, and less errors near zero. On the

	$\phi =$ **-0.3**	$\phi =$ **-0.1**	$\phi =$ **+0.1**	$\phi =$ **+0.3**
DbC	0.000	0.063	0.060	0.000
DbC-α	0.000	0.065	0.062	0.000
SLEXbC	0.000	0.131	0.127	0.000
Contamination 1				
DbC	0.000	0.077	0.074	0.000
DbC-α	0.000	0.064	0.062	0.000
SLEXbC	0.000	0.175	0.172	0.000
Contamination 2				
DbC	0.000	0.300	0.513	0.001
DbC-α	0.000	0.065	0.062	0.000
SLEXbC	0.001	0.377	0.491	0.002
Contamination 3				
DbC	0.001	0.512	0.300	0.000
DbC-α	0.000	0.064	0.062	0.000
SLEXbC	0.002	0.490	0.377	0.001

Table 1. Empirical misclassification rates (1000 runs) in setting 1 with and without contaminations.

other side, DbC-α is the only method maintaining the same pattern (with and without contamination) and a considerable amount of errors near zero.

	$\phi =$ **-0.3**	$\phi =$ **-0.1**	$\phi =$ **+0.1**	$\phi =$ **+0.3**
DbC	0.027	0.027	0.027	0.027
DbC-α	0.045	0.045	0.044	0.044
SLEXbC	0.678	0.724	0.713	0.670

Table 2. Mean times (in seconds per run) for setting 1

With respect to the computation times, from the simulations we can extract some qualitative conclusions (the quantitative depends more on the implementation, not on the method itself). For the SLEXbC method we have used an implementation provided by the authors. To select the parameters for this method, we have done an small optimization for each simulation exercise; the results were similar to the values that authors recommended us.

Since chronometer is called after generating series, it can be expected the computation times not to depend on the parameters of the stochastic processes. This is what we observe for our algorithms, but not for the SLEXbC method. Perhaps this is because this method need to select a basis of the SLEX library explaining best the differences between series, while our method works only with graphs. For our procedure times increase with k.

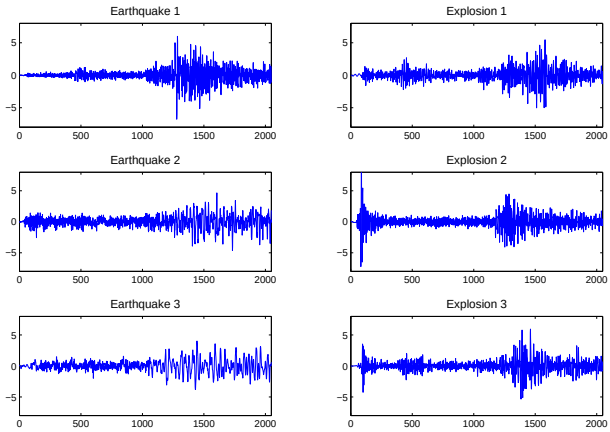


Fig. 2. Three earthquakes and three explosions.

The computation of depth is moderately time-consuming with the sample size and series length involved in exercises 1 and 3. This computation time should increase quite with sample size and a little with series length, since slowness cames from the number of comparisons the evaluation of depths includes. Nevertheless, it is possible to do these comparisons only once by implementing López-Pintado and Romo (2006)'s method conveniently; such an implementation allows using their methods with middle samples sizes. Finally, for our approach computation time depends on sample size but slightly on series length, while the SLEXbC method gets slower when any N or T increases.

4 Real data example

We have evaluated our proposal in a benchmark data set containing 8 explosions, 8 earthquakes and an extra series—known as NZ event—not classified (but being an earthquake or an explosion). Each series is made up of 2048 points in two different parts: the first half is the P wave, and the second is the S wave. For each series we have considered the curve formed by merging the non-normalized integrated periodograms of parts P and S. Considering the 8 earthquakes as group 1 and the 8 explosions as group 2, and applying leave-one-out cross validation, both of our algorithms misclassify only the first series of the group 2. With respect to the NZ event, both algorithms agree on assigning it to the explosions group, as other authors do, for example, Kakizawa et al. (1998) and Huang et al. (2004).

Now we have carried out an additional exercise. Since many methods classify the NZ event as an explosion, we consider an artificial data set con-

structed by the 8 earthquakes plus the NZ event as group 1, and the 8 explosions as group 2. In this situation, the result for DbC is that it misclassifies the first and the third—not only the first—elements of the group 2. But DbC–α again misclassifies only the first series of group 2, even though the NZ event was included in group 1 (earthquakes). This illustrates the robustness of our proposed second algorithm.

Acknowledgments: This research was partially supported by the projects SEJ2005-06454, CCG06-UC3M/ESP-0856 and SEJ2007-64500.

References

ALONSO, A.M., CASADO, D., LÓPEZ–PINTADO, S. and ROMO, J. (2008): A Functional Data Based Method for Time Series Classification. *Preprint.*

DAHLHAUS, R. (1997): Fitting Time Series Models to Nonstationary Processes. *The Annals of Statistics 25 (1), 1-37.*

HUANG,H., OMBAO, H. and STOFFER, D. (2004): Discrimination and Classification of Nonstationary Time Series Using the SLEX Model. *Journal of the American Statistical Association 99 (467), 763–774.*

KAKIZAWA, Y., SHUMWAY, R.H. and TANIGUCHI, M. (1998): Discrimination and Clustering for Multivariate Time Series. *Journal of the American Statistical Association 93 (441), 328–340.*

LÓPEZ-PINTADO, S. and ROMO., J. (2006): Depth-based Classification for Functional Data. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science.* American Mathematical Society 72.

MAHARAJ, E.A. and ALONSO, A.M. (2007): Discrimination of locally stationary time series using wavelets. *Computational Statistics and Data Analysis 52 (2), 879–895.*

SAKIYAMA, K. and TANIGUCHI, M. (2004): Discriminant analysis for locally stationary processes. *Journal of Multivariate Analysis 90 (2), 282-300.*

SHUMWAY, R.H. (2003): Time-frequency clustering and discriminant analysis. *Statistics and Probability Letters 63 (3), 307-314.*

From Quasi-Arithmetic Means to Parametric Families of Probability Distributions for Functional Data

Etienne Cuvelier and Monique Noirhomme-Fraiture

Facultés Universitaires Notre-Dame de la Paix
Faculté d'Informatique
21, rue grandgagnage 5000 Namur, Belgium,
ecu@info.fundp.ac.be, mno@info.fundp.ac.be

Abstract. Probability distributions are central tools for probabilistic modeling in data mining, and they lack in functional data analysis (FDA). In this paper we propose a probability distribution law for functional data. We build it using jointly Quasi-arithmetic means and generators of Archimedean copulas. We also define a density adapted to the infinite dimension of the space of functional data. For this we use the Gâteaux differential. We use these concepts in supervised classification.

Keywords: functional data analysis, probability distributions, Archimedean copulas, quasi-arithmetic means

1 Introduction

Probability distributions and their densities have proved their usefulness in data mining, more particularly for classification tasks. Functional data are also stochastic process, and in this field the probability distributions have been studied largely, but with rather strong hypotheses : Markov process, Wiener process or Brownian motion. In functional data analysis such assumptions are not made on the behaviour of functions, and in this case, few studies have been done. Ferraty and Vieu (2000) and Dabo-Niang (2002) use a non parametric estimations of the density : the naive estimator and the kernel estimator. In this later estimation the kernel function is computed on the semi-norm : $K(||X_i - x||)$. Diday (2003) have proposed the use of archimedean copulas to build a finite approximation for the distribution of functional data in the special case when the functional are univariate cumulative distribution functions. Vrac et al. (2001) have used this technique with an finite approximation in $\mathbb{R}^2$. In Cuvelier and Noirhomme-Fraiture (2005) we have proposed to use the Clayton copula for higher dimensions $\mathbb{R}^n$. In this paper we propose a family of distribution functions directly defined in the $\mathbb{R}^\infty$ space, i.e. without any approximation. We propose also an adapted density function, and thus we are able to compute a distribution and a density for a functional random variable. We illustrate the utility of these tools with a supervised classification application.

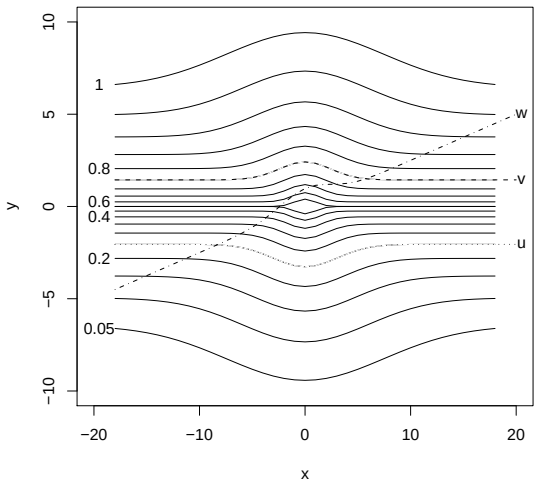


Fig. 1. A simple example of functional data.

2 QAMML distributions

Let $(\Omega, \mathcal{A}, P)$ a probability space and $\mathcal{D}$ a closed real interval. A *functional random variable (frv)* is any function from $\mathcal{D} \times \Omega \to \mathbb{R}$ such for any $t \in \mathcal{D}, X(t,.)$ is a real random variable on $(\Omega, \mathcal{A}, P)$. Let $L^2(\mathcal{D})$ be the space of square integrable functions $u(t)$ defined on $\mathcal{D}$. If $f, g \in L^2(\mathcal{D})$, then the pointwise order between f and g on $\mathcal{D}$ is defined as follows :

$$\forall t \in \mathcal{D}, f(t) \leq g(t) \iff f \leq_{\mathcal{D}} g. \tag{1}$$

It is easy to see that the pointwise order is a partial order over $L^2(\mathcal{D})$, and not a total order. We define the *functional cumulative distribution function (fcdf)* of a *frv* $\underline{X}$ on $L^2(\mathcal{D})$ computed at $u \in L^2(\mathcal{D})$ by :

$$F_{\underline{X},\mathcal{D}}(u) = P[\underline{X} \leq_{\mathcal{D}} u]. \tag{2}$$

Figure 1 shows a simple synthetic dataset with 20 functions (u, v and solid lines). It is simple to empirically estimate the probability distribution for a function of this dataset, but it is less easy for the function w, which is not in this sample. To compute this probability in all cases, let us remark that, it is easy to calculate the probability distribution of the value of $X(t)$ for a specific value of t, and this for any $t \in \mathcal{D}$. Then we define respectively the *surface of distributions* and the *surface of densities* as follow :

$$G : \mathcal{D} \times \mathbb{R} \to [0,1] : (t, y) \mapsto P[X(t) \leq y] \tag{3}$$

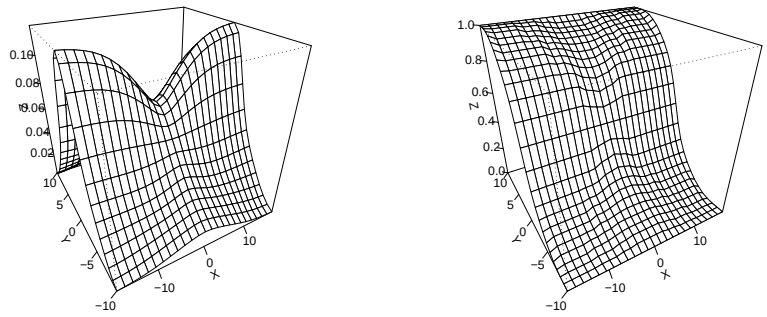


Fig. 2. Surfaces of distributions and densities for the data set shown in fig. 1.

$$g : \mathcal{D} \times \mathbb{R} \to [0,1] : (t,y) \mapsto \frac{\partial}{\partial t} G(t,y). \tag{4}$$

We can use various methods for determining suitable g and G for a chosen value of $\underline{X}$. Thus for example, if $\underline{X}$ is a Gaussian process with mean value $\mu(t)$ and standard deviation $\sigma(t)$, then, for any $(t,y) \in \mathcal{D} \times \mathbb{R}$, we have :

$$G(t,y) = F_{\mathcal{N}(\mu(t),\sigma(t))}(y) \tag{5}$$

$$g(t,y) = f_{\mathcal{N}(\mu(t),\sigma(t))}(y). \tag{6}$$

In other cases, if we have a sample of realizations $\{x_1, \ldots, x_N\}$, then we can use the empirical cumulative distribution function and the kernel density estimation to estimate $\hat{G}$ and $\hat{g}$:

$$\hat{G}(t,y) = \frac{Card\{x_i(t) \leq y\}}{N} \tag{7}$$

$$\hat{g}(t,y) = \frac{1}{N \cdot h(t)} \sum_{i=1}^{N} K\left(\frac{y - x_i(t)}{h(t)}\right) \tag{8}$$

where $Card$ is the cardinal of the set and $h(t)$ the smoothing parameter. Figure 2 shows the surfaces of distributions and densities for the data set shown in the figure 1.

In the following we will always use the function G with a function u of $L^2(\mathcal{D})$, so, for the ease of the notations, we will write : $G[t;u] = G[t,u(t)]$. We will use the same notation for g. In what follows we define our parametric families of probability distributions.

Let $\underline{X}$ be a frv, $u \in L^2(\mathcal{D})$ and G its *Surface of Distributions*. Let also ϕ be a continuous strictly decreasing function from $[0,1]$ to $[0,\infty]$ such that $\phi(0) = \infty$, $\phi(1) = 0$, where $\psi = \phi^{-1}$ must be completely monotonic on $[0,\infty[$ i.e. $(-1)^k \frac{d^k}{dt^k}\psi(t) \geq 0$ for all t in $[0,\infty[$ and for all k. We define the *Quasi-Arithmetic Mean of Margins Limit (QAMML)* distribution of $\underline{X}$ by :

$$F_{\underline{X},\mathcal{D}}(u) = \psi\left[\frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \phi\left(G\left[t;u\right]\right)dt\right]. \tag{9}$$

In fact the expression (9) can be seen as the limiting (or continuous) case of two other expressions. The first expression, which is obvious and gives its name to (9), use a quasi-arithmetic mean M :

$$F_{\underline{X},\mathcal{D}}(u) = \lim_{n\to\infty} M\left\{G[t_1;u],\ldots,G[t_n;u]\right\} \tag{10}$$

where $\{t_1,\ldots,t_n\} \subset \mathcal{D}$ is a subset of points in $\mathcal{D}$, preferably equidistant. In the discrete case, a quasi-arithmetic mean is a function $M : [a,b]^n \to [a,b]$ defined as follows:

$$M(x_1,\ldots,x_n) = \psi\left(\frac{1}{n}\sum_{i=1}^{n}\phi\left(x_i\right)\right) \tag{11}$$

where ϕ is a continuous strictly monotonic real function and $\psi = \phi^{-1}$.
The second limiting case links the *QAMML* distributions to the classical approximation : $P\left[\underline{X} \leq_{\mathcal{D}} u\right] = H\left(u(t_1),\ldots,u(t_n)\right)$, using the archimedean copulas:

$$F_{\underline{X},\mathcal{D}}(u) = \lim_{n\to\infty} \psi\left[\sum_{i=1}^{n}\phi\left(G^*\left[t_i;u\right]\right)\right] \tag{12}$$

where * is the following transformation, applied to margins:

$$G^*(x) = \psi\left(\frac{1}{n}\phi\left(G(x)\right)\right). \tag{13}$$

Let us remind that a copula is a multivariate cumulative distribution function defined on the n-dimensional unit cube $[0,1]^n$ such that every marginal distribution is uniform on the interval [0, 1]. The interest of copulas comes from the fact that (Sklar's theorem), if H is an n-dimensional distribution function with margins $F_1, ..., F_n$, then there exists an n-copula C such that for all $x \in \mathbb{R}^n$,

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)). \tag{14}$$

The copula captures the dependence structure of the distribution. An important family of copulas is the family of Archimedean copula, given by the following expression :

$$C(u_1, ..., u_n) = \psi\left[\sum_{i=1}^{n}\phi(u_i)\right] \tag{15}$$

where ϕ, called the generator, has the same properties that a *QAMML* generator.
This second limiting case shows that *QAMML* shares the properties and limitations of archimedeans copulas in the modeling of an *frv* $\underline{X}$ (see the GQAMML section).

3 Gateaux density

A *fcdf* is an incomplete tool without an associate density, but as the *QAMML* distributions deal directly with infinite nature of functional data, we cannot use the classical multivariate density function:

$$h(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} H(x_1, \dots, x_n). \tag{16}$$

To solve this problem we propose to use a concept of the functional analysis : the *Gâteaux differential* which is a generalization of directional derivative. Let $\underline{X}$ be a *frv*, $F_{\underline{X},\mathcal{D}}$ its *fcdf* and u a function of $L^2(\mathcal{D})$. Then for $h \in L^2(\mathcal{D})$ we define the *Gâteaux density of* $F_{\underline{X},\mathcal{D}}$ at u and in the direction of h by:

$$f_{\underline{X},\mathcal{D},h}(u) = \lim_{\epsilon \to 0} \frac{F_{\underline{X},\mathcal{D}}\left(u + h \cdot \epsilon\right) - F_{\underline{X},\mathcal{D}}\left(u\right)}{\epsilon} = DF_{\underline{X},\mathcal{D}}(u; h) \tag{17}$$

where $DF_{\underline{X},\mathcal{D}}(u; h)$ is the *Gâteaux differential* of $F_{\underline{X},\mathcal{D}}$ at u in the direction $h \in V$.
It is easy to show that, if $F_{\underline{X},\mathcal{D}}$ is a *QAMML fcdf*, u and h are two functions of $L^2(\mathcal{D})$, then the corresponding *Gâteaux density of* $F_{\underline{X},\mathcal{D}}$ computed in u, in direction of h is given by:

$$f_{\underline{X},\mathcal{D},h}(u) = \frac{1}{|\mathcal{D}|} \cdot \psi' \left[\frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \phi\left(G\left[t; u\right]\right)\ dt \right] \cdot \left\{ \int_{\mathcal{D}} \phi'\left(G\left[t; u\right]\right) \cdot g\left[t; u\right] \cdot h(t)\ dt \right\}. \tag{18}$$

We can show that, if we use the statistical dispersion $\sigma(t)$ of the functional data, then $f_{\underline{X},\mathcal{D},\sigma}(u) = P[\underline{X} = u]$.

4 GQAMML distributions

QAMML shares the limitations of archimedeans copulas (see section 1), but the archimedean copulas of dimension $n > 2$, can capture dependence structures from independence until the complete positive dependence between variables. Thus, if for $s, t \in \mathcal{D}$, there is a negative dependence between $X(s)$ and $X(t)$, the *QAMML* will not be able to model the situation. But the bidimensional archimedean copulas can deal with this kind of dependence, using the same generator, but with larger domain for the parameter. Then we define the *Generalized Quasi-Arithmetic Mean of Margins Limit (GQAMML)* $\mathbb{F}_{\underline{X},\mathcal{D}}(u)$ as follows. Let $\underline{X}$ be a *frv* defined on $\mathcal{D}$, $u \in L^2(\mathcal{D})$, $\{\mathcal{D}_p, \mathcal{D}_n\}$ a partition of $\mathcal{D}$ such :

- $\forall s, t \in \mathcal{D}_p$, there is a positive dependence between $X(s)$ and $X(t)$,
- $\forall s, t \in \mathcal{D}_n$, there is a positive dependence between $X(s)$ and $X(t)$,
- $\forall s \in \mathcal{D}_p$ and $\forall t \in \mathcal{D}_n$, there is a negative dependence between $X(s)$ and $X(t)$.

Then

$$\mathbb{F}_{\underline{X},\mathcal{D}}(u) = \psi\left(\frac{|\mathcal{D}_p|}{|\mathcal{D}|}\phi\left[F_{\underline{X},\mathcal{D}_p}(u)\right] + \frac{|\mathcal{D}_n|}{|\mathcal{D}|}\phi\left[F_{\underline{X},\mathcal{D}_n}(u)\right]\right) \tag{19}$$

where ϕ is the generator of an bidimensional archimedean copulas.
Of course, using the chain rule, the *Gâteaux density of* $\mathbb{F}_{\underline{X},\mathcal{D}}$ is given by

$$\mathbf{f}_{\underline{X},\mathcal{D},\sigma}(u) = \psi'\left(\frac{|\mathcal{D}_p|}{|\mathcal{D}|}\phi\left[F_{\underline{X},\mathcal{D}_p}(u)\right] + \frac{|\mathcal{D}_n|}{|\mathcal{D}|}\phi\left[F_{\underline{X},\mathcal{D}_n}(u)\right]\right)$$
$$\left\{\frac{|\mathcal{D}_p|}{|\mathcal{D}|}\phi'\left[F_{\underline{X},\mathcal{D}_p}(u)\right] f_{\underline{X},\mathcal{D}_p,\sigma}(u) + \frac{|\mathcal{D}_n|}{|\mathcal{D}|}\phi'\left[F_{\underline{X},\mathcal{D}_n}(u)\right] f_{\underline{X},\mathcal{D}_n,\sigma}(u)\right\} . \tag{20}$$

5 CQAMML distributions

In functional data analysis, we know that, some times, when we treat smooth data, there is a lot of information in the derivatives of the data. Of course we can apply the *GQAMML* distributions to the concerned derivative, but we can also consider jointly the distribution of the different derivatives. Then we define the *Complete Quasi-Arithmetic Mean of Margins Limit (CQAMML)* ${\mathbb{F}^j}_{i\,\underline{X},\mathcal{D}}(u)$ (with $i < j$) as follows. Let $\underline{X}$ be a *frv* defined on $\mathcal{D}$ with j successive derivatives, $u \in L^2(\mathcal{D})$ with j successive derivatives:

$${\mathbb{F}^j}_{i\,\underline{X},\mathcal{D}}(u) = C\left(\mathbb{F}_{\underline{X}^{[i]},\mathcal{D}}\left(u^{[i]}\right), \ldots, \mathbb{F}_{\underline{X}^{[j]},\mathcal{D}}\left(u^{[j]}\right)\right) \tag{21}$$

where :

- $\underline{X}^{[i]}$ and $u^{[i]}$ are the ith derivatives for $\underline{X}$ and u,
- C is a n-dimensional copula.

Note that the copula C is not necessarily an archimedean copula. The density of the *CQAMML* distribution is a classical joint density used with the *Gâteaux densities* of the different *GQAMML* distributions.

6 Supervised classification

To illustrate the interest of the *QAMML* families of distribution we propose to use it in a supervised classification application. To perform the classification we use the *Gâteaux density of a QAMML distribution* to build a bayesian classifier:

$$P(\omega_i|u) = \frac{\mathbf{f}_{\omega_i,\mathcal{D},h}(u) \cdot P(\omega_i)}{P(u)} \tag{22}$$

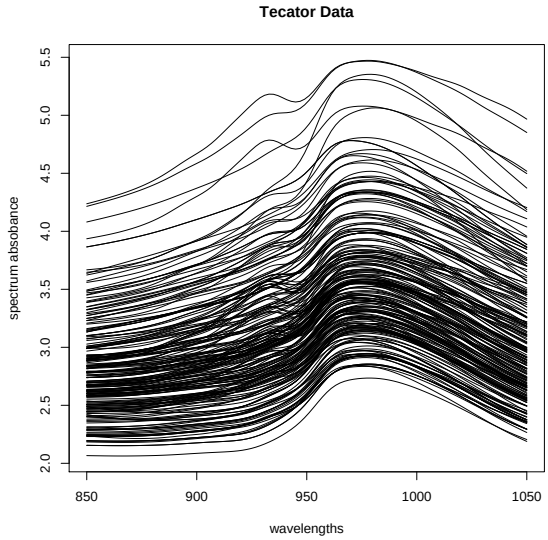


Fig. 3. The Tecator data set.

where $P(\omega_i|u)$ is the probability that u belong to the ith group, $\mathbf{f}_{\omega_i,\mathcal{D},h}(u)$ the adequate *Gâteaux density*, and $P(u)$ the probability of u (but this latter is constant for all cluster, so it is not necessary to compute it). We compute the parameters of each cluster using the classical maximum likelihood, and the cluster of u is the cluster with the highest probability $P(\omega|u)$. The chosen dataset is the well known spectrometric data from Tecator. The data consist in 100 channels of spectrum absorbance (wavelength from 850 nm to 1050 nm). The goal is to distinguish the data with more than 20% of fat content, from the data with less than 20% of fat content. We have performed a 10-fold cross validation on the data, the first derivative, the second derivative using the GQAMML distributions, and jointly on the different derivatives using the CQAMLL distributions, and this with the following parametrization :

- Surface of distributions G : Normal distribution,
- QAMML and GQAMML generators : Clayton generator,
- CQAMML copula : Normal copula.

The table 1 shows the results, and we can see that the best results are given using the distribution of the second derivative, and also considering jointly the distribution of the first and second derivative, but it is well known that the second derivative of these data contains the more interesting information to distinguish the clusters. We can also remark that when we use directly the functionnal data jointly with the derivatives, the quality of the classification decrease, but we know that original functions contain only slight differences between the two groups.

Table 1. Results of the 10-fold cross validations.

Distributions	**misclassifications**
$F_{\underline{X},\mathcal{D}}$	31.4%
$F_{\underline{X}',\mathcal{D}}$	9.4%
$F_{\underline{X}'',\mathcal{D}}$	5.5%
$\mathbb{F}^1_{0\underline{X},\mathcal{D}}$	16.5%
$\mathbb{F}^2_{1\underline{X},\mathcal{D}}$	4%
$\mathbb{F}^2_{0\underline{X},\mathcal{D}}$	9.4%

7 Conclusions

In this paper we have proposed a new probabilistic tool for functional data and shown that when we use it with an existing algorithm of supervised classification we can obtain good results. The probability distributions are very important tools in multivariate data analysis, and we think that the QAMML family of probability distributions can be used in conjunction with many other existing techniques to extend these techniques to functional data.

References

ACZEL, J. (1966): *Lectures on Functional Equations and Their Applications*. Academic Press, Mathematics in Science and Engineering, New York and London.

CUVELIER, E. and NOIRHOMME-FRAITURE, M. (2005): M. Clayton copula and mixture decomposition. In: *ASMDA 2005*. Brest, 699-708.

CUVELIER, E. and NOIRHOMME-FRAITURE, M. (2007): Classification de fonctions continues l'aide d'une distribution et d'une définies dans un espace de dimension infinie. In: M. Noirhomme-Fraiture and G. Venturini (Eds.) : *RNTI : Extraction et Gestion des Connaissances 2007*. Cépaduès, Paris, 679–690.

DABO-NIANG, S. (2002): Estimation de la densité dans un espace de dimension infinie : Application aux diffusions. *C. R. Acad. Sci. Paris Ser., 334, 213-216*.

DIDAY, E. (2003): Mixture decomposition of distributions by copulas Classification In : *Clustering and Data Analysis*. Springer, Verlag, 297-310.

FERRATY, F. and VIEU, P. (2000): Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Comptes Rendus de l'Académie des Sciences de Paris, Série I, 2000, 330, 139-142*.

JOE, H. (1997): *Multivariate models and dependence concepts*. Chapman and Hall, London.

KOLMOGOROV, A. (1930): Sur la notion de moyenne. *Rendiconti Accademia dei Lincei, 12 (6), 388-391*.

LUSTERNIK, L.A. and SOBOLEV, V.J. (1974): *Elements of Functional Analysis* Hindustan Publishing Corpn., Delhi.

NELSEN, R.B. (1999): *An introduction to copulas*. Springer, London.

VRAC, M., DIDAY, E., CHEDIN, A. and NAVEAU, P. (2001): Mélange de distributions de distributions, décomposition de mélange de copules et application la climatologie. In : *Actes du VIIIème congrés de la Société Francophone de Classification*. Guadeloupe, 348-355.

Pyramidisation Procedure for a Hierarchy of Time Series Based on the Kullback Leibler Divergence

Mireille Gettler Summa[(1),(2)] and Kutluhan Kemal Pak[(3)]

[1] Université Paris-Dauphine, Ceremade, F-75016 Paris, France,
mireille.summa@dauphine.fr
[2] *CNRS, UMR7534, F-75016 Paris, France,*
summa@ceremade.dauphine.fr
[3] *ISTHMA, Paris, France, pak@isthma.fr*

Abstract. We propose in this paper a new unsupervised classification approach for multivariate functional data. It consists in a piramidisation of an ascending hierarchical clustering. For this purpose we define new indices which generalize the usual ones to the functional context. The initial dissimilarity measures are based on the Kullback Leibler Divergence.
This method was applied in order to guide the variables selection in a modelling phase of time series when correlations failed. The final issue was to forecast the railways traffic in France for machine and metal products. The endogenous data are three volumes per distance along 21 years of the internal French railways traffic; the exogenous variables are macroeconomic aggregates, market costs, prices and product characteristics and eventually infrastructure development. The result of the research provides an ordered list of cofactors displayed by decreasing similarity to the target time series.

Keywords: time series, Kullback Leibler divergence, pyramidal order, functional clustering

1 Introduction

Clustering time series is an important challenge today because of the growing of time data bases, moreover of flows data (Romano et al (2007)).
Our first purpose was a necessary preprocessing of time series in order to help a classifying phase in a multivariate approach in econometrics. We were looking for ordering the curves in some way to obtain a guided procedure in the variables selection for the final models and also for their cardinality.
We propose therefore a pyramidisation of an ascending hierarchy on the initial functions, based on the *Kullback Leibler Divergence*, after having transformed the times series according to the requirements of the methods.

2 Hierarchical clustering of time series based on the Kullback Leibler divergence

Let F $= \{f_i(t), i \in \{1, \ldots, k\}, t \in T\}$ be a finite set C, $Card(C) = k$, of time series defined on the same time domain T with uniform scaling on the time axis.
Let T_i be an allowed transformation of f_i as defined in the literature Bagnal and Janacek (2005) such as: moving averages, spline or Gaussian smoothing, polynomial or wavelet bases, SARIMA modeling, SAX transformation.
In order to obtain only positive or null values, let us consider the common translation Tl of the F $= \{T_i[f_i(t)], i \in \{1, \ldots, k\}, t \in T\}$, which subtracts from each ordinate the $\min_{i,t} \{T_i[f_i(t)], i \in \{1, \ldots, k\}, t \in T, T_i[f_i(t)] \leq 0\}$; when there are no negative values, Tl is the null translation.
Let us now consider the homothetic transformation Th for each function which consists in dividing it by its L_1 norm.
Let Tr be the resulting transformation $Tr = Th \quad o \quad Tl$
Let G $= \{g_i(t), i \in \{1, \ldots, k\}, t \in T\} = \{(Tr \quad o \quad T_i)[fi(t)], i \in \{1, \ldots, k\}, t \in T\}$ be the finite set of functions resulting of the various transformations performed on the f_i.
We have:

$$\begin{cases} \forall i \in \{1, \ldots, k\}, \forall t \in T, g_i(t) \geq 0 \\ \forall i \in \{1, \ldots, k\}, \int_T g_i(t)dt = 1 \end{cases}$$

Definition 2. The Kullback-Leibler divergence between two G functions is:

$$DKL(g_i, g_j) = \int_T g_i(t) Log[\frac{g_i(t)}{g_j(t)}]dt + \int_T g_j(t) Log[\frac{g_j(t)}{g_i(t)}]dt$$

Theorem 1. *The Kullback-Leibler divergence is a dissimilarity index.*

Note that the Kullback-Leibler divergence is not a distance because the triangle inequality is not verified.
Our purpose is to extend the hierarchical clustering procedure on time series by the pyramidization option and by using the DKL in the initial dissimilarities matrix.
We must then define an index for measuring the dissimilarity between two sets of functions. Usual indices may be generalized to the functional context and specialized to the Kullback Leibler dissimilarity.

Definition 3. Let Cent be the notation for the mean function of a set of functions as defined in(Ramsay and Sulveman(1997)).

The mean linkage CL between two nodes L^m and L^r for a clustering procedure on a set of functions, based on the DKL is:

$$CL(L^m, L^r) = \int_T Cent^m(t) Log\left[\frac{Cent^m(t)}{Cent^r(t)}\right] dt + \int_T Cent^r(t) Log\left[\frac{Cent^r(t)}{Cent^m(t)}\right] dt$$

Definition 4. The average linkage AL between two nodes L^m and L^r for a clustering procedure based on the DKL is:

$$AL(L^m, L^r) = |L^m|^{-1} |L^r|^{-1} \sum_{i \in L^m} \sum_{i \in L^r} \left[\int_T g_i(t) Log[\frac{g_i(t)}{g_j(t)}]dt + \int_T g_j(t) Log[\frac{g_j(t)}{g_i(t)}]dt\right]$$

Definition 5. The simple linkage δ^*_{min} between two nodes L^m and L^r for a clustering procedure based on the DKL is:

$$\delta^*_{min}(L^m, L^r) = Min_{i \in L^m, j \in L^r} \left[\int_T g_i(t) Log[\frac{g_i(t)}{g_j(t)}]dt + \int_T g_j(t) Log[\frac{g_j(t)}{g_i(t)}]dt\right]$$

Definition 6. The complete linkage δ^*_{max} between two nodes L^m and L^r for a clustering based on the DKL is:

$$\delta^* max(L^m, L^r) = Max_{i \in L^m, j \in L^r} \left[\int_T g_i(t) Log[\frac{g_i(t)}{g_j(t)}]dt + \int_T g_j(t) Log[\frac{g_j(t)}{g_i(t)}]dt\right]$$

A classic definition can be used for indexing the hierarchy:

$$I[L(L_i \cup L_j)] = Max\left[\delta(L_i, L_j), I(L_i), I(L_j)\right]$$

3 The pyramidasation procedure

Hierarchies are the most widely used among ascending clustering approaches, even in the context of functional data, specifically time series (Ferraty and Vieu (2006)). Nevertheless, pyramids have the advantage of inducing an order on the initial units to be clustered (Bertrand and Diday (1985), Brito (1995), Durand and Fichet (1988)).
On the other hand growing a hierarchy is faster and less complex than building a pyramid.
The 'pyramidisation' of a hierarchy consists of reordering the terminals and the nodes of a hierarchy in such a way that the result induces an order on the terminals as compatible as possible with the initial dissimilarities matrix. Moreover the complexity must be similar to the hierarchy one, $O(n^2)$. A detailed algorithm is presented in (Pak (2005)).
One of the fundamental steps consists in a permutation of the terminals of two contiguous levels in order to put one next to one another the two closest ones. On figure 1, the fusions that can be done from the hierarchy are:

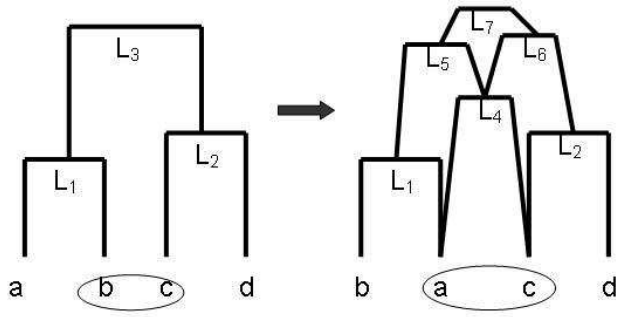


Fig. 1. Example of a hierarchy pyramidisation.

(L_1, L_2), (b, c), (a, c), (a, d), (b, d), (L_1, c), (L_1, d), (L_2, b), (L_2, a). Let suppose that the coupe (a,c) is the one that has the best index. The component including L_1 (or L_2) must be permuted in order to allow the creation of L_4.

On high dimensional functional data, a hierarchical procedure requires anyway further transformations of the initial data. The G functions are for example to be replaced by symbolic representations with a common alphabet (Huguenay (2006), Yang and Jia (2000)).

4 Application

This method was applied in order to guide the variables selection in a modelling phase of time series, especially when correlations failed.
The final issue is to forecast the railways traffic in France for machine, metal, and other products in 2040. The models research phase was the first step of a huge economic analysis. The data base consists in four endogenous variables and about 30 exogenous variables along 21 years and for nine different products aggregates. The initial data array is thus a three way data table which is transformed in a two way table but considering a whole time series as the value of a single cell. The framework becomes for this reason the multivariate functional analysis one.
The endogenous data are the volumes per distance of the domestic French railways traffic from 1985 to 2005 (path: one year), which are broken down into three variables according to a spatial clustering of the network: region 'within', regions 'contiguous', regions 'far away'. A fourth endogenous variable is the total traffic.
The exogenous variables are macroeconomic aggregates, market costs, prices and product characteristics and eventually infrastructure development descriptors.
The time series appear to have no seasonality: a piece of the data can be observed on Figure 2 and Figure 3.

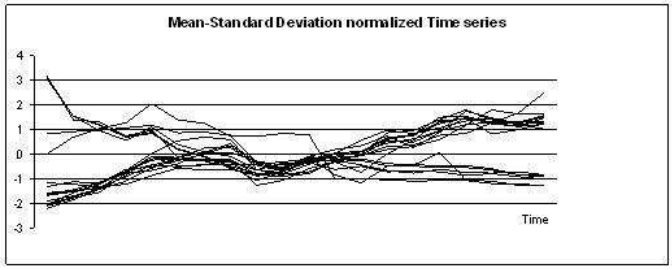


Fig. 2. French railways L2 normalized time series.

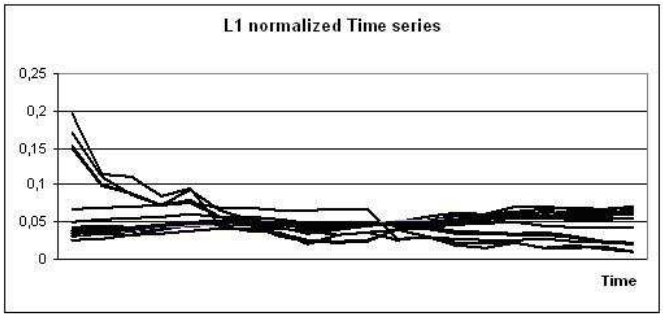


Fig. 3. French railways L1 normalized time series.

The clustering approach is performed on a table which is the merging of the initial time series table and the first differences data table of the same times series. By taking into account the two tables simultaneously, values and shape are both acting in the functions grouping.
Expert rules were also integrated at the end of the process, in order to cancel some of the variables from the final ordered set of terminals. In fact, final models had to contain only few exogenous variables (maximum four) and only one for each economical context: production -employment etc.-, product characteristic -prices, logistic etc.-, and infrastructure development of the railway network -distance etc.
Figure 4 shows the dendrogram of the hierarchy based on the DKL, after

pyramidistation of some of the terminals.

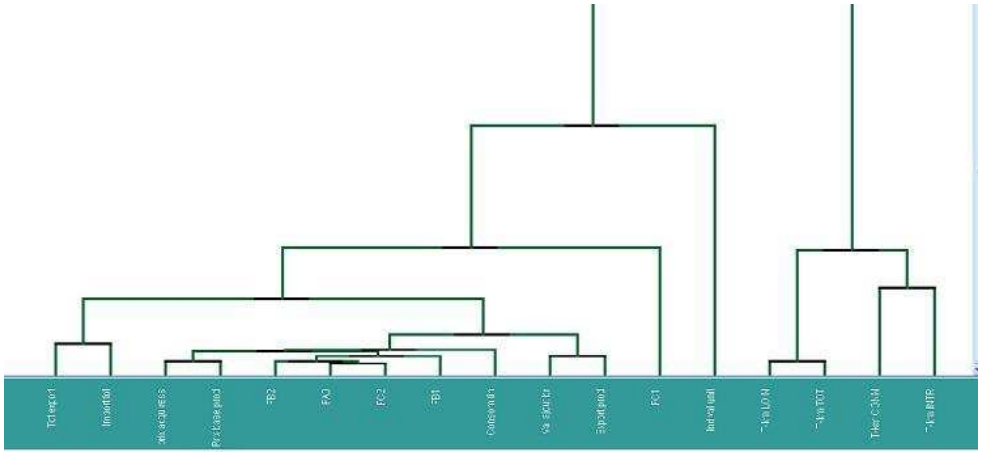


Fig. 4. Pyramidized dendrogram of French railways time series hierarchy.

The final retained order on the terminals, which are a subset of the initial clustered time series, is the following:
'Tot export, Import tot, prix acqu ress, Prix base prod FB2, FA0 FC2, FB1, Consom fin, Val ajout br, Export prod, FC1, Ind val unit, T-Km LOIN, Tkm TOT, T-km CONN, T-Km INTR'. We see for example that the four endogenous variables (LOIN i.e. region 'far away', TOT, i.e. region total, INTR i.e. region 'within', and CONN i.e. region 'contiguous') lay next to one another in a certain order which is quite coherent with the results of the first factorial plan of the Principal Component Analysis of Figure 5.

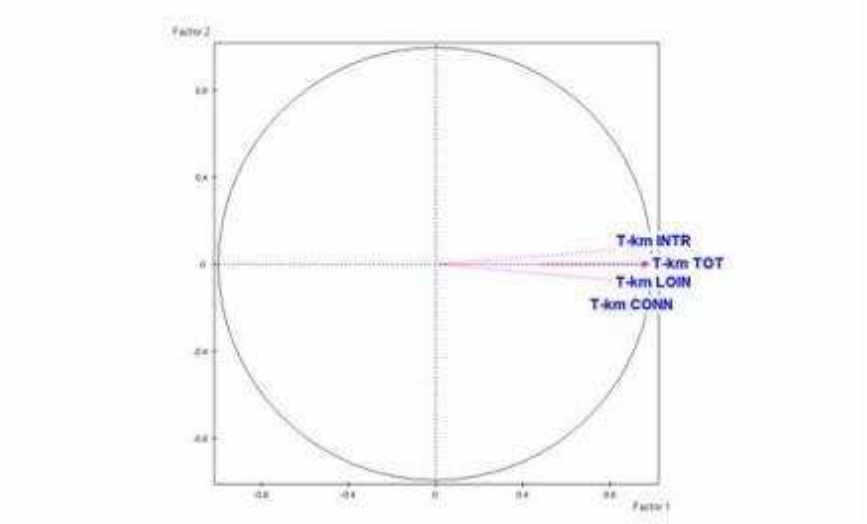


Fig. 5. Principal Component Analysis on the four endogenous variables.

Nevertheless the PCA does not lead very often to easily interpretable results whereas the result of the clustering approach always provides an ordered list of endogenous and exogenous variables, particularly for the cofactors which are displayed by decreasing similarity to the target time series.
Another application in epidemiology has been performed with the Deltametrics software©on epidemiological data (Gettler Summa and al. (2007)).

5 Discussion

Similarity in time, similarity in shape, and similarity in change are the guide lines for evaluating the goodness of a similarity coefficient on time series (Klapakis, Gada, Puttagunta (2001)).
Therefore, hierarchies of the same data, based on the L2 norm and the DKL should be compared, same for correlations values and DKL induced order.
A first evaluation must be the cost of the different transformations applied on the initial time series for the dissimilarity measures. We see on figure 1 and figure 2 two different final features of the same data which thus penalize in a different way the final interpretations (standard deviation or L1 homothetic coefficient per each curve).
The correlations are not capable of producing a complete order on the times series but on the first factorial plan, a correct interpretation of the relative

positions of the vectors suggests decisions similar to the ones allowed by the ordered terminals on a pyramidized hierarchy.
Anyway, in the French railway traffic forecasting application, by using the results of the pyramidized hierarchy, we shortened the search for entering the variables in the models and we reduced the number of models.
A specific interest of the Kullback Leibler Divergence is that it is possible to obtain a ranking of the couples of the time series, according to the significance of their DKL respect to an associated test (Broniatowski (2003), Keziou (2002)). A future development will thus be the possibility of measuring the homogeneity of a cluster. It will then be also possible to compare two clusters by evaluating their consistency.

References

BAGNAL, A. and JANACEK, G. (2005): Clustering Times series with Clipped Data, *Time Series Data Mining, Machine Learning , 58 (2/3) Springer Ed.*

BERTRAND, P. and DIDAY, E. (1985): A visual representation of the compatibility between an order and a dissimimilarity index. In: *the pyramids in Computational Statistics Quarterly, 2(1), 31–42*

BRITO, P. (1995): Symbolic Objects : Order Structure and Pyramidal Clustering. In: *Annals of Operations Research, 55, 277–297.*

BRONIATOWSKI, M. (2003): Estimation through Kullback-Leibler divergence. In: *Mathematical Methods of Statistics, 12, 4391–409.*

DURAND, C. and FICHET, B. (1988): One to one correspondence in pyramidal representation: an unified approach in Classification and Related Methods of Data Analysis. In: Ed. H.Bock Elsevier Science Publishers B.V. (North Holland).

FERRATY, F. and VIEU, P. (2006): Nonparametric Functional Data Analysis Theory and Practice. *Springer Series in Statistics.*

GETTLER SUMMA, M., STEYAERT, J.M., VAUTRAIN, F. and WEITKUNAT, R.(2007): A new clustering method for times series for discovering geographical cancer trends from 1960 to 2000. *Annals of epidemiology ,17 (9), 723–751.*

HUGUENAY, B. (2006): Cadre général et algorithmes de constructions pour des représentations symboliques adaptatives de séries temporelles, *Revue MODULAD,34, 1–12.*

KEZIOU, A. (2002): Sur l'estimation de l'entropie des lois support dénombrable. *Compte Rendu Acadmie des Sciences, Paris, 335 (9), 763–766.*

KLAPAKIS, K., GADA, D. and PUTTAGUNTA, V.(2001) : Distance measures for effective clustering of ARIMA times series. In proceedings of the IEEE Int'l Conference on Data Mining. San Jose, CA, Nov 29-Dec 2, 273–280.

PAK, K.K.(2005): Classifications hiérarchique et Pyramidale Spatiales : nouvelles techniques d'interprétation. *Thse de doctorat en Informatique, Universit Paris Dauphine France.*

RAMSAY, J.O. and SULVERMAN, B.W. (1997): Functional Data Analysis, *Springer series in statistics.*

ROMANO, E. and BLAZANELL, A. and VERDE, R. (2007): Knowledge extraction by dynamical clustering of sea waves streaming data proceedings of the European workshop on data stream analysis, Caserta, Italy.

YANG, E. and JIA, Y. (2000): Universal lossless coding of sources with large or unbounded alphabets. *Numbers, Information and Complexity*, (Ingo Althfor, et al, eds.), Kluwer Academic Publishers, 421–442.

Part XI

Graphical Models and Bayes Nets

A Wald's Test for Conditional Independence Skew Normal Graphs

Antonella Capitanio[1] and Simona Pacillo[2]

[1] Department of Statistical Sciences 'P. Fortunati', University of Bologna
Via Belle Arti 41, 40126 Bologna, Italy, *antonella.capitanio@unibo.it*
[2] Department PE.ME.IS., Statistical section, University of Sannio
Piazza Arechi II, 82100 Benevento, Italy, *simona.pacillo@unisannio.it*

Abstract. In this paper we present some results on model selection in conditional independence graphs when the variables have an extended Skew Normal distribution. This family is a slight modification of the Skew Normal one, that extends the class of Normal distributions through the addition of a shape parameter that regulates the skewness. A test for a single edge exclusion/inclusion is proposed which is based on a Wald-type statistic. Its performances, in finite samples, are assessed through numerical experiments.

Keywords: edge inclusion/exclusion test, graphical model, skew-normal

1 Introduction

The analysis of the conditional independence structure within the componentes of a d-dimensional multivariate random variable Y is usually performed through graphical models (Lauritzen (1996) and Whittaker (1990)). A graph $G = (V, E)$ consists of a pair of sets (V, E), where $V = \{1, \ldots, d\}$ is the set of vertices and $E \subset V \times V$ is the set of edges. Each variable of the multivariate distribution is associated with a vertex contained in V and the links between the vertices represent the absence of marginal or conditional independence. A graphical Gaussian model is a family of Normal distributions for Y with mean vector μ equal to zero and covariance matrix Σ, which is assumed positive definite. Here pairwise conditional independence relationships are signalled by a zero off-diagonal element in the inverse of the covariance matrix. More specifically, if Σ^{ij} denotes the (i, j)th entry of Σ^{-1}, then $\Sigma^{ij} = 0$ means conditional independence between the variables Y_i and Y_j given all the remaining ones. Such independence structure among the variables is represented by a conditional independence graph, that is an undirected graph where missing edges correspond to zero off-diagonal entries in Σ^{-1}.
In this context it is essential to have a model selection procedure to identify the graph underlying the data. The standard approach to model selection is the well-known stepwise method. It is based on the backward edge exclusion with a deviance difference stopping rule. The reader is referred for example

to Edwards (2000) for a detailed presentation of the different approaches to model selection in graphical Gaussian models. When Y has a $N_d(\mu, \Sigma)$ distribution, the pairwise conditional independence relationship between Y_i and Y_j implies also that the corresponding partial correlation coefficient $\rho_{ij|rest}$ is zero. This fact suggests that the model selection can be equivalently performed by carrying out the following test:

$$H_0 : \rho_{ij|rest} = 0 \qquad H_1 : \rho_{ij|rest} \neq 0$$

for each of the possible couples of vertices $\{i, j\}$. If Y does not follow a Gaussian distribution this approach is no longer suitable. Actually "out of the Normal context" incorrelation does not imply independence. In this paper we explore the situation where there is a departure from normality caused by a lack of symmetry and propose the use of the so called extended Skew Normal distribution as defined in Capitanio et al. (2003) to cope with skewness in the data. Within this context we present a test for a single edge inclusion/exclusion in conditional independence graphs.
Section 2 reviews some properties of the extended Skew Normal distribution. The proposed test is outlined in Section 3 while its performances in finite sample are assessed numerically in Section 4.

2 Skew normal conditional independence graphs

When the deviation from normality is due to a lack of symmetry the class of Skew Normal (SN) distributions defined by Azzalini and Dalla Valle (1996) provides a useful model to represent the data. This family extends the Gaussian distribution by adding a skewness parameter α; the Normal model is obtained as a special case when $\alpha = 0$. The SN distribution allows to carry out inference based on the likelihood function while dealing with skewness, and shares many properties with the Normal one, such as closure under marginalization and linear transformations.
The extended Skew Normal (ESN) distribution is a slight extension of the SN distribution that in addition achieves closure under conditioning, and hence it is suitable for the analysis of conditional independence relationships in the context of graphical models.
The density function of a d-dimensional ESN variate is:

$$f(y) = \frac{1}{\Phi(\tau)} \phi_d(y - \xi; \Omega) \Phi\left(\tau(1 + \alpha^T \overline{\Omega} \alpha)^{1/2} + \alpha^T \omega^{-1}(y - \xi)\right) \qquad (1)$$

where $\phi_d(y; \Omega)$ is the density of a d-dimensional $N_d(0, \Omega)$ variate, Φ is the distribution function of a $N(0, 1)$, Ω is a full rank covariance matrix, ω is a diagonal matrix such that $\overline{\Omega} = \omega^{-1} \Omega \omega^{-1}$ is the corresponding correlation matrix, α is a parameter regulating skewness, ξ is the location parameter and $\tau \in \Re$ is an additional shape parameter. When $\tau = 0$ the SN density is

recovered.
The mean vector and the covariance matrix of Y are:

$$E(Y) = \xi + \zeta_1(\tau)\omega\delta \qquad Var(Y) = \Omega + \zeta_2(\tau)\omega\delta\delta^T\omega$$

where $\zeta_m(\cdot)$ is the mth derivative of $log\,[2\Phi(\cdot)]$ and $\delta = (1 + \alpha^T\overline{\Omega}\alpha)^{-1/2}\overline{\Omega}\alpha$.
A density having form (1) arose in Azzalini and Capitanio (1999, Section 4, expression (13)) from a conditioning operation on a SN variate. These authors stated the conditions for independence among blocks of linear transformations of Skew Normal random variables (see their Proposition 6), and showed (see Section 6.3) how they can be extended to the case of the conditional SN density. Arnold and Beaver (2000) also examined densities of type (1), and noticed the closure under conditioning. Capitanio et al. (2003) investigated the relationships of conditional independence among the components of an ESN variate, as well as other issues related to the use of this model in the context of graphical models. Actually, if $Y = (Y_1, \ldots, Y_d)$ has density (1), pairwise conditional independence between Y_i and Y_j occurs if and only if the following conditions hold simultaneosly:

$$(a)\ \Omega^{ij} = 0 \quad \text{and} \quad (b)\ \alpha_i\alpha_j = 0 \tag{2}$$

where Ω^{ij} denotes the (i, j)th entry of Ω^{-1}.
Condition (a) shows that for an ESN variate the matrix Ω^{-1} plays the same role of the concentration matrix Σ^{-1} used in the Normal context, nevertheless a further condition, given by (b), on the elements of the shape vector α needs to be considered.
After some algebra the inverse of the covariance matrix of Y turns out to be

$$\Omega^{-1} - \zeta_2(\tau)\frac{\omega^{-1}\alpha\alpha^T\omega^{-1}}{(1 + \zeta_2(\tau)\alpha^T\overline{\Omega}\alpha + \alpha^T\overline{\Omega}\alpha)}. \tag{3}$$

It is clearly evident that if Y_i and Y_j fulfill conditions (2) then the (i, j)th entry of (3) is equal to zero, whilst the fact that (3) has a zero entry does not imply that a pairwise conditional independence relationship occurs between the corresponding components of Y. As a consequence the approach described in Section 1 for testing conditional independence appears to be inadequate.
The conditional independence graph $G = (V, E)$ used to specify the association structure among the components of Y is obtained by connecting two vertices $\{i\}$ and $\{j\}$ if Ω^{ij} and/or the product $\alpha_i\alpha_j$ are different from zero. On the contrary there is a missing edge between the vertices $\{i\}$ and $\{j\}$ if conditions (a) and (b) hold simultaneously.
Suppose for example, that the 3-dimensional variate Y has an ESN distribution with parameters

$$\Omega^{-1} = \begin{pmatrix} \Omega^{11} & \Omega^{12} & 0 \\ \Omega^{21} & \Omega^{22} & \Omega^{23} \\ 0 & \Omega^{32} & \Omega^{33} \end{pmatrix} \qquad \alpha = \begin{pmatrix} 0 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}.$$

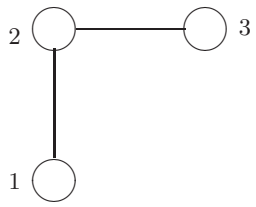


Fig. 1. Conditional independence graph for the ESN_3

Since $\Omega^{13} = 0$ and $\alpha_1\alpha_3 = 0$, Y_1 and Y_3 are independent conditionally on Y_2, and the edge that connects vertices $\{1\}$ and $\{3\}$ is missing. Furthermore, since Ω^{12} and Ω^{23} are different from zero, there are two edges which connect vertices $\{1\}$ and $\{2\}$, and $\{2\}$ and $\{3\}$, respectively. The resulting conditional independence graph is shown in Figure 1.

3 Methodology

In this section we present a Wald test useful for model selection for the extended Skew Normal conditional independence graphs. By exploiting the pairwise conditional independence conditions (2), the model selection procedure can be based on a test for a single edge exclusion/inclusion whose null hypothesis is:

$$H_0 : g(\theta) = \begin{pmatrix} \Omega^{ij} \\ \alpha_i\alpha_j \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{4}$$

To our purposes it is useful to consider the parametrization $\theta = (\xi, \Omega^{-1}, \alpha, \tau)$. Consider an observed random sample $y_1, \ldots, y_n$ drawn from Y having density (1). The log-likelihood function for θ is $\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta)$ where $\ell_i(\theta)$ is the log-likelihood based on a single observation:

$$\ell_i(\theta) = -\frac{d}{2}log(2\pi) + \frac{1}{2}log\left|\Omega^{-1}\right| - \frac{1}{2}(y_i - \xi)^T\Omega^{-1}(y_i - \xi)$$

$$+log\left[\Phi\left\{\tau(1 + \alpha^T\overline{\Omega}\alpha)^{1/2} + \alpha^T\omega^{-1}(y_i - \xi)\right\}\right] - log\left[\Phi(\tau)\right].$$

The log-likelihood function cannot be maximized in closed form: therefore the maximization of $l(\theta)$ requires the use of numerical methods.
Let $\hat{\theta}$ be the maximum likelihood estimator of θ and $g(\hat{\theta})$ the estimator of $g(\theta)$. By applying the delta method the covariance matrix of $g(\hat{\theta})$ is $\Sigma_g(\theta) = g'(\theta)I(\theta)g'(\theta)^T$ where $I(\theta)$ is the information matrix of $\hat{\theta}$ and $g'(\cdot)$ denotes

the matrix of the first derivatives of g with respect to θ. In obvious notation we have

$$\Sigma_g(\theta) = \begin{bmatrix} Var(\hat{\Omega}^{ij}) & Cov(\hat{\Omega}^{ij}, \hat{\alpha}_i\hat{\alpha}_j) \\ \cdot & Var(\hat{\alpha}_j\hat{\alpha}_i) \end{bmatrix}$$

with

$$Cov(\hat{\Omega}^{ij}, \hat{\alpha}_i\hat{\alpha}_j) = \alpha_j Cov(\hat{\Omega}^{ij}, \hat{\alpha}_i) + \alpha_i Cov(\hat{\Omega}^{ij}, \hat{\alpha}_j)$$

and

$$Var(\hat{\alpha}_j\hat{\alpha}_i) = \alpha_i^2 Var(\hat{\alpha}_j) + \alpha_j^2 Var(\hat{\alpha}_i) + 2\alpha_i\alpha_j Cov(\hat{\alpha}_i\hat{\alpha}_j).$$

The elements of $\Sigma_g(\theta)$ can be estimated by replacing the parameters by their estimates and by replacing the variances and covariances by the corresponding elements of the inverse observed information matrix. We shall indicate the sample version of $\Sigma_g(\theta)$ by $\hat{\Sigma}_g$.
The test concerning the null hypothesis (4) can be based on the Wald-type statistic

$$W_n(Y) = \left(g(\hat{\theta}) - g(\theta)\right)^T \hat{\Sigma}_g^{-1} \left(g(\hat{\theta}) - g(\theta)\right).$$

Under the null hypothesis, which implies that variables Y_i and Y_j are conditionally independent given the remaining ones, the statistic $W_n(Y)$ has a χ^2_2-distribution.

4 Simulation study

In this section the finite sample performance of the test has been explored via Monte Carlo experiments. The objective of the analysis is twofold. On one hand we are interested in the properties of the test itself; on the other hand we wish to investigate its ability in detecting the presence of non linear dependence relationships which may occur in skew normal data. To this purpose, we carried out two experiments organized as follows.
We generated 10000 random samples of size $n = 100, 200, 500$ and 1000 from each of two 3-dimensional ESN variates with parameters

$${}_{(1)}\Omega^{-1} = \begin{pmatrix} 0.81 & -0.756 & 0 \\ -0.756 & 1.44 & 1.344 \\ 0 & 1.344 & 2.56 \end{pmatrix} \quad {}_{(2)}\Omega^{-1} = \begin{pmatrix} 2.076 & -0.7920 & 0 \\ -0.7920 & 1.7424 & 0.6633 \\ 0 & 0.6633 & 1.2636 \end{pmatrix}$$

$${}_{(1)}\alpha = {}_{(2)}\alpha = \begin{pmatrix} 0 \\ -3 \\ 3.5 \end{pmatrix} \quad {}_{(1)}\xi = {}_{(2)}\xi = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad {}_{(1)}\tau = {}_{(2)}\tau = 0.4$$

where the subscript on the left hand side indicates the experiment in which they are used. Note that the parameters values are such that in both the experiments the conditional independence relationship $Y_1 \perp Y_3|Y_2$ holds, so that the corresponding graph is the one shown in Figure 1, where the edge

set is $E = \{(1,2),(2,3)\}$. It is also straightforward to check that the two populations from which the data are drawn differ in the magnitude of the partial correlation coefficients. More specifically, in experiment 1 ${}_{(1)}\rho_{12|3} = 0.698$ and ${}_{(1)}\rho_{23|1} = -0.686$ while in experiment 2 ${}_{(2)}\rho_{12|3} = 0.381$ and ${}_{(2)}\rho_{23|1} = -0.155$.

We computed the percentage of rejections of the null hypothesis in (4) on each edge when the nominal level of the test is 0.10 and 0.05, respectively. The parameters estimates have been obtained using the reparametrization described in Capitanio et al. (2003). The observed information matrix $I(\hat{\theta})$ has been evaluated numerically. The results are displayed in Tables 1 and 3. In addition we investigated what happens if, regardless of the actual distribution, the usual test on the null hypothesis of vanishing partial correlations were used. This test can be quickly computed using the `pcor.test` function available in the `library(ggm)` implemented in `R` software. The corresponding results are shown in Tables 2 and 4.

Note that the null hypothesis is true for the edge connecting vertices $\{1\}$ and $\{3\}$ so that the corresponding entries in the tables approximate the actual level. On the contrary the null hypothesis is false for the edges connecting vertices $\{1\}$ and $\{2\}$, and $\{2\}$ and $\{3\}$ respectively, hence the corresponding entries are to be interpreted as estimates of the power of the test.

The proposed test appears to be somewhat conservative. However when n increases the actual level becomes closer to the nominal one and the power is satisfactory. By comparing the results of the two experiments we notice that, as the partial correlation coefficients decrease, both tests become less powerful. This behavior is definitely more critical for the test based on the partial correlation coefficients. It suggests that when the data come from a skew normal distribution the usual test on partial correlation coefficients could be quite inadequate for detecting non linear dependence relationships, and in this sense the proposed Wald's-type test seems to be preferable.

	$\alpha = 0.10$ *Edges*			$\alpha = 0.05$ *Edges*		
n	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
100	99.860%	1.904%	99.892%	99.772%	0.796%	99.792%
200	100.00%	3.19%	100.00%	100.00%	1.38%	100.00%
500	100.00%	4.13%	100.00%	100.00%	1.77%	100.00%
1000	100.00%	5.7%	100.00%	100.00%	2.63%	100.00%

Table 1. Experiment 1 - percentage of rejections of the null hypothesis (4) using a nominal level equal to 0.10 and 0.05.

	$\alpha = 0.10$ *Edges*			$\alpha = 0.05$ *Edges*		
n	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
100	100.0%	9.21%	100.0%	100.0%	4.72%	100.0%
200	100.0%	10.05%	100.0 %	100.0%	5.29%	100.0%
500	100.0%	9.7%	100.0%	100.0%	4.56%	100.0%
1000	100.0%	9.7%	100.0%	100.0%	4.88%	100.0%

Table 2. Experiment 1 - percentage of rejections of the null hypothesis when a test for zero partial correlation is used.

	$\alpha = 0.10$ *Edges*			$\alpha = 0.05$ *Edges*		
n	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
100	85.92%	2.87%	73.45%	77.01%	1.09%	64.28%
200	99.68%	3.85%	97.14%	99.41%	1.70%	95.97%
500	100.00%	5.30%	99.95%	100.00%	2.40%	99.94%
1000	100.00%	6.65%	100.00%	100.00%	2.77%	100.00%

Table 3. Experiment 2 - percentage of rejections of the null hypothesis (4) using a nominal level equal to 0.10 and 0.05.

	$\alpha = 0.10$ *Edges*			$\alpha = 0.05$ *Edges*		
n	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
100	98.89%	10.07%	45.47%	97.59%	5.36%	33.88%
200	100.00%	9.43%	69.94%	99.98%	4.86%	58.91%
500	100.00%	9.89%	95.84%	100.00%	4.90%	92.68%
1000	100.00%	9.84%	99.90%	100.00%	4.82%	99.73%

Table 4. Experiment 2 - percentage of rejections of the null hypothesis when a test for zero partial correlation is used.

5 Conclusion

This paper provides a simple way to carry out model selection in conditional independence graphs when the variables have an extended Skew Normal distribution. A test is proposed, whose null hypothesis takes into account the pairwise conditional independence property for the considered class of distributions. A Wald-type statistic, with an asymptotic χ^2 distribution, is obtained. The main advantage of this procedure is the possibility to work with a family which shares many properties with the Normal model while it is able

to fit distributions of the data which are in a neighbourhood of the Gaussian one.
A simulation study shows that when the sample size increases, the actual level of the test is reasonably close to the nominal one and the power is satisfactory.

6 Acknowledgments

Work supported by PRIN 2006, grant No. 2006132978, from MIUR, Italy.
We are grateful to Adelchi Azzalini for kindly providing the numerical routines used in the estimation of the parameters. We also wish to thank an anonymous referee, whose constructive comments greatly enhanced the paper.

References

ARNOLD, B.C. and BEAVER, R.J. (2000): Hidden truncation models. *Sankhyā Series A 62, 23-35.*

AZZALINI, A. and CAPITANIO, A. (1999): Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 61 (3), 579-602.*

AZZALINI, A. and DALLA VALLE, A. (1996): The multivariate skew-normal distribution. *Biometrika 83 (4), 715-726.*

CAPITANIO, A., AZZALINI, A. and STANGHELLINI, E. (2003): Graphical models for skew-normal variates. *Scandinavian Journal of Statistics 30, 129-144.*

EDWARDS, D. (2000): *Introduction to Graphical Modelling.* Springer, New York.

LAURITZEN, S.L. (1996): *Graphical Models.* Oxford Science Publications, Oxford.

WHITTAKER, J. (1990): *Graphical Models in Applied Multivariate Statistics.* Wiley, New York.

Package giRaph for Graph Representation in R

Luca La Rocca[1], Claus Dethlefsen[2], and Jens Henrik Badsberg[3]

[1] Economics and Communication Sciences, University of Modena and Reggio E.
Viale Allegri 9, 42100 Reggio Emilia, Italy, *luca.larocca@unimore.it*
[2] Center for Cardiovascular Research, Aalborg Hospital
Sdr. Skovvej 15, 9000 Aalborg, Denmark, *cld@rn.dk*
[3] Quality Assurances, Statens Serum Institut
Artillerivej 5, 2300 Copenhagen S, Denmark, *jhd@ssi.dk*

Abstract. This paper presents the giRaph package for R, providing formal classes and methods to handle a broad family of graphs, including graphs with loops, multiple edges and hyperedges (i.e. edges involving more than two vertices) both directed and undirected. Since there is no unique way to represent a graph that is optimal for all computations, four different representations are considered: incidence list, incidence matrix, adjacency list and adjacency matrix. Simple methods to set and retrieve information using these representations are made available.

Keywords: graphical model, hypergraph, multigraph, simple graph

1 Introduction

Different flavours of *graphs* have proven useful in disparate fields of science, providing scientists with an appealing modelling tool. In particular, graphs can be used to define statistical models (*graphical models*) and in this context *hypergraphs* and *simple graphs* are the most interesting flavours; see Lauritzen (1996). Other important flavours include the *directed hypergraphs* studied by Gallo et al. (1993) as a tool to deal with some classes of problems arising in operations research and in computer science, and the *multigraphs* used by Gomes et al. (2006) in the field of transportation analysis. The picture is barely sketched, but already suffices to provide motivation for the broad family of graphs implemented by giRaph (Badsberg et al. (2007)). This is an R (R Development Core Team (2007)) package providing formal, i.e. S4, classes and methods (Chambers (1998)) for graph representation and manipulation. In particular, giRaph is intended as a contribution to the gR project put forth by Lauritzen (2002) for the development of graphical model facilities in R; the CRAN (Comprehensive R Archive Network) Task View on gRaphical models, available at http://cran.r-project.org/web/views/gR.html, contains an annotated list of R packages dealing with graphical models.

The giRaph package provides classes for four different graph representations: *incidence list*, *incidence matrix*, *adjacency list* and *adjacency matrix*; see e.g. Ahuja et al. (1993). Each representation exhibits its own computational advantages and disadvantages, and is suited to represent a different

graph family. Classes for four such families are available in `giRaph`, with methods for handling alternative representations transparently with respect to the user; see Section 3. This richness and flexibility in graph representation is a distinctive feature of `giRaph`, which is not shared by other `R` packages for graphs such as `igraph` by Csardi (2007), `graph` by Gentleman et al. (2007) and `mathgraph` by Burns et al. (2007). The `giRaph` package also provides formal classes for edges and vertices, so that simple graph operations such as adding an edge or extracting an induced subgraph can be performed via overloaded operators.

For each class the `giRaph` package provides a robust `initialize` method, that takes care of producing valid output from varied input, an user-friendly `show` method, adopting typical graph notation, and methods to set and retrieve information. In addition, conversions between different graph representations, and between graphs of different families, are implemented by means of `coerce` methods. Finally, an interface to the `mathgraph` package and to `dynamicGraph` by Badsberg (2007) is available; the latter is an interactive graphical tool that is also part of the `gR` project.

The rest of the paper is organised as follows. Section 2 presents the graph families and representations implemented by `giRaph`. Section 3 illustrates the facilities provided by `giRaph` for handling alternative representations.

2 Graphs and their representations

We consider a broad notion of graph: any graph $\mathcal{G} = (V, E)$ consists of a finite set V of *vertices* together with a finite multiset E of *edges*, representing some kind of relationship between vertices. This is our *abstract definition* of graph, and we need to specify what an edge is, in order to concretely define a family of graphs. We consider the two following types of edge.

Definition 7. Given a vertex set V, an *undirected edge* of V is a subset of V. An undirected edge $e \subseteq V$ is *proper* if it is non-empty ($e \neq \emptyset$).

Definition 8. Given a vertex set V, a *directed edge* of V is an ordered sequence $(V_1, \ldots, V_k)$ of disjoint non-empty subsets of V (proper undirected edges of V) called its *components*. A directed edge is *proper* if it consists of at least two components ($k \geq 2$).

For simplicity, we say that an edge e contains the vertex v, or equivalently that v belongs to e, both if e is undirected and $v \in e$ and if $e = (V_1, \ldots, V_k)$ is directed and $v \in V_j$ for some $j \in \{1, \ldots, k\}$. Then, we are interested in the following classification of edges.

Definition 9. A *hyperedge* is an edge containing more than two vertices, while an *ordinary edge* is an edge containing at most two vertices. An ordinary edge containing just a single vertex is called a *loop*.

The giRaph package provides a formal class, vertexSet, to represent V. We assume that vertices are identified by name, and let the vertexSet class inherit from the character class. Vertex names are restricted to be unique syntactically valid names, which is guaranteed by the initialize method; they can be retrieved (as a character object) by means of the names method. A short-hand function, v, makes vertex set construction immediate:

```
> v("a","b","c")
{a,b,c}
```

Methods for vertexSet objects include: a card method to retrieve their cardinality, an isEmpty method to check whether this is zero, and a comparison method, areTheSame, to check whether two vertexSet objects represent the same V (disregarding storage order). A multiple extractor method, [, extracting vertex subsets, and a single extractor method, [[, accessing individual vertex names, are also available; notice that here storage order matters. Finally, pairwise union, intersection and asymmetric difference of vertex sets are implemented by overloading the +, * and - operators, respectively.

Turning to edge representation, the giRaph package provides two different classes for the two types of edge under consideration: undirectedEdge and directedEdge. They both inherit from a *virtual* edge class, intended as a superclass of all edge classes, including those that might be implemented in the future. Since every edge is defined with respect to a given vertex set, we efficiently identify the vertices contained in an edge by means of numbers referring to an understood vertexSet object (here storage order matters). Thus, we let the undirectedEdge class inherit from the integer class, and the directedEdge class inherit from the list class (with integer elements). We provide two short-hand functions, u and v, for undirected and directed edge construction, respectively, and also a short-hand function, r, for "reverse" construction of directed edges:

```
> u(1,2)          > d(1,2)          > r(1,2)
1--2              1->2              2->1
```

The above displayed edges are $\{a, b\}$, $(\{a\}, \{b\})$ and $(\{b\}, \{a\})$, respectively, if the understood vertex set is $\{a, b, c\}$. A showRel method is available to display an edge using the vertex names in a given vertexSet object. Even if the initialize method guarantees that the numeric identifiers in an edge object are strictly positive integers, these are not necessarily meaningful with respect to any given vertexSet object: e.g. u(1,4) makes no sense for v("a","b","c"). For this reason, we provide a maxId method returning the maximum numeric identifier in an edge object, which should be compared with the the cardinality of the vertex set. Since vertex storage order matters, a recode method is available, in case we need to consider an edge with respect to a different vertexSet object (e.g. after subsetting). The cardinality of an edge, returned by the card method, is the number of vertices belonging to the edge. The same number is returned by length, for undirected edges,

whereas `length` returns the number of components for directed edges. The `areTheSame` methods comply to the following criteria: two undirected edges are equal when they contain the same vertices, two directed edges are equal when they have the same components in the same order, and two edges of different type are never equal. Finally, extractor methods behave as follows: the `[` method extracts an object of the same class, whereas the `[[` method extracts a single numeric identifier from an undirected edge and a component (in the form of an undirected edge) from a directed edge.

The `edgeList` class is designed to represent the multiset of edges, E, of a graph $\mathcal{G} = (V, E)$. This class inherits from the `list` class, so that the easiest way to construct an `edgeList` object is to coerce a list of `edge` objects. Just like for individual edge classes, there are `showRel`, `maxId` and `recode` methods to handle the relationship with the understood `vertexSet` object. Moreover, we furnish an `isPresent` method to check whether a specific edge occurs at least once in a given multiset. The `card` method returns the total number of edge occurrences. The `areTheSame` method checks whether two `edgeList` objects contain the same edges with the same multiplicity (disregarding storage order). The multiple extractor method, `[`, returns a multiset of edges, while the single extractor method, `[[`, returns a single edge. Finally, the possibility of adding and removing individual edge occurrences is implemented by overloading the `+` and `-` operators:

```
> new("edgeList")+u(1,2)+d(1,2)
{
1--2
1->2
}
```

We are now in a position to introduce the `incidenceList` class, which is intended as a framework to represent any graph fitting our abstract definition. An `incidenceList` object consists of two slots: a `vetexSet` slot, `V`, and an `edgeList` object, `E`. In principle, if new types of edge are defined in the future, it will be possible to accommodate them in this class. However, for the time being, our most general family of graphs will be the following.

Definition 10. A *general graph* $\mathcal{G} = (V, E)$ consists of a vertex set V together with a finite multiset of proper directed and undirected edges for V.

An example of general graph is drawn in Figure 1, and its incidence list representation is reported in Table 1. Note the two hyperedges: the directed hyperedge $(\{f\}, \{e\}, \{b, d\}, \{a, c\})$, represented by the first element of `E`, and the undirected hyperedge $\{b, d, e\}$, represented by the second element of `E`. All other edges are ordinary. Also note the loop $\{i\}$, depicted in Figure 1 with two arrows so as to point out its special status, and the three occurrences of the edge $(\{e\}, \{h\})$. In the following, for the sake of clarity, we will stick to the graphical edge notation of Table 1.

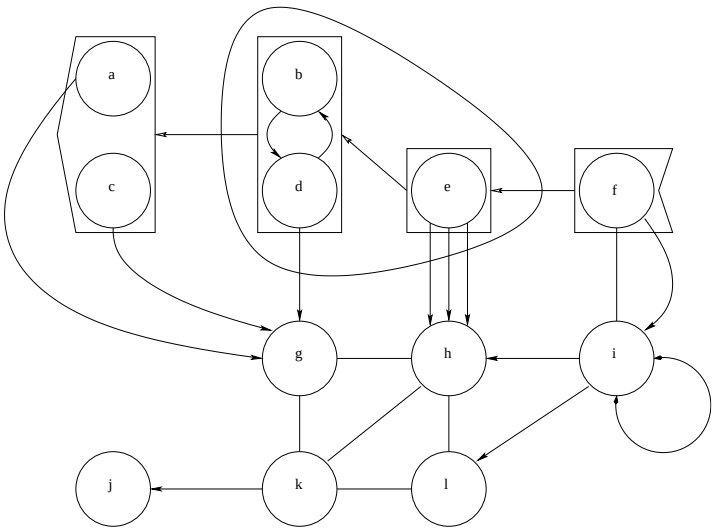


Fig. 1. Example of general graph.

```
An object of class "incidenceList"
V = {a,b,c,d,e,f,g,h,i,j,k,l}
E = {f->e->b--d->a--c, b--d--e, b->d, d->b, a->g, c->g, d->g, e->h,
     e->h, e->h, f--i, f->i, i<>i, i->h, i->l, g--h, h--l, k--l, k--g,
     k--h, k->j}
```

Table 1. Incidence list representation of the general graph in Figure 1. Undirected edges are denoted by `--`, directed edges by `->`, and loops by `<>`. Notice that, due to space reasons, the `R` output for the multiset of edges has been compressed.

An alternative representation for general graphs is implemented by the `incidenceMatrix` class, inheriting from the `matrix` class. The data part of an `incidenceMatrix` object representing $\mathcal{G} = (V, E)$ stores a matrix I with a row for each occurrence of an edge in E and a column for each vertex in V: if e_j is undirected, then $I_{jv} = 1$ if $v \in e_j$, and $I_{jv} = 0$ otherwise; if $e_j = (V_1, \ldots, V_m)$ is directed, then $I_{jv} = k$ if $v \in V_k$ for some $k \in \{1, \ldots, m\}$, and $I_{jv} = 0$ otherwise. An `incidenceMatrix` object representing the general graph of Figure 1 is shown in Table 2 (left); it was obtained via the coercion `as(G,"incidenceMatrix")` from the incidence list, `G`, of Table 1. An incidence matrix allows to check in constant time whether a given vertex belongs to a given edge, whereas an incidence list saves storage memory.

We now consider two subfamilies of graphs that admit other interesting representations. The first one is obtained by ruling out hyperedges.

Definition 11. A *multigraph (MG)* is a general graph $\mathcal{G} = (V, E)$ such that all edges in E are ordinary.

```
An object of class "incidenceMatrix"
      a b c d e f g h i j k l
 [1,] 4 3 4 3 2 1 0 0 0 0 0 0
 [2,] 0 1 0 1 1 0 0 0 0 0 0 0
 [3,] 0 1 0 2 0 0 0 0 0 0 0 0
 [4,] 0 2 0 1 0 0 0 0 0 0 0 0
 [5,] 1 0 0 0 0 0 2 0 0 0 0 0
 [6,] 0 0 1 0 0 0 2 0 0 0 0 0
 [7,] 0 0 0 1 0 0 2 0 0 0 0 0
 [8,] 0 0 0 0 1 0 0 2 0 0 0 0
 [9,] 0 0 0 0 1 0 0 2 0 0 0 0
[10,] 0 0 0 0 1 0 0 2 0 0 0 0
[11,] 0 0 0 0 0 1 0 0 1 0 0 0
[12,] 0 0 0 0 0 1 0 0 2 0 0 0
[13,] 0 0 0 0 0 0 0 0 1 0 0 0
[14,] 0 0 0 0 0 0 0 0 2 1 0 0 0
[15,] 0 0 0 0 0 0 0 0 0 1 0 0 2
[16,] 0 0 0 0 0 0 0 1 1 0 0 0 0
[17,] 0 0 0 0 0 0 0 0 1 0 0 0 1
[18,] 0 0 0 0 0 0 0 0 0 0 0 1 1
[19,] 0 0 0 0 0 0 0 1 0 0 0 1 0
[20,] 0 0 0 0 0 0 0 0 1 0 0 1 0
[21,] 0 0 0 0 0 0 0 0 0 0 2 1 0
```

```
An object of class "adjacencyMatrix"
  a b c d e f g h i j k l
a 0 0 0 0 0 0 1 0 0 0 0 0
b 0 0 0 1 0 0 0 0 0 0 0 0
c 0 0 0 0 0 0 1 0 0 0 0 0
d 0 1 0 0 0 0 1 0 0 0 0 0
e 0 0 0 0 0 0 0 1 0 0 0 0
f 0 0 0 0 0 0 0 0 1 0 0 0
g 0 0 0 0 0 0 0 1 0 0 1 0
h 0 0 0 0 0 0 1 0 0 0 1 1
i 0 0 0 0 0 1 0 1 0 0 0 1
j 0 0 0 0 0 0 0 0 0 0 0 0
k 0 0 0 0 0 0 1 1 0 1 0 1
l 0 0 0 0 0 0 0 1 0 0 1 0
```

Table 2. Incidence matrix representation of the general graph in Figure 1 (left) and adjacency matrix representation of the simple graph obtained from it as described in the text (right).

For an MG $\mathcal{G} = (V, E)$ we define the notions of *neighbour*, *parent* and *child* of a vertex $v \in V$: u is a neighbour of v if u—$v \in E$, u is a parent of v if $u \to v \in E$, and u is a child of v if $v \to u \in E$. Clearly, since multiple edges are possible, the neighbours of v in $\mathcal{G}$, $\mathrm{ne}_{\mathcal{G}}(v)$, the parents of v in $\mathcal{G}$, $\mathrm{pa}_{\mathcal{G}}(v)$, and the children of v in $\mathcal{G}$, $\mathrm{ch}_{\mathcal{G}}(v)$, are not necessarily disjoint. Moreover, their elements come with an associated multiplicity. The `adjacencyList` class, inheriting from the `list` class, implements a *local* representation of MGs. The generic element of the data part of an `adjacencyList` object, corresponding to some $v \in V$, is a `list` object with three named elements, `ne`, `pa` and `ch`, representing the vertices of $\mathrm{ne}_{\mathcal{G}}(v)$, $\mathrm{pa}_{\mathcal{G}}(v)$ and $\mathrm{ch}_{\mathcal{G}}(v)$, respectively (each vertex replicated according to its own multiplicity). The adjacency list shown in Table 3 was obtained from the incidence matrix, `I`, of Table 2 by dropping the two hyperedges, i.e. via the coercion `as(I,"adjacencyList")`.

```
An object of class "adjacencyList"
a <- {}
  -- {}
  -> {g}
b <- {d}
  -- {}
  -> {d}
c <- {}
  -- {}
  -> {g}
d <- {b}
  -- {}
  -> {b,g}
e <- {}
  -- {}
  -> {h,h,h}
f <- {}
  -- {i}
  -> {i}
g <- {a,c,d}
  -- {h,k}
  -> {}
h <- {e,e,e,i}
  -- {g,l,k}
  -> {}
i <- {f}
  -- {f,i}
  -> {h,l}
j <- {k}
  -- {}
  -> {}
k <- {}
  -- {l,g,h}
  -> {j}
l <- {i}
  -- {h,k}
  -> {}
```

Table 3. Adjacency list representation of the multigraph obtained from the general graph of Figure 1 by dropping the two hyperedges. Empty sets are denoted by {}, the symbols `<-`, `--` and `->` denote parents, neighbours and children, respectively.

Another useful family of graphs is obtained from the family of MGs by ruling out loops and multiple edges.

Definition 12. A *simple graph (SG)* is a general graph $\mathcal{G} = (V, E)$ such that every edge in E contains exactly two vertices, every edge in E has multiplicity one, and for any two different vertices u and v in V at most one among u—v, $u \to v$ and $v \to u$ belongs to E.

For SGs the adjacencyMatrix class, inheriting from the matrix class, offers a representation that allows to check in constant time whether there is or not an edge containing a given pair of vertices, at the price of storing many zeros in case the answer is in the negative for most pairs. The data part of an adjacencyMatrix object representing $\mathcal{G} = (V, E)$ stores a matrix X with a row and a column for each vertex in V: $X_{uv} = X_{vu} = 1$ if $u—v \in E$, $X_{uv} = 1$ and $X_{vu} = 0$ if $u \to v \in E$, and $X_{uv} = X_{vu} = 0$ otherwise. Table 2 (right) shows the adjacency matrix obtained from the adjacency list, A, of Table 3 by dropping $i \leftrightarrow i$, keeping just one copy of $e \to h$, letting $f—i$ prevail on $f \to i$ and replacing the pair $b \to d$ and $d \to b$ with the single edge $b—d$, i.e. via the coercion as(A,"adjacencyMatrix").

The suite of methods available for representation objects includes the following methods: the names method to retrieve vertex names, together with the replacement method names<- to set them; the card method, returning as a named list the number of vertices and the total number of edge occurrences; the isEmpty method, answering the question whether there are no vertices in the represented graph; the isPresent method, checking whether a given edge is present in the represented graph; the areTheSame method, telling whether two objects (of the same class) represent the same graph; the [method, to extract representations for induced subgraphs, and the [[method to extract individual vertex names. In addition, the +, - and * operators have been overloaded so that representation + vertexSet adds isolated vertices, representation - vertexSet removes vertices and drops edges containing them, representation * vertexSet gets an induced subgraph, representation ± edge adds/removes an edge (occurrence). Note that here representation stands for any of the four possible representation classes. Special care is needed when adding an edge to an adjacencyMatrix object, since a->b + b->a == a--b and a->b + a--b == a--b.

3 Graph classes and methods

Package giRaph provides four high level classes: anyGraph for any graph fitting our abstract definition (vertex set plus multiset of edges), generalGraph for general graphs (Definition 10), multiGraph for MGs (Definition 11), and simpleGraph for SGs (Definition 12). Objects of class anyGraph consist of a single slot of class incidenceList, as this is the only possible representation if no assumptions are made on the edges. Objects of class generalGraph have a second slot of class incidenceMatrix, handling proper directed and undirected edges. Objects of class multiGraph have a third slot of class adjacencyList, and objects of class simpleGraph a fourth slot of class adjacencyMatrix. Each class inherits from the preceding one, so that for instance an SG can be seen as an MG, if no specific method for SGs is available, reflecting the fact that SGs form a subfamily of MGs.

Alternative representations for graphs of a given family are handled transparently with respect to the user, although the latter retains full control. The idea is that one ore more non-empty representations are present at any time, unless the graph is empty (in which case all representations are empty) and all of them represent the same graph. Therefore, methods are entitled to retrieve information from any non-empty slot and required to set information in all non-empty slots. In this way, the suite of methods available for representation classes is also made available for graph classes. The different representations can be set and retrieved via the following straightforward syntax, where `gg` is an existing (possibly empty) `generalGraph` object:

```
> incidenceList(gg) <- G # the G of Table 1
> areTheSame(I, incidenceMatrix(gg)) # the I of Table 2 (left)
[1] TRUE
```

By default the replacement method `incidenceMatrix<-` empties all slots except the one that is set. This is controlled by the optional argument `force`, which is `TRUE` by default, meaning that the representation should be set even if this amounts to changing the graph. The same happens for all other representations. If force is `FALSE`, then the representation is set only if it represents the same graph, in which case all other non-empty representations are preserved. Thus, alternative representations can coexist.

References

AHUJA, R.K., MAGNATI, T.L. and ORLIN, J.B. (1993): *Network Flows*. Prentice-Hall, Upper Saddle River.

BADSBERG, J.H. (2007): *dynamicGraph: dynamicGraph*. R package version 0.2.2.4, CRAN.

BADSBERG, J.H., DETHLEFSEN, C. and LA ROCCA, L. (2007): *giRaph: The giRaph package for graph representation in R*. R package version 0.0.1.3, CRAN.

BURNS, P.J., EFTHYMIOU, N. and DETHLEFSEN, C. (2007): *mathgraph: Directed and undirected graphs*. R package version 0.9-8, CRAN.

CHAMBERS, J.M. (1998): *Programming with Data*. Springer, New York.

CSARDI, G. (2007). *igraph: Routines for simple graphs, network analysis*. R package version 0.4.4, CRAN.

GALLO, G., LONGO, G. and PALLOTTINO, S. (1993): Directed hypergraphs and applications. *Discrete Applied Mathematics 42 (2–3) 177–201*.

GENTLEMAN, R., WHALEN, E., HUBER, W. and FALCON, S. (2007): *graph: A package to handle graph data structures*. R package version 1.15.6, CRAN.

GOMES, M.C., CAVIQUE, L and THEMIDO, I. (2006): The crew timetabling problem: An extension of the crew scheduling problem. *Annals of Operations Research 144 (1) 111–132*.

LAURITZEN, S.L. (1996): *Graphical Models*. Oxford University Press, Oxford.

LAURITZEN, S.L. (2002): gRaphical models in R: A new initiative within the R project. *R News 2 (3) 39*.

R DEVELOPMENT CORE TEAM (2007): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Part XII

Image and Signal Processing

A Method of Trend Extraction Using Singular Spectrum Analysis

Theodore Alexandrov[1]

Center for Industrial Mathematics, University of Bremen
Bibliothekstr. 1, 28359 Bremen, Germany, *theodore@math.uni-bremen.de*

Abstract. The paper presents a method of trend extraction in the framework of the Singular Spectrum Analysis (SSA) approach. This method is easy to use, does not need specification of models of time series and trend, allows to extract trend in the presence of noise and oscillations and has only two parameters (besides basic SSA parameter called window length). One parameter manages scale of the extracted trend and other is a method specific threshold value. We propose strategies for their choice. The presented method is evaluated on the seasonally adjusted monthly data of unemployment level in Alaska for the period 1976/01-2006/09.

Keywords: time series, trend extraction, Singular Spectrum Analysis

1 Introduction

Trend extraction is an important task in applied time series analysis, in particular in economics and engineering. We present a new method of trend extraction in the framework of the Singular Spectrum Analysis approach.

Trend is usually defined as a smooth additive component containing information about time series global change. This definition is rather vague (which type of smoothness is used? which kind of information is contained in the trend?). It may sound strange, but there is no more specific definition of the trend accepted by the majority of researchers and practitioners. Each approach to trend extraction defines trend with respect to the used mathematical tools (e.g. using Fourier transformation or derivatives). Thus in the corresponding literature one can meet various specific definitions of the trend.

Singular Spectrum Analysis (SSA) is a general approach to time series analysis and forecast. Algorithm of SSA is similar to those of Principal Components Analysis (PCA) of multivariate data. On the contrary SSA is applied to time series and provides representation of the given time series in the form of eigenvalues and eigenvectors of a matrix made of the time series. SSA has originated in diverse fields independently. Firstly, it was invented for attractor reconstruction in dynamical systems, see historical review in Ghil et al. (2002). Parallel, the so-called Caterpillar approach, based on PCA, was developed by a group of statisticians in Russia, see Golyandina et al. (2001) and the references therein. SSA and the Caterpillar approach appeared to be very close to each other. However, Caterpillar ap-

proach is probably less familiar to American and European statisticians for being poorly presented in top-level statistical journals.

SSA can be used for a wide range of tasks: trend or quasi-periodic component detection and extraction, denoising, forecasting, change-point detection. The present bibliography on SSA includes two monographes, several book chapters, and over a hundred papers. For more details see references at the website SSAwiki: http://www.math.uni-bremen.de/∼theodore/ssawiki/.

The method presented in this paper has been firstly proposed in Alexandrov and Golyandina (2005) and is studied in details in the author's unpublished Ph.D. thesis (Alexandrov (2006)) available only in Russian (at http://www.pdmi.ras.ru/∼theo/autossa/). We do not cite these works but only refer to the latter for proofs since it is impossible to reproduce the proofs here. We apologize for any inconvenience this may cause.

The proposed method is easy to use (has only two parameters), does not need specification of models of time series and trend, allows to specify desired trend scale, and extracts trend in the presence of noise and oscillations.

The outline of this paper is as follows. Section 2 introduces SSA. Section 3 describes properties of trends in SSA. In section 4, we propose the method of trend extraction and, in section 5, the strategies for selecting its parameters are introduced. In section 6, the method is demonstrated on a simulated time series with polynomial trend and on the unemployment level in Alaska.

2 SSA

Let us have a time series $X = (x_0, \ldots, x_{N-1})$, $x_n \in \mathbb{R}$, of length N, and we are looking for some specific additive component of X (e.g. trend). The central idea of SSA is to embed X into high-dimensional euclidean space, then find a subspace corresponding to the sought-for component and, finally, reconstruct the time series component corresponding to this subspace. The choice of the subspace is a crucial question in SSA. The base SSA algorithm has one parameter, the window length L $(1 < L < N)$ and consists of decomposition of a time series and reconstruction of a desired additive component. For the detailed description, see page 16 of Golyandina et al. (2001).

Decomposition. The decomposition starts with constructing the so-called trajectory matrix $\mathbf{X} \in \mathbb{R}^{L \times K}$, $K = N - L + 1$, with stepwise taken portions of the original time series X as columns:

$$X = (x_0, \ldots, x_{N-1}) \to \mathbf{X} = [C_1 : \ldots : C_K], \quad C_j = (x_{j-1}, \ldots, x_{j+L-2})^{\mathrm{T}}. \quad (1)$$

Note that $\mathbf{X}$ is a Hankel matrix and (1) defines one-to-one correspondence between series of length N and Hankel matrices of size $L \times K$. Then Singular Value Decomposition (SVD) of $\mathbf{X}$ is applied, where j'th component of SVD is specified by j'th eigenvalue λ_j and eigenvector U_j of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$:

$$\mathbf{X} = \sum_{j=1}^{L} \sqrt{\lambda_j} U_j {V_j}^{\mathrm{T}}, \quad V_j = \mathbf{X}^{\mathrm{T}} U_j \Big/ \sqrt{\lambda_j}.$$

The SVD components are numbered in the decreasing order of eigenvalues λ_j. We define j'th Empirical Orthogonal Function (EOF) as the sequence of elements of the j'th eigenvector U_j.

Reconstruction. The reconstruction stage combines (i) selection of a subgroup $\mathcal{J} \subset \{1, \ldots, L\}$ of SVD components; (ii) hankelization (averaging along entries with indices $i + j = const$) of the matrix formed from the truncated SVD; (iii) reconstruction of a time series component from the Hankel matrix by the mentioned one-to-one correspondence (like in (1) but in the reverse direction):

$$\mathbf{X}_{\mathcal{J}} = \sum_{j \in \mathcal{J}} \sqrt{\lambda_j} U_j {V_j}^{\mathrm{T}} \xrightarrow{\text{hankelization}} \widetilde{\mathbf{X}}_{\mathcal{J}} \in \mathbb{R}^{L \times K} \rightarrow Z = (y_0, \ldots, y_{N-1}).$$

The trend extraction problem is reduced to (i) the choice of window length L and (ii) the selection of the subgroup $\mathcal{J}$ of SVD components used for reconstruction. The first problem is thoroughly discussed in section 1.6 of Golyandina et al. (2001). In this paper, we propose a solution for the latter problem.

Note that for the reconstruction of a time series additive component, SSA considers the whole time series, as its algorithm uses SVD of the trajectory matrix built from all parts of the time series. Thus, SSA is not a local method in contrast to the linear filtering or wavelet methods. On the other hand, this property makes SSA robust to outliers.

An essential disadvantage of SSA is its computational complexity for the calculation of SVD. This shortcoming can be reduced by using parallel algorithms for SVD. For trend revision in case of receiving new data points, algorithms for updating SVD can be used.

3 Trend in SSA

SSA is a nonparametric approach which does not need a priori specification of models of time series and trend, neither deterministic nor stochastic ones. The classes of trends and residuals which can be successfully separated by SSA are characterized as follows.

First, since we extract any trend by selecting a subgroup of all L SVD components, this trend should generate only d of them (for some $d < L$). For infinite time series, a class of such trends coincides with the class of time series governed by finite difference equations, see Golyandina et al. (2001). This class can be described explicitly as linear combinations of products of polynomials, exponentials and sines, see Buchstaber (1995). An element of this class suits well for representation of a smooth and slowly varying trend.

Second, a residual should belong to the class of time series which can be separated from a trend. The separability theory due to Nekrutkin (1996) allows to determine this class and postulates that (i) any deterministic function can be asymptotically separated from any ergodic stochastic noise as the time

series length and window length tend to infinity; (ii) under some conditions any trend can be separated from any quasi-periodic component, see Golyandina et al. (2001). These properties of SSA allow to apply this approach to trend extraction in the presence of noise and quasi-periodic components.

Finally, as trend is a smooth and slow varying component, it generates SVD components with smooth and slow varying EOFs. Eigenvectors represent an orthonormal basis of trajectory vector space spanned on the columns of trajectory matrix. Thus each EOF is a linear combination of portions of the corresponding time series and inherits its global smoothness properties. This idea is considered in details in Golyandina et al. (2001) where the cases of polynomial and exponential trends are thoroughly examined.

3.1 Present methods of trend extraction in SSA

A naive approach to trend extraction in SSA is to reconstruct trend from several first SVD components. Despite of its simplicity, this approach works in many real-life cases. An eigenvalue represents contribution of the corresponding SVD component into the form of the time series, see section 1.6 of Golyandina et al. (2001). Since trend usually characterizes time series, its eigenvalues are larger than the other ones, that implies small order numbers of the trend SVD components. However, the selection procedure fails when the values of the trend are small enough as compared with the residual, or when the trend has complicated structure (e.g. a high-order polynomial) and is characterized by many (not only the first ones) SVD components.

A smarter way of selection of trend SVD components is to choose the components with smooth and slowly varying EOFs (we have explained this fact above). At present, there exist only one parametric method which can be referred to as following this approach, see Vautard et al. (1992). Vautard et al. (1992) proposed using the Kendall correlation coefficient for testing for monotonic growth of an EOF. Unfortunately, this method is problematic since it is not clear which trends can be extracted by its means. Certainly, this method extracts monotonic trends as their EOFs are typically monotonic. However, in general, nonmonotonic polynomials of low order can have monotonic EOFs, see e.g. Golyandina et al. (2001) for details.

4 Proposed method

In this section, we present our method of trend extraction. Firstly we need to introduce the periodogram $I_Y^M(\omega)$ of a vector $Z \in \mathbb{R}^M$, $Z = (z_0, \ldots, z_{M-1})^{\mathrm{T}}$:

$$I_Z^M(\omega) = \frac{1}{M}\left|\sum_{n=0}^{M-1} e^{-i2\pi\omega n} z_n\right|^2, \quad \omega \in \{k/M\}_{k=0}^{\lfloor M/2 \rfloor},$$

which can be treated as the contribution of the frequency ω. The cumulative contribution is evaluated as $\pi_Z^M(\omega) = \sum_{k:0\le k/M\le\omega} I_Z^M(k/M)$, $\omega \in [0, 0.5]$.

Given $\omega_0 \in (0, 0.5)$, the contribution of low frequencies $[0, \omega_0]$ to $Z \in \mathbb{R}^M$ is defined as

$$\mathcal{C}(Z, \omega_0) = \pi_Z^M(\omega_0)/\pi_Z^M(0.5).$$

Then, given parameters $\omega_0 \in (0, 0.5)$ and $\mathcal{C}_0 \in [0, 1]$, we propose to select those SVD components whose eigenvectors satisfy the following criterion:

$$\mathcal{C}(U_j, \omega_0) \geq \mathcal{C}_0,$$

where U_j is the corresponding j'th eigenvector. One may interpret this method as selection of SVD components with EOFs characterized mostly by low-frequency fluctuations.

5 Choice of the parameters

Low-frequency boundary ω_0. The low-frequency boundary ω_0 manages the scale of the extracted trend: the lower is ω_0, the slower varies the extracted trend. One can prespecify the desired scale using ω_0 but we also present instructions for the choice of ω_0 based on the information about the given time series.

Firstly, if we assume to have a quasi-periodic component with known period T, then $\omega_0 < 1/T$. For extraction of trend from monthly data with possible seasonal oscillations (of period 12), we suggest $\omega_0 = 0.075 < 1/12$.

Secondly, a reasonable choice of ω_0 can be carried out by examination of periodogram of the original time series. To explain this approach, we give the following facts referring to page 31 of Alexandrov (2006) for the proofs.

Proposition 1. *Let us have two time series* $G = (g_0, \ldots, g_{N-1})$ *and* $H = (h_0, \ldots, h_{N-1})$ *of length* N*, then for each* k*:* $0 \leq k \leq \lfloor N/2 \rfloor$*, holds:*

$$\left| I_{G+H}^N(k/N) - I_G^N(k/N) - I_H^N(k/N) \right| \leq 2\sqrt{I_G^N(k/N) I_H^N(k/N)}.$$

Corollary 2. *For two time series* G *and* H *whose periodogram supports are nearly disjoint, the periodogram of their sum* $G + H$ *is close to the sum of their periodograms.*

In many applications, the given time series can be modelled as made of trend with large periodogram values at low-frequency interval $[0, \omega_0]$, oscillations with periods smaller than $1/\omega_0$, and noise whose frequency contribution spread over all the frequencies $[0, 0.5]$ but is relatively small. Therefore periodogram supports of the trend and the residual can be considered as nearly disjoint. Examining the trend periodogram, we can guess ω_0 as a value bounding the interval of large periodogram values close to zero frequency.

Low-frequency contribution $\mathcal{C}_0$. For choosing the second parameter $\mathcal{C}_0$, we propose the following heuristic procedure. Given the time series X and its trend $T(\omega_0, \mathcal{C}_0)$ extracted with some ω_0 and $\mathcal{C}_0$, we define the normalized contribution of low-frequency oscillations in the residual as:

$$\mathcal{R}_{X,\omega_0}(\mathcal{C}_0) = \mathcal{C}\big(X - T(\omega_0, \mathcal{C}_0), \omega_0\big)\mathcal{C}(X, \omega_0)^{-1}.$$

As discussed, trend EOFs have varies slowly. As easily shown on page 47 of Alexandrov (2006), this property is inherited by elementary reconstructed components, each reconstructed from one SVD component. Having trend in the original time series, we expect that only the elementary components corresponding to the trend have large contribution of low frequencies. Thus, the maximal values of $\mathcal{C}_0$ which lead to selection of trend-corresponding SVD components should generate jumps of $\mathcal{R}_{X,\omega_0}(\mathcal{C}_0)$.

Based on this idea, we propose the following way for choosing $\mathcal{C}_0$:

$$\mathcal{C}_0^{\mathcal{R}} = \min\big\{\mathcal{C}_0 \in [0,1] : \quad \mathcal{R}_{X,\omega_0}(\mathcal{C}_0 + \Delta\mathcal{C}) - \mathcal{R}_{X,\omega_0}(\mathcal{C}_0) \geq \Delta\mathcal{R}\big\}, \qquad (2)$$

where $\Delta\mathcal{C}$ is a search step and $\Delta\mathcal{R}$ is the given threshold. On the one hand, this strategy is heuristic and requires selection of $\Delta\mathcal{R}$, but on the other hand, the simulation results and application to different time series showed its ability to choose reasonable $\mathcal{C}_0$ in many cases with prespecified $\Delta\mathcal{R}$. This experience allows us to suggest using $0.05 \leq \Delta\mathcal{R} \leq 0.1$. The step $\Delta\mathcal{C}$ is to be chosen as small as possible to discriminate identifications occurring at different values of $\mathcal{C}_0$. To reduce computational time, we commonly take $\Delta\mathcal{C} \geq 0.01$ and suggest a default value of $\Delta\mathcal{C} = 0.01$.

6 Examples

Simulated example with polynomial trend. The first example illustrates the choice of parameters ω_0 and $\mathcal{C}_0$. We simulated a time series of length $N = 300$, shown in Figure 1, containing a polynomial trend, exponentially-modulated sine wave, and white gaussian noise, whose n'th element is expressed as $x_n = 10^{-11}(n-10)(n-70)(n-160)^2(n-290)^2 + \exp(0.01n)\sin(2\pi n/12) + \varepsilon_n$, ε_n is $iidN(0, 5^2)$. The period of the sine wave is assumed to be unknown.

We have chosen the window length $L = N/2 = 150$ for achieving better separability of trend and residual. The value $\omega_0 = 0.02$ was selected according to section 5, as the large periodogram values next to zero frequency are concentrated in $[0, 0.02]$. The search of $\mathcal{C}_0$ using (2) has been done with step $\Delta\mathcal{C} = 0.01$ and $\Delta\mathcal{R} = 0.05$. As shown in Figure 1, despite of the strong noise and oscillations, the extracted trend approximates the original one very well. The achieved mean square error is 0.79. For example, the ideal low pass filter with cutoff frequency 0.02 produced the error of 3.14. This superiority is achieved mostly due to better approximation at the first and last 50 points of the time series. All the calculations were performed using our Matlab-based software AutoSSA available at http://www.pdmi.ras.ru/∼theo/autossa.

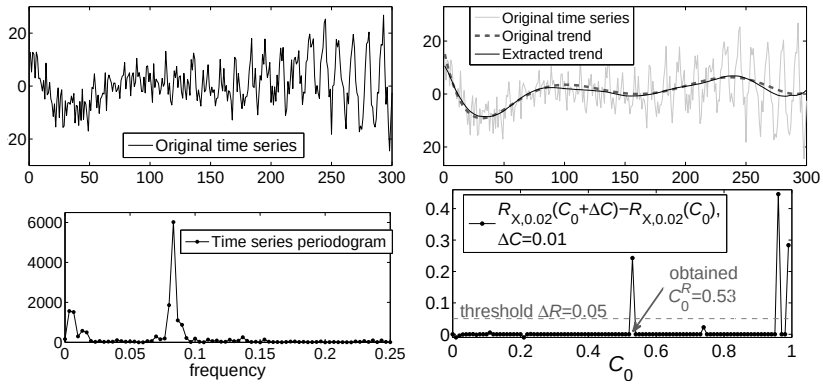


Fig. 1. Simulated example with polynomial trend: original time series, original and extracted trends ($L = 180$, $\Delta\mathcal{C} = 0.01$, and $\Delta\mathcal{R} = 0.05$), zoomed time series periodogram (inside $\omega \in [0, 0.25]$), the values of $\mathcal{R}_{X,\omega_0}(\mathcal{C}_0 + \Delta\mathcal{C}) - \mathcal{R}_{X,\omega_0}(\mathcal{C}_0)$ used for the choice of $\mathcal{C}_0$ (resulted in a value $\mathcal{C}_0^{\mathcal{R}} = 0.53$).

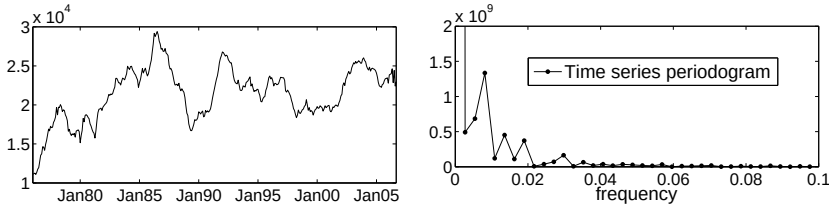


Fig. 2. Unemployment level in Alaska: original data, zoomed periodogram.

Trends of the unemployment level. Let us demonstrate extraction of trends of different scale. We consider the unemployment level (unemployed persons) in Alaska for the period 1976/01-2006/09 (monthly data, seasonally adjusted), provided by the Bureau of Labor Statistics at http://www.bls.gov under the identifier LASST02000004, see Figure 2. This time series is typical for economical applications, where data contain relatively little noise and are subject to abrupt changes. Economists are often interested in the "short" term trend which includes cyclical fluctuations and is referred to as trend-cycle.

The length of the data is $N = 369$. For achieving better separability of trend and residual we selected L close to $N/2$ but divisible by the period $T = 12$ of probably existing seasonal oscillations: $L = 12\lfloor N/24 \rfloor = 180$.

Having considered the periodogram of the data, see Figure 2, we extracted trends of different scale with the following ω_0: 0.01, 0.02, 0.05, and 0.075, see Figure 3. The value 0.02 was selected according to section 5, as the periodogram is mostly concentrated inside the interval $[0, 0.02]$. The value 0.075 is the default value for monthly data. Other values (0.01 and 0.05) were considered for better illustration of how the value of ω_0 influences the scale of the extracted trend. The search of $\mathcal{C}_0$ was performed as described in section 5 in interval $[0.5, 1]$ with step $\Delta\mathcal{C} = 0.01$ and $\Delta\mathcal{R} = 0.05$.

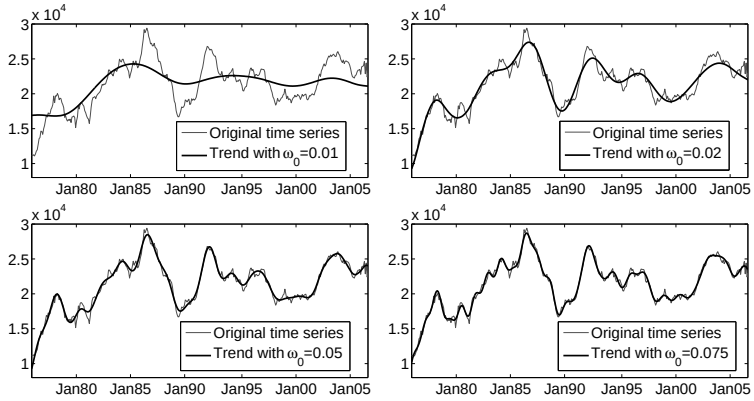


Fig. 3. Unemployment level in Alaska: extracted trends of different resolution with $\omega_0 = 0.01, 0.02, 0.05$, and 0.075 ($L = 180$, $\Delta\mathcal{C} = 0.01$, and $\Delta\mathcal{R} = 0.05$).

7 Conclusions

SSA is an attractive approach to trend extraction because it: (i) requires no model specification of time series and trend, (ii) extracts trend of noisy time series containing oscillations of unknown period, (iii) is robust to outliers. In this paper, we presented a method which inherits these properties and easy to use since it requires selection of only two parameters.

Acknowledgements. The helpful comments of two anonymous reviewers are gratefully acknowledged.

References

ALEXANDROV, T. (2006): Software package for automatic extraction and forecast of additive components of time series in the framework of the Caterpillar-SSA approach. PhD thesis, St.Petersburg State University.

ALEXANDROV, T. and GOLYANDINA, N. (2005): Thresholds for methods of automatic extraction of time series trend and periodical components with the help of the "Caterpillar"-SSA approach. In: *Proc. 4th Conf. System Identification and Control Problems*. Inst. of Control Sci., Moscow, 1849–1864.

BUCHSTABER, V. M. (1995): Time series analysis and grassmannians. *Transactions of the American Mathematical Society 162, 1–17.*

GHIL, M., ALLEN, R. M., DETTINGER, M. D., IDE, K., KONDRASHOV, D., MANN, M. E., ROBERTSON, A., SAUNDERS, A., TIAN, Y., VARADI, F. and YIOU, P. (2002): Advanced spectral methods for climatic time series. *Reviews of Geophysics, 40 (1), 1–41.*

GOLYANDINA, N., NEKRUTKIN, V. and ZHIGLJAVSKY, A. (2001): *Analysis of time series structure: SSA and related techniques*. Chapman&Hall/CRC.

NEKRUTKIN, V. (1996): Theoretical properties of the "Caterpillar" method of time series analysis. In: *Proc. 8th IEEE Signal Proc. Workshop on Stat. Signal and Array Processing*. IEEE Computer Society, Washington, DC, 395–397.

VAUTARD, R., YIOU, P. and GHIL, M. (1992): Singular-spectrum analysis: a toolkit for short, noisy chaotic signals. *Physica D 58, 95–126.*

QRS Complex Boundaries Location for Multilead Electrocardiogram

Rute Almeida[1], Juan Pablo Martínez[1], Ana Paula Rocha[2], and Pablo Laguna[1]

[1] Communications Technology Group, Aragón Institute of Engineering Research (I3A), University of Zaragoza, María de Luna 1, 50018 Zaragoza, CIBER-BBN, Spain *e-mail: {rbalmeid, jpmart, laguna}@unizar.es*

[2] Departamento de Matemática Aplicada, Faculdade de Ciências, Universidade do Porto and Centro de Matemática da Universidade do Porto (CMUP), Rua Campo Alegre 687, 4169-007 Porto, Portugal *e-mail: aprocha@fc.up.pt*

Abstract. In this paper a multilead methodology regarding QRS complex boundaries location is proposed and validated. It was established from a single-lead based system previously developed and attends to the spatial characteristics of the different leads, aiming to achieve a more robust delineation. It provides more robust and accurate boundaries locations than any electrocardiographic lead by itself.

Keywords: ECG wave delineation, multilead, wavelets

1 Introduction

The electrocardiogram (ECG) is the record of the cardiac electrical activity as a function of time, by means of electrodes placed on the skin. It is a noninvasive and painless procedure and an indispensable diagnostic tool for many cardiac and non cardiac conditions. By using several electrodes it is possible to access simultaneous recording directions, known as *electrocardiographic leads*, providing a spacial perspective. Each heart beat is produced by an electric wavefront that crosses the different cardiac structures; the activation/inactivation of those correspond to different waves in the ECG, known as P wave, Q, R and S waves (QRS complex) and T wave. In particular, the waves in the QRS complex reflect the activation of both ventricles. In spite of the general characteristics (as smoothness and relative polarity), the waves' morphology depends on several factors, especially on the recorded lead. ECG delineation consists on detecting peaks and boundaries (onset and end) of those waves and provides fundamental features to derive clinically useful information, namely about the duration of the phenomena and their beat-to-beat evolution. As there are not standard clear rules to locate the waves' boundaries, systematizing the delineation is a difficult task. Clinical ECG often present relevant levels of noise that mask the signal information.

This group has previously proposed an automatic single-lead (SL) delineation system (Martínez, J. P. et al (2004)), that generalizes the wavelet

transform (WT) based methodology of Li, C. et al (1995). The WT provides a description of the signal in the time-scale domain, allowing the representation of its temporal features at different resolutions according to their frequency content. Thus, regarding the purpose of locating different waves with typical frequency characteristics, the WT is a suitable tool for ECG delineation.

Each lead is characterized by a lead vector giving the direction from one electrode to the other. According to the dipole hypothesis, the electrical activity of the heart can be approximated by a time-variant electrical dipole, called the *electrical heart vector* (EHV). Thus, the voltage measured at a given lead is merely the projection of the EHV into the unitary vector defined by the lead axis (Malmivuo, J. and Plonsey, R. (1995)). Nevertheless, the lead set most widely used in clinical practice is not an orthogonal system, but rather the somewhat redundant standard 12-lead system, considered to contain 8 truly independent leads describing dipolar and non dipolar components. A set of linear transformations between the 12-lead and the most used orthogonal system, the Frank leads (X, Y, Z) is given by the **Dower matrix** (Dower, G. E. (1984)). This quite old linear transformation has been object of many criticisms but no wide accepted alternative has been proposed yet.

Using a particular lead for ECG delineation determines a point of view over the cardiac phenomena, thus different latencies on the waves's onsets and ends are found in different leads. Combining adequately the information provided by multiple leads is essential for the correct location of lead-independent waves' boundaries. The SL system in Martínez, J. P. et al (2004) includes post-processing decision rules to deal with multilead records, by choosing global marks based on the single-lead derived locations. Nevertheless, in spite of the satisfactory performance, this system is not truly multilead and it requires to apply SL delineation to each one of the leads.

In this paper is proposed and validated an actually multilead (ML) methodology regarding QRS complex boundaries location. The ML approach was established from the SL system and attends to the spatial characteristics of the different available leads, aiming to achieve a more robust delineation.

2 Methods

2.1 Single-lead based delineation

The SL based delineation system is described in detail in Martínez, J. P. et al (2004) and only general features are here referred. The detection of the fiducial points is carried out across the adequate WT scales, attending to the dominant frequency components of each ECG wave. The prototype wavelet used allows to obtain a WT at scale 2^m, $w_{x,m}[n]$, proportional to the derivative of the version of the digitalized signal $x[n]$ filtered with a smoothing impulse response at scale 2^m. Thus, ECG wave peaks correspond to zero crossings in the WT and ECG maximum slopes correspond to WT's maxima and minima. The onset [end] of a wave (n_o [n_e]), occurs before [after] the

first [last] maximum of $|w_{x,m}[n]|$, at sample n_f $[n_l]$. Boundaries are located by selecting the sample nearest to n_f $[n_l]$ satisfying a threshold based criteria.

To deal with multiple leads a robust post-processing decision rule over SL derived locations (SLR) is used: the SL annotations are ordered and the onset [end] of a wave is selected as the first [last] annotation whose 3 nearest neighbours lay within a δ ms interval with $\delta = 10$ ms for QRS end and $\delta = 12$ ms for QRS onset.

2.2 Multilead System

The ML delineation system proposed considers three simultaneous orthogonal leads (X,Y,Z), taking advantage of the spacial information by them represented. For a scale $2^m\big|_{m\in\{1,2,3,\ldots\}}$, a spatial WT *loop* can be defined as

$$\mathbf{w}_m[n] = [w_{x,m}[n], w_{y,m}[n], w_{z,m}[n]]^T. \tag{1}$$

The WT prototype used produces a WT loop $\mathbf{w}_m[n]$ proportional to the ECG derivative and describes the EHV evolution. Therefore, the director vector of the best straight line fit to all points in $\mathbf{w}_m[n]$, $n \in W$ gives the main direction $\mathbf{u} = [u_X, u_Y, u_Z]^T$ of EHV variations in a scale 2^m for a time interval of interest W.

Considering the ECG loop $[x[n], y[n], z[n]]^T$, a *generated* lead $d[n]$ defined by axis $\mathbf{u}$ and combining the information provided by the 3 leads, can be obtained by projecting over $\mathbf{u}$ the points of the ECG loop defined on an extended interval containing one beat. Instead, the WT loop $\mathbf{w}_m[n]$ can be projected and the WT of the *derived* signal, $w_{d,m}[n]$, obtained .

The strategy proposed for ML boundary delineation using WT loops is based in a multi-step iterative search for an improved spatial lead for delineation (with *steepest* slopes). Multilead location of the QRS boundaries is performed as illustrated for QRS onset in Figure 1. Let's define $n^{(0)}_{\mathrm{QRS,o}}$ $\left[n^{(0)}_{\mathrm{QRS,e}}\right]$ as the earliest [latest] QRS onset [end] location given by the SL methods (over each orthogonal lead) and $n^{(0)}_{\mathrm{QRS,f}}$ $\left[n^{(0)}_{\mathrm{QRS,l}}\right]$ is the earliest [latest] significant maximum modulus location.

The multilead delineation strategy for QRS boundaries is described by the following algorithm, which for each beat and boundary, consists in an initialization and a variable number of iterations:

INITIALIZATION

a_0) an initial search window adequate to find the EHV's main direction in the boundary is defined respectively for QRS onset and end, as

$$Q_{[1]} = [n^{(0)}_{\mathrm{QRS,o}} - 4s_{\mathrm{CSE}}(\mathrm{QRS}_{on}),\ n^{(0)}_{\mathrm{QRS,f}}];\ \ S^{(1)} = [n^{(0)}_{\mathrm{QRS,l}}\ ,\ n^{(0)}_{\mathrm{QRS,e}} + 4s_{\mathrm{CSE}}(\mathrm{QRS}_{end})]$$

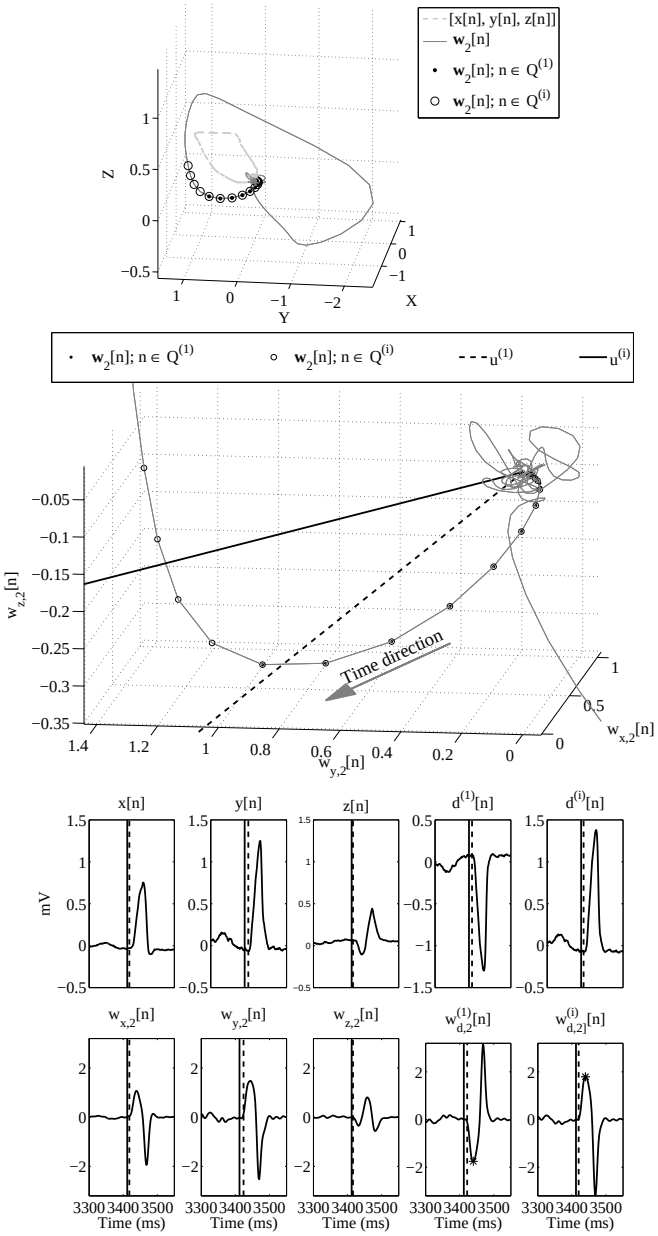


Fig. 1. Example of ML delineation (file from CSE database): initial and final steps. Upper panel: ECG and WT loops. Middle panel:WT loops and the directions of the best line fit. Lower panel: ECG in orthogonal leads, WT signals, derived ECG and WT signals, delineation mark found in the respective lead (vertical dashed lines), *median referee* marks (solid line) and first significant maximum modulus in the constructed lead (stars). ECG in mV and final step $i = 2$.

where $s_{\mathrm{CSE}}(\mathrm{QRS}_{on}) = \frac{6.5}{2} f_s$ samples $[s_{\mathrm{CSE}}(\mathrm{QRS}_{end}) = \frac{11.6}{2} f_s$ samples$]$ correspond to the standard deviation tolerance values provided by The CSE Working Party (1985), with f_s the sampling frequency;

b_0) the initial main direction of EHV variations $\mathbf{u^{(1)}}$ is estimated as the best line fit in total least squares (TLS) sense to $\mathbf{w}_m[n]\,|_{n\in Q^{(1)}}$ or $\mathbf{w}_m[n]\,|_{n\in S^{(1)}}$;

c_0) the loop $\mathbf{w}_m[n]\Big|_{n\in\left[n^{(0)}_{\mathrm{QRS},k-1}\,,\,n^{(0)}_{\mathrm{QRS},k+1}\right]}$, for $n^{(0)}_{\mathrm{QRS},k}$ the median of SL derived locations for the QRS complex in the k^{th} beat, is projected over $\mathbf{u}^{(1)}$ to construct the new derived WT signal $w^{(1)}_{d,m}[n]$;

d_0) SL delineation is performed over $w^{(1)}_{d,m}[n]$ to locate $n^{(1)}_{\mathrm{QRS,o}}$ or $n^{(1)}_{\mathrm{QRS,e}}$.

ITERATION - STEP (i)

a) the search window is updated as

$Q^{(i)} = [n^{(i-1)}_{\mathrm{QRS,o}} - 4s_{\mathrm{CSE}}(\mathrm{QRS}_{on}),\ n^{(i-1)}_{\mathrm{QRS,f}}];\ S^{(i)} = [n^{(i-1)}_{\mathrm{QRS,l}}\ ,\ n^{(i-1)}_{\mathrm{QRS,e}} + 4s_{\mathrm{CSE}}(\mathrm{QRS}_{end})]$

where $n^{(i-1)}_{\mathrm{QRS,o}}$ $\left[n^{(i-1)}_{\mathrm{QRS,e}}\right]$ is the QRS onset [end] position found in the step $(i-1)$ and $n^{(i-1)}_{\mathrm{QRS,f}}$ $\left[n^{(i-1)}_{\mathrm{QRS,l}}\right]$ is the location of the first [last] significant maximum modulus of $w^{(i-1)}_{d,m}[n]$;

b) the main direction of EHV variations $\mathbf{u}^{(i)}$ is estimated as the TLS best line fit to $\mathbf{w}_m[n]\,|_{n\in Q^{(i)}}$ or $\mathbf{w}_m[n]\,|_{n\in S^{(i)}}$;

c) the new derived WT signal $w_{d,m,[g]}[n]$ is constructed by projecting $\mathbf{w}_m[n]\Big|_{n\in\left[n^{(0)}_{\mathrm{QRS},k-1}\,,\,n^{(0)}_{\mathrm{QRS},k+1}\right]}$

d) IF $n^{(i)}_{\mathrm{QRS,f}}$ $\left[n^{(i)}_{\mathrm{QRS,l}}\right]$ has the same polarity than $n^{(i-1)}_{\mathrm{QRS,f}}$ $\left[n^{(i-1)}_{\mathrm{QRS,l}}\right]$, equal or lower amplitude and QRS complex morphology includes a Q [S] wave (the lead constructed at step (i) is not better for QRS onset [end] location than the constructed in the step $(i-1)$)
OR no significant maximum of $w^{(i)}_{d,m}[n]$ was found (the lead is not adequate for boundary location)
THEN $n^{(i-1)}_{\mathrm{QRS,o}}$ $\left[n^{(i-1)}_{\mathrm{QRS,e}}\right]$ is adopted as ML mark; STOP;
ELSE SL delineation of the boundary is performed over $w^{(i)}_{d,m}[n]$ to find $n^{(i)}_{\mathrm{QRS,o}}$ $\left[n^{(i)}_{\mathrm{QRS,e}}\right]$ updated marks;

e) IF the same location is achieved for 3 (possible nonconsecutive) iterations
THEN $n^{(i)}_{\mathrm{QRS,o}}$ $\left[n^{(i)}_{\mathrm{QRS,e}}\right]$ is adopted as ML mark; STOP;
ELSE REPEAT from a).

It must be remarked that the choice of *basing the lead direction* in the WT loop, instead of taking directly the ECG loop is relevant, as it allows to avoid the high frequency noise contamination and thus produces a more accurate selection.

3 Results and discussion

The evaluation of the automatic delineation strategies was performed over real files from the CSE multilead measurement database (CSEDB, Willems, J. L. et al (1987), 42 short signals in 15 leads at 500 Hz) which include referee marks for 32 QRS onsets and 26 QRS ends. The delineation error (ε) was taken as the *automatically detected boundary minus the respective referee mark* and the mean (m_ε) and standard deviation (s_ε) of ε were evaluated across files; the mean $\left(m_{|\varepsilon|}\right)$ and standard deviation $\left(s_{|\varepsilon|}\right)$ of the absolute error $|\varepsilon|$ were also calculated. Additionally, the above mentioned parameters were calculated after the exclusion of the 5% most *extreme cases* in each tail.

Different orthogonal lead systems were considered:
lead set F - defined by recorded orthogonal Frank leads (X,Y,Z);
lead set M - defined by leads *V5*, *aVF* and *V2*, a subset of 3 mutually orthogonal leads out of the standard 12-lead system;
lead set D - defined by the synthesised orthogonal leads from the standard 12-lead system, by using the coefficients provided by the *Dower Matrix*;
lead set PC1 - defined by the first 3 principal components calculated from the whole 12-lead signal;
lead set PC2 - defined by the first 3 principal components calculated from the 8 truly independent leads and based in the segment of interest QRS onset to T wave end according to SL delineation over lead II.

For the sake of comparison, SL was applied to each available lead and the post processing rules (SLR) described in Subsection 2.1 were applied over 12 or 15 leads. Results are presented in Figure 2 and 3.

It was found that a relative low number of extreme cases was causing a large fraction of the global error. The exclusion of the 10% more extreme measurements in each approach allowed a generalised improvement in the errors dispersion, with bias increase in some cases. In particular ML over lead sets F, PC1 and PC2 after the exclusion of the most extreme files performs closely as well as the best SL based delineation. Nevertheless still far from the error dispersion obtained using SLR over the 12 leads or all the 15 leads together. It should be remarked that the ML proposed method requires the WT calculation of 3 leads, with delineation procedures involving a variable number of signals. Thus, even considering fitting and projecting features, the ML strategy is likely to be more efficient than applying SL to 12 or 15 leads, as the number of iterations needed is not very high. Thus, the results presented denote a clear improvement in terms of computational complexity with comparable performance. The lead sets PC1 or PC2 are good alternatives to Dower matrix, since ML performs better over those than over D.

The bias found can be due to the referee annotation protocol itself. Referees were required to look for the earliest onset and latest end signs of the waves, in order to detect the whole electric activation/inactivation phenomena reflected in the QRS complex. This rule is risky, especially in automatic strategies, as it likely to be affected by outliers resulting from noise contami-

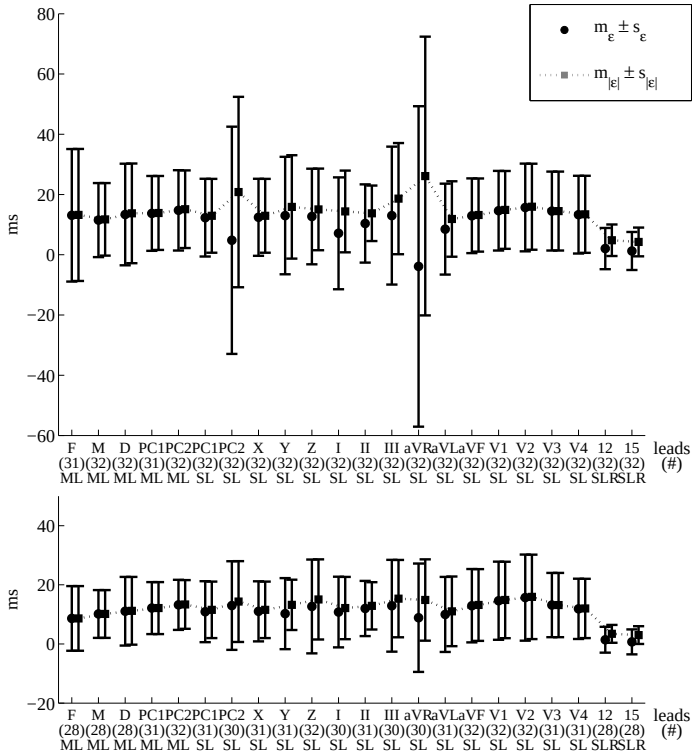


Fig. 2. Delineation results for QRS onset: upper panel corresponds to results in all true positive detections (# denotes the number of detections out of 32 reference marks provided), lower panel corresponds to results in after excluding 10% extreme cases in each approach (# denotes the number of beats considered).

nation. Furthermore, these early activation / late inactivation signs can reflect local properties not related to the whole myocardium. Thus, a more global rule, as the one proposed in this work, is better suited for global myocardium activation/inactivation in applications like evaluation drug cardiotoxicity or others where the effect of interest comes from the global myocardium.

In this work we focused on the problem of QRS boundaries location, that is, the delineation of the higher frequency component of the ECG. The boundaries of the waves P and T, which reflect lower components of the signal, can be also located by similar, although adapted, strategies.

4 Conclusions

A novel ML WT based automatic system for ECG boundaries delineation was here proposed and evaluated with respect to the QRS boundaries. The results pointed out that both SL and ML methodologies are adequate for ECG

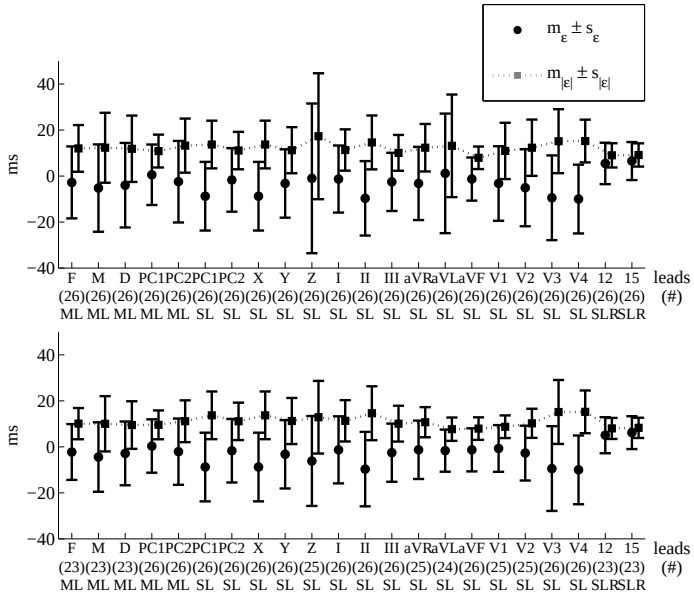


Fig. 3. Delineation results for QRS end: upper panel corresponds to results in all true positive detections (# denotes the number of detections out of 26 reference marks provided), lower panel corresponds to results in after excluding 10% extreme cases in each approach (# denotes the number of beats considered).

waves delineation. ML provided more robust and more accurate boundaries locations than any electrocardiographic lead by itself. ML over lead set PC1 or PC2 performs better than lead set D on QRS onset, being a good alternative to Dower matrix when Frank leads are not available.

References

DOWER, G.E. (1984): The ECGD: a derivation of the ECG from VCG leads. *J. Electrocardiol. 17 (2), 189-191.*

LI, C., ZHENG, C. and TAI, C. (1995): Detection of ECG Characteristic Points Using Wavelet Transforms. *IEEE Transactions on Biomedical Engineering 42 (1), 21-28.*

MALMIVUO, J. and PLONSEY, R. (1995): *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields.* Oxford University Press

MARTÍNEZ, J.P., ALMEIDA, R., OLMOS, S., ROCHA, A.P. and LAGUNA, P. (2004): Wavelet-based ECG delineator: evaluation on standard databases. *IEEE Transactions on Biomedical Engineering 51 (4), 570-581.*

The CSE Working Party (1985): Recomendations for measurement standards in quantitative electrocardiography. *Eur. Heart J. 6, 815-825.*

WILLEMS, J.L., ARNAUD, P., VAN BEMMEL and et al (1987) A reference data base for multilead electrocardiographic computer measurement programs. *J. Am. Coll. Cardiol. 10 (6), 1313-1321.*

On the Equivalence of the Weighted Least Squares and the Generalised Least Squares Estimators

Alessandra Luati[1] and Tommaso Proietti[2]

[1] University of Bologna, Department of Statistics,
via Belle Arti 41, 40126 Bologna, Italy, *alessandra.luati@unibo.it*
[2] University of Rome "Tor Vergata", S.E.F. e ME. Q.,
via Columbia 2, 00133 Roma, Italy, *tommaso.proietti@uniroma2.it*

Abstract. This paper is concerned with the equivalence of the weighted least squares estimators (WLSE) and the generalised least squares estimators (GLSE). Necessary and sufficient conditions for the WLSE to be best linear unbiased estimators are derived, generalising Anderson (1948, 1971) theorem on the equivalence between the ordinary least squares estimators and the GLSE. Procedures for obtaining the optimal kernel for a given covariance structure are also described, where optimality is to be intended in the Gauss-Markov sense. The results are illustrated in the context of local polynomial regression methods for the estimation of the underlying trend of a time series.

Keywords: Local polynomial regression, kernel smoothing, Epanechnikov kernel, Henderson filters, trend estimation

1 Introduction

Let us consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$, $p < n$. Throughout the paper we will assume that $\mathbf{X}$ is a deterministic matrix with full column rank and that the covariance matrix $\boldsymbol{\Sigma}$ is positive definite and non singular. We can relax both the assumption of normality and of deterministic regressors and replace it by the weak exogeneity assumption, $\mathrm{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \boldsymbol{\Sigma}$.

A well-known result (Aitken theorem) is that, if $\boldsymbol{\Sigma}$ is known, the best linear unbiased estimator (BLUE) of the regression parameters is the generalised least squares estimator (GLSE)

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}. \tag{2}$$

Much attention has been devoted in the literature to the search of conditions for which the ordinary least squares estimator (OLSE),

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{3}$$

is equivalent to the GLS estimator (2), and thus it is BLUE.

Anderson was the first who faced this problem, stating (1948, p. 48) and proving (1971, pp. 19 and 560) that equality between (2) and (3) holds if and only if there are p linear combinations of the columns of $\mathbf{X}$ that are eigenvectors of $\mathbf{\Sigma}$. The relevance of this result is self-evident, although Anderson's condition is not easy to verify in practice, i.e. for given matrices $\mathbf{X}$ and $\mathbf{\Sigma}$. Later developments in this field concerned the search of equivalent conditions for OLSE to be BLUE. A relevant contribution in this sense was that of Zyskind (1967), who derived eight equivalent conditions, among which the commutativity relation between the covariance matrix and the orthogonal projection matrix onto the column space of $\mathbf{X}$. An excellent and exhaustive review of these results and their further developments is that of Puntanen and Styan (1989).

This paper is concerned instead with establishing the conditions under which there exists a diagonal matrix $\mathbf{K}$ such that the GLSE is equivalent to the weighted least squares estimator (WLSE)

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}\mathbf{y}. \tag{4}$$

When these conditions are met, the diagonal elements of $\mathbf{K}$ provide the optimal kernel weights corresponding to a given covariance structure $\mathbf{\Sigma}$, where optimality is to be intended in the Gauss-Markov sense. The interest in this issue arises in local polynomial modelling. It turns out that the Epanechnikov kernel is the optimal kernel in local polynomial regression with strictly non-invertible first order moving average errors. Similarly, the Henderson kernel is optimal in the presence of non-invertible third order moving averages.

2 Main results

This section contains the main results of the paper. To state them, some additional notation is required. Let us denote by $\mathcal{C}(\mathbf{X})$ the column space of $\mathbf{X}$, also called its range, and by $\mathcal{N}(\mathbf{X})$ its null space. If $\mathbf{W} \in \mathbb{R}^{n\times n}$ and $\text{rank}(\mathbf{W}) = n$, then $\mathbf{H}_W = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ is the (oblique) projection matrix onto $\mathcal{C}(\mathbf{X})$ along $\mathcal{N}(\mathbf{X}'\mathbf{W})$. The subspaces $\mathcal{C}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X}'\mathbf{W})$ are complementary, in the sense that they have null intersection and their union is $\mathbb{R}^n$.

In the following theorem, a necessary and sufficient condition for equality between $\hat{\boldsymbol{\beta}}_{GLS}$ and $\hat{\boldsymbol{\beta}}_{WLS}$ is stated. For the proof see Luati and Proietti (2008).

Theorem 1 *Equality between the GLS estimator (2) and the WLS estimator (4) holds if and only if* $\mathbf{X} = \mathbf{V}^*\mathbf{M}$ *where the* p *columns of* $\mathbf{V}^*$ *are eigenvectors of* $\mathbf{\Sigma}\mathbf{K}$ *and* $\mathbf{M}$ *is a non singular matrix.*

The theorem states that if there are p linear combinations of the columns of $\mathbf{X}$ that are eigenvectors of $\mathbf{\Sigma}\mathbf{K}$ then the GLSE with covariance matrix $\mathbf{\Sigma}$

is equal to the WLSE with kernel $\mathbf{K}$. If the conditions of the theorem hold, the equality is true for all $\mathbf{y} \in \mathbb{R}^n$, i.e.

$$(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1} = (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}$$

from which follows that $\mathbf{X}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1} = \mathbf{X}(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}$. The latter equality states that the projection matrix onto $\mathcal{C}(\mathbf{X})$ along $\mathcal{N}(\mathbf{X}'\mathbf{\Sigma}^{-1})$ is equal to the projection matrix onto $\mathcal{C}(\mathbf{X})$ along $\mathcal{N}(\mathbf{X}'\mathbf{K})$, i.e. $\mathbf{H}_{\Sigma^{-1}} = \mathbf{H}_K$. By uniqueness of the projection and complementarity of the spaces which it acts onto and along, it follows that $\mathcal{N}(\mathbf{X}'\mathbf{\Sigma}^{-1}) \equiv \mathcal{N}(\mathbf{X}'\mathbf{K})$. This allows to generalise Zyskind (1967) most famous equivalent condition to Anderson theorem in the following corollary, that is proved in Luati and Proietti (2008).

Corollary 1 *A necessary and sufficient condition for equality between the GLS estimator (2) and the WLS estimator (4) is that* $\mathbf{\Sigma}\mathbf{K}\mathbf{H} = \mathbf{H}\mathbf{\Sigma}\mathbf{K}$ *where* $\mathbf{H} = \mathbf{H}_{\Sigma^{-1}} = \mathbf{H}_K$.

For $\mathbf{K} = \mathbf{I}$, the identity matrix, we find Zyskind condition for OLSE to be BLUE. The generalisation is not straightforward, given that Zyskind proof is based on the symmetry of both $\mathbf{\Sigma}$ and $\mathbf{H}_I$, the orthogonal projection matrix onto $\mathcal{C}(\mathbf{X})$, that consents to show that the two matrices have the same eigenvectors and therefore commute. When $\mathbf{K}$ is not the identity or more generally a scalar matrix, then neither $\mathbf{H}$ nor $\mathbf{\Sigma}\mathbf{K}$ are symmetric. In any case the corollary establishes that the matrices $\mathbf{\Sigma}\mathbf{K}$ and $\mathbf{H}$ commute and therefore have the same eigenvectors. Given that a complete set of eigenvectors of $\mathbf{H}$ spans $\mathbb{R}^n$, the matrix $\mathbf{\Sigma}\mathbf{K}$ can be reduced to a diagonal form through the same matrix that diagonalises $\mathbf{H}$. This provides a further condition to verify if equality holds between (2) and (4).

These conditions enable to easily verify if, given $\mathbf{\Sigma}$ and $\mathbf{K}$, the respective GLS and WLS estimators will be equal or not, but actually they do not provide any practical information on how to obtain the optimal (in the Gauss-Markov sense) kernel for a given covariance structure. To do that, further considerations are required, that will be discussed in next section.

3 Discussion

On a purely theoretical viewpoint, there are infinite sets $\{\mathbf{X}, \mathbf{\Sigma}, \mathbf{K}\}$ that satisfy theorem 1. In fact, for any given covariance structure $\mathbf{\Sigma}$, kernel $\mathbf{K}$ and non singular matrix $\mathbf{M}$, a design satisfying $\mathbf{X} = \mathbf{V}^*\mathbf{M}$ exists. On the other hand, for a given design, which is the case that happens in practice, it is rather complicated to find criteria to establish if there exist $\mathbf{\Sigma}$ and $\mathbf{K}$ that fulfill the hypotheses of the theorem and how they are related. Since the case when $\mathbf{\Sigma}$ is given is mathematically more attractive than its opposite, we first consider this direction in looking at the problem.

One way to delineate procedures for either obtaining $\mathbf{K}$ or assessing that the theorem cannot hold consists in advancing some hypotheses on possible

structures for $\mathbf{X}$ and $\mathbf{M}$. Concerning the latter, it comes from theorem 1 that $\mathbf{M} = \mathbf{Q}'\mathbf{E}^{-1}$ where $\mathbf{Q}$ is the orthogonal matrix that diagonalises $\mathbf{E}'\mathbf{X}'\mathbf{K}\mathbf{X}\mathbf{E}$ and $\mathbf{E}^{-1}$ can be chosen as the upper triangular Cholesky factor of $\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}$ (see Luati and Proietti, 2008). As long as $\mathbf{X}$ and $\mathbf{\Sigma}$ are given and $\mathbf{K}$ is diagonal, to advance hypotheses on some structure for $\mathbf{M}$ is sensible.

For example, consider the case of any regression problem where there is an intercept, i.e. the first column of the matrix $\mathbf{X}$ is a vector of ones denoted by $\mathbf{i}$. Let us suppose that the design and covariance structure allow $\mathbf{M}$ to be upper triangular. Then, the first column of $\mathbf{X}$ is itself an eigenvector of $\mathbf{\Sigma}\mathbf{K}$ corresponding to an eigenvalue, say, d_1, so that $\mathbf{\Sigma}\mathbf{K} = d_1\mathbf{i}$. It therefore follows that a necessary condition for $\mathbf{K}$ to satisfy theorem 1 is that, up to the factor d_1,

$$\mathbf{K}\mathbf{i} \propto \mathbf{\Sigma}^{-1}\mathbf{i} \tag{5}$$

which means that the elements of $\mathbf{K}$ are (proportional to) the sum of the row elements of $\mathbf{\Sigma}^{-1}$. For local constant estimators such as those of the Nadaraya-Watson type, the condition is also sufficient.

Another case that may serve as an example arises when $\mathbf{M}$ is diagonal. The theorem is then satisfied if and only if (up to constant factors given by the elements of $\mathbf{M}^{-1}$) each column of $\mathbf{X}$, let us denote it by $\mathbf{x}_r$, $r = 1, 2, ..., p+1$ is itself an eigenvector of $\mathbf{\Sigma}\mathbf{K}$, or equivalently

$$\mathbf{K}\mathbf{x}_r \propto \mathbf{\Sigma}^{-1}\mathbf{x}_r.$$

If, in addition, the columns of $\mathbf{X}$ have some special structure (e.g. vectors of the canonical basis, constant vectors, sparse vectors, etc.), then conditions on the elements of $\mathbf{K}$ can be easily derived. An example that will be illustrated in the following section is that of a local linear regression in time series.

4 Local polynomial regression

Let us assume that y_t is a time series, measured at discrete and equally spaced time points, that can be decomposed as $y_t = \mu_t + \varepsilon_t$, where μ_t is the signal (trend) and $\varepsilon_t \sim \mathrm{N}(0, \sigma^2_\varepsilon)$ is the noise. The signal is approximated locally by a polynomial of degree d, so that in the neighbourhood of time t, $y_{t+j} = m_{t+j} + \varepsilon_{t+j}$, $m_{t+j} = \beta_0 + \beta_1 j + \beta_2 j^2 + \cdots + \beta_d j^d, j = 0, \pm 1, \cdots, \pm h$. In matrix notation, the local polynomial approximation can be written as (1), where $\mathbf{y} = [y_{t-h}, \cdots, y_t, \cdots, y_{t+h}]'$, $\boldsymbol{\varepsilon} = [\varepsilon_{t-h}, \cdots, \varepsilon_t, \cdots, \varepsilon_{t+h}]'$, the $r+1$-th column of $\mathbf{X}$ is $[(-h)^r, [-(h-1)]^r, \cdots, (h-1)^r, (h)^r]', r = 0, ..., d$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_d]'$, and $\mathbf{\Sigma} = \{\sigma_{ij}, i, j = -h, \ldots, h\}$.

Using this design, the value of the trend at time t is simply given by the intercept, $m_t = \beta_0$. Provided that $2h \geq d$, the $d+1$ unknown coefficients $\beta_k, k = 0, \ldots, d$, can be estimated by the method of generalised least squares, giving $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{y}$. In order to obtain $\hat{m}_t = \hat{\beta}_0$, we need to select the first element of the vector $\hat{\boldsymbol{\beta}}_{GLS}$. Hence, denoting by $\mathbf{e}_1$ the

$d+1$ vector $\mathbf{e}_1 = [1, 0, \ldots, 0]'$, $\hat{m}_t = \mathbf{e}_1'\hat{\boldsymbol{\beta}}_{GLS} = \mathbf{e}_1'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} = \mathbf{w}'\mathbf{y} = \sum_{j=-h}^{h} \mathrm{w}_j y_{t-j}$, which expresses the estimate of the trend as a linear combination of the observations with coefficients

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{e}_1. \tag{6}$$

It is immediate to show that (6) is the solution of the constrained minimisation problem:

$$\min_{\mathbf{w}}\{\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}\} \text{ subject to } \mathbf{w}'\mathbf{X} = \mathbf{e}_1', \tag{7}$$

where the linear constraints $\mathbf{w}'\mathbf{X} = \mathbf{e}_1'$ enforce the condition that the trend estimate reproduces a polynomial of degree d (i.e. if $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, $\hat{m}_t = \mathbf{w}'\mathbf{y} = \beta_0$).

Estimates of $\boldsymbol{\beta}$ can be also obtained by the method of weighted least squares which consists of minimising with respect to the β_k's the objective function $\sum_{j=-h}^{h} \kappa_j \left(y_{t+j} - \hat{\beta}_0 - \hat{\beta}_1 j - \hat{\beta}_2 j^2 - \cdots - \hat{\beta}_d j^d\right)^2$, where $\kappa_j \geq 0$ is a set of weights that define, either explicitly or implicitly, a kernel function. In general, kernels are chosen to be symmetric and non increasing functions of j, in order to weight the observations differently according to their distance from time t; in particular, larger weight may be assigned to the observations that are closer to t. As a result, the influence of each individual observation is controlled not only by the bandwidth h but also by the kernel. In matrix notation, setting $\mathbf{K} = \mathrm{diag}(\kappa_{-h}, \ldots, \kappa_{-1}, \kappa_0, \kappa_1, \ldots, \kappa_h)$, the WLS estimate of the coefficients is $\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}\mathbf{y}$ and the elements of the vector $\mathbf{w} = \mathbf{K}\mathbf{X}(\mathbf{X}'\mathbf{K}\mathbf{X})^{-1}\mathbf{e}_1$ constitute the so called equivalent kernel. Note that the notation $\mathbf{w}$ is used both for the GLS coefficients (6) and for the equivalent kernel arising from WLS estimation, since we will mainly focus on the case when their elements are identical. If this should not be the case, then which one of the two meanings is to be intended will be clear from the context.

In the local polynomial regression problem described so far, the matrix $\mathbf{M}$ of theorem 1 can be chosen as upper triangular with further zeros along the secondary, fourth, and so on, (upper) diagonals. This follows by algebraic considerations on the structure of $\mathbf{X}'\mathbf{K}\mathbf{X}$ and $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$. In fact, $\mathbf{X}'\mathbf{K}\mathbf{X}$ is a Hankel matrix whose elements are the values $S_r = \sum_{j=-h}^{h} j^r \kappa_j$, for $r = 0, 1, ..., 2d$, from S_0 to S_d in the first row and from S_d to S_{2d} in the last column. Note that for symmetric kernel weights satisfying $\kappa_j = \kappa_{-j}$, $S_r = 0$ for odd r and therefore $\mathbf{X}'\mathbf{K}\mathbf{X}$ has null elements along the secondary, fourth, and so on, diagonals. The matrix $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ has not Hankel structure but has zeros along the secondary, fourth, and so on diagonals as well, that is a consequence of the fact that the covariance matrix of a stationary stochastic process is a symmetric Toeplitz matrix. If we choose $\mathbf{M}$ to be the upper triangular Cholesky factor of $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ (or of $\mathbf{X}'\mathbf{K}\mathbf{X}$, according to which one of the two matrices $\mathbf{K}$ or $\boldsymbol{\Sigma}$ is given), then $\mathbf{M}^{-1} = \mathbf{E}$

and $\mathbf{M}^{-1'}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{M}^{-1} = \mathbf{I}$. If equality holds between GLSE and WLSE, then $\mathbf{M}^{-1}$ also satisfies $\mathbf{M}^{-1'}\mathbf{X}'\mathbf{K}\mathbf{X}\mathbf{M}^{-1} = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{K}\mathbf{X})$ corresponding to the (eigenvectors) columns of $\mathbf{M}^{-1}$, and $\mathbf{\Sigma}\mathbf{K}\mathbf{X}\mathbf{M}^{-1} = \mathbf{X}\mathbf{M}^{-1}\mathbf{D}$ so that the linear combinations of the columns of $\mathbf{X}$ yielding p eigenvectors of $\mathbf{\Sigma}\mathbf{K}$ are known. This gives an operative procedure to get $\mathbf{K}$ by $\mathbf{\Sigma}$, formalised in $d+1$ conditions that directly follow by the sparse upper triangular structure of $\mathbf{M}$. Here in the following, we provide explicit conditions in terms of the generic elements of $\mathbf{\Sigma}^{-1}$ and of $\mathbf{K}$ for $d = 0, 1, 2$, which are the most frequently encountered degrees for the fitting polynomial. For $d \geq 3$ we give an example.

4.1 Local constant regression

When the degree of the fitting polynomial is equal to zero, then $\mathbf{X} = \mathbf{i}$ (and $\mathbf{M}$ is a scalar), so that the necessary and sufficient condition that $\mathbf{K}$ and $\mathbf{\Sigma}$ must satisfy for the WLSE to equal the GLSE is that $\mathbf{\Sigma}^{-1}\mathbf{i} = \mathbf{K}\mathbf{i}$, i.e. if we denote by ς_{ij} the generic element of the symmetric matrix $\mathbf{\Sigma}^{-1}$,

$$\sum_{i=-h}^{h} \varsigma_{ij} \propto \kappa_j \quad \text{for } j = -h, ..., h.$$

An example is provided by the first order moving average process $\varepsilon_t = \eta_t + \theta\eta_{t-1}$ and $\eta_t \sim \text{wn}(0, \sigma_\eta^2)$; then $\mathbf{\Sigma}$ is proportional to a symmetric tridiagonal matrix with $(1+\theta^2)$ on the diagonal and θ on the subdiagonal. The kernel then follows the second order difference equation $\theta\kappa_{j-1} + (1+\theta^2)\kappa_j + \theta\kappa_{j+1} = 1$ which for $\theta = 0$ yields the uniform kernel and for $\theta = -1$ yields the Epanechnikov kernel,

$$\kappa_j \propto \frac{3}{4}\left[1 - \left(\frac{j}{h+1}\right)^2\right], \; j = -h, \ldots, h.$$

The latter is optimal for all d, see section 4.4.

4.2 Local linear regression

If $d = 1$, then $\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}$ and $\mathbf{X}'\mathbf{K}\mathbf{X}$ are diagonal, and so is $\mathbf{M}$ satisfying $\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X} = \mathbf{M}'\mathbf{M}$ and $\mathbf{M}^{-1'}\mathbf{X}'\mathbf{K}\mathbf{X}\mathbf{M}^{-1} = \mathbf{D}$ as well as $\mathbf{V}^* = \mathbf{X}\mathbf{M}^{-1}$. It therefore follows that necessary and sufficient conditions for $\hat{\boldsymbol{\beta}}_{GLS} = \hat{\boldsymbol{\beta}}_{WLS}$ are $\mathbf{\Sigma}^{-1}\mathbf{x}_r \propto \mathbf{K}\mathbf{x}_r$ for $r = 1, 2$, i.e.

$$\sum_{i=-h}^{h} \varsigma_{ij} \propto \kappa_j \quad \text{for } j = -h, ..., h \tag{8}$$

and

$$\sum_{i=-h}^{h} i\varsigma_{ij} \propto j\kappa_j \quad \text{for } j = -h, ..., h. \tag{9}$$

As it is well known, the estimate of β_0 obtained with a local linear regression is equal to that obtained with a local constant regression. So, whenever the interests lies on $\hat{\beta}_0$ only, the necessary condition (8) for optimal (Gauss-Markov) kernel estimation is also sufficient and can be derived, alternatively, based on the reproducing kernel Hilbert space (RKHS) representation of $\hat{\beta}_{0,WLS}$ and by the Cramer rule for the explicit solution of $\hat{\beta}_{0,GLS}$. In fact, within the RKHS methodology, the equivalent kernel of a linear estimator of order d can be obtained as $K_d(t) = R_d(t,0)f_0(t)$, where $R_d(t,0)$ is the reproducing kernel of a Hilbert space of polynomials up to degree $d \geq 1$ with inner product defined with respect to a density function $f_0(t)$. The reproducing kernel is so called because it reproduces any function in the Hilbert space in the sense that $\langle g, R_d(t,.)\rangle_{\mathcal{H}} = g(t), \forall t \in T, g \in \mathcal{H}$, from which many inferential properties can be derived. When $d = 1$ and t is discrete, $R_1(j,0)$ is constant with respect to j and $f_0(j) \propto \kappa_j$ (for the proof see Proietti and Luati, 2007, where the result was obtained in the case of asymmetric kernels) so that, finally, $\mathrm{w}_j \propto \kappa_j$. On the other hand, let us consider the linear system $\mathbf{X'\Sigma^{-1}X}\boldsymbol{\beta} = \mathbf{b}$, where $\mathbf{b} = \mathbf{X'\Sigma^{-1}y}$. Given that $\mathbf{b} = \sum_{j=-h}^{h} \mathbf{x}_j \sum_{i=-h}^{h} \varsigma_{ij} y_{t+i}$, where $\mathbf{x}_j = [1, j, j^2, ..., j^d]'$, then $\det(\mathbf{X'\Sigma^{-1}X}[1,\mathbf{b}]) = \sum_{j=-h}^{h} \det(\mathbf{X'\Sigma^{-1}X}[1,\mathbf{x}_j]) \sum_{i=-h}^{h} \varsigma_{ij} y_{t+i}$ so that, by the Cramer rule,

$$\hat{\beta}_0 = \sum_{j=-h}^{h} \frac{\det(\mathbf{X'\Sigma^{-1}X}[1,\mathbf{x}_j])}{\det(\mathbf{X'\Sigma^{-1}X})} \sum_{i=-h}^{h} \varsigma_{ij} y_{t+i}.$$

Since $\mathbf{x}_j = [1, j]'$, $j = -h, ..., h$ and $\mathbf{X'\Sigma^{-1}X}$ is diagonal, the ratio of the two determinants is constant with respect to j and therefore $\mathrm{w}_j \propto \sum_{i=-h}^{h} \varsigma_{ij}$ so that, reminding the above expression for the RKHS representation of w_j, equality holds between the WLS and GLS estimates of β_0 if and only if $\sum_{i=-h}^{h} \varsigma_{ij} = \kappa_j$.

4.3 Local quadratic regression

When $d = 2$, the expressions for $\mathbf{M}$ and $\mathbf{M^{-1}}$ are

$$\mathbf{M} = \begin{bmatrix} m_{11} & 0 & m_{13} \\ 0 & m_{22} & 0 \\ 0 & 0 & m_{33} \end{bmatrix} \quad \mathbf{M}^{-1} = \begin{bmatrix} m^{(11)} & 0 & m^{(13)} \\ 0 & m^{(22)} & 0 \\ 0 & 0 & m^{(33)} \end{bmatrix}$$

and therefore a further condition besides (8) and (9) is required:

$$\sum_{i=-h}^{h} i^2 \varsigma_{ij} = \frac{1}{d_3} \left[j^2 + \frac{m^{(13)}}{m^{(33)}} \left(1 - \frac{d_3}{d_1} \right) \right] \kappa_j \quad \text{for } j = -h, ..., h \tag{10}$$

where d_1 and d_3 are elements of $\mathbf{D}$.

4.4 An example holding for any d

Let us denote by $\mathbf{\Sigma}_q$ the symmetric Toeplitz matrix whose first row (column) has j-th element equal to $(-1)^j \binom{2q}{q+j}$ for $j = 0, ..., q < 2h$ and equal to zero otherwise. This is the covariance matrix of the process

$$\varepsilon_t = (1 - B)^q \eta_t, \quad \eta_t \sim \text{wn}(0, \sigma_\eta^2),$$

where B is the backshift operator such that $B^k \varepsilon_t = \varepsilon_{t-k}$. Given the regression model (1) with $\mathbf{\Sigma} = \mathbf{\Sigma}_q$, the GLSE $\hat{\boldsymbol{\beta}}_{GLS}$ is equal to the WLSE $\hat{\boldsymbol{\beta}}_{WLS}$ with kernel K_q that has weights equal to

$$\kappa_j \propto [(h+1)^2 - j^2][(h+2)^2 - j^2] \dots [(h+q)^2 - j^2], \tag{11}$$

for $j = -h, ..., h$. In other words, the design-covariance-kernel set $\{\mathbf{X}_d, \mathbf{\Sigma}_q, \mathbf{K}_q\}$, where $\mathbf{X}_d$ explicitly indicates the degree of the fitting polynomial and $\mathbf{K}_q$ denotes the diagonal matrix associated to K_q, satisfies theorem 1.

When $q = 1$, $\varepsilon_t = (1 - B)\eta_t$ and K_1 is the Epanechnikov kernel.

When q (the order of integration of the noise) is equal to d (the degree of the polynomial trend), the GLS and WLS estimators are optimal in the sense of minimising the sum of the square d-th differences of the estimates. The case $d = q = 3$ gives rise to the Henderson filters. The reason for equality, in this specific case, lies in the fact that $\mathbf{\Sigma}_q$ can be interpreted as the matrix associated to the difference operator $(1 - B)^{2q}$ subject to symmetric Toeplitz boundary conditions. If the kernel follows a polynomial of order $2q$ subject to the same $2q$ (forward and backward) boundary conditions, then the product $\mathbf{\Sigma}_q \mathbf{K}_q$ leaves unchanged any constant vector. It follows that (5) is necessarily satisfied and so are the other higher order conditions. It is crucial that symmetric Toeplitz boundary conditions are respected: polynomial kernels like the biweight, triweight, or tricube cannot be optimal for stochastic processes with covariance structures like $\mathbf{\Sigma}_q$.

References

ANDERSON, T.W. (1948): On the theory of testing serial correlation, *Skandinavisk Aktuarietidskrift, 31, 88-116.*

ANDERSON, T.W. (1971): *The Statistical Analysis of Time Series*, John Wiley and Sons, New York.

LUATI, A. and PROIETTI, T. (2008): On the equivalence of the weighted least squares and generalised least squares estimators, with applications to kernel smoothing, *Working paper.*

PROIETTI, T. and LUATI, A. (2007): Real time estimation in local polynomial regression, with applications to trend-cycle analysis, *Working paper.*

PUNTANEN, S. and STYAN, G.P.H. (1989): The equality of the ordinary least squares estimator and the best linear unbiased estimator, *The American Statistician, 43, 3, 153-161.*

ZYSKIND, G. (1969): Parametric argumentations and error structures under which certain simple least squares and analysis of variance procedures are also best, *Journal of the American Statistical Association, 64, 1353-1368.*

Bayesian Image Segmentation by Hidden Markov Models

Roberta Paroli[1] and Luigi Spezia[2]

[1] Dipartimento di Scienze Statistiche, Università Cattolica SC
Largo Gemelli 1, Milano, Italy, *roberta.paroli@unicatt.it*
[2] Biomathematics & Statistics Scotland, Macaulay Institute
Craigiebuckler, Aberdeen, UK, *luigi@bioss.ac.uk*

Abstract. We consider hidden Markov models for the segmentation, i.e. the classification, of the pixel intensities of digital images in a small set of colours, whose cardinality is unknown. New Reversible jump Markov chain Monte Carlo algorithms for estimating both the dimension and the unknown parameters of the model are introduced. Parameters are updated by random walk Metropolis-Hastings moves, accepting in block the whole set of proposals, without the updating of the sequence of the hidden Markov chain. Our image segmenters are based on the dynamics of the hidden Markov chain and the two-dimensional images are transformed into a vector through the Peano-Hilbert scan of the image. We make experiments on synthetic images, and then, the MCMC algorithms are applied to a brain magnetic resonance image.

Keywords: brain magnetic resonance images, clustering, label switching, Peano-Hilbert scan, reversible jump Markov Chain Monte Carlo

1 Introduction

Hidden Markov models (HMMs) are widely used tools in dealing with time series with incomplete data: we have a sequence of unobserved variables $X^T = (X_1, \ldots, X_T)'$ which can be analysed only through the realizations of an auxiliary observed process $Y^T = (Y_1, \ldots, Y_T)'$, by assuming that the observed variables are conditionally independent given the observed ones and that their distributions depend only on the contemporary realizations of the latent, or hidden, process. HMMs rise when the hidden process is assumed a priori to be a finite-state Markov chain.

HMMs with unknown number of regimes are considered here and the implementations of new efficient algorithms for estimating both the dimension and the unknown parameters of the model are pursued. The segmentation of the hidden regimes is another aim. The segmentation problem consists in estimating the unobserved realization $X^T = x^T$ from the observed realization $Y^T = y^T$, given the parameters of the conditional distributions.

We apply our Bayesian inference and segmentation tools to digital images. Complex stochastic image segmenters are based on hidden Markov random fields, which have the drawback of the cumbersome computation of the

normalizing constant of the Gibbs density. Computational advantages can be obtained by replacing the hidden Markov random field with a hidden Markov chain, by transforming the two-dimensional (2D) Markov random field into a one-dimensional (1D) Markov chain through the Peano-Hilbert scan of the image; the segmented 2D image is then reconstructed from x^T, by using the inverse Peano-Hilbert scan. By segmenting a sequence of pixel intensities, we show that HMMs can cluster the observations efficiently, even when the number of groups is unknown and no classification threshold is fixed.

We perform segmentation by means of four competing algorithms, whose performances we compare through simulation experiments. The four segmenters we consider are: Viterbi Algorithm (VA, MacDonald and Zucchini, 1997), Iterated Conditional Modes (ICM, Besag, 1986), Forward Filtering-Backward Sampling (FF-BSa, Carter and Kohn, 1994; Frühwirth-Schnatter, 1994), Forward Filtering-Backward Smoothing (FF-BSm, Kim, 1993). Finally, segmentation algorithms are applied to the classification of the pixel intensities of a brain magnetic resonance image.

2 Gaussian hidden Markov models

Let us consider a pair of stochastic processes $\{X_t\}$ and $\{Y_t\}$, taking values in $S_X = \{1, \ldots, m\}$ and in $\mathbb{R}$, respectively. Process $\{X_t\}$ is a priori a discrete-time, first-order, homogeneous Markov chain on S_X; the transition matrix is $\Gamma = [\gamma_{i,j}]$, where $\gamma_{i,j} = P(X_t = j \mid X_{t-1} = i)$, for any $i; j \in S_X$ and for any $t = 2, \ldots, T$, with $0 < \gamma_{i,j} < 1$. Process $\{Y_t\}$, given $\{X_t\}$, is an observed sequence of conditionally independent random variables, whose conditional distributions depend on $\{X_t\}$ only through the contemporary X_t's and they are assumed to be Gaussian.

Due to these hypotheses, the stochastic process $(\{X_t\}; \{Y_t\})$ is a Gaussian hidden Markov model (GHMM) and it can be represented as a "signal plus noise" model

$$Y_t = \mu_i + \sigma_i E_t,$$

where $\{E_t\}$ is a standardized Gaussian white noise process, with $E_t \sim \mathcal{N}(0; 1)$, so that $(Y_t | X_t = i) \sim \mathcal{N}(\mu_i; \sigma_i^2)$, for any $i \in S_X$ and for any $t = 1, \ldots, T$. However, the inferential procedures and the computational tools we introduce can be applied also when another conditional distribution is hypothesized.

We reparameterize $\Gamma = [\gamma_{i,j}]$ by $\Omega = [\omega_{i,j}]$, according to the equality

$$\gamma_{i,j} = \omega_{i,j} \Big/ \sum_{j=1}^{m} \omega_{i,j},$$

with $\omega_{i,j} > 0$, for any $i; j \in S_X$, to facilitate the random walk Metropolis-Hastings moves of the MCMC algorithm.

Vector $\left(\mu, \sigma^2, \Omega, m\right)'$ contains the unknown parameters of the GHMM to be estimated, where μ is the vector of the m signals μ_i and σ^2 is the vector of the m variances σ_i^2.

The sequence of the observations is denoted by $y^T = (y_1, \ldots, y_T)'$; so, the joint density of all the variables included in the model is

$$p(m, \mu, \sigma^{-2}, \Omega, y^T) = p\left(y^T \mid \mu, \sigma^2, \Omega, m\right) p(\mu \mid m) p(\sigma^{-2} \mid m) p(\Omega \mid m) p(m).$$

3 MCMC algorithms

We use two Markov chain Monte Carlo (MCMC) algorithms for simulating from the posterior density: the first allows to obtain the posterior estimates of the parameters, when the number of regimes is fixed; the second allows to compute the number of regimes, when it is a random variable.

The main problem to tackle is label switching, due to the multimodality of the posterior density: when we have m regimes, we have $m!$ ways to label them; so, if the priors are invariant to the relabelling of the regimes and they do not set artificial identifiability constraints, the posterior density is defined on $m!$ subspaces; hence, when we sample from the unconstrained posterior density, $m!$ labellings can alternate during the MCMC iterations. If label switching occurs, inference cannot be performed by taking the ergodic averages of the simulated values. This problem is called label switching and we tame it through the post-processing algorithm by Marin et al. (2005), by which one of the $m!$ modal regions of the posterior density is selected ex-post and then the proximity of the entries of the MCMC sample to this region manages the relabelling. Our algorithms are encouraged to visit all the $m!$ subspaces by random permutation sampling (Frühwirth-Schnatter, 2001): each draw is concluded by generating one of the $m!$ ways of labelling the regimes and then all the parameters are permuted according to the random ordering. Note that, when $m > 8$, the algorithms by Marin et al. (2005) and Frühwirth-Schnatter (2001) can take a large amount of time to run jointly.

Instead of exploiting the conjugacy of the prior distributions to implement Gibbs sampling, we prefer to update the parameters, after having mapped them on the real line, by random walk Metropolis-Hastings moves, accepting in block the whole set of proposals, given that the Gibbs sampler is less able to traverse the posterior surface and to escape local modes.

All the sweeps of our MCMC algorithm are characterized by the updating of the parameters without the updating of the sequence of the hidden Markov chain, which is never simulated, in order to eliminate unnecessary randomness from the procedure and to reduce the dimension of the parameter space. This device, called "no completion", improves the precision of the simulation and accelerates the convergence of the algorithm.

When the number of hidden states is a random variable, we consider a Reversible Jump MCMC algorithm in which the dimension of the model changes in split-and-merge and birth-and-death moves.

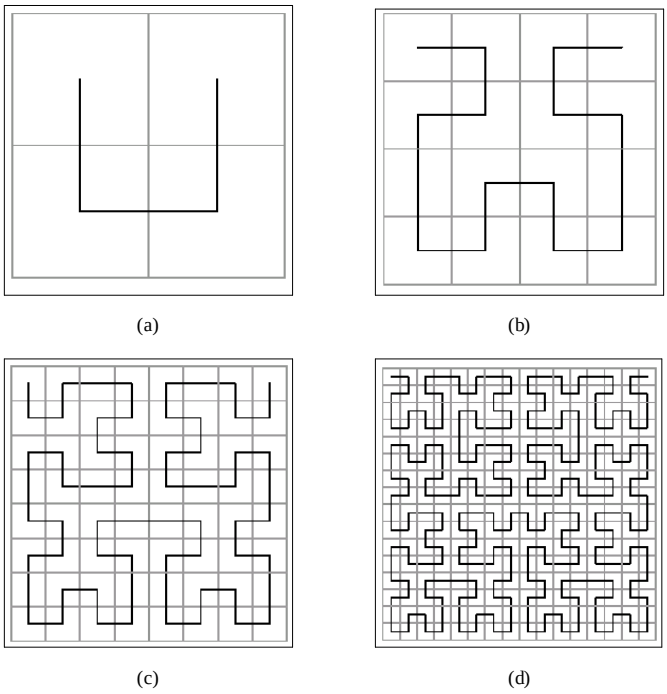


Fig. 1. Construction of the Peano-Hilbert scan.

4 The Peano-Hilbert space-filling curve

In order to use HMMs as segmenters of digital images, we have to transform the 2D set of pixels into a 1D set through the Peano-Hilbert space-filling curve on the image. This scan recursively traverses each quadrant of the image entirely before moving to the next quadrant and, thus, increases the pixels similarities among neighbouring pixels in the scan. Peano-Hilbert scan is more efficient than the simpler linear scan, where the pixels are traversed horizontally line by line, which looses the most of the spatial similarities among nearby pixels. The linearization of the image, even if it reduces the complexity of models and algorithms with respect to those that maintain the spatial structure, produces satisfying results, competitive with those obtained through hidden Markov random fields (Giordana and Pieczynski, 1997). The Peano-Hilbert curve is obtained by a recursive procedure and the first four stages of its construction are presented in Figure 1, starting with a four pixel image and, at each step, multiplying the number of pixels of the image by four. In any figure, the curve starts in the centre of the north-west square and arrives in the centre of the north-east square. The curve always joins the

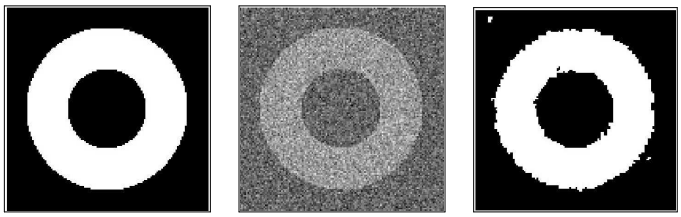


Fig. 2. Two colours images: (a) true, (b) blurred with $\sigma_i = 50$, (c) segmented with FF-BSm algorithm.

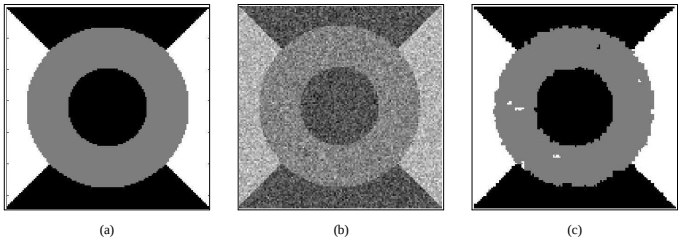


Fig. 3. Three colours images: (a) true, (b) blurred with $\sigma_i = 50$, (c) segmented with FF-BSm algorithm.

centres of two contiguous squares. Continuation of this sequence creates the Peano-Hilbert scan on an $N \times N$ image, where $N = 2^k$, $k \in \mathbb{N}_+$ and $T = N^2$.

5 Experiments on synthetic images

In this section, the analysis of three simulated examples will be presented, by considering synthetic images with two (Figure 2), three (Figure 3) and four (Figure 4) colours. We consider square synthetic images of size 128×128 pixels. Each pixel (i, j), for any $i, j = 1, \ldots, 128$, has intensity $I_{i,j}$ which is an integer belonging to the interval $[0; 255]$, where 0 is black and 255 is white. By the Peano-Hilbert scan, the 2D image is transformed into a 1D image and, then, pixel intensities are standardized.

To validate the performances of the four HMM segmenters, we make the comparison between any segmented image with the corresponding syntethic hidden image we created for the various experiments. Comparisons are made through the computation of the misclassification ratio (MR), which is

$$\text{MR} = \frac{\text{number of misclassified pixels}}{\text{total number of pixels}}.$$

The true images are shown in Figure 2a, 3a and 4a; the blurred images, produced by adding independent Gaussian noises with standard deviations

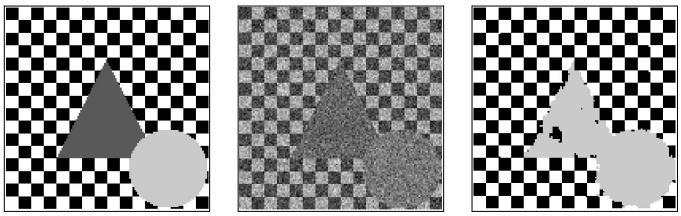


Fig. 4. Four colours images: (a) true, (b) blurred with $\sigma_i = 50$, (c) segmented with FF-BSm algorithm.

$\sigma_i = 50$, $i = 1, \ldots, m$, are shown in Figures 2b, 3b and 4b; finally, the segmented images are reproduced in Figure 2c, 3c and 4c. We also made experiments with $\sigma_i = 25$ and $\sigma_i = 80$, $i = 1, \ldots, m$.

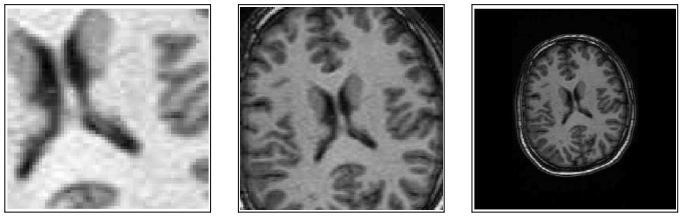
In all the experiments, the RJMCMC algorithm visits many models and always strong support is given to the correct models. The results of the segmentation experiments are collected in Table 1. By simulation results, we obtain the best performances through FF-BSm algorithm, which has been originally developed to compute the smoothed probabilities of state variables in state-space models with regime switching.

	FF-BSm	VA	ICM	FF-BSa
$m = 2$				
$\sigma_i = 25$	0.47%	0.78%	0.56%	0.53%
$\sigma_i = 50$	1.67%	7.45%	3.62%	2.63%
$\sigma_i = 80$	3.15%	27.20%	11.65%	7.12%
$m = 3$				
$\sigma_i = 25$	1.56%	5.29%	4.91%	2.20%
$\sigma_i = 50$	2.01%	38.35%	4.08%	2.40%
$\sigma_i = 80$	5.16%	42.00%	23.68%	7.43%
$m = 4$				
$\sigma_i = 25$	12.06%	12.08%	12.72%	12.74%
$\sigma_i = 50$	11.92%	58.53%	13.94%	12.16%
$\sigma_i = 80$	15.36%	34.29%	24.05%	17.86%

Table 1: Misclassification ratios for $m = 2$, $m = 3$, $m = 4$.

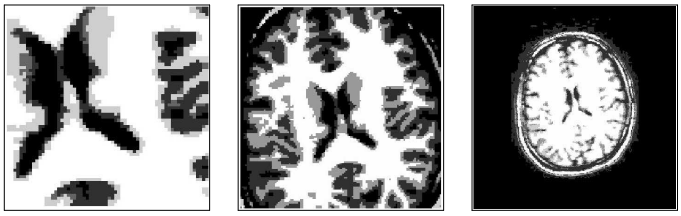
6 Analysis of a real image

Image segmentation is an important task in Computational Neuroimaging that is devoted to the development of efficient and automated techniques for brain images interpretation. In particular, in recent years, the problem of automatig the segmentation of brain images using Magnetic Resonance Images (MRI) has received special attention. MRIs provide much informations on



(a) (b) (c)

Fig. 5. MRI images: (a) 64x64; (b) 128x128; (c) 256x256.



(a) (b) (c)

Fig. 6. Segmented images: (a) 64x64; (b) 128x128; (d) 256x256.

the human soft tissue anatomy and their analysis plays an important role in diagnosing of various neural diseases. In medical literature, the brain tissue is classified in three types: Gray Matter (GM), White Matter (WM) and Cerebro-Spinal Fluid (CSF), while other structures (such as scalp-bone or non-brain structures) are classified as background. In our image it is possible to see also skull, scalp and other non-brain tissues that we are interested to segment. Furthermore GW, WM and CSF can assume different intensities in the MRI, also for degenerations of brain tissues due to some disease: so we segment our MRI by more than three or four colours. We analyse a real 2D MRI of size 256×160. To be able to apply the Peano-Hilbert scan, we must have an image of size $2^k \times 2^k$; so we analyse two sub-images, of size 64×64 and 128×128 (Figures 5a and 5b), respectively, and a 256×256 image, obtained by adding background pixels to the original MRI (Figure 5c). The RJMCMC algorithms indicate that the number of hidden states for the three images is $m = 6$. After that, the segmentations are done by means of the FF-BSm algorithm, due to its best performances, as shown in the results of Section 5. The segmented images are presented in Figure 6.

7 Conclusions

We applied hidden Markov models (HMMs) to the segmentation of the pixel intensities of a brain magnetic resonance image in a small set of colours,

whose cardinality was unknown and has been estimated by new reversible jump Markov chain Monte Carlo algorithms.

Complex stochastic image segmenters are based on hidden Markov random fields, which have the drawback of the cumbersome computation of the normalizing constant of the Gibbs density. By contrast, our image segmenters present computational advantages because we replaced the hidden Markov random field with a hidden Markov chain, by transforming the two-dimensional Markov random field into a one-dimensional Markov chain, through the Peano-Hilbert scan of the image. By segmenting a sequence of pixel intensities, we showed that HMMs can cluster the observations efficiently, even when the number of groups is unknown and no classification threshold is fixed.

We also made experiments on synthetic images and performed segmentation by means of four competing algorithms, whose performances we compared through simulation experiments: according to our results, the Forward Filtering-Backward Smoothing (Kim, 1993) has been chosen as the best segmenter.

Our future researches on this class of models can take two directions: first, we want to use hidden Markov random fields (HMRFs) as segmentation tool and compare their performances with those of HMMs; then, we want to consider Multivariate Gaussian HMMs and HMRFs to be able to segment multi-colour satellite images.

References

BESAG, J. (1986): On the Statistical Analysis of Dirty Pictures (with Discussion). *Journal of the Royal Statistical Society, Series B, 48, 259-302.*

CARTER, C.K. and KOHN, R. (1994): On Gibbs sampling for state space models. *Biometrika, 81, 541-53.*

FRÜHWIRTH-SCHNATTER, S. (1994): Data Augmentation and Dynamic Linear Models. *Journal of Time Series Analysis, 15, 183-202.*

FRÜHWIRTH-SCHNATTER, S. (2001): Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models. *Journal of the American Statistical Association, 96, 194-209.*

GIORDANA, N. and PIECZYNSKI, W. (1997): Estimation of Generalised Multisensor Hidden Markov Chains and Unsupervised Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19, 465-475.*

KIM, C.-J. (1993): Dynamic Linear Models with Markov-Switching. *Journal of Econometrics, 60, 1-22.*

MACDONALD, I.L. and ZUCCHINI, W. (1997): *Hidden Markov and Other Models for Discrete-valued Time Series.* Chapman & Hall, London.

MARIN, J.M., MENGERSEN K.L., ROBERT C.P. (2005): Bayesian Modelling and Inference on Mixtures of Distributions. In D. Dey and C.R. Rao (Eds.): *Hanbooks of Statistics 25.* Elsevier Science, Amsterdam, 459-507.

Part XIII

Multivariate Data Analysis and Dimensionality Reduction

Efficient l_α Distance Approximation for High Dimensional Data Using α-Stable Projection

Peter Clifford and Ioana Ada Cosma

Department of Statistics, University of Oxford
1 South Parks Road, Oxford OX1 3TG, United Kingdom
{*clifford,cosma*}*@stats.ox.ac.uk*

Abstract. In recent years, large high-dimensional data sets have become commonplace in a wide range of applications in science and commerce. Techniques for dimension reduction are of primary concern in statistical analysis. Projection methods play an important role. We investigate the use of projection algorithms that exploit properties of the α-stable distributions. We show that l_α distances and quasi-distances can be recovered from random projections with full statistical efficiency by L-estimation. The computational requirements of our algorithm are modest; after a once-and-for-all calculation to determine an array of length k, the algorithm runs in $O(k \log k)$ time for each distance, where k is the reduced dimension of the projection.

Keywords: random projections, stable distribution, L-estimation

1 Introduction

The efficient estimation of distances between high-dimensional data vectors is an increasingly important objective in modern statistical analysis. This paper is concerned with the recovery of l_α distances (quasi-distances) from data sketches based on α-stable random projections (Indyk, 2006). The cases $\alpha = 1$ and $\alpha = 2$, corresponding to l_1 and l_2 distances respectively, are of special interest, as is the limiting case $\alpha \to 0$ which yields the Hamming distance. By projecting m-dimensional vectors into a lower k-dimensional space, the time complexity for calculating all pair-wise distances between n vectors is potentially reduced from $O(n^2 m)$ to $O(nmk + n^2 k)$, provided the original distances can be adequately estimated from the projections. Important applications of distance-preserving dimension reduction are in clustering and classification of high-dimensional data sets, and computations over streaming data, such as Hamming distance approximations and other measures of distributional dissimilarity in stream comparisons. For motivation and further examples of the use of l_α quasi-distances in machine-learning and computational statistics, see Li and Hastie (2008), Li, Hastie and Church (2007) and the references therein.

In Section 2 and 3 we define α-stable random projections, and show that l_α distance recovery from such projections reduces to estimation of the scale parameter of the symmetric, strictly stable law. The main contributions of the paper (in Section 4) are (i) to give a simple L-estimator for the scale

parameter (and hence the l_α distance) that out-performs the estimators proposed by Li and Hastie (2008) and (ii) to show that the estimator achieves the smallest possible asymptotic standard error. Numerical illustrations of the superior performance are given in Section 5.

2 Random projections

We consider the problem of preserving l_α distances (quasi-distances) defined by $d_\alpha(u,v) = \sum_{i=1}^m |u_i - v_i|^\alpha$, for $(u_1,\ldots,u_m)$ and $(v_1,\ldots,v_m) \in \mathbb{R}^m$, for $\alpha \in (0,2]$. We remark that $[d_\alpha(u,v)]^{1/\alpha}$ is a distance measure for $\alpha \geq 1$, but not for $\alpha < 1$, and that the Hamming distance is obtained as $\lim_{\alpha\to 0} d_\alpha(u,v)$. In data stream applications, u_i and v_i are the cumulative numbers of data elements of type i in two separate streams.

A random variable X with distribution F is said to be *strictly stable* if for every $n > 0$, and independent variables $X_1,\ldots,X_n \sim F$, there exist constants $a_n > 0$ such that $X_1 + \ldots + X_n \stackrel{\mathcal{D}}{=} a_n X$, where $\mathcal{D}$ denotes equality in distribution. The only possible norming constants are $a_n = n^{1/\alpha}$, where $0 < \alpha \leq 2$; the parameter α is known as the *index* of stability (Feller, 1971). The densities of stable distributions are not available in closed form, except in a few cases: Cauchy($\alpha = 1$), Normal($\alpha = 2$) and Lévy($\alpha = 0.5$).

We are interested in symmetric, strictly stable random variables of index α and parameter $\theta > 0$, with characteristic function $\mathbb{E}\exp(itX) = e^{-\theta|t|^\alpha}$, defined for t real. Let $f(x;\alpha,\theta)$ and $F(x;\alpha,\theta)$ be the density and distribution function of X. Of particular interest is the following property. Suppose that $X_1,\ldots,X_m$ are independent variables with distribution function $F(x;\alpha,1)$ and that $u_1,\ldots,u_m$ are real constants, then $\sum_{i=1}^m u_i X_i \sim F(x;\alpha,\theta)$ where $\theta = \sum_{i=1}^d |u_i|^\alpha$. If $v_1,\ldots,v_m$ is another sequence of real constants, then it follows that $\sum_{i=1}^m (u_i - v_i)X_i \sim F(x;\alpha,\theta)$ with $\theta = d_\alpha(u,v)$.

We assume that the data V is arranged into a matrix $\mathbf{V}$ with n rows and m columns, i.e. one row for each of the n data points. Let $\mathbf{X} \in \mathbb{R}^{m\times k}$ be a matrix whose entries are independent symmetric, strictly stable random variables with index α, and $\theta = 1$ for fixed $0 < \alpha \leq 2$. We term $\mathbf{X}$ a *random projection matrix* mapping from $\mathbb{R}^m$ to $\mathbb{R}^k$ via the map $\mathbf{V} \mapsto \mathbf{VX}$.

Let $\mathbf{B} = \mathbf{VX}$ and consider u and v, the ith and jth rows of $\mathbf{V}$, $i \neq j$, corresponding to the ith and jth data points in V. Let a and b be the corresponding rows of $\mathbf{B}$. Then, for $z = 1,\ldots,k$, we have

$$a_z - b_z = \sum_{l=1}^m (u_l - v_l)X_{lz} \sim F(x;\alpha,d_{ij}), \quad \text{independently for } z = 1,\ldots,k,$$

where $d_{ij} = d_\alpha(u,v)$. Our aim is to recover $d_\alpha(u,v)$ from (a,b). Since $\{a_z - b_z : z = 1,\ldots,k\}$ provides a sample of values from a distribution with parameter $d_\alpha(u,v)$ we are in a position to apply the usual repertoire of statistical estimation techniques to obtain estimators with specified accuracy.

This is of particular relevance in the context of streaming data, where d_α, for $\alpha \leq 1$, is a meaningful measure of the pairwise distance between streams; in the extreme case of $\alpha \to 0$, d_α tends to the Hamming distance, the number of mismatches between two sequences. We point out that for fast and efficient processing of data streams, the k-dimensional representation $(a_1, \ldots, a_k)$ is directly stored as the stream is processed, thus providing a substantial dimension reduction from m to k dimensions.

When $\alpha \in [1, 2]$, the l_α distance is given by $d_\alpha^{1/\alpha}$ with potential interest for clustering in high dimensional spaces, and when $\alpha < 1$ interest focuses on d_α. In both cases the statistical problem reduces to estimating the scale parameter of the symmetric, strictly stable law.

3 Estimation of the scale parameter

The problem of parameter estimation of the stable law is particularly challenging due to the fact that the density function does not exist in closed form for most values of $\alpha \in (0, 2]$. The cases $\alpha = 1$ and $\alpha = 2$ have been extensively studied. See for example Li et al.(2007) for references. DuMouchel (1973) showed that maximum likelihood estimators (MLEs) of the parameters are both consistent and asymptotically normal, and computed estimates of the asymptotic standard deviations and correlations. Matsui and Takemura (2006) improved upon these estimates by providing accurate approximations to the first and second derivatives of the stable densities. Nolan (2001) proposes an iterative approach to maximum likelihood estimation of the parameters, implemented in his software package STABLE, available at http://www.robustanalysis.com/. Furthermore, the package STABLE implements five additional methods for estimating stable parameters, including the empirical characteristic function method and fractional moments; see references therein.

We compute approximations to the second derivative of the stable density and the logarithm of a transformed density by a second order finite difference scheme with grid width $h = 0.01$ using the integral form of the density function given in Zolotarev (1986), as implemented in the contributed package fBasics to R; Figure 1 displays the approximations. We obtained similar estimates using the expressions in Matsui and Takemura (2006).

More recently, Li (2008) proposes the harmonic mean estimator for $\alpha \leq 0.344$ and the geometric mean estimator for $0.344 < \alpha < 2$ to estimate θ. An important property of a given estimator is the asymptotic relative efficiency (ARE), defined as the ratio of the asymptotic variance of the best possible estimator, the MLE, to the asymptotic variance of the estimator under study. Combined, the estimators of Li (2008) have an ARE exceeding 70% and increasing to 100% as $\alpha \to 0$. Furthermore, Li and Hastie (2008) propose a unified estimator based on fractional powers with ARE no smaller than 75%, out-performing the combined harmonic and geometric mean estimators, and with good small sample performance for values of k as small as 10;

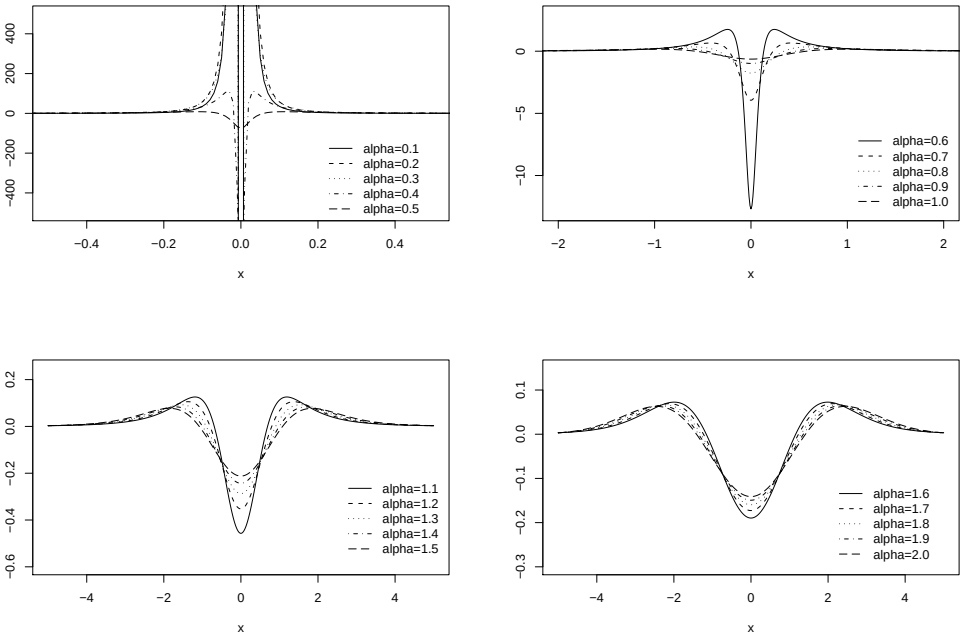


Fig. 1. Approximations to the second derivative of $f(x;\alpha,1)$ for $\alpha \in [0.1, 2]$.

the fractional power estimator was proposed previously by Nikias and Shao (1995). Our approach is to use L-estimation to estimate the logarithm of the scale parameter. We will show that the method is simple and practical, involving only a precalculated table and then a subsequent sum of products to achieve asymptotic efficiency of 100%.

4 The approach of L-estimation

Consider a random sample $x_1, \ldots, x_k \sim f(x;\alpha,\theta)$ and let $\gamma = \theta^{1/\alpha}$. Define

$$y_i := \log|x_i| \stackrel{\mathcal{D}}{=} \mu + z_i,\ i = 1, \ldots, k,$$

where z_i is distributed as the logarithm of the absolute value of a symmetric, strictly stable random variable of index α and $\theta = 1$, and $\mu = \log\gamma$. Let $f_0(z)$ and $F_0(z)$ denote the p.d.f. and distribution function of z_i, respectively. So, $(y_1, \ldots, y_k)$ is a random sample of variables with p.d.f. $f_0(y-\mu)$, where

$$f_0(z) = 2e^z f(e^z;\alpha,1), -\infty < z < \infty.$$

The problem reduces to that of estimating the location parameter μ for the family of distributions $\{f_0(y-\mu), \mu \in \mathbb{R}\}$, based on a random sample $(y_1, \ldots, y_k)$ from $f_0(y-\mu)$.

The method of L-estimation defines the estimate $\hat{\mu}$ as a weighted linear combination of order statistics $y_{(1)}, \ldots, y_{(k)}$. Chernoff et al. (1967) prove that when the weights are suitably chosen , $\sqrt{k}\,(\hat{\mu} - \mathbb{E}(\hat{\mu}))$ is asymptotically normal with mean 0 and variance I_μ^{-1}. Consequently the estimator $\hat{\mu}$ is asymptotically efficient.

In large samples, the weights can be approximated by

$$w_{ik} = -\frac{1}{kI_\mu}\ell''\Big(F_0^{-1}\Big(\frac{i}{k+1}\Big)\Big), \tag{1}$$

where $\ell(y) = \log f_0(y)$. Furthermore, the systematic bias-correction term is given by

$$BC = \mathbb{E}(\hat{\mu}) - \hat{\mu} = -\frac{1}{I_\mu}\int_{-\infty}^{\infty} z\ell''(z)f_0(z)dz,$$

so, the corresponding bias-corrected estimator is $\hat{\mu}_{BC} = \sum_{i=1}^{k} w_{ik}y_{(i)} - BC$.

Table 1 gives the Fisher information and the bias for various values of α, obtained numerically by making use of approximations to the stable densities and quantiles in the R package fBasics. The values of Fisher information agree with those presented by Matsui and Takemura (2006) to within 3-4 significant digits for most values of α.

α	I_μ	BC	α	I_μ	BC	α	I_μ	BC	α	I_μ	BC
0.14	0 0.0183	-1.5253	0.6	0.2325	-0.4380	1.1	0.5774	0.0762	1.6	1.0780	0.4183
0.15	0.0210	-1.4522	0.65	0.2626	-0.3658	1.15	0.6182	0.1119	1.65	1.1459	0.4497
0.2	0.0363	-1.1956	0.7	0.2937	-0.2995	1.2	0.6604	0.1466	1.7	1.2198	0.4741
0.25	0.0547	-1.0420	0.75	0.3256	-0.2388	1.25	0.7042	0.1804	1.75	1.3011	0.4874
0.3	0.0755	-0.9331	0.8	0.3585	-0.1834	1.3	0.7499	0.2138	1.8	1.3920	0.4875
0.35	0.982	-0.8438	0.85	0.3924	-0.1324	1.35	0.7976	0.2470	1.85	1.4968	0.4743
0.4	0.1226	-0.7611	0.9	0.4272	-0.0852	1.4	0.8476	0.2804	1.9	1.6270	0.4480
0.45	0.1483	-0.6790	0.95	0.4631	-0.0412	1.45	0.9002	0.3142	1.95	1.7882	0.4122
0.5	0.1753	-0.5965	1.0	0.5	0	1.5	0.9558	0.3487	1.99	1.8861	0.3912
0.55	0.2034	-0.5154	1.05	0.5379	0.0390	1.55	1.0148	0.3838	2.0	2.0	0.3687

Table 1. Fisher information I_μ for the parameter μ and the systematic bias (BC) in estimating μ by efficient L-estimation, tabulated for values of $\alpha \in [0.14, 2]$.

In the case $\alpha > 1$, we will be interested in estimating $\gamma = e^\mu$, corresponding to the l_α^α norm. We propose the estimator $\hat{\gamma} = \exp(\hat{\mu}_{BC})$. It follows that $\sqrt{k}\big(\hat{\gamma} - \gamma\big)$ is asymptotically normal with mean 0 and variance $1/I_\gamma$, where I_γ is the Fisher information about the scale parameter γ contained in $(x_1, \ldots, x_k)$, or equivalently $(y_1, \ldots, y_k)$. By second order Taylor expansion, we show that the bias incurred by exponentiating is approximately

$$\mathbb{E}(\hat{\gamma}) \approx \gamma + \frac{1}{2}\gamma\mathbb{E}\,(\hat{\mu}_{BC} - \mu)^2 = \gamma\left(1 + \frac{1}{2kI_\mu}\right),$$

so the bias-corrected estimator $\hat{\gamma}_{BC} = \hat{\gamma}\Big(1 - \frac{1}{2kI_\mu}\Big)$ is unbiased up to terms of order $O(1/k^2)$.

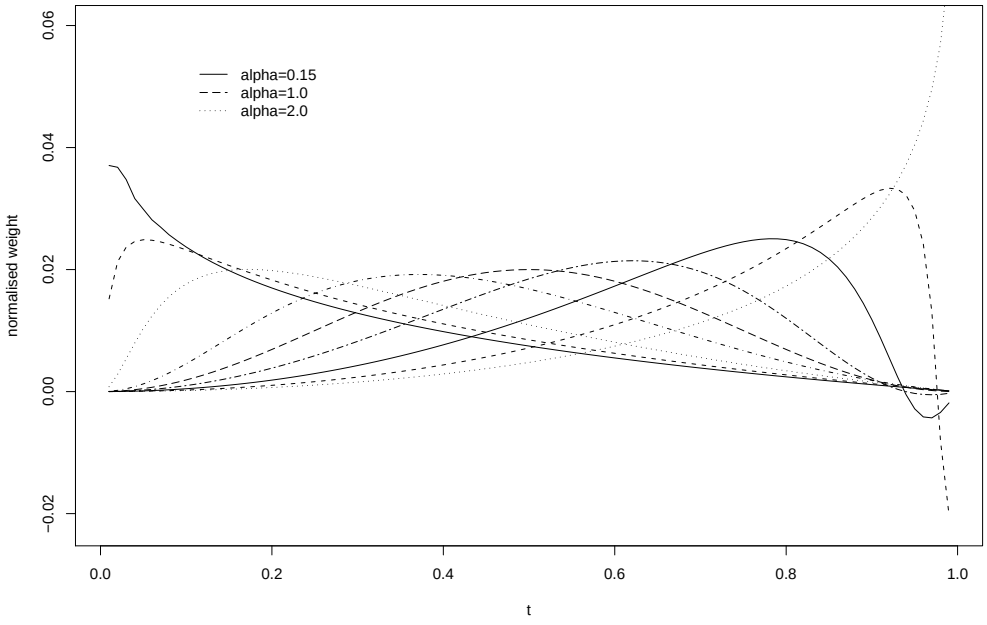


Fig. 2. Approximate weights w_{ik} for $t := \frac{i}{k+1} \in (0.01, 0.99)$ and $\alpha = 0.15, 0.3, 0.5, 0.8, 1.0, 1.2, 1.5, 1.8, 2.0$ (from left to right, in order of the peaks).

In practice, we use the following approximation for the weights in (1)

$$w_{ik} \approx \frac{\ell''\big(F_0^{-1}\big(\frac{i}{k+1}\big)\big)}{\sum_{j=1}^{k} \ell''\big(F_0^{-1}\big(\frac{j}{k+1}\big)\big)},$$

normalised to sum to 1; Figure 2 displays the weights for various values of α. For α small, the weighted sum in the formulation of the L-estimator places significant weight on the small order statistics, and negligible weight on the large order statistics, gradually shifting the weight balance towards large order statistics as $\alpha \to 2$. The bias-corrected estimator of γ is computed as follows:

$$\hat{\gamma}_{BC} = \exp\Big\{ \sum_{i=1}^{k} w_{ik} \Big(y_{(i)} - F_0^{-1}\Big(\frac{i}{k+1}\Big)\Big)\Big\} \Big[1 + \frac{1}{2\sum_{j=1}^{k} \ell''\big(F_0^{-1}\big(\frac{j}{k+1}\big)\big)}\Big].$$

Similar calculations provide an asymptotically efficient estimator for θ; a more relevant parameter for values of α less than 1.

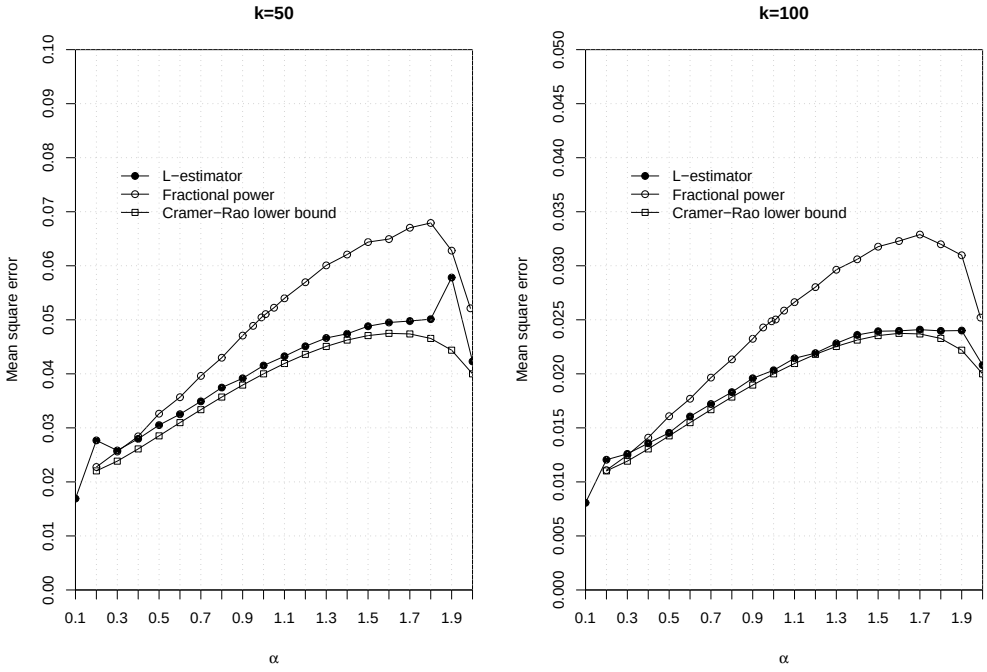


Fig. 3. Comparison in terms of mean square error (m.s.e.) of the L-estimator of θ with the fractional power estimator of Li and Hastie (2008) (10^5 replicates). The Cramér-Rao lower bound is plotted for comparison. The equivalent plot for estimators of $\gamma = \theta^{1/\alpha}$ shows a similar pattern. The perturbation in the m.s.e.for the L-estimator at $\alpha = 1.9$ is caused by an oscillation in the weight function; it can be minimised by selective trimming.

5 Numerical results

The L-estimator is easily computable as the weights depend only on α and k, and can be tabulated once-and-for-all for any required value of α. The calculation of these terms depends on accurate approximations to the quantiles and the density of the symmetric, strictly stable distribution. Whereas it is possible to obtain a good approximation to the MLE via an iterative procedure with a suitably large table of pre-calculated derivatives for fixed α, the L-estimation procedure has the advantage of achieving the same asymptotic performance without iteration. The L-estimator has modest computing requirements; it has $O(k \log k)$ running time and $O(k)$ storage requirement given a table of pre-calculated weights for given α.

To confirm the superior performance of our L-estimator, we have simulated its mean square error for various sample sizes and various values of α. Figure 3 shows that, as expected, the L-estimator has smaller mean square

error than the estimator of Li and Hastie (2008). The perturbations in the m.s.e. of the L-estimator at $\alpha = 1.9$ are caused by an oscillation of the weight function which becomes negative when $\frac{i}{k+1}$ is close to 1 (see Figure 2). The effect can be minimised by using a trimmed version of the L-estimator. This is work in progress and will be reported elsewhere.

In summary, we have shown that l_α distances and quasi-distances can be recovered from α-stable random projections with full statistical efficiency by a non-iterative L-estimator with small computational cost. In numerical experiments, we confirm that the mean square errors of our estimators are consistently smaller than the estimators of Li and Hastie (2008) for the entire range of $\alpha \in (0, 2]$.

References

CHERNOFF, H., GASTWIRTH, J. L. and JOHNS, Jr., M. V. (1967): Asymptotic Distribution of Linear Combinations of Functions of Order Statistics with Applications to Estimation. *Ann. Math. Stat. 38 (1), 52-72.*

DUMOUCHEL, W.H. (1973): On the asymptotic normality of the maximum likelihood estimate when sampling from a stable distribution. *Ann. Stat. 1 (5), 948-957.*

FELLER, W. (1971): *An Introduction to Probability Theory and Its Applications.* John Wiley & Sons, New York.

INDYK, P. (2006): Stable distribution, pseudorandom generators, embeddings, and data stream computation. *Journal of ACM, 53 (3), 307-323.*

LI, P. (2008): Estimators and Tail Bounds for Dimension Reduction in l_α $(0 < \alpha \leq 2)$ Using Stable Random Variables. In: *SODA.* San Francisco, CA.

LI, P. and HASTIE, T.J. (2008): A Unified Near-Optimal Estimator for Dimension Reduction in l_α $(0 < \alpha \leq 2)$ Using Stable Random Variables. In: J. C. Platt, D. Koller, Y. Singer and S. Roweis (Eds.): *Advances in Neural Information Processing Systems 20.* MIT Press, Cambridge, MA.

LI, P., HASTIE, T.J. and CHURCH, K.W. (2007): Nonlinear Estimators and Tail Bounds for Dimension Reduction in l_1 Using Cauchy Random Projections. *Journal of Machine Learning Research 8, 2497-2532.*

MATSUI, M. and TAKEMURA, A. (2006): Some Improvements in Numerical Evaluation of Symmetric Stable Density and Its Derivatives. *Communications in Statistics: Theory and Methods 35 (1), 149-172.*

NIKIAS, C.L. and SHAO, M. (1995): *Signal Processing with Alpha-Stable Distributions and Applications.* Wiley, New York.

NOLAN, J.P. (2001): Maximum likelihood estimation of stable parameters. In: O. E. Barndorff-Nielsen, T. Mikosch and S. I. Resnick (Eds.): *Lévy Processes: Theory and Applications.* Birkhäuser, Boston, MA, 379-400.

ZOLOTAREV, V. M. (1986): *One-dimensional stable distributions.* American Mathematical Society, Providence, RI.

Analysis of Consensus Through Symbolic Objects

José M. García-Santesmases[1] and M. Carmen Bravo[2]

[1] Universidad Complutense de Madrid, Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Matemáticas, 28040 Madrid, Spain, *jsantes@mat.ucm.es*

[2] Universidad Complutense de Madrid, Servicio Informático de Apoyo a Docencia e Investigación, Edificio Real Jardín Botánico Alfonso XIII, 28040 Madrid, Spain, *mcbravo@pas.ucm.es*

Abstract. This paper addresses the problem of analysing the existence of different patterns of consesus when data come from several observers who separately evaluated several issues on a rating scale of ordered categories. We propose a method that uses clustering techniques and symbolic objects to identify and describe groups of individuals with a high agreement on several questions, **consensus groups**. For each consensus group, we find out how many individuals should change their opinions and to what extent (jump) to belong to the consensus group. The consensus solution we propose is the minimum number of consensus groups that cover all the individuals with a maximum fixed jump.

Keywords: symbolic objects, rater agreement, measure of agreement, consensus measure

1 Introduction

A new approach is proposed for the analysis of consensus over a group of individuals. One common meaning of consensus is a general agreement among the members of a given group and can be seen as a function of shared team feelings towards an issue. To analyse it, two main steps can be distinguished:

a) The use of consensus measures to evaluate the strength of consensus in a class of individuals (Tastle et al. (2005)).

b) The evaluation of each individual to propose changes in his opinion in order to increase the strength of the consensus.

Usually the analysis is based on the responses to a single variable. When there is more than one question the analysis is made separately in each question.

We propose to extend this type of analysis to several issues and use concepts developed in the context of Symbolic Data Analysis (SDA) (Bock and Diday (2000), Diday and Noirhomme-Fraiture (2007)) to identify and describe groups of individuals with a high strength of consensus. Assertion objects will represent the consensus groups which extensions are usually an

overlapping clustering of the individuals. For a fixed maximun number of allowed changes in individual opinions in order to belong to a consensus group, a two-step integer programming problem based on δ-extensions of symbolic objects is used to reduce the number of consensus groups needed to cover all individuals.

2 Basic concepts and notation

Let $E = \{u_1, u_2, ..., u_n\}$ be the set of individuals or experts that answer p questions $y_1, y_2, ..., y_p$ on an ordinal scale $Y_l = \{r_1, r_2, ..., r_{t_l}\}$ that represents the ratings or rankings of each individual preference. We shall consider an ordinal level rating for example like the rating levels on a Likert-type scale. The rating measures the extent to which a person agrees or disagrees with the question. For example with five possible values: 1 strongly disagree, 2 somewhat disagree, 3 undecided, 4 somewhat agree, 5 strongly agree, or with 10 possible values from 1 to 10.

Let s be an assertion symbolic object. Then $s = \bigwedge_{j=1,...p} [y_j R_j D_j]$, with D_j, a subset of consecutive values of Y_j, and R_j, a boolean relation between descriptions given by the $\in$ operation. The assertion s is the mapping $s : E \to \{0, 1\}$ defined by:

$$s(u) := \prod_{j=1,...p} [y_j(u) R_j D_j] \qquad for \quad u \in E \tag{1}$$

Let A denote the set of assertion objects and g the boolean extension mapping $g : A \to P(E)$ defined by:

$$g(s) := EXT(s) = \{u \in E \mid s(u) = 1\} \tag{2}$$

The sum $\sum_{j=1,...p} [y_j(u) R_j D_j]$ is the number of variables for which $u \in E$ verifies the description of s and

$$j_s(u) = p - \sum_{j=1,...p} [y_j(u) R_j D_j] \tag{3}$$

is the number of variables for which u should change in order to belong to $g(s)$. This is the jump of u to belong to $g(s)$.

For any given threshold $\delta \in [0, 1]$ the extension of level δ of s is defined by:

$$g^\delta(s) := EXT^\delta(s) = \left\{ u \in E \mid \frac{j_s(u)}{p} \leq 1 - \delta \right\} \tag{4}$$

The set $g^\delta(s)$ contains the individuals of E that need a jump less or equal to $p(1 - \delta)$ to belong to $g(s)$. We have $g(s) = g^1(s) \subseteq g^\delta(s) \subseteq g^0(s) = E$.

The symbolic object builder $f : P(E) \rightarrow A$, associates with any subset $C \subset E$, an assertion symbolic object $f(C)$ defined as:

$$f(C) := \bigwedge_{j=1,...p} [y_j R_j D_j]\,,\ D_j = \{c_j, c_j + 1, ..., c_j^{'}\}, \qquad (5)$$
$$c_j = \min_{u \in C} y_j(u),\ c_j^{'} = \max_{u \in C} y_j(u)$$

Example 2. Let $C = \{(1,2,5,2,5),(2,2,4,3,5),(1,5,2,6,9),(2,4,2,3,7)\}$ be a set of four individuals described by their values on $p = 5$ variables, with $Y_j = \{0,1,...9\}\,,\ j = 1,...,5$. Then $f(C) = [y_1 \in \{1,2\}] \wedge [y_2 \in \{2,3,4,5\}] \wedge [y_3 \in \{2,3,4,5\}] \wedge [y_4 \in \{2,3,...,6\}] \wedge [y_5 \in \{5,6,...,9\}]$ that can be represented in a more simple way as: $f(C) = /1,2/ \wedge /2,5/ \wedge /2,5/ \wedge /2,6/ \wedge /5,9/$.

3 Consensus description by symbolic objects

Symbolic objects are a well known tool to represent a class of individuals. We propose to use symbolic objects to explain the consensus of individuals. Given a set of individuals E, for every subset $C \subset E$, we build $f(C)$ the associated assertion object. We propose the 2D zoom star of Noirhomme-Fraiture and Rouard (2000) as an explanatory graphical representation of $f(C)$.

The consensus measure between two individuals can be based on a concordance meausure like the Cohen's Kappa statistic (Cohen (1960)) calculated over a contingency table. This measure expresses the extent to which the observed amount of agreement between the two raters exceeds what would be expected when both made their ratings completely randomly.

Example 3. Let $u_1 = (1,2,5,2,2,5,3,2,3,1,2,3,4,4,5)$ and $u_2 = (2,2,4,3,3,4,3,4,2,1,2,2,5,4,5)$ be two individuals described by their values on $p = 16$ variables, $Y_j = \{1,2,...,5\}\,,\ j = 1,...,16$. The contingency table that is obtained crossing responses of u_1, u_2 to the common variable scale is:

$u_1 \backslash u_2$	1	2	3	4	5
1	**1**	**2**	0	0	0
2	**0**	**2**	**2**	0	0
3	0	**1**	**1**	**0**	0
4	0	0	**1**	**1**	**1**
5	0	0	0	**1**	**1**

(6)

The more elements on the diagonal the stronger the ageement is.

For any number of individuals in C a graphical representation of $s = f(C)$ is a 2D zoom star.

Example 4. For individuals for example 1, the 2D zoom star of $s = f(C)$ is shown in Figure 1.

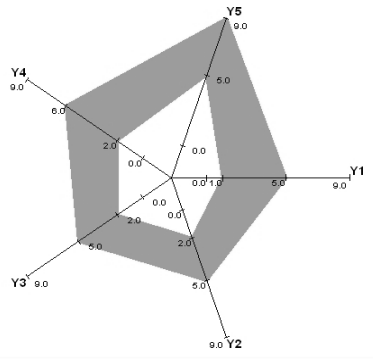


Fig. 1. 2D zoom star for data of example 1.

The star representation provides a good idea of the type and strength of the consensus in C by the width and the shape of the star. The narrower the star is, the stronger the consensus is. Given $s = \bigwedge_{j=1,...p} [y_j R_j D_j]$, the star width in one axis can be measured by $d_j(s) = \frac{|D_j|}{|Y_j|}$ (as in Brito (2000)) and the **star width** by $\mathbf{d}(s) = (d_1(s), d_2(s), ..., d_p(s))$, with $|.|$ the cardinal. A consensus measure can be $1 - V_i(s)$, $i = 1, 2$ with:

$$V_1(s) = (\sum_{j=1,...,p} d_j(s))/p \tag{7}$$

$$V_2(s) = \prod_{j=1,...p} d_j(s) \tag{8}$$

These measures are the width axis average and the star volume. In case of two raters, these measures are equivalent to the diagonal band width of the contingency table. In Table (6), numbers in bold type show a band width of 2/5. When there are more raters, it is very easy to incorporate more individuals to a star representation (Figure 1) whereas no such easyness can be attained by contingency table representations. In the sequel, we will refer V_i by the volume and it will be denoted by V. For s of example 3, $V_1(s) = 2/5$ and $V_2(s) = 0.8 * 10^{-2}$.

For E we can analyse the consensus of E elements with $f(E)$ and with the volume $V(f(E))$. This generalization process of E into $f(E)$ usually gives over-generalisation, that is, too big values for $V(f(E))$. A way to reduce over-generalisation is to define a specialization step as in Stephan et al. (2000). This step builds from $f(E)$ a new assertion symbolic object $r(f(E))$ such that $V(r(f(E))) < V(f(E))$ for which the extension is not very different to E.

This over-generalisation also means a low consensus measure. We propose a clustering based solution to reduce this over-generalization by identifying

symbolic objects $S = \{s_1, s_2, ..., s_L\}$ such that $\{g(s_1), g(s_2), ..., g(s_L)\}$ is a clustering of E, not necessarily a partition. Desired properties of S in order to increase the consensus measure are those of any clustering process, based in minimising the values of the volume of each $s \in S$, $V(s)$, the total volume, $\mathbf{V}(S) = \sum_{s \in S} V(s)$ and the number of clusters, L.

Our approach to the analysis of consensus on the data is to obtain a solution $S = \{s_1, s_2, ..., s_L\}$, with s_l a symbolic object that describes a class of individuals $g(s_l)$ with a strong consensus, given by $1 - V(s_l)$. Each $s \in S$ shall be a consensus group and we will be interested in identifying the individuals that should change their opinions and to what extent to belong to $g(s)$. The set $g^{\delta}(s) - g(s)$ is the subset of E that should change their opinion with a positive jump lower or equal to $p(1 - \delta)$.

4 Algorithms and solutions

Our attention will be focused on solutions $S = \{s_1, s_2, ..., s_L\}$ such that the width of each $s \in S$ will be below a fixed vector value $\mathbf{h}$. The S is a solution of level $\mathbf{h}$ when $\mathbf{d}(s) \leq \mathbf{h}$, $\forall s \in S$. In algorithm 1 (in 4.1) we obtain solutions of level $\mathbf{h}$ applying a method similar to Hartigan's (1975, p. 74-78) leader algorithm. In algorithm 2 (in 4.2), from a given solution we obtain a stable solution with lower volume. Finally in 4.3, q-solutions are obtained.

4.1 Solutions of level h

Algorithm 1. The initial clusters are $L \leq n$ individuals. The maximum star width is $\mathbf{h}$.

Step 1 (Initialization)
Fix $\mathbf{h}$ and L
Let $\{u_1, u_2, ..., u_L\} \subset E$ chosen at random.
Build $S = \{s_1, s_2, ..., s_L\}$; $s_l = f(\{u_l\})$

Step 2
$\forall u \in E$ find $s' \in S \mid V(f(g(s') \cup \{u\})) - V(s') =$
$\min_{s \in S}(V(f(g(s) \cup \{u\}) - V(s))$
If $\mathbf{d}(f(g(s') \cup \{u\})) \leq \mathbf{h}$ then $s' \leftarrow f(g(s') \cup \{u\})$
Else find next s'
If not modification of any s' then $s_{L+1} \leftarrow f(\{u\})$, $S \leftarrow S \cup \{s_{L+1}\}$

4.2 Stable solutions of level h

A solution S is a **stable solution** when there are no possible movements between clusters of S. Given a solution $S = \{s_1, s_2, ..., s_L\}$, $u \in g(s_l)$ **can move** from s_l to s_r when:

$$V(f(g(s_l) - \{u\})) + V(f(g(s_r) \cup \{u\})) < V(s_l) + V(s_r) \tag{9}$$

If S is a solution of level **h** and the following condition:

$$\mathbf{d}(f(g(s_r) \cup \{u\})) \leq \mathbf{h} \tag{10}$$

is added to movement condition (9), then S is a **stable solution of level h**. The border of $s \in S$ is defined by:

$$Br(s) := \{u \in g(s) | V(f(g(s) - \{u\})) < V(s)\} \tag{11}$$

The following properties are easily proven:

-If $u \in g(s) - Br(s)$ then u cannot move to any $s' \in S$
-If $u \notin g(s)$ then $V(s) < V(f(g(s) \cup \{u\}))$
-If $u \in g(s)$ then $V(s) = V(f(g(s) \cup \{u\}))$
-If $u \in g(s_l) \cap g(s_r)$ and $u \in Br(s_l)$ then u can move from s_l to s_r
-If $|g(s)| = 2$ then $Br(s) = g(s)$

To obtain a stable solution of level **h** we propose the following algorithm. It starts either from a partition of E, $\{C_1, C_2, ..., C_L\}$ or from a solution S. In the case of a partition, the solution $\{f(C_1), f(C_2), ..., f(C_L)\}$ is built applying the object builder function.

Algorithm 2. Let $S = \{s_1, s_2, ..., s_L\}$ be a solution:
Step 1 (Initialization)
Build $\mathbf{h} = \max_{l=1,...L} \mathbf{d}(s_l)$
Step 2
For $i = 1$ *TO* n *DO*
For $l = 1$ *TO* L *DO*
$u_i \in Br(s_l)$; $L_l = \{1, ..., L\} - \{l\}$
Find $r \in L_l \mid V(f(g(s_r) \cup \{u_i\})) - V(s_r) =$
$\min_{t \in L_l}(V(f(g(s_t) \cup \{u_i\}) - V(s_t))$
If u_i can move from s_l to s_r and $\mathbf{d}(f(g(s_r) \cup \{u_i\})) \leq \mathbf{h}$ then
Update $S : s_l \leftarrow f(g(s_l) - \{u_i\})$; $s_r \leftarrow f(g(s_r) \cup \{u_i\})$
Else $L_l \leftarrow L_l - \{r\}$
Find next r
END DO
END DO

4.3 q-solutions

Given a stable solution $S = \{s_1, s_2, ..., s_L\}$, L is the minimum number of clusters needed to cover all the individuals of E with jump 0 ($\bigcup_{l=1,...,L} g(s_l) = E$). A **q-solution** $S^q \subset S$ is a class of elements of S with the minimum number of clusters needed to cover E with maximum jump q, $0 \leq q \leq p$ ($\bigcup_{s_l \in S^q} g^{1-q/p}(s_l) = E$).

Let x_l^q be a binary variable that represents the presence/absence of cluster s_l in the q-solution and $a_{s_l}^q(u_i) = 0$ when $j_{s_l}(u_i) > q$; and, $a_{s_l}^q(u_i) = p -$

$j_{s_l}(u_i)$, otherwise. The function $a_s^q(.)$ is the complementary to p of $j_s(.)$ when jump is lower or equal to q.

The solution of the following integer programing problem gives m, the minimum number of clusters in a **q-solution**.

$$Min \sum_{l=1,...,L} x_l^q$$

subject to:

$$\sum_{l=1,...,L} x_l^q a_{s_l}^q(u_i) > 0,\ i = 1, ..., n; \qquad x_l^q \in \{0,1\},\ l = 1, ..., L \quad (12)$$

This problem (12) also gives one of the q-solutions. Among all possible q-solutions we will choose the one that mimimizes a function of the total number of jumps. The final q-solution is the solution of this integer programing problem:

$$Max \sum_{l=1,...,L} x_l^q \sum_{i=1,...,n} a_{s_l}^q(u_i)$$

subject to:

$$\sum_{l=1,...,L} x_l^q = m; \quad \sum_{l=1,...,L} x_l^q a_{s_l}^q(u_i) > 0, i = 1, ..., n; \quad x_l^q \in \{0,1\}, l = 1, ..., L \quad (13)$$

5 Example

To illustrate the proposed method we apply it to the data set that represents the ratings given by 50 experts to 7 quality indexes. The ratings are on a 1 to 10 scale with 1 meaning very low quality and 10 meaning very strong quality. The indexes were: y_1, health; y_2, educational and cultural; y_3, employment; y_4, household; y_5, economical resources; y_6, safety; y_7, environment. The data table is:

expert	y_1	y_2	y_3	y_4	y_5	y_6	y_7	expert	y_1	y_2	y_3	y_4	y_5	y_6	y_7	expert	y_1	y_2	y_3	y_4	y_5	y_6	y_7
exp1	6	7	8	6	6	6	5	exp19	5	8	9	9	4	5	6	exp37	5	8	5	5	4	6	8
exp2	4	9	6	8	3	1	8	exp20	3	10	7	9	3	5	7	exp38	3	5	8	6	3	4	10
exp3	6	10	10	10	1	1	10	exp21	8	2	4	5	8	7	2	exp39	2	6	4	5	4	8	9
exp4	9	10	9	9	2	1	6	exp22	10	6	7	6	7	9	3	exp40	5	7	7	7	6	4	3
exp5	8	10	10	10	4	2	5	exp23	9	1	4	2	10	7	6	exp41	7	3	7	4	8	9	1
exp6	3	4	7	7	6	5	3	exp24	5	2	3	4	7	7	3	exp42	4	7	9	7	7	6	4
exp7	7	9	7	9	4	4	3	exp25	8	4	6	6	8	7	3	exp43	6	5	3	7	6	5	3
exp8	4	6	4	3	4	6	10	exp26	6	4	2	4	7	8	3	exp44	5	4	4	5	4	4	6
exp9	2	3	6	1	5	8	6	exp27	8	9	6	8	6	5	3	exp45	7	2	1	5	7	7	7
exp10	1	1	6	1	6	9	7	exp28	5	6	8	2	10	10	6	exp46	5	2	3	5	6	6	7
exp11	2	7	4	3	4	7	9	exp29	8	5	6	6	9	7	5	exp47	5	6	3	5	4	4	2
exp12	6	3	2	4	7	7	6	exp30	7	7	3	3	7	6	5	exp48	5	3	1	5	7	7	4
exp13	8	3	4	5	7	6	4	exp31	4	6	5	5	8	8	4	exp49	6	5	2	6	4	7	5

expert	y_1	y_2	y_3	y_4	y_5	y_6	y_7	expert	y_1	y_2	y_3	y_4	y_5	y_6	y_7	expert	y_1	y_2	y_3	y_4	y_5	y_6	y_7
exp14	3	6	7	7	4	4	5	exp32	4	5	5	4	4	7	6	exp50	7	4	7	5	6	5	5
exp15	4	4	7	6	6	4	5	exp33	6	7	7	5	7	6	4	exp18	6	6	9	7	5	6	5
exp16	5	4	5	6	5	4	6	exp34	7	2	3	2	8	7	5	exp36	6	6	5	4	3	5	7
exp17	5	7	8	7	4	3	6	exp35	7	9	6	6	5	8	9								

We have applied algorithm 1 with $\mathbf{h}$=(0.6,0.6,0.6,0.6,0.6,0.6,0.6) and initial value of L=1. We have obtained a solution of level $\mathbf{h}$, $S = \{s_1, ..., s_7\}$. The total volume (using V_1 of (7)) is $\mathbf{V}(S) = 3.42857$. The interesections between the extensions of elements of S are:

$g(s_1) \cap g(s_4) = \{exp1\}$
$g(s_1) \cap g(s_6) = \{exp1, exp19\}$
$g(s_2) \cap g(s_3) = \{exp32\}$
$g(s_2) \cap g(s_4) = \{exp15, exp16, exp32, exp44, exp50\}$
$g(s_2) \cap g(s_5) = \{exp25, exp29\}$
$g(s_3) \cap g(s_4) = \{exp8, exp36, exp46\}$
$g(s_4) \cap g(s_5) = \{exp34\}$

When we apply algorithm 2 to this solution we obtain a stable solution of level $\mathbf{h}$, $S' = \{s'_1, ..., s'_7\}$ that reduces the total volume to $\mathbf{V}(S') = 2.92$. The elements of S' are:

$s'_1 = /6, 9/ \wedge /10, 10/ \wedge /9, 10/ \wedge /9, 10/ \wedge /1, 4/ \wedge /1, 2/ \wedge /5, 10/$
$s'_2 = /3, 8/ \wedge /4, 9/ \wedge /2, 7/ \wedge /4, 9/ \wedge /4, 8/ \wedge /4, 8/ \wedge /2, 6/$
$s'_3 = /1, 6/ \wedge /1, 3/ \wedge /1, 6/ \wedge /1, 5/ \wedge /5, 7/ \wedge /6, 9/ \wedge /4, 7/$
$s'_4 = /2, 7/ \wedge /5, 9/ \wedge /3, 8/ \wedge /3, 6/ \wedge /3, 7/ \wedge /4, 8/ \wedge /5, 10/$
$s'_5 = /5, 10/ \wedge /1, 6/ \wedge /3, 8/ \wedge /2, 6/ \wedge /7, 10/ \wedge /6, 10/ \wedge /1, 6/$
$s'_6 = /3, 6/ \wedge /6, 10/ \wedge /6, 9/ \wedge /7, 9/ \wedge /3, 7/ \wedge /1, 6/ \wedge /4, 8/$
$s'_7 = /7, 7/ \wedge /2, 2/ \wedge /1, 1/ \wedge /5, 5/ \wedge /7, 7/ \wedge /7, 7/ \wedge /7, 7/$

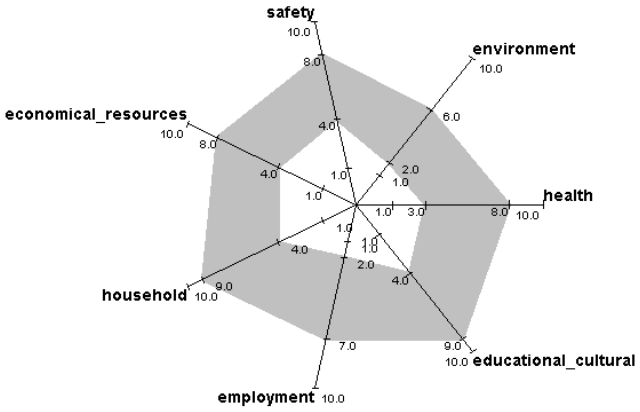


Fig. 2. 2D zoom star for the consensus group s'_2.

The cardinal of the intersections are also reduced.

$g(s'_2) \cap g(s'_6) = \{exp14\}$
$g(s'_2) \cap g(s'_4) = \{exp32\}$
$g(s'_2) \cap g(s'_5) = \{exp25\}$

Figure 2 is a graphic representation of s'_2. The extension of s'_2 is given by $g(s'_2) = \{exp6, exp7, exp14, exp15, exp16, exp25, exp26, exp27, exp31, exp32, exp33, exp40, exp43, exp44, exp47, exp49, exp50\}$. These are the individuals of E that need a jump 0 to belong to s'_2. For a jump of 1, 12 more experts are added, who are the experts that should change their opinion in just one question. For a jump of 2, 8 more experts are added. For a jump of 3, 10 more experts are added. For a jump of 4, $\{exp5, exp23, exp28\}$ are added. For a jump of 5, $\{exp4, exp10\}$ are added. And, for a jump of 6 the last expert, $exp3$, is added.

A 3-solution is obtained. It is composed of 4 consensus groups (solution of minimisation problem (12)), given by $S'^3 = \{s'_1, s'_2, s'_3, s'_4\}$ (solution of maximisation problem (13)), with volume $\mathbf{V}(S'^3) = 1.8286$.

6 Conclusion

We have described how assertion symbolic objects can be used to analyse consensus over a set of individuals. Stable solutions of fixed level and q solutions are also presented and algoritms are given to obtain them. We have used the zoom star representation as a graphical representation of the consensus groups. An example is given to illustrate a complete analysis.

References

BOCK, H.H. and DIDAY, E. (Eds.) (2000): *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heidelberg.

BRITO, P. (2000): Hierarchical and Pyramidal Custering with Complete Symbolic Objects. In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data.* Springer Verlag, Heidelberg, 312-323.

COHEN, J. (1960): A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20, 37-46.*

DIDAY, E. and NOIRHOMME-FRAITURE, M. (Eds.) (2008): *Symbolic Data Analysis and the SODAS Software*. Wiley & Sons, Chichester.

HARTIGAN, J. A. (1975): *Clustering Algorithms*. Wiley & Sons, New York.

NOIRHOMME-FRAITURE, M. and ROUARD, M. (2000): Visualizing and Editing Symbolic Objects. In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data.* Springer Verlag, Heidelberg, 125-138.

STÉPHAN, V., HEBRAIL, H. and LECHEVALLIER, Y. (2000): Generation of Symbolic Objects from Relational Databases. In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data.* Springer Verlag, Heidelberg, 78-105.

TASTLE, W.J., WIERMAN, M.J. and DUMDUM, U.R. (2005): Ranking Ordinal Scales Using the Consensus Measure. *Issues in Information Systems VI (2), 96-102.*

Visualizing Exploratory Factor Analysis Models

Sigbert Klinke[1,2] and Cornelia Wagner[2]

[1] Institute for Statistics and Econometrics, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Strasse 1, 10178 Berlin, Germany, *sigbert@wiwi.hu-berlin.de*

[2] Department of Business education, Institute of Education, Faculty of Arts IV, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, *cornelia.wagner@staff.hu-berlin.de*

Abstract. Exploratory factor analysis (EFA) is an important tool in data analyses, particularly in social science. Usually four steps are carried out which contain a large number of options. One important option is the number of factors and the association of variables with a factor. Our tools aim to visualize various models with different numbers in parallel of factors and to analyze which consequences a specific option has. We apply our method to data collected at the School of Business and Economics for evaluation of lectures by students. These data were analyzed by Zhou (2004) and Reichelt (2007).

Keywords: factor analysis, visualisation, questionnaire, evaluation of teaching

JEL classification: C39, C45, C63

1 Introduction

The exploratory factor analysis of a dataset consists of four steps:

1. **estimating the correlation matrix** $\hat{R}$ between the observed p variables. The Bravais-Pearson correlation is the one usually used. For ordinal data Kendall's τ_b, Spearmans rank correlation or polychoric correlation (underlying variable approach, see e.g. Bartholomew, Steele, Moustaki and Galbraith, 2002) can be used.
2. **estimating the loadings matrix** $\hat{A}$ of the common factors. Depending on the beliefs about the data, several extraction methods can be used: principal component (PC), principal axis (PA), maximum likelihood (ML), unweighted least squares (ULS).
3. **estimating the number of common factors** $k < p$. Various criteria are used to find the number of factors: Kaiser (eigenvalues larger than 1), Parallel analysis of Horn (1965), 90% of explained variance and Elbow-criterion.

4. rotating the loadings to improve interpretability. Different rotation methods have been developed, e.g. the varimax rotation if the rotated factors should be uncorrelated and the promax rotation if the rotated factors can be correlated.

2 Visualizations

We have used four plots to obtain information about our variables and factor models:

correlation plot which visualizes the underlying correlation matrix of the variables (see Figure 1). White here represents a small correlation whereas black represents a large absolute correlation. The number of colors and the colors used to represent the entries of the correlation matrix can be chosen freely; default is a color palette ranging from green to blue via white with eleven colors. If we group the highly correlated variables together then we can see which variables will become a factor.

scree plot which is a simple Scree plot added with the decision criteria (see Figure 2) mentioned before. The horizontal grey line represents the Kaiser criterion, the (nearly) horizontal falling line represent the Horn criterion and the vertical lines the 10%, ..., 90% variance criterion. This plot indicates how many factors should be chosen.

factor model plot we can see for each factor model which variables are explained by the same factor (see Figure 3). These variables are combined by a grey horizontal line. A grey square indicates that the absolute factor loading is smaller than a cut-off value (default: 0.5), but still has its absolute maximum loading at this factor. The black square indicates that the absolute factor loading is above the cut-off value. The colored plot version allows to differentiate between between small and large loadings above the cut-off value.

communality plot where each curve represents one factor model and shows how much "variance" is explained by it (see Figure 4). For each model we can see where it improves the variance explanation of a variable.

In all graphics except the scree plot we can choose the order of the variables. In the correlation plot the variables are arranged in such a way that variables with the largest absolute correlation are near to each other. In the factor model plot and communality plot we count, in all computed models, how often variables are explained by the same factor (based on the grey *and* black squares). The variables with the highest counts are placed near to each other.

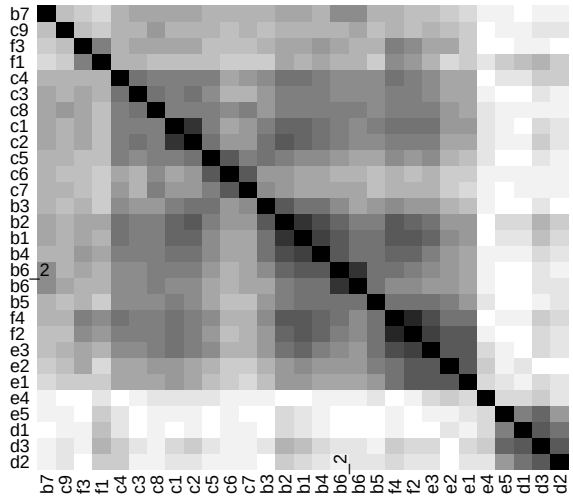


Fig. 1. Left: Tetrachoric correlation of the 29 items in the evaluation data. A darker square means a higher absolute correlation between the items (white: between -0.05 and 0.05, ..., black: below -0.95 or above 0.95).

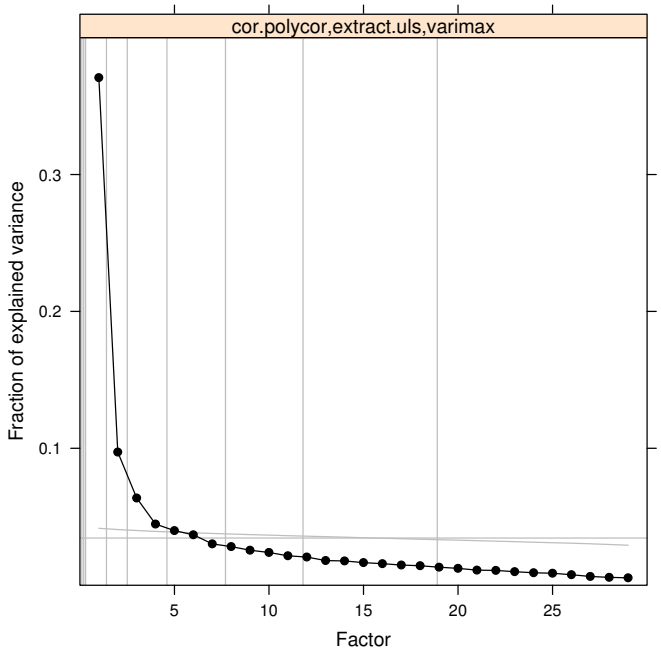


Fig. 2. Scree plot for the evaluation data. The horizontal grey line indicates the Kaiser criterion, the slowly falling, nearly horizontal, grey line the Horn criterion and the vertical grey lines indicate the 10%, 20%, ... up to 90% explained variance lines.

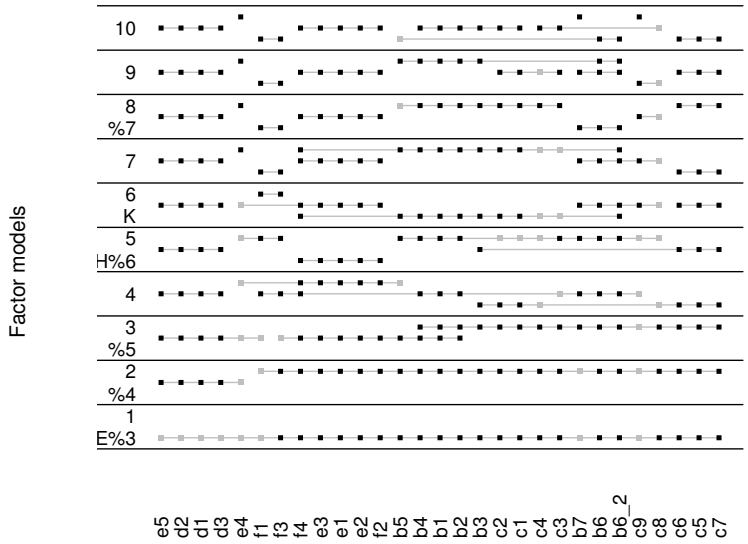


Fig. 3. We see the factor models starting from one up to ten factors. Black squares indicate an absolute loading larger than 0.5, grey squares indicate the largest absolute loading of a variable. Variables explained by the same common factor are connected by a grey line. The letters on the left incorporate information from the scree plot. For example, `%5` below the number `3` means that a factor model with 3 or more factors explains at least 50% of the total variance, `H` stands for Horn criterion, `K` for Kaiser criterion and `E` for (first) ellbow criterion.

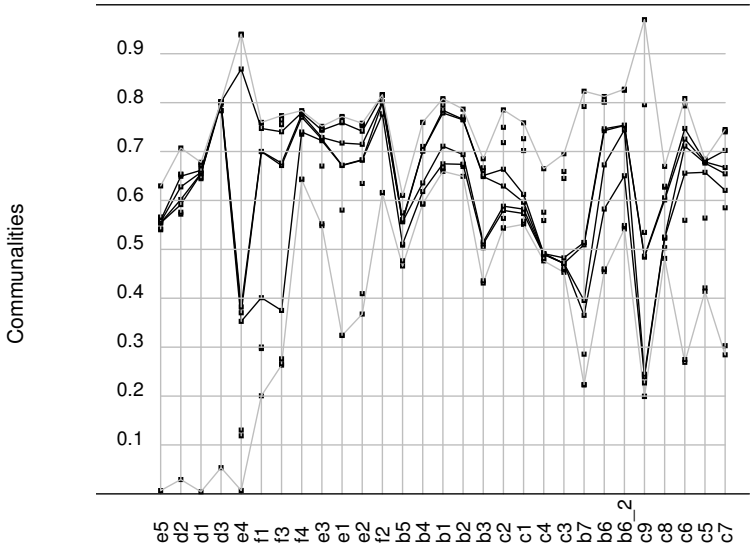


Fig. 4. Communalities of items explained by several factor models. The lower grey line represents the one-factor model, the upper grey line the ten-factor model, the black lines the four-, five-, six- and seven-factor model (from lower to upper).

3 Application

For more than ten years, each semester students of the School of Business and Economics have been asked (see appendix A) to evaluate by questionnaire the lectures they have attended. Questionnaire data for the summer term 2002, 2003, 2005 and 2006 was analyzed by Zhou (2004) and Reichelt (2007). Here we reanalyze the data from lectures in the summer term of 2003 where the questionnaire consisted of 29 questions, each item with five answers ranging from good to bad. The missing values have been replaced by the maximum likelihood for categorical data method as described in Schafer (1997, p. 239ff).

The correlation plot, Figure 1, indicates that we should expect around five factors. One with the variables e5-d3; this is pretty much uncorrelated with all other factors. The other four groups of variables (c4-c5, c6-c7, b3-b6, b5-e1) seem to be correlated to each other.

In the scree plot in Figure2 we identify five (Horn) or sixfactors (Kaiser). Both explain between 60% and 70% of the total variance. Zhou (2004) identified five factors: "communication skill" (b1, b2, b3, b4, c1, c2, c3), "lecture notes" (c5, c6, c7), "course attributes" (d1, d2, d3, e5), "question answering" (b6, b6_2) and "student reactions" (e1, e2, e3, f2, f4). It might also be interesting to look at the seven factor model since the eigenvalue curve here falls down a fraction.

For the factor model plot we therefore decided to visualize all models starting from a one factor model up to a ten factor model. We are currently looking for a set of items which form a factor and which is stable to about several factor models.

Looking for the models, especially the four till seven factor model, we see that

- the items e5, d1, d2 and d3 (course attributes) form a stable factor over nearly all models. Since the questionnaire was carried out four weeks before the exams, the student could also appreciate the speed and difficulties of a course.
- Another set of items is f2, f4, e1, e2 and e3 (student reactions). However, as the item f4 also loads on a different factor, it might be better to exclude it from the factor.
- c5, c6 and c7 also form a stable factor (lecture notes).
- In the eight factor model the items b1-b5 and c1-c4 form one factor (communications skill). However some items turn out to be problematic in earlier models: b3 and b5 belong either to different factors or also load on a different factor.
- Finally, we have two factors (f1, f3 and "question answering": b6, b6_2, b7) with a small number of items.

We end up with four or six factors depending whether we want factors with a small number of items or not. This complies with Zhou's result (2004) that a six factor model is appropriate; her final choice of a five factor model is

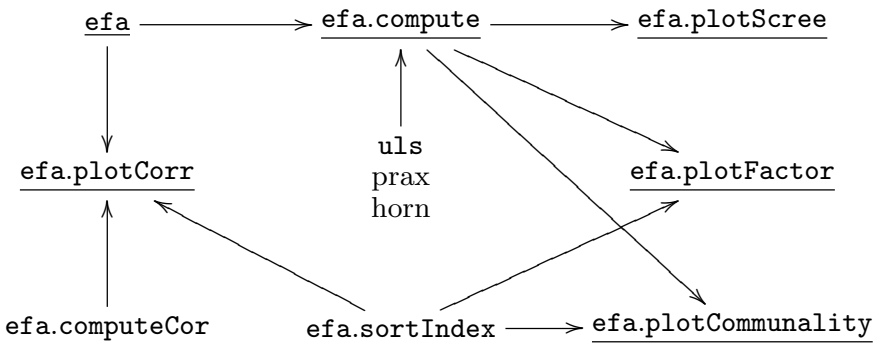


Fig. 5. Shows the relationship of the `efa` functions. The underlined functions are the ones which are usually used.

due to the fact that in the other three datasets she found a five-factor model. The visual analysis of several factor models provides us, via stability analysis and variance explanation for each item, with a more informative and reliable result.

However, an analysis of later data in Reichelt (2007) for the lectures in 2005 and 2006 revealed a high correlation (≈ 0.7) between the factors "communication skills" and "lecture notes" in a promax rotated model. This is also reflected to some extent in the factor models: later split between items means higher correlation between them, see for example the factor "course attributes". This clearly indicates that students tend to make a general judgment about a lecture rather than differentiating its characteristics which in Reichelt (2007) led to a two factor model:

(i) Did the student like the course?
(ii) Did the student consider the course difficult?

4 R functions

To produce our plots we have written several R functions. Figures 5 and 6 show the order how the functions should be applied:

`efa` generates from a data set a R object of class `efa` and computes a correlation matrix. Possible correlations are: `pearson` (default) for Bravais-Pearson correlation, `kendall` for Kendalls τ, `spearman` for Spearmans rank correlation, `cov` for covariance and `uv` for tetrachoric correlation (slow).

`efa.plotCorr` visualizes the correlation between the variables (see Figure 1).

```
vl03_full <- read.csv2("vl03_impute.csv")
vl03      <- vl03_full[,8:36] # extract question answers
efa_vl03  <- efa(vl03, "uv")
efa.plotCorr (efa_vl03)
efa_vl03  <- efa.compute(efa_vl03, factors=10,
              extract="uls", horn=T)
efa.plotScree (efa_vl03)
efa.plotFactor (efa_vl03)
efa.plotCommunality (efa_vl03, modelsep=c(1,4:7,10),
                     col=c("grey", "black", "black",
                       "black", "black", "grey"))
```

Fig. 6. Basic R program to generate the graphics in the paper.

`efa.compute` computes the factor models based on the correlation matrix. Options are `none`, `promax` and `varimax` (default) for rotation, `pc` (principal component), `uls` (unweighted least squares), `mle` (maximum likelihood) and `prax` principal axis for extraction. The parameter `factors` determines the maximal number of factors and can either be a text (`kaiser`, `elbow` or `horn`) or a number. Numbers between zero and one are interpreted as minimal percentage of variance explained and numbers larger than one give the maximal number of factors to be extracted.

`efa.plotFactors` visualizes the computed factor models (see Figure 3),

`efa.plotScree` shows the scree plot with the selection criteria (see Figure 2) and

`efa.plotCommunality` shows the explained variance per variable (see Figure 4).

Additionally some helper functions have been written to realize specific extraction methods etc.:

`uls` unweighted least squares method to compute the factor loadings,

`prax` principal axis method to compute the factor loadings (like in SPSS),

`horn` computes the eigenvalues for a parallel analysis (Horn, 1965),

`efa.computeCor` computes the correlation for an R object of class `efa` and

`efa.sortIndex` computes an order of variables based on a square matrix, e.g. the correlation matrix.

All R functions are still in development, for example the correlation plot needs to be improved, e.g. better standard color palette, legend for connecting colors to correlations values, amongst others.

5 Conclusion

The factor model plot, in particular, will simplify the task of understanding how many factors we can identify with an exploratory factor analysis and

which variables should belong to a factor. It also incorporate steps from a more traditional approach (computing a model, creating scales and computing reliability). If a variable in a scale leads to a too small Cronbachs α, it may load on different factors in different factor models.
With a different questionnaire we were able, based on the factor model plot, to analyze the effect of missing value treatment and to provide a better interpretable factor model.

References

BARTHOLOMEW, D.J., STEELE, F., MOUSTAKI, I. and GALBRAITH, J.I. (2002), *The analysis and interpretation of multivariate data for social Scientists*, Chapman & Hall

HORN, J. L. (1965). *A rationale and test for the number of factors in factor analysis. Psychometrika*, 30, 179-185.

REICHELT, M. (2007), Bewertung von Lehrveranstaltungen mit Hilfe der Evaluationsdaten, Master thesis (in german) at Humboldt-Universität zu Berlin `http://edoc.hu-berlin.de/docviews/abstract.php?id=28130`

SCHAFER, J.L. (1997), *Analysis of incomplete multivariate data*, Chapman & Hall

ZHOU, Y. (2004), Basic Statistical Analysis and Modelling of Evaluation Data for Teaching, Master thesis at Humboldt-Universität zu Berlin `http://edoc.hu-berlin.de/docviews/abstract.php?id=26957`

A Evaluation questionnaire

Lecturer		
b1	Explain ability	
b2	Content clarity	
b3	Transparency quality	
b4	Didactical ability	
b5	Stimulation of independent thought	
b6	Willingness to answer questions	
b6_2	Quality of answered questions	
b7	Time allowed after course	
Lecture Concept		
c1	Aspects covered deepness	
c2	Topic structure clarity	
c3	Related topics reference	
c4	Practical example application	
c5	Choice of lecture notes	
c6	Availability of lecture notes	
c7	Presence in the internet	
c8	Content update	
c9	Relevance between lecture and exercise	
Course attributes		
d1	Lecture speed	
d2	Mathematical level	
d3	Difficulty	
Self assessment		
e1	Interest degree	
e2	Attention span	
e3	Knowledge increase	
e4	Preparation level	
e5	Challenging feeling	
Course atmosphere		
f1	Atmosphere-stress level	
f2	Atmosphere-interest degree	
f3	Atmosphere-disciplined degree	
f4	Atmosphere- motivation level	

For the questionnaire form and coding see Zhou (2004), page 64 and 70.

A Cluster-Based Approach for Sliced Inverse Regression

Vanessa Kuentz[1] and Jérôme Saracco[2]

[1] Universités Bordeaux 1 et 2, IMB, UMR CNRS 5251,
351 Cours de la Libération, 33405 Talence Cedex, France,
vanessa.kuentz@math.u-bordeaux1.fr
[2] Université Montesquieu - Bordeaux 4, GREThA, UMR CNRS 5113,
Avenue Léon Duguit, 33608 Pessac Cedex, France,
jerome.saracco@u-bordeaux4.fr

Abstract. In the theory of sufficient dimension reduction, Sliced Inverse Regression (SIR) is a famous technique that enables to reduce the dimensionality of regression problems. This semiparametric regression method is based on a linearity condition on the marginal distribution of the predictor $\mathbf{x}$, which appears to be a limitation. Using an idea of Li et al. (2004), we propose to cluster the predictor space so that this condition approximately holds in the different partitions. We estimate the dimension reduction subspace by combining the individual estimates of the clusters. We give asymptotic properties of the corresponding estimator and show with a simulation study the numerical performances of cluster-based SIR.

Keywords: Sliced Inverse Regression (SIR), effective dimension reduction (e.d.r.) space, clustering, linearity condition

1 Introduction

Many dimension reduction tools in regression context assume that the features of a covariable $\mathbf{x} = (x^1, \ldots, x^p)'$, with $\mathbb{E}(\mathbf{x}) = \mu$ and $\mathbb{V}(\mathbf{x}) = \Sigma$, can be captured in a lower K-dimensional projection subspace (with $K < p$), such as Sliced Inverse Regression (SIR) methods introduced by Li (1991). The underlying semiparametric model assumes that the dependency between the predictors and the response variable y is described by linear combinations of the predictors. It is written:

$$y = f(\mathbf{x}'\beta_1, \ldots, \mathbf{x}'\beta_K, \varepsilon), \tag{1}$$

where f is an unknown function, ε is an unknown random error independent of $\mathbf{x}$, and $\beta_1, \ldots, \beta_K$ are K unknown vectors in $\mathbb{R}^p$, assumed to be linearly independent. As none condition on the form of f is imposed, it is only possible to estimate the space spanned by the vectors β_k, called the effective dimension reduction (e.d.r.) space, which will be denoted by E. When K is small ($K << p$), the goal of reduction theory is achieved and we can project

the p-dimensional regressor $\mathbf{x}$ onto this K-dimensional space without loss of information on the feature of y given $\mathbf{x}$. Then it will be easier to study the relationship between $\mathbf{x}$ and y via a nonparametric estimation of the regression of y on the corresponding K-dimensional variable.

The basic principle of SIR methods is to reverse the role of y and $\mathbf{x}$ and to study the property of the conditional moments of $\mathbf{x}$ given y. In this paper, we will only focus on the SIR-I method which is based on the first conditional moment. Li (1991) has shown that the eigenvectors associated with the non-null K eigenvalues of the matrix $\Sigma^{-1}M_I$, where $M_I = \mathbb{V}(\mathbb{E}(\mathbf{x}|T(y)))$ and T denotes a slicing on the variable y, are e.d.r. directions. One important point in SIR-I theory is the underlying crucial linearity condition:

$$\mathbb{E}(\mathbf{x}'b|\mathbf{x}'\beta_1, \ldots, \mathbf{x}'\beta_K) \text{ is linear in } \mathbf{x}'\beta_1, \ldots, \mathbf{x}'\beta_K \text{ for any } b. \tag{2}$$

This condition is hard to verify in practice since it involves the unknown directions of the e.d.r. space. However it can be proved that (2) is verified when $\mathbf{x}$ follows an elliptically symmetric distribution, condition which is stronger in theory but easier to verify in practice.

If the collected data set does not follow an elliptically distribution, solutions exist to force data to behave as if they were issued from such a distribution, see for instance Cook and Nachtsheim (1994). Problems with this technique are that it can severely reduce the sample size and that it is difficult to put into practice if $\mathbf{x}$ is high-dimensional. So from a theoretical and practical point of view, the linearity condition appears to be a limitation.

In this paper we propose to cluster the predictor space, which will force the linearity condition to hold approximately in each cluster. The idea is inspired by the work of Li et al. (2004), who proposed a cluster-based Ordinary Least Squares (OLS) approach for single index models ($K = 1$). It consists in partitioning the predictor space with a k-means algorithm, evaluating the OLS estimate of each cluster and finally pooling them so as to provide an efficient estimation of the central mean subspace. In our approach, we also partition the predictor space into disjoint clusters with a k-means algorithm, which aims at constructing approximately elliptical clusters. Then we estimate the e.d.r directions in each cluster and combine them to produce an efficient estimation of the e.d.r space of model (1). The proposed approach will be refered in the rest of the paper as cluster-based SIR.
In Section 2, we consider the case of single index model, we describe the population and sample approaches of the cluster-based SIR. We show the convergence in probability and the asymptotic distribution of the corresponding estimator of the e.d.r. direction. We extend this approach to multiple indices models in Section 3, where the dimension K is assumed to be known. A simulation study is carried out in Section 4 in order to show the numerical performance of the approach and to compare it to SIR. Finally concluding remarks are given in Section 5.

2 Approach for single index model

We consider in this section single index model ($K = 1$). The corresponding model is:

$$y = f(\mathbf{x}'\beta, \varepsilon). \tag{3}$$

So we focus on the estimation of only one e.d.r. direction b colinear to β.

The idea of the proposed approach is to partition the predictor space into a fixed number c of clusters. By doing that, the linearity condition will approximately hold in each cluster. For each one, we compute the e.d.r. direction with SIR. Finally we combine these directions to find the e.d.r direction of model (3) taking into account the whole space.

2.1 Population version

Let us consider a fixed number c of clusters and let us assume that $\mathbf{x}$ is partitioned into c clusters $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(c)}$. Accordingly to the partitioning scheme of $\mathbf{x}$, we get the partition $(\mathbf{x}^{(j)}, y^{(j)}), j = 1, \ldots, c$ of $(\mathbf{x}, y)$. Let us assume that the linearity condition holds in each cluster:

(LC) For $j = 1, \ldots, c$, $\mathbb{E}(\mathbf{x}^{(j)\prime} b | \mathbf{x}^{(j)\prime}\beta)$ is linear in $\mathbf{x}^{(j)\prime}\beta$ for any b.

In each cluster j, let $T^{(j)}$ be the slicing of $y^{(j)}$ into $H^{(j)}$ fixed slices, $s_1^{(j)}, \ldots, s_{H^{(j)}}^{(j)}$, with $H^{(j)} > 1$. From this slicing, the matrix $M_I^{(j)}$ can be written as $M_I^{(j)} = \sum_{h=1}^{H^{(j)}} p_h^{(j)} (m_h^{(j)} - \mu^{(j)})(m_h^{(j)} - \mu^{(j)})'$, where $p_h^{(j)} = P(y^{(j)} \in s_h^{(j)})$, $m_h^{(j)} = \mathbb{E}(\mathbf{x}^{(j)} | y^{(j)} \in s_h^{(j)})$ and $\mu^{(j)} = \mathbb{E}(\mathbf{x}^{(j)})$. Let $\Sigma^{(j)} = \mathbb{V}(\mathbf{x}^{(j)})$. The eigenvector $b^{(j)}$ associated with the largest eigenvalue of the matrix $(\Sigma^{(j)})^{-1} M_I^{(j)}$ is an e.d.r. direction. We define the matrix $B = [b^{(1)}, \ldots, b^{(c)}]$ and we note b the first left singular vector of this matrix. Then Theorem 1 guarantees that this vector is an e.d.r. direction.

Theorem 1. *Assuming the linearity condition (LC) and model* (3)*, the major eigenvector b of the matrix BB' is colinear with β.*

PROOF. Since each $b^{(j)}$ is colinear with β, we have: $B = \alpha' \otimes \beta$ for a non-null vector $\alpha \in \mathbb{R}^c$, and then $BB' = \|\alpha\|^2 \beta\beta'$. Therefore the eigenvector b associated with the largest eigenvalue of BB' is colinear with β.

2.2 Sample version

Let $\mathcal{S} = \{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$ be a sample from the reference model (3). We partition these observations into c clusters using a k-means approach. So for $j = 1, \ldots, c$, we get samples $\mathcal{S}^{(j)} = \{(y_i^{(j)}, \mathbf{x}_i^{(j)\prime}), i = 1, \ldots, n^{(j)}\}$, where $n^{(j)}$ denotes the number of observations in the jth cluster. We use empirical means, variances and proportions to estimate the covariance matrix

$\widehat{M_I}^{(j)}$. Then the eigenvector $\hat{b}^{(j)}$ associated with the largest eigenvalue of $(\widehat{\Sigma}^{(j)})^{-1}\widehat{M_I}^{(j)}$ is the estimated e.d.r. direction in the jth cluster. We construct the matrix $\widehat{B} = [\hat{b}^{(1)}, \ldots, \hat{b}^{(c)}]$. The major eigenvector $\hat{b}$ of the matrix $\widehat{B}\widehat{B}'$ is then the e.d.r. estimated direction in model (3).

2.3 Asymptotic theory

The assumptions that are necessary to state our results are gathered below for easy reference.
(A1) The sample $\mathcal{S}$ is a sample of independent observations from the single index model (3) or the multiple indices model (1).
(A2) $\mathbf{x}$ is partitioned into c fixed clusters $\mathbf{x}^{(j)}, j = 1, \ldots, c$, such that $\cup_{j=1}^{c}\mathcal{S}^{(j)} = \mathcal{S}$ and $\forall j \neq l, \mathcal{S}^{(j)} \cap \mathcal{S}^{(l)} = \emptyset$.
(A3) The support of $y^{(j)}$ is partitioned into a fixed number $H^{(j)}$ of slices such that $p_h^{(j)} \neq 0, h = 1, \ldots, H^{(j)}$.
(A4) For $j = 1, \ldots, c$, $n^{(j)} \to \infty$ as $n \to \infty$.

Theorem 2. *Under the linearity condition (LC) and the assumptions **(A1)-(A4)**, we have:*
(a) $\hat{b} = b + O_p(n^{-1/2})$*, where* b *is an e.d.r. direction (colinear with* β*),*
(b) $\sqrt{n}(\hat{b} - b) \longrightarrow_d U \sim \mathcal{N}(0, \Gamma_U)$*, where the expression of* Γ_U *can be found in Kuentz and Saracco (2007).*

2.4 Optimal number of clusters

In practice, a crucial step in the proposed method is the choice of the number c of clusters for the partitioning of the predictor space. The choice of an optimal number c^* of clusters can be defined through the following optimization problem:

$$c^* = \arg \min_{c=1,\ldots,C} \mathbb{E}((y - \mathbb{E}(y|\mathbf{x}'\hat{b}_{[c]}))^2) \tag{4}$$

where $\hat{b}_{[c]}$ denotes the estimator of the e.d.r. direction when the number of clusters is c.
From a practical point of view, we consider an empirical smoothed version of this minimization problem:

$$\hat{c}^* = \arg \min_{c=1,\ldots,C} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,[c]})^2 \tag{5}$$

where $\hat{y}_{i,[c]} = \sum_{j=1}^{n} y_j \mathcal{K}((\mathbf{x}_i'\hat{b}_{[c]} - \mathbf{x}_j'\hat{b}_{[c]})/h_c) / \sum_{j=1}^{n} \mathcal{K}((\mathbf{x}_i'\hat{b}_{[c]} - \mathbf{x}_j'\hat{b}_{[c]})/h_c)$ is a kernel estimation of $\mathbb{E}(y|\mathbf{x}_i'\hat{b}_{[c]})$, for which h_c is the bandwidth parameter for a partitioning into c clusters and $\mathcal{K}$ is a kernel (the density of the standard univariate normal distribution for instance). The bandwidth parameters $h_c, c = 1, \ldots, C$, can be chosen by cross validation.

3 Extension to multiple indices model

In this section, we extend the proposed approach to multiple indices model ($K > 1$). We assume here that the dimension K is known. The corresponding model is given in (1). We search for a basis that spans the e.d.r. space $E = \text{Span}(\beta_1, \ldots, \beta_K)$.

3.1 Population version

As for the single index model, we partition the predictor space $\mathbf{x}$ into c clusters. We get the partitions $(\mathbf{x}^{(j)}, y^{(j)}), j = 1, \ldots, c$. For each cluster, we seek with SIR a basis of the e.d.r. space. Let us assume that the following linearity condition (LC*) holds:

(LC*) For $j = 1, \ldots, c$, $\mathbb{E}(\mathbf{x}^{(j)\prime} b | \mathbf{x}^{(j)\prime}\beta_1, \ldots, \mathbf{x}^{(j)\prime}\beta_K)$ is linear in $\mathbf{x}^{(j)\prime}\beta_1, \ldots, \mathbf{x}^{(j)\prime}\beta_K$ for any b.

The eigenvectors $b_1^{(j)}, \ldots, b_K^{(j)}$ associated with the largest K eigenvalues of the matrix $(\Sigma^{(j)})^{-1} M_I^{(j)}$ are e.d.r. directions, where matrices $\Sigma^{(j)}$ and $M_I^{(j)}$ have been defined in Section 2. We define the matrix $B^{(j)} = [b_1^{(j)}, \ldots, b_K^{(j)}]$ containing these e.d.r. directions, which form a $\Sigma^{(j)}$-orthogonal basis of E. Then the first K eigenvectors of the matrix $B^{(j)} B^{(j)'}$, denoted by $\tilde{b}_1^{(j)}, \ldots, \tilde{b}_K^{(j)}$, form an I_p-orthonormal basis of E. We store these vectors in the matrix $\tilde{B}^{(j)} = [\tilde{b}_1^{(j)}, \ldots, \tilde{b}_K^{(j)}]$. We can now pool the matrices $\tilde{B}^{(j)}$ in the matrix $\mathbb{B}^{(c)} = [\tilde{B}^{(1)}, \ldots, \tilde{B}^{(c)}]$. The first K eigenvectors of the matrix $\mathbb{B}^{(c)} \mathbb{B}^{(c)'}$ are denoted by $\tilde{\tilde{b}}^{(1)}, \ldots, \tilde{\tilde{b}}^{(K)}$.

Theorem 3. *Assuming the linearity condition (LC*) and model* (1)*, the vectors* $\tilde{\tilde{b}}^{(1)}, \ldots, \tilde{\tilde{b}}^{(K)}$ *form an* I_p*-orthogonal basis of the e.d.r. space* E*.*

PROOF. Since $\tilde{b}_1^{(j)}, \ldots, \tilde{b}_K^{(j)}$ form an I_p-orthonormal basis of E, we have $\text{Span}(\mathbb{B}^{(c)}) = E$. Then the eigenvectors associated with the K largest eigenvalues of $\mathbb{B}^{(c)} \mathbb{B}^{(c)'}$ form an I_p-orthonormal basis of E.

3.2 Sample version

As for the single index model, we estimate in each cluster a basis of the e.d.r space: the first K eigenvectors of the matrix $(\widehat{\Sigma}^{(j)})^{-1} \widehat{M_I}^{(j)}$, defined in Section 2. These vectors form a $\widehat{\Sigma}^{(j)}$-orthogonal basis of the estimated e.d.r. space. We store them in the matrix $\widehat{B}^{(j)} = [\hat{b}_1^{(j)}, \ldots, \hat{b}_K^{(j)}]$. Then the first K eigenvectors of the matrix $\widehat{B}^{(j)} \widehat{B}^{(j)'}$, denoted by $\hat{\tilde{b}}_1^{(j)}, \ldots, \hat{\tilde{b}}_K^{(j)}$, form an I_p-orthogonal basis of the estimated e.d.r. space. We store them in the matrix $\hat{\tilde{B}}^{(j)} = [\hat{\tilde{b}}_1^{(j)}, \ldots, \hat{\tilde{b}}_K^{(j)}]$. Let $\hat{\mathbb{B}}^{(c)} = [\hat{\tilde{B}}^{(1)}, \ldots, \hat{\tilde{B}}^{(c)}]$. Finally the first

K eigenvectors of the matrix $\hat{\mathbb{B}}^{(c)}\hat{\mathbb{B}}^{(c)'}$, denoted by $\hat{\tilde{b}}^{(1)}, \ldots, \hat{\tilde{b}}^{(K)}$, form an I_p-basis of the estimated e.d.r. space.

Theorem 4. *Under the linearity condition (LC*) and the assumptions **(A1)**-**(A4)**, we have $\hat{\tilde{b}}^{(k)} = \tilde{\tilde{b}}^{(k)} + O_p(n^{-1/2})$, then the estimated e.d.r. basis converges to an e.d.r. basis at root n rate.*

As for the single index model, using Delta-method and asymptotic results of Saracco (1997) and Tyler (1981), the asymptotic normality of the eigenprojector onto the estimated e.d.r. space can be obtained, as well as the asymptotic distribution of the estimated e.d.r. direction, associated with eigenvalues assumed to be different.

Optimal number of clusters. To choose the optimal number of clusters, we can use the method proposed for single index models, where the kernel $\mathcal{K}$ is replaced by a multidimensional one. Note that this approach is sound for $K = 1$ or 2. For higher dimension, this kernel approach suffers from the curse of dimensionality. Note that in practice, the choice of the dimension is often lower than $K \leq 2$.

4 Simulation study

A simulation study has been carried out to evaluate the numerical performance of the proposed method and to compare it to SIR. We first recall the definition of the efficiency measure. In our simulations, we first consider a single index model and then a multiple indices model (with K=2).

4.1 Efficiency measure in simulation study

Let $\check{b}_1, \ldots, \check{b}_K$ be the K estimated e.d.r. directions. We note $\check{B} = [\check{b}_1, \ldots, \check{b}_K]$ and $\check{E} = \text{Span}(\check{B})$ the linear subspace spanned by the $\check{b}_k$'s. Let $B = [\beta_1, \ldots, \beta_K]$ be the matrix of the true directions and let $E = \text{Span}(B)$. Let P_E (resp. $P_{\check{E}}$) be the I_p-orthogonal projector onto E (resp. $\check{E}$). Since with cluster-based SIR approach we construct an I_p-orthonormal basis of E and if we assume the β_k's such that $B'B = I_K$, the expression of the projectors reduces to: $P_E = BB'$ and $P_{\check{E}} = \check{B}\check{B}'$. Since in simulation B is known, we can measure the quality of the estimate $\check{E}$ of E by:

$$m(E, \check{E}) = \text{Trace}(P_E P_{\check{E}})/K. \tag{6}$$

Then the closer this value is to one, the better is the estimation. When $K = 1$ (single index model), this measure is the squared cosine of the angle formed by the vectors β and $\check{b}$. Note that this measure can not be used in a real case study, when the true e.d.r. space is unknown.

4.2 Single index model

First we define the simulated model, then we describe our approach on multiple data replications for which the linearity condition may be seriously violated or not.

We consider the following regression model:

$$y = \exp(x_1 - x_2) + \varepsilon, \tag{7}$$

with $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)'$, where $x_j \sim (1-\theta) \times Exp(1) + \theta \times \mathcal{N}(0,1)$ and $\varepsilon \sim \mathcal{N}(0, 0.5^2)$. The variables x_j are mutually independent and the error term ε is independent of $\mathbf{x}$. In this model, the true normalized direction is $\beta = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0, 0, 0, 0, 0)'$. In our simuations, the parameter θ will belong to the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The value 0 corresponds to non elliptical distribution (in this case the linearity condition is not verified) and 1 to multinormal distribution.

In this study, we compare SIR and cluster-based SIR on $N = 100$ data replications of model (7). The number n of observations will be 100, 200, 500 and 1000. For each simulated sample, the e.d.r. direction is estimated with SIR and cluster-based SIR. Cluster-based SIR was implemented with a number of clusters c varying from 1 to 10 (or 20 for $n = 1000$). In this simulation study, the optimal number of clusters was chosen according to criterion (6). We could also use the criterion (5) which gives very similar results but is computationally expensive. The quality measure presented for cluster-based SIR is the one obtained with the optimal number of clusters. Note that the best number may sometimes be equal to 1 (especially for $n = 100$), corresponding then to classical SIR.

Figure 1 shows the mean of the $N = 100$ squared cosines obtained for each of 6 values of θ (from 0 to 1) with SIR and cluster-based SIR estimation methods.

In each case, both methods give reliable results. For the four sample sizes, the performances of both methods increase as θ increases, that is as the data are close to be elliptically distributed ($\theta = 1$). This shows that cluster-based SIR is above all helpfull in case of non ellipticity, which was the aim of the proposed work. Moreover nothing is lost in case of elliptical distribution. We also see that the performances of both methods increase as the sample size gets higher. Concerning cluster-based SIR, this comes from the fact that the proposed approach partitions the predictor space. Indeed with large samples, the clustering is better: clusters are better defined and bigger. Therefore the slicing in SIR step occurs on a large number of observations. On the contrary with a small number of observations, the clustering is not so clear and provides sometimes clusters too small for the slicing to be computed.

Two indices model. We also made a simulation study with a two indices model. Cluster-based SIR always provides better estimations than classical

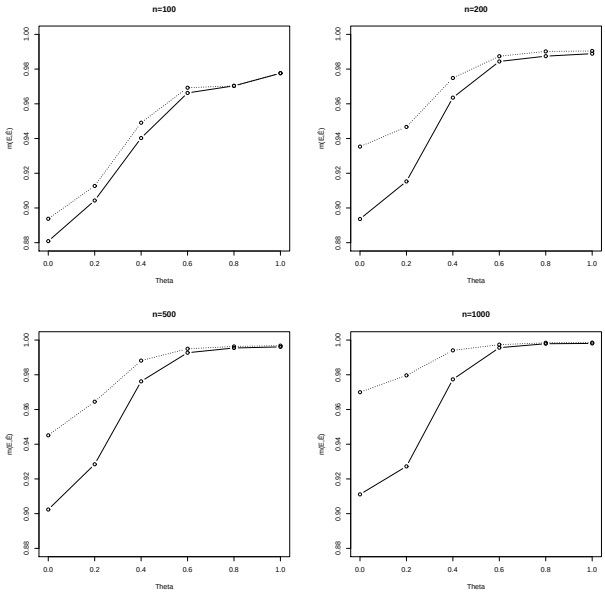


Fig. 1. Plots of the mean of the squared cosines for different values of θ and n (solid line: SIR, dotted line: cluster-based SIR).

SIR. Compared to SIR, cluster-based SIR is less sensitive to violation of the linearity condition (or elliptical distribution). The detailed results can be found in Kuentz and Saracco (2007).

{Conclusion We have proposed an extension of the well-known dimension reduction method SIR, called cluster-based SIR, which can be used when the crucial linearity condition is not verified. Asymptotic properties of the estimator have been obtained. A simulation study has shown the good numerical behaviour of the proposed approach. The optimal number of clusters can be computed from a minimization criterion. We have observed that cluster-based SIR is less sensitive than SIR to violation of the linearity condition. Thus it opens future prospects for a broader use of SIR.

References

COOK, R.D. and NACHTSHEIM, C.J. (1994): Re-weighting to achieve elliptically contoured covariates in regression *Journal of the American Statistical Association 89, 592-599.*

KUENTZ, V. and SARACCO, J. (2007): Cluster-based Sliced Inverse Regression, *submitted paper.*

LI, K.C. (1991): Sliced inverse regression for dimension reduction, with discussion, *Journal of the American Statistical Association 86, 316-342.*

LI, L., COOK, R.D. NACHTSHEIM, C.J., (2004): Cluster-based estimation for sufficient dimension reduction, *Computational Statistics & Data Analysis 47, 175-193.*

SARACCO, J. (1997): An asymptotic theory for Sliced Inverse Regression, *Communications in Statistics - Theory and Methods 26, 2141-2171.*

TYLER, D.E. (1981): Asymptotic inference for eigenvectors, *The Annals of Statistics 9, 725-736.*

New Selection Criteria and Interface in a Variable Selection Environment VASMM

Yuichi Mori[1], Liang Zhang[2], Kaoru Fueda[2] and Masaya Iizuka[2]

[1] Faculty of Informatics, Okayama University of Science. 1-1 Ridai-cho, Okayama 700-0005, Japan, *mori@soci.ous.ac.jp*
[2] Graduate School of Environmental Science, Okayama University. 3-1-1, Tsushima Naka, Okayama 700-8530, Japan, *zhang@stud.ems.okayama-u.ac.jp*, *fueda@ems.okayama-u.ac.jp*, *iizuka@ems.okayama-u.ac.jp*

Abstract. A statistical environment VASMM (VAriable Selection in Multivariate Methods) provides useful computational tools and information for selecting a reasonable subset of variables in multivariate methods without external variables. Currently new selection criteria have been proposed for some practical applications and implemented in VASMM and a new interface has been developed for Excel so that a variety of users can perform variable selection easily. The paper gives an overview of VASMM including selection criteria implemented in VASMM, flow of computation and three versions of VASMM at first and shows practical works of VASMM by illustrating how the Excel interface selects variables when one of new criteria is applied to a real data set.

Keywords: statistical tools using R, interface for Excel, RExcel, estimation of principal component scores

1 Introduction

We sometimes meet the problem of variable selection in multivariate methods (MMs) such as dimension reduction methods including principal component analysis (PCA), factor analysis (FA) and correspondence analysis (CA), for example, where we wish to select a subset of variables in a particular test so as to produce a simplified version of the test without losing the amount of information of whole variables. In practical data analysis in such situations, various ad hoc methods are used. Because MMs such as dimension reduction methods generally specify no external (response or dependent) variable in the analysis, the results usually differ depending on which method is applied. For that reason, it will be valuable to provide objective selection methods/criteria and the corresponding software to select variables in MMs without external variables (we hereafter abbreviate "without external variables" as "–ev").

Several selection methods and criteria have been proposed so far and new criteria are studied continuously. However, we had no device to perform any variable selection method easily except MMs with external variables. Against this background, we developed a statistical program for variable selection in

MMs–ev; in 1999 we developed a web system VASpca (VAriable Selection in PCA) using R as a statistical engine (see, e.g., Mori et al. (2000)), and in 2002 we developed VASfa (VAriable Selection in FA) and VAScorres (VAriable Selection in CA) and integrated them as VASMM (VAriable Selection in MMs) (see, e.g., Iizuka et al. (2002)). Since this integration we have included a local version (an R function with interactive interface) as well as a web version. Recently we have developed an interface for Microsoft Excel (Excel version of VASMM) to perform variable selection via only Excel with R in the background. This allows users to apply any selection method/criterion directly to Excel data and to make the most use of Excel functions for outputs.

In this paper, we give an overview of VASMM including new implemented criteria and a new interface in Section 2 and illustrate an example of variable selection in PCA using a new criterion and interface in Section 3.

2 Statistical environment VASMM

VASMM site is `http://mo161.soci.ous.ac.jp/vasmm/`. This site provides various information on variable selection in MMs–ev and links to three subsystems, VASpca, VASfa, and VAScorres.

2.1 Implemented selection methods/criteria

Users can perform variable selection in PCA, FA and CA based on any method/criterion among the followings:

- VASpca (variable selection in PCA)
  - (i) Proportion P / RV-coefficient (criteria in Modified PCA by Tanaka and Mori (1997))
  - (ii) Principal variables (McCabe (1984))
  - (iii) Procrustes analysis (Krzanowski (1987))
  - (iv) RV-coefficient (Robert and Escoufier (1976))
  - (v) PRESS (prediction error) (Mori et al. (2000))
  - (vi) Loadings (B2 and B4 in Jolliffe (1972))
  - (vii) Influence analysis of variables (Mori et al. (2000))
  - (viii) Estimation of principal component scores (Mori et al. (2004b))
  - (ix) Proportion P / RV-coefficient for qualitative variables (Mori et al. (2007))
- VASfa (variable selection in FA)
  - (i) Tanaka's D / Q / RV-coefficient (configurations of factor scores) (Tanaka (1983); Iizuka et al. (2002))
  - (ii) Using estimation of factor scores (Mori et al. (2004b))
- VAScorres (variable selection in CA)
  - (i) PCE / PCS / PCO (goodness of fit criteria) (Mori et al. (2004a))
  - (ii) Estimation of individual scores (Mori et al. (2004b))

New criteria implemented currently in VASMM are No.8 in VASpca, No.2 in VASfa and No.2 in VAScorres, which are based on the same selection idea, and No.9 in VASpca.

Here we briefly show the former idea which will be used in Section 3. This idea was originally proposed in Mori et al. (2004b). We computed using Matlab at the beginning but we trasferred from Matlab to R in 2005 and implemented the idea into VASMM recently. The idea is as follows. (See Mori et al. (2007) for the latter idea.)

Selection criterion – using global scores estimation
Consider a situation where observations have been evaluated using global scores such as principal component (PC) scores in PCA, based on all variables/items/questions in some survey but the smaller number of variables would be used in successive surveys to reduce the measurement cost. We can consider similar situations for factor scores in FA and individual scores in CA. In such cases, it is expected that the small number of variables should be selected reasonably from the previous survey so that the global scores based on the selected variables, to the greatest degree possible, represent information based on all preceding variables. Here we propose the following idea as a suitable selection criterion for the situation.

Suppose we observe p variables $\mathbf{y}_1, \ldots, \mathbf{y}_p$ and focus on the first r ($1 \leq r < p$) global scores $\mathbf{z}_1, \ldots, \mathbf{z}_r$. Using q variables $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_q}$, which is a subset of $\mathbf{y}_1, \ldots, \mathbf{y}_p$, we compute the r global scores $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_r$ that can approximate the original r global scores $\mathbf{z}_1, \ldots, \mathbf{z}_r$ as much as possible in the context of least squares. When $\mathbf{z}_1, \ldots, \mathbf{z}_r$ are uncorrelated, $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_r$ are restricted to being uncorrelated with each other. Under this restriction, we can take the following three computational steps to estimate $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_r$.

Step 1 : $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_r$ can be given as the least-squares estimator of $\mathbf{z}_1, \ldots, \mathbf{z}_r$ in the same way as ordinary regression analysis, in which $\mathbf{z}_1, \ldots, \mathbf{z}_r$ are regarded as dependent variables and $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_q}$ independent ones. For $r = 1$ or in the case where the correlation between global scores is not considered, the solution in this step is sufficient.

Step 2 : For $r \geq 2$ in Step 1, because $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_r$ are not guaranteed to be uncorrelated with each other, we consider an additional computational step including the orthogonal problem. We first find the best estimator for the most important score (normally the 1st 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. We can compute the estimator exactly by applying Gram-Schmidt orthogonalization to the estimator given in Step 1.

Step 3 : The column-orthogonality proceeds successively, as in Step 2. No order of importance exists in r scores in many cases. For that reason, we must find the best solution that provides an estimator matrix whose columns are mutually uncorrelated. To do this, we apply an iteration technique, Givens transformation, to the estimators found in Step 2.

Since the above computation naturally includes variable selection, we can use this idea as a selection criterion. We compute the residual sum of squares (RSS) for all possible subsets of q variables and find a subset providing the smallest RSS among them. The subset is the best subset of size q by which the global scores based on the whole variables can be represented as much as possible. However, Step 3 engenders a high computational cost. Therefore, Step 1 or Steps 1 + 2 might be applicable as a convenient means to save the cost, although neither retains orthogonality completely. The cost-saving selection procedures are applicable to further reduce computational time.

2.2 Three versions of VASMM

VASMM provides three types of computational tools; a web version, a local version and a new version for Excel. All versions are programmed with R.

Web version of VASMM: This version is an online analysis system on the web (Figure 1). The system is controlled by CGI with R server as a statistical engine. The web version has advantages in that the developers can maintain the system and update information rapidly and timely as well as in that users can perform variable selection through the web anytime and anywhere. On the other hand, closed data sets and large data sets are not suitable for online computation and we set some restrictions to the version because of saving the computational resources; single-user mode, limited data/file size and upper limit to the parameter values. So we currently think it better that users use this version for a trial or a small data set, although all of the functions for the implemented methods/criteria can be available.

Local version of VASMM: This is a set of R functions to select subsets of variables (Figure 2). The main function `vasmm()` includes not only functions necessary for variable selection but also step-by-step and interactive interfaces using GUI functions in R so that users obtain the results by minimum inputs in specifying a data set and parameters. Using this function users can perform variable selection without difficulties on their own local computers. The R image file and source codes are provided in the VASMM site.

Excel version of VASMM: This is an new version of VASMM - an interface to Microsoft Excel, which has been developed using RExcel (Baier and Neuwirth, `http://sunsite. univie.ac.at/rcom/`). RExcel allows to use R as a helper application or a statistical engine for Excel, that is, RExcel transfers data from Excel to R, calls R functions from Excel and then transfers the results to Excel. We have developed an Excel interface for variable selection using VBA, in which RExcel is installed. This allows users, especially who are not familiar with R and/or who want to use Excel functions in data handling and output processing, to apply any selection method/criterion directly to Excel data and to obtain the results directly on Excel sheets. The Excel version of VASMM is provided as an Excel add-in file in the VASMM site.

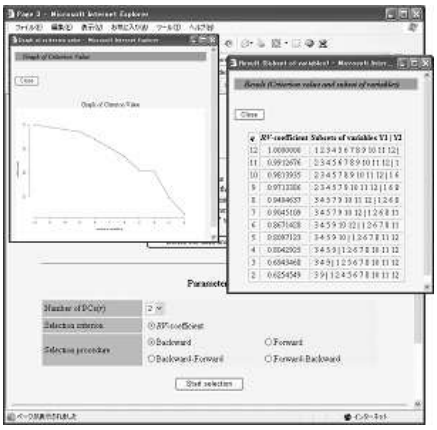


Fig. 1. Web version of VASMM.

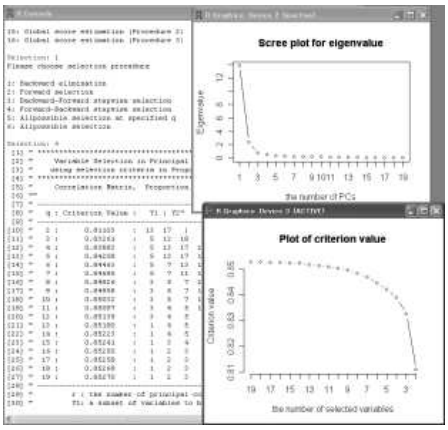


Fig. 2. Local version of VASMM.

3 An example: Variable selection using global scores estimation – Excel version

A Japanese newspaper company sends questionnaires to enterprises in Japan and applies PCA to the collected answers to rank enterprises using the first PC score as an environmental management index and to draw a scatter plot using the first two PC scores as an environmental management map (Nikkei (1997–2006)). The number of items have been changed every year according to trends and interests. In practice, when the 5th survey was conducted, four items were deleted from the 4th survey and one item was added.

We apply the criterion No.8 in VASpca described in Section 2.1 to 11 variables, {V1: reducing, V2: resource conservation, V3: waste control, V4: prevention of global warming, V5: chemical product control, V6: pollution prevention, V7: products and distribution control, V8: organization system, V9: environmental management system, V10: environmental report and accounting, V11: environmental education}, on 791 enterprises in the 4th survey to find the best subset of size 7 to be assigned to the 5th survey.

Here we use Excel version of VASMM to obtain a reasonable subset.

Preparation:

() At first we install RExcel and Excel version of VASMM to Excel. They are recognized as add-in software in Excel.

Initial stage:

(i) We open the data in usual way and select [VASpca] from [VASMM] menu in the tool bar. A dialogue box appears to ask us to select data range, matrix to be used and output format of graphs (Figure 3).

(ii) We click [Next] button and then Excel version of VASMM applies ordinary PCA to the data. A new Excel sheet is created including eigenvalues,

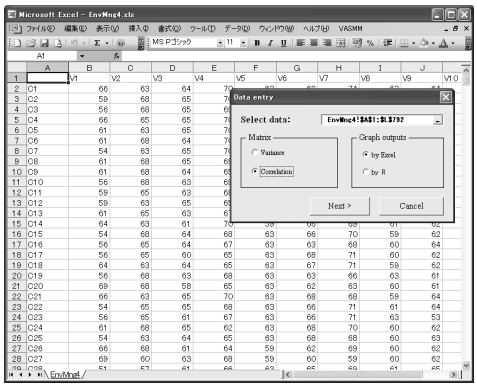


Fig. 3. Selection of data, matrix and output format of graphs.

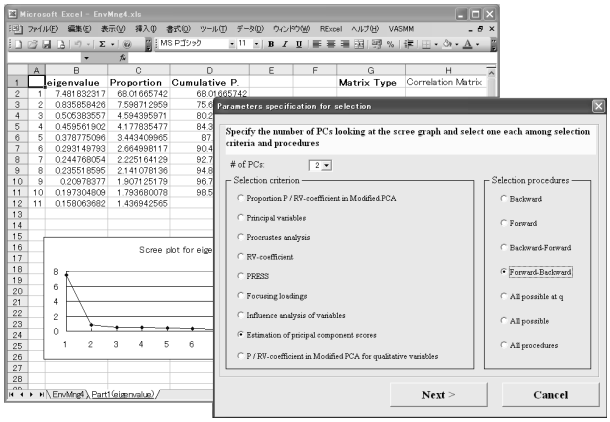


Fig. 4. Results of PCA and selection of # of PCs, criterion and procedure.

proportions and a scree graph of eigenvalues and a further dialogue box appears to ask us to specify the number of PCs and to select a selection criterion and selection procedure (Figure 4).

(iii) Looking at the results we assign the number of PCs; here "2" is selected.

Selection stage:

(iv) Next we select [Estimation of principal components scores] as a criterion and [Forward-Backward] as a selection procedure in the same dialogue box. Since this criterion requires one more specification on computational step, we choose [Step 3] in the next dialogue box (Figure 5).

(v) According to 3 and 4, Excel version of VASMM starts to calculate.

(vi) After a few seconds, the results of the variable selection are displayed on another new sheet; criterion values (RSS), selected variables (deleted variables) and a graph of criterion values (Figure 6).

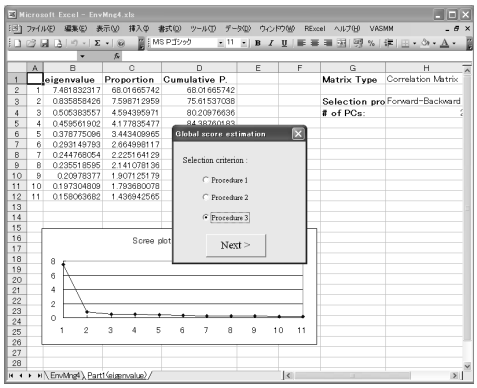


Fig. 5. Sub-selection for variable selection using global scores estimation.

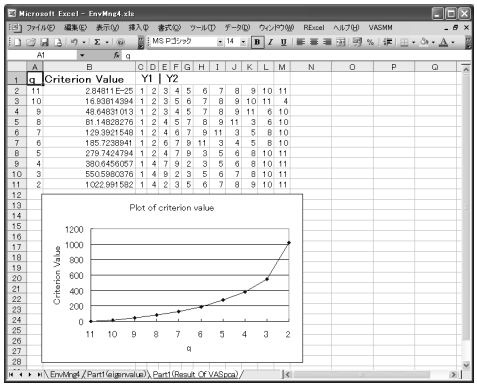


Fig. 6. Results of selection (criterion values, selected/deleted variables and graph of criterion values).

All operations are very easy as the above and in Figures. You can obtain useful information from the output sheet to select a reasonable subset at any q and to decide the number of variables to be used, based on RSS with Step 3 and forward-backward stepwise selection. As for the purpose to select reasonable 7 among 11 variables in this example, the results show that a subset {V1, V2, V4, V6, V7, V9, V11} should be assigned to the 5th survey in the context of the greatest degree to represent information based on all preceding variables by the two PCs based on the selected variables, while the 5th survey consisted of a subset {V2, V4, V6, V7, V9, V10, V11} in practice.

4 Concluding remarks and future works

We have developed and maintain a statistical environment VASMM for variable selection in MMs-ev. We proposed some new criteria and implemented

them in VASMM, and created a new version of VASMM, an interface for Excel, by which users can execute variable selection only using Excel functions without the knowledge of R. This version as well as other two can give useful means and chances for selecting variables in MMs-ev to a variety of users.

We have to propose methods/criteria corresponding to various applications continuously and implement a method/criterion into VASMM once it has been proposed. As for the operability of VASMM, more interactive (or visual) interface for variable selection can be considered.

References

BAIER, T. and NEUWIRTH, E.: R (D)COM Server and RExcel. `http://sunsite.univie.ac.at/rcom/`

IIZUKA, M., MORI, Y., TARUMI, T. and TANAKA, Y. (2002): Statistical software VASMM for variable selection in multivariate methods. In: Härdle, W. and Rönz, B. (eds): *COMPSTAT2002 Proceedings*, Springer, 563–568.

JOLLIFFE, I.T. (1972): Discarding variables in a principal component analysis I – Artificial data –. *Appl. Statist.*, 21, 160–173.

KRZANOWSKI, W.J. (1987): Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist.*, 36, 22–33.

McCABE, G.P. (1984): Principal variables. *Technometrics*, 26, 137–44.

MORI, M., DU, X. and IIZUKA, M. (2004a): Considering variable selection criteria in correspondence analysis. *Bulletin of Faculty of Environmental Science and Technology in Okayama University*, 10(2), 49–56. (in Japanese)

MORI, Y., FUEDA, K. and IIZUKA, M. (2004b): Orthogonal score estimation with variable selection in multivariate methods. In: Antoch, J. (ed.): *COMPSTAT2004 Proceedings*, Springer, 1527–1534.

MORI, Y., IIZUKA, M., TARUMI, T. and TANAKA, Y. (2000): Statistical software "VASPCA" for variable selection in principal component analysis. *The 14th Symposium on Computational Statistics*, *Short Communications*, 73-74.

MORI, Y., IIZUKA, M., TANAKA, Y. and TARUMI, T. (2006): Variable Selection in Principal Component Analysis. In: Härdle, W., Mori, Y. and Vieu, P. (eds): *Statistical Methods for Biostatistics and Related Fields*, Springer, 265-283.

MORI, Y., MATSUMOTO, Y., IIZUKA, M. and TANAKA, Y. (2007): A variable selection in modified principal component analysis for qualitative data. *The 56th Session of the International Statistical Institute. Abstract Book*, 337.

MORI, Y., TARUMI, T. and TANAKA, Y. (1998): Principal Component analysis based on a subset of variables –Numerical investigation on variable selection procedures–. *Bulletin of Computational Statistics of Japan*, 11(1), 1–12.

NIKKEI RESEARCH INC. (1997–2006): *Environmental Management Survey* (1st in 1997 to 9th in 2006), Nihon Keizai Shimbun.

ROBERT, P. AND ESCOUFIER, Y. (1976): A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Appl. Statist.*, 25, 257–265.

TANAKA, Y. (1983): Some criteria for variable selection in factor analysis. *Behaviormetrika*, 13, 31–45.

TANAKA, Y. and 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.

Measuring the Importance of Variables in Kernel PCA

Victor Muñiz, Johan Van Horebeek, and Rogelio Ramos

Centro de Investigación en Matemáticas
Apdo. Postal 402, Guanajuato, Gto. 36000, Mexico,
{*victor_m,horebeek,rramosq*}*@cimat.mx*

Abstract. Recently, Kernel Principal Component Analysis (Kernel PCA) has become a popular technique to extend PCA using implicit transformations. The standard solution of Kernel PCA is given as a function of inner products of the observations and not in terms of the variables (features). An interesting question is how to get insight about the importance and influence of the variables used in the underlying projections of Kernel PCA. To this end, we propose a solution based on a particular parameterization of ANOVA kernels and we formulate it as a restricted optimization problem. The proposed method is illustrated with microarray data and the segmentation of fringe patterns.

Keywords: principal component analysis, kernel PCA, ANOVA kernel

1 Introduction

Without any doubt, Principal Component Analysis (PCA) is one of the most popular dimension reduction techniques based on projecting data in directions of maximal variance. It is well known that these directions are defined by the eigenvectors of the covariance matrix. We refer to Jolliffe (1986) for a general introduction.

Among the many extensions and variants that have been published, Kernel PCA (Scholkopf and Smola (2002)) is probably the most recent one. It is based on the property that the solution of PCA can be completely expressed in terms of inner products of the data. To this end, we define the so called Gram matrix K:

$$K = XX^T, \text{ with } X \text{ the data matrix,}$$

For simplicity, we first suppose that the data are centered.
As a consequence of the Singular Value Decomposition Theorem, if we denote by (e_k, λ_k) a normalized eigenvector-eigenvalue pair of the empirical covariance matrix, one can show that the projection of a data point x in the direction of e_k (score function) is given by:

$$\langle x, e_k \rangle = \sum_i \frac{\alpha_k^i}{\sqrt{\lambda_k}} \langle x, x_i \rangle, \text{ with } (\alpha_k, \lambda_k) \text{ an eigenvector-eigenvalue pair of } K. \tag{1}$$

We observe that the projection depends only on the data through inner products.

Next, suppose we transform x into $\Phi(x)$ and we apply PCA on the transformed data. Equation (1) remains true for $\langle \Phi(x), e_k \rangle$ and - again - depends only on the data through inner products of the transformed data defined by the so called kernel function:

$$k(x, y) := \langle \Phi(x), \Phi(y) \rangle. \tag{2}$$

with the corresponding Gram matrix K where $K_{i,j} = k(x_i, x_j)$. The underlying idea of Kernel PCA is to define explicitly $k(\cdot, \cdot)$ and avoiding the calculation of $\Phi(\cdot)$. In this way we can construct non linear extensions of PCA mapping x into a higher dimensional space but keeping the number of parameters α_k bounded by the number of data points.

Two popular choices of k are the polynomial kernel:

$$k(x, y) = (\langle x, y \rangle + c)^t, \text{ with } t, c > 0 \text{ free parameters}$$

and the radial basis kernel:

$$k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}, \text{ with } \sigma \text{ a free parameter.}$$

In case the (transformed) data are not centered, we use the centered Gram matrix $\tilde{K}$:

$$\tilde{K} = K - \frac{1}{n}JK - K\frac{1}{n}J + \frac{1}{n^2}JKJ = (I - \frac{1}{n}J)K(I - \frac{1}{n}J)$$

where n is the number of observations of dimension d, and J is the $n \times n$ matrix with all entries equal to 1.

In the following we will work with ANOVA kernels, based on particular combinations of the previous kernels. In general, an ANOVA kernel of degree R based on a set of kernel functions $\{k^i(\cdot, \cdot)\}$ is defined as:

$$k_R(x, y) = \sum_{1 \le i_1 \le \cdots \le i_R \le d} \prod_{r=1}^{R} k^{i_r}(x_{i_r}, y_{i_r})$$

where d denotes the dimension of the original space. As the solution of Kernel PCA given by (1) is parametrized in terms of the data points and not in terms of the variables, opposite to ordinary PCA, no inmediate insight is obtained about the loadings of each variable in the corresponding score function. In the next section we propose a solution for the particular case of an ANOVA kernel. In this way we obtain results for kernel PCA similar to the ones obtained in Lee et al. (2006) for support vector machine based on Gunn and Kandola (2002). In section 3 we include four examples: the first is a toy example for illustration purposes, the second is a comparison of the proposed method with PCA by using the Fisher's iris dataset, the third is a segmentation problem of fringe patterns (large number of data in a low dimensional space) and the fourth is an application on microarray data (few number of data in a high dimensional space).

2 Weighted ANOVA kernel PCA

In order to obtain information about which variables (or interactions) are important in the score function, we introduce weights β_i in the ANOVA kernel function. E.g., for $R = 1$:

$$k_1(\beta; x, y) = \beta_1 k^1(x_1, y_1) + \beta_2 k^2(x_2, y_2) + \cdots + \beta_d k^d(x_d, y_d).$$

While searching for a direction of maximal variance we will optimize at the same time over the weights β. As we restrict them to be positive and having norm 1, at the end, each β_i will reflect the importance of a particular interaction in the score function. For the case of $R = 1$ where each term involves one variable, β_i reflects the importance of variable i. To favor sparse solutions we will add a L_1 regularization term to the score function:

$$\operatorname*{Var}_{\{x_j\}} \left(\sum_i \alpha_i \tilde{k}(\beta; x_i, x_j) \right) - \lambda \|\beta\|_{L_1}. \tag{3}$$

For computational reasons instead of optimizing simultaneously over α and β, we optimize alternatingly over one parameter while keeping the other fixed. The (constrained) optimization of (3) over α is a standard kernel PCA problem for a particular kernel determined by β. The optimization of (3) over β can be rewritten as:

$$\max_{\beta} \beta^T S \beta - \lambda \|\beta\|_{L_1} \quad \text{subject to} \quad \beta \geq 0, \quad \|\beta\|_{L_2} = 1 \tag{4}$$

where the matrix S is defined by

$$S_{i,j} = \frac{1}{n} \alpha^T \tilde{K}_i \tilde{K}_j^T \alpha \tag{5}$$

and $\tilde{K}_i$ is the centered Gram matrix based on the kernel k^i from the weighted ANOVA kernel expansion of degree R.
This leads to a constrained quadratic optimization problem and is solved by an Augmented Lagrangian method replacing the original constrained problem by a sequence of unconstrained subproblems and solving them by a modified Newton method (see Nocedal and Wright (1999) for details).
The resulting algorithm is given in Figure 1.

As will be shown in example 3 of the next section, for the case of many observations, it is worthwhile to use in steps 3 and 4 of each iteration a (different) subsample of the dataset to speed up the algorithm.
A further modification consists in decoupling in (3) the estimation of the score function and the calculation of its variance, using for each one a different subsample to decrease the variability and data dependency. Denoting by n_1, n_2 the size of the first and second sample, this will modify (4) into

$$\max_{\beta} \ \beta^T (S - T) \beta - \lambda \|\beta\|_{L_1} \tag{6}$$

1: Given starting β^0.
2: **for** $k = 1, 2, \ldots$ **do**
3: Obtain α^k through Kernel PCA by using $\tilde{k}_R(\beta^{k-1}; \cdot, \cdot)$ with fixed β^{k-1}.
4: Obtain β^k by solving (4) with fixed α^k.
5: **end for**

Fig. 1. Algorithm for weighted ANOVA Kernel PCA.

with

$$S_{i,j} = \frac{1}{n_1}\alpha^T \tilde{K}_i \tilde{K}_j^T \alpha, \quad T_{i,j} = \alpha^T \tilde{K}_i 1_{n_2} 1_{n_2}^T \tilde{K}_j^T \alpha \tag{7}$$

Here, 1_{n_2} is the column vector with all entries equal to $1/n_2$.

3 Experiments

In the following experiments we will always use a weighted ANOVA radial basis kernel of degree $R = 1$. The parameter σ was chosen such that it maximizes the variance of the scores of the first projection.

3.1 Toy dataset

We generated 30 observations (x_1, x_2) from a mixture of two independent bivariate gaussians $\mathcal{N}((3,3),\ 0.25 * I)$ and $\mathcal{N}((4,4),\ 0.25 * I)$, and we added 6 noise variables $(x_3, \cdots, x_8)$ from independent standard normal distributions.

The obtained β_i's using 3 iterations and with λ ranging from 0.1 to 1, are shown in Figure 2. We observe systematically higher values for β_1 and β_2, showing that the two non noise variables are correctly identified.

It is important to observe that the proposed procedure also improves the projections. In Figure 3, the corresponding plot is shown using $\lambda = 0.5$ (left) and compared to the plot of unweighted ANOVA Kernel PCA (center) and ordinary Kernel PCA with radial basis kernel (right). The plotting symbols refer to the true group of each observation. We observe that only in the left figure the original clusters are recovered.

3.2 Iris dataset

In this example we used Fisher's iris dataset to compare the proposed method with PCA. Observe that PCA coincides with ANOVA Kernel PCA if $\sigma \to \infty$ and $\lambda \to 0$ for the particular case of a radial base kernel; for small values of σ, Kernel PCA looks at directions maximizing differences in the density.

The projections of the data on the first two PC are shown in Figure 4 for weighted ANOVA Kernel PCA with $\lambda = 0.1$, $\sigma = 0.5$ (left), for ordinary Kernel PCA with $\sigma = 0.5$ (center) and PCA (right). One observes a superior separation of the clusters in the left plot. The corresponding β weights equal

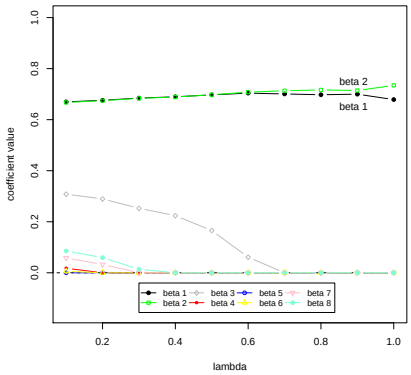


Fig. 2. Variable weighting for the toy dataset based on 3 iterations of the algorithm in Figure 1.

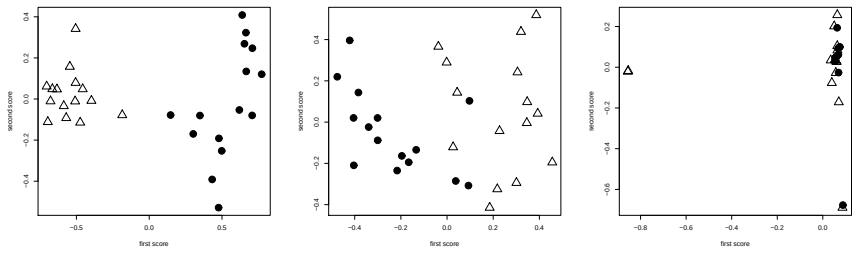


Fig. 3. Projections of the data on the first and second component based on the proposed method (left), using unweighted ANOVA Kernel PCA (center) and using ordinary Kernel PCA with a radial basis kernel (right) for the toy dataset.

$(0.416, 0.092, 0.596, 0.680)$ and show that the third and fourth variables are the most influential ones for the first PC; in PCA these variables get also the largest loadings.

3.3 Segmentation of fringe patterns

In this example we use mixtures of synthetic fringe patterns as shown in Figure 5: a fringe pattern is superposed on top of a background fringe pattern and contaminated by gaussian noise. The goal is to identify the location of the foreground fringe pattern; i.e., assigning each pixel to which pattern it belongs. We used images of 128×128 pixels.

We applied sixteen filters, each one tuned at a different frequency, those filters form together a complete filter bank as was used in Guerrero et al. (2005). We use the magnitude of the response to those filters at each pixel

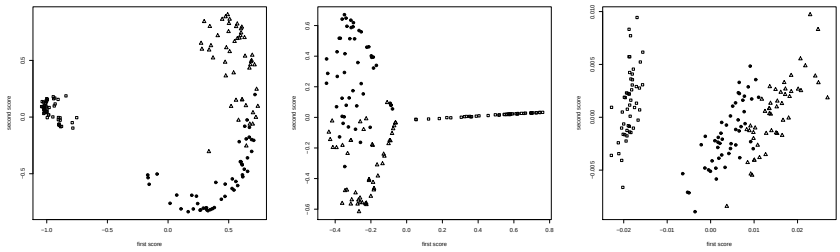


Fig. 4. Projections on the first and second component using the proposed method (left), using ordinary Kernel PCA with radial basis kernel (center) and using the proposed method with $\sigma = 100$ and $\lambda = 0$ for the iris dataset.

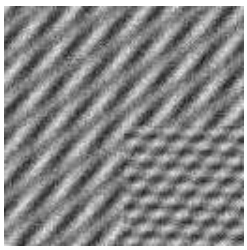


Fig. 5. Mixture of synthetic fringe patterns with additive gaussian noise.

to define the variables $(x_1, \cdots, x_{16})$. The results of applying the proposed method using subsampling based on 90 groups of size 182 pixels are shown in Figure 6. On the left, we observe that two values β_2 and β_5 are clearly dominating. The orientation of the filters with the two highest β_i's effectively corresponds to orientations present in the forward pattern and not in the background pattern. The fact that the important variables are correctly identified is confirmed by the biplot at the right side of Figure 6. Again, we used two different plotting symbols to indicate the true group of each pixel. One observes the agreement between the two observed groups of points and the true groups, suggesting a good segmentation.

3.4 Microarray data

In this experiment, we used microarray data used from Khan et al. (2001) based on the expression level of 2308 genes from different tumors. In order to obtain an unsupervised problem, we limited the data to the 24 samples belonging to the Ewing family of tumors (EWS). Using the above method with only one iteration and taking $\lambda = 5$, the values of the β coefficients are shown in Figure 7. To validate these results, we ordered the variables according to their importance estimated by the β_i coefficients. We perturbed the observations of the 500 most important variables with uniform noise and,

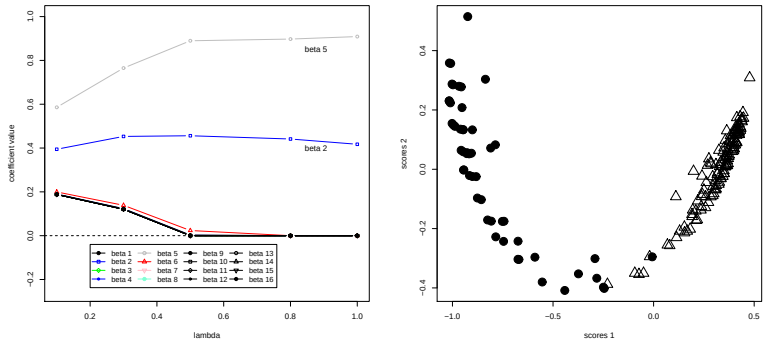


Fig. 6. Left: Variable weighting using subsampling for the fringe pattern image in Figure 5. We used the algorithm shown in Figure 1 with subsampling based on 90 random samples of size 182 pixels. Right: Projections in the first and second component obtained for the same data.

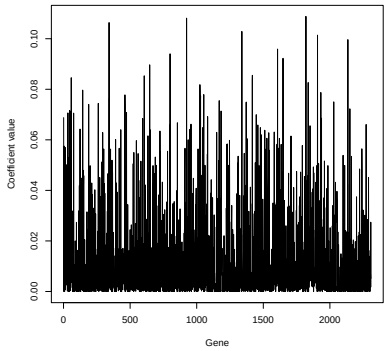


Fig. 7. Variable weighting for the EWS group in the microarray data.

as a reference, we also perturbed the observations of the 500 least important variables. For both datasets we applied weighted ANOVA Kernel PCA and plotted for each variable its corresponding weight using the original versus the weight obtained by perturbing the corresponding observations. The results are shown in Figure 8. As it should be, we observe that perturbing the most important variables is much more destructive than perturbing the least important ones.

4 Conclusions

In this paper we presented a way to obtain insight about variable importance in Kernel PCA. This leads to a restricted optimization problem that we

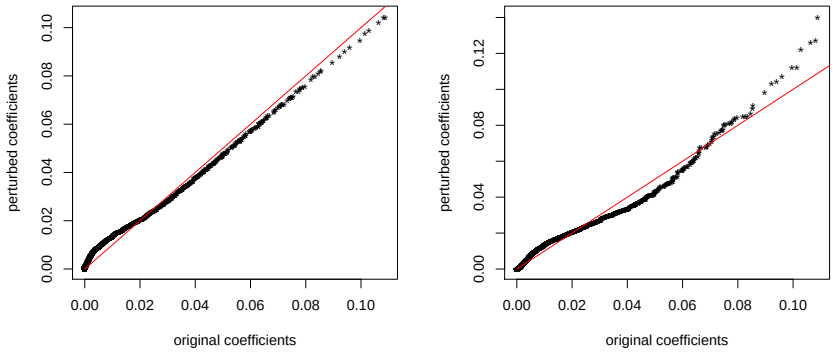


Fig. 8. Plot of the original β coefficients versus the coefficients obtained by perturbing the last 500 variables (left) and the first 500 variables (right) according to their importance given by the β values.

solve with an augmented Lagrangian method. The experiments confirm the usefulness of the proposed algorithm.

References

BRADLEY, P.S. and MANGASARIAN, O.L. (1998): Feature selection via concave minimization and support vector machines. In: *Proceedings of the 15th Conference on Machine Learning*, 82–90.

GUERRERO, J.A., MARROQUIN, J.L., RIVERA, M. and QUIROGA, J.A. (2005): Adaptive monogenic filtering and normalization of ESPI fringe patterns. *Opt. Lett. 30, 3018-3020.*

GUNN, S.R. and KANDOLA, J.S. (2002): Structural modelling with sparse kernels. *Machine Learning 48, 137-163.*

JOLLIFFE, I.T. (1986): *Principal Component Analysis.* Springer Series in Statistics, New York.

KHAN, J., WEI, J., RINGER, M., SAAL, L., LADANYI, M., WESTERMANN, F., BERTHHOLD, F., SCHWAB, M., ATONESCU, C., PETERSON, C. and MELTZER, P. (2001): Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine 7, 673-679.*

LEE, Y., KIM, Y., LEE, S. and KOO, J. (2006): Structured multicategory support vector machines with analysis of variance decomposition. *Biometrika 93, 555-571.*

NOCEDAL, J. and WRIGHT, S.J. (1999): *Numerical Optimization.* Springer Series in Operations Research.

SCHOLKOPF, B. and SMOLA, A.J. (2002): *Learning With Kernels. Support Vector Machines, Regularization, Optimization And Beyond.* MIT press.

A New Approach to Data Fusion Through Constrained Principal Component Analysis

Alfonso Piscitelli

Department of Sociology
Federico II University of Naples, Italy,
alfonso.piscitelli@unina.it

Abstract. Data Fusion consists of merging information coming from two different surveys. The first is called "reference" or "donor survey" while the second is called "punctual" or "receptor survey". The aim is to complete the receptor matrix exploiting information acquired from the donor matrix. The two independent surveys have a block of common variables used as a bridge between them. In this work a Data Fusion methodology based on the *Constrained Principal Component Analysis* (*CPCA*) technique is presented. The proposed method allows to impute the missing information into the second survey taking into account knowledge about non-symmetric relationship structure among variables.

Keywords: file grafting, missing values imputation, non symmetrical exploratory data analysis

1 Introduction

Data Fusion aims at matching two already held surveys in order to make it possible to transferring part of information contained in the first survey to the second. The first survey is called reference survey (donor matrix); the second is called punctual survey (receptor matrix). The need for socio-economical Data Fusion arises in market studies (Baker et al., 1989) especially in media and consumption surveys.

Data Fusion allows us to treat data coming from the two distinct surveys as whole. The aim is to determine the unobserved values of q variables $\mathbf{Y}$ included in a first survey, but not in a second. This is usually treated as a missing value imputation problem. Missing data of the receptor matrix will be imputed by exploiting information coming from the donor matrix. A necessary condition to perform such a imputation is the presence a set of p variables $\mathbf{X}$ in common to both surveys.

Different methodologies have been proposed in literature (Little and Rubin, 1987). *Explicit model-based estimation* consists in finding a *model* relating the variables $\mathbf{Y}$ with the variables $\mathbf{X}$ in the donor survey and applying this model to the *receptor* survey. *Implicit models for imputation* consist in finding for each individual of the receptor survey one or more donor individuals that are as closest as possible. According to some statistical technique

the values of the variables $\mathbf{Y}$ will be transferred to the *receptor* individual. This latter is known as the "donor" principle.

A commonly used implicit method is File Grafting (Aluja-Banet et al., 1995) which is based on Principal Component Analysis (*PCA*). In order to find a common subspace onto which to project the statistical units coming from the two surveys, a *PCA* is performed of the common variables of the reference survey. As it is well none the *PCA* analyzes the correlation structure, and in this sense the variables play the symmetric role, assuming an interdependence structure among them (D'Ambra and Lauro, 1982). However, in sociological and economic theories some relationships are given and well-known, and hence some *a* priori knowledge on dependency structure among the $\mathbf{X}$ and $\mathbf{Y}$ variables is available. In such a case non symmetrical data analysis approach could be more profitably used.

In this paper we propose a *Non Symmetrical Grafting* (*NSG*) technique that exploit the non symmetrical *PCA* to explore the dependence structure of data. We use the *Constrained Principal Component Analysis* (*CPCA*) technique (D'Ambra and Lauro, 1982). The *NSG* algorithm projects individuals belonging to different surveys onto the same subspace, determined through the non symmetrical *PCA*. In such a space, distances among individuals belonging to the different surveys are evaluated. Finally, for each subject of the receiver survey, the "missed" values are taken by the nearest neighbour donors. The proposed procedure has been applied to a simulated data sets in order to show how it works and to compare it with the standard file grafting.

The paper is organized as follow. In the second section we briefly present main ideas about the file grafting; in section 3 we present the proposed Non Symmetrical Grafting and introduce the *Constrained Principal Component Analysis* in Data Fusion framework; section 4 offers details on imputation methods and its validation; finally in section 5 we present the main results of the simulation study.

2 File grafting for data fusion

The file grafting uses descriptive factorial analysis techniques with the aim of connecting information coming from two distinct surveys: the first one containing information about a set of $p+k=q$ variables observed on n_0 subjects; the second one, containing information about a set of $p+j=z$ variables observed on n_1 subjects. In both surveys a set of p variables $\mathbf{X}$ is in common. Let us to denote with $\mathbf{X}_0$ the set of common variables referred to the donor survey, and with $\mathbf{X}_1$ the other one referred to the receiver survey. Analogously we denote with $\mathbf{Z}_1$ the j specific variables of the receptor survey, with $\mathbf{Y}_0$ the k specific variables of the donor survey and with $\mathbf{Y}_1$ the specific variables to be imputed. The aim is to fill the hyphened part of the second data matrix. We use the donor survey ($\mathbf{X}_0$;$\mathbf{Y}_0$) to impute a set of unknown variables $(p + k)$ of the receiver survey (Fig.1).

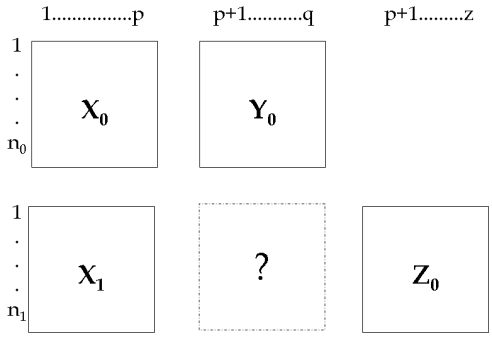


Fig. 1. Blocks of shared and unshared information.

To perform file grafting the assumption of stability of the relationships among variables is required (Bonnefous et al., 1986). This latter allows to define a common space on which to represent the whole information of both data sets.

Specifically, the file grafting technique essentially consists of two phases (Rius et al., 1996): pre-grafting and grafting.

The former consists of studying the common variables and testing the common space stability in order to ensure the grafting feasibility. In this phase the problem is to identify a subset of common variables defining a similar subspace of representation for both data sets, and ensuring the stability. Such variables represent the "bridge" to transfer information from one data set to the other (namely, projecting on it). In second phase the actual graft is performed by projecting the whole information as additional elements. First we perform a singular value decomposition of $\mathbf{X}_0$, $\mathbf{X}_0 = V_0 \Lambda_0 U_0'$, and then we represent the row elements in the U_0 basis with coordinates $\Psi_0 = \mathbf{X}_0 U_0$. Hence, the *grafting* consist on positioning the elements of the second data set, $\mathbf{X}_1$, upon the same reference basis U_0. Naturally, the individuals of $\mathbf{X}_1$ are projected as supplementary points with coordinates $\Psi_1 = \mathbf{X}_1 U_0$.

2.1 Grafting for imputation

Once all the individuals of the two surveys have been projected on the previously defined subspace, for each individual of receiver matrix $\mathbf{X}_1$ we look for a donor(s) having the close profile with the common variables (*nearest neighbours*). The *nearest neighbours* of the *i-th* unit of the receptor survey are individuals of the donor survey having the minimum distance in the common space. In the Data Fusion original proposal *nearest neighbour* (*nn*) algorithm have been applied (Baker et al., 1989); a modified version (Aluja-Banet et al.,

2001) exploits and applies the *K-nearest neighbours* (*knn*) algorithm (Fukunaga and Narendra, 1975). Finally, missing data are imputed by hot deck imputation (Ford, 1980).

After the imputation it is necessary to measure the precision of the performed Data Fusion. One way consists of carrying out a self-imputation of $\mathbf{Y}_0$ variables upon the same individuals $\mathbf{X}_0$. In such a case to compare the observed values with the imputed ones by the index R_y (Aluja-Banet et al., 2001):

$$R_y = \frac{tr\left[(\mathbf{Y}_0 - \tilde{\mathbf{Y}}_0)'(\mathbf{Y}_0 - \tilde{\mathbf{Y}}_0)\right]}{tr\left[(\mathbf{Y}_0 - \bar{\mathbf{Y}}_0)'(\mathbf{Y}_0 - \bar{\mathbf{Y}}_0)\right]} \tag{1}$$

R_y can be interpreted as the proportion of error we are producing compared with the error we would produce when imputing with the simple mean of the variable. Such an index could be exploited to determine the value of k, when the *knn* algorithm is used in the fusion process. Evaluating R_y for the increasing k and plotting R_y as a function of k, we look for the k corresponding to the minimum value of R_y.

3 Non symmetrical grafting for data fusion

Descriptive factorial analysis used for file grafting (e.g. PCA, MCA,) do not imply any *a priori* knowledge about the phenomenon under study. For sample survey data, *a priori* information about different roles of the variables of the same survey are available. Often in the same survey there is a dependence structure between two sets of variables (e.g. income and number of the family member affect the consumptions and savings). If a set of variables (dependent variables) depend on another set of variables (independent variables) we can use this information to improve Data Fusion process. To build a common space on which projecting information from the two surveys, we propose the use of the *Constrained Principal Component Analysis* (*CPCA*) technique. *CPCA* consists of carrying out a PCA of the Y's image projected onto the common variables subspace through a suitable orthogonal projection operator. Let $\mathbf{X}$ and $\mathbf{Y}$ be two blocks of centered and scaled variables observed on the same n units which identify two sub-sets. the goal of CPCA analysis is to analyze the relationship structure of $\mathbf{Y}$ block in respect to $\mathbf{X}$ block in terms of principal components which are associated with the latter block.

Let $\Re^q$ be the $p + k$ dimensional vectorial space, and let $\Re^p$ be the vectorial sub-space of $\Re^q$ generated by the columns of $\mathbf{X}$, and consider the image of $\mathbf{Y}$ in the sub-space $\Re^p$:

$$\mathbf{Y}^* = P_X \mathbf{Y}, \tag{2}$$

i.e $\mathbf{Y}^*$ is the projection of $\mathbf{Y}$ on $\Re^p$ through the orthogonal projection operator $P_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$. The *CPCA* analysis consists on a singular value decomposition of $\mathbf{Y}^*$,

$$\mathbf{Y}^* = V^* \Lambda^* U'^*. \tag{3}$$

In such case we represent the row elements in the U^* basis, with coordinate

$$\Psi^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}\mathbf{U}^* = \mathbf{Y}^* U^* \tag{4}$$

3.1 Building the basic matrix

To use a priori information when doing *grafting*, we need to define the optimal number of common variables influencing the specific variables to be imputed. In other words, variables on which we will perform the *CPCA* have to be selected. An appropriate criterion for variables selection is the *backward elimination*. In other words, we fit a regression model for each variable belonging to $\mathbf{Y}$ block on $\mathbf{X}$, and we select the predictors through backward elimination. Then, in performing *CPCA* we use the set of predictors that have been selected by the backward elimination procedure, and that are in common to the two surveys.

3.2 Graft in CPCA

Once the common space is built, the process of grafting consists of projecting the whole information in the common space in order to represent the two data clouds. Starting from the principal components of the *CPCA*, it is possible to perform the projection of additional individual which is described by the matrix $[\mathbf{Y}_s|\mathbf{X}_s]$ with coordinates

$$\Psi_s^* = \mathbf{X_s}(\mathbf{X_s}'\mathbf{X_s})^{-\mathbf{1}}\mathbf{X_s}'\mathbf{Y_s}\mathbf{U}^* = \mathbf{Y_s}^* U^* \tag{5}$$

In such a case, the individuals to be supplementary projected are lines of the receptor matrix, and hence the $\mathbf{Y}_s$ values we need in equ. (5) are missed. To overcome this problem we propose to estimate them trough a regression model for each variable, starting from the reference survey's data. The usual OLS estimate $\hat{\beta}_0 = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Y}_0$ is used to perform a first imputation of the specific variables $\hat{\mathbf{Y}}_1$ in the punctual survey, with

$$\hat{\mathbf{Y}}_1 = \mathbf{X}_1\hat{\beta}_0. \tag{6}$$

For every individual from the matrix of the receivers we replace $\mathbf{Y}_s$ at equ. (5) with the values $\hat{\mathbf{Y}}_1$ obtained from the application of the regression model. Finally, in the case of Non Symmetrical Grafting, the coordinates of the supplementary points of the receptor matrix are as:

$$\Psi_1^* = \mathbf{X_1}(\mathbf{X_1}'\mathbf{X_1})^{-\mathbf{1}}\mathbf{X_1}'\hat{\mathbf{Y}}_\mathbf{1}\mathbf{U_0^*} = \mathbf{P_{X_1}}\hat{\mathbf{Y}}_\mathbf{1}\mathbf{U_0^*} \tag{7}$$

where U_0^* are the base of the *CPCA* for $\mathbf{Y_0}^* = P_{X_0}\mathbf{Y_0}$. With the equ. (7) we solve the problem of the projection in supplementary of the receptor matrix individuals.

4 Imputation and validation

Once all the individuals of the two surveys are projected in the subspace with *CPCA* technique, we calculate, for every individual of the receptor matrix, the distance from the donor matrix. We use the Euclidean metric and for each subject of the receptor matrix, we define the *nearest neighbour*, or the group of *k-nearest neighbour*. The imputation is deterministic, using the *hot deck imputation*.

If we use the *nearest neighbour* algorithm, ($k = 1$), we copy the specific values of the response variables given by the best donor and imputing (pasting) them to the considered receiver (T1DM imputation).

If we use the *k nearest neighbour* algorithm, ($k > 1$), we calculate the average on the specific variables values given by the optimal *knn* donors and imputing it to the considered receptor (TKDU imputation). This process is performed for every single variable.
To evaluate the optimal number of the nearest neighbour, we proceed to an auto-imputation of the variables $\mathbf{Y}_0$ on the same $\mathbf{X}_0$, individuals, in order to be able to measure the produced error. (Aluja-Banet et al., 2001). Once the imputation is completed, it is necessary to validate the imputed data.

We have three validation levels of the quality of the imputed data. The first level consist of a global statistics comparison. We perform a two Sample Means Test between the block of the imputed variables $\tilde{\mathbf{Y}}_1$ and the block of the donor matrix specific variables $\mathbf{Y}_0$. The second level tends to assess the homogeneity of imputations evaluating the internal coherency and the external coherency of the imputed variables. The former is based on comparison of the correlation matrix of $\tilde{\mathbf{Y}}_1$ and the corresponding correlation matrix of $\mathbf{Y}_0$. The latter tends to verify the similarity of the two cross correlation matrix of $\mathbf{X}_0$ with $\mathbf{Y}_0$ and of $\mathbf{X}_1$ with $\tilde{\mathbf{Y}}_1$.

In order to evaluate both internal and external coherency we use the Fisher transformation z_r ,

$$z_r = 0,5 \ln\left(\frac{1+r}{1-r}\right), \tag{8}$$

and we perform a set of significance tests based on the Z distribution to verify the pairwise correlation coefficient's homogeneity.The imputed variables are coherent when the difference among the couples of the transformed correlation coefficients is not significant for a given p-value.

Finally, the last level of the validation process is given by the accuracy of the imputation, whereas with accuracy we mean the degree of correspondence between the imputed variables values and the "real values". Accuracy can

be measured by the calculation of the root mean square error between the imputed variables values $\tilde{\mathbf{Y}}_1$ and the "real values" $\mathbf{Y}_1$, that we should have had if we observed those variables in the punctual survey.

$$RMSE = \sqrt{n_1^{-1} tr\left[(\tilde{\mathbf{Y}}_1 - \mathbf{Y}_1)'(\tilde{\mathbf{Y}}_1 - \mathbf{Y}_1)\right]}. \tag{9}$$

This level of validation is possible only in simulation study where the real values are known.

5 Simulation study

In this section we present a little simulation study to compare the *NSG* algorithm with respect to the classical file grafting methodology based on the *PCA*. In the simulation we follow the *missing data at random* (MAR) approach. The simulated data set consists of 12 variables (7 common variables and 5 specific variables) with sample size $n = 350$. In Table 1 we report the scheme of the data generating process: the simulated variables mimic socio-economical variables like, age, income, consumptions and savings.

Table 1. Simulation study: scheme of the data generating process.

Common variables	*Specific variables*
$X_1 \sim U(18, 65)$	$Y_1 = \alpha + 3,5X_1 + e$
$X_2 \sim U(800, 3100)$	$Y_2 = \alpha - 2,3X_1 + 1,6X_2 + e$
$X_3 \sim N(120, 70)$	$Y_3 = \alpha + 0,4X_7 - 0,8X_1 + e$
$X_4 \sim N(100, 20)$	$Y_4 \sim N(66, 17)$
$X_5 \sim N(65, 15)$	$Y_5 = \alpha + 1,9X_5 - 0,4X_6 + e$
$X_6 \sim N(250, 10)$	
$X_7 \sim U(10, 1000)$	

In table (1), α is a constant and $e \sim N(0, 1)$ is a usual white noise. The data set has been randomly divided in two subset with sizes $n_0 = 200$ and $n_1 = 150$. In the second subset the group of specific variables have been "deleted" . The "deleted" block of matrix is called "control block". The imputation method we use the *TKDU* both for the algorithm *NSG* and for the "classic" procedure. Therefore the imputed values come from the group of *nearest neighbour* donors (*knn*). The validation of the imputation is performed for all the three levels previously described. For the comparison of the global statistics, the averages of the donor specific variables and the averages of the imputed variables do not meaningfully differ. This is true for both imputation procedure. Concerning the internal coherence, in the case of the *NSG* algorithm the correlation coefficient tests are not significant (except for the couples of variables Y_1-Y_4 and Y_3-Y_5). On the contrary, the classical file

grafting procedure 7 of the 10 performed tests are significant. Also in the case of the external coherence, the *NSG* algorithm presents better results than the imputation performed through the *PCA*. The data suggest that *NSG* algorithm better reconstructs the correlation structure the whole correlation structure. Finally, being available the "control block", we calculate the *RMSE* for the specific variables of the second survey (Table 2). The *NSG* algorithm yields lower *RMSE* values with respect to the *PCA* algorithm for all specific variables.

Table 2. *RMSE* values for the *NSG* and *PCA* algorithms.

RMSE					
	Y_1	Y_2	Y_3	Y_4	Y_5
NSG	39,892	13531	44,729	10,635	55,689
PCA	44,148	14062	104,39	13,782	57,064

References

ALUJA -BANET T., NONELL R., RIUS R., MARTÍNEZ M., (1995): File grafting. *In F. Mola and A. Morineau (eds) Actes du IIIme Congrés International d'Analyses Multidimensionnelles des Données - NGUS'95, Centre Int. de Statistique et d'Informatique Appliquées, CISIA-CERESTA, 23-32.*

ALUJA -BANET T., MORINEAU A., RIUS R., (1999): La greffe de fichiers et ses conditions d'application. Méthode et exemple. *In G. Brossier and A.M. Dussaix (eds) Enquêtes et sondages, Dunod, 94-102.*

ALUJA-BANET T., THIO S., (2001): Survey data fusion, *Bulletin of Sociological Methodology, 72, 20-36.*

BAKER K., HARRIS P., O'BRIEN J., (1989): Data fusion: an appraisal and experimental evaluation, *Journal of the Market Research Society, 31(2), 153-212.*

BONNEFOUS S., BRENOT J., PAGES J.P., (1986): Methode de la greffe et communication entre enquetes. *In: E. Diday et al., Data Analysis and Informatics IV, North Holland, 603-617.*

D'AMBRA L., LAURO N.C., (1982): Analisi in componenti principali in rapporto a un sottospazio di riferimento. *Rivista di Statistica applicata, 15, (1), 51-67.*

FORD B., (1980): An Overview of Hot-Deck Procedures. *In: Madow, W. et. al. (eds), Incomplete data in sample surveys, Vol 2., Academic Press, NY., 185-206.*

FUKUNAGA K., NARENDRA P.M., (1975): A branch and bound algorithm for computing k-nearest neighbours. *IEEE Trans. Computers, C-24, (7), 750-753.*

LITTLE J.A., RUBIN D.B., (1987): *Statistical analysis with missing data*. Second edition, Wiley & Sons.

RIUS R., NONELL R., ALUJA -BANET T., (1996): File grafting: a data sets communication tool, *COMPSTAT '96, Physica Verlag, 417-422.*

Part XIV

Non-Parametric Statistics and Smoothing

Prewhitening-Based Estimation in Partial Linear Regression Models

Germn Aneiros-Pérez[1] and Juan Manuel Vilar-Fernández[2]

[1] Facultad de Informática, Universidade da Coruña
Campus de Elviña s/n, 15071, A Coruña, España, *ganeiros@udc.es*
[2] Facultad de Informática, Universidade da Coruña
Campus de Elviña s/n, 15071, A Coruña, España, *eijvilar@udc.es*

Abstract. A regression model whose regression function is the sum of a linear and a nonparametric component is presented. The design is random and the response and explanatory variables satisfy mixing conditions. A new local polynomial type estimator for the nonparametric component of the model is proposed and its asymptotic normality is obtained. Specifically, this estimator works on a prewhitening transformation of the dependent variable, and the results show that it is asymptotically more efficient than the conventional estimator (which works on the original dependent variable) when the errors of the model are autocorrelated.

Keywords: mixing processes, nonparametric regression

1 Introduction

A special class of semiparametric regression models which are flexible, overcome (or reduce) the "curse of dimensionality" and allow easy interpretation of the effect of each explanatory variable on the response variable was proposed by Engle et al. (1986). This class of models, known as Partial Linear Regression (PLR) models, assumes that the regression function is the sum of a linear and a nonparametric component, that is,

$$Y_i = \mathbf{X}_i^T \beta + m(\mathbf{T}_i) + \varepsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id_0})^T$ and $\mathbf{T}_i = (T_{i1}, \ldots, T_{id_1})^T$ ($d_0 \geq 1$ and $d_1 \geq 1$) are vectors of explanatory variables, $\beta = (\beta_1, \ldots, \beta_{d_0})^T$ is a vector of unknown real parameters, m is an unknown smooth real function and $\{\varepsilon_i\}$ are the random errors satisfying

$$E(\varepsilon_i \mid \mathbf{X}_i, \mathbf{T}_i) = 0 \quad (i = 1, \ldots, n). \tag{2}$$

The PLR model has been studied extensively for i.i.d. data (see, for example, Speckman (1988) and Robinson (1988)) as well as for dependent data (see, for example, Gao (1995) and Aneiros-Pérez et al. (2004)). In general, these works propose different estimators for β and/or m in (1) and study

their consistency and asymptotic normality. In addition, PLR models have demonstrated their usefulness in many fields of applied sciences, such as economics, environmental studies, medicine, ... (see Härdle et al. (2000), for a monograph and applications of the PLR model).

This paper deals with the estimation of the nonparametric component m of the PLR model (1), assuming both random design on $\{(\mathbf{X}_i, \mathbf{T}_i)\}$ and mixing conditions on $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}$. For pure nonparametric regression models with random design (that is, model (1) where β is known and $\mathbf{T}_i$ are random vectors), Masry (1996) and Xiao et al. (2003), among others, reported their findings. Masry obtained the asymptotic normality of the (conventional) local polynomial estimator, while Xiao et al. obtained that of a local polynomial estimator applied to a prewhitening transformation of the dependent variable, this transformation being estimated from the data. In conclusion, the estimator proposed by Xiao et al. takes into account the correlation structure of the error process and is asymptotically more efficient than that studied by Masry. Given this result, we will estimate m in the model (1) following the procedure proposed by Xiao et al., but applied to PLR models instead of to pure nonparametric models. As we will show, the asymptotic result proven by Xiao et al. holds for our estimator.

The construction of the estimator is presented in Section 2, while the asymptotic results and the conditions used to obtain these are given in Section 3.

2 The estimator

Let $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)^T$, $\mathbf{Y} = (Y_1, ..., Y_n)^T$ and $\widetilde{\mathbf{A}}_h = (\mathbf{I} - \mathbf{W}_h)\mathbf{A}$, where $\mathbf{W}_h = (w_{h,j}(\mathbf{T}_i, \mathbf{T}_1, ..., \mathbf{T}_n))_{i,j}$ is an $n \times n$ matrix with $w_{h,j}(\cdot, \mathbf{T}_1, ..., \mathbf{T}_n)$ being a weight function, and $\mathbf{A}$ and $h > 0$ are any $(n \times s)$-matrix ($s \geq 1$) and real number, respectively. Throughout this paper, we will denote $w_{h,j}(\cdot) \equiv w_{h,j}(\cdot, \mathbf{T}_1, ..., \mathbf{T}_n)$.

From (1) and (2), we have that

$$Y_i - E(Y_i \mid \mathbf{T}_i) = (\mathbf{X}_i - E(\mathbf{X}_i \mid \mathbf{T}_i))^T \beta + \varepsilon_i \quad (i = 1, \ldots, n). \tag{3}$$

Assuming that the conditional moments included in (3) are smooth functions of $\mathbf{T}_i$, these can be estimated using nonparametric estimators. Hence, an estimator of β can be obtained applying Ordinary Least Squares (OLS) estimation to the model (3) after estimating those moments. This estimation gives

$$\widehat{\beta}_{h_0} = \left(\widetilde{\mathbf{X}}_{h_0}^T \widetilde{\mathbf{X}}_{h_0}\right)^{-1} \widetilde{\mathbf{X}}_{h_0}^T \widetilde{\mathbf{Y}}_{h_0}. \tag{4}$$

Finally, a nonparametric estimator for m in the PLR model (1) can be obtained by smoothing the points $\left\{\left(Y_j - \mathbf{X}_j^T \widehat{\beta}_{h_0}, \mathbf{T}_j\right)\right\} \subset \mathbb{R} \times \mathbb{R}^{d_1}$, this esti-

mator taking the expression

$$\widehat{m}_{h_0,h_1}(\mathbf{t}) = \sum_{j=1}^{n} w_{h_1,j}(\mathbf{t}) \left(Y_j - \mathbf{X}_j^T \widehat{\beta}_{h_0} \right). \tag{5}$$

(In (4) and (5), $h_0 > 0$ and $h_1 > 0$ are smoothing parameters or bandwidths that typically appear in nonparametric or semiparametric estimations.) The estimators (4) and (5) were proposed by Speckman (1988) and Robinson (1988) for i.i.d. data. Under suitable conditions, their asymptotic results (normality and rates of convergence) hold if the weights take the form

$$w_{h,j}(\mathbf{t}) = \frac{K\left(\dfrac{\mathbf{t} - \mathbf{T}_j}{h}\right)}{\sum_{i=1}^{n} K\left(\dfrac{\mathbf{t} - \mathbf{T}_i}{h}\right)},$$

where $\mathbf{t} = (t_1, ..., t_{d_1})^T \in \mathbb{R}^{d_1}$ and $K(\mathbf{t}) = \prod_{j=1}^{d_1} k(t_j)$, with $k(\cdot)$ an univariate kernel function.

In this paper we assume that the error process $\{\varepsilon_i\}$ in (1) is autocorrelated. Then, we will propose and study an estimator like (5), but using prewhitening observations instead of original observations Y_j (so, the proposed estimator considers the autocorrelation structure in $\{\varepsilon_i\}$). Furthermore, we will use local p-order polynomial type weights (see Stone (1977)).

The purpose of the prewhitening transformation is to obtain a regression model with uncorrelated errors. Let us assume that the errors $\{\varepsilon_i\}$ in (1) follow the invertible linear process

$$\varepsilon_i = \sum_{j=0}^{\infty} c_j e_{i-j}, \text{ where } c_0 = 1 \text{ and } e_i \text{ are i.i.d. r.v. with } E(e_i) = 0. \tag{6}$$

Let $c(L) = \sum_{j=0}^{\infty} c_j L^j$, where L is the lag operator, and

$$a(L) = c(L)^{-1} = a_0 - \sum_{j=1}^{\infty} a_j L^j \text{ with } a_0 = 1. \tag{7}$$

Applying $a(L)$ to the original PLR model (1) and rewriting the corresponding equation, we obtain the new PLR model

$$\underline{Y}_i = \mathbf{X}_i^T \beta + m(\mathbf{T}_i) + e_i \quad (i = 1, \ldots, n), \tag{8}$$

where $\underline{Y}_i = Y_i - \sum_{j=1}^{\infty} a_j \left(Y_{i-j} - \mathbf{X}_{i-j}^T \beta - m(\mathbf{T}_{i-j}) \right) = Y_i - \sum_{j=1}^{\infty} a_j \varepsilon_{i-j}$. The regression function in the new PLR model (8) is the same as that in the original PLR model (1), but in (8) the errors are i.i.d. Now, we propose to construct an estimator for m based on this new PLR model. Because in

practice the response variable $\underline{Y}_i$ in (8) is unknown, the first step in constructing such an estimator should be to propose a "reasonable" approximation for $\underline{Y}_i$. Following our findings and the work by Xiao et al. (2003) based on pure nonparametric regression, we propose to use the residuals $\widehat{\varepsilon}_i = Y_i - \mathbf{X}_i^T\widehat{\beta}_{h_0} - \widehat{m}_{h_0,h_0}(\mathbf{T}_i)$ of the original PLR model (1) to construct an estimate of $A_{\mathcal{T}} = (a_1, ..., a_{\mathcal{T}})^T$, $\mathcal{T}$ being a truncation parameter large enough to avoid problems with the bias. Specifically, this estimator for $A_{\mathcal{T}}$ is constructed using the OLS method applied to the model

$$\widehat{\varepsilon}_i = a_1\widehat{\varepsilon}_{i-1} + \cdots + a_{\mathcal{T}}\widehat{\varepsilon}_{i-\mathcal{T}} + residual_i \quad (i = \mathcal{T}+1, \ldots, n).$$

Hence, the estimator

$$\widehat{A}_{\mathcal{T}} = \left(\widehat{\varepsilon}_{\mathcal{T}}^T\widehat{\varepsilon}_{\mathcal{T}}\right)^{-1}\widehat{\varepsilon}_{\mathcal{T}}^T\widehat{\varepsilon}$$

is obtained, where $\widehat{\varepsilon} = (\widehat{\varepsilon}_{\mathcal{T}+1}, \ldots, \widehat{\varepsilon}_n)^T$ and $\widehat{\varepsilon}_{\mathcal{T}} = (\widehat{\varepsilon}_{i,j})_{\substack{1\le i\le n-\mathcal{T}\\ 1\le j\le \mathcal{T}}}$ with $\widehat{\varepsilon}_{i,j} = \widehat{\varepsilon}_{i-j+\mathcal{T}}$. Now, using $\widehat{A}_{\mathcal{T}}$ together with $\widehat{\beta}_{h_0}$ and $\widehat{m}_{h_0,h_0}(\cdot)$, we define

$$\underline{\widehat{Y}}_{\mathcal{T},i} = Y_i - \sum_{j=1}^{\mathcal{T}}\widehat{a}_j\left(Y_{i-j} - \mathbf{X}_{i-j}^T\widehat{\beta}_{h_0} - \widehat{m}_{h_0,h_0}(\mathbf{T}_{i-j})\right) \quad (i = \mathcal{T}+1, \ldots, n). \tag{9}$$

Finally, based on (8) and (9), we construct the new estimator

$$\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\mathbf{t}) = \sum_{i=\mathcal{T}+1}^{n} w_{h_1,i}(\mathbf{t})(\underline{\widehat{Y}}_{\mathcal{T},i} - \mathbf{X}_i^T\widehat{\beta}_{h_0}). \tag{10}$$

Remark 1. Observe that, as in the construction of $\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\mathbf{t})$, one could construct an estimator $\underline{\widehat{\beta}}_{\mathcal{T},h_0}$ for β based on (8) and (9), and then introduce $\underline{\widehat{\beta}}_{\mathcal{T},h_0}$ instead of $\widehat{\beta}_{h_0}$ in (10). Nevertheless, as we will see in the proof of Theorem 2(b) below, the really important feature of the estimator for β introduced in (10) is its root-consistency, which is reached by $\widehat{\beta}_{h_0}$ (see Theorem 1 below). Given the expression of $\widehat{\beta}_{h_0}$ is simpler than that of $\underline{\widehat{\beta}}_{\mathcal{T},h_0}$, we have preferred to introduce $\widehat{\beta}_{h_0}$.

For simplicity, the dependence of $\underline{\widehat{Y}}_{\mathcal{T},i}$ and $\underline{\widehat{m}}_{\mathcal{T},h_0,h_1}(\cdot)$ on $\mathcal{T}$ will be suppressed.

3 Asymptotic properties

In this section, we present the asymptotic normality of the estimators $\widehat{\beta}_{h_0}$, $\widehat{m}_{h_0,h_1}(\mathbf{t})$ and $\underline{\widehat{m}}_{h_0,h_1}(\mathbf{t})$ defined in (4), (5) and (10), respectively. Then, the asymptotic behaviors of $\widehat{m}_{h_0,h_1}(\mathbf{t})$ and $\underline{\widehat{m}}_{h_0,h_1}(\mathbf{t})$ will be compared. As stated above, the weight functions $w_{h_0,i}(\cdot)$ and $w_{h_1,i}(\cdot)$ used in these estimators are local p-order polynomial type weight functions.

Let $r_j(\mathbf{t}) = E(X_{ij} \mid \mathbf{T}_i = \mathbf{t})$, $\eta_{ij} = X_{ij} - r_j(\mathbf{T}_i)$, $\eta_i = (\eta_{i1}, \ldots, \eta_{id_0})^T$ $(1 \le i \le n$ and $1 \le j \le d_0)$ and $\eta = (\eta_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le d_0}}$. Furthermore, $\underline{k}$, $|\underline{k}|$, $\mathbf{u}$ and $\mathbf{u}^{\underline{k}}$ mean $(k_1, ..., k_{d_1})$ (with $k_i \ge 0$), $\sum_{i=1}^{d_1} k_i$, $(u_1, \ldots, u_{d_1})^T$ and $u_1^{k_1} \times \cdots \times u_{d_1}^{k_{d_1}}$, respectively. The following assumptions will be used to obtain our asymptotic results.

(A1) The kernel $K : \mathbb{R}^{d_1} \to \mathbb{R}$ satisfies $K(\mathbf{u}) = \prod_{j=1}^{d_1} k(u_j)$, where the function $k : \mathbb{R} \to \mathbb{R}$ is bounded, has compact support $[-1, 1]$, is symmetric about zero and $\int k(u)du = 1$. Furthermore, the functions $H_{\underline{j}}(\mathbf{u}) = \mathbf{u}^{\underline{j}} K(\mathbf{u})$ $(0 \le |\underline{j}| \le 2p + 1)$ are Lipschitz continuous.

(A2) For some $L_0 > 4$, $E|\eta_{11}|^{L_0} + \cdots + E|\eta_{1d_0}|^{L_0} < \infty$.

(A3) For some $L_1 > 4$, $E\left(|\varepsilon_1|^{L_1}\right) < \infty$.

(A4) $\mathbf{V}_{\eta,\varepsilon} = \lim_{n\to\infty} n^{-1} E\left(\eta^T \mathbf{V}_\varepsilon \eta\right)$ is a positive definite matrix, where we have denoted $\mathbf{V}_\varepsilon = E\left(\varepsilon\varepsilon^T\right)$ with $\varepsilon^T = (\varepsilon_1, \ldots, \varepsilon_n)$.

(A5) $\mathbf{V}_\eta = E\left(\eta_1 \eta_1^T\right)$ is a positive definite matrix.

(A6) η_i, ε_i and $\mathbf{T}_i$ are independent among themselves $(i = 1, \ldots, n)$.

(A7) $\left\{(Y_i, X_{i1}, \ldots, X_{id_0}, T_{i1}, \ldots, T_{id_1})^T\right\}$ is a stationary strongly mixing process whose mixing coefficients $\alpha(\cdot)$ satisfy $n^x \alpha(n) \to 0$ as $n \to \infty$ for some $x > 7/2$. In addition, the density $f_{\mathbf{T}}$ of $\mathbf{T}_i$ and the joint densities of $(\mathbf{T}_i, \mathbf{T}_{i+l})$, $(\mathbf{T}_i, \mathbf{T}_{i+l}, \mathbf{T}_{i+j})$ and $(\mathbf{T}_i, \mathbf{T}_{i+l}, \mathbf{T}_{i+j}, \mathbf{T}_{i+s})$ are uniformly bounded away from zero on their supports.

(A8) The random vector $\mathbf{T}_i$ is valued in some given compact subset $\mathbb{T}$ of $\mathbb{R}^{d_1}$.

(A9) The process $\{\varepsilon_i\}$ is a stationary and invertible linear process representable in the form (6), and has inverse (7). In addition, there exists $\lambda \in (0, 1)$ such that $|a_j| = O\left(\lambda^j\right)$.

(A10) The functions $m(\cdot), r_1(\cdot), \ldots, r_{d_0}(\cdot)$ are $(p+1)$ times partially differentiable, and their $(p+1)$st-order partial derivatives are Lipschitz continuous on $\mathbb{T}$. Furthermore, the first-order partial derivatives of $f_{\mathbf{T}}$ exist and are continuous on $\mathbb{T}$.

(A11) The bandwidths h_0 and h_1 satisfy that $nh_0^{4q} \to 0$, $n^{-1/2} h_0^{-d_1} \log n \to 0$, $h_0^{2q-d_1} \log n \to 0$, $n^{1/L-1/2} h_0^{-d_1/2} (\log n)^{3/2} \to 0$, $n^{1/\delta_0} h_1^{d_1/2} \to 0$, $h_0/h_1 \to 0$ and $n^{1/2} h_1^{d_1/2} h_0^{2q} \log n \to 0$ as $n \to \infty$, where $q = p + 1$, $4 < \delta_0 < L_0$ and $L = \min\{L_0, L_1\}$.

(A12) There exists a sequence $\{v_n\}$ of positive integers satisfying $v_n \to \infty$ and $v_n = o\left(\left(nh_1^{d_1}\right)^{1/2}\right)$ such that $\left(n/h_1^{d_1}\right)^{1/2} \alpha(v_n) \to 0$ as $n \to \infty$.

(A13) The truncation parameter $\mathcal{T}$ satisfies $\mathcal{T}(n) = c \log n$ for some $c > 0$.

Remark 2. The assumptions above are commonly used for pure nonparametric regression and/or PLR models (see Masry (1996), Xiao et al. (2003) and Aneiros-Pérez et al. (2004), among others). Note that the condition on the

mixing coefficients in (A7) is more restrictive than the corresponding condition for a pure nonparametric regression (i.e., there exist some $v > 2$, $L_1 \geq v$ and $\delta > 1 - 2/v$ such that $E\left(|\varepsilon_1|^{L_1}\right) < \infty$ and $\sum_{i=1}^{\infty} i^{\delta} \left(\alpha(i)\right)^{1-2/v} < \infty$; see Xiao et al. (2003)). This is motivated by the fact that in this last setting, one considers that β is known, while in the PLR model (1) one needs a root-consistent estimator $\widehat{\beta}_{h_0}$. As seen in our Theorem 1, to obtain this consistence it is required that $n^x \alpha(n) \to 0$ as $n \to \infty$ for some $x > 7/2$ (this condition is general enough, satisfied by causal ARMA processes with continuous innovations). A similar comment could be made about the conditions on the bandwidths in Assumption (A11).

Remark 3. Let us assume that $h_0 \sim n^{-a}$ and $h_1 \sim n^{-b}$. Then, if we consider $(4q)^{-1} < a < (2d_1)^{-1}$ and $2\left(d_1\delta_0\right)^{-1} < b < a$, we have that Assumption (A11) holds. However, if $0 < b < d_1^{-1}$, a sufficient condition for (A12) is $\alpha(j) = O\left(j^{-\overline{b}}\right)$ with $\overline{b} > (1 + bd_1)(1 - bd_1)^{-1}$ (then $v_n = \left[\left(nh_1^{d_1}\right)^{1/2} / \log n\right]$ could be considered in (A12)). Thus, under the condition on the mixing coefficients in (A7), we have that sufficient conditions for both (A11) and (A12) are $(4q)^{-1} < a < (2d_1)^{-1}$ and $2\left(d_1\delta_0\right)^{-1} < b < a$.

Following the notation of Masry (1996), let $N_i = (i + d_1 - 1)!/i!(d_1 - 1)!$ be the number of distinct d_1-tuples $\underline{j}$ with $\left|\underline{j}\right| = i$. Arrange these N_i d_1-tuples as a sequence in a lexicographic order (with highest priority to last position so that $(0, \ldots, 0, i)$ is the first element in the sequence and $(i, 0, \ldots, 0)$ is the last element) and let ϕ_i^{-1} denote this one-to-one map. For each $\underline{j}$ with $0 \leq \left|\underline{j}\right| \leq 2p$, let $\mu_{\underline{j}}(K) = \int \mathbf{u}^{\underline{j}} K(\mathbf{u}) d\mathbf{u}$ and $v_{\underline{j}}(K) = \int \mathbf{u}^{\underline{j}} K^2(\mathbf{u}) d\mathbf{u}$, and define the $N \times N$ dimensional matrices $\mathbf{M}$ and $\overline{\mathbf{\Gamma}}$ and the $N \times 1$ vector $\mathbf{B}$, where $N = \sum_{i=0}^{p} N_i$, by

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{0,0} & \mathbf{M}_{0,1} & \cdots & \mathbf{M}_{0,p} \\ \mathbf{M}_{1,0} & \mathbf{M}_{1,1} & & \mathbf{M}_{1,p} \\ \vdots & & & \vdots \\ \mathbf{M}_{p,0} & \mathbf{M}_{p,1} & \cdots & \mathbf{M}_{p,p} \end{bmatrix}, \ \mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_{0,0} & \mathbf{\Gamma}_{0,1} & \cdots & \mathbf{\Gamma}_{0,p} \\ \mathbf{\Gamma}_{1,0} & \mathbf{\Gamma}_{1,1} & & \mathbf{\Gamma}_{1,p} \\ \vdots & & & \vdots \\ \mathbf{\Gamma}_{p,0} & \mathbf{\Gamma}_{p,1} & \cdots & \mathbf{\Gamma}_{p,p} \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} \mathbf{M}_{0,p+1} \\ \mathbf{M}_{1,p+1} \\ \vdots \\ \mathbf{M}_{p,p+1} \end{bmatrix},$$

where $\mathbf{M}_{i,j}$ and $\mathbf{\Gamma}_{i,j}$ are $N_i \times N_j$ dimensional matrices whose (a, b)-th elements are $\mu_{\phi_i(a)+\phi_j(b)}(K)$ and $v_{\phi_i(a)+\phi_j(b)}(K)$, respectively. In addition, we arrange the N_s real numbers

$$\frac{1}{k_1! \cdots k_{d_1}!} \frac{\partial^{\underline{k}} m(\mathbf{t})}{\partial t_1^{k_1} \cdots \partial t_{d_1}^{k_{d_1}}}, \text{ where } \underline{k} = (k_1, \ldots, k_{d_1}) \text{ with } |\underline{k}| = s,$$

as a column vector $\mathbf{m}^{(s)}(\mathbf{t})$ using the lexicographical order introduced above.

Now, we present our results.

Theorem 1. *Under Assumptions (A1)-(A11) we have that*

$$\sqrt{n}\left(\widehat{\beta}_{h_0}-\beta\right)\xrightarrow{d}N\left(\mathbf{0},\mathbf{V}_{\eta}^{-1}\mathbf{V}_{\eta,\varepsilon}\mathbf{V}_{\eta}^{-1}\right).$$

Theorem 2. *Let* $\mathbf{t}$ *be an interior point of* $\mathbb{T}\subset\mathbb{R}^{d_1}$.

(a) Under Assumptions (A1)-(A12) we have that

$$\sqrt{nh_1^{d_1}}\left(\widehat{m}_{h_0,h_1}(\mathbf{t})-m(\mathbf{t})-h_1^q\left[\mathbf{M}^{-1}\mathbf{B}\mathbf{m}^{(q)}(\mathbf{t})\right]_{0,0}\right)\xrightarrow{d}N\left(0,\frac{\sigma_{\varepsilon}^2}{f_T(\mathbf{t})}\mathbf{V}\right).$$

(b) Under Assumptions (A1)-(A11) and (A13) we have that

$$\sqrt{nh_1^{d_1}}\left(\underline{\widehat{m}}_{h_0,h_1}(\mathbf{t})-m(\mathbf{t})-h_1^q\left[\mathbf{M}^{-1}\mathbf{B}\mathbf{m}^{(q)}(\mathbf{t})\right]_{0,0}\right)\xrightarrow{d}N\left(0,\frac{\sigma_{e}^2}{f_T(\mathbf{t})}\mathbf{V}\right),$$

where $\sigma_{\varepsilon}^2=Var(\varepsilon_i)$, $\sigma_{e}^2=Var(e_i)$, $\mathbf{V}=\left[\mathbf{M}^{-1}\mathbf{\Gamma}\mathbf{M}^{-1}\right]_{0,0}$ *and* $[\mathbf{A}]_{0,0}$ *signifies the upper-left element of matrix* $\mathbf{A}$.

Remark 4. From Theorem 1 we have that $\widehat{\beta}_{h_0}$ is a root-n consistent estimator for β, this kind of consistency being a key result in the proofs corresponding to Theorem 2. In fact, Theorem 2 holds if $\widehat{\beta}_{h_0}$ in the expressions of $\widehat{m}_{h_0,h_1}(\mathbf{t})$ and $\underline{\widehat{m}}_{h_0,h_1}(\mathbf{t})$ is replaced with any root-n consistent estimator for β. However, Theorem 2 extends the results existing in pure nonparametric regression models (Xiao et al. (2003)) to the case of PLR models. From this, two interesting conclusions are found: the existence of a linear component does not change the asymptotic distribution of the nonparametric estimator (that is, from an asymptotic perspective, the fact of knowing or not the value of β in (1) is irrelevant), and the estimator $\underline{\widehat{m}}_{h_0,h_1}(\mathbf{t})$ is asymptotically more efficient than the estimator $\widehat{m}_{h_0,h_1}(\mathbf{t})$ (note that both estimators asymptotically have the same bias but different variances, the variance of $\widehat{m}_{h_0,h_1}(\mathbf{t})$ relative to the variance $\underline{\widehat{m}}_{h_0,h_1}(\mathbf{t})$ being $\sigma_{\varepsilon}^2/\sigma_{e}^2=\sum_{j=0}^{\infty}c_j^2\geq 1$, equally holding if and only if $\{\varepsilon_i\}$ is i.i.d.).

Remark 5. In cases of fixed design for $\mathbf{T}_i$, prewhitening a pure nonparametric regression model has no (asymptotically) effect on the efficiency of the corresponding estimator of the regression function (see Francisco-Fernández and Vilar-Fernández (2001)), this holding when estimating the nonparametric part in a PLR model (see Aneiros-Pérez and Quintela-del-Río (2001)). From the work of Xiao et al. (2003) and our Theorem 2, in the case of random design for $\mathbf{T}_i$, the results are completely different.

Remark 6. Both estimators (proposed and conventional) were compared using a simulation study, and the better performance of the new estimator was apparent from the curve estimation perspective as well as from the point estimation perspective. Finally, both the usefulness of the PLR model and the competitiveness of the prewhitening transformation were illustrated by application to a financial time series. We hope to have the opportunity to present both analysis at the congress.

Acknowledgements:
Research was supported by MEC Grant (ERDF included) MTM2005-00429 and Xunta de Galicia Grant PGIDIT07PXIB105259PR.

References

ANEIROS-PÉREZ, G., GONZÁLEZ-MANTEIGA, W. and VIEU, P. (2004): Estimation and testing in a partial linear regression model under long-memory dependence. *Bernoulli 10, 49-78.*

ANEIROS-PÉREZ, G. and QUINTELA-DEL-RÍO, A. (2001): Asymptotic properties in partial linear models under dependence. *Test 10, 333-355.*

ENGLE, R., GRANGER, C., RICE, J. and WEISS, A.(1986): Nonparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association 81, 310-320.*

FRANCISCO-FERNÁNDEZ, M. and VILAR-FERNÁNDEZ, J.M. (2001): Local polynomial regression estimation with correlated errors. *Communications in Statististics-Theory and Methods 30, 1271-1293.*

GAO, J.T. (1995): Asymptotic theory for partly linear models. *Communications in Statististics-Theory and Methods 24, 1985-2009.*

HÄRDLE, W., LIANG, H. and GAO, J.T. (2000): *Partially Linear Models.* Physica-Verlag, Heidelberg.

MASRY, E. (1996): Multivariate regression estimation: local polynomial fitting for time series. *Stochastic Processes and their Applications 65, 81-101.*

ROBINSON, P. (1988): Root-n-consistent semiparametric regression. *Econometrica 56, 931-954.*

SPECKMAN, P. (1988): Kernel smoothing in partial linear models.*Journal of the Royal Statistical Society, Series B 50, 413-436.*

STONE, C. (1977): Consistent nonparametric regression. *Annals of Statistics 5, 595-645.*

XIAO, Z., LINTON, O.B., CARROLL, R.J. and MAMMEN, E. (2003): More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association 98, 980-992.*

Stochastic Orders Based on the Percentile Residual Life Function

Alba María Franco Pereira[1], Rosa Elvira Lillo Rodriguez[1], Juan Romo[1], and Moshe Shaked[2]

[1] Faculdad de Ciencias Sociales y Jurídicas, Universidad Carlos III de Madrid
Calle Madrid, 126-28903 Getafe (Madrid), España, *alba.franco@uc3m.es*, *rosaelvira.lillo@uc3m.es*, *juan.romo@uc3m.es*
[2] The University of Arizona
617 N. Santa Rita Ave., P.O. Box 210089 Tucson AZ 85721-0089, USA, *shaked@math.arizona.edu*

Abstract. In this paper we introduce and study a family of stochastic orders indexed by $\alpha \in (0,1)$. Fixed $\alpha \in (0,1)$ the α-percentile residual life order compares pointwise the α-percentile residual life functions of two random variables. The meaning of these stochastic orders, their properties, and relationship to other common stochastic orders are examined and investigated.

Keywords: percentile residual life function, stochastic orders, reliability theory

1 Introduction

Let X be a random variable, and let u_X be the right endpoint of its support. For any $t < u_X$, the *residual life* at time t, that is associated with X, is any random variable that has the conditional distribution of $X - t$ given that $X > t$. We denote it by

$$X_t = [X - t | X > t], \qquad t < u_X. \tag{1}$$

If F_X denotes the distribution function of X, and $\bar{F}_X = 1 - F_X$ denotes the corresponding survival function, then the survival function of X_t is given by

$$\bar{F}_{X_t}(x) = \frac{\bar{F}_X(t+x)}{\bar{F}_X(t)}, \qquad x \geq 0.$$

The residual life is of interest in many areas of applied probability and statistics such as actuarial studies, biometry, survivorship analysis, and reliability; see, for example, Lillo (2005) for a list of references.

The mean residual life function m_X that is associated with X is given by

$$m_X(t) = \begin{cases} E[X - t | X > t], & t < u_X; \\ 0, & t \geq u_X, \end{cases}$$

provided the expectation exists. It is a useful tool for analyzing important properties of X when it exists. However, the mean residual life function may not exist. Even when it exists it may have some practical shortcomings, especially in situations where the data are censored, or when the underlying distribution is skewed or heavy-tailed. In such cases, either the empirical mean residual life function cannot be calculated, or a single long-term survivor can have a marked effect upon it which will tend to be unstable due to its strong dependence on very long durations.

An alternative to the mean residual life function is the α-percentile residual life function $q_{X,\alpha}$, where α is some number between 0 and 1. This function is defined for any $t < u_X$ by letting $q_{X,\alpha}$ be the α-percentile of X_t. A formal definition of $q_{X,\alpha}$ will be given in Section 2, but here we note that such a function describes, for example, the value that will be survived, by $(1-\alpha)\%$ of items (in reliability theory) or of individuals (in biology), that survived up to time t. The α-percentile residual life functions were studied in some detail by Arnold and Brockett (1983), Gupta and Langford (1984), Joe and Proschan (1984a), and Joe (1985), as well as by Haines and Singpurwalla (1974).

The purpose of this paper is to introduce and study a family of stochastic orders indexed by $\alpha \in (0, 1)$. Fixed $\alpha \in (0, 1)$ the αth order compares pointwise $q_{X,\alpha}$ with $q_{Y,\alpha}$, where the latter is the α-percentile residual life function of a random variable Y. These stochastic orders were introduced in Joe and Proschan (1984b), but their properties were not extensively studied.

Some conventions that we use in this paper are the following. By "increasing" and "decreasing" we mean "nondecreasing" and "nonincreasing", respectively. For any distribution function F we let function F^{-1} be the left continuous version of the inverse of F, that is

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}, \qquad p \in [0, 1]. \tag{2}$$

2 Definition

Let X be a random variable. The α-percentile residual life function $q_{X,\alpha}$ is defined by

$$q_{X,\alpha}(t) = \begin{cases} F_{X_t}^{-1}(\alpha), & t < u_X; \\ 0, & t \geq u_X, \end{cases} \tag{3}$$

A straightforward computation shows that

$$q_{X,\alpha}(t) = \bar{F}_X^{-1}((1-\alpha)\bar{F}_X) - t, \qquad t < u_X.$$

Similar expressions can be found in Joe and Proschan (1984b). Note that, unlike Joe and Proschan (1984a,b), we do not assume here that X is a non-negative random variable.

Now let Y be another random variable, and let $q_{Y,\alpha}$ be its α-percentile residual life function. If

$$q_{X,\alpha}(t) \leq q_{Y,\alpha}(t) \qquad \text{for all } t, \tag{4}$$

then we say that X is smaller that Y in the *α-percentile residual life order*, and we denote it as $X \leq_{\alpha-rl} Y$. This inequality defines a family of stochastic orders, indexed by $\alpha \in (0,1)$.

3 Relationship to other stochastic orders

It is interesting to study the relationship between this new family of stochastic orders and some other well-known stochastic orders. Recall the following definitions (for more details see, for example, Shaked and Shanthikumar (2007)):

A random variable X is said to be smaller or equal than the random variable Y in the ordinary stochastic order (denoted as $X \leq_{st} Y$) if $\bar{F}_X(x) \leq \bar{F}_Y(x)$, for all $x \in \mathbb{R}$.

A random variable X is said to be smaller or equal than the random variable Y in the hazard rate order (denoted as $X \leq_{hr} Y$) if $\bar{F}_X(x)\bar{F}_Y(y) \geq \bar{F}_X(y)\bar{F}_Y(x)$, for all $x \leq y$.

A random variable X is said to be smaller or equal than the random variable Y in the mean residual life order (denoted as $X \leq_{mrl} Y$) if $m_X(x) \leq m_Y(x)$, for all $x \in \mathbb{R}$.

A continuous random variable X with density f_X is said to be smaller or equal than the continuous random variable Y with density f_Y in the likelihood ratio order (denoted as $X \leq_{lr} Y$) if $f_X(x)f_Y(y) \geq f_X(y)f_Y(x)$, for all $x \leq y$.

For any $\alpha \in (0,1)$, the following diagram reflects the relationship between the α-percentile residual life order and other common stochastic orders:

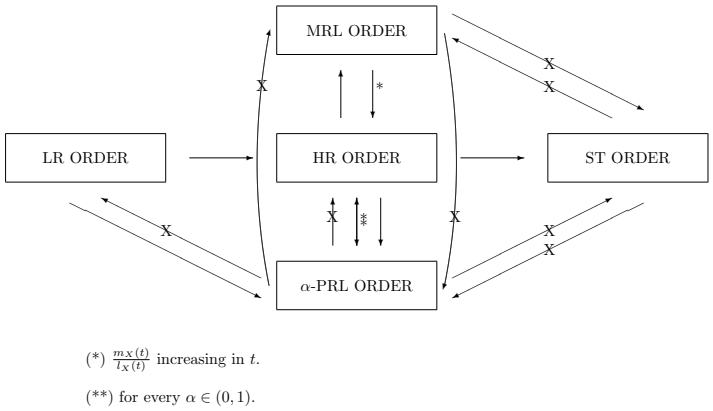


Fig. 1. Relationship among some stochastic orders.

4 Closure properties

The α-percentile residual life orders satisfy some desirable closure properties. These properties are described in this section.

The α-percentile residual life orders are preserved under strictly increasing transformations. That is,

$$X \leq_{\alpha-rl} Y \Leftrightarrow \phi(X) \leq_{\alpha-rl} \phi(Y), \text{ for every strictly increasing function } \phi.$$

The α-percentile residual life orders are closed under mixtures. That is, let X, Y , U, and V be random variables with continuous distribution functions, and let W be a random variable with distribution function $F_W = pF_X + (1-p)F_Y$, for some $p \in [0,1]$. Then,

(i) If $U \leq_{\alpha-rl} X$ and $U \leq_{\alpha-rl} Y$ then $U \leq_{\alpha-rl} W$.
(ii) If $X \leq_{\alpha-rl} V$ and $Y \leq_{\alpha-rl} V$ then $W \leq_{\alpha-rl} V$.

The possible preservation of a stochastic order under the formation of coherent systems is a useful property that has important applications in reliability theory (see, for example, Barlow and Proschan (1975) for the definition and the use of coherent systems). Thus it is of interest to ask whether the α-percentile residual life orders are closed under this formation. Boland, El-Neweihi, and Proschan (1994) showed that the hazard rate order is not preserved under the formation of coherent systems. In this case we have that, for all α, the α-percentile residual life order is not closed under this formation. In fact, unlike the hazard rate order, for every $\alpha \in (0,1)$, the α-percentile

residual life order is not even closed under the formation of series systems (that is, under the minimum operation).

5 Some applications of the median residual life order

In this section we present an application of the median residual life order. The data were collected from the book of Kalbfleisch and Prentice (1980) and provides data for a part of a large clinical trial carried out by the Radiation Therapy Oncology Group in the United States. The full study included patients with squamous carcinoma of 15 sites in the mouth and throat, with 16 participating institutions, though only data on three sites in the oropharynx reported by the six largest institutions are considered here. Patients entering the study were randomly assigned to one of two treatment groups, radiation therapy alone or radiation therapy together with a chemotherapeutic agent. Our objective is to compare the two treatment policies with respect to patient survival.

Approximately 30% of the survival times are censored owing primarily to patients surviving to the time of analysis. Some patients were lost to follow-up because the patient moved or transferred to an institution not participating in the study, though these cases were relatively rare. From a statistical point of view, the main feature of these data that distinguishes this example from others is the considerable lack of homogeneity among the individuals being studied.

In order to explain which treatment is more efficient in the sense of the survival time of the patients after each treatment, the mean residual life functions of the survival time under both kinds of treatment can be compared. However, since the densities of the survival time of the patients are, in both cases, asymmetric, the use of the mean residual life function may not be appropriate in this case. The mean residual life function present a strong dependence to the underlying distribution when it is asymmetric. As an alternative, the median residual life functions can be compared. The percentile residual life function is more robust than the mean residual life function in the sense that it does not depend as much on the underlying distribution. Therefore, if we compare the median residual life functions under both kinds of treatment, we will get a more reliable conclusion.

In this case we see that comparing the mean residual life functions and the median residual life functions for the data of Kalbfleisch and Prentice (1980) lead to different conclusions. We analyze the advantages of using the median residual life order in this context and similar contexts.

References

ARNOLD, B.C. and BROCKETT, P.L. (1983): When does the βth percentile residual life function determine the distribution? *Operations Research 31, 391-396.*

BARLOW, R.E. and PROSCHAN, F. (1975): *Statistical Theory of Reliability and Life Testing: Probability Models.* Holt, Rinehart & Winston, New York.

BOLAND, J. ,EL-NEWEIHI, E. and PROSCHAN, F. (1994): Applications of the hazard rate ordering in reliability and orders tatistics. *Journal of the Applied Probability 31,180-195.*

GUPTA, R.C. (1975): On the characterization of distributions by conditional expectations. *Communications in Statistics 4, 99-103.*

HAINES, A.L. and SINGPURWALLA, N.D. (1974): Some contributions to the stochastic characterization of wear. In: F. Proschan and R. J. Serfling (Eds.): *Reliability and Biometry, Statistical Analysis of Lifelength.* SIAM, Philadelphia, 47–80.

JOE, H. (1985): Characterizations of life distributions from percentile residual lifetimes. *Annals of the Institute of Statistical Mathematics 37, 165-172*

JOE, H. and PROSCHAN, F. (1984a): Percentile residual life fucntions. *Operational Research 32, 668-678*

JOE, H. and PROSCHAN, F. (1984b): Comparison of two life distributions on the basis of their percentile residual life functions. *Canadian Journal of Statistics 12, 91-97*

KALBFLEISCH, J.D. and PRENTICE, R.L. (1980): *The statistical analysis of failure time data.* John Wiley, New York.

LILLO, R.E. (2005): On the median residual lifetime and its aging properties: A characterization theorem and applications. *Naval Research Logistics 52, 370-380*

SHAKED, M. and SHANTHIKUMAR, J.G. (2007): *Stochastic Orders.* Springer, New York.

An Improved Estimator for Removing Boundary Bias in Kernel Cumulative Distribution Function Estimation

Jan Koláček

Department of Mathematics and Statistics, Masaryk University
Janáčkovo nám. 2a, 602 00 Brno, Czech Republic, *kolacek@math.muni.cz*

Abstract. In this paper we focus on kernel estimates of cumulative distribution functions in case that random variables $X_1, \dots X_n$ are nonnegative. It is well known that kernel distribution estimators are not consistent when estimating a distribution function near the point $x = 0$. This fact is regrettable in many applications, for example in kernel ROC curve estimation (Koláček and Karunamuni (2007)). In order to avoid this problem we propose a bias reducing technique which is a kind of generalized reflection method. Our method is based on ideas of Karunamuni and Alberts (2005) and Zhang et al. (1999) developed for boundary correction in kernel density estimation. Finally, the proposed estimator is compared with the traditional kernel estimator and with the estimator based on "classical" reflection method using simulation studies.

Keywords: kernel estimation, reflection, distribution estimation

1 Introduction

The most commonly used nonparametric estimate of a cumulative distribution function F is an empirical distribution function F_n. But F_n is a step function even in case that F is continuous. Another type of nonparametric estimators for F is derived from kernel smoothing methods. Kernel smoothing is most widely used because it is easy to derive and has good properties. Kernel smoothing has received a lot of attention in density estimation. Good references in this area are Gasser at al. (1985), Silverman (1986) and Wand and Jones (1995). However, results in kernel distribution function estimation are relatively few. Theoretical properties of kernel distribution function estimator have been investigated by Nadaraya (1964), Reiss (1981) and Azzalini (1981). Although there is a vast literature on boundary correction in density estimation context, boundary effects problem in distribution function context has been less studied.

In this paper, we develop a new kernel type estimator of the cumulative distribution function that removes boundary effects near the end points of the support. Our estimator is based on a new boundary corrected kernel estimator of distribution functions and it is based on ideas of Karunamuni

and Alberts (2005) and Zhang et al. (1999) developed for boundary correction in kernel density estimation. The basic technique of construction of the proposed estimator is kind of a generalized reflection method involving reflecting a transformation of the observed data. In fact, the proposed method generates a class of boundary corrected estimators. We derive expressions for the bias and variance of the proposed estimator. Furthermore, the proposed estimator is compared with the traditional estimator and with the estimator based on "classical" reflection method using simulation studies. We observe that the proposed estimator successfully remove boundary effects and performs considerably better than the others two.

Kernel smoothing in distribution function estimation and boundary effects are discussed in the next section. The proposed estimator is given in Section 3. Simulation results are given in Section 4. Finally, some concluding remarks are given in Section 5.

2 Kernel distribution estimator and boundary effects

Let f denote a continuous density function with support $[0, a]$, $0 < a \leq \infty$, and consider nonparametric estimation of the cumulative distribution function F of f based on a random sample X_1, ..., X_n from f. Suppose that $F^{(j)}$, the j-th derivative of F, exists and is continuous on $[0, a]$, $j = 0, 1, 2$, with $F^{(0)} = F$ and $F^{(1)} = f$. Then the traditional kernel estimator of F is given by

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \qquad W(x) = \int_{-1}^{x} K(t)dt$$

where K is a unimodal symmetric density function with support $[-1, 1]$ and h is the bandwidth ($h \to 0$ as $n \to \infty$). Set $\beta_2 = \int_{-1}^{1} t^2 K(t)dt$. The basic properties of $\widehat{F}_{h,K}(x)$ at interior points are well-known (e.g. Lejeune and Sarda (1992)), and under some smoothness assumptions these include, for $h \leq x \leq a - h$,

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = \frac{1}{2}\beta_2 f^{(1)}(x)h^2 + o(h^2)$$

$$n\mathrm{Var}(\widehat{F}_{h,K}(x)) = F(x)(1 - F(x)) + hf(x)\int_{-1}^{1} W(t)(W(t) - 1)dt + o(h).$$

The performance of $\widehat{F}_{h,K}(x)$ at boundary points, i.e., for $x \in [0, h) \cup (a-h, a]$, however, differs from the interior points due to so-called "boundary effects" that occur in nonparametric curve estimation problems. More specifically,

the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ instead of $O(h^2)$ at boundary points while the variance of $\widehat{F}_{h,K}(x)$ is of the same order. This fact can be clearly seen by examining the behavior of $\widehat{F}_{h,K}$ inside the left boundary region $[0, h]$. Let x be a point in the left boundary, i.e., $x \in [0, h]$. Then we can write $x = ch,\ 0 \leq c \leq 1$. The bias and variance of $\widehat{F}_{h,K}(x)$ at $x = ch$ are of the form

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = hf(0) \int_{-1}^{-c} W(t)dt + h^2 f^{(1)}(0) \left\{ \frac{c^2}{2} + c \int_{-1}^{-c} W(t)dt - \int_{-1}^{c} tW(t)dt \right\} + o(h^2) \tag{1}$$

$$n\mathrm{Var}(\widehat{F}_{h,K}(x)) = F(x)(1 - F(x)) + hf(0) \left\{ \int_{-1}^{c} W^2(t)dt - c \right\} + o(h). \tag{2}$$

From the expression (1) it is now clear that the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ instead of $O(h^2)$.To remove this boundary effect in kernel distribution estimation we investigate a new class of estimators in the next section.

3 The proposed estimator

In this section we propose a class of estimators of the distribution function F of the form

$$\widetilde{F}_{h,K}(x) = \frac{1}{n} \sum_{i=1}^{n} \left\{ W\left(\frac{x - g_1(X_i)}{h} \right) - W\left(-\frac{x + g_2(X_i)}{h} \right) \right\}, \tag{3}$$

where h is the bandwidth, K is a symmetric density function with support $[-1, 1]$ and g_1 and g_2 are two transformations that need to be determined. We assume that $g_i,\ i = 1, 2$, are nonnegative, continuous and monotonically increasing functions defined on $[0, \infty)$. Further assume that g_i^{-1} exists, $g_i(0) = 0,\ g_i^{(1)}(0) = 1$, and that $g_i^{(2)}$ exists and is continuous on $[0, \infty)$, where $g_i^{(j)}$ denotes the j-th derivative of g_i, with $g_i^{(0)} = g_i$ and g_i^{-1} denoting the inverse function of $g_i,\ i = 1, 2$. We will choose g_1 and g_2 so that $\widetilde{F}_{h,K}(x) \geq 0$ everywhere. Note that the i-th term of the sum in (3) can be expressed as

$$W\left(\frac{x - g_1(X_i)}{h} \right) - W\left(-\frac{x + g_2(X_i)}{h} \right) = \int_{\frac{-x+g_1(X_i)}{h}}^{\frac{x+g_2(X_i)}{h}} K(t)dt.$$

The preceding integral is non-negative provided the inequality $\frac{-x+g_1(X_i)}{h} \leq \frac{x+g_2(X_i)}{h}$ holds. Since $x \geq 0$, the preceding inequality will be satisfied if g_1 and g_2 are such that $g_1(X_i) \leq g_2(X_i)$ for $i = 1, \ldots, n$. Thus we will assume that g_1 and g_2 are chosen such that $g_1(x) \leq g_2(x)$ for $x \in [0, \infty)$ for our proposed estimator. Now, we can obtain the bias and variance of (3) at $x = ch, 0 \leq c \leq 1$, as

$$\begin{aligned} \mathrm{E}(\widetilde{F}_{h,K}(x)) - F(x) = h^2 \Bigg\{ & f^{(1)}(0) \left(\frac{c^2}{2} + 2c \int_{-1}^{-c} W(t)dt - \int_{-c}^{c} tW(t)dt \right) \\ & - f(0) g_1^{(2)}(0) \int_{-1}^{c} (c-t) W(t)dt \\ & - f(0) g_2^{(2)}(0) \int_{-1}^{-c} (c+t) W(t)dt \Bigg\} + o(h^2). \end{aligned} \tag{4}$$

$$\begin{aligned} n\mathrm{Var}(\widetilde{F}_{h,K}(x)) = F(x)(1-F(x)) + hf(0) \Bigg\{ & \int_{-1}^{c} W^2(t)dt \\ & - 2 \int_{-1}^{c} W(t) W(t-2c)dt + \int_{-1}^{-c} W^2(t)dt \Bigg\} + o(h). \end{aligned} \tag{5}$$

The proofs of (4) and (5) are given in Koláček and Karunamuni (2007). Similarly we could express the bias and variance of (3) at "interior" points $x = c > 1$. Note that the contribution of g_2 on the bias vanishes as $c \to 1$. By comparing expressions (1), (4), (2) and (5) at boundary points we can see that the variances are of the same order and the bias of $\widehat{F}_{h,K}(x)$ is of order $O(h)$ while the bias of $\widetilde{F}_{h,K}(x)$ is of order $O(h^2)$. So our proposed estimator removes boundary effects in kernel distribution estimation since the bias at boundary points is of the same order as the bias at interior points.

It is clear that there are various possible choices available for the pair (g_1, g_2). However, we will choose g_1 and g_2 so that the condition $\widetilde{F}_{h,K}(0) = 0$ will be satisfied because of the fact that $F(0) = 0$. A sufficient (but not necessary) condition for the preceding to be satisfied is that g_1 and g_2 must be equal. Thus we need to construct a single transformation function g such that $g = g_1 = g_2$. Other important properties that are desirable in the estimator $\widetilde{F}_{h,K}$ are the local adaptivity, that is the transformation function g depends on c.

Some discussion on the choice of g_c and other various improvements that can be made would be appropriate here. The trivial choice is $g_c(y) = y$, which represents the "classical" reflection method estimator. However, it is possible to construct functions g_c's that improve the bias further under some

additional conditions. For instance, if one examines the right hand side of bias expansion (4) then it is not difficult to see that the terms inside bracket (i.e., the coefficient of h^2) can be made equal to zero if g_c is appropriately chosen. Set

$$A_c = \begin{cases} d_1 \frac{\frac{c^2}{2}+2cI_1-I_2}{c^2+2cI_1-I_2}, & \text{for } 0 \le c < 1 \\ d_1 \frac{\beta_2}{c^2+\beta_2}, & \text{for } c > 1 \end{cases}$$

where $d_1 = \frac{f^{(1)}(0)}{f(0)}$, $I_1 = \int_{-1}^{-c} W(t)dt$, $I_2 = \int_{-c}^{c} tW(t)dt$.

If g_c is chosen such that $g_c^{(2)}(0) = A_c$ then the bias of $\widetilde{F}_{h,K}(x)$ would be theoretically of order $O(h^3)$. For such a function g_c, the second derivative at zero $g_c^{(2)}(0)$ will be dependent on the ratio $d_1 = \frac{f^{(1)}(0)}{f(0)}$. Then the problem of estimation of d_1 naturally arises as in the paper Karunamuni and Zhang (2007). The ratio $d_1 = \frac{f^{(1)}(0)}{f(0)}$ is estimated there as the first derivative of natural logarithm of f at zero. For more details, especially for the exact formula for $\widehat{d_1}$ and for some statistical properties, especially for the asymptotic convergence rate, see the preceding paper.

Summarizing all the assumptions, it is clear now that g_c should satisfy the following conditions:

(i) $g_c : [0, \infty) \to [0, \infty)$, g_c is continuous, monotonically increasing and $g_c^{(i)}$ exists, $i = 1, 2$,
(ii) $g_c^{-1}(0) = 0$, $g_c^{(1)}(0) = 1$
(iii) $g_c^{(2)}(0) = A_c$.

Functions satisfying conditions (i) – (iii) are easy to construct. We will consider the following transformation. For $y \ge 0$, let us define

$$g_c(y) = y + \frac{1}{2}\widehat{A}_c y^2 + \lambda \widehat{A}_c^2 y^3, \tag{6}$$

where $\widehat{A}_c$ is an estimator of A_c based on the estimator $\widehat{d_1}$ and λ is a positive constant such that $\lambda > \frac{1}{12}$. This condition on λ is necessary for $g_c(y)$ to be an increasing function of y. Based on extensive simulations, we find that this transformation adapts well to various shapes of distributions with setting $\lambda = 0.1$.

4 A simulation study

To test the effectiveness of our estimator, we simulated its performance against the classical reflection method. The simulation is based on 1 000 replications. In each replication, the random variables $X \sim Exp(0.005)$ were generated and the estimate of cumulative distribution function was computed.

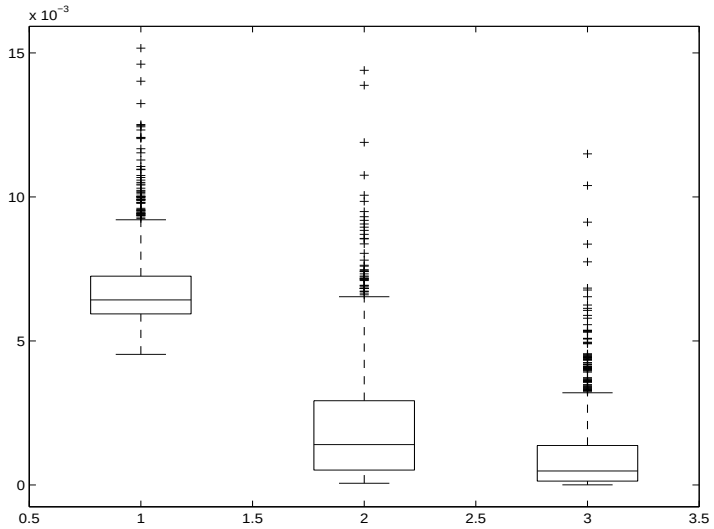


Fig. 1. MISE for estimates of CDF for the classical estimator with boundary effects (1), the reflection method (2) and for our proposed method (3).

The parameter 0.005 was considered in connection with the work of Dette and Weissbach (2007) with applications to credit risk.

In all replications the sample size of $n = 100$ was used. In this case, the actual global optimal bandwidth (see Azzalini (1981)) for F is $h_F = 231.36$. For the kernel estimation of cumulative distribution we have used the quartic kernel $K(x) = \frac{15}{16}(1 - x^2)^2 I_{[-1,1]}$, where I_A is the indicator function on the set A.

For each distribution function we have calculated the Mean Integrated Squared Error (MISE) on the interval $[0, h_F]$ over all 1 000 replications and have displayed the results in a boxplot in Figure 1. The variance of each estimator can be accurately gauged by the whiskers of the plot. The values of means and standard deviations for MISE of each method are given in Table 2. As we can see the reflection method gives the smaller values of MISE than the classical estimator, but the variance is not so small. From this point of view the proposed estimator seems to be better.

Method	*Mean*	*STD*
Classical	0.0068	0.0014
Reflection	0.0020	0.0020
Proposed	0.0010	0.0014

Table 1. Means and STD's for MISE.

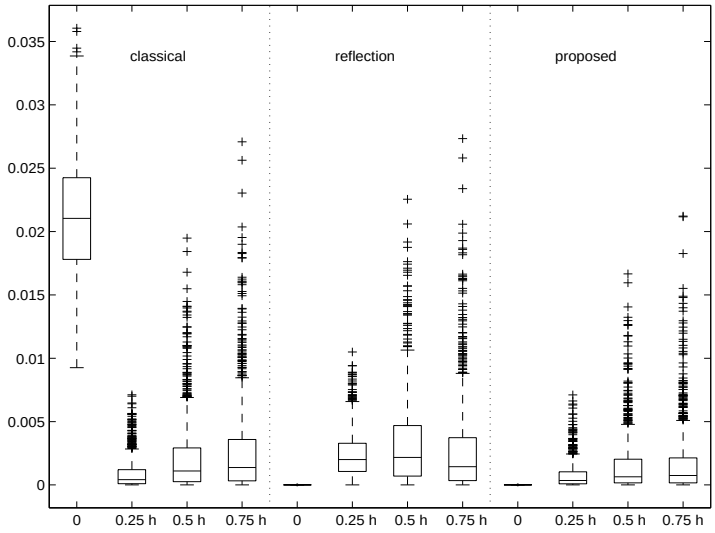


Fig. 2. MSE at points $x = ch_F$, $c = 0$, 0.25, 0.5, 0.75 for the classical estimator with boundary effects, the reflection method and for our proposed method.

To get more detailed information about estimators we have calculated the Mean Squared Error (MSE) at four points in the boundary region $x = ch_F$, $c = 0$, 0.25, 0.5, 0.75. The boxplot of MSE for each estimator over all 1 000 replications is illustrated in Figure 2. The values of means and standard deviations for MSE at each point for each method are given in Table 3. These values describe the performance of our proposed method with respect to MSE. The values of mean and also of the variance were smallest in the case of our proposed estimator. This is caused by a local adaptivity of our estimator and by the fact that the bias is theoretically of order $O(h^3)$ instead of $O(h^2)$ while the variance is of the same order.

	Classical		*Reflection*		*Proposed*	
c	*Mean*	*STD*	*Mean*	*STD*	*Mean*	*STD*
0.00	0.0215	0.0048	0.0000	0.0000	0.0000	0.0000
0.25	0.0009	0.0013	0.0023	0.0017	0.0008	0.0010
0.50	0.0021	0.0025	0.0032	0.0032	0.0016	0.0021
0.75	0.0026	0.0033	0.0027	0.0034	0.0017	0.0024

Table 2. Means and STD's for MSE at $x = ch_F$.

5 Conclusion

In this paper we proposed a new kernel-type distribution estimator to avoid the difficulties near the boundary. The technique implemented is a kind of generalized reflection method involving reflecting a transformation of the data. The proposed method generates a class of boundary corrected estimators and it is based on ideas of boundary corrections for kernel density estimators presented in Karunamuni and Alberts (2005). We showed some good properties of our proposed method (e.g., local adaptivity). Furthermore, it is shown that bias of the proposed estimator is better than that of the "classical" case.

Acknowledgements: The research was supported by The Jaroslav Hájek center for theoretical and applied statistics (MSMT LC 06024).

References

AZZALINI, A. (1981): A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika 68, 326–328.*

DETTE, H., WEISSBACH, R. (2007): Kolmogorov-Smirnov-type testing for the partial homogeneity of Markov processes – with application to credit risk. *Applied Stochastic Models in Business and Industry, Vol. 23, No. 3, 223–234.*

GASSER, T., MÜLLER, H.G. and MAMMITZSCH, V. (1985): Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B, Vol. 47, No. 2, 238–252.*

KARUNAMUNI, R.J. and ALBERTS, T. (2005): On boundary correction in kernel density estimation. *Statistical Methodology 2, 191–212.*

KARUNAMUNI, R.J. and ZHANG, S. (2007): Some improvements on a boundary corrected kernel density estimator. *Statist. Probab. Lett., in print.*

KOLÁČEK, J. and KARUNAMUNI, R.J. (2007): On boundary correction in kernel estimation of ROC curves. *Austrian Journal of Statistics, in review process.*

LEJEUNE, M. and SARDA, P. (1992): Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis 14, 457–471.*

NADARAYA, E.A. (1964): Some new estimates for distribution functions. *Theory Prob. Appl. 15, 497–500.*

REISS,, R.D. (1981): Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics 8, 116–119.*

SILVERMAN, W.R. (1986): *Density estimation for statistics and data analysis.* Chapman and Hall, London.

WAND, M.P. and JONES, M.C. (1995): *Kernel smoothing.* Chapman and Hall, London.

ZHANG, S., KARUNAMUNI, R.J. and JONES, M.C. (1999): An improved estimator of the density function at the boundary. *Journal of the American Statistical Association 94, 1231–1241.*

Additive Models with Missing Data

Rocío Raya-Miranda[1], María Dolores Martínez-Miranda, Wenceslao González-Manteiga[2] and Andrés González-Carmona[1]

[1] Department of Statistics and R.O., University of Granada
Granada, Spain, *rraya@ugr.es, mmiranda@ugr.es, andresgc@ugr.es*
[2] Department of Statistics and R.O., University of Santiago de Compostela
Santiago de Compostela, Spain,
wenceslao@usc.es

Abstract. This paper deals the nonparametric estimation of additive models in the presence of missing values on the response variable and specifically, in the case of additive models estimated by the backfitting algorithm. Two estimators are presented, one based on the available data and another based on a complete sample from imputation techniques. The performance of the estimators is evaluated, making comparisons between them, in a simulation study considering an additive regression model.

Keywords: additive model, backfitting, nonparametric regression, missing, imputation

1 Introduction

Some observations in samples are often incomplete in practice. For example, let us think about the appealing nonresponse in sample surveys, in clinical essays, in studies with longitudinal and ecological data, etc. Rubin (1976) distinguished among three different missing data: *missing completely at random* (MCAR) whether the missingness does not depend on the variable with missing data nor the observed data; *missing at random* (MAR) when lost data can depend on the observed data but not on the missing data, and finally, *nonignorably missing* (NINR) if the dependence is on both missing and observed data.

The standard nonparametric regression estimation methods consider complete samples. The usual way to deal with missing data is to drop the incomplete observations or substitute/impute the missing values with the mean or a simple parametric regression estimation (over the available variables). Some of these methods are only lightly theoretical supported and suffer from bias problems. Other imputation methods more refined are those of Dempster et al. (1977), which develop the EM algorithm and the method of Rubin (1987) based on multiple imputation. Recently, multiple imputation methods have been proposed based on nonparametric and semiparametric estimation, such as those proposed by Aerts et al. (2002). The paper of Gómez-García et al.

(2006) provides a simulation study evaluating the effect of some imputation methods with different causes for missing data.

Several inferential problems have been considered in the presence of missing data: linear regression, Little (1992); logistic regression, Vach (1994); log-linear models, Fuchs (1982); generalized linear models, Ibrahim (1990), Ibrahim et al. (1999, 2001), etc. However, the nonparametric additive models with missing data have not been paid special attention. In multivariate regression, the missing data can arise on the response and/or the covariates in the model. In case that the loss of data is in the covariates, Wang et al. (1997-1998) considered the problem under a semiparametric or generalized linear model. However, Nittner (2002) estimated an additive model with missing values at the covariates. And in case of missing data in the response, it stands out, on the one hand, the papers of Ibrahim et al. (2001), for semiparametric models, and on the other, the papers of Chu and Cheng (1995) and González-Manteiga and Pérez-González (2004), for general multivariate nonparametric regression models. Chen and Ibrahim (2006) estimated semiparametric models considering missing data for the covariates and the response variable.

For estimating the unconditional mean of the response, $E[Y]$, under a nonparametric treatment, it stands out the following papers: Cheng and Wei (1986), Cheng (1990), (1994), and Nielsen (2001), which considered simple imputation techniques. Also Aerts et al. (2002) and González-Manteiga and Pérez-González (2004), solved the problem using multiple imputation.

It has been considered other problems such as the density estimation with missing data by Titterington and Mill (1983) and Titterington and Sedransk (1989); the estimation of other functions such as the unconditional variance, $Var(Y)$ or the distribution function of Y by Cheng (1994); and the problem of choosing the best model for incomplete samples by Hens et al. (2006).

In this paper we have considered the nonparametric additive estimation of the regression function in the presence of missing data in the response variable. We propose two nonparametric estimators based on the backfitting algorithm with local polynomial smoothers, Opsomer (2000). The first estimator consists of using only the complete observations and the second uses a simple imputation method to complete the sample.

Others alternative methods have been proposed to estimate the additive models, for example: the marginal integration method by Linton and Nielsen (1995) and Tjϕstheim and Auestad (1994); the smooth backfitting by Mammen et al. (1999) and a two step procedure by Horowitz et al. (2003). The estimators presented in the paper can be easily extended for these additive estimation methods.

2 The additive model

For $i = 1, \ldots, n$, it is assumed for one-dimensional response variables $Y_1, \ldots, Y_n$ that

$$Y_i = m_0 + m_1(X_{1i}) + \ldots + m_d(X_{di}) + \varepsilon_i, \tag{1}$$

where ε_i are error variables, $m_1, \ldots, m_d$ are unknown functions satisfying $E[m_j(X_{ji})] = 0$, m_0 is an unknown constant and $\mathbf{X}_i = (X_{1i}, \ldots, X_{di})$ are random variables $\mathtt{R}^d$ $(i = 1, \ldots, n,\ j = 1, \ldots, d)$. Throughout the paper we make the assumption that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent identically distributed (i.i.d.) and that X_{ji} takes its values in a bounded interval I_j. Furthermore, the error variables, $\varepsilon_1, \ldots, \varepsilon_n$, are assumed to be i.i.d. with mean zero and being independent of $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

In the case where no observations are missing, a sample of i.i.d. vectors with respect to the random vector $\{(\mathbf{X}_i^t, Y_i)\}_{i=1}^n$, is available.

In our case it may be possible that Y_i is not observed for any index i. In order to check whether an observation is complete or not, a new variable δ is introduced into the model as an indicator of the missing observations. Thus $\delta_i = 1$ if Y_i is observed and zero if Y_i is missing, for $i = 1, \ldots, n$.

Following the patterns in literature (see Little and Rubin, 2002), we need to establish whether the loss of an item of data is independent or not of the value of the observed data and/or the missing data. In this paper we suppose that the data are missing at random (MAR), i.e.

$$P[\delta = 1|Y, \mathbf{X}] = P[\delta = 1|\mathbf{X}] = p(\mathbf{X}).$$

3 Simple imputation

If there are not missing data, the nonparametric estimation of the component additive functions, $\mathbf{m}_j$, can be given by solving the system of normal equations:

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \cdots & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_d \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} \mathbf{Y}. \tag{2}$$

where $\mathbf{S}_j$ represents the $n \times n$ smoother matrix with respect to de jth covariate vector. The smoother matrices for local polynomial regression are $\mathbf{S}_j = (\mathbf{s}_{j,X_{d1}}, \ldots, \mathbf{s}_{j,X_{dn}})^T$, where $\mathbf{s}_{j,x_j}$ represents the equivalent kernel for the jth covariate at the point x_j:

$$\mathbf{s}_{j,x_j} = \mathbf{e}_1^T \left(\mathbf{X}_{j,x_j}^T \mathbf{K}_{x_j} \mathbf{X}_{j,x_j} \right)^{-1} \mathbf{X}_{j,x_j} \mathbf{K}_{x_j},$$

with $\mathbf{e}_i$ a vector with a one in the ith position and zeros elsewhere, the matrix $\mathbf{K}_{x_j} = \text{diag}\{K_{h_j}(X_{j1} - x_j), \ldots, K_{h_j}(X_{jn} - x_j)\}$ for some kernel function K and bandwidth h_j,

$$\mathbf{X}_{j,x_j} = \begin{bmatrix} 1 & (X_{j1} - x_j) & \cdots & (X_{j1} - x_j)^{p_j} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_{jn} - x_j) & \cdots & (X_{jn} - x_j)^{p_j} \end{bmatrix},$$

and p_j the degree of the local polynomial for fitting $\mathbf{m}_j$.

The backfitting algorithm, Buja et al. (1999) provides an iterative solution of (2). Opsomer (2000) wrote the estimators directly as

$$\begin{bmatrix} \widehat{\mathbf{m}}_1 \\ \widehat{\mathbf{m}}_2 \\ \vdots \\ \widehat{\mathbf{m}}_d \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \cdots & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} \mathbf{Y} \equiv \mathbf{M}^{-1}\mathbf{C}\mathbf{Y},$$

provided the inverse of $\mathbf{M}$ exists. This expression shows that the additive model estimators are also linear smoothers, and it is possible to define the additive smoother matrix $\mathbf{W}_j$ as

$$\mathbf{W}_j = \mathbf{E}_j\mathbf{M}^{-1}\mathbf{C}, \tag{3}$$

where $\mathbf{E}_j$ is a partitioned matrix of dimension $n \times nd$ with an $n \times n$ identity matrix as the jth block and zeros elsewhere, so that $\widehat{\mathbf{m}}_j = \mathbf{W}_j\mathbf{Y}$, for $j = 1, \ldots, d$.

If there are missing observations in the response variable, a very simple way to estimate the regression function is the Simplified Backfitting (SB), which consists of using only complete observations, in other words, those where $\delta_i = 1$. Thus the SB can be obtained as

$$\widehat{\mathbf{m}}_{SB,j} = \mathbf{W}_j^{\delta}\mathbf{Y}, \quad j = 1, \ldots, d, \tag{4}$$

where

$$\mathbf{W}_j^{\delta} = \mathbf{E}_j(\mathbf{M}^{\delta})^{-1}\mathbf{C}^{\delta}. \tag{5}$$

And the smoother matrices $\mathbf{S}_j^{\delta} = \left(\mathbf{s}_{j,X_{d1}}^{\delta}, \ldots, \mathbf{s}_{j,X_{dn}}^{\delta}\right)^T$, where $\mathbf{s}_{j,x_j}^{\delta}$:

$$\mathbf{s}_{j,x_j}^{\delta} = \mathbf{e}_1^T \left(\mathbf{X}_{j,x_j}^T \mathbf{K}_{x_j}^{\delta} \mathbf{X}_{j,x_j}\right)^{-1} \mathbf{X}_{j,x_j} \mathbf{K}_{x_j}^{\delta},$$

the matrix $\mathbf{K}_{x_j}^{\delta} = diag\{K_{h_j}(X_{j1} - x_j)\delta_1, \ldots, K_{h_j}(X_{jn} - x_j)\delta_n\}$.

Another option is the Imputed Backfitting (IB), which is constructed in two stages. In the first stage, the SB is used to estimate the missing

observations so as to complete the sample. In this way a completed sample, $\left\{(\mathbf{X}_i, \widehat{Y}_i),\ i = 1, \ldots, n\right\}$, is obtained where $\widehat{Y}_i = \delta_i Y_i + (1 - \delta_i)\widehat{\mathbf{m}}_{SB}(\mathbf{X}_i)$, with $\widehat{\mathbf{m}}_{SB}(\mathbf{X}_i) = \widehat{\mathbf{m}}_{SB,0} + \sum_{j=1}^{d} \widehat{\mathbf{m}}_{SB,j}(X_{ji})$ being the SB estimation of the additive function m, evaluated on X_i.

Once the sample is completed, Backfitting is applied to the data $\{(\mathbf{X}_i, \widehat{Y}_i),$ $i = 1, \ldots, n\}$, where $(\widehat{Y}_1, \ldots, \widehat{Y}_n)^t$ is the imputed response vector.

4 Simulations

We next present a simulation study in which both estimators (SB and IB) are compared using the mean squared error over 500 pseudosamples $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, with $n = 100$ and 500 from the model

$$Y_i = sin(X_{i1}) + sin(X_{i2}) + \varepsilon_i,$$

where ε_i are distributed from a $N(0, 0.1)$. The covariates were generated from an uniform in $[-\pi, \pi]^2$. The kernel function was the gaussian kernel $K(x) = (2\pi)^{(-1/2)} \exp\{-x^2/2\}$ and the missing data mechanism was $p(\mathbf{x}) = 1/(1 + \exp\{-x_1^2\})$. The bandwidths h_j used for these results were chosen using the plug-in selector defined by Opsomer and Ruppert (1998). We have used the software available in Matlab language for additive models given by Opsomer and Ruppert (1998), that had been adapted to the problem with missing observations.

Figures 1 and 2 and Table 1 summarize the results. Figures 1 and 2 depict the bias, the variance and the mean squared error curves of the Simplified Backfitting and the Imputed Backfitting estimates, which are based on 500 pseudosamples for sample sizes, $n = 100$ and $n = 500$, respectively. Table 1 shows the integrated squared biases, integrated variances and integrated mean squared errors. It is observed from Figures 1 and 2 that the bias property of the Imputed Backfitting is nearly the same as that of the Simplified Backfitting estimate in the interior and on the boundary. In the variance properties of the two estimates, the Imputed Backfitting estimate is seem to be slightly more stable. Because of this, the Imputed Backfitting estimate has a slightly improved mean integrated squared error property, as shown in Table 1.

References

AERTS, M., CLAESKENS, G., HENS, N. and MOLENBERGHS, G. (2002): Local multiple imputation. *Biometrika 89 (2), 375–388.*

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989): Linear smoothers and additive models (with discussion). *The Annals of Statistics 17, 453–555.*

CHEN, Q. and IBRAHIM, J.G. (2006): Semiparametric models for missing covariate and response data in regression models. *Biometrics 62, 177–184.*

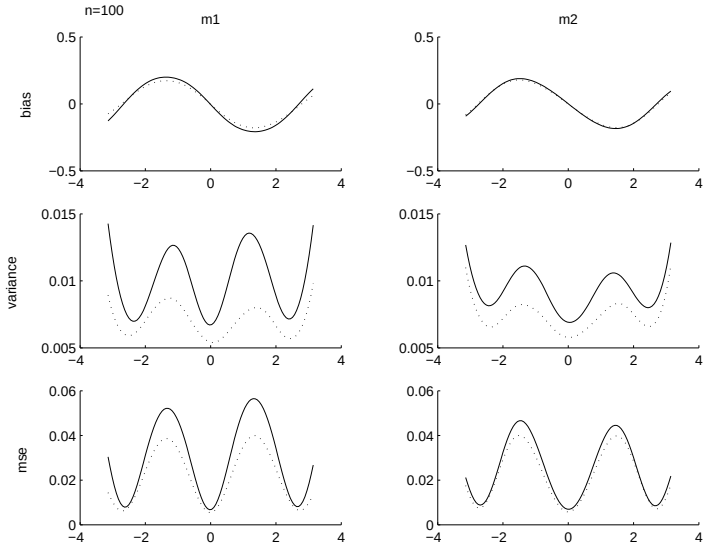


Fig. 1. Bias, variance and mean squared error curves. The solid curves correspond to the Simplified Backfitting, and the dashed curves are for the Imputed Backfitting. In each row, the left panel corresponds to m_1 and the right panel is for m_2. These are based on 500 samples of size 100.

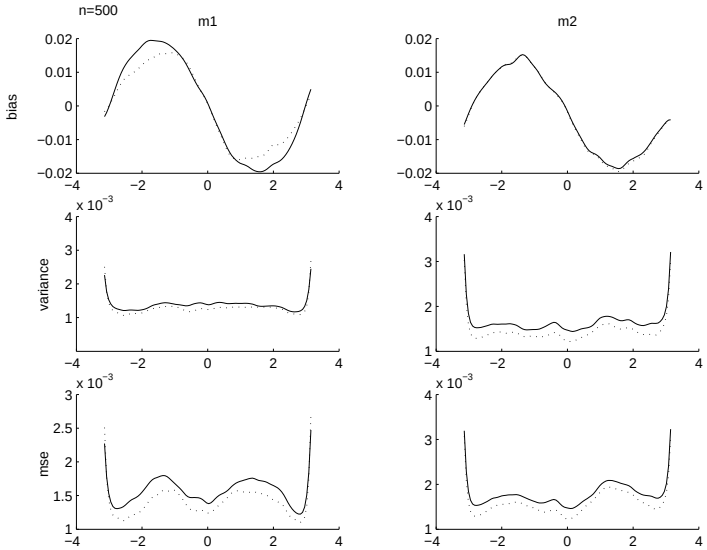


Fig. 2. Bias, variance and mean squared error curves. The solid curves correspond to the Simplified Backfitting, and the dashed curves are for the Imputed Backfitting. In each row, the left panel corresponds to m_1 and the right panel is for m_2. These are based on 500 samples of size 500.

Sample size	Target function	Estimate	Integrated sq. bias	Integrated variance	Integrated MSE
$n = 100$	m_1	$\widehat{m}_{1,SB}$	0.0194	0.0099	0.0294
		$\widehat{m}_{1,IB}$	0.0141	0.0070	0.0209
	m_2	$\widehat{m}_{2,SB}$	0.0153	0.0093	0.0246
		$\widehat{m}_{2,IB}$	0.0141	0.0074	0.0214
$n = 500$	m_1	$\widehat{m}_{1,SB}$	0.0019	0.0014	0.0016
		$\widehat{m}_{1,IB}$	0.0012	0.0013	0.0014
	m_2	$\widehat{m}_{2,SB}$	0.0013	0.0016	0.0018
		$\widehat{m}_{2,IB}$	0.0013	0.0015	0.0016

Table 1. Integrated squared bias, integrated variance and integrated mean squared error of the SB and the IB estimates based on 500 pseudosamples.

CHENG, P.E. (1990): Applications of kernel regression estimation: a survey. *Communications in Statistics, Theory and Methods 19 (11), 4103–4134.*

CHENG, P.E. (1994): Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association 89 (425), 81–87.*

CHENG, P.E. and WEI, L.J. (1986): Nonparametric inference under ignorable missing data process and treatment assignment. *International Statistical Symposium 1, Taipei, ROC, 97–112.*

CHU, C.K. and CHENG, P.E. (1995): Nonparametric regression estimation with missing data. *Journal of Statistical Planning and Inference 48, 85–99.*

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 39, 1–38.*

FUCHS, C. (1982): Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association 77, 270–278.*

GÓMEZ-GARCÍA, J., PALAREA-ALBALADEJO, J. and MARTIN-FERNÁNDEZ, J.A. (2006): Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones. *Estadística Espaola 48 (162), 241–270.*

GONZÁLEZ-MANTEIGA, W. and PÉREZ-GONZÁLEZ, A. (2004): Nonparametric mean estimation with missing data, *Communications in Statistics, Theory and Methods 33 (2), 277–303.*

HENS, H., AERTS, M. and MOLENBERGHS, G. (2006): Model selection form incomplete and design-based samples. *Statistics in Medicine 25, 2502–2520.*

HOROWITZ, J., KLEMELÄ, J. and MAMMEN, E. (2006): Optimal estimation in additive regression model. *Bernoulli 12, 271–298.*

IBRAHIM, J.G. (1990): Incomplete data in generalized linear models. *Journal of the American Statistical Association, Theory and Methods 85 (411), 765–770.*

IBRAHIM, J.G., LIPSITZ, S.R. and CHEN, M.-H. (1999): Missing covariates in generalised linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Series B 61 (1), 173–190.*

IBRAHIM, J.G., CHEN, M.-H. and LIPSITZ, S.R. (2001): Missing responses in generalised linear mixed models when the missing data mechanism is noningnorable. *Biometrika 88 (2), 551–564.*

LINTON, O. and NIELSEN, J.P. (1995): A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika 82, 93–100.*

LITTLE, R.J.A. (1992): Regression with missing X's: A review. *Journal of the American Statistical Association 87 (420), 1227–1237.*

MAMMEN, E., LINTON, O. and NIELSEN, J.P. (1999): The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics 27, 1443-1490.*

NIELSEN, S.F. (2001): Nonparametric conditional mean imputation, *Journal of Statistical Planning and Inference 99, 129–150.*

NITTNER, T. (2002): The additive model with missing values in the independent variable - Theory and simulation. Online http://epub.ub.uni-muenchen.de

OPSOMER, J.D. (2000): Asymptotic properties of backfitting estimators. *Journal of the Multivariate Analysis 73, 166-179.*

OPSOMER, J.D., and RUPPERT, D. (1998): A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association 93, 605-619.*

RUBIN, D.B. (1976): Inference and missing data. *Biometrika 63, 581–592.*

RUBIN, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys.* Willey & Sons.

TITTERINGTON, D.M. and MILL, G.M. (1983): Kernel-based density estimates from incomplete data. *Journal of the Royal Statistical Society, Serie B 45, 258–266.*

TITTERINGTON, D.M. and SEDRANSK, J. (1989): Imputation of missing values using density estimation. *Statistical Probability Letters 8, 411-418.*

TJϕSTHEIM, D. and AUESTAD, B.H. (1994): Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association 89, 1398–1409.*

VACH, W. (1994): Logistic regression with missing values and covariates. *Lecture Notes in Statistics 86, Springer-Verlag, Berlin.*

WANG, C.Y., WANG, S., ZHAO, L.-P. and OU, S.-T. (1997): Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association 92 (438), 512–525.*

WANG, C.Y., WANG, S., CARROL, R.J. and GUTIÉRREZ, R.G. (1998): Local linear regression for generalized linear models with missing data. *The Annals of Statistics 26 (3), 1028–1050.*

Nonparametric Test for Latent Root of Covariance Matrix in Multi-Population

Shin-ichi Tsukada[1] and Hidetoshi Murakami[2]

[1] Meisei University, School of Science and Engineering
2-1-1 Hodokubo, Hino-City, Tokyo 191-8506, Japan, *tsukada@ge.meisei-u.ac.jp*
[2] Chuo University, Department of Industrial and Systems Engineering
1-13-27 Kasuga, Bunkyo Ward, Tokyo 112-8551, Japan,
murakami@indsys.chuo-u.ac.jp

Abstract. We deal with a testing hypothesis for a latent root of covariance matrix in multi-population. Though the criterion for testing the hypothesis can be constructed by parametric procedure, large samples are necessary to keep a significance level. We may consider the nonparametric procedure using variance comparison. By simulation, we investigate the actual significance level and the power of the procedure using several tests for variance, and find that the bootstrap test is superior to our method.

Keywords: nonparametric test, latent root, covariance matrix

1 Introduction

We compare an actual significance levels and a power of nonparametric test for the hypothesis that the α-th largest latent root of covariance matrix is equivalent in multi-population. In principal component analysis (PCA), the α-th largest latent root of covariance matrix represents a contribution of the α-th principal component. Although there are many books on PCA, we have hardly seen the hypothesis that the latent root in multi-population is equivalent. In two populations, Sugiyama and Ushizawa (1998) propose a procedure applying Ansari-Bradley test which is testing the equivalence of variances, and simulate the accuracy.

There are several procedures to test the equivalence of variances. See Tsai *et al.* (1975), Hollander and Wolfe (1999), Good (2005) and Manly (2007). Conover *et al.* (1981) investigate the robustness of tests among 56 procedures and recommend three procedures. Tsukada (2006) shows that the procedures using the test recommended by Conover *et al.* (1981) are superior to the test applying Ansari-Bradley test in two populations. Murakami *et al.* (2007a) propose the test procedure applying Mood test in multi-populations and show that the test is superior to the test applying Ansari-Bradley test. We treat the testing hypothesis for latent root as the equivalence of variances in multi-population, and investigate the suitability for procedures. We compare the actual significance levels and the power of the above procedures and our

procedure which uses permutation test, and show that the procedure applying bootstrap test is superior by simulation.

2 Test procedure

Suppose that $\{\boldsymbol{x}_i^{(g)}; i = 1, \ldots, N_g\}(g = 1, \ldots, k)$ are random observations from p-variate population $\mathit{\Lambda}_p(\boldsymbol{\mu}_g, \mathit{\Sigma}_g)$ with mean $\boldsymbol{\mu}_g$ and covariance matrix $\mathit{\Sigma}_g$. Let $\lambda_\alpha^{(g)}$ and $\boldsymbol{\gamma}_\alpha^{(g)}$ be the α-th largest latent root of $\mathit{\Sigma}_g$ and the latent vector corresponding to $\lambda_\alpha^{(g)}$, respectively. We consider the following hypothesis:

$$H_0 : \lambda_\alpha^{(1)} = \lambda_\alpha^{(2)} = \cdots = \lambda_\alpha^{(k)} (= \lambda_\alpha),$$
$$H_1 : \text{not } H_0.$$

We calculate the sample mean $\bar{\boldsymbol{x}}_g$, the latent root $l_\alpha^{(g)}$ and the latent vector $\boldsymbol{h}_\alpha^{(g)}$ for the unbiased sample covariance matrix S_g. Let

$$Y_i^{(g)} = \boldsymbol{h}_\alpha^{(g)\prime}(\boldsymbol{x}_i^{(g)} - \bar{\boldsymbol{x}}_g) \equiv (y_{\alpha i}^{(g)} - \bar{y}_\alpha^{(g)}), \qquad (i = 1, \ldots, N_g).$$

Under the null hypothesis, the variance and the covariance of $Y^{(g)}$ are as follows:

$$\mathrm{Var}[Y^{(g)}] = \lambda_\alpha + O(N_g^{-1}), \qquad \mathrm{Cov}[Y^{(g_1)}, Y^{(g_2)}] = 0, (g_1 \neq g_2).$$

The scores $Y^{(g)}$ have asymptotically the same distribution and do not correlate each other.

Under the alternative hypothesis, the variance and the covariance are as follows:

$$\mathrm{Var}[Y^{(g)}] = \lambda_\alpha^{(g)} + O(N_g^{-1}), \qquad \mathrm{Cov}[Y^{(g_1)}, Y^{(g_2)}] = 0.$$

The scores $Y^{(g_1)}$ and $Y^{(g_2)}$ do not have asymptotically the same distribution and do not correlate each other. Therefore we deal with the above testing hypothesis as the equivalence for variance of $Y^{(g)}$. Sugiyama and Ushizawa (1998) adopt Ansari-Bradley test for the equivalence of variances in two populations. Murakami *et al.* (2007b) propose procedures applying permutation test using the ratio of variance. We improve the criterion as follows:

$$\sum_{g_1 < g_2}^{k} \frac{N r_{g_1} r_{g_2}}{2r} \left(\log l_\alpha^{(g_1)} - \log l_\alpha^{(g_2)} \right)^2, \tag{1}$$

where $r_g = N_g/N$, $r = \sum_{g=1}^k r_g$ and $N = \sum_{g=1}^k N_g$.

There are several tests for the equivalence of variances, that is, Mood test, test by chi-squared criterion using sample mean in Fligner *et al.* (1976), bootstrap test by Boos *et al.* (1989), randomization test by Wludyka and

Sa (2004), Levene test by Levene (1960), Brown-Forsythe test using 10% trimmed mean by Brown *et al.* (1974) and O'Brien's Test (1979, 1981). We compare the actual significance level and the power of test applying these tests.

Hayes (1997) indicates that randomization test of the F ratio of the sample variance is not valid when the population means differ. Since the mean of principal component scores $\boldsymbol{Y}_i^{(g)}$ are adjusted in this testing hypothesis, we also adopt the randomization test.

Nonparametric test requires the independence of each sample, but $y_{\alpha i}$ and $y_{\alpha j}$ are no longer independent. Now we evaluate the degree of dependence. We omit the suffix representing the population and let $E(\boldsymbol{x}_i) = 0$ without less of generality. When $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and λ_k is simple, the covariance of $y_{\alpha i}$ and $y_{\alpha j}$ is as follows:

$$
\begin{aligned}
E[y_{\alpha i} y_{\alpha j}] &= E\left[\sum_{u=1}^{p}\sum_{v=1}^{p} h_{u\alpha} h_{v\alpha} x_{ui} x_{vj}\right] = -\frac{2}{n^2}\sum_{l\neq\alpha}^{p} \lambda_{l\alpha}^2 \kappa_{\alpha l}^{21}\kappa_{\alpha l}^{21} \\
&-\frac{1}{n^2}\sum_{u\neq\alpha}^{p} \lambda_{u\alpha}^2(\kappa_{\alpha u}^{21}\kappa_{\alpha u}^{21} + \kappa_{\alpha}^{3}\kappa_{\alpha u}^{12}) + \frac{1}{n^2}\sum_{\substack{l,u\neq\alpha \\ u\neq l}}^{p} \lambda_{u\alpha}\lambda_{l\alpha}(\kappa_{ul}^{21}\kappa_{\alpha l}^{21} + \kappa_{ul\alpha}^{111}\kappa_{ul\alpha}^{111}) \\
&-\frac{1}{n^2}\sum_{v\neq\alpha}^{p} \lambda_{v\alpha}^2(\kappa_{\alpha}^{3}\kappa_{\alpha v}^{12} + \kappa_{\alpha v}^{21}\kappa_{\alpha v}^{21}) + \frac{1}{n^2}\sum_{\substack{v,l\neq\alpha \\ v\neq l}}^{p} \lambda_{v\alpha}\lambda_{l\alpha}(\kappa_{vl\alpha}^{111}\kappa_{vl\alpha}^{111} + \kappa_{vl}^{21}\kappa_{\alpha l}^{21}) \\
&+\frac{1}{n^2}\sum_{u,v\neq\alpha}^{p} \lambda_{u\alpha}\lambda_{v\alpha}(\kappa_{\alpha u}^{12}\kappa_{\alpha v}^{12} + \kappa_{\alpha uv}^{111}\kappa_{\alpha uv}^{111}) + O(n^{-3}) \qquad (2)
\end{aligned}
$$

where $n = N - 1$ and $\lambda_{\alpha\beta} = (\lambda_\alpha - \lambda_\beta)^{-1}$. The third moments denote $\kappa_i^3 = E(x_i x_i x_i)$, $\kappa_{ij}^{21} = E(x_i x_i x_j)$, $\kappa_{ij}^{12} = E(x_i x_j x_j)$ and $\kappa_{ijk}^{111} = E(x_i x_j x_k)$. Though we do not express terms of higher order for the above expansion, the expansion consists of odd-order moments. For a symmetric population,

$$E[y_{\alpha i} y_{\alpha j}] = 0,$$

the degree of dependence is very weak. This expansion shows that the degree of dependence is weak when the sample size is sufficiently large. Therefore, for large sample we may ignore the influence of dependence. But the degree of dependence is influenced by the third moments for an asymmetric population.

3 Simulation

3.1 Significance levels

We simulate the actual significance levels in three populations. We set $\alpha = 1$ and $\alpha = 2$, the sample size as $N_1 = N_2 = N_3 = 25, 50, 100$ and $N_1 =$

$150, N_2 = 100, N_3 = 50$. The number of simulation is a hundred thousand and the number of permutation or bootstrap is five thousand. As the population we select the multivariate normal distribution:

$$N(\mathbf{0}, \Sigma^{(g)})$$

the contaminated normal distribution:

$$0.05N(\mathbf{0}, 9\Sigma^{(g)}) + 0.95N(\mathbf{0}, \Sigma^{(g)}),$$

and the skew normal distribution:

$$SN(\mathbf{0}, \Omega^{(g)}, \alpha^{(g)}),$$

where $\Sigma^{(g)} = \mathrm{diag}(6.0, 3.0, 1.0)$, g=1, 2, 3, $\alpha^{(1)} = (-0.92, 1.84, 49.78)'/100$, $\alpha^{(2)} = (-2.05, 1.08, 48.17)'/100$, $\alpha^{(3)} = (-2.83, 0.69, 38.93)'/100$, $\Omega^{(1)} = \begin{pmatrix} 35.84 & 4.34 & 1.34 \\ 4.34 & 8.84 & 0.59 \\ 1.34 & 0.59 & 0.84 \end{pmatrix}$, $\Omega^{(2)} = \begin{pmatrix} 35.12 & 5.00 & 2.26 \\ 5.00 & 8.86 & 0.90 \\ 2.26 & 0.90 & 0.96 \end{pmatrix}$ and $\Omega^{(3)} = \begin{pmatrix} 33.04 & 5.26 & 3.45 \\ 5.26 & 8.85 & 1.37 \\ 3.45 & 1.37 & 1.37 \end{pmatrix}$.

Table 1 and Table 2 represent the actual significance levels for the largest latent root ($\alpha = 1$) and the second largest latent root ($\alpha = 2$), respectively. We indicate AB as Ansari-Bradley test, MO as Mood test, TM as test using criterion (1), FK as test by chi-squared criterion using sample mean in Fligner *et al.* (1976), BO as Bootstrap test by Boos *et al.* (1989), WS as randomization test by Wludyka and Sa (2004), LE as Levene's test, BF as Brown-Forsythe's test and OB as O'Brien's test in each Table. The significance level is 0.05.

For the largest latent root, all tests are conservative except Ansari-Bradley test and Mood test in the contaminated normal population and Bootstrap test in the skew normal population. Ansari-Bradley test and Mood test are liberal in large sample size for the contaminated normal population and Bootstrap test is also liberal for the skew normal population.

There is a similar tendency in the case that the sample sizes are unbalanced and the case that all sample sizes are a hundred. On the contaminated normal population, there is a difference for FK test, BF test and O'Brien's test. These may arise from the influence that the sample sizes are unbalanced.

For the second largest latent root, Ansari-Bradley test, Mood test and our test are liberal in the contaminated normal population and FK test, Levene's test, BF test and O'Brien's test are liberal in the case that the sample sizes are unbalanced. Bootstrap test is also liberal in all populations. As a whole, the actual significance levels for the second largest root have a similar tendency to those for the largest root, and are closer to 0.05 than those for the largest root.

Through the simulation, Ansari-Bradley test, Mood test and our test may be influenced by the kurtosis, and FK test, Levene's test, BF test and

Table 1. Actual Significance Levels ($\alpha = 1$, Significance Level 5%j.

Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.038	.033	.021	.031	.038	.018	.032	.031	.024
$N_1 = N_2 = N_3 = 50$	.043	.041	.028	.040	.047	.029	.041	.040	.035
$N_1 = N_2 = N_3 = 100$	.047	.046	.034	.046	.051	.037	.046	.042	.043
$N_1 = 150, N_2 = 100, N_3 = 50$	.046	.044	.041	.043	.049	.035	.044	.044	.043
Contaminated Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.050	.041	.024	.031	.039	.018	.032	.032	.024
$N_1 = N_2 = N_3 = 50$	.060	.059	.036	.040	.048	.030	.041	.040	.035
$N_1 = N_2 = N_3 = 100$	.062	.062	.043	.046	.051	.038	.046	.046	.042
$N_1 = 150, N_2 = 100, N_3 = 50$	.060	.060	.041	.055	.051	.036	.043	.041	.033
Skew Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.041	.034	.026	.035	.042	.019	.034	.033	.026
$N_1 = N_2 = N_3 = 50$	.045	.044	.039	.043	.050	.032	.043	.042	.037
$N_1 = N_2 = N_3 = 100$	.048	.048	.047	.048	.053	.040	.048	.047	.045
$N_1 = 150, N_2 = 100, N_3 = 50$	.048	.048	.044	.048	.056	.039	.048	.048	.049

O'Brien's test may be influenced by the fact that the sample sizes are unbalanced. Bootstrap test and WS test may be not influenced by the kurtosis and the unbalanced sample size. All tests are not influenced by the skewness.

3.2 Power

In this section, we investigate the power of several procedures. We set the alternative hypothesis as follows:

$$\Sigma^{(3)} = \mathrm{diag}\left(6.0 + \frac{70}{\sqrt{N_3}}, 3.0 + \frac{35}{\sqrt{N_3}}, 1.0 + \frac{20}{\sqrt{N_3}}\right),$$

$$\alpha^{(3)} = (-0.19, 0.37, 2.66)'/100, \quad \Omega^{(3)} = \begin{pmatrix} 33.04 & 5.26 & 3.45 \\ 5.26 & 8.85 & 1.37 \\ 3.45 & 1.37 & 1.37 \end{pmatrix},$$

$\Sigma^{(g)}$, $\alpha^{(g)}$ and $\Omega^{(g)}$ $(g = 1, 2)$ are same under the null hypothesis. Table 3 denotes the power of test for the largest latent root ($\alpha = 1$) and Table 4 does for the second largest latent root ($\alpha = 2$).

Table 2. Actual Significance Levels ($\alpha = 2$, Significance Level 5%j.

Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.048	.043	.036	.041	.052	.027	.042	.042	.032
$N_1 = N_2 = N_3 = 50$	.048	.047	.045	.046	.056	.037	.047	.047	.041
$N_1 = N_2 = N_3 = 100$	.049	.050	.048	.049	.054	.041	.050	.049	.046
$N_1 = 150, N_2 = 100, N_3 = 50$	.049	.050	.047	.048	.055	.037	.048	.048	.043
Contaminated Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.079	.080	.073	.041	.053	.027	.043	.042	.032
$N_1 = N_2 = N_3 = 50$	.082	.090	.084	.046	.055	.036	.047	.047	.041
$N_1 = N_2 = N_3 = 100$	.075	.081	.073	.049	.054	.041	.049	.049	.046
$N_1 = 150, N_2 = 100, N_3 = 50$	.076	.083	.049	.084	.056	.040	.078	.077	.054
Skew Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.049	.045	.039	.045	.055	.029	.045	.044	.034
$N_1 = N_2 = N_3 = 50$	.049	.048	.045	.049	.058	.037	.049	.048	.042
$N_1 = N_2 = N_3 = 100$	.051	.050	.050	.051	.057	.043	.050	.050	.047
$N_1 = 150, N_2 = 100, N_3 = 50$	.052	.052	.054	.052	.061	.042	.052	.051	.045

Since the power of test that the actual significance level is different can not be simply compared, we compare the power of test that the actual significance level is close.

For the largest latent root, we compare Ansari-Bradley test, Mood test, test by chi-squared criterion using sample mean, bootstrap test and Levene test. It is found that there is a tendency that AB $<$ MO $<$ TM given in Murakami *et al.* (2007b). The power of bootstrap test is superior as a whole. In the unbalanced sample sizes there is a similar tendency for the case that the balanced sample size is a hundred, but the power of FK test and Levene's test is small in the contaminated normal population.

For the second largest latent root, there are similar tendencies for the largest latent root. The power of O'Brien test is superior in the unbalanced sample sizes, but is inferior in the contaminated normal population. In the balanced case, the power of bootstrap test and O'Brien test are largest. There are similar tendencies for the largest latent root and the second largest latent root in the balanced case, but the powers of test for the second largest latent root are smaller than the powers of test for the largest latent root.

Table 3. Power of tests ($\alpha = 1$, Significance Level 5%j.

Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.630	.733	.804	.777	.894	.814	.828	.827	.841
$N_1 = N_2 = N_3 = 50$	.763	.860	.934	.900	.957	.936	.924	.923	.949
$N_1 = N_2 = N_3 = 100$	.866	.935	.979	.962	.984	.979	.970	.970	.984
$N_1 = 150, N_2 = 100, N_3 = 50$	.855	.932	.980	.967	.988	.984	.977	.977	.992
Contaminated Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.579	.655	.820	.776	.894	.814	.827	.826	.840
$N_1 = N_2 = N_3 = 50$	.716	.799	.885	.900	.957	.937	.924	.923	.949
$N_1 = N_2 = N_3 = 100$	.830	.896	.957	.962	.984	.980	.970	.970	.984
$N_1 = 150, N_2 = 100, N_3 = 50$	.803	.880	.980	.904	.988	.984	.905	.903	.757
Skew Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.642	.743	.809	.790	.897	.820	.834	.832	.844
$N_1 = N_2 = N_3 = 50$	.780	.874	.940	.911	.961	.942	.932	.931	.954
$N_1 = N_2 = N_3 = 100$	.884	.946	.983	.969	.988	.983	.976	.976	.988
$N_1 = 150, N_2 = 100, N_3 = 50$	.866	.939	.981	.971	.989	.985	.980	.980	.993

4 Conclusions

We compare the actual significance level and the power of tests for latent root by simulation. The test proposed by us is superior compared with the rank test as Ansari-Bradley test and Mood test. But the test applying bootstrap method is superior to our method in the alternative hypothesis which we adopt this time. In the case that the sample size is unbalanced, the procedure applying O'Brien test is also superior. The actual significance level and the power of Levene test and Brown-Forsythe's test have a similar tendency as a natural result. As a whole, the test applying bootstrap method is superior in the point of the actual significance level and the power.

We need to investigate the actual significance level and the power under other alternative hypothesis.

References

BOOS, D.D. and BROWNIE, C. (1989): Bootstrap methods for testing homogeneity of variances. *Technometrics 31 (1), 69-82.*

Table 4. Power of tests ($\alpha = 2$, Significance Level 5%j.

Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.649	.750	.820	.791	.902	.830	.841	.840	.853
$N_1 = N_2 = N_3 = 50$	.770	.865	.934	.903	.958	.936	.926	.925	.950
$N_1 = N_2 = N_3 = 100$	.868	.936	.980	.962	.985	.980	.970	.970	.984
$N_1 = 150, N_2 = 100, N_3 = 50$	.833	.914	.972	.956	.983	.977	.968	.968	.987
Contaminated Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.634	.719	.740	.791	.902	.828	.840	.839	.852
$N_1 = N_2 = N_3 = 50$	.735	.819	.890	.903	.959	.937	.926	.925	.950
$N_1 = N_2 = N_3 = 100$	.833	.899	.957	.962	.984	.980	.970	.970	.984
$N_1 = 150, N_2 = 100, N_3 = 50$	.810	.887	.970	.909	.982	.976	.910	.909	.808
Skew Normal population									
Sample size	AB	MO	TM	FK	BO	WS	LE	BF	OB
$N_1 = N_2 = N_3 = 25$	.633	.732	.799	.779	.887	.807	.823	.822	.836
$N_1 = N_2 = N_3 = 50$	.750	.849	.923	.891	.949	.924	.914	.913	.939
$N_1 = N_2 = N_3 = 100$	.848	.922	.973	.953	.979	.973	.962	.962	.979
$N_1 = 150, N_2 = 100, N_3 = 50$	.815	.903	.964	.947	.976	.970	.960	.961	.983

BROWN, M.B. and FORSYTHE, A.B. (1974): The small sample behavior of some statistics which test the equality of several means. *Technometrics 16, 129-132.*

CONOVER, W.J., JOHNSON, M.E. and JOHNSON, M.M. (1981): A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics 23 (4), 351-361.*

FLIGNER, M.A. and KILLEEN, T.J. (1976): Distribution-Free Two-Sample Tests for Scale. *Journal of the American Statistical Association 71, 210-213.*

GOOD, P.I. (2005): *Permutation, Parametric, and Bootstrap Tests of Hypotheses.* 3rd ed., Springer-Verlag, New York.

HAYES, A.F. (1997): Cautions in testing variance equality with randomization tests. *J. Statist. Comput. Simul. 59, 25-31.*

HOLLANDER, M. and WOLFE, D.A. (1999): *Nonparametric Statistical methods.* 2nd ed., John Wiley & Sons, New York.

MANLY, B.F.J. (2007): *Randomization, Bootstrap and Monte Carlo Methods in Biology.* 3rd ed., Chapman & Hall, London.

LEVENE, H. (1960): Robust Tests for Equality of Variances. I. Olkin, Palo Alto ed. *in Contributions to Probability and Statistics.* CA: Stanford Univ. Press.

MURAKAMI, H., HINO, E. and TSUKADA, S. (2007a): Nonparametric test for eigenvalues of covariance matrix in multipopulation. *J. Japan Statist. Soc. 37, 299-306.*

MURAKAMI, H., TSUKADA, S. and TAKEDA, Y. (2007b): A new statistic for testing the equality of eigenvalue of covariance matrix on multipopulation. *J. Jpn. Comp. Statist.* (Submit).

O'BRIEN, R.G. (1979): A general ANOVA method for robust test of additive models for variance. *Journal of the American Statistical Association 74, 877-880.*

O'BRIEN, R.G. (1981): A simple test for variance effects in experimental designs. *Psychological Bulletin 89, 570-574.*

SUGIURA, N. (1976): Asymptotic expansions of the distributions of the latent roots and the latent vector of the Wishart and multivariate F matrices. *J. Multivariate Anal. 6, 500-525.*

SUGIYAMA, T. and USHIZAWA, K. (1998): A non-parametric method to test equality of intermediate latent roots of two populations in a principal component analysis. *Journal of Japan Statistical Society 28 (2), 227-235.*

TSAI, W.S., DURAN, B.S. and LEWIS, T.O. (1975): Small-sample behavior of some multisample nonparametric tests for scale. *Journal of the American Statistical Association 70, 791-796.*

TSUKADA, S. (2006): Power Comparison of Nonparametric Test for Latent Root of Covariance Matrix in Two Populations. Rizzi, A. and Vichi, M. Eds. *Compstat2006: Proceedings in Computational Statistics.* Physica-Verlag/Springer, 1713–1720.

WLUDYKA, P.S. and SA, P. (2004): Robust I-Sample Analysis of Means Type Randomization Tests for Variances for Unbalanced Designs. *Journal of Statistical Computation and Simulation 74, 701-726.*

Part XV

Optimization and Random Search Algorithms

A Random Decision Model for Reproducing Heavy-Tailed Algorithmic Behavior

Alda Carvalho[1,2], Nuno Crato[2], and Carla Gomes[3]

[1] Instituto Superior de Engenharia de Lisboa
Rua Conselheiro Emídio Navarro 1, 1959-007 Lisboa, Portugal,
acarvalho@dem.isel.ipl.pt
[2] Cemapre, Instituto Superior de Economia e Gestão
Rua Miguel Lupi 20, 1200 Lisboa, Portugal, *ncrato@iseg.utl.pt*
[3] Department of Computer Science
Cornell University, Ithaca, NY 14850, U.S.A., *gomes@cs.cornell.edu*

Abstract. Some random search algorithms display heavy-tailed distributions. The reason for the appearance of these statistical characteristics is still not well understood and researchers have made efforts to create models that could help understanding the phenomenon. In this paper, we provide a tree search model that follows very simple rules and is capable of displaying a similar behavior. The decisions are equiprobable at each node and so this model mimics a key characteristic of real random algorithms.

Keywords: heavy tails, decision trees, random search

1 Introduction

During the last years there has been much interest in the statistical study of random algorithm's computing costs (Gomes et al. (2005)). Such costs can be measured as the number of backtracks performed until a search is terminated or by any other measure independent of the machine's performance. A search is considered to end when a solution is found or when the algorithm proves that there is no solution.

Some of these statistical studies exhibited search times with heavy-tailed distributions, i.e., with a high probability of finding extremely long costs. Such heavy-tailed phenomena were first identified in combinatorial search, specifically, in a study of Latin square completion problems (Gomes et al. (1997), Gomes et al. (2000)). A Latin square is a $N \times N$ table of N symbols in which each symbol occurs once in each row and once in each column. Latin squares translate what is known as quasigroup completion problems, in particular, what in computer science is better known as constraint satisfaction problems (CSP). These problems were introduced as a benchmark for evaluating combinatorial search methods, since their structure is similar to the ones found in real-world problems such as scheduling, timetabling, routing and designing of statistical experiments.

In general, CSP problems exhibit an easy-hard-easy pattern of search costs, depending on the constrainedness of the problem (Hogg et al. (1996)). Given a Latin square of order N, the constrainedness is measured by the number of assigned cells. If the table is either almost empty or almost full, the problem is usually very easy to solve. If the table is filled at some intermediate levels, the problem may be very hard to solve.

Gomes et al. (1997) studied the runtime distributions of randomized backtrack search algorithms and clarified that the source of extreme variance observed in exceptional hard instances was not due to the inherent hardness of the instances. They showed that the runtime distributions of random search methods quite often exhibit heavy tails. The understanding of the heavy-tailed nature of the distributions underlying such search methods has led to the design of more efficient backtrack search techniques (Gomes (2003)).

Heavy-tailed distributions were first introduced by Pareto in the context of income distributions and were extensively studied by Lévy, when investigating additive stable laws, established that infinite variance stable laws have Paretian tails. Until Mandelbrot's work on fractals, Paretian heavy tail distributions were often considered pathological cases. Since then heavy-tailed distributions have been used to model phenomena in areas as diverse as economics, physics, geophysics, biology (a review of the literature can be seen in Mitzenmacher (2004)) and more recently in computer science (Adler et al. (1998)).

So far, evidence for heavy-tailed behavior of randomize backtrack search procedures on concrete instance models has been largely empirical. Moreover, it is clear that not all problem instances exhibit heavy tails. Gomes et al. (2005) observe two regions with dramatically different statistical regimes of the runtime distributions and provide a better characterization of when this heavy-tailed behavior occurs and when it does not.

On Figure 1 we may observe these two regimes. Each line corresponds to a simulation of the computing costs of a particular constrained instance, for a given search problem. The log-log plot is particularly helpful for distinguishing heavy-tailed from non-heavy-tailed behavior. The cumulative density function (CDF) grows rapidly for normal distribution tails and grows slower, in a sense made precise below, for a distribution with heavy tails. This behavior is better observed with the complementary cumulative density function (CCDF = 1 – CDF). In the log-log plot, heavy tails appear as straight lines, and non heavy tails appear as curved lines, displaying a faster decay.

For the particular CSP model in the figure, each case corresponds to a different constraining parameter, p, which sets the percentage of assigned cells in the Latin square. It is possible to see two dramatically different statistical regimes of the runtime distributions. In the first regime ($p \leq 0.07$), we see heavy-tailed behavior: instances are easy to solve, but in some runs the search method explore large subtrees with no solution. Increasing the constrained-

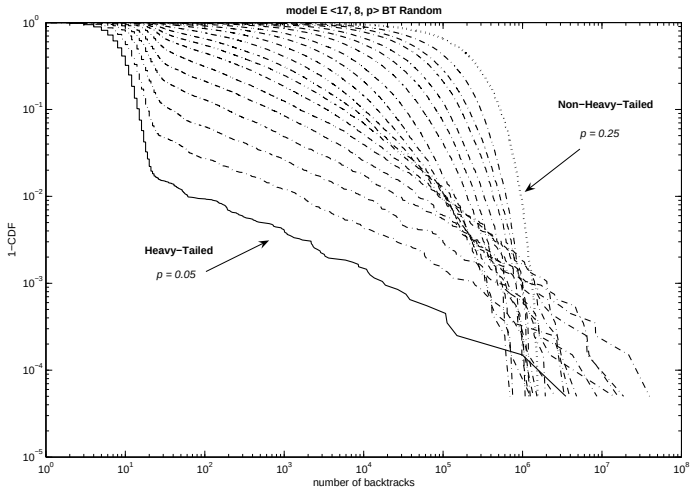


Fig. 1. The progression from heavy-tailed regime to a non-heavy-tailed regime. CDF stands for cumulative density function (Gomes et al. (2005)).

ness of the model ($p \uparrow$), a different statistical regime appears, in which the heavy-tails disappear. In this latter regime, the instances become inherently hard and all runs become homogeneously long—variance decreases and tails decay exponentially.

The reason why these long tails appear is not apparent. Efficient random algorithms are intrinsically complex and it is hard to visualize a behavior that generates extremely long runs. So far, no one has been able to explain such mechanisms. In order to get an insight on the problem, we decided to explore simpler models that replicate the essential behavior of search algorithms.

As a model for the random search algorithms we set up a random tree. Solutions are fixed nodes distributed on the tree. The search consists on descending the tree according to a random path, until a solution is found. Each descending choice leads to a solution with probability 1. The random variable, X, is the total number of visited tree nodes until a solution is found. Most search trees display an exponential decay, which is the standard non heavy-tailed behavior. We need to build special structures in order to find a heavy-tailed decay for this random variable.

2 Trees that generate exponential decay

Many standard probability distributions have exponentially decreasing tails,

$$P(X > x) \sim Ce^{-x} \ , \ C > 0. \tag{1}$$

This is the case of the exponential, the normal and other common distributions, in which extreme values are very rare.

We start by considering the simplest possible model, which is an infinite binary tree with one solution at each level. The search proceeds by descending the tree and randomly choosing at each level the edge to descend through next (Figure 2). The search stops when a solution is found. In this tree, the probability that the search continues after n choices is 2^{-n} and so

$$P(X > n) = 2^{-n} \to 0 \text{ as } n \to \infty.$$

The whole tail of the distribution is exponential (geometric).

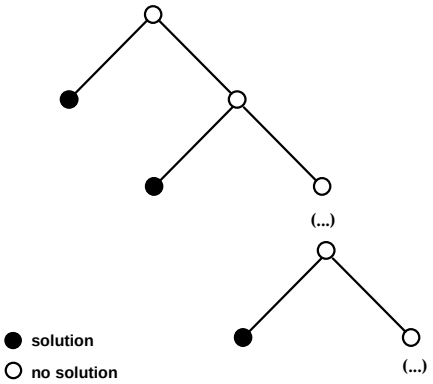


Fig. 2. Binary tree.

This tree can be easily generalized to a tree with branching factor k and a fixed number s ($s < k$) of solutions below each node. For such a tree, $P(X > n) = (\frac{k-s}{k})^n$ Again, we have an exponentially decaying tail.

It is not difficult to generate trees that display heavy tails provided we relax the assumption of equiprobable nodes. Such a construction is provided in Carvalho et al. (2006). What is not so easy to generate is a equiprobable nodes tree with simple rules for both the branching factor and the number of solutions.

In order to have a tree that exhibits such hyperbolic tail decay, the probability of continuing the search must increase slowly as we descend the tree. In the following sections we present a construction of a class of trees with this property. As we will see, the described class of trees provides an appropriate model for CSP algorithms.

3 Trees that generate heavy tails

Heavy-tailed distributions have tails that decay like a power function,

$$P(X > x) \sim Cx^{-\alpha}\ ,\ x > 0, \tag{2}$$

where $0 < \alpha \leq 2$ and $C > 0$ are constants. One or both tails of these distributions have hyperbolic decay. Without loss of generality, we will discuss the right tail behavior and assume that the distribution has support on the positive half line only.

The α in the equation (2) is called the tail index or index of stability of the distribution. The lower the index, the heavier the tail is. The existence of moments depends on the α parameter; for $\alpha < 2$, moments of X of order less than α are finite while all higher order moments are infinite, i.e., $\alpha = sup\{b > 0 : E|X|^b < \infty\}$. So, when $1 < \alpha < 2$, the distribution has finite mean but infinite variance. With $\alpha \leq 1$, the distribution has infinite mean and variance. The CCDF log-log plot of a distribution with heavy tails exhibits linear behavior with slope $-\alpha$, as we can see in (2).

We now develop two examples of trees with heavy-tailed distributions. For both examples, we set up a tree with the number of edges increasing with the level.

On the first example, each node at level n originates $n + 1$ edges. Each corresponding descending edge is chosen with equal probability. From each node, there is only a solution below.

The tree starts with two edges with equal probability; one is a solution and the other not. So, in the second level there is one more edge, three choices, and one solution. The pattern continues, i.e., at each level an additional edge is added and below each node only a solution exists (Figure 3, left).

At each level n, the probability that a solution is found and then the search stops is

$$P(X = n) = \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \ldots \times \frac{n-2}{n-1} \times \frac{1}{n} = \frac{1}{n-1} \times \frac{1}{n},$$

and

$$P(X > n) = \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \ldots \times \frac{n-2}{n-1} \times \frac{n-1}{n} = \frac{1}{n},$$

which means that the CCDF has a hyperbolic decay with $\alpha = 1$.

On the second example, we considerer a modification of this tree: instead of increasing one edge in each level, we add two edges (Figure 3, right). Then,

$$P(X > n) = \frac{1}{2} \times \frac{3}{4} \times \frac{5}{6} \times \ldots \times \frac{2n-3}{2n-2} \times \frac{2n-1}{2n}. \tag{3}$$

If we square (3),

$$[P(X > n)]^2 = \frac{1}{2} \times \frac{1}{2} \times \frac{3}{4} \times \frac{3}{4} \times \frac{5}{6} \times \frac{5}{6} \times \ldots \times \frac{2n-3}{2n-2} \times \frac{2n-3}{2n-2} \times \frac{2n-1}{2n} \times \frac{2n-1}{2n},$$

and use the well known John Wallis product formula for π,

$$\frac{2}{1} \times \frac{2}{3} \times \frac{4}{3} \times \frac{4}{5} \times \ldots = \frac{\pi}{2},$$

we conclude that $2n[P(X > n)]^2 \rightarrow \frac{2}{\pi}$. So,

$$P(X > n) \sim \frac{1}{\sqrt{\pi}} n^{-\frac{1}{2}},$$

which means that the CCDF of this tree has a hyperbolic decay with $\alpha = \frac{1}{2}$. This is a very interesting result. This particular tree could even be useful to generate randomly the value of π.

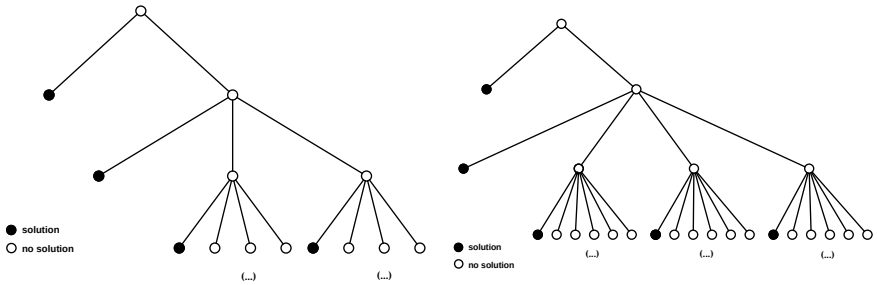


Fig. 3. Search tree with heavy-tail: $\alpha = 1$ (left) and $\alpha = \frac{1}{2}$ (right).

4 Generalization for any rational index

Now we will propose a construction for a random search tree with hyperbolic decay with index of stability $\alpha \in \mathbb{Q}$. For the given construction, we will show that $P(X > n) \sim Cn^{-\alpha}$, with $\alpha = \frac{p}{q}, p, q \in \mathbb{N}$.

Construction of the tree related with $\alpha = \frac{p}{q}$:

In the first level the tree has k edges and p solutions ($p < k$). In a search algorithm there are $k - p$ hypothesis to continue to the second level. In the following levels, for each no solution node, we add q edges and keep p solutions: the second level has $k + q$ edges and p solutions, the third level has $k + 2q$ edges with p solutions and so on... As an example we can see the three first levels on Figure 4 ($k = 3, p = 2, q = 3 \rightarrow \alpha = \frac{2}{3}$).

Proposition 2. *For $k, p, q \in \mathbb{N}$ and the tree constructed above, we have*

$$P(X > n) \sim Cn^{-\frac{p}{q}},$$

where $C = \frac{\Gamma(\frac{k}{q})}{\Gamma(\frac{k-p}{q})}$, and $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, x > 0$ is the Gamma function.

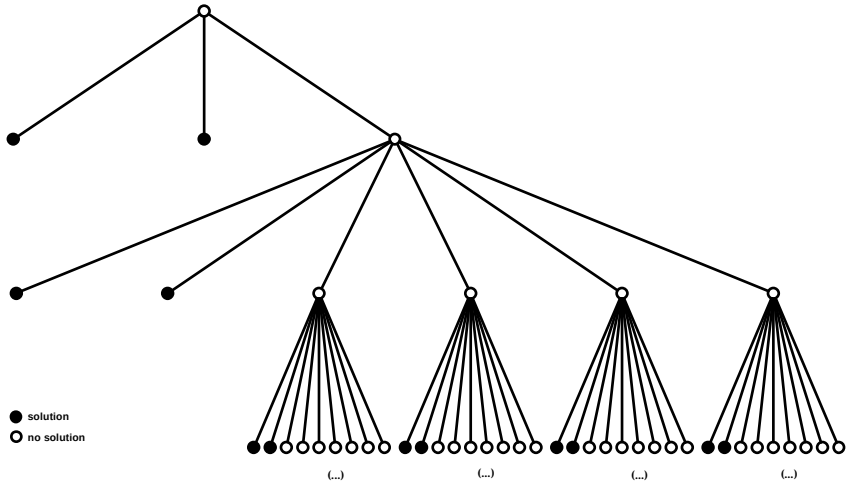


Fig. 4. Search tree with heavy-tail with $\alpha = \frac{2}{3}$.

Proof. We have

$$P(X > n) = \frac{k-p}{k} \times \frac{k-p+q}{k+q} \times \frac{k-p+2q}{k+2q} \times \ldots \times \frac{k-p+(n-1)q}{k+(n-1)q} = \\ = \frac{(k-p) \times (k-p+q) \times (k-p+2q) \times \ldots \times (k-p+(n-1)q)}{k \times (k+q) \times (k+2q) \times \ldots \times (k+(n-1)q)},$$

and then

$$P(X > n) = \frac{q^n \times \frac{\Gamma\left(n+\frac{k-p}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)}}{q^n \times \frac{\Gamma\left(n+\frac{k}{q}\right)}{\Gamma\left(\frac{k}{q}\right)}} = \frac{\Gamma\left(\frac{k}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)} \times \frac{\Gamma\left(n+\frac{k-p}{q}\right)}{\Gamma\left(n+\frac{k}{q}\right)},$$

where we used the fact $\Gamma(z+n) = (n-1+z) \times \ldots \times z \times \Gamma(z)$.

Now, to prove that $P(X > n) \sim n^{-\frac{p}{q}}$, we can calculate

$$\lim_{n \to \infty} P(X > n) \times n^{\frac{p}{q}}.$$

So,

$$P(X > n) \times n^{\frac{p}{q}} = \frac{\Gamma\left(\frac{k}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)} \times \frac{n^{\frac{p}{q}} \times \Gamma\left(n+\frac{k-p}{q}\right)}{\Gamma\left(n+\frac{k}{q}\right)}.$$

We can observe that $\frac{n^{\frac{p}{q}}\Gamma\left(n+\frac{k-p}{q}\right)}{\Gamma\left(n+\frac{k}{q}\right)} \to 1$, because

$$\lim_{n \to \infty} \frac{\Gamma(n+a)}{\Gamma(n+b)} n^{b-a} = 1.$$

So, we can conclude

$$P(X > n) \times n^{\frac{p}{q}} \rightarrow \frac{\Gamma\left(\frac{k}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)}$$

and

$$P(X > n) \sim \frac{\Gamma\left(\frac{k}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)} n^{-\frac{p}{q}}$$

□

Remarks

(i) If $k = 2, q = 1$ and $p = 1$, we have the $\alpha = 1$ tree mentioned above. In this case $C = \frac{\Gamma(\frac{k}{q})}{\Gamma(\frac{k-p}{q})} = \frac{\Gamma(\frac{2}{1})}{\Gamma(\frac{2-1}{1})} = \frac{\Gamma(2)}{\Gamma(1)} = 1$.

(ii) If $k = 2, q = 2$ and $p = 1$, we have the Wallis tree mentioned above. In this case $C = \frac{\Gamma(\frac{k}{q})}{\Gamma(\frac{k-p}{q})} = \frac{\Gamma(\frac{2}{2})}{\Gamma(\frac{2-1}{2})} = \frac{\Gamma(1)}{\Gamma(\frac{1}{2})} = \frac{1}{\sqrt{\pi}}$.

References

ADLER, R., FELDMAN, R. and TAQQU, M. (1998): *A Practical guide to heavy tails*. Birkauser.

CARVALHO, A., CRATO, N. and GOMES, C. (2006): A Generative Power-Law Search Tree Model. In: *CORS/Optimization Days Joint Conference*.

GOMES, C. (2003): Complete Randomized Backtrack Search. *Constraint and Integer Programming: Toward a Unified Methodology, Milano, M., (ed.), Kluwer* 233-283.

GOMES, C., FERNANDEZ, C., SELMAN, B. and BESSIERE, C., (2005): Statistical Regimes Across Constrainedness Regions. *Constraints*. 10(4):317-337.

GOMES, C. and SELMAN, B. and CRATO, N. (1997): Heavy-Tailed Phenomena in Combinatorial Search. In: *Proceedings CP'97, Linz, Austria*, 121-135.

GOMES, C., SELMAN, B., CRATO, N. and KAUTZ,H. (2000): Heavy-Tailed Phenomena in Satisfiability and Constraint Satisfaction Problems. *J. Automated Reasoning*, 24: 67-100.

HOGG, T., HUBERMAN, B. and WILLIAMS, C. (2004): Phase Transition and Search Problems. *Artificial Intelligence* 81 (1-2), 1-15.

MITZENMACHER, M. (2004): A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, Vol. I, No. 2, 226-251.

An Evolutionary Algorithm for LTS–Regression: A Comparative Study

Oliver Morell[1], Thorsten Bernholt[2], Roland Fried[1], Joachim Kunert[1], and Robin Nunkesser[2]

[1] Fakultät Statistik, Technische Universität Dortmund
Vogelpothsweg 87, 44227 Dortmund, Germany,
morell@statistik.uni-dortmund.de
[2] Fakultät Informatik, Technische Universität Dortmund
Otto-Hahn-Str. 14, 44227 Dortmund, Germany

Abstract. Least Trimmed Squares (LTS) regression is one of the most popular highly robust regression techniques. As the exact computation of LTS is very demanding, particularly in the case of a large number of regressors, search heuristics are commonly applied. Here a new evolutionary algorithm for LTS is introduced and compared to the popular Fast–LTS procedure using designed experiments. It turns out that the evolutionary algorithm requires a higher computation time, but can deliver considerably better solutions in challenging data situations.

Keywords: evolutionary algorithm, search heuristics, robust regression, design of experiments

1 Introduction

Least Trimmed Squares (LTS) regression (Rousseeuw and Leroy, 1987) is a frequently applied robust regression technique. Its popularity is due to the fact that it can be designed to have optimal breakdown point with higher efficiency as measured by the variance than e.g. the Least Median of Squares and better maxbias behavior than other estimators with high breakdown point. LTS is defined by the non-convex minimization problem of finding a subset of predetermined size for which the least squares distance to a regression hyperplane is minimized. However, since it is an NP–hard problem (Bernholt, 2005), the exact computation of LTS is very demanding, particularly in the case of a large number of explanatory variables. There have been several attempts to use heuristic algorithms like tabu search (Woodruff and Rocke, 1994), genetic algorithms (Chakraborty and Chaudhuri, 2003) and simulated annealing (Todorov, 1992) for solving such computational problems in robust statistics. Such heuristics search for homogeneous subsets instead of considering all subsets of a given size. Nowadays, the Fast–LTS procedure of Rousseeuw and van Driessen (2006) is commonly used.
We propose a new evolutionary algorithm (EA) to find homogeneous subsets, introducing a new problem specific mutation operator. Similar as Fast–LTS,

this algorithm needs the proper choice of several design parameters. The EA is compared to the Fast–LTS algorithm implemented in R (2007). Performance is measured by the value of the LTS–criterion and the computation time, which are jointly optimized by desirability indices. We search optimal parameter settings for the algorithms using factorial designs and response surface methods.

2 Algorithms for LTS–regression

2.1 Least Trimmed Squares regression

We consider the linear regression model

$$Y_i \;=\; \beta_0 + \sum_{j=1}^{p-1} x_{i,j}\beta_j \;+\; E_i \;, \quad i = 1, \ldots, n \quad . \tag{1}$$

Y_i is an observable random variable and $x_{i,j}$ is the value of the deterministic regressor $x_{\bullet j}$ for individual i, $j = 1, \ldots, p-1$. The unknown parameters are the intercept β_0 and the slope parameters $\beta_1, \ldots, \beta_{p-1}$, which describe the effect of $x_{i,1}, \ldots, x_{i,(p-1)}$ on Y_i. $E_1, \ldots, E_n$ are random disturbances. Define $Y = (Y_1, \ldots, Y_n)'$ and the design matrix $\mathbf{X}$ to contain the element $x_{i,j-1}$ at position (i, j) with $x_{i,0} = 1$. Model (1) reads in matrix notation $Y = \mathbf{X}\beta + E$. The parameter vector $\beta \in \mathbb{R}^p$ comprises β_0 and $\beta_1, \ldots, \beta_{p-1}$, while E includes the random disturbances $E_1, \ldots, E_n$. It is well known that the best unbiased estimator of β is the Ordinary Least Squares (OLS) estimator $\widehat{\beta}_{\mathrm{OLS}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Y$, if $E_1, \ldots, E_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. If outliers occur robust regression methods like the Least Trimmed Squares (LTS) estimator can lead to much better results than OLS. The LTS–estimator of β is defined as $\widehat{\beta}_{\mathrm{LTS}} = \underset{\widehat{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{h} \widehat{E}_{(i)}^{\,2}$ where $\widehat{E}_{(1)}^{\,2}, \ldots, \widehat{E}_{(n)}^{\,2}$ are the order statistics of the squared residuals $\widehat{E}_1^{\,2}, \ldots, \widehat{E}_n^{\,2}$. This means that the sum of the h smallest squared residuals is minimized, with a properly chosen $h > n/2$. We denote this LTS–criterion by Q. For $h = n$ LTS and OLS coincide, while LTS achieves the optimal finite sample breakdown point

$$\epsilon_n\left(\widehat{\beta}_{LTS}, Z\right) = \frac{\left(\lfloor \frac{n-p}{2} \rfloor + 1\right)}{n} \quad \text{for} \quad h = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{p+1}{2} \right\rfloor \tag{2}$$

(Rousseeuw and Leroy, 1987). Exact minimization of the LTS–criterion is NP–hard (Bernholt, 2005), meaning that for larger p exact computation of LTS is very demanding. Search heuristics are commonly used therefore.

2.2 Fast–LTS and PROGRESS

The Fast–LTS algorithm (Rousseeuw and van Driessen, 2006) `ltsReg` implemented in the `R`–package `robustbase` (2007) works as follows:
For a given sample of data Z with

$$Z \;=\; \big(z_1,\;\ldots,\;z_n\big) \text{ and } z_i \;=\; \big(y_i,\;x_{i,1},\;\ldots,\;x_{i,(p-1)}\big)\,,\; i=1,\ldots,n \quad , \tag{3}$$

a random subset J_0 of size p is drawn from the indexset I_Z of Z, $J_0 \subset I_Z = \{1,2,\ldots,n\}$. With a larger size than p, e.g. h as in (2), there is a higher risk of getting an outlier in J_0, causing a bad initialization of the iterations. If the observations in J_0 do not span a $(p-1)$–dimensional hyperplane, additional data points are drawn until the OLS–estimator $\widehat{\beta}^{\,0}$ becomes unique. The resulting h smallest squared residuals $\widehat{e}^{\,2}_{(1)},\;\ldots,\;\widehat{e}^{\,2}_{(h)}$ of $\widehat{e}_1^{\,2},\ldots,\widehat{e}_n^{\,2}$ establish a new index subset J_1 with

$$J_1 \;=\; \Big\{ i \;\in\; I_Z : \;\; \widehat{e}_i^{\,2} \;\in\; \Big\{ \widehat{e}^{\,2}_{(1)},\;\ldots,\;\widehat{e}^{\,2}_{(h)} \Big\} \;\Big\} \quad , \tag{4}$$

and an approximation $\widehat{Q}_1$ of the true LTS–criterion Q, with $\widehat{Q}_1 = \sum\limits_{i\in J_1} \widehat{e}_i^{\,2}$.
Using the data points in J_1, a new OLS–estimation $\widehat{\beta}^{\,1}$ is obtained, which delivers a new index subset J_2 corresponding to the h smallest squared residuals and a new LTS–criterion value $\widehat{Q}_2$. These steps are iterated until the LTS–criterion no longer changes. The sequence $\widehat{Q}_\nu$, $\nu \in \mathbb{N}$, of LTS–values necessarily converges because of its monotonicity and the finite number of index subsets.
One choice in the Fast–LTS procedure is the intercept adjustment (IA). When IA is used, the LTS–hyperplane of the initial estimation is vertically shifted, which means that a new estimate of the intercept

$$\widetilde{\beta}_0^{\,\nu} \;=\; \underset{\widehat{\beta}_0^{\,\nu}\;\in\;\mathbb{R}}{\arg\min} \sum_{i=1}^{h} \big(y_i \;-\; \widehat{\beta}_0^{\,\nu} \;-\; \sum_{j=1}^{p-1} x_{ij}\widehat{\beta}_j^{\,\nu}\;\big)^2_{(i)} \tag{5}$$

is obtained, keeping the current estimates $\widehat{\beta}_1^{\,\nu},\ldots,\widehat{\beta}_{p-1}^{\,\nu}$. IA often reduces the value of $\widehat{Q}_\nu$ and is controlled by the binary parameter `adjust` (`ADJ`).
To reduce the danger of finding only a local optimum, the procedure is repeated using n_I initial subsets J_0 and the vector $\widehat{\beta}$ leading to the smallest value of the LTS–criterion found is used as final solution. The value n_I can be controlled with the parameter `nsamp` (`NSP`).
Another approximate LTS–algorithm based on Rousseeuw and Hubert (1997) is `lqs` implemented in the `R`–package `MASS` (2007). There are three parameters: `nsamp` (`NSP`) specifies the number of the random subsets, `psamp` (`PSP`) is the size of the random subsets and `adjust` (`ADJ`) gives the possibility of using the IA of (5). Every initial subset yields an LTS–estimation and the subset with the smallest criterion value becomes the final solution.

2.3 An evolutionary algorithm for LTS–regression

The evolutionary algorithm `ltsEA` for LTS proposed here works as follows: First a subset Ψ_0 called individual (the 0th generation) with

$$\Psi_0 \;=\; \{\, \psi_1^0,\ \ldots,\ \psi_p^0 \,\} \;\subset\; \{\, z_1,\ \ldots,\ z_n \,\} \;=\; Z \qquad (6)$$

is drawn from Z. Again p is the number of unknown parameters in (1) and usually much smaller than h. If the resulting hyperplane defined by $\psi_1^0, \ldots, \psi_p^0$ is unique, the h smallest squared residuals of all n data points are summed to get an initial LTS–value $\widehat{Q}_0$ of Q, otherwise $\widehat{Q}_0$ is set to infinity.
We use two mutation operations, which give new values of β und Q by generating a new individual: point replacement and move operation. A point replacement replaces one point ψ_d^0, $d \in \{1, \ldots, p\}$ in Ψ_0 by a point φ_q^0, $q \in \{p+1, \ldots, n\}$ not in Ψ_0 to create a new individual Ψ_0'. A move operation shifts the estimated hyperplane parallel through a point ϕ_r^0, $r \in \{p+1, \ldots, n\}$ not in Ψ_0. The new individual Ψ_0' includes ϕ_r^0 and those $p-1$ observations with the smallest squared distances to the shifted hyperplane. The indices d, q and r are taken uniformly at random. The used mutation is determined randomly with the parameter `percentagemove` (`PER`) specifying the probability of a move operation. Note that as opposed to previous approaches we swap one point out of p, and not out of h in the point replacement. This is motivated by the success of the Fast–LTS initialized with p–elemental subsets. The goodness of fit (fitness) of the individuals is measured by the LTS–criterion. The LTS–value $\widehat{Q}_0'$ of the new individual Ψ_0' is compared to the previous value $\widehat{Q}_0$. The individual with smaller LTS–value is taken into the next generation and is denoted by Ψ_1.
The LTS–criterion value can be further reduced in each step by fitting a new hyperplane to the h data points with the smallest squared residuals to the current hyperplane, setting `hyperplane adjust` (`HYP`) to true. These steps of generating new individuals $\psi_1, \psi_2, \ldots$ are continued until n_G generations Ψ_t have been obtained, $t \in \{1, \ldots, n_G\}$. The parameter `generations` (`GEN`) controls n_G. The second stop criterion `wait for improvement` (`WFI`) specifies the number n_W of subsequent generations after which the algorithm stops if there is no improvement. To reduce the probability of reaching only a local minimum, we repeat the whole procedure n_I times. The number of starts n_I is controlled by the parameter `number of starts` (`NOS`).
To further reduce the risk of getting stuck in a local optimum, we include the possibility of using Simulated Annealing. Simulated Annealing allows individuals with a worse criterion value to become the new individual if we set the parameter `annealing` (`ANN`) to true. With $f(t)$ being a monotonic decreasing function of the current number of generations t, the probability of Ψ_t' to become the new individual Ψ_{t+1} is

$$P\ (\Psi_t' \;=\; \Psi_{t+1}) \quad = \quad \begin{cases} 1 & , \text{ if } \widehat{Q}_t - \widehat{Q}_t' \;\geq\; 0 \\ \exp\left(\frac{\widehat{Q}_t - \widehat{Q}_t'}{f(t)}\right) & , \text{ if } \widehat{Q}_t - \widehat{Q}_t' \;<\; 0 \ . \end{cases} \qquad (7)$$

3 Design of experiments for a comparison

We use Central Composite Designs (CCD) (Draper, 1982) to design our experiments, with which we want to estimate linear and quadratic effects and interactions of second order of all parameters on the LTS–criterion and the computation time for all three algorithms. Each design has 72 runs with 32 runs from a factorial design. For `ltsReg` we use an eight times repeated 2^2 factorial design and for `lqs` a four times repeated 2^3 factorial design. For `ltsEA` a 2^{6-1}–design with resolution VI can be found. So for both algorithms all linear effects and interactions are estimable without bias. With 32 star runs and 8 center points the quadratic effects (of the non–binary parameters) can also be estimated without bias for both algorithms. As factorial runs and star runs are repeated, the star run value α is set to (Draper, 1982) $\alpha = \sqrt[4]{\frac{m_f \cdot n_f}{m_s}}$, where n_f is the number of factorial runs, m_f is the number of their reccurences and m_s is the number of recurrences of the star runs. This ensures the rotability of the design, which allows to predict the response with the same variance at every point at the same distance from the center of the design, which is the setting of the center points. Let $\eta = \left\lfloor \frac{p+h}{2} \right\rfloor$. To compare `ltsReg`, `lqs` and `ltsEA` the parameter settings in Table 1 are chosen. The setting of `PSP` depends on the number of parameters p. We use the same settings for `NSP` and `NOS` as both parameters define the number of starts of the procedure.

	$-\alpha$	-1	0	$+1$	$+\alpha$
`NSP` (`ltsReg`)	50	122	500	878	950
`NSP` (`lqs`)	50	182	500	818	950
`ADJ`		`FALSE`		`TRUE`	
`PSP`	p	$\left\lfloor \eta - \frac{h-\eta}{\alpha} \right\rfloor$	η	$\left\lfloor \eta + \frac{h-\eta}{\alpha} \right\rfloor$	h
`NOS`	50	232	500	768	950
`GEN`	500	1108	2000	2892	3500
`WFI`	50	131	250	369	450
`PER`	0	20.27	50	79.73	100
`ANN`		`FALSE`		`TRUE`	

Table 1. Parameter settings in the designed experiments.

To find an optimal setting which minimises the LTS–criterion Υ^1 and the computation time Υ^2, the desirability index for a minimisations of Derringer and Suich (1980) is chosen. After executing the runs of the CCD, the effects of all parameters are estimated by OLS and predictions $\widehat{v}^\gamma$ for Υ^γ, $\gamma = 1, 2$,

are made for each adjustable parameter–combination. The transformation

$$\varrho^\gamma \;=\; \begin{cases} 0 & , \text{ if } \widehat{v}^\gamma \;\leq\; \xi_\gamma \\ \left(\dfrac{\widehat{v}^\gamma \;-\; \xi_\gamma}{\chi_\gamma \;-\; \xi_\gamma} \right)^{\lambda_\gamma} & , \text{ if } \widehat{v}^\gamma \;\in\; (\, \chi_\gamma,\, \xi_\gamma \,) \\ 1 & , \text{ if } \widehat{v}^\gamma \;\geq\; \chi_\gamma \end{cases} \tag{8}$$

assigns a desirability value $\varrho^\gamma \in [0,1]$ to each $\widehat{v}^\gamma$, where $\varrho^\gamma = 1$ means maximal and $\varrho^\gamma = 0$ means minimal desirability. χ_γ and ξ_γ are reasonable lower and upper bounds for Υ^γ. The transformation parameter $\lambda_\gamma > 0$ is used to weight the importance of Υ^γ. The geometric mean $R = \left(\prod_{\gamma=1}^{2} \varrho^\gamma \right)^{\frac{1}{2}}$ is called desirability index for the prediction $\widehat{v}^1$ and $\widehat{v}^2$. The combination of the parameters, which gives the highest value for R is considered to be a simultaneously optimal parameter setting for Υ^1 and Υ^2 for this choice of χ_γ, ξ_γ and λ_γ, $\gamma = 1, 2$.

4 Comparative study

For a comparison we simulate data with $n = 500$ data points from two different models. In both models the independent regressors $x_{\bullet 1}, \ldots, x_{\bullet(p-1)}$ stem from a uniformly distributed random design on the interval (0,1) and $E_1, \ldots, E_n$ are i.i.d. $\mathcal{N}(0,1)$. The first model is for $i = 1, \ldots, n$ given by

$$Y_i \;=\; \beta_0 + \beta_1(x_{i,1} + \Gamma_i^x) + \ldots + \beta_{(p-1)}(x_{i,(p-1)} + \Gamma_i^x) + E_i + \Gamma_i^y \quad , \tag{9}$$

and contains outliers in y and $x_{\bullet 1}, \ldots, x_{\bullet(p-1)}$ with $p = 1, \ldots, 30$,

$$\Gamma_i^x = \begin{cases} 0 & , \text{if } x_{i,j} \text{ is no outlier} \\ 1.5 & , \text{if } x_{i,j} \text{ is an outlier} \end{cases} \quad \text{and} \quad \Gamma_i^y = \begin{cases} 0 & , \text{if } y_i \text{ is no outlier} \\ 5 & , \text{if } y_i \text{ is an outlier} \end{cases} \tag{10}$$

$\Gamma_i^x = 1.5$ means that for each $j = 1, \ldots, (p-1)$, $x_{i,j}$ is an outlying value. There are exactly 40% outliers: Γ_i^x and Γ_i^y are each for 25% of the data points set to the non–zero value and for 10% of all cases both together turn to an outlying value.
In the second model we simulate a structural change for $p = 1, \ldots, 30$ from

$$Y_i \;=\; \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{(p-1)} x_{i,(p-1)} + E_i \quad , \; i = 1, \ldots, n \tag{11}$$

The slope β_1 is 1, while $\beta_2, \ldots, \beta_{p-1}$ are 0, i.e. $x_{\bullet 2}, \ldots, x_{\bullet(p-1)}$ add only noise to the problem. The structural change is in the intercept, which is

$$\beta_0 \;=\; \begin{cases} 0 & , \text{ if } x_{i,1} \leq x_{(300),1} \\ 10 & , \text{ if } x_{i,1} > x_{(300),1} \end{cases} \tag{12}$$

with $x_{(300),1}$ as 60%–percentile of $x_{1,1}, \ldots, x_{n,1}$. A good robust estimate of β should give β_0 close to 0 and the slopes as above.
As the realized observed computation times are heteroskedastic, we transform this response with the logarithm to detect the true influences of the parameters. In (8) we set $\lambda_1 = 5$ as we are interested in small changes of the LTS–value to get closer to the optimal criterion value Q. The bounds χ_1 and ξ_1 are chosen as the minimum and the maximum of the observed LTS–values of both algorithm for each data set. For the computation time the upper bound $\xi_2 = 300$ seconds is chosen. For the lower bound χ_2 we take the minimum of all observed computation times for each model and for fixed p. We set $\lambda_2 = \lambda_1$ treating both criteria the same way.
For most of the data sets an optimal setting for `ltsReg` and `lqs` is found with large `NSP` and activated `ADJ`. The setting of `PSP` from `lqs` depends on p: for a small dimension it is set to p, for higher dimensions it is mostly set to $\lfloor \eta + (h - \eta)/\alpha \rfloor$. So for large p the "curse of dimensionality" seems to have a higher influence on the choice of `PSP` than the outliers in this data situation.
An optimal setting for `ltsEA` is found with `NOS` and `WFI` set to the smallest value. `ADJ` is activated while `ANN` is not. Both binary parameters increase the computation time, but only `ADJ` decreases the criterion value. For small p `GEN` is set to a low level, for high p on a high level. The parameter `PER` is for small p equal to 0, so only point replacements are needed there. For most data sets 50% or sometimes up to 80% move operations are useful.
For both models `ltsEA` delivers better criterion values than `ltsReg` and than `lqs`, especially for a high dimension p, but it needs higher computation times. We take a larger value for `NSP` in `ltsReg` and `lqs` to achieve a computation time for both algorithms similar to the one of `ltsEA`. Figure 1 shows boxplots of this situation for both models. It can be seen that even in this situation `ltsEA` obtains better LTS–criterion values than the Fast–LTS and the `lqs` algorithm. The variance of the LTS–criterion values can be reduced with a higher setting of `NOS`.

5 Conclusion

Design of experiments is useful to compare algorithms regarding their output and computation time. In this study we found that the evolutionary algorithm `ltsEA` delivers better LTS–criterion values than `ltsReg` and `lqs` for the same number of starts and comparable computation times in cases of a high dimensional regressor space and a high percentage of contamination. We have also applied `ltsEA` to the data sets from Rousseeuw and van Driessen (2006) and obtained similarly good results as by `ltsReg` except for two examples. Advantages of `ltsEA` are the move operation and the hyperplane adjustment, which are helpful to avoid local optima. A further improvement of `ltsEA` could be another parameter to set the number of start points, as in `lqs` the "curse of dimensionality" seems to have a higher influence on this choice than the outliers. We plan to make `ltsEA` available in `R` soon.

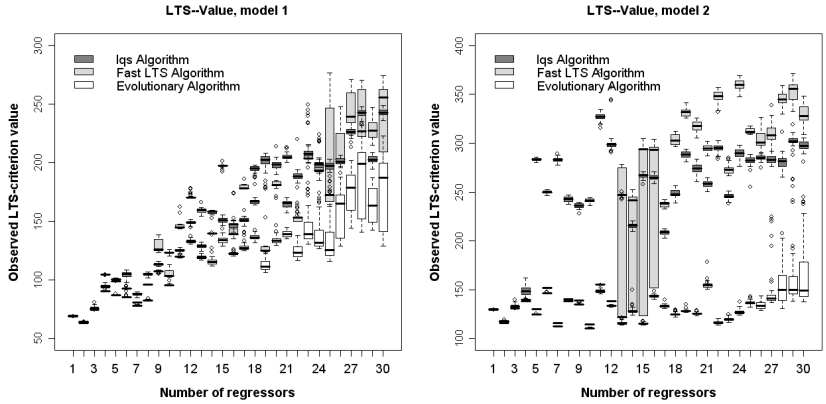


Fig. 1. Boxplots of the LTS–criteria for a comparable computation time.

Acknowledgements

Financial support of the DFG (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

BERNHOLT, T. (2005): Robust estimators are hard to compute. *Technical report (52/2005), SFB 475, University Dortmund.*

CHAKRABORTY, B. and CHAUDHURI, P. (2003): On the use of genetic algorithm with elitism in robust and nonparametric multivariate analysis. *Austrian Journal of Statistics 32 (1), 13–27.*

DERRINGER, G. and SUICH, R. (1980): Simultaneous optimization of several response variables, *Journal of Quality Technology 12, 214–219.*

DRAPER, N.R. (1982): Center points in second–order response surface designs. *Technometrics 24 (2), 127–133.*

R DEVELOPEMENT CORE TEAM (2007): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

ROUSSEEUW, P.J. and HUBERT, M. (1997): Recent developments in PROGRESS. In: Dodge, Y. (Edr): L_1-*statistical procedures and related topics.* Hayward, Calif., Institute of Mathematical Statistics, 201–214.

ROUSSEEUW, P.J. and LEROY, A.M. (1987): *Robust regression and outlier detection.* John Wiley & Sons, New York.

ROUSSEEUW, P.J. and VAN DRIESSEN, K. (2006): Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery 12 (1), 29–45.*

TODOROV, V. (1992): Computing the minimum covariance determinant estimator (MCD) by simulated annealing. *Computational Statistics & Data Analysis 14 (4), 515–525.*

WOODRUFF, D.L. and ROCKE, D.M. (1994): Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association, 89, 888–896.*

Convergence of Componentwise Aitken δ^2 Acceleration of the EM algorithm

Michio Sakakihara[1] and Masahiro Kuroda[2]

[1] Department of Information Science, Okayama University of Science
Okayama 700-0005, Ridai-cho 1-1, Japan, *sakaki@mis.ous.ac.jp*
[2] Department of Socio-Information, Okayama University of Science
Okayama 700-0005, Ridai-cho 1-1, Japan, *kuroda@soci.ous.ac.jp*

Abstract. The EM algorithm of Dempster, Laird and Rubin (1977) is a very general and popular iterative computational algorithm for finding maximum likelihood estimates from incomplete data and broadly used to statistical analysis with missing data, because of its numerical stability, computational simplicity and flexibility in interpreting the incompleteness of data. However, the EM algorithm converges slowly when there is a relatively large proportion of missing data. In this paper, we propose the componentwise Aitken δ^2 acceleration for the EM algorithm. The accelerated EM algorithm is a kind of extension of Aitken δ^2 method for scalar cases. We discuss the formulation of the acceleration from the secant method in Banach spaces and prove the convergence theorem.

Keywords: Aitken δ^2 method, EM algorithm, acceleration of convergence

1 Introduction

The EM algorithm of Dempster, Laird and Rubin (1977) is a very general and popular iterative computational algorithm for finding maximum likelihood estimates from incomplete data and broadly used to statistical analysis with missing data, because of its numerical stability, computational simplicity and flexibility in interpreting the incompleteness of data. However, the EM algorithm converges slowly when there is a relatively large proportion of missing data. In order to speed up the convergence of the EM algorithm, various acceleration algorithms have been proposed. Louis (1982) suggested the hybrid EM algorithm incorporating Aitken's acceleration method. Jamshidian and Jennrich (1993) proposed an acceleration algorithm based on conjugate gradients. Lange (1995) used a quasi-Newton algorithm to accelerate the EM algorithm. Their algorithms are based on the Newton-Raphson algorithm and then are potentially including unavoidable problems. Firstly, it requires, at each iteration, the computation of the information matrix that is the matrix of the negative of the second-order partial derivations of the log-likelihood function. Then its computation is likely to become rapidly complicated as the number of parameters is increasing. Secondly, the Newton-Raphson algorithm may be sensitive for an initial value than the EM algorithm. Because

of such possible computational difficulties, their accelerated EM algorithm are lost the attractive features such as the stability, simplicity of the EM algorithm.

Recently, we proposed an alternative acceleration of the EM algorithm via the vector ε algorithm (Kuroda and Sakakihara (2006B)): The vector ε acceleration of EM algorithm that accelerates the convergence of the sequence of EM iterates using the vector ε algorithm. Then the vector ε acceleration of the EM algorithm preserves the stability, simplicity of the EM algorithm. The ε algorithm has some scalar products and then is a kind of extension of the Aitken δ^2. The aim of this paper is to propose a simpler acceleration method for EM algorithm by applying Aitken δ^2 method. For a special case that it is possible to transform a multi-dimensional problem, we discussed the application of Aitken δ^2 to into a sequence of one-dimensional one (Kuroda and Sakakihara (2006A)). We also examined the componentwise Aitken δ^2 acceleration for the EM algorithm and discussed with the efficiency of the approach by numerical experiments (Sakakihara and Kuroda (2005)). However, the rigorous formulation for the componentwise Aitken δ^2 acceleration for EM algorithm does not illustrate in that paper. In this paper, we present the formulation of the componentwise Aitken δ^2 acceleration and prove the convergence theorem for the acceleration.

2 The EM algorithm incorporating acceleration methods

This section presents the EM algorithm incorporating an acceleration method. Let y be observed data with a sample space Ω_Y and x be complete data augmented by y with a sample space Ω_X. We assume that there exists some function $h(x) = y$ that relates x to y. Let $f(\cdot|\theta)$ denote a probability density function depending on an unknown parameter vector $\theta = (\theta_1, \ldots, \theta_d)^T$ with a parameter space Θ. Define the conditional expectation of $\log f(x|\theta)$ given y and θ' as

$$Q(\theta|\theta') = \mathrm{E}[\log f(X|\theta)|y, \theta'].$$

Starting from an initial value $\theta^{(0)} \in \Theta$, the EM algorithm chooses

$$\theta^{(t)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(t-1)}),$$

at each iteration $t = 1, 2, \ldots$. Let $\theta[t]$ be the vector subsequence of the sequence generated by the EM algorithm up to the t-th iteration. In order to accelerate the convergence of the EM algorithm, we introduce an accelerator $A(\theta[t])$ which transforms the sequence generated by the EM algorithm into the sequence which converge more rapid than the original one.

Then the EM algorithm adding an acceleration process performs the following steps:

E-step : Calculate

$$Q(\theta|\theta^{(t-1)}) = \mathrm{E}[\log f(X|\theta)|y, \theta^{(t-1)}].$$

M-step : Choose $\theta^{(t)}$ such that

$$Q(\theta^{(t)}|\theta^{(t-1)}) \geq Q(\theta|\theta^{(t-1)})$$

for all $\theta \in \Theta$.

Acceleration-step : Obtain $\dot{\theta}$ by using

$$\dot{\theta} = A(\theta[t])$$

and check the convergence using a desired accuracy.

The accelerated EM algorithm with $A(\theta[t])$ is an extension of the EM algorithm without affecting its simplicity and stability. Moreover, the local convergence properties of the EM algorithm are preserved, because the accelerated algorithm does not improve the updating equations in the E- and M-steps in themselves but only adding the accelerating process. The number of arguments for the accelerator depends on the acceleration method.

3 Accelerators

In this section we illustrate some examples of acceleration methods for the EM algorithm described in Section 2.

3.1 Vector ε algorithm by Wynn (1961)

The ε algorithm of Wynn (1956) is a nonlinear method for accelerating the convergence of sequences. It is known that its algorithm is powerful for the sequence converging linearly. Wynn (1962) extended the ε algorithm to the vector sequence. We present the ε algorithm for the vector case.

For the acceleration of convergence of a sequence $\{\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(t)}, \ldots\}$, the rule of the vector ε algorithm is

$$\begin{aligned} \varepsilon^{(t,-1)} &= 0, \\ \varepsilon^{(t,0)} &= \theta^{(t)}, \\ \varepsilon^{(t,k+1)} &= \varepsilon^{(t+1,k-1)} + \left[\varepsilon^{(t+1,k)} - \varepsilon^{(t,k)}\right]^{-1}, \end{aligned}$$

where

$$\left[\varepsilon^{(t+1,k)} - \varepsilon^{(t,k)}\right]^{-1} = \frac{\varepsilon^{(t+1,k)} - \varepsilon^{(t,k)}}{\left\|\varepsilon^{(t+1,k)} - \varepsilon^{(t,k)}\right\|^2}$$

and $\|\cdot\|$ is usual Euclid norm for vectors. For the initializations $\varepsilon^{(t,0)} = \theta^{(0)}$ and $\varepsilon^{(t,-2)} = \infty$, we obtain the sequence generated by

$$\dot{\theta}^{(t)} = \varepsilon^{(t,2)} = \theta^{(t+1)} + \left[\left(\theta^{(t)} - \theta^{(t+1)}\right)^{-1} + \left(\theta^{(t+2)} - \theta^{(t+1)}\right)^{-1}\right]^{-1}. \quad (1)$$

Then $\dot{\theta}^{(t)}$ converges faster than $\theta^{(t)}$ to the limit of the vector sequence $\theta^{(\infty)}$. Therefore we have

$$\theta^{(\infty)} = \lim_{t\to\infty} \theta^{(t)} = \lim_{t\to\infty} \dot{\theta}^{(t)}. \quad (2)$$

Note that, at each iteration, the vector ε algorithm is achieved at a cost of $O(d^2)$ while the Newton-Raphson algorithm requires $O(d^3)$, so that the computational cost is likely to become more expensive as d becomes large.

3.2 Componentwise Aitken δ^2 acceleration

Set $x = (x_1, \ldots, x_d)^T$ and $F(x) = (f_1(x), \ldots, f_d(x))^T$. Let us consider the fixed point equation and the generated iteration:

$$x = F(x), \quad (3)$$
$$x^{(t+1)} = F(x^{(t)}), \quad (4)$$

where $F : R^d \to R^d$. An accelerated sequence for Equation (4) is able to generate by applying a secant method as follows:

$$y^{(t+2)} = x^{(t+1)} - J_a^{-1}(x^{(t)}, x^{(t+1)})(x^{(t+1)} - F(x^{(t)})) \quad (5)$$

where $J_a(x^{(t)}, x^{(t+1)})$ is an approximation of the Jacobian matrix arising in the Newton method for Equation (5) and is a dense matrix in general. If we take the diagonal matrix such as

$$\begin{aligned} J_a(x^{(t)}, x^{(t+1)}) &= A_D(x^{(t)}, x^{(t+1)}) \\ &= \left(\delta_{ij}\left[\frac{x_i^{(t+1)} - f_i(x^{(t+1)}) - x_i^{(t)} + f_i(x^{(t)})}{x_i^{(t+1)} - x_i^{(t)},}\right]\right) \end{aligned} \quad (6)$$

where δ_{ij} Kronecker's delta, we obtain the following componentwise accelerated sequence for Equation (4):

$$\begin{aligned} y_i^{(t+2)} &= x_i^{(t+1)} \\ &\quad - \left[\frac{x_i^{(t+1)} - f_i(x^{(t+1)}) - x_i^{(t)} + f_i(x^{(t)})}{x_i^{(t+1)} - x_i^{(t)}}\right]^{-1}\left(x_i^{(t+1)} - f_i(x^{(t+1)})\right) \end{aligned} \quad (7)$$

since

$$A_D^{-1}(x^{(t)}, x^{(t+1)}) = \left(\delta_{ij}\left[\frac{x_i^{(t+1)} - f_i(x^{(t+1)}) - x_i^{(t)} + f_i(x^{(t)})}{x_i^{(t+1)} - x_i^{(t)}}\right]^{-1}\right). \quad (8)$$

Equation (7) reduces to the following equation with the computed values $x_i^{(t)}$:

$$y_i^{(t+2)} = \frac{x_i^{(t)} x_i^{(t+2)} - [x_i^{(t+1)}]^2}{x_i^{(t)} - 2x_i^{(t+1)} + x_i^{(t+2)}}. \tag{9}$$

By applying Equation (9) to the sequence $\theta^{(t)}$ we have the transformed sequence $\bar{\theta}^{(t)}$. Summarizing these steps we give the following algorithm:

1) Set an initial vector $\theta^{(0)}$
2) Compute $\theta^{(1)}$ and $\theta^{(2)}$ with the E- and M-steps
3) Compute transformed vector $\bar{\theta}^{(2)}$ by

$$\bar{\theta}_i^{(2)} = \frac{\theta_i^{(0)} \theta_i^{(2)} - [\theta_i^{(1)}]^2}{\theta_i^{(0)} - 2\theta_i^{(1)} + \theta_i^{(2)}} \tag{10}$$

4) Compute $\theta^{(t+2)}$ with the E- and M-steps and obtain next transformed vector $\bar{\theta}^{(t+2)}$ with

$$\bar{\theta}_i^{(t+2)} = \frac{\theta_i^{(t)} \theta_i^{(t+2)} - [\theta_i^{(t+1)}]^2}{\theta_i^{(t)} - 2\theta_i^{(t+1)} + \theta_i^{(t+2)}} \tag{11}$$

5) Check the convergence with $|\bar{\theta}_i^{(t+2)} - \bar{\theta}_i^{(t+1)}| < \epsilon$ for each of them. If a component satisfies the criteria then we stop the next evaluation for the component. Set $t = t + 1$ and go to Step 4). If all components satisfy the criteria then we stop the iteration.

4 Convergence

Let us consider the convergence of the accelerated EM algorithm proposed in the Section 2 with accelerators illustrated in Section 3.

4.1 Vector ε algorithm

We show the following result:

Theorem 1. *The sequence $\{\dot{\theta}^{(t)}\}$ generated by the vector ε acceleration of the EM algorithm converges to the stationary point θ^* of the EM sequence.*

Proof. See the proof in the paper (Kuroda and Sakakihara 2006B). □

4.2 Componentwise Aitken δ^2 acceleration

In the formulation of componentwise Aitken δ^2 acceleration, we apply the secant method by Schmit(1963) and Ulm(1967). The convergence of the secant method is described for Banach spaces. We illustrate some results by Solak and Strus(1976). Let B denote Banach space with the norm $\|\cdot\|$. Consider the equation

$$F(x) = (f_1(x), \ldots, f_d(x))^T = 0, \tag{12}$$

where $f : B \to B$ is continuous. Let $A(u, w)$, $u, w \in B$ be the family of linear continuous mapping of B. Suppose that $A(u, w)$ has the inverse $A^{-1}(u, w)$. The problem solving Equation (12) is equivalent to solving the equation

$$x = x - A^{-1}(u, w)F(x). \tag{13}$$

Let $V \subset B$ be the open ball containing the fixed point for Equation (13) which is also the solution of Equation (12). We choose arbitrary the sequence $\{u^{(t)}\}, \{w^{(t)}\} \subset V$ and $x^{(0)} \in V$ and compute a sequence $\{x^{(t)}\}$ given by the iterate:

$$x^{(t+1)} = x^{(t)} - A^{-1}(u^{(t)}, w^{(t)})F(x^{(t)}), \tag{14}$$

where $x^{(t)} \to x^*$ as $t \to \infty$. Then we have:

Theorem 2. *Compute the transformed sequence* $\{y^{(t)}\}$ *by*

$$y^{(t+1)} = x^{(t)} - A^{-1}(u^{(t)}, w^{(t)})F(x^{(t)}). \tag{15}$$

Assume that the fixed point for Equation (13) is the solution of Equation (12) and

$$\exists > 0 \forall u, w \in V : \|A(u, w)\| < K, \tag{16}$$

where K *is a positive constant. Then* $y^{(t)} \to x^*$ *as* $t \to \infty$.

Proof. From Equation (15) and the assumption we obtain

$$\|y^{(t+1)} - x^{(t)}\| \leq \|A^{-1}(u^{(t)}, w^{(t)})F(x^{(t)})\| \leq KF(x^{(t)}). \tag{17}$$

From the assumption that $x^{(t)} \to x^*$ as $t \to \infty$, we have $F(x^{(t)}) \to 0$ as $t \to \infty$. Therefore we proved the theorem. □

For the componentwise Aitken δ^2 we have the following result by applying Theorem 2:

Theorem 3. *Let* $B = R^d$ *and* $\{\theta^{(t)}\}$ *be the EM sequence. When the transformed sequence* $\bar{\theta}^{(t)}$ *is generated by Equation (10), we have* $\bar{\theta}^{(t)} \to \theta^*$ *where* θ^* *denotes the fixed point for the EM iteration.*

5 Numerical experiments

Consider a 2×2 contingency table with completely and partially classified observations. Let X and Y be dichotomous variables and $\theta = \{p_{ij}\}_{i,j=1,2}$ be a set of joint probabilities of X and Y. Denote the cross-classified data of X and Y as $n_{XY} = \{n_{XY}(i,j)\}_{i,j=1,2}$, and the partially classified data of X as $n_X = \{n_X(i)\}_{i=1,2}$ and Y as $n_Y = \{n_Y(j)\}_{j=1,2}$. Assume that the datasets have a multinomial distribution with an unknown parameter θ.

Table 1. Contingency table with completely and partially classified data.

	n_Y		n_X		n_{XY}			
	i missing		$i=1$	$i=2$	$i=1$		$i=2$	
	$j=1$	$j=2$	j missing		$j=1$	$j=2$	$j=1$	$j=2$
(a)	50	30	70	100	5	4	2	1
(b)	100	60	70	100	5	4	2	1
(c)	300	100	70	100	5	4	2	1
(d)	1000	600	70	100	5	4	2	1

Table 2. The number of iterations for $\epsilon = 10^{-8}$.

	$\{n_Y(j)\}_{j=1,2}$			
	(a)	(b)	(c)	(d)
EM	201	252	442	996
vector ε acceleration	57	71	192	198
Aitken δ^2 acceleration	66	76	213	201

The datasets are shown in Table 1. For these data patterns, the convergence of the EM algorithm is quite slow, because its convergence is deeply associated with the proportion of missing data. In Table 2, we summarize the number of iterations for the EM and the vector ε and componentwise Aitken δ^2 acceleration of the EM algorithm for $\epsilon = 10^{-8}$ and the datasets (a) to (d). As shown in Table 2, the convergence of the accelerated EM algorithms is significantly faster than the EM algorithm for all datasets.

6 Concluding remarks

In this paper, we discuss the EM algorithm with the vector ε and componentwise Aitken δ^2 accelerators. Both accelerators are very simple computational

procedure. The EM algorithm with these accelerators is an interesting approach within the framework of the EM algorithm without losing its good properties. According to the papers of Kuroda and Sakakihara (2006B) and Sakakihara and Kuroda (2005), we found that these accelerators effectively speed up the convergence of the EM algorithm by numerical experiments. For the vector ε algorithm we presented that the accelerated sequence also converges to the fixed point of EM iterates (Kuroda and Sakakihara (2006B)).

We proved the convergence of the componentwise Aitken δ^2 acceleration by the use of some results on the secant method in Banach spaces. Moreover, numerical experiments are shown to examine the practical utility of the presented methods. Table 2 indicates that the Aitken δ^2 acceleration requires the number of iterations slightly greater than that of the vector ε acceleration. However, the vector ε algorithm contains some inner products to proceed the computation. Therefore, the computational cost of the componentwise Aitken δ^2 acceleration is less expensive than that of the vector ε algorithm. Our future work is to give the theoretical evaluation for the rate of convergence of the EM algorithm with the componentwise Aitken δ^2 acceleration.

References

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B.* **39**, 1 – 22.

JAMSHIDIAN, M. and JENNRICH, R.I. (1993): Conjugate gradient acceleration of the EM algorithm. *J. Amer. Statist. Assoc.* **88**, 221 – 228.

KURODA, M. and SAKAKIHARA, M. (2006A): Acceleration of the EM and ECM algorithms for log-linear models with missing data, *Proceedings in Computational Statistics 2006*, 591–598.

KURODA, M. and SAKAKIHARA, M. (2006B): Accelerationthe convergence of the EM algorithm using the vector ε algorithm, *Comput. Statist. Data Anal.*, **51**, 1549–1561.

LANGE, K. (1995): A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **57**, 425 – 437.

LOUIS, T.A. (1982): Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. B.* **44**, 226 – 233.

MCLACHLAN, G.J.. and KRISHNAN, T. (1997): *The EM algorithm and extensions*, Wiley, New York.

SAKAKIHARA, M. and KURODA, M. (2005): Improving convergence rate of EM algorithm via componentwise Aitken Δ^2 acceleration, *Information*, **6**, 215–230.

SCHMIDT, J.W. (1963): Eine Übertragung der Regula Falsi auf Gleichungen in Banachräumen, *Z.angew Math. Mech.*, **43**, 1-8,97-110.

SOLAK, W. and STRUS, M. (1976): The secant method in Banach spaces, *Computing*, **16**, 201-209.

ULM, S. (1967): On generalization divided differences, Esti nsu teaduste Akademin toimetised, *Fusika Mathematik*, **16/1**, 13-26.

WYNN, P. (1956): On a procrustean technique for the numerical transformation of slowly convergent sequences and series. *Proc. Cambridge Phil. Sco.* **52**, 663 – 671.

WYNN, P. (1961): The epsilon algorithm and operational formulas of numerical analysis. *Math. Comp.* **15**, 151 – 158.

Part XVI

Partial Least Squares and Structural Equations Models

From Multiblock Partial Least Squares to Multiblock Redundancy Analysis. A Continuum Approach.

Stéphanie Bougeard[1], Mohamed Hanafi[2], Coralie Lupo[1] and El Mostafa Qannari[2]

[1] $AFSSA$, Department of Epidemiology - Zoopole, BP53, 22440 Ploufragan, France, *s.bougeard@afssa.fr, c.lupo@afssa.fr*

[2] $ENITIAA-INRA$, Department of Chemometrics and Sensometrics - Rue de la Géraudière BP82225, 44322 Nantes Cedex, France, *hanafi@enitiaa-nantes.fr, qannari@enitiaa-nantes.fr*

Abstract. Multiblock PLS regression is devoted to investigate the relationships between one or several data sets and several predictive blocks of variables. An extension of redundancy analysis to the multiblock setting is proposed. From the solutions of both these approaches, *i.e.* multiblock PLS regression and multiblock redundancy analysis, it turns out that they are the two end points of a continuum approach that we propose to investigate.

Keywords: multiblock PLS, multiblock redundancy analysis, continuum approach, multicollinearity, multiway data analysis, cross-validation, regularization parameter.

1 Introduction

In a previous paper (Bougeard et al., *In Press*), we have shown a connection between $PLS2$ regression (Wold, 1966) and redundancy analysis (Rao, 1964; Van den Wollenberg, 1977) of two data tables X and Y. Indeed we have shown that these two methods aim at maximizing the same criterion, namely $cov^2(t,u)$ with $t = Xw$ and $u = Yv$. The difference between $PLS2$ regression and redundancy analysis lies in the constraints imposed on the components to be determined. More precisely, $PLS2$ regression imposes the constraints $||w|| = ||v|| = 1$ whereas redundancy analysis imposes the constraints $||t|| = ||v|| = 1$. It is well known that the vector of loadings w is given by the eigenvector associated with the largest eigenvalue of matrix $(X'YY'X)$ for $PLS2$ regression and $(X'X)^{-1}(X'YY'X)$ for redundancy analysis. Thereafter, we have introduced a continuum approach by considering a combination of the two matrices. We have demonstrated interesting properties which pinpoint the rationale behind the continuum approach.

In this paper we start by extending redundancy analysis to the multiblock setting and we highlight its connection with multiblock PLS (Wold,

1984; Wangen and Kowalski, 1988). It turns out that multiblock PLS regression ($MPLS$) and multiblock Redundancy Analysis (MRA) are the two end points of a continuum approach that we propose to investigate. The properties highlight the rationale of the continuum approach and how it handles the multicollinearity problem. As the continuum establishes a bridge between $MPLS$ and MRA, we shall refer to it as "Multiblock Continuum Redundancy PLS regression" ($MCR-PLS$). The interest of both methods and the properties of the continuum are illustrated on the basis of a data set pertaining to veterinary epidemiology.

2 Methods

2.1 Multiblock *PLS* regression

Consider the multiblock setting where we have $(K+1)$ data sets: a data set Y to be predicted from K data sets X_k $(k=1,\ldots,K)$. The Y table contains Q variables and each table X_k contains p_k variables. All these variables are measured on the same N individuals and supposed to be column centred.

In the case where only one block of variables is explained by several blocks of explanatory variables, Westerhuis et al. (1998), Qin et al. (2001) and then Vivien (2002) show that the solution obtained from the iterative algorithm of $MPLS$ is equivalent to the solution obtained from a PLS regression of Y and X, where the merged data set X is defined as $[X_1|\ldots|X_K]$. More precisely, Vivien (2002) proved that $MPLS$ seeks in a first step a component $t^{(1)}=Xw^{(1)}$ which is highly related to a component $u^{(1)}=Yv^{(1)}$ and which sums up partial components $t_k^{(1)}$ respectively associated with the blocks X_k. More formally, $MPLS$ consists in maximizing the following criterion (1).

$$cov^2(u^{(1)},t^{(1)}) \quad \text{with} \quad t^{(1)}=\sum_{k=1}^{K} a_k^{(1)} t_k^{(1)},\ u^{(1)}=Yv^{(1)}, \qquad (1)$$
$$t_k^{(1)}=X_k w_k^{(1)}, \quad \textstyle\sum_{k=1}^{K} a_k^{(1)2}=1, \quad ||w_k^{(1)}||=||v^{(1)}||=1$$

The optimal vector of loadings $w^{(1)}$ is given by the eigenvector of the matrix $(X'YY'X)$ associated with the largest eigenvalue $\lambda_{MPLS}^{(1)}$. The vector $v^{(1)}$ is given by the eigenvector of $M_{MPLS}=(Y'XX'Y)$ associated with the same eigenvalue. Then the partial axes $w_k^{(1)}$ are given by $w_k^{(1)}=w_k^{(1)*}/||w_k^{(1)*}||$ where $w_k^{(1)*}$ are the blocks vectors of $w^{(1)}$, namely $w^{(1)}=[w_1^{(1)*}|\ldots|w_K^{(1)*}]'$. It is clear that $a_k^{(1)}=||w_k^{(1)*}||$ which indeed fulfills the constraint $\sum_k a_k^{(1)2}=1$.

In order to perform the second order solution, we consider the residuals of the orthogonal projections $X_k^{(1)}$ (respectively $Y^{(1)}$) of X_k (respectively Y) onto the subspace spanned by the first global component $t^{(1)}$ (Westerhuis, 1997). $X^{(1)}$ is defined as $X^{(1)}=[X_1^{(1)}|\ldots|X_K^{(1)}]$. The optimal vector of loadings $w^{(2)}$ is given by the eigenvector of $(X^{(1)'}Y^{(1)}Y^{(1)'}X^{(1)})$ associated with

the largest eigenvalue. The components $u^{(2)}$ and $t_k^{(2)}$ are computed as previously. This process can be reiterated in order to find higher order solutions.

2.2 Proposition of a multiblock redundancy analysis

For the purpose of exploring and modelling the relationships between a single X and Y, we know that redundancy analysis (RA) and PLS regression are based on the same criterion to maximize associated with different norm constraints (Burnham et al., 1996; Bougeard et al., *In Press*). As $MPLS$ is a direct extension of PLS regression to the multiblock setting, we propose an extension of redundancy analysis in order to improve the fitting ability of $MPLS$. The new problem consists in maximizing the criterion (2).

$$cov^2(u^{(1)}, t^{(1)}) \quad \text{with} \quad t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)},\ u^{(1)} = Y v^{(1)}, \tag{2}$$
$$t_k^{(1)} = X_k w_k^{(1)}, \quad \textstyle\sum_{k=1}^{K} a_k^{(1)2} = 1, \quad ||t_k^{(1)}|| = ||v^{(1)}|| = 1$$

As previously, the method derives a global component $t^{(1)} = Xw^{(1)}$ oriented towards the explanation of Y, that sums up partial components $t_k^{(1)}$ for $k = (1, \ldots, K)$ respectively associated with the blocks X_k. In the case where there is only one block of explanatory variables ($K = 1$), multiblock redundancy analysis leads to RA.

The solution is given by $v^{(1)}$ the normalized eigenvector of the matrix $M_{MRA} = \sum_k Y' X_k (X_k' X_k)^{-1} X_k' Y$ associated with the largest eigenvalue $\lambda_{MRA}^{(1)}$ (Bougeard et al., 2007). The partial components $(t_1^{(1)}, \ldots, t_K^{(1)})$ are directly derived from the eigenvector $v^{(1)}$: $t_k^{(1)} = P_k u^{(1)} / ||P_k u^{(1)}||$, where $P_k = X_k (X_k' X_k)^{-1} X_k'$ is the projector onto the subspace spanned by the X_k variables. The coefficients $a_k^{(1)}$ are given by $a_k^{(1)} = cov(u^{(1)}, t_k^{(1)}) / \sqrt{\sum_l cov^2(u^{(1)}, t_l^{(1)})} = ||P_k u^{(1)}|| / \sqrt{\sum_l ||P_l u^{(1)}||^2}$. This implies that the global component $t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)} = \sum_k P_k u^{(1)} / \sqrt{\sum_k ||P_k u^{(1)}||^2}$.

We recall that the optimal solution of the maximization of the criterion (2) is based on the eigenvector of the matrix $M_{MRA} = \sum_k Y' X_k (X_k' X_k)^{-1} X_k' Y$. Because projectors $P_k = X_k (X_k' X_k)^{-1} X_k'$ are symmetric and idempotent, it follows that $M_{MRA} = \sum_k (P_k Y)' (P_k Y)$. From this point of view, MRA appears as a principal component analysis of the table obtained by the vertical concatenation of the projection of Y onto each subspace spanned by the X_k blocks.

Higher order solutions are obtained by considering the residuals of the orthogonal projections of X_k and Y onto the subspace spanned by the global components $(t^{(1)}, \ldots, t^{(h)})$, as it is previously described for multiblock PLS regression.

2.3 Continuum between multiblock *PLS* and multiblock redundancy analysis

It turns out that MRA and $MPLS$ regression are respectively based on the eigenstructure of matrices $M_{MRA} = \sum_k Y'X_k(X_k'X_k)^{-1}X_k'Y$ and $M_{MPLS} = Y'XX'Y = \sum_k Y'(X_k'X_k)Y$. Thus, it appears that $MPLS$ corresponds to a shrinkage of the matrices $(X_k'X_k)^{-1}$ towards the identity matrices I_{p_k} for $k = (1, \ldots, K)$. From this standpoint, we can adopt a gradual shrinkage of the matrices $(X_k'X_k)^{-1}$ towards I_{p_k} by considering a convex combination of these matrices. More precisely, we propose to determine the solution in such a way so as to maximize the criterion (3).

$$cov^2(u_\gamma^{(1)}, t_\gamma^{(1)}) \quad \text{with} \quad t_\gamma^{(1)} = \sum_{k=1}^{K} a_{k,\gamma}^{(1)} t_{k,\gamma}^{(1)},\ t_{k,\gamma}^{(1)} = X_k w_{k,\gamma}^{(1)},\ u_\gamma^{(1)} = Y v_\gamma^{(1)} \quad (3)$$
$$\textstyle\sum_{k=1}^{K} a_{k,\gamma}^{(1)2} = 1,\ ||v_\gamma^{(1)}|| = 1,\ \gamma||w_{k,\gamma}^{(1)}||^2 + (1-\gamma)||t_{k,\gamma}^{(1)}||^2 = 1,\ 0 \leq \gamma \leq 1$$

It is clear that the case ($\gamma = 0$) corresponds to MRA applied to the data sets $(Y, X_1, \ldots, X_K)$ whereas the case ($\gamma = 1$) corresponds to $MPLS$. We shall refer to this strategy of analysis as Multiblock Continuum Redundancy PLS regression ($MCR - PLS$). We prove that $v_\gamma^{(1)}$ is the eigenvector of the matrix $M_\gamma = \sum_k Y'X_k[(1-\gamma)X_k'X_k + \gamma I_{p_k}]^{-1}X_k'Y$ associated with the largest eigenvalue $\lambda_\gamma^{(1)}$.

As previously, higher order solutions are obtained by considering the residuals of the orthogonal projections of each X_k and Y onto the subspace spanned by the global components $(t_\gamma^{(1)}, \ldots, t_\gamma^{(h)})$. This set of components can be used for the purpose of exploring the relationships between $(X_1, \ldots, X_K)$ and Y.

2.4 Optimal number of components

For all the methods described, the prediction of Y can be obtained by regressing the Y variables onto the orthogonal global components $(t^{(1)}, \ldots, t^{(h)})$. Moreover, the components can be expressed as linear combinations of X. This leads to the model $Y = X[w^{*(1)}c^{(1)'} + \cdots + w^{*(h)}c^{(h)'}] + Y^{(h)}$, $Y^{(h)}$ being the matrix of residuals, and w^* and c which are defind as in PLS regression (Wold et al., 1983). From a practical point of view, the final model may be obtained by choosing the appropriate number h of components to be introduced in the model, and for the continuum approach the γ parameter can be selected by a validation technique such as cross-validation (Stone, 1974). This consists in splitting the whole dataset into two sets, namely a calibration set and a validation set. The calibration set is used to estimate the parameters of the model and the validation set is used to compute the root mean square error of validation ($RMSE_V$) which reflects the prediction ability of the model under consideration. Thereafter, this procedure is

repeated several times. For each number h of components to be introduced in the model, the optimal value of γ is determined by minimizing $RMSE_V$. Eventually, among all these models corresponding to the various values of h, the model with the smallest value of $RMSE_V$ is retained.

2.5 Properties

The sensitivity to multicollinearity can be reflected by the condition index (Belsley et al., 1980). The condition index η_k of matrix $(X_k'X_k)$ is the ratio of the largest eigenvalue $\lambda_k^{(1)}$ to the smallest eigenvalue $\lambda_k^{(p_k)}$ of matrix $(X_k'X_k)$. A large value of η_k flags the presence of multicollinearity among X_k which is likely to lead to an unstable model. The condition index of each matrix $[(1-\gamma)X_k'X_k + \gamma I_{p_k}]$ is $\eta_{k,\gamma} = [(1-\gamma)\lambda_k^{(1)} + \gamma]/[(1-\gamma_k)\lambda^{(p_k)} + \gamma]$ for $k = (1,\ldots,K)$. It is easy to prove, by considering its derivative with respect to γ, that each $\eta_{k,\gamma}$ decreases when γ increases. Within $MCR-PLS$, $MPLS$ corresponds to the smallest values of $\eta_{k,\gamma}$ whereas MRA corresponds to the largest ones. The γ parameter stands as a regularization parameter as it improves the conditioning of each matrix $(X_k'X_k)$.

3 Application

3.1 Epidemiological multiblock data and objectives

The data set consists in the measurements of several variables on 404 broiler chicken flocks that were studied during farming, transport and at slaughterhouse (Lupo et al., *In press*). The Y table to be explained contains two variables which refer to the official reasons for the condemnation at slaughterhouse grouped in two classes, *i.e.* infectious ($INFECT$) or traumatic ($TRAUMA$) reasons. The X explanatory table is organized in three blocks. Table X_1 contains 26 variables pertaining to the farming features. Table X_2 contains 11 variables which refer to the transport condition between farm and slaughterhouse. Table X_3 contains 4 variables pertaining to the slaughtering conditions. The condition index performed on each explanatory table flags the presence of multicolinearity among X_1. Indicator (dummy) variables are considered for the categorical variables. Variables are column centred and scaled to unit variance. The two aims of the statistical analysis are to describe which variables differentiate the broiler chicken flocks and to assess the risk factors for the cases of condemnation.

3.2 Illustration of the properties of the continuum approach

In a first stage, we focus on the first component $t_\gamma^{(1)}$. The aim is to exhibit some of its properties and to pinpoint the role of the tuning parameter γ. When parameter γ varies from 0 to 1, we depict in Figure 1 the evolution of

the variance of $t_\gamma^{(1)}$, the variance of Y explained by $t_\gamma^{(1)}$ and the norm of the regression estimator length $||\beta_\gamma^{(1)}||$. The curves indicate that we are moving away from directions in the X space which may reflect noise only (small variance) and investigating new directions which are more linked to the Y variables (relatively large explained variance). The last curve indicates that the norm of the matrix of estimated coefficients is gradually shrunk when γ increases.

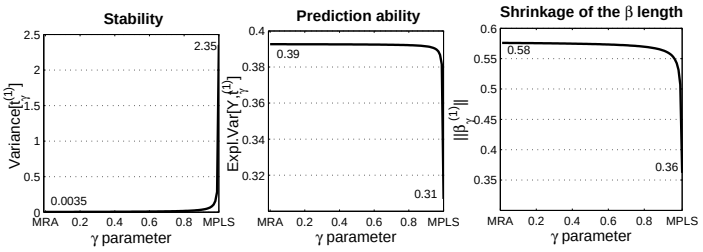


Fig. 1. Properties of the first dimension of the continuum $MCR - PLS$.

3.3 Graphical displays

The relationships between (X_1, X_2, X_3) and Y can be highlighted using the components $(t_\gamma^{(1)}, \ldots, t_\gamma^{(h)})$. The graphical displays in Figure 2 depict the loadings associated with the first two components $t_\gamma^{(1)}$ and $t_\gamma^{(2)}$ for MRA ($\gamma = 0$) and $MPLS$ regression ($\gamma = 1$). It highlights the relationships among the explanatory variables from X_1, X_2 and X_3, and makes it possible to identify some risk factors associated with the condemnation reasons (Y table).

3.4 Explanation and prediction from multiblock data sets

The choice of the optimal model (*i.e.* appropriate number of components and γ parameter) is a compromise achieved by both minimizing the root

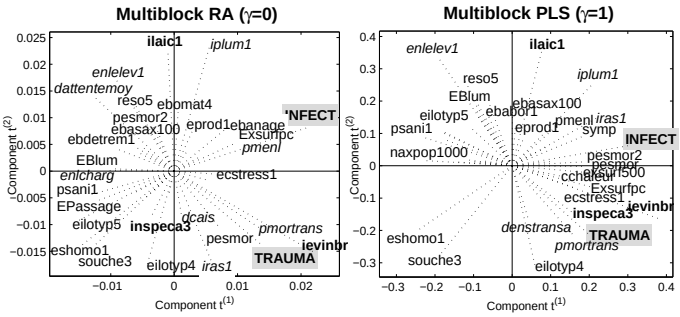


Fig. 2. Plots of the variable loadings associated with the first two components, for multiblock redundancy analysis (MRA) and multiblock PLS regression. 13 (resp. 14) variables that were not deemed important for the interpretation of the MRA (resp. $MPLS$) graphical display were discarded from the plot (although these variables were included in the analysis). Y variables are bold with a grey background, X_2 variables are slanted and X_3 variables are bold.

mean square error of calibration ($RMSE_C$) and validation ($RMSE_V$), which respectively reflect the fitting ability and the prediction ability of the model under consideration. The cross-validation procedure is repeated ($m = 200$) times and the γ value varies from 0 to 1 with an increment of 0.01. The optimal values of the tuning parameter γ for $MCR - PLS$ are chosen by cross-validation, by leaving one third of the calibration data set out, with minimizing $RMSE_V$. We undertake a comparison of $MCR-PLS$, MRA and $MPLS$ regression on the basis of the $RMSE_C$ and $RMSE_V$ criteria. Figure 3 shows the $RMSE_C$ and the $RMSE_V$ criteria as functions of the number h of components $(t^{(1)}, \ldots, t^{(h)})$ introduced in the model. It can be seen in Figure 3 that $MCR - PLS$, MRA and $MPLS$ have a comparable fitting ability, although MRA outperforms the other methods for a reduced number of components. On this case study, the continuum approach $MCR - PLS$ outperforms MRA and $MPLS$ especially for the first four dimensions.

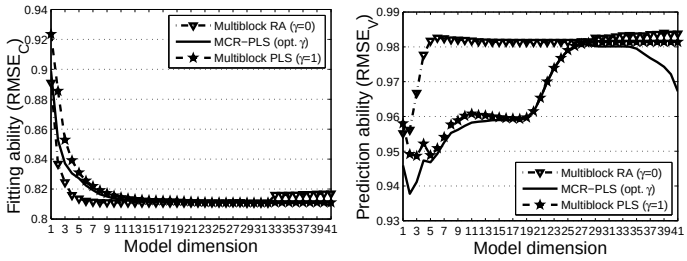


Fig. 3. Fitting ability ($RMSE_C$) and prediction ability ($RMSE_V$), as a function of the number of components introduced in the model. Comparison of $MCR-PLS$ (optimal tuning parameter), MRA ($\gamma = 0$) and $MPLS$ ($\gamma = 1$).

4 Concluding remarks

For the purpose of exploring and modelling the relationships between one block of variables Y to be explained by several blocks of explanatory variables $(X_1, \ldots, X_K)$, we propose an extension of redundancy analysis in order to improve the fitting ability of multiblock PLS regression. As we show that $MPLS$ and MRA are based on the same criterion to maximize associated with different norm constraints, we investigate a continuum approach, called Multiblock Continuum Redundancy PLS regression ($MCR-PLS$). The key feature of this approach is the shrinkage of the variance-covariance matrices $(X_k'X_k)$ towards the identity matrices, which leads to explore a continuum of methods ranging between MRA and $MPLS$. Moreover, the continuum is easy to understand and implement because the solutions are derived from an eigenanalysis of a matrix. However, it is clear that this is done at the cost of introducing a new parameter which needs to be customized to the data at hand. Obviously, there is a connection between the appropriate tuning parameter and the number of components to be introduced in the model.

We prove that the γ parameter stands as a regularization parameter as it improves the conditioning of each matrix $(X_k'X_k)$. From the case study discussed in this paper, we illustrate that $MCR-PLS$ stands as a compromise between having a model with a good fitting ability (MRA for $\gamma = 0$) and a

stable model ($MPLS$ for $\gamma = 1$). The properties pinpoint the idea that when γ increases, the strategy of analysis investigates more stable directions in the X space, but the fitting ability of the model decreases. From the case study, we found that $MCR - PLS$ outperforms $MPLS$ and MRA. A large study based on simulations should be undertaken in order to assess the performance of the various methods.

References

BELSLEY, D.A., KUH, E. and WELSCH, R.E. (1980): *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley (Ed.).

BOUGEARD, S., HANAFI, M. and QANNARI, E.M. (*In Press*): Continuum redundancy PLS regression: a simple continuum approach. Application to epidemiological. *Computational Statistics and Data Analysis*.

BOUGEARD, S., HANAFI, M. and QANNARI, E.M. (2007): $ACPVI$ multibloc. Application à des données d'épidémiologie animale. *Journal de la Société Française de Statistique*, 148: 77–94.

BURNHAM, A.J., VIVEROS, R. and MACGREGOR, J.F. (1996): Framework for latent variable multivariate regression. *Journal of Chemometrics*, 10: 31-45.

LUPO, C., CHAUVIN, C., BALAINE, L., PETETIN, I., PERASTE, J., COLIN, P. and LE BOUQUIN, S. (*In Press*): Post mortem condemnation of processed broiler chickens in Western France. *The Veterinary Record*.

QIN, S.J., VALLE, S. and PIOVOSO, M.J. (2001): On unifying multiblock analysis with application to decentralized process monitoring. *Journal of Chemometrics*, 15: 715–742.

RAO, C.R. (1964): *The use and interpretation of principal component analysis in applied research*. Sankhya, A., 26: 329–358.

STONE, M. (1974): Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36: 111-147.

VAN DEN WOLLENBERG, A. (1977): Redundancy analysis: an alternative for canonical correlation analysis. *Psychometrika*, 42: 207–219.

VIVIEN, M. (2002): *Approches PLS Linéaires et Non-linéaires pour la Modélisation de Multi-tableaux : Théorie et Applications. PhD* Thesis, University of Montpellier 1.

WANGEN, L.E. and KOWALSKI, B.R. (1988): A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, 3: 3–20.

WESTERHUIS, J.A., KOURTI, T. and MACGREGOR, J.F. (1998): Analysis of multiblock and hierarchical PCA and PLS model. *Journal of Chemometrics*, 12: 301–321.

WESTERHUIS, J.A. and COENEGRACHT, P.M.J. (1997): Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics*, 11: 379–392.

WOLD, H. (1966): Estimation of Principal Components and Related Models by Iterative Least Squares. *Multivariate analysis*, Krishnaiah (Ed.), Academic press, New York, 391–420.

WOLD, S. (1984): Three PLS algorithms according to SW. *Symposium $MULDAST$ (Multivariate analysis in science and technology)*, Umea University, Sweden, 26–30.

WOLD, S., MARTENS, H. and WOLD, H. (1983): The multivariate calibration problem in chemistry solved by the *PLS* method. *Proceedings of the Conference on Matrix Pencils*, Lecture Notes in Mathematics, Heidelberg, Springer Verlag, 286-293.

Free Model for Generalized Path Modelling and Comparison with Bayesian Networks

Christian Derquenne

Electricité de France, Research and Development
1, avenue du Général de Gaulle, 92141 Clamart Cedex, France,
christian.derquenne@edf.fr

Abstract. This paper introduces a new method to build a graphical model of categorical variables (a Free model) in the frame of structural equation models. Firstly a clustering of variables is applied, then for each cluster a numeric latent variable is calculated. After that, links between latent variables are searched for and expert decision issued to position these links. Finally, this Free model is estimated by the Regression on First Generalized Principal Components approach (RFGPC). We have applied this new approach on real data in marketing, listing the advantages and drawbacks of this approach with respect to others, notably the Bayesian networks. Lastly, we discuss on potential researches.

Keywords: graphical model, Bayesian networks, structural equation models, categorical data, marketing application

1 Introduction

The frame applications (marketing, econometrics, psychology,) of this paper is relative to the actors' behavior. Firstly this complex behavior is modeled theoretically by experts. The aim of such conceptual models is to represent the "reality". These must be validated in order to apply them. Usually, structural equation models are used to estimate these expert models. The structural relations are defined by "causal" links between the unobservable variables, named latent variables and the observed variables called manifest variables. For instance, the latent variables correspond to the aspects or components of customers' satisfaction: the manifest variables relate to the customers' answers of satisfaction survey questionnaires. The links study between the variables is based on expert marketing models coming from the decision theory process of customers. The choice of the model is essential, because it conditions the analysis of relations. The two statistical methods traditionally used to estimate these models are LISREL - LInear Structural RELationships (Jöreskog and Sörbum (1979)) and the PLS approach - Partial Least Squares (Wold (1982)). Another approach named RFPC - Regression on First Principal Components has been introduced (Derquenne and Hallais (2004)). The two latter methods have one main advantage with respect to the first one: to converge without difficulty in the step of the estimation of model.

These models are fixed at four main levels: choice of the manifest variables corresponding to questions of the survey, choice of blocks constituting latent variables, presence or absence of links between blocks, and positioned links. In this case, the confirmatory statistical methods are used. However, if we consider that the first level is imposed, the two others can be examinated. We can let the data talk and use an exploratory analysis. In this frame, we have developed an approach to build "Free models" (Derquenne and Hallais (2004), Jakobowicz and Derquenne (2007)), when the manifest variables are measured on numeric scale. Such approach is based on four steps: clustering of manifest variables with oblique rotation in Factor Analysis to build the blocks, selection of the first principal component of Principal Component Analysis (PCA) to obtain one latent variable by block, searching the links between latent variables with partial linear correlation, expert decision to position these previous links. Finally, this Free model is estimated by the RFPC approach, or by the PLS approach. The goal of this paper is to generalize this approach to other scales of manifest variable (binary, ordinal, nominal, ...). We apply this new approach on real data in marketing, listing the advantages and drawbacks of this approach with respect to others, notably the Bayesian networks. Lastly, we discuss on potential researches.

2 Generalized Free models approach : an overview of the four steps

2.1 Step 1: Clustering variables

This step builds links between the manifest variables, in order to obtain the external model. Let $(X_1, ..., X_j, ..., X_p)$ be p manifest variables measured on different scales. Each variable is modelled by the $p-1$ others, through the Generalized Linear Models. For instance, if X_j is nominal dependent variable with R_j categories, then a multinomial logit model is preferred:

$$\Pr[X_j = r_j / X_k, k = 1, ..., p, k \neq j] = \frac{\exp(\alpha_{r_j} + \sum_{k=1}^{p} x_k \beta_k^{(r_j)})}{1 + \sum_{r_j=1}^{R_j - 1} \exp(\alpha_{r_j} + \sum_{k=1}^{p} x_k \beta_k^{(r_j)})} \quad (1)$$

Usual marginal test of likelihood ratio is used to evaluate the contribution of each explanatory variable X_k $(k = 1, ..., p, k \neq j)$ to X_j, such as:

$$\Lambda(X_j / X_k) = -2[l(\hat{\beta}_{(k)}, X_j) - l(\hat{\beta}, X_j)] \quad (2)$$

where $\hat{\beta}$ is a vector of the parameters with dimension m associated to the full model and $\hat{\beta}_{(k)}$ is a vector of the parameters (dim=$m - m_k$) without X_k. Under the null hypothesis (non significant contribution of X_k) this statistic

is distributed as χ^2 with $\nu_k = (R_j - 1)m_k$ degrees of freedom (number of estimated parameters for X_k) and we obtain a p-value: p-val(X_j/X_k)=Pr$[\chi^2_{\nu_k} > \Lambda_{obs}(X_j/X_k)]$. For each variable X_j there are the $(p-1)$ p-values corresponding to $(p-1)$ explanatory variables. In the end, we have $p(p-1)$, p-values associated to all the couples of variables (X_j, X_k). However, each couple has two p-values: p-val(X_j/X_k) and p-val(X_k/X_j). Then the minimum of these values is chosen in order to evaluate the dissimilarity between X_j and X_k. With this rule, we can build a matrix of dissimilarities between the p variables. Another way to calculate the dissimilarities consists in using a stepwise regression to obtain the p-values. In both cases, dissimilarity is noted : $d_{j,k}$. The number of clusters of variables is chosen with the Ward's hierarchical ascending clustering on the dissimilarities-matrix. The simplest rule cuts the clustering tree where there is a big loss of inter-cluster inertia. However, this rule can be improved with a cutoff depending on a significant level α of a previous test, for instance $\alpha = 0.01$ or $\alpha = 0.05$. Indeed, the Ward's distance minimizes the lost of between-clusters inertia from G clusters to $G-1$ clusters. The Ward's distance between two clusters c_s and c_l is as follows :

$$D_W(c_s, c_l) = \frac{(p_s + p_t)D_W(c_s, c_t) + (p_s + p_u)D_W(c_s, c_u) - p_s D_W(c_t, c_u)}{p_s + p_t + p_u} \tag{3}$$

where $c_l = c_t \bigcup c_u$ and p_s, p_t and p_u are the associated numbers of variables in each cluster.

For G clusters, within-cluster inertia W_G corresponds to the sum of agregated Ward's distance. The proportion of within-cluster inertia $R^2_w(G)$ is equal to ratio: W_G/T, where $T = \sum_{j,k} d^2_{j,k}/p^2$ is the total inertia. Then the cutoff rule is given by: if $R^2_w(G) > 1 - \frac{\alpha^2(p-G)}{T}$ then G clusters can be selected. Indeed, second term in the inequality is obtained in replacing each $D_W(c_s, c_t)$ by α^2 in (3). The advantage of this rule consists in taking into account a significant level of test and gives a good estimation of maximum number of clusters.

2.2 Step 2: Building the latent variables

Each cluster contains p_g variables which can be measured on different scales. It is not reasonable to apply a Principal Components Analysis to each block of manifest variables as in Derquenne and Hallais (2004), because they are not all on a numeric scale. Hence, we use the Generalized Factorial Analysis (GFA) (Derquenne (2005)) developed in the framework of generalized path modelling. This technique deals with missing data according to the same principle as the NIPALS algorithm (Wold (1973)). The categorical NIPALS has been developed when a block contains only binary or ordinal variables and missing data. This method is based on binary logit or odds proportional logit

model in the first step and on ANOVA model in the second step. Furthermore, if there is no missing data, this method is exactly a MCA (Multiple Correspondence Analysis), as NIPALS is a PCA. Lastly, a generalized NIPALS method has been developed to the nominal, count and mixed variables in each block and to missing data. This method is very powerful because it allows treating all type of variables. Finally, we obtain G numeric latent variables, $(Z_1, ..., Z_G)$. The following table gives the methods to apply in case of different types of variables and if there are missing data or not.

	Numeric	**Binary or ordinal**	**Nominal**	**Count**	**Mixed**
Not missing	PCA	MCA	MCA	Generalized NIPALS	Generalized NIPALS
Missing	NIPALS	Categorical NIPALS	Categorical NIPALS	Generalized NIPALS	Generalized NIPALS

Table 1 : Results of link between latent variables.

2.3 Step 3: Searching for links between latent variables

As all the G latent variables are on a numeric scale, we calculate the partial linear correlation between each couple (Z_s, Z_q) giving the $G-2$ other Z_j: $\rho(Z_s, Z_q/\forall Z_j, j \neq s, q)$. The associated test statistic follows a Student's distribution with $n-G$ degrees of freedom under H_0 (non-correlation between Z_s and Z_q). The link between two latent variables is considered significant if the p-value is lower than a fixed threshold. This procedure leads to $K \in [0, G(G-1)/2]$ links. The final result is a non-directed graph.

2.4 Step 4: Position of links between latent variables

It is usually preferable to use an expert decision to avoid a non efficient model. However, different statistical solutions can be proposed. Indeed, if there are K significant links, then the number of different models is equal to 2^K. Lastly, each model can be evaluated through a score, (GoF, global R^2, ...), then an expert can choose among the ten highest scores. However, these models must not have cycles in the directed graph.

2.5 Estimation of the Free model

The Free model is estimated through the RFGPC approach (Regression on the First Generalized Principal Components) introduced in Derquenne (2005), in the framework of generalized path modelling. The RFPC approach allows estimating the measurement (external), without taking into account

the structural (internal) model. Indeed, each block of manifest variables is estimated by the first principal component of PCA, if there is no missing data, otherwise NIPALS is used. As in PLS approach, the latent variables are built in two main steps. The first step for each block consists in building an initial latent variable, with aid of a reference manifest variable into this block, whereas in the PLS approach, the reference variables are into the blocks connected to the block to estimated. The second step corresponds to an iterative process until convergence of each block. In this case, the structural model is only estimated by linear multiple regressions between endogenous and exogenous latent variables, as in PLS approach.

As in the PML approach (Derquenne (2005)) which generalizes the PLS approach, the RFGPC approach generalizes the RFPC approach. Indeed, this method allows building a latent variable with different types of manifest variables. Then the both main steps are modified in function of the type of variables. The internal model is estimated by means of multiple linear regression between latent variables and the external model is estimated by means regression between each latent variable and its own manifest variables.

3 Application to marketing data

In this section, we build our Free model on marketing data and we compare its to a Bayesian network. Briefly, Bayesian networks are represented in graphical form where the nodes correspond to the model's variables and the edges represent direct probability relations. They are powerful tools for non-supervised learning (Jensen (2001)) and can be used to find links between observed variables in order to create a robust outer model. Customer satisfaction surveys are handled using questions with nominal or ordinal scales; these questions can be directly processed as variables using structure learning heuristics based on network scores (minimum description length) and a search based on a global characterization of the data and on the exploitation of the properties of equivalent Bayesian networks (i.e. the networks representing the same set of dependences or independences) with SopLEQ algorithm (Jouffe (2004)). Then a segmentation of variables with aid of this graph is applied to obtain clusters of nodes (manifest variables). This step corresponds to clustering variables (step 1) in our process of building the Free model. For each cluster, a categorical latent variable is built by clustering the individuals, as a new node (step 2) and the external model is created. We apply SopLEQ again on these new nodes, then new relations are found (step 3). To build the structural model, the position of links (step 4) is directly given by SopLEQ, but a different position of links can be chosen by experts among equivalent Bayesian networks. Lastly, the estimation of Free model is given by making inference on full model (external and internal models) in selecting a target node among latent variables.

The marketing Department would like to evaluate the leverage of loyalty and satisfaction with respect to central heating. But a specific marketing conceptual model is not available. However a satisfaction and loyalty survey has been conducted, but the questionnaire was not explicitly built to be used for a path modelling. This department asks the R&D to construct a first statistical model as an initial marketing model. We have applied our Free model approach on these data. It is composed of several parts including socio-economics questions (age, type of household, ...) and specific questions (satisfaction, expectations, loyalty, perceived value, ...). The answers are distributed on semantic scales (binary, ordinal and nominal): 6 binary variables, 5 nominal variables and 20 ordinal variables. We have selected 6899 questionnaires. Some contain missing data.

We obtain 8 variables clusters: expectation #1 (EXPEC#1: p=5 variables), expectation #2 (EXPEC#2: p=4), socio-demographics (SOC-DEM: p=4), housing (HOUSING: p=4), customer's contract (CONTRA: p=2), loyalty (LOYALTY: p=2), satisfaction/recommandation (SAT-REC: p=3) and perceived value (PER-VAL: p=7). The contents of each group is logical enough given the initial variables. Then, for each cluster, a principal component has been estimated by means of GFA. 19 significant links coming from partial correlation (p-value$<$ 0.01) between latent variables have been found, then an expert has decided the position of the link (table 2). Lastly, we have applied the RFGPC approach on this Free model. The graph (figure 1) shows an interesting and significant structure. Indeed, a target (LOYALTY) appears clearly and corresponds to 2 questions (modification and choice of future product), satisfaction and recommandation (SAT-REC: satisfaction, advise to friends and settlement date) show a high influence on loyalty (t-value = -30.6). On the other hand, the perceived value (PER-VAL) and customer contract (CONTRA) act upon PER-VAL. In addition, the first expectation block has a strong correlation with the second (t-value = 42.8). Perhaps, these two clusters could be merged. Finally exogenous block socio-demographics (SOC-DEM: chidren attendance, age, electric comsumption and tariff) have a stronger influence housing structure (HOUSING: flat/house, tenant/householder, electric heating or not and fireplace or not). In the Bayesian network, only five clusters are discovered: socio-demographics and housing (SOC-DEM-HOU: p=8), customer's contract (CONTRA: p=2), expectations (EXPECT p=9), perceived value (PER-VAL: p=7) and, loyalty and satisfaction (SAT-LOY p=5). But these clusters are very closed to Free model clusters. Indeed, SOC-DEM-HOU contains exactly the both clusters SOC-DEM and HOUSING of our Free model, CONTRA and PER-VAL are the same, EXPECT owns EXPEC#1 and EXPEC#2, lastly SAT-LOY includes SAT-REC and LOYALTY. In the Bayesian network, the links are similar too. There is a very high relation (Kullback Leibler's distance) between target loyalty and the socio-demographics dimension (40.6%), and another with perceived value (28.6%). Then the expectations play a role too on per-

ceived value. Lastly, a link between socio-demographics and services has been found. The position of links has been decided by a marketing expert.

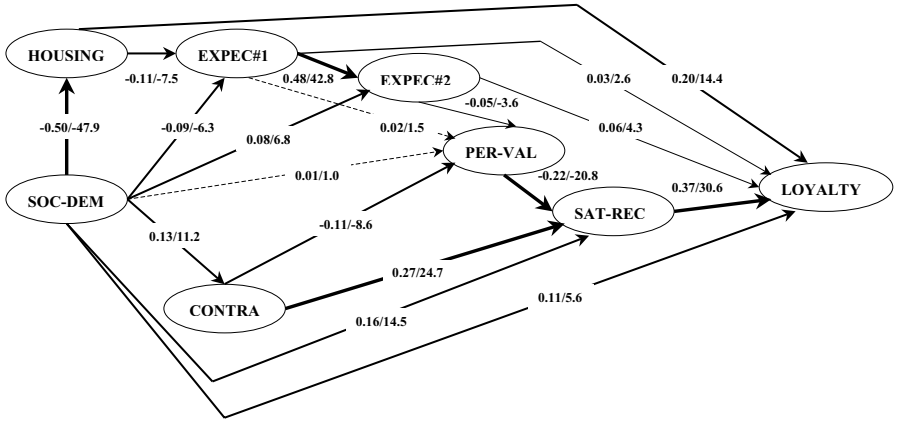
Couple of latent variables	**Partial correlation**	p**-value**	**Position of links**
(EXPEC#1,EXPEC#2)	+0.6311	< 0.0001	EXPEC#1→EXPEC#2
(LOYALTY,SAT-REC)	+0.5024	< 0.0001	SAT-REC→LOYALTY
(SOC-DEM,HOUSING)	+0.4004	< 0.0001	SOC-DEM→HOUSING
(SAT-REC,PER-VAL)	-0.2560	< 0.0001	PER-VAL→SAT-REC
(SAT-REC,HOUSING)	+0.1555	< 0.0001	HOUSING→SAT-REC
(EXPEC#1,SOC-DEM)	-0.1449	< 0.0001	SOC-DEM→EXPEC#1
(LOYALTY,HOUSING)	-0.1290	< 0.0001	HOUSING→LOYALTY
(SOC-DEM,LOYALTY)	+0.1217	< 0.0001	SOC-DEM→LOYALTY
(CONTRA,SAT-REC)	+0.1022	< 0.0001	CONTRA→SAT-REC
(EXPEC#2,SOC-DEM)	+0.1005	< 0.0001	SOC-DEM→EXPEC#2
(SOC-DEM,PER-VAL)	+0.0893	< 0.0001	SOC-DEM→OPINION
(SOC-DEM,SAT-REC)	+0.0828	< 0.0001	SOC-DEM→SAT-REC
(EXPEC#1,LOYALTY)	-0.0735	< 0.0001	EXPEC#1→LOYALTY
(EXPEC#1,HOUSING)	+0.0697	< 0.0001	EXPEC#1→HOU-EDF
(EXPEC#2,OPINION)	+0.0649	< 0.0001	EXPEC#2→OPINION
(EXPEC#2,LOYALTY)	+0.0471	0.0001	EXPEC#2→LOYALTY
(CONTRA,SOC-DEM)	+0.0469	0.0001	SOC-DEM→CONTRA
(CONTRA,PER-VAL)	-0.0394	0.0011	CONTRA→PER-VAL
(EXPEC#1,PER-VAL)	-0.0353	0.0034	EXPEC#1→PER-VAL

Table 2 : Results of link between latent variables.

Fig. 1. Free model estimated by RFGPC Approach.

4 Contribution, applications and future works

Our Free model approach lets the data talk and uses an exploratory data analysis. This model includes the external and internal models as a classical structural equations model, and has the big advantage to work with variables measured on mixed scales. In addition, it stays in classical statistical inference



to evaluate the links between latent variables. Other methods for creating models exist: stepwise structural equation modelling (Hui (1982)) or Bayesian networks. The first only builds an internal model for variables on numeric scale, whereas the second only makes an external model for semantic scales. But we have introduced a solution to build structural model with Bayesian network. However statistical inference on full model (external and internal) is poor enough compared to our approach. In addition our Free model approach can easily be applied in marketing, chemiometrics, policy, human capital, ... Future researches concern the improvement of the clustering step, for example by using immediately the GFA method and to improve the inference tools in Bayesian networks.

References

DERQUENNE, Ch. and HALLAIS, C. (2004): Une méthode alternative l'approche PLS : comparaison et application aux modéles conceptuels marketing. *Revue de Statistique Appliquée*. LII, 37-72. Paris, France.

DERQUENNE, Ch. (2005): Generalized Path Modelling based on the Partial Maximum Likelihood Approach, PLS'05, *Fourth International Symposium on PLS and Related Methods*. Barcelona, Spain.

JAKOBOWICZ, E. and DERQUENNE, C. (2007): A modified PLS path modeling algorithm handling reflective categorical variables and a new model building strategy. *Computational Statistics & Data Analysis*, 51(8): 3666-3678, Elsevier.

JENSEN, F.V. (2001): *Bayesian Networks and Decision Graphs*. Springer, New-York.

JÖRESKOG, K.G. and SÖRBUM, D. (1979): *Advanced in Factor Analysis and Structural Equation Models*. Abt Books, Cambridge.

JOUFFE, L. (2004): Bayesialab: the decision support and data mining tools, *ISBA Bulletin*. Vol 11(4), 7-10.

HUI, B.S. (1982): On Building Partial Least Squares Models with Interdependent Inner Relations, In: K.G. Jöreskog & H. Wold (eds.), *Systems under Indirect*

Observations: Causality, Structure, Prediction. Vol 2, 249-271, Amsterdam, North-Holland.

WOLD, H. (1973): Nonlinear Iterative Partial Least Squares (NIPALS) Modelling some Current Developments, In: Krishnaiah P.R. Editor, *Multivariate Analysis*, III, 391-420, Academic Press, New-York.

WOLD, H. (1982): Soft modeling: the basic design and some extensions, In: K.G. Jöreskog & H. Wold (eds.), *Systems under Indirect Observations: Causality, Structure, Prediction*. Vol 2, 1-54, Amsterdam, North-Holland.

A Robust Method Applied to Structural Equation Modeling

Alina Matei[1] and Petroula Mavrikiou[2]

[1] Institute of Statistics, University of Neuchâtel,
Pierre à Mazel 7, 2000 Neuchâtel, Switzerland, *alina.matei@unine.ch*
[2] Department of Business Administration, School of Economic Sciences and Administration, Frederick University,
P.O. Box 24729, 1303 Nicosia, Cyprus, *bus.mp@fit.ac.cy*

Abstract. In structural equation modeling, the presence of outliers in data can affect the covariance matrix and the estimates. We develop a method for handling outliers based on the BACON algorithm, Blocked Adaptive Computationally-efficient Outliers Nominators, see Billor et al. (2000). The latter is a very efficient outlier detection method in multivariate data with elliptical distribution. The method also provides a robust covariance matrix. Replacing the ordinary sample covariance by the BACON covariance matrix is advantageous when outliers are influential. We study the benefits of the BACON approach in terms of robustness in structural equation modeling. The proposed method is exemplified using simulated and real data.

Keywords: outliers, robust covariance, structural equation modeling

1 Introduction

Structural equation modeling (SEM) is widely used in social and behavioral sciences for exploring how latent variables are related to the observed ones. Using SEM, one can test whether variables are interrelated through a set of linear relationships by examining their covariances and variances. Classical methods for SEM are characterized by the assumption of multivariate normal data; see Bollen (1989). If the observed variables are multivariate normal, one obtains maximum likelihood (ML) estimates. The ML method is the most commonly used approach. Inference and departures from the normality of the data have been studied and asymptotically distribution free procedures have been also proposed in the literature; see Browne (1984).

In practice, data may often contain outliers. Most of the multivariate analysis methods assume the normality of the data and use estimates of the location and scale parameters of the distribution. Outliers can distort the values of these estimates and thus affect the results of these techniques. In structural equation models, according to Bollen (1989), page 31: 'outliers can increase, decrease, or have no effect on covariances and on the estimates based on these covariances.' Outliers can also lead to heavy tailed distributions, and

thus the assumption of multivariate normal data is not appropriate. According to Bollen and Arminger (1991), the potential causes of the outliers in SEM are many: 'coding errors, incorrect functional form, omitted variables, structural shifts, etc.' The literature in the detection of the outliers and construction of the robust covariance matrix estimation for multivariate data is rich. In SEM, contributions on this subject are due to Huba and Harlow (1987), Yuan and Bentler (1998), Lee and Xia (2006), etc.

The BACON (Blocked Adaptive Computationally-efficient Outliers Nominators) algorithm of Billor et al. (2000) is a method for handling outliers in multivariate analysis with elliptical distribution. We propose the use of the BACON algorithm to construct and apply a robust covariance matrix in SEM. The objective is to obtain accurate estimates, by downweighting the effect of the outliers.

2 Structural equation models with latent variables

The first component of the general structural equation model is the latent variable model, where the relation between the latent variables η_i and ξ_i is performed by regressing the dependent vector η_i on the explanatory vector ξ_i as follows

$$\eta_i = \alpha + \mathbf{B}\eta_i + \mathbf{\Gamma}\xi_i + \varsigma_i, \tag{1}$$

for $i = 1, \ldots, n$ observations. The matrix $\mathbf{B}_{m\times m}$ describes the relationships among latent variables in η_i and has its diagonal elements zero. The $m \times n$ matrix $\mathbf{\Gamma}$ quantifies the influence of ξ_i on η_i. The $m \times 1$ vectors α and ς_i represent the intercept and the unexplained parts of η_i, respectively.

The second component of the general structural equation model is the following measurement model:

$$\mathbf{y}_i = \mu_y + \mathbf{\Lambda}_y\eta_i + \delta_i^y; \tag{2}$$

$$\mathbf{x}_i = \mu_x + \mathbf{\Lambda}_x\xi_i + \delta_i^x, \tag{3}$$

where model (3) relates the vector of indicators $\mathbf{y}_i = (y_{i1}, .., y_{ip})'$ to the vector of latent variables $\eta_i = (\eta_{i1}, .., \eta_{im})', m \leq p$, through the $p \times m$ factor loadings matrix $\mathbf{\Lambda}_y$. Similarly, model (5) relates $\mathbf{x}_i = (x_{i1}, .., x_{iq})'$ to the vector of latent variables $\xi_i = (\xi_{i1}, .., \xi_{in})'$, $n \leq q$, through the $q \times n$ matrix $\mathbf{\Lambda}_x$. The vectors δ_i^y and δ_i^x are the errors of measurement for $\mathbf{y}_i$ and $\mathbf{x}_i$, respectively, and are assumed uncorrelated with η_i and ξ_i. The vectors μ_y, $p \times 1$, and μ_x , $q \times 1$, are the intercept terms of the measurement models. Common assumptions in SEM are: i) $\xi_i \sim N_n(\mu_\xi, \mathbf{\Omega}_\xi)$ and $\varsigma_i \sim N_m(\mathbf{0}, \mathbf{\Omega}_\varsigma)$ are independent; ii) $\delta_i^y \sim N_p(\mathbf{0}, \mathbf{\Sigma}_y)$ and $\delta_i^x \sim N_q(\mathbf{0}, \mathbf{\Sigma}_x)$ are independent; iii) $\delta' = (\delta_i^{y'}, \delta_i^{x'}), Cov(\varsigma_i, \delta') = 0, Cov(\xi_i, \delta') = 0, Cov(\xi_i, \varsigma_i^{'}) = 0$, and iv) $(\mathbf{I} - \mathbf{B})$ is nonsingular. In the ML approach, the parameters are estimated by minimizing the function

$$F(\mathbf{S}, \mathbf{\Sigma}) = \log|\mathbf{\Sigma}| - \log|\mathbf{S}| + trace(\mathbf{S}\mathbf{\Sigma}^{-1}) - (p + q), \tag{4}$$

where $p+q$ is the number of the observed variables, $\mathbf{S}$ is the sample covariance matrix, and $\mathbf{\Sigma}$ is the expected covariance matrix. We replace the standard sample covariance matrix $\mathbf{S}$ in (4) by a robust covariance matrix. Thus we use a plug-in ML approach for analyzing the proposed method and we test its performances on simulated and real data. We assume that both observed and latent variables are continuous.

2.1 The BACON algorithm

Let $\mathbf{S}$ denote the ordinary sample covariance matrix among the observed variables computed directly from the data. The BACON algorithm for outlier detection in multivariate data uses as input a matrix $\mathbf{X}_{n\times r}$, with $r = p + q$ in our case. It starts by forming an initial basic subset of m observations that is presumably free from multivariate outliers. The initial subset can be either specified by the analyst (who can indicate the number m of the 'clean' observations) or by the algorithm, which computes $m = cn$. The value c is a constant, and it assures that at least c observations for each parameter are available. Simulations conducted in Billor et al. (2000) have shown that $c = 4$ or 5 perform quite well. The initial basic subset can be found algorithmically in one of the following two ways:

(i) using the Mahalanobis distance

$$d(\bar{\mathbf{x}}, \mathbf{S})_i = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})}, i = 1, \ldots, n, \tag{5}$$

where $\mathbf{x}_i$ is the i^{th} row of $\mathbf{X}$, and $\bar{\mathbf{x}}$ is the sample mean;

(ii) using the distances from the medians

$$d(\mathbf{m})_i = ||\mathbf{x}_i - \mathbf{m}||, i = 1, \ldots, n, \tag{6}$$

where $\mathbf{m}$ is a vector containing the coordinatewise medians, and $||.||$ is the vector norm.

The former case is not a robust one, but it is affine equivariant. The latter is robust, but it is not affine equivariant.

The BACON algorithm computes iteratively a basic subset of observations, starting with the initial one. Observations that are consistent with the current basic subset are added to it. The algorithm stops when the current basic subset does not change any more. If all the observations are added to the final basic subset, the data set is declared to be free from outliers; otherwise the observations that are not consistent with the final basic subset are declared as multivariate outliers. As output, the BACON algorithm provides a set of observations nominated as outliers and the discrepancies for all observations relative to the final basic subset. The discrepancies (distances) used in the algorithm are based on the formula

$$d(\bar{\mathbf{x}}_b, \mathbf{S}_b)_i = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}_b)'\mathbf{S}_b^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_b)}, i = 1, \ldots, n, \tag{7}$$

where $\bar{\mathbf{x}}_b$ and $\mathbf{S}_b$ are the mean and covariance matrix of the observations in the current basic subset b. The distances at the final step can be used as robust distances.

2.2 Robust covariance using the BACON algorithm

There are two approaches to use the BACON algorithm for SEM to construct a robust covariance matrix. A first idea is to apply the BACON algorithm to detect outliers, which are eliminated from the data set. A robust covariance matrix is then constructed without taking into account these outliers. This corresponds to weighting observation by 0/1 values. The loadings are estimated, afterwards, using this robust covariance matrix. A second idea is to construct unequal weights associated to cases based on the distances provided by the BACON algorithm, and finally to construct a robust weighted covariance matrix. The outliers receive small weights and their effect on the loadings is controlled. This method is in contrast to the standard approach, which uses $\mathbf{S}$ and where the cases have equal weights. The BACON algorithm is used for two reasons: (a) it is effective in the detection of outliers and (b) it is computationally very efficient. The covariance matrix taking into account the BACON distances is defined as: $\mathbf{S}_{BACON} = \sum_{i=1}^{n} w_i(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' / \sum_{i=1}^{n} w_i$, where $\boldsymbol{\mu} = \sum_{i=1}^{n} w_i \mathbf{x}_i / \sum_{i=1}^{n} w_i$, and w_i are the weights computed as $1/d_i; d_i$ is the final distance provided by the BACON algorithm for the observation i, and given in expression (7). Attention should be paid since the weights w_i may be unstable. Note that $\mathbf{S}_{BACON}$ does not provide necessarily unbiased and consistent estimates when the data come from a multivariate normal distribution. Following Yuan and Bentler (1998), 'a robust estimator generally does not converge to the population covariance matrix; it converges instead to a constant times the population covariance matrix [..] However, the term - structural equation modeling with robust covariances - is still well defined as long as the structural model is invariant under a constant scaling factor (ICSF) [..] So, if the model is ICSF, the model that holds for a covariance matrix Σ also holds for a rescaled version of the covariance matrix $\alpha\Sigma$.'

3 Applications of the proposed method

We estimate two sets of parameters corresponding to the standard sample covariance and robust covariance matrix $\mathbf{S}_{BACON}$, respectively. The matrix $\mathbf{S}_{BACON}$ is computed using $c = 4$ and the initial subset is computed using the distance to the median given in expression (6). Departure of the estimated parameters from the true values is measured by the the root mean square error (RMSE) given by $\text{RMSE}_\beta = \left(\sum_{i=1}^{sim}(\widehat{\beta}_i - \beta)^2)/sim\right)^{1/2}$, where $\widehat{\beta}_i$ is the estimated parameter value in the i^{th} simulation, sim is the number of simulations, and β is the true parameter value. This quantity is computed on

the basis of 100 simulations. We also compute the relative efficiency (RE) of the BACON algorithm as the ratio between the RMSE of the BACON estimates and ordinary estimates, respectively. Values of RE between 0 and 0.5 indicates high performance of the BACON algorithm relative to the standard method; values close to one shows no improvement by using the BACON algorithm in SEM. In the following table, we denote by 'Standard RMSE' and 'BACON RMSE' the values of RMSE using the standard covariance matrix and the BACON one, respectively.

Examples

I) ***Simulated data***: Our simulation study is based on the simulation setting (I) given in Lee and Xia (2006). The known model is fitted, and the variables are generated from the normal distribution. There are three latent variables η, ξ_1 and ξ_2, related by the linear equation $\eta_i = \gamma_1\xi_{1i} + \gamma_2\xi_{2i} + \zeta_i$, where $\gamma_1 = \gamma_2 = 0.5$, and $i = 1, \ldots, 300$. The population parameter values in the measurement model are:

$$\mathbf{\Lambda}' = \begin{pmatrix} 1^* & 0.8 & 0.8 & 0^* & 0^* & 0^* & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 1^* & 0.8 & 0.8 & 0^* & 0^* & 0^* \\ 0^* & 0^* & 0^* & 0^* & 0^* & 0^* & 1^* & 0.8 & 0.8 \end{pmatrix}, \mathbf{\Omega}_\xi = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix},$$

$\Sigma = (\Sigma_y, \Sigma_x) = (\sigma_1, \ldots, \sigma_9) = diag(1, \ldots, 1)$ and $\mathbf{\Omega}_\zeta = (\omega_1) = 1$. The values with an asterisk are fixed. The other parameters are denoted by: $\lambda_{21}, \lambda_{31}, \lambda_{52}, \lambda_{62}, \lambda_{83}, \lambda_{93}$. We contaminate the data, letting $\zeta \sim N(5, 10)$ for some randomly selected observations among the 300 observations, which represent 5%, 10% and 15% of the data, respectively. This is done before calculating the covariance matrix and estimating the model. We repeat this for 100 samples that had convergent solutions. We apply the BACON algorithm for the first 3 observed variables, in order to compute a 3×3 robust covariance matrix. We expect that at least ω_1 should be protected using the BACON approach. For space reason only the results for 10% and 15% contamination are reported here in Table 1. The results for 0% contamination (not showed here) indicate similar behavior for the standard method and the BACON, with RE close to 1. This indicates that the BACON algorithm is consistent when outliers are not presented. For 5% contamination (results not showed here), the BACON algorithm performs very well for two parameters: ω_1 and σ_1. For these two cases, RE is equal to 0.288 and 0.335, respectively. For the other parameters, RE has a value close to 1. In Table 1, for 10% contamination, the BACON algorithm performs well for four parameters: $\omega_1, \sigma_1, \sigma_2$ and σ_3, with values of RE 0.250, 0.314, 0.584, 0.665, respectively. In Table 1, for 15% contamination, we emphasize the values of RE for $\omega_1, \sigma_1, \sigma_2$ and σ_3: 0.178, 0.417, 0.684, 0.423, respectively. The results for 20% contamination level and more (not shown here) indicate that the empirical breakdown point of the BACON algorithm in the SEM context is about 20%.

Parameter	Standard RMSE	BACON RMSE	RE	Standard RMSE	BACON RMSE	RE
	10% contamination			15% contamination		
$\lambda_{21}= 0.8$	0.292	0.282	0.967	0.289	0.297	1.027
$\lambda_{31}= 0.8$	0.290	0.283	0.976	0.300	0.296	0.985
$\lambda_{52}= 0.8$	0.323	0.287	0.889	0.451	0.291	0.645
$\lambda_{62}= 0.8$	0.330	0.288	0.875	0.366	0.290	0.793
$\lambda_{83}= 0.8$	0.288	0.291	1.011	0.315	0.294	0.936
$\lambda_{93}= 0.8$	0.291	0.284	0.977	0.328	0.301	0.916
$\gamma_1= 0.5$	0.326	0.238	0.731	0.454	0.307	0.677
$\gamma_2= 0.5$	0.341	0.244	0.715	0.421	0.287	0.681
$\omega_{11}= 0.3$	0.181	0.177	0.978	0.267	0.192	0.718
$\omega_{12}= 1.0$	0.373	0.327	0.877	0.468	0.343	0.731
$\omega_{22}= 1.0$	0.426	0.334	0.783	0.393	0.368	0.936
$\omega_1= 1.0$	1.890	0.473	0.250	3.591	0.639	0.178
$\sigma_1= 1.0$	1.230	0.386	0.314	2.130	0.888	0.417
$\sigma_2= 1.0$	0.454	0.265	0.584	0.445	0.304	0.684
$\sigma_3= 1.0$	0.393	0.261	0.665	0.694	0.293	0.423
$\sigma_4= 1.0$	0.366	0.315	0.859	0.428	0.327	0.765
$\sigma_5= 1.0$	0.334	0.321	0.962	0.336	0.321	0.956
$\sigma_6= 1.0$	0.327	0.319	0.975	0.353	0.322	0.914
$\sigma_7= 1.0$	0.366	0.322	0.880	0.360	0.319	0.887
$\sigma_8= 1.0$	0.343	0.316	0.920	0.317	0.321	1.012
$\sigma_9= 1.0$	0.339	0.319	0.940	0.323	0.331	1.025

Table 1. RMSE values computed using simulated normal data with 10% and 15% levels of contamination.

II) ***Real data***: The second example uses a real data set, the cloud cover data, described in Bollen (1989), p. 30–31. The data are three judges' estimates of the percentage of the sky that contains clouds of 60 photographic slides. The data is known to contain outliers, and it was previously analyzed in Bollen and Arminger (1991), Yuan and Bentler (1998). Let us consider the centered data. Applying the multivariate BACON algorithm on the cloud data results in the detection of 22 outlying observations. Cases 40, 52 and 51 are the most extreme outliers, as indicated by the BACON final robust distances. A plot of the latter versus index is given in Figure 1. The BACON algorithm turns out to be severe in this example. This is mainly due to the fact that the data set does not provide evidence of a joint elliptical distribu-

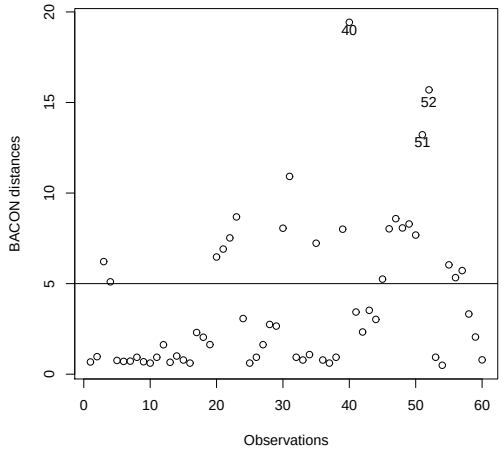


Fig. 1. The BACON distances for cloud data.

tion. Consequently, the χ^2 distances inside the BACON algorithm over reject the hypothesis that an observation i belongs to the basic subset. One way to deal with this problem might be a data transformation to achieve some kind of normality. Yet, another approach is to use the BACON distances (or a function of them) to form a weighted covariance estimate, which is more efficient in this example. Thus, the detected outliers are not removed from the data set, but their effect is downweighted using the inverse of the BACON distances. The observations 40, 52 and 51 are essentially neglected, since their weights (equal to 0.05, 0.064, 0.076, respectively) are small compared to the other weights. In Table 2, we give the results of the ML method on the whole data set, the ML results after removing observations 40, 52 and 51, as well as the BACON results. These three solutions are referred to as 'Standard', '3 outliers' and 'BACON', respectively. Parameters' estimates and their standard errors are denoted by 'Estimates' and 'SE', respectively. Note that no test statistic is available, since there are 0 degrees of freedom. In the ML case, the sample covariance based on all the 60 observations leads to an improper solution (a negative error variance). The estimates of the factor loadings and their standard errors are similar for the first two solutions, and the BACON solution preserves the order of their magnitude. Differences are visible for the error variance estimates. In the first solution, $\widehat{\sigma_3}$ is not significant at 5% level, but it is significant for the other two solutions. The BACON solution provides estimates with smaller standard errors than the first two solutions.

	Standard		3 outliers		BACON	
	Estimate	SE	Estimate	SE	Estimate	SE
λ_1	32.717	3.702	32.129	3.350	19.924	1.938
λ_2	31.608	4.105	36.774	3.879	22.338	2.479
λ_3	38.384	3.488	36.090	3.581	21.129	2.086
σ_1	249.492	61.185	105.357	28.427	19.788	9.502
σ_2	472.788	94.840	157.140	39.856	109.100	22.920
σ_3	-51.639	56.529	58.273	27.865	29.821	11.294

Table 2. Parameter estimates and their standard errors for cloud data.

4 Conclusions

We proposed a method for handling outliers and controlling their influence in structural equation modeling based on the BACON algorithm. Our experience based on simulated and real data indicates that the BACON algorithm provides a reliable tool for handling outliers in SEM. As we have seen, the BACON approach is superior in simulated normal data with contamination level less than 20%. As in most statistical methods, attention should be paid before the proposed method is applied on real data. By that we emphasize on assumptions related to the BACON algorithm and elliptical underlying distribution. If these assumptions are violated, the proposed method may not be appropriate, unless until some action is taken in advance. Nevertheless, the BACON approach seems to be robust in this context, since its final weights provide a tool to correct for departures from the underlying assumption as seen in the cloud data. For this reason, we sustain the idea of downweighting the detected outliers, and not to remove them from the data. Finally, it should be noted that other robust methods for covariance estimation may be used in the SEM context, with high breakdown point and which are affine equivariant procedures (see Huber et al. (2007)).

Acknowledgements: This project was supported by a grant from European Science Foundation for Quantitative Methods in Social Science.

References

BILLOR, N., HADI, A.S. and VELLEMAN, P.F. (2000): BACON: blocked adaptive computationally efficient outlier nominators, *Computational Statistics & Data Analysis, 34, 279–298.*

BOLLEN, K.A. and ARMINGER, G. (1991): Observational residuals in factor analysis and structural equation models, *Sociological Methodology, 21, 235–262.*

BOLLEN, K.A. (1989): *Structural equations with latent variables*, Wiley, New York.

BROWNE, M. W. (1982): Covariance structures. In: Hawkins, D.M. (Eds.): *Topics in applied multivariate analysis.* Cambridge University, UK, 72–141.

HUBA, G.J. and HARLOW, L.L. (1987): Robust Structural Equation Models: Implications for Developmental Psychology, *Child Development, 58 (1), 147–166.*

HUBERT, M., ROUSSEEUW, P.J. and VAN AELST, S. (2007): High-breakdown robust multivariate methods. *Statistical Science. in press.*

LEE, S.-Y. and XIA, Y.-M. (2006): Maximum likelihood methods in treating outliers and symmetrically heavy-tailed distributions for nonlinear structural equation models with missing data, *Psychometrika, 71 (3), 565–585.*

YUAN, K.-H., and BENTLER, P.M. (1998): Structural Equation Modeling With Robust Covariances, *Sociological Methodology, 28 (1), 363–396.*

Second-Order Model of Patent and Market Value

Alba Martinez-Ruiz[1] and Tomas Aluja-Banet[2]

[1] Departament of Statistics and Operations Research, Technical University of Catalonia
c. Jordi Girona 1-3, Campus Nord C5 204, 08034 Barcelona, Spain,
alba.martinez-ruiz@upc.edu

[2] Departament of Statistics and Operations Research, Technical University of Catalonia
c. Jordi Girona 1-3, Campus Nord C5 204, 08034 Barcelona, Spain,
tomas.aluja@upc.edu

Abstract. This research is a first attempt in a definition of a unique replicable model that allows one an estimate of the influence of the patent and market information in patent value, using a multidimensional and international approach. Partial least square path modelling and multiple imputation technique have been used to estimate the model and to solve the problem of missing values, respectively. Results suggest that the patent value is a second-order latent variable, international scope is the most important variable in the prediction of patent value and this can be decomposed in a potential and recognized value.

Keywords: patent value, SEM, partial least square path modelling

1 Introduction

Until now research in the patent field has been associated to the analysis of: (a) information contained in the document of a patent through its indicators, and (b) the relation between patents and R&D, innovation or economic growth. In recent years patent indicators have been used to study the economical value of the patent. In most cases patent indicators are developed from data contained in the patent document. For instance, family size is the number of countries where the patent is granted for the same invention (Reitzig 2004). Backward citations is the number of citations made by the patent examiner. Forward citations is the number of times that each patent has been cited by another patent (Trajtenberg 1990). Claims is a numbered list in the patent document, where it is specified what is protected. On the other hand, each one of these indicators has been related to another variable of more conceptual order. Number of the inventors and assignees, backward citations and number of claims have been related with the novelty of the patent. Protection level of the patent or its technological scope or its breadth can be measured by the number of claims or through the number of classes IPC (International

Patent Classification) into which the patent is protected. Moreover, it has been studied the relation between patent citations and patent value. Griliches (1981) and Hall et al. (2005) found a significant relationship between market value of a company, book value of research and development (R&D) expenditures, number of patents and citations. Relation between patent value and patenting strategy, technological diversity, domestic and international R&D collaborations and/or co-applications and the mix of designated states for protection, have been studied by Guellec et al. (2000).

In this research is rather considered the technological value of the patent, in the sense of how it is useful for the future technological development, and how it relates or influences the economical value of the company. This is a first attempt in a definition of a unique replicable model that allows one an estimate of the relationship between variables that determine the patent value, considering the different component of technology and market information and using Partial Least Square Path Modelling (Tenenhaus et al. 2005). This paper is organized as follows. First, the models are set. Then, the patent and economical data are given and the models analysed. Finally the results and conclusions are presented.

2 Model

Four models are tested. First we are interested in knowing the relationships between the patent indicators, the patent value and the different constructs and to know if it is possible to relate each other in a suitable way in a unique model. Therefore the first-order model considers four latent variables: patent value, knowledge stock of the patent, technological scope and international scope (see Figure 1).

The knowledge stock represents the base of knowledge contributed by the applicant. Therefore this latent variable is formed through of number of applicants and the number of inventors. Likewise number of backwards citations because is the knowledge that must exist previously to the birth of an invention. The technological scope is related with applying an invention in different technological fields. The manifest variables are, in a formative sense, the number of IPC where the patent is protected and the number of claims. The international scope refers to the geographic zones where the invention is protected after application. Therefore it is defined, in a formative sense, a dummy variable that considers if the patent had been protected in United States of America (USA), in Japan (JP), through of European Patent Office (EPO) or through WIPO (WO in the model). On the other hand, the importance that a patent has for the future technological developments will be reflected in the number of occasions that patent is cited. Also this importance will be reflected in the patent strategy of the company. We take into account the size of patent family. All variables on the left side of the model give an "a priori" value of the patent, i.e. the intrinsic characteristics of the patent

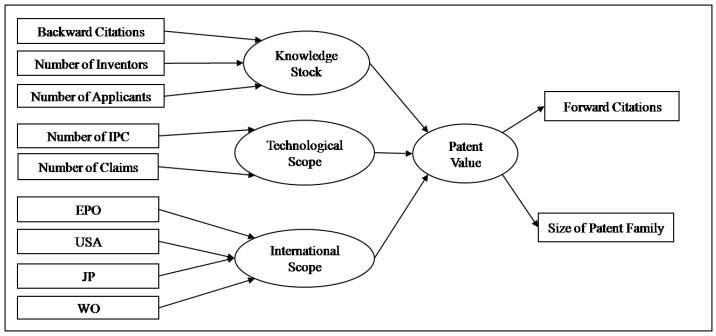


Fig. 1. Model 1.

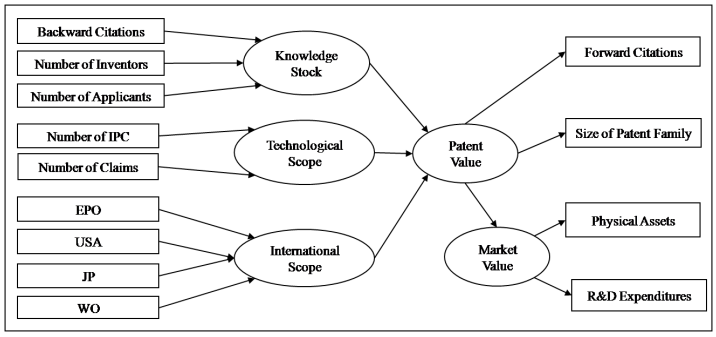


Fig. 2. Model 2.

at the time of its application and the patenting strategy of the company in the first 12 months can give an idea of a priori value of the patent. On the other hand, the variables in the right side of the model give an "a posteriori" value of the patent.

The second model considers the model 1 related to the market value of the company through the book value of total assets and R&D expenditures (see Figure 2). Given the results obtained in model 1 (see below) and the reflective nature of forward citations and patent family, it has been considered that in fact this variables represent a fifth latent variable, which is related to the "usefulness" of the patent, because the more useful is, more cited and more important the patenting strategy of the company. We propose that the patent value is a second-order latent variable. In its value not only contribute the prior value of the patent, but also its usefulness in the future. In figure 3 and 4 is shown a second-order model of patent value and one that relates it with market value.

PLS Path Modelling and bootstrapping are carried out in SmartPLS with centroid weighting scheme and individual sign changes. In reflective blocks, loadings indicate how well the indicators reflect its latent variable. To assess the internal consistency, it makes use of the Cronbach's alpha coefficient

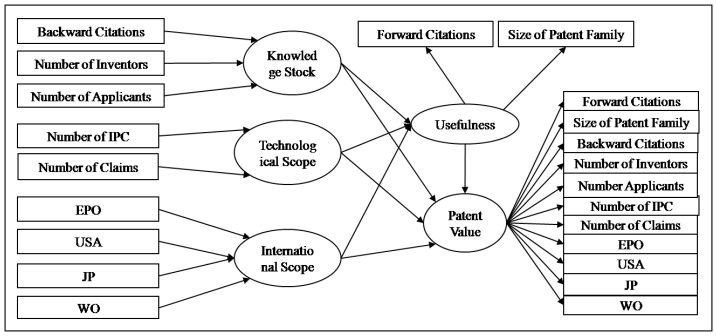


Fig. 3. Model 3.

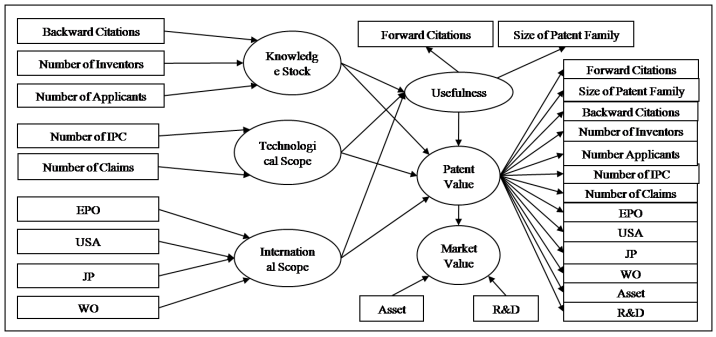


Fig. 4. Model 4.

(should be > 0.7), PCA (first eigenvalue must be larger than 1 and the second one smaller than 1) and composite reliability (should be > 0.7). Moreover, it is calculated the Average Variance Extracted (AVE) (should be > 0.5). In the case of formative blocks, weights/loadings allow us to determine the extent to which each indicator contributed to the formation of the constructs. The inner model is assessed examining the path coefficients (standardized beta), t-values and determination coefficients. Second-order models are estimated using Hierarchical Component Model.

3 Patent and economical data

We have a sample of 750 patents obtained from Derwent Innovation Index. The following indicators are calculated: family size, number of inventors, number of applicant, backward citations, number of IPC codes and the dummy variable: USA, JP, EPO and WO. Table 1 shows the descriptive statistics for the variables used in the analysis. The data indicate that variables are very heterogeneous and asymmetric, and also these exhibit large variance. Hence we work with the square root of the data.

Variable	Mean	Median	Minimum	Maximum	Standard Deviation	% of Missing Values
Backward	14.8	10	0	143	17.11	0
Inventors	2.63	2	1	12	1.81	0.53
Applicant	1.26	1	1	5	0.61	0
Number IPC	4.02	3	1	24	3.27	0
EPO	0.49	0	0	1	0.50	0
USA	0.80	1	0	1	0.39	0
JP	0.41	0	0	1	0.49	0
WO	0.10	0	0	1	0.30	0
Claims	14.05	11	1	77	10.30	2.00
Forward	11.20	7	0	126	13.06	0
Family	4.71	4	1	27	3.66	0

Table 1. Descriptive statistics for patent data.

The 77.8% of applicants are companies. Book value of total assets and the R&D expenditures are gathered from Annual Reports from 1990 to 2006. It has a sample of 553 patents with economical data but with high percentage of missing values by year, ranging from 67.45% in 1990 to 2.53% in 2006 to total assets (decreasing almost linearly through the years), and from 76.4% in 1990 to 9,76% in 2006 to R&D expenditures. Due to this number of missing value of economical data, we are driven to use multiple imputation techniques to not exlcude cases and to take into account the uncertainty of imputed value. Predictive Mean Matching is used with ten imputations (Schafer 1999). Therefore it has a complete matrix from 2000 to 2006 with economical data for 526 patents. We work with the logarithm of this data.

4 Results

For patent value in model 1 and 2, Cronbach's alpha coefficient is 0.5, the first eigenvalue is 1.27 and the second one 0.72, and the composite reliability is 0.77. Cronbach's alpha coefficient of patent value is 0.76 and 0.73 in model 3 and 4, respectively. Usefulness in model 3 and 4 has the same properties that patent value in models 1 and 2. Therefore, it is assumed that both, the patent value and usefulness, are unidimensional in all models. AVE is 0.64 in model 1 and 2. This value drops to 0.34 in model 3 and to 0.28 in model 4. AVE of usefulness is 0.64 in model 3 and 4. Table 2 gives the loadings and weights of PLS estimation for each model. In the case of models 2 and 4, the average values of ten imputations are shown and the coefficients incorporate

the uncertainty of the economical data. Significance levels are calculated by bootstrapping in models 1 and 3, and by Rubin's rule in the multiple imputations (models 2 and 4). The only variables that are not significant are the number of inventors, number of claims and WO in models 1 and 3. All other indicators are significant at 1% level, ranking from 2.05 (number of applicants) to 50.93 (family patent).

In all models the results suggest that backward citation is the most important variable in the formation of the base of knowledge and afterward the number of applicants; in the case of technology scope, the number of IPC is significant, that is the number of technological field where the patent have potential applicability. It is a bit odd that the number of claims is not appearing as a significant variable, it is not what one would expect. Perhaps the number of claims is not related with the technological scope but rather with a construct to measure the "quality" of the invention, in the sense of how this invention have an impact in a given technological field. Likewise international scope seems formed by its indicators. On the other hand, patent value is always reflected in the forward citations and patent family. Model 3 shows that in fact, we can relate these variables with the usefulness of the patent. Moreover, R&D expenditures appear more related with market value than the total assets of the company (models 2 and 4).

Table 3 shows the result for the inner model (standardized beta coefficients and significance levels). Coefficient of knowledge stock, technological scope and international scope to patent value are significant at 0.01 levels in the four models. Therefore, the patent value is formed by latent variables built through indicators of the patent document. These latent variables are not only reflected in the patent value, but also in its future usefulness (model 3 and 4). Patent value and its usefulness are related strongly and significantly. In addition, patent value is reflected significantly in the market value. The determination coefficient of patent value (model 1 and 2) is 0.7, i.e. the model fit the data in a 70%. In models 3 and 4, this coefficient rises up to 0.9. Usefulness has a determination coefficient of 0.6. Only the market value shows a poor fit of the data. More generally PLS estimation suggests that patent value is a second-order latent variable and the international scope is the most important variable in the prediction of its value.

5 Conclusions

Findings provide evidence that suggests that international scope, technological scope and knowledge stock are latent variables that determine the value of patent and they can be related into a unique replicable model which "feeds" of available public data. The magnitude of this relation is also known. The patent value can be decomposed in "a priori" and intrinsic value that the patent have at the moment of application, it is a potential value, and in "a posteriori" value that the patent acquired through time by the action of the

Construct	Indicator	Model 1	Model 2	Model 3	Model 4
Knowledge	Backward	0.9001	0.8982	0.8831	0.8862
	Inventors	0.1411	0.1430	0.1688	0.1762
	Applicants	0.2213	0.2244	0.2367	0.2201
Technological	Number IPC	0.9596	0.9608	0.9456	0.9462
	Claims	0.1691	0.1654	0.2097	0.2080
International	EPO	0.6060	0.6085	0.5951	0.5972
	USA	0.3067	0.2994	0.3181	0.3179
	JP	0.4265	0.4285	0.4282	0.4230
	WO	0.1169	0.1159	0.1259	0.1348
Usefulness	Family	-	-	0.9301	0.9308
	Forward	-	-	0.6567	0.6554
Patent value	Family	0.9371	0.9416	0.9029	0.8924
	Forward	0.6420	0.6318	0.4986	0.4830
	Backward	-	-	0.7524	0.7543
	Inventors	-	-	0.3377	0.3471
	Applicants	-	-	0.3026	0.2787
	Number IPC	-	-	0.7216	0.7130
	Claims	-	-	0.2861	0.2809
	EPO	-	-	0.7271	0.7234
	USA	-	-	0.4284	0.4241
	JP	-	-	0.6966	0.6863
	WO	-	-	0.2402	0.2527
	R&D	-	-	-	0.2564
	Assets	-	-	-	0.2292
Market Value	R&D	-	0.8318	-	0.8868
	Assets	-	0.2975	-	0.2292

Table 2. Loadings and weights of outer model according to the type of latent variable.

Latent Variable	Model 1	Model 2	Model 3	Model 4
Knowledge stock to Patent value	0.2679	0.2672	0.2644	0.2753
	(3.6401)	(2.3263)	(11.1624)	(2.3263)
Technological scope to Patent value	0.2827	0.2852	0.2361	0.2361
	(3.9159)	2.3263)	(11.5623)	(2.3263)
International scope to Patent value	0.4492	0.4504	0.3586	0.3673
	(5.7066)	(2.3263)	(14.4187)	(2.3263)
Knowledge stock to Usefulness	-	-	0.2661	0.2652
			(3.6288)	(2.3263)
Technological scope to Usefulness	-	-	0.2762	0.2778
			(3.6787)	(2.3263)
International scope to Usefulness	-	-	0.4513	0.4513
			(5.7354)	(2.3263)
Usefulness to Patent Value	-	-	0.3163	0.2880
			(13.0610)	(2.3263)
Patent value to Market value	-	0.0849	-	0.2632
		(2.8214)		(2.3263)
R^2 of patent value	0.7054	0.7120	0.9967	0.9773
R^2 of usefulness	-	-	0.6982	0.6993
R^2 of market value	-	0.0049	-	0.0627

Table 3. Path coefficients of the inner model with t-value in parenthesis. For models 2 and 4, they are mean value of 10 imputations.

company itself and others, it is the value that is recognized. The potential value depends on the characteristics itself of the patent. First the patenting strategy of the company, second the technological applicability (potential) of the patents and third the base of knowledge that is necessary for the creation of this new invention. When time passes, there is a recognition of the patent potentiality which is reflected in the number of times that is cited and the number of countries where it is protected. This is a reflection of the usefulness of the patent that also contributes to form its value. Moreover, this transfer is reflected in the market value of the company. Some limitations of this study are the availability of financial information. We tested the model with economical data of 2000, but it is likely that the value of these patents is reflected much earlier in market value.

Finally PLS Path Modelling shows to be a good tool for analyzing this kind of data. Other authors have already showed that the distribution of the patent value is highly skew. PLS Path Modelling has advantages over other

analysis methods to be an iterative algorithm and not make assumptions about data distribution.

6 Acknowledgement

This research is partly funded by Universidad Católica de la Ssma. Concepción, Chile.

References

GRILICHES, Z. (1981): Market value, R-and-D, and patents. *Economics Letters 7(2), 183-187.*

GUELLEC, D. and VAN POTTELSBERGUE, B. (2000): Applications, grants and the value of patent. *Economics Letters 69(1), 109-114.*

HALL, B.H., JAFFE, A. and TRAJTENBERG, M. (2005): Market value and patent citations. *The RAND Journal of Economics 36(1), 16-38.*

REITZIG, M. (2004): Improving patent valuations for management purposes - validating new indicators by analyzing application rationales. *Research Policy 33(6-7), 939-957.*

SCHAFER, J.L. (1999): Multiple imputation: a primer. *Statistical Methods in Medical Research(8), 3-15.*

TENENHAUS, M.,ESPOSITO, V., CHATELIN, Y.M. and LAURO, C. (2005): PLS path modeling. *Computational Statistics & Data Analysis 48, 159-205.*

TRAJTENBERG, M. (1990): A penny for your quotes - patent citations and the value of innovations. *The RAND Journal of Economics 21(1), 172-187.*

Part XVII

Resampling

A Robust Approach for Treatment Ranking Within the Multivariate One-Way ANOVA Layout

Rosa Arboretti Giancristofaro,[1] Dario Basso,[2] Stefano Bonnini[1] and Livio Corain[2]

[1] Department of Mathematics, University of Ferrara; Via Machiavelli, 35 - 44100 Ferrara, Italy, *rbrrso@unife.it*
[2] Department of Management and Engineering, University of Padova; Stradella S. Nicola, 3 - 3600 Vicenza, Italy, *livio.corain@unipd.it*

Abstract. Our goal is to define a robust treatment ranking criterion in the framework of multivariate one-way ANOVA layout. This is an extension of the method proposed by Corain and Salmaso (2007). In this paper we evaluate and compare via simulation study our proposed method, which directly make use of p-values, with the sum of score method proposed by Corain and Salmaso. Our proposed method is theoretically more robust and performs better than the sum of partial scores because p-values are in general more informative. Finally, we propose an application of our robust approach to industrial washing performance testing.

Keywords: dependent rankings, multiple comparisons, nonparametric combination

1 Introduction and motivations

The literature of multiple comparison methods addresses the problem of ranking the treatment groups from worst to best (Westfall et al., 1999), but it does not seem to provide clear indications on how dealing with the information from pairwise multiple comparisons especially in case of blocking (or stratification) and in case of multivariate response variable. This problem is not only of theoretical interest but also it has a very practical relevance. In fact, especially for industrial research, a global ranking from many performance indicators of all investigated products/prototypes is a very natural goal. The topic of defining a global preference ranking of industrial products has been addressed by Corain and Salmaso (2007). Authors propose to apply a two-stage nonparametric procedure: firstly they perform a set of C-sample testing procedures, followed by multiple comparisons. In this way they evaluate a set of partial preference rankings. In the second stage of the procedure they synthesise the partial rankings by combining them into a global ranking that provides a general product preference rule. In this paper we proceed in that direction by extending the procedure for the multivariate one-way ANOVA layout. As we will show, the modified procedure performs better than that proposed by Corain and Salmaso.

2 A robust approach for treatment ranking

Let Y be the multivariate numeric variable related to the response of any experiment of interest and let us assume, without loss of generality, that high values of each Y univariate element correspond to better performance and therefore to a higher degree of treatment preference. The experimental design of interest is defined by the comparison of C groups or treatments with respect to S different variables where n replicates of a single experiment are performed by a random assignment of a statistical unit to a given group. The C-group multivariate statistical model (with fixed effects) can be represented as follows:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim \text{IID}(0, \sigma_{ij}^2), i = 1, ..., C; j = 1, ..., S; \quad k = 1, ..., n;$$

where, in the case of a balanced design, n is equal to the number of replicates and subscripts i and j are related with the groups (treatments) and the univariate response variable respectively. The resulting inferential problem of interest is concerned with a set of S hypothesis testing procedures $H_{0j} : \mu_{1j} = \mu_{2j} = \cdots = \mu_{Cj}$ vs. $H_{1j} : \exists \mu_{ij} \neq \mu_{hj}$, $i, h = 1, \ldots, C$, $i \neq h$, $j = 1, \ldots, S$. If H_{0j} is rejected a further possible set of $C \times (C-1)/2$ all pairwise comparisons are performed:

$$\begin{cases} H_{0ih|j} : & \mu_{ij} = \mu_{hj} \\ H_{1ih|j} : \exists & \mu_{ij} \neq \mu_{hj} \end{cases}.$$

In the framework of parametric methods, when assuming the hypothesis of normality for random error components, the inferential problem can be solved by means of the ANOVA F test and a further set of pairwise tests using Fisher's LSD or Tukey procedures, which are two of most popular multiple comparison procedures (Montgomery, 2005). On the basis of inferential results achieved at the univariate C-group comparison stage, the next step consists in producing a ranking of the treatments from the less to the more preferred. For this goal, Corain and Salmaso (2007) propose to sum some meaningful scores from inferential results at the univariate C-group comparison and then to apply the NonParametric Combination (NPC) of partial rankings (Lago and Pesarin, 2000). In this way they acquire a unique preference criterion which jointly takes into account all performances achieved for every response variable. To illustrate the method, let us suppose H_{0j} has been rejected for all $j = 1, ..., S$, so that for each univariate response there is some treatment that significantly differs for some other. In order to suitably synthesise the pair-wise comparison results for each response variable j, $j = 1, ..., S$, let us define a set of S score matrices of dimension $C \times C$, where each element $[x_{ih|j}]$ is related to the comparison between the treatments i and h for each response variable j, giving the value of $+1$ to the significantly better treatment and -1 to the other, while both score are 0 if the comparison

is not significant. Formally,

$$\begin{cases} \text{if } H_{0ih|j} : \mu_{ij} = \mu_{hj} \text{ is not rejected then } x_{ih|j} = x_{hi|j} = 0; \\ \text{if } H_{0ih|j} : \mu_{ij} = \mu_{hj} \text{ is rejected then } \begin{cases} \text{if } \bar{y}_{ij} > \bar{y}_{hj} \text{ then } x_{ih|j} = +1 \text{ and } x_{hi|j} = -1; \\ \text{if } \bar{y}_{ij} < \bar{y}_{hj} \text{ then } x_{ih|j} = -1 \text{ and } x_{hi|j} = +1; \end{cases} \end{cases}$$

where $\bar{y}_{ij}$ and $\bar{y}_{hj}$, $i, h = 1, \ldots, C$, $i \neq h$, are the sample means of groups i and h for response variable Y_{ij}, $i = 1, \ldots, C$, $j = 1, \ldots, S$. Note that pairwise comparisons and the valid score assignments are performed only when the C-sample test has rejected the null hypothesis H_{0j}, $j = 1, ..., S$. For each response variable $j = 1, ..., S$ once this assignment has been performed for each pairwise comparison, it is easily feasible to obtain a set of S $X_j = [x_{1|j}, x_{2|j}, \ldots, x_{C|j}]'$ score vectors, $j = 1, ..., S$, where the C elements $x_{i|j}$, $i = 1, \ldots, C$, $j = 1, ..., S$, of the score vector X_j, are calculated by summing all the obtained scores for each treatment, i.e.:

$$x_{i|j} = \sum_{h=1, h\neq i}^{C} x_{ih|j}, \qquad i = 1, ..., C, j = 1, ..., S.$$

Instead of using ± 1 summation as proposed by Corain and Salmaso (2007), we suggest to make directly use of p-values. For this goal let us consider the set of S one-sided p-value matrices of dimension $C \times C$, where each element $[p_{ih|j}]$ is related to the comparison between treatments i and h for response variable j. For each response variable j, $j = 1, ..., S$, it is possible to obtain an alternative set of S score vectors X_j, $j = 1, ..., S$, where each element of X_j is calculated as follows:

$$x_{i|j} = -2 \sum_{h=1, h\neq i}^{\mathrm{C}} \log\, p_{ih|j},\ i = 1, \ldots, C,\ j = 1, ..., S.$$

Note that we use as p-value synthesis criterion the Fisher's combining function. It is worth noting that the Fisher's combining function is nonparametric with respect to the underlying dependence structure among p-values from different univariate variable response, in that all kinds of monotonic dependences are implicitly captured. Indeed, no explicit model for this dependence structure is needed and no dependent coefficient has to be estimated directly from the data. Then, in order to suitably synthesise the scores for each response variable j, $j = 1, ..., S$, Corain and Salmaso suggest to use the NonParametric Combination (NPC) of partial rankings (Lago and Pesarin, 2000) to acquire a unique preference criterion which jointly takes into account all response variables. For this goal, consider the rank transformations R_{ij} (partial rankings):

$$\left\{ R_{i|j} = R(x_{i|j}) = \#(x_{i|j} \geq x_{h|j}), \quad i, h = 1, \ldots, C, i \neq h, j = 1, \ldots, S \right\}.$$

Associated with these ranks are the scores:

$$\left\{\lambda_{i|j} = \frac{R_{i|j}+0.5}{\mathrm{C}+1},\ i = 1, ..., \mathrm{C},\ j = 1, ..., \mathrm{S}\right\}.$$

Once a combining function ψ (for more details refer to Pesarin, 2001 and Arboretti et al., 2005) has been chosen, we compute the transformation:

$$\psi : \left\{ Q_i = \psi(\lambda_{i|1}, \ldots, \lambda_{i|\mathrm{S}}; w_1, \ldots, w_{\mathrm{S}}), i = 1, \ldots, \mathrm{C}\right\}$$

and finally, applying the rank transformation, we obtain the global combined ranking G:

$$\{G_i = R(Q_i) = \#(Q_i \geq Q_h), i, h = 1, \ldots, \mathrm{C}\}.$$

In the global ranking G, each treatments is ranked in a unique way, by taking into consideration the whole set of the S informative response variables. Some of the combining functions most frequently used are Fisher, Logistic, Lancaster, Liptak, Tippett, Mahalanobis (Pesarin, 2001). In the overall analysis it is possible to include an eventual *a priori* importance of response variable by using appropriate weights opportunely fixed: $w_j \geq 0, j = 1, ..., S$; hence Fisher and Logistic combining functions become:

$$\psi_i^F = -\sum_{j=1}^{\mathrm{S}} w_j \cdot \log(1-\lambda_{i|j}),\ \psi_i^L = \sum_{j=1}^{\mathrm{S}} w_j \cdot \log\left[\lambda_{i|j}/(1-\lambda_{i|j})\right].$$

3 Simulation study

In order to evaluate their degrees of accuracy in detecting the true 'unknown' treatment ranking, in this section we perform a comparative simulation study between the p-value score method and the sum of score method. Let us consider the following simulation setting:

- 3 positive numeric response variables (C=3), with "hypothetical" maximum value at 100, 6 treatments (S=6) and 4 experimental replicates (n=4)
- normal random errors (ε_{ijk} ~IIN(0,σ_{ij}^2), i=1,2,3, j=1,...,6, k=1,...,4), where σ_{ij}^2 follows one of the three settings below;
- three variance/covariance settings:
  - (i) Setting 1: variables are independent ($\sigma_{ih}^2 = 0, i,\ h = 1, 2, 3,\ i \neq h$) and their variances are homoschedastic ($\sigma_{ii}^2 = 1,\ i = 1, 2, 3$);
  - (ii) Setting 2: variables are independent ($\sigma_{ih}^2 = 0, i,\ h = 1, 2, 3,\ i \neq h$) and their variances are heteroschedastic ($\sigma_{11}^2 = 1,\ \sigma_{22}^2 = 4,\ \sigma_{33}^2 = 2.25$);
  - (iii) Setting 3: variables are not independent ($\sigma_{12} = 0.3,\ \sigma_{13} = 0.5,\ \sigma_{23} = 0.8$) and their variances are homoschedastic ($\sigma_{ii}^2 = 1,\ i = 1, 2, 3$).
- ANOVA F test for C-sample testing and pairwise Fisher's LSD for multiple comparisons;
- Fisher's combining function (as suggested by Corain and Salmaso, 2007) and 1,000 independent simulations;

Table 1. Setting of treatment mean value for simulation study.

Treatment	μ_1	μ_2	μ_3	Distance from (100,100,100)	True global ranking
1	90	90	90	17.3	1
2	89	90	88	19.1	2
3	88	87	89	20.8	3
4	87	86	88	22.6	4
5	86	87	85	24.3	5
6	85	85	85	26.0	6

- a fixed structure of true treatment mean values, as in Table 1.

Note that the true global ranking follows the label treatment ordering and can be obtained by calculating the Euclidean distance of each treatment from the perfect ideal treatment, that is from the 3-dimensional point (100,100,100). Table 2 summarizes counting of obtained global rankings for p-value score method and sum of score method, with reference to setting 1. In our simulation we rejected the global null hypothesis H_{0j}, $j = 1,2,3$, all 1000 times. In fact, our setting of treatment means is quite strength against H_{0j}, $j = 1,2,3$. In case we did not rejected H_{0j}, $j = 1,2,3$, all treatment would had the same ranking (all equal to 1). Note that even if both methods perform very good, p-value score method is more precise. We can also evaluate and compare the accuracy of the two procedures by counting the rates they are able to perform a global ranking which is exactly the same as the true ranking. For p-value scores method we get a rate of exact correspondence with the true ranking of 96.0% while it is only 69.4% for sum of score method. In general, note that

Table 2. Global rankings for variance setting 1.

true global	p-value scores method						Sum of scores ranking					
ranking	1	2	3	4	5	6	1	2	3	4	5	6
1	993	7	0	0	0	0	931	67	2	0	0	0
2	7	985	8	0	0	0	15	929	54	2	0	0
3	0	8	986	6	0	0	0	23	905	70	2	0
4	0	0	6	979	15	0	0	0	13	915	71	1
5	0	0	0	15	975	10	0	0	0	25	904	71
6	0	0	0	0	10	990	0	0	0	1	18	981
% of true rank.	99.3	98.5	98.6	97.9	97.5	99.0	93.1	92.9	90.5	91.5	90.4	98.1

p-value scores method does produce a lower rate of misclassification with a lower misclassification variability as well. With reference to the more realistic second third variance setting (1,000 independent simulations), results are displayed in Table 3. Also for setting 2 and 3, in our simulation we rejected

Table 3. Global rankings for variance setting 2 and 3.

true global	p-value scores method						Sum of scores ranking					
ranking	1	2	3	4	5	6	1	2	3	4	5	6
Setting 2												
1	904	92	4	0	0	0	828	153	12	2	2	
2	95	812	89	4	0	0	77	674	175	50	21	
3	1	95	790	112	2	0	5	80	587	261	63	
4	0	1	113	775	108	3	0	3	68	645	267	
5	0	0	4	106	788	102	0	0	3	104	717	
6	0	0	0	3	102	895	0	0	0	3	92	
% true rank.	90.4	81.2	79.0	77.5	78.8	89.5	82.8	67.4	58.7	64.5	71.7	
Setting 3												
1	954	45	1	0	0	0	872	122	6	0	0	
2	46	906	48	0	0	0	51	827	112	9	1	
3	0	49	903	46	2	0	0	45	819	128	8	
4	0	0	47	913	40	0	0	0	39	834	126	
5	0	0	1	41	917	41	0	0	1	44	833	
6	0	0	0	0	41	959	0	0	0	0	34	
% true rank.	95.4	90.6	90.3	91.3	91.7	95.9	87.2	82.7	81.9	83.4	83.3	

the global null hypothesis H_{0j}, $j = 1,2,3$, all 1000 times. With reference to second and third variance settings, the percentage of right classification of the global ranking is 74.2% for p-value scores method against only 24.6% for sum of score method in setting 2, and 79.0% for p-value scores method against 49.2% for sum of score method in setting 3. The better performance and the lower rate of misclassification of p-value scores method with respect to sum of scores can be promptly assessed by Spearman correlation coefficient and Shannon heterogeneity coefficient (Agresti, 2002). Both coefficients, reported in Table 4, are calculated (along with the corresponding p-value) from counts of Table 2 and 3. Hence, Spearman correlation coefficient evaluates the 'strength' of the correlation between true and performed rankings, while Shannon heterogeneity coefficient assesses the 'degree of dispersion' of counts of each of the two procedures. As expected, since p-value score method

performs better than sum of scores, it has an higher value of Spearman coefficient and a lower value of the Shannon coefficient which is a measure of the misclassification variability.

Table 4. Spearman and Shannon coefficients and related p-values.

	p-value score method		Sum of score method	
	Spearman	Shannon	Spearman	Shannon
Setting	corr. coeff.	heter. coeff.	corr. coeff.	heter. coeff.
1	0.997 (.000)	0.524 (.000)	0.987 (.000)	0.581 (.000)
2	0.968 (.000)	0.657 (.000)	0.929 (.000)	0.717 (.000)
3	0.987 (.000)	0.585 (.000)	0.974 (.000)	0.633 (.000)

In general, with reference to all simulation settings, we can conclude that it is evident the observed better performance of p-value score in comparison with sum of score method. Moreover, the p-value score method is more robust than sum of scores method with respect to violations of homoschedasticity and of independence.

4 Application to industrial products

In this section we illustrate the application of the proposed methodology to a real case study concerned with the development of a new detergent. The R&D division of a chemical company wants to assess the degree of preference of a set of 6 prototypes (C=6), taking into account their performances on 4 types of soil (S=4) belonging to bleach type. A suitable experiment has been designed: for each of the 6 considered products 5 washing machines (there are n=5 replicates) are used to wash a piece of fabric soiled with the 4 soils. The experimental response variable is the reflectance, i.e. the percentage of removed soil for the 4 types of soil. Table 5 and 6 display the sample statistics from observed experimental data. Considering a significance level of 5%, parametric C-sample F-test and pairwise Fisher's LSD, adjusted using the Bonferroni procedure, have been performed. We have rejected the global null hypothesis for all 4 types of soil.

Then we have calculated the global ranking (Fisher combining function) by applying both global ranking methods (Table 7). Results do not change using Liptak or Tippet combining functions instead of Fisher combining function.

It is worth noting that the two global rankings are similar but not identical. Due to its robustness showed by previous simulations we can be confident that the p-value score ranking is more reliable than sum of score ranking.

5 Conclusions

As highlighted by the simulation study presented in this paper, the proposed p-value score ranking method performs better than the sum of partial scores. The proposed method is able to 'include' more useful information from experimental data than the sum of score method. Moreover, p-value

Table 5. Sample mean and std. deviation from observed experimental data.

Soil	Sample	Product					
	Mean/Std. Dev.	1	2	3	4	5	6
CFT	Mean	78.2	84.8	84.8	97.8	84.0	82.9
CS 103	Srd. Dev.	1.4	1.0	1.0	0.8	0.2	1.1
CFT	Mean	63.1	69.8	72.0	67.7	65.4	89.3
BC 03	Srd. Dev.	1.1	0.5	0.6	0.5	0.4	0.4
CFT	Mean	79.1	82.5	84.1	83.4	82.3	89.1
BC 02	Srd. Dev.	0.8	0.5	0.4	0.3	0.3	0.2
CFT	Mean	79.1	82.5	84.1	83.4	82.3	89.1
CS 15	Srd. Dev.	0.8	0.5	0.4	0.3	0.3	0.2

Table 6. Sample soil correlation and related p-value.

Soil	Sample	Soil					
	Corr./p-value	CFT	BC03	CFT	BC02	CFT	CS15
CFT	Correlation	.085		.003		.352	
CS 103	p-value	.656		.985		.057	
CFT	Correlation			.975		.943	
BC 03	p-value			.000		.000	
CFT	Correlation					.911	
BC 02	p-value					.000	

Table 7. Global rankings.

	Product
Global ranking method	1 2 3 4 5 6
p-value scores	6 3 1 4 5 2
Sum of scores	6 4 2 3 5 1

score ranking can implicitly capture the dependence structure within partial univariate rankings if it exists. This result confirms the findings of Arboretti G. et al. (2005), where authors highlight that in case of heavy-tailed random distributions NPC ranking performs better than the simple arithmetic mean.

References

AGRESTI, A. (2002): *Categorical Data Analysis.* John Wiley and sons, Chichester.

ARBORETTI, G.R., MAROZZI, M. and SALMASO, L. (2005): Nonparametric Pooling and Testing of Preference Ratings for Full-Profile Conjoint Analysis Experiments. *Journal of Modern Applied Statistical Methods, 4, 2, 353-628.*

CORAIN, L. and SALMASO, L. (2007): A nonparametric method for defining a global preference ranking of industrial products. *Journal of Applied Statistics, 34, 2, 203-216.*

LAGO, A. and PESARIN, F. (2000): Non Parametric Combination of Dependent Rankings with Application to the Quality Assessment of Industrial Products. *Metron, LVIII, 39-52.*

MONTGOMERY, D.C. (2005): *Design and Analysis of Experiments, 6th Edition.* John Wiley and sons, Chichester.

PESARIN, F. (2001): *Multivariate permutation tests with applications in biostatistics.* John Wiley and sons, Chichester.

WESTFALL, P.H., TOBIAS, R.D., ROM, D., WOLFINGER, R.D. and HOCHBERG, Y. (1999): *Multiple Comparisons and Multiple Tests using the SAS.* SAS Books by Users.

Permutation Testing for Alternative Nonlinear Models with Application to Aging Curves of Refrigerated Vehicles

Rosa Arboretti Giancristofaro[1], Livio Corain[2], Samuela Franceschini[2], Andrey Pepelyshev[3], and Stefano Rossi[4]

[1] Department of Mathematics, University of Ferrara; Via Machiavelli, 35 - 44100 Ferrara, Italy, *rbrrso@unife.it*
[2] Department of Management and Engineering, University of Padova; Stradella S. Nicola, 3 - 3600 Vicenza, Italy, *livio.corain@unipd.it*
[3] Department of Stochastic Simulation, St. Petersburg State University; Bibliotechnaya sq.2, St.Petersburg - 198904, Russia
[4] Italian National Research Council, Construction Technologies Institute; Corso Stati Uniti, 4 - 35127 Padova, Italy

Abstract. Testing of model fitting for alternative nonlinear model comparisons within the parametric approach is traditionally a difficult topic due to the complexity of studying the null distribution of test statistics. The nonparametric permutation approach is a flexible method, suitable to be implemented for nonlinear models. In this paper, we introduce a novel algorithm within the nonparametric permutation framework able to perform proper inference on parameters of any specified nonlinear model. The algorithm, although general in its kind, offers a well-rounded approach to make inference via permutation test. Finally, we show the usefulness of the proposed method by applying it for making inference on parameters of a nonlinear aging curve for refrigerated vehicles.

Keywords: aging curve, nonlinear models, permutation tests

1 Nonparametric permutation inference on linear models

Permutation tests are conditional inferential procedures where conditioning is performed with respect to the sub-space associated with the set of sufficient statistics under the null hypothesis for all nuisance entities, including the underlying, known or unknown, distribution. For details, see Edgington (1995) and Pesarin (2001). The observed dataset is always a set of sufficient statistics under the null hypothesis for whatever underlying distribution. Therefore, permutation tests can be viewed as nonparametric inferential procedures, conditioned to the space generated by all possible data assignments. Provided that the null hypothesis implies the exchangeability of data, in the framework of permutation tests, the reference distribution of a relevant test statistic is then constructed by calculating its value for all possible permutations (re-orderings) of the observations. Thus a p-value can be computed as the proportion of the permutation values of the statistic that are

equal to or greater than the observed value. For a more detailed introduction on permutation tests, we refer the reader to Pesarin (2001).

With regard to inference on linear models, permutation tests applied to multiple regression analysis have been proposed in the literature by Cade and Richards (1995) and by Kennedy and Cade (1996). These authors suggested to permute the residuals which are calculated with respect to estimated regression models. The model parameters can be estimated using the least squares method (ter Braak, 1992; Kennedy and Cade,1996) or the least absolute deviations method (Cade and Richards, 1995, Mielke and Berry, 2001). Cade and Richards (1996) proposed a permutation test for hypothesis testing on LAD (Least Absolute Deviation) regression models, based on permutation of the observed data. The test statistic was drawn from the F test, used in the least squares regression to evaluate the goodness of the estimated models. Kennedy and Cade (1995) employed the permutation test in the comparison of nested models evaluated using the least squares method. They showed that this permutation strategy is valid only when the effects which are not under testing are null. Stapel and ter Braak (1994), however, showed that the method is valid when estimating the largest possible effect, since the other effects influence only marginally such estimation.

2 A general algorithm for permutation inference in nonlinear models

In this section, we present a novel general residual-based algorithm for inference on a single parameter or on a subset of parameters within a nonparametric permutation approach. The proposed technique is suitable for both linear and nonlinear models.

Let us consider any specified nonlinear model

$$Y_i = f(X_i; \beta) + \varepsilon_i, \quad i = 1, \ldots, n$$

where Y_i is the response variable, $f(\bullet; \beta)$ is the nonlinear link function, X_i is the vector of explicative variables and ε_i are exchangeable random errors with zero mean and unknown continuous distribution P, $i = 1, \ldots, n$. Let us suppose an appropriate method is available to calculate $\hat{\beta}$, i.e. the estimate of the parameter vector β. The null hypothesis of interest is

$$\boldsymbol{H_0 : \tilde{\beta} = 0} \text{ vs. } \boldsymbol{H_1 : \tilde{\beta} \neq 0}$$

where $\tilde{\beta}$ is a single parameter or a subset of parameters from β.

The proposed algorithm is defined by the following steps:

(i) estimate the parameter vector β from two models: the first estimate $\hat{\beta}_0$ related to the first model M_0, i.e. under H_0 (with the constraint $\tilde{\beta} = 0$), and the second estimate $\hat{\beta}_1$ related to the second model M_1, i.e. under H_1 (without the constraint $\tilde{\beta} = 0$);

(ii) calculate two vectors of estimated response values: the first $\widehat{Y}_0 = f(X, \hat{\beta}_0)$ (from the model M_0) and the second $\widehat{Y}_1 = f(X, \hat{\beta}_1)$ (from the model M_1);

(iii) calculate two vectors of residuals: the first $\boldsymbol{R_0} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_{\boldsymbol{0}}$ (under H_0) and the second $\boldsymbol{R_1} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_{\boldsymbol{1}}$ (under H_1);

(iv) calculate S_0, that is the observed value of an appropriate statistic $S(R_0, R_1)$, based on R_0 and R_1. As test statistic we propose for example

$$S = (\mathrm{SSE}_0 - \mathrm{SSE}_1)/\mathrm{SSE}_1$$

where SSE_0 and SSE_1 are respectively the sum of square of residuals under H_0 and under H_1. Other residual-based statistics related to alternative model comparison (Burnham and Anderson, 2002) may be suitable;

(v) randomly permute the paired elements of R_0 and R_1 to obtain R_0^* and R_1^*. Note that the null hypothesis H_0 implies the exchangeability of random errors ε_i, $i = 1, \ldots, n$, with respect to the models M_0 and M_1. Thus, if H_0 holds, we can randomly permute residuals;

(vi) calculate $\boldsymbol{Y_0^*} = \widehat{\boldsymbol{Y}}_{\boldsymbol{0}} + \boldsymbol{R_0^*}$ and $\boldsymbol{Y_1^*} = \widehat{\boldsymbol{Y}}_{\boldsymbol{1}} + \boldsymbol{R_1^*}$;

(vii) from Y_0^* and Y_1^*, re-estimate the parameter vector $\hat{\beta}_0$ and $\hat{\beta}_1$ the two model M_0 and M_1 and the corresponding residuals;

(viii) re-calculate the value of S so that we have $S^* = S(R_0^*, R_1^*)$;

(ix) carry out $B - 1$ independent repetitions of steps (5)-(8), so that we have S_j^*, $j = 1, \ldots, B$, (i.e. a random sampling from the permutation distribution of S);

(x) the permutation estimated p-value $\hat{p}$ for H_0 vs. H_1 is given by

$$\hat{p} = \#(S_j^* \geq S_0)/B;$$

(xi) if $\hat{p} < \alpha$, the null hypothesis H_0 is rejected at the significance level α.

3 Simulation study

In this section, we evaluate the appropriateness of the proposed method through the use of a Monte Carlo simulation study. Let us to consider the well-known three-parameter logistic model:

$$Y_i = \frac{\theta}{1 + \beta_2 e^{-\beta_3 X_{1i}}} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Moreover, we assume that

$$\theta = \beta_1 + \beta_4 X_{2i} + \beta_5 X_{3i}$$

where X_2 is a dummy variable representing some sort of possible fixed effects in Y while X_3 is a numerical covariate. We set the value of parameters as follows $\beta_1 = \beta_2 = \beta_3 = 1$, $\beta_4 = 0.1$, $\beta_5 = 0.07$ and we generate the value of numerical covariates and random errors as follows: X_{1i} are i.i.d from Uniform$[0, 4]$, X_{2i} are i.i.d from Bernoulli$[1/2]$, X_{3i} are i.i.d from a $N(0, 1)$ and ε_i are i.i.d from a $N(0, 0.1)$ since such distributions seem appropriate to represent real data configurations. Hence, possible hypotheses of interest are

- $H_{01} : \beta_4 = \beta_5 = 0$ vs. H_{11} : at least one β_i, $i = 4, 5$ is different from 0;
- $H_{02} : \beta_4 = 0$ vs. $H_{12} : \beta_4 \neq 0$;
- $H_{03} : \beta_5 = 0$ vs. $H_{13} : \beta_5 \neq 0$.

Note that the hypotheses of interest are related to any possible nonlinear model which is alternative to the more simple three-parameter logistic model. Hence, parameters β_4 and β_5 represent a possible different model which we would like to identify using the proposed permutation testing procedure.

Suitable MatLab routines were implemented in order to numerically estimate the parameters of the nonlinear model using the Nelder-Mead algorithm (Lagarias et.al., 1999) and to execute the proposed permutation test. These programs are available upon request by authors.

The considered simulation setting consists of 1000 Monte Carlo simulations for the generation of 100 observations ($n = 100$), where the true values are added to standard normally distributed random errors. For each one of the 1000 simulated data we separately estimated the permutation p-values (with 1000 random permutations) following the proposed algorithm for each one of the hypotheses of interest.

Simulations under H_0, which are reported here for hypothesis H_{01} in the last row of Table 1, show that the test distribution follows the achievable nominal levels. The rejection rates under each specific alternative are displayed in Table 1.

Table 1. Permutation test rejection rates under H_{11}, H_{12} and H_{13}.

Hypothesis	nominal level α				
	0.01	0.05	0.1	0.2	0.3
H_{11}(all param.)	0.642	0.945	1.000	1.000	1.000
$H_{12}(\beta_4)$	0.079	0.424	0.782	0.970	1.000
$H_{13}(\beta_5)$	0.242	0.703	0.939	1.000	1.000
H_{01} (all param.)	0.007	0.027	0.082	0.178	0.260

Note that the proposed permutation tests show in general a very good power. since they allow us to identify the true alternative for each of the three hypotheses of interest. For example, when setting the significance level α at 0.05, we reject the false null hypothesis H_{11} 94.5% times, H_{12} 42.4% times and H_{13} 70.3% times. Note that, the The procedure is more powerful in presence of numerical covariate (H_{12}) than for categorical variable (H_{13}).

4 Application to aging curve of refrigerated vehicles

The Accord Transport Perishable (ATP), established in 1970 among some European states and ratified in Italy in 1977, defines the precise structural characteristics of isothermal units at controlled temperature to be used in the transport of perishable products. Over time the insulated capacity of refrigerated transportation vehicles tends to diminish, thus allowing for an increase in the so-called overall coefficient of heat transfer K which represents the insulating capacity of the equipment and is defined as $K = W/(S \triangle T)$ where W is the thermal capacity required in a body of mean surface area S

to maintain the absolute difference $\triangle T$ between the mean inside temperature T_i and the mean outside temperature T_e, during continuous operation, when the mean outside temperature T_e is constant. The mean surface area S of the body is the geometric mean of the inside surface area S_i and the outside surface area S_e (United Nations Economic Commission for Europe, 1970).

Several factors contribute to the increase in K: some maybe regarded as structural deformation due to wear and tear while others are related to an increase of water in the polyurethane slab. A mathematical formulation of the aging curve was derived to study the effects that structural characteristics, as well as operation and maintenance practices, have on the life of refrigerated transport units. For details see Sicuro (2006). This theoretical model was derived as a combination of the physical processes involved in the heat transfer within the insulating panel and was calibrated with respect to the data available through the ATP database. This model formulation allows comparing the effective age of different type of refrigeration units independently from their manufacturing, structure or use.

Based on this model, the theoretical aging, Y_t of a refrigeration unit at its time-life t can be computed according to (1):

$$Y_t = 100\left(\frac{TC_t}{TC_0} - 1\right)\theta \tag{1}$$

where TC_0 and TC_t are the thermal conductivities respectively computed at time $t = 0$ (i.e. when the refrigeration unit is new) and at time t during the life of the unit (i.e. while the refrigeration unit is in use). These thermal conductivities are obtained using (2) and (3).

$$TC_0 = \frac{2}{3}(1 - K_1)\left(1 - \frac{K_2}{2}\right)K_3 + K_1\frac{K_4K_5 + K_4K_5}{K_4 + K_5} + \frac{16}{3}\frac{K_8K_9K_{10}}{\sqrt{3.68\frac{K_{11}}{K_{12}}}} \tag{2}$$

$$\begin{aligned} TC_t = {} & \tfrac{2}{3}(1 - K_1)\left(1 - \tfrac{K_2}{2}\right)K_3 + K_1(K_{13}\cdot t)K_{14} + \\ & K_1(1 - K_{13}\cdot t)\left[\left(\tfrac{P_{air,t}}{P_{tot,t}}\right)K_4 + \left(\tfrac{P_{R,t}}{P_{tot,t}}\right)K_6\right] + \tfrac{16}{3}\tfrac{K_8\cdot K_9\cdot K_{10}}{\sqrt{3.68\frac{K_{11}}{K_{12}}}} \end{aligned} \tag{3}$$

where (in brackets we report the value of constants K_i, $i = 1, \ldots, 15$) K_1 (0.97) is the porosity of the polyurethane slab, which is given by the ratio of the volume occupied by the air and gas and the total volume of the slab; K_2 (0.85) the fraction of volume occupied by the solid in the selected geometrical representation of the polyurethane structure. For this particular derivation, the insulating material is represented as in line cubic cells of equal dimensions surrounded by a layer of gas. The constants K_3 (0.29 W/mK), K_4 (0.026 W/mK), K_6 (0.0078 W/mK) and K_{14} (0.6163 W/mK) are the thermal conductivity coefficients of the solid, the air, the compressed gas (R) and the water in the polyurethane slab, respectively, K_5 (1000 Pa) and K_7 (90000 Pa) are the initial partial pressures of the air and of the compressed gas, respectively, $P_{tot,t}$ is the total pressure in the polyurethane slab at time t [Pa], and is given by the sum of $P_{air,t}$ and $P_{R,t}$, the partial pressures of the

air and gas at time t, respectively, $P_{air,t}$ and $P_{R,t}$ are computed according to (4) and (5),

$$P_{air,t} = K_5 + (K_{15} - K_5)\,(1 - P\,(t, A_a))\,, \tag{4}$$

$$P_{R,t} = K_7 + P\,(t, A_R)\,, \tag{5}$$

K_{13} (3.0e-12 mc/mc-s) is the flux of condensed water in the slab; K_8 (5.6704E-8 W/mK3) is the Stephan-Boltzmann constant; K_9 (298 K) is the mean temperature of two faces of polyurethane slab; K_{10} (0.0005 m) is the mean equivalent diameter of the cells and K_{11} (35 Kg/mc) and K_{12} (1200 Kg/mc) are the densities of the foam and the solid, respectively.

In (4), K_{15} (101325 Pa) is the partial pressure of the air outside the refrigeration unit and $P(t, A)$ is a function of (6) that computes the partial pressure of a gas at time t based on the parameter A, given by (7).

$$P\,(t, \beta_i) = \frac{8\sqrt{2}}{1.01895\pi^2}\left(e^{-\beta_i t} - \frac{e^{-9\beta_i t}}{9}\right) \tag{6}$$

$$\beta_i = \left(\frac{\pi}{2K_{16}}\right)^2 \tilde{\beta}_i, \quad i = 2, 3 \tag{7}$$

where K_{16} (0.1 m) is the mean thickness of the polyurethane layer and β_i is a calibrated unknown parameter corresponding to the coefficient of diffusivity of the air or of the compressed gas respectively for $\tilde{\beta}_2$ and $\tilde{\beta}_3$. Equation (6) represents the mean pressure over time and across the polyurethane layer and was derived as a simplification of the diffusivity processes of the gases present in the polyurethane layer, based on Fick's Law (Smith, 2004).

From an engineering point of view, the parameter θ in (1) can be interpreted as the aging velocity of the refrigeration unit. To account for structural characteristics and specifications that might contribute to the overall aging of the isothermal unit, we can represent θ as an additional linear model such as in (8). This allows evaluating the aging results of refrigerated transportation systems that might differ by structural factors or by method of employment.

$$\begin{aligned}\theta = \beta_1 + \beta_4 X_1 + \textstyle\sum \beta_{5j} X_{2j} + \beta_6 X_3 + \beta_7 X_4 + \beta_8 X_5 + \beta_9 X_6 \\ + \beta_{10} X_7 + \beta_{11} X_8 + \beta_{12} X_9 + \beta_{13} X_{10}\end{aligned} \tag{8}$$

where β_1 is a constant, β_i, $i = 4, 6 \ldots, 13$, and β_{5j}, $j = 1, \ldots, 10$, are parameters related to several possible relevant factors potentially affecting the aging velocity. The variables under study are: X_1 : type of use, X_{2j}, $j = 1, \ldots, 10$: type of transported perishables (j = catering, dairy, deep frozen foods, dry, fish, fruit and vegetables, general perishable, ice cream, meat, poultry), X_3 : number of leafs for the second door, X_4 : number of leafs for the first door, X_5 : total perimeter doors, X_6 : presence of refrigerating unit in the vehicle, X_7 : presence of meat rails in the roof of the vehicle, X_8 : average thermal thickness, X_9 : average geometrical thickness, X_{10} : full working status.

In Table 2 the permutation p-values (with 1000 independent random permutations) are obtained for each factor, from a database of nearly 4,000 records of measurements and real aging data, available from 1998 to 2007

at the *Laboratories of Chill Techniques* (LCT) within the Italian National Research Council, Construction Technologies Institute, Padova, Italy. The LCT is one of the centers certified to measure the overall coefficient of heat transfer in transported refrigeration systems.

Table 2. Permutation p-values associated with the analyzed factors affecting the aging curve of refrigerated vehicles.

Factor	Permutation p-value
All factors	0.001
Type of use	0.098
Type of transported perishables	0.040
Number of leafs for the second door	0.010
Number of leafs for the first door	0.001
Total perimeter doors	0.050
Refrigerating unit	0.159
Meat rails in the roof	0.003
Average thermal thickness	0.038
Average geometrical thickness	0.088
Full working	0.001

Results in Table 2 may suggest several practical conclusions. In fact, the more relevant factors affecting the aging curve of refrigerated vehicles are those with a smaller permutation p-value: Number of leafs for the first door and Full working, followed by Meat rails in the roof. When these structural characteristics are introduced in the refrigerated vehicle we can expect a great changing in the related aging curve.

As illustration of partial results, presented in Figure 1 we have four estimated aging curves considering the contribution of two significant detected factors. More precisely:

- A: Meat rails in the roof = NO, Number of leafs for the first door = 0
- B: Meat rails in the roof = YES, Number of leafs for the first door = 0
- C: Meat rails in the roof = NO, Number of leafs for the first door = 2
- D: Meat rails in the roof = YES, Number of leafs for the first door = 2

5 Conclusions

In this work we have introduced a novel algorithm within the nonparametric permutation framework able to perform proper inference on parameters of any specified nonlinear model. As suggested by the simulation study and by

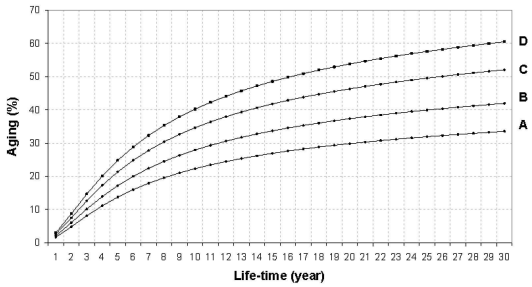


Fig. 1. Four estimated aging curves by considering the contribution of few significant detected factor.

the application to a real case study, we can state that the proposed methods offers a well-rounded approach to make inference on nonlinear models. Therefore, in each situation where the normality assumption is hard to justify or where the null distribution of test statistics is too hard to cope with, the proposed nonparametric procedure can be considered as a valid solution. We believe that in many experimental and observational studies this permutation approach may provide a significant contribution to successful research related to nonlinear and also linear models.

References

BURNHAM, K.P. and ANDERSON, D.R. (2002): *Model selection and multimodel inference: a practical information-theoretic approach (2nd edn)*. Springer-Verlag, New York.

CADE, B.S. and RICHARDS, J.D. (1996): Permutation tests for least absolute deviation regression. *Biometrics, 52, 886–902*.

EDGINGTON, E.S. (1995): *Randomization tests (3rd edn)*. Marcel Dekker.

LAGARIAS, J.C.; REEDS, J.A.; WRIGHT, M.H. and WRIGHT, P.E. (1999): Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM J. Optim. 9 , no. 1, 112–147*.

KENNEDY, P.E. and CADE, B.S. (1996): Randomization tests for multiple regression. *Communications in Statistics. Simulation and Computation, 25, 923-936*.

MIELKE, P.W.JR. and BERRY, K.J. (2001): *Permutation Methods: A distance Function Approach*. Springer Series in Statistics.

PESARIN, F. (2001): *Multivariate Permutation Tests: With Applications in Biostatistics*. Wiley, Chichester.

SICURO, A. (2006): *Modello analitico dell'invecchiamento dei pannelli isolanti, mezzi isotermi e confronto con dati sperimentali*. Degree Thesis, supervisor Prof. C. Bonacina. Dipartimento di Fisica Tecnica, Università di Padova.

SMITH, W.F. (2004): *Foundations of Materials Science and Engineering* 3rd ed., McGraw-Hill.

STAPEL, M. and TER BRAAK, C.F.J. (1994): Randomization and boostrap test in factorial experiments: Does analysis follow from design? *Munster: Dutch-German Biometrics Meeting, 15-18 May, 1994*.

TER BRAAK, C.F.J. (1992): *Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: K.H. Jöckel, G. Rothe and W. Sendler (Eds.): Bootstrapping and related resampling techniques*. Springer, 79–86.

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (1970): *Agreement on the International Carriage of Perishable Foodstuffs and on the Special Equipment to be used for such Carriage (ATP)*. Transportation Division.

Bootstrap Methods for Finding Confidence Intervals of Mahalanobis Distance

Parameshwaran S. Iyer[1] and Anil Kumar Maddulapalli[2]

[1] Researcher, General Motors R&D, India Science Lab, 3rd Floor, Creator Building, ITPB, Whitefield Road, Bangalore - 560008, India.
Phone:(+91)-80-41984350, Fax: (+91)-80-41158562.
parameshwaran.iyer@gm.com
[2] Senior Researcher, General Motors R&D, India Science Lab.
anil.maddulapalli@gm.com

Abstract. In this work, we address the problem of finding the Confidence Intervals (CI) for the Mahalanobis (MH) distance of a new point from the mean of a given dataset. We have just an instance of the dataset and do not have any other information on the underlying probability distributions of the dataset. We look into different methods and in particular investigate the non parametric naïve Bootstrap method for finding the CI of MH distance. We propose modifications to the naïve Bootstrap method, namely, the use of a ratio based pivot function and a "uniform resampling" scheme developed by Barton and Schruben (1993) & (2001). Our results show that the proposed method gives a more consistent and robust performance, in terms of coverage probability, for the scenarios we considered.

Keywords: Mahalanobis distance, confidence intervals & bootstrap

1 Introduction

Often in engineering design optimization, we are faced with the problem of statistically estimating the closeness of a new design concept to a multi dimensional (where each dimension corresponds to an attribute of a design) cluster of existing designs. Mahalanobis (MH) distance is a popular metric, to assess the closeness, as it takes into account the correlation between different dimensions of the existing designs. However the nature of engineering design is such that there is always uncertainty in the data collected for the existing designs. Hence it becomes essential to not only estimate the MH distance but also quantify the uncertainty in the MH distance of the new design concept. Developing a two-sided Confidence Interval (CI) for the MH distance is a good way to capture the uncertainty associated and is especially useful in an optimization scenario. In this work, we assume that we have an instance of the dataset which forms the cluster of existing designs and we look into methods for constructing the two-sided CI's of the MH distance without any information on the underlying probability distributions of the dataset. For finding the CI of MH distance we first used a method developed by Reiser

(2001). The aim herein was to study whether the parametric assumptions (normality of the dataset) of the Reiser's method provide a good approximation for the problem in hand. We also investigate the application of the most popular non parametric approach, the naïve Bootstrap (Efron (1979)) to our problem. We identified some issues with the naïve Bootstrap that results in poor coverage probabilities for the CI of MH distance (Maddulapalli and Iyer (2008)). To overcome these issues, we propose to use a 'ratio pivot function' and 'uniform resampling' scheme (Barton and Schruben (1993)) as modifications to the naïve Bootstrap (see Section 2 for details). We conduct simulation studies to compare three methods: naïve Bootstrap with ratio pivot, uniform resampling Bootstrap with ratio pivot, and Reiser's method. For conducting these simulations, we generate different datasets where the true MH distance is known and then check for the coverage probabilities of CI obtained by the different methods. We designed two data generators (DGI & DGII, refer appendix for details) with known true MH distance for our simulation studies.

2 Bootstrap methodology

The Bootstrap methodology developed by Efron (1979) provides the inferentist with an accurate tool to construct near to exact CI. It is based on the plug in principle (Efron and Tibshirani (1993)) wherein the statistic of interest θ can be estimated using an approximation $\hat{F}$ to the unknown Cumulative Distribution Function(CDF) F of the probability distribution. Let X denote a random p-dimensional dataset of size n and let $t(X)$ be the function used to estimate the statistic $\theta(F)$ (MH distance in our case). Given we observe dataset $X = x$, the naïve Bootstrap algorithm is as follows:

(i) For the parameter of interest, construct a pivotal quantity $R(X, F)$ whose asymptotic distribution does not depend on the unknown F. For some parameters it may not be possible to construct a pivotal quantity, when this pivotal quantity exists its use is recommended.
(ii) Construct the empirical CDF $\hat{F}$ by putting mass $1/n$ at each observation.
(iii) Draw a bootstrap random sample of size n with replacement $X^* = x^*$ from $\hat{F}$.
(iv) Approximate the sampling distribution of R^* by obtaining its value at each bootstrap resample from $R^*(X^*, \hat{F})$

In a complex scenario where a closed form expression for the distribution of R^* does not exist, the steps 3-4 are repeated B times to generate B resamples of X^* and R^*. Plug-in principle states that, assuming $R(X, F)$ exists and has an asymptotic distribution, we can approximate the distribution of R using the Bootstrap distribution of R^*. Using this information we can now construct approximate CI for $t(X)$ from the quantiles of R^*. The choice of R, $\hat{F}$, B and methods to determine quantiles of R^* have been discussed frequently in literature for a wide range of statistics.

2.1 Choice of pivot function

Efron (1979) presents the bias estimator, $student - t$ statistic and $t(X)$ as choices for R. These choices of pivot functions have been extremely popular in the literature and have been used in a variety of applications (DiCiccio and Efron (1996) & Efron and Tibshirani (1993)). Bickel and Freedman (1981) and Babu (1995) provide counter examples where these pivot functions fail, in particular they conclude that these pivot forms work well with statistics having pivotal quantities in the form of the t-statistic. Due to the asymptotic normality properties of the bias estimator and $student - t$ pivots, these could result in negative values for CI for the MH distance which is not desirable. Choice of the observed MH distance as a pivot function, $R = t(X)$, will always result in positive CI's unlike the bias and $student - t$ based pivot functions. However our results show that this choice fails to produce the desired coverage (Maddulapalli and Iyer (2008)). Hence, we suggest using the ratio of MH distance of the new point from the mean of each Bootstrap resample (i.e., $t(X^*)$) and the nominal MH distance (i.e., $\hat{\theta}$) as the pivot function as in equation 1.

$$R(X, F) = \frac{t(X)}{\theta(F)} \qquad R^*(X^*, \hat{F}) = \frac{t(X^*)}{\hat{\theta}} \tag{1}$$

Choice of pivot of the form in equation 1 assures us of the non-negativity of the two sided CI for the MH distance. When the underlying unknown process from which the data is generated follows a normal distribution, Krazanowski (1988) showed that the observed MH distance of a new point independent of its sample mean and covariance follows a scaled f-distribution as in equation 2.

$$t(X) \ \sim \ \frac{p(n^2-1)}{n(n-p)} f_{p,n-p} \tag{2}$$

Note that in Reiser's method (Reiser (2001)), the new point is sampled from a normal distribution with a similar true covariance as the cluster of existing points resulting in a scaled non-central f-distribution for $t(X)$. Irrespective of whether the new point is independent to the existing cluster or not, the observed MH distance follows some form of f-distribution. Also, the true MH distance θ follows a χ^2_p distribution (Krazanowski (1988)) when the underlying distribution is normal. Hence, the distribution of R in equation 1 is a ratio of f distribution and χ^2_p distribution. As $n \to \infty$ the limiting distribution of R is a f-distribution, when the dataset of interest follows a normal distribution. In such cases with this choice of pivot function, naïve Bootstrap resampling should result in the desired coverage for the CI. When the underlying distribution is non-normal, we recommend a different choice of the resampling distribution in Section 2.2.

2.2 Choice of $\hat{F}$

The empirical $\hat{F}$ used in the bootstrap algorithm by Efron (1979) is obtained by placing a probability mass of $1/n$ on each observation. Efron suggests the use of other estimates of $\hat{F}$ based on problem contexts. There have been various contributions in literature for identifying the optimal weighting scheme. Importance resampling methods are implemented in Johns (1988) to reduce the amount of Bootstrap resamples required to generate CI. Advantages of using weighted Bootstrap schemes have been shown in Hall and Presenell (1999). Most of the approaches discussed above associate non random weights to each observation in each Bootstrap iteration. In what follows next, we describe the Uniform Resampling Methodology, which assigns a random weight to each observation in every Bootstrap iteration. This provides a kind of flexibility to the Bootstrap algorithm by allowing resampling to occur under a family of F. The uniform resampling method proposed by Barton and Schruben (1993) & (2001) can be explained as follows: Given a set of observations X^i, we know the unknown CDF $F(X^i) \sim U(0,1)$. Let $X^{(k)}$ be the n order statistics of the random variable X^i $(k = 1 \cdots n)$, then even when the true value of $F(X^{(k)})$ is unknown due to the unknown F, its distribution is known. The joint distribution of $(F(X^{(1)}), F(X^{(k)}), \cdots, F(X^{(n)}))$ is an n order statistics of a uniform distribution given by a multivariate uniform distribution. The marginal distribution of $F(X^{(k)})$ is a $beta(k, n-k+1)$. The expectation of $F(X^{(k)})$ which follows a *beta* distribution is $k/(n+1)$ which corresponds in structure to the standard empirical form of CDF. Using these results we modify the second step of the Bootstrap algorithm (recall section 2) as follows:

2 *We then construct the empirical CDF* $\hat{F}$ *as follows:*
   (a) *Generate* n $U(0,1)$ *and obtain the* n *order statistics* $u^{(1)}, u^{(2)}, \cdots, u^{(n)}$ *(Generating from the Beta distributions for each order would also yield the same results).*
   (b) *Obtain the order statistics for* X^i*, and set* $\hat{F}(X^{(i)}) = u^{(i)}$

Obtaining order statistics $X^{(i)}$ is trivial in the univariate case (as done in Barton and Schruben (1993) & (2001)) and not so trivial in the multivariate case. *We propose to use the MH distance of a point from the sample mean as a criterion to obtain the multivariate order statistics, for our problem.*

2.3 Selecting confidence intervals and the choice of B

Confidence intervals are delicate tools and hence require additional computations to achieve the required precision, thereby requiring a large B. With MH distance resamples, one can construct the empirical CDF of the MH distance distribution and from the quantiles of this CDF construct a CI. This method is referred to as the percentile method and it is found to be first order correct (Efron and Tibshirani (1993)). When a monotone transformation that

transforms the unknown θ to normality exists, it is found that the CI from the percentile method performs accurately (Efron and Tibshirani (1993)). In our case for the form of pivot function we suggest, $R > 0$ as it is a ratio of distances. Hence a log transformation which is monotone would transform the pivot function to approximate normality. Such transformations to approximate normality have been used constantly in literature, e.g. Box-Cox, log and square root. An advantage of the percentile method is that it is not required to know what the transformation actually is as it is incorporated automatically. In our applications we use the percentile method to obtain the CI and fix the number of resamples to be 500 (Efron and Tibshirani (1993) recommend 250 as a good choice).

3 Simulation study and discussion of results

We conduct various simulation studies to evaluate the performance of the two different resampling schemes (naïve & uniform resampling schemes) and choice of our proposed pivot function (recall equation 1). For our experiments, we generate datasets using the data generators DGI and DGII. We fix $B = 500$ and obtain 90% (i.e. $\alpha = 0.1$) coverage probabilities for 5000 datasets generated for each data generator. We use the percentile method to construct the CI. In our simulation studies, we also implement the Reiser's method for obtaining CI (Reiser (2001)) to compare a parametric approach with the Bootstrap technique. Table 1 shows the results of our first simulation

Table 1. True Coverage of MH Distance from New Concept Point Defined by $scale = 2$, from datasets obtained by DGI.

Dims (True MH Dist)	Number of Sample Points	Bootstrap Coverage	Bootstrap CI Length	Uniform Resampling Bootstrap Coverage	Uniform Resampling Bootstrap CI Length	Reiser Coverage	Reiser CI Length
2(1.10)	20	0.83	4.02	0.90	7.20	0.98	7.79
2(1.10)	30	0.84	3.05	0.91	5.67	0.99	6.94
3(1.12)	30	0.80	3.02	0.91	4.19	0.94	5.82
3(1.12)	45	0.81	2.32	0.91	3.38	0.95	5.09
4(1.16)	40	0.78	4.32	0.87	5.52	0.85	6.19
4(1.16)	60	0.80	3.07	0.88	4.07	0.84	4.71

study using DGI and a new point with scale 2 (refer Appendix for details). We note that the dataset generated from DGI is non-normal. From Table 1, we observe that uniform resampling methodology generally gives better coverage than the naïve Bootstrap resampling methodology. Also the average CI lengths from uniform resampling methodology are generally less than the CI lengths from Reiser's method. Naïve Bootstrap resampling methodology generally gives less than required nominal coverage and Reiser methodology generally gives way more or less than the required nominal coverage depending on the number of dimensions. From the above results, taking into account

both coverage probability and CI lengths, we can conclude that for this simulation study, uniform resampling methodology is better than naïve Bootstrap resampling and Reiser methodologies. Table 2 shows the results of our second simulation study. In this simulation study, we use DGII with $Z_{\cdot i} \sim U(1, 20)$ and a new point with scale 5 (refer Appendix for details). We note that the datasets generated in the above mentioned way are close to multi-variate normality. From Table 2 we observe that the naïve Bootstrap resampling methodology gives close to required nominal coverage. The uniform resampling methodology generally gives more than the required nominal coverage. Reiser methodology is a very inconsistent giving more than required nominal coverage for lower dimensions and less than required nominal coverage for higher dimensions. So, we conclude that Reiser's methodology is outperformed by other two Bootstrap methodologies in this simulation study. From Table 2, we can see that the average CI lengths from the naïve Boot-

Table 2. True Coverage of MH Distance from New Concept Point Defined by $scale = 5$, from datasets obtained by DGII with $Z_{\cdot i} \sim U(1, 20)$.

Dims (True MH Dist)	Number of Sample Points	Bootstrap Coverage	Bootstrap CI Length	Uniform Resampling Bootstrap Coverage	Uniform Resampling Bootstrap CI Length	Reiser Coverage	Reiser CI Length
4(6.37)	40	0.89	5.39	0.95	7.13	1.00	14.44
4(6.37)	60	0.90	4.31	0.96	5.85	1.00	14.06
8(11.86)	80	0.89	6.88	0.94	8.83	0.99	18.24
8(11.86)	120	0.90	5.53	0.95	7.32	1.00	17.98
12(17.33)	120	0.89	8.11	0.92	10.34	0.87	21.58
12(17.33)	180	0.89	6.48	0.94	8.51	0.91	21.30
16(22.80)	160	0.89	9.22	0.91	11.70	0.67	24.77
16(22.80)	240	0.90	7.32	0.94	9.60	0.66	24.35

strap methodology are generally smaller than the CI lengths from the uniform Bootstrap methodology. So, from the results of this simulation study we can conclude that naïve Bootstrap resampling methodology out-performs the uniform resampling methodology. However, note that the datasets generated in this simulation study are close to normality. We mentioned in Section 2.1 that when the underlying process is close to normality the pivot function we proposed would work well with the naïve Bootstrap resampling scheme because of the asymptotic properties of the pivot function. Since in practice we cannot guarantee the normality of underlying process, we still like to think that uniform resampling methodology is better than naïve Bootstrap methodology even though uniform resampling methodology could be conservative (i.e., more than required nominal coverage) for datasets close to normality. Table 3 shows the results of our third simulation study. For the third simulation study, we use datasets generated from DGII with $Z_{\cdot i} \sim Weibull(0.5, 0.5)$ and a new point with scale 7. We note that the datasets generated in the above mentioned way are non-normal. From Table 3, we can see that uniform re-

sampling methodology gives close to required nominal coverage. The naïve Bootstrap resampling methodology gives consistently less than the required nominal coverage. On the other hand, the Reiser's methodology is very inconsistent with the method resulting in more than required nominal coverage for higher dimensions and less than required nominal coverage for lower dimensions. This behavior is expected from Reiser's method because the dataset used in this study is not normal and the new concept point is fixed. From the

Table 3. True Coverage of MH Distance from New Concept Point Defined by $scale = 7$, from datasets obtained by DGII with $Z_{\cdot i} \sim Weibull(0.5, 0.5)$.

Dims (True MH Dist)	Number of Sample Points	Bootstrap Coverage	Bootstrap CI Length	Uniform Resampling Bootstrap Coverage	Uniform Resampling Bootstrap CI Length	Reiser Coverage	Reiser CI Length
4(20.66)	40	0.68	52.64	0.83	61.21	0.59	59.44
4(20.66)	60	0.69	38.02	0.83	45.70	0.69	47.28
8(38.44)	80	0.67	48.62	0.86	54.09	0.70	63.79
8(38.44)	120	0.71	36.36	0.88	41.86	0.87	54.58
12(56.42)	120	0.71	48.21	0.89	53.56	0.84	69.72
12(56.42)	180	0.70	36.76	0.88	42.38	0.94	61.64
16(74.17)	160	0.70	48.88	0.89	54.77	0.90	75.74
16(74.17)	240	0.69	37.77	0.89	43.78	0.97	67.98

above three simulation studies, we conclude that uniform resampling with the pivot function given in equation 1 is the best method given no information on the underlying probability distributions of the datasets. Hence we recommend these choices of R and $\hat{F}$ for computing the confidence intervals of the unknown Mahalanobis distance. We refer readers to Maddulapalli and Iyer (2008) for more simulation studies which support our conclusions.

4 Summary

In this paper, we analyzed different Bootstrap methods for finding the confidence intervals of the Mahalanobis distance of a new point from the mean of a given instance of the dataset. We recommend the use of a ratio pivot function given by equation 1 and uniform resampling scheme (Barton and Schruben (1993) & (2001)) as modifications to the naïve Bootstrap. We have conducted simulation studies to compare three methods: namely, naïve Bootstrap with a ratio pivot, uniform resampling Bootstrap with a ratio pivot and Reiser's method (Reiser (2001)). Our simulation studies clearly show that uniform resampling Bootstrap with a ratio pivot gives robust performance when compared to the other two methods.

5 Appendix

DGI can generate non-normal datasets having up to 4 dimensions. The structure of DGI is as follows:

$$X_1 \sim Exp(a); \quad X_2 \sim Exp(b); \quad X_3 \sim U(0, X_1 + X_2); \quad X_4 \sim U(0, X_1 \times X_2)$$

a & b are set to 1 & 1.5 for the simulation study. DGII can generate observations for any number of dimensions. The structure of DGII for generating n points in p dimensions is as follows:

$$Z_{\cdot i} \sim g(A); \quad X_{\cdot i} = Z_{\cdot i} + \rho \frac{Z_{\cdot (i+1)}}{(1+\rho)^2}, \ \forall \ i = 1 \cdots p$$

$Z'_{\cdot i}s$ are independent, $g(A)$ is a known distribution with fixed parameters A. Also ρ is a fixed correlation constant and we assume a value of 0.5 for it in our experiments. The different forms of $g(A)$ we consider in our experiments are $U(1, 20)$ and a $Weibull(0.5, 0.5)$. For the uniform case DGII produces normal data and non-normal data for the weibull case. We investigate $Q-Q$ plots to check for the normality of the data (Maddulapalli and Iyer (2008)). We define the new point y as $y = scale \times 1_{dim \times p}$ and obtain the true MH distance using a Monte Carlo simulation. The scale determines how close the new concept point is to the mean of the observed dataset and hence the magnitude of the MH distance.

References

BABU, G. (1995): Bootstrap for nonstandard cases. *Journal of Statistical Planning and Inference 43, 197-203.*

BARTON, R.R., and SCHRUBEN, L.W. (1993): Uniform and bootstrap resampling of empirical distributions. In: *WSC '93: Proceedings of the* 25^{th} *conference on Winter simulation.* ACM, New York, NY, USA, 503–508.

BARTON, R.R., and SCHRUBEN, L.W. (2001): Resampling methods for input modeling In: *WSC '01: Proceedings of the* 33^{rd} *conference on Winter simulation.* IEEE Computer Society, Washington, DC, USA, 372–378.

BICKEL, P., and FREEDMAN, D. (1981): Some asymptotic theory for the bootstrap. *Annals of Statistics 9(6), 1196-1217.*

DICICCIO, T.J., and EFRON, B. (1996): Bootstrap confidence intervals. *Statistical Science 11, 189-228.*

EFRON, B. (1979): Bootstrap methods: another look at the jackknife. *Annals of Statistics 7(1), 1-26.*

EFRON, B., and TIBSHIRANI, R. (1993): *An introduction to bootstrap: monographs on statistics and applied probability.* Chapman and Hall.

HALL, P., and PRESENELL, B. (1999): Intentionally biased bootstrap methods. *Journal of Royal Statistical Society, Series B Statistical Methodology 61, 143-158.*

JOHNS, M. (1988): Importance sampling for bootstrap confidence intervals. *Journal of American Statistical Association 83 (403), 709-714.*

KRAZANOWSKI, W. (1988): *Principles of Multivariate Analysis.* Oxford: Oxford University Press.

MADDULAPALLI, A.K., and IYER, P.S. (2008): Methods for finding confidence intervals of mahalanobis distance. *General Motors R&D Report 11081.*

REISER, B. (2001): Confidence intervals for the mahalanobis distance. *Communications in Statistics - Simulation and Computation 30, 37-45.*

Test of Mean Difference for Paired Longitudinal Data Based on Circular Block Bootstrap

Hirohito Sakurai[1] and Masaaki Taguri[2]

[1] Division of Computer Science, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo 060–0814, Japan,
sakurai@main.ist.hokudai.ac.jp
[2] National Center for University Entrance Examinations
2–19–23 Komaba, Meguro-ku, Tokyo 153–8501, Japan, *taguri@rd.dnc.ac.jp*

Abstract. This paper proposes a testing method for detecting the difference of two means or mean curves in paired longitudinal data based on the circular block bootstrap. For the detection of mean difference, we utilize four types of test statistics. Monte Carlo simulations are carried out in order to examine the sizes and powers of the proposed test.

Keywords: circular block bootstrap, longitudinal data, resampling, comparison of mean curves

1 Introduction

Suppose that there are paired two samples given by $\{(Y_i(t), X_i(t))\}_{i=1}^{q}$ for $t = 1, \ldots, n$, where $Y_i(t)$ and $X_i(t)$ are continuous in t, and assume that, for fixed t, $Y_1(t), \ldots, Y_q(t)$ are independent over q subjects, and that $X_1(t), \ldots, X_q(t)$ are independent over q subjects. Then we consider the model

$$Y_i(t) = f(t) + \varepsilon_i(t), \quad X_i(t) = g(t) + \eta_i(t), \quad i = 1, \ldots, q, \ t = 1, \ldots, n, \quad (1)$$

where $f(t)$ and $g(t)$ are unknown regression functions, and $\varepsilon_i(t)$ and $\eta_i(t)$ are the error terms having means 0 and finite variances. More general formulations for (1) can be found, for example, in Hall and Hart (1990) and references therein. Our problem is then to test

$$\begin{cases} H_0 : f(t) = g(t) \text{ for all } t, \\ H_1 : f(t) \neq g(t) \text{ for some } t, \end{cases} \quad (2)$$

where H_0 and H_1 are the null and alternative hypotheses, respectively.

For example, Figure 1 shows the wind velocity data which are obtained by an artificial satellite (left panel) and a radar (right panel) on the earth, and are measured at altitudes from 80 km to 90 km every 1 km ($(q =)11$

subjects) during ($n =$)13 days. For these data, we want to know whether the mean behavior of the two devices in measuring wind velocity is equal or not. Then the problem is formulated as (2) and the significant difference between them is detected by some methods, which is briefly explained in Section 4.

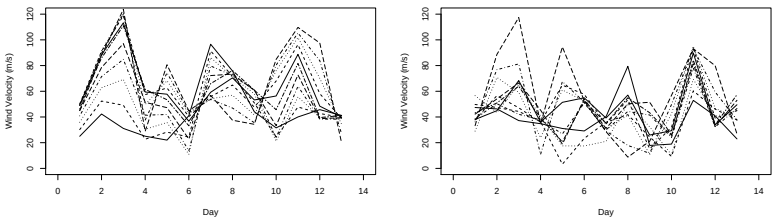


Fig. 1. Wind velocity data (left: satellite, right: radar).

Several methods about the comparison of two means or regression curves have been proposed, and most of them assume that the error terms are independent and identically distributed (i.i.d.), whose distributions are normal. However, when we analyze a real dataset, it may be unrealistic to put such assumptions. If we cannot assume the normality, the nonparametric approach, for example by Bowman and Young (1996), is available. Another possible approach would be an application of resampling such as nonparametric bootstrap.

The bootstrap proposed by Efron (1979) is a widely used resampling method in statistics and other research fields. However, in some settings of dependent data, for example, in time series analysis, the naive bootstrap usually fails to capture the dependent structure of data because it ignores the order of observations. In order to overcome this problem, two kinds of bootstrap methods are proposed. One is a model-based approach which resamples from approximately i.i.d. residuals, and the other is a nonparametric, purely model-free bootstrap scheme, which resamples from blocks of observations; see, for example, Bühlmann (2002), Härdle et al. (2003), Lahiri (2003), and references therein.

This paper concerns with the case where paired longitudinal data from two groups are given, and proposes a testing method for detecting the difference between two means or two mean curves in longitudinal data based on the circular block bootstrap approach (Politis and Romano (1992)). As seen in the numerical examinations given below, our approach is superior to Bowman and Young's (1996) test.

The rest of this paper is constructed as follows. Section 2 proposes a testing procedure which generates the resamples corresponding to two samples by circular block bootstrap and calculates a p-value (achieved significance level). In order to investigate the properties of sizes and powers of the pro-

posed testing method, Monte Carlo simulations are carried out in Section 3, and some concluding remarks are summarized in Section 4.

2 Testing method

To detect the difference between $f(t)$ and $g(t)$ in (1), we also focus attention on the area-difference given by $A = \int |f(t) - g(t)|\, dt$. Note here that the quantity A is 0 under H_0 and positive under H_1. Thus, the hypothesis of interest reduces to testing

$$H_0 : A = 0 \quad \text{vs.} \quad H_1 : A > 0. \tag{3}$$

There are some approaches to detecting the difference between two mean curves, $f(t)$ and $g(t)$, in (1). In this paper, our interest concentrates on the behavior of four types of test statistics given below. The following statistic is proposed by Hall and Hart (1990):

$$S_n = S_n(D_1, \ldots, D_n) = \left[\sum_{j=0}^{n-1} \left(\sum_{t=j+1}^{j+h} D_t \right)^2 \right] \left[n \sum_{t=1}^{n-1} \frac{(D_{t+1} - D_t)^2}{2} \right]^{-1}, \tag{4}$$

where $D_t = Y_t - X_t$ for $t = 1, \ldots, n$ or $D_t = Y_{t-n} - X_{t-n}$ for $t = n+1, \ldots, n+h$, $Y_t = \sum_{i=1}^{q} Y_i(t)/q$, $X_t = \sum_{i=1}^{q} X_i(t)/q$, $h = [np]$ is the integer part of np, and p is a tuning constant satisfying $0 < p < 1$ which is determined by the fully data-driven approach; the second approach described in Hall and Hart (1990, pp.1043–1044). The statistic (4) is essentially based on kernel estimators of $f(t)$ and $g(t)$. As another type of test statistics, we can consider

$$T_{1n} = T_{1n}(D_1, \ldots, D_n) = \sum_{t=1}^{n} |D_t|, \quad T_{2n} = T_{2n}(D_1, \ldots, D_n) = \sum_{t=1}^{n} D_t^2. \tag{5}$$

In addition to (4) and (5), we here also focus attention on area-difference, $A = \int |f(t) - g(t)|\, dt$, and then consider the following test statistic:

$$T_{3n} = T_{3n}(D_1, \ldots, D_n) = \frac{1}{2} \sum_{t=1}^{n-1} (|D_t| + |D_{t+1}|) I_1 + \frac{1}{2} \sum_{t=1}^{n-1} \frac{|D_t|^2 + |D_{t+1}|^2}{|D_t| + |D_{t+1}|} I_2, \tag{6}$$

where $I_1 = I\{D_t D_{t+1} \geq 0\}$, $I_2 = I\{D_t D_{t+1} < 0\}$ and $I\{\cdot\}$ is the indicator function. The test statistic T_{3n} seems to have a complicate form, however it is an estimator of A in (3) constructed by the trapezoidal rule with linear interpolations of adjacent observation values.

The values of S_n and T_{rn} $(r = 1, 2, 3)$ may be small when H_0 is true, and large when H_0 is false. Therefore, the above four statistics enable us to measure the discrepancy between $f(t)$ and $g(t)$.

Let $D_{0,t} = D_t - \bar{D} = D_t - \sum_{t=1}^{n} D_t/n$ for $t = 1, \ldots, n$, and d_t and $d_{0,t}$ denote realizations of D_t and $D_{0,t}$. In this section, we propose a nonparametric testing method for the problem (3) using (4), (5) and (6). The main ideas of our testing method is that we apply the circular block bootstrap to the centered observations $\{d_{0,1}, \ldots, d_{0,n}\}$ in order to approximate the null distribution of the test statistics.

For simplicity, let T be a generic notation for S_n, T_{1n}, T_{2n} or T_{3n}. For a given significance level α, the proposed testing algorithm together with Monte Carlo method is described as follows.

(i) For the centered observations $\{d_{0,1}, \ldots, d_{0,n}\}$ define n blocks each with length l, in the manner of Politis and Romano (1992), where each block is

$$\xi_j = \begin{cases} \{d_{0,j}, \ldots, d_{0,j+l-1}\}, & j = 1, \ldots, n-l+1, \\ \{d_{0,j}, \ldots, d_{0,n}, d_{0,1}, \ldots, d_{0,j+l-n-1}\}, & j = n-l+2, \ldots, n. \end{cases}$$

(ii) Draw m blocks $\{\xi_1^{*b}, \ldots, \xi_m^{*b}\}$ randomly with replacement from $\{\xi_1, \ldots, \xi_n\}$ to obtain a resample $\{d_1^{*b}, \ldots, d_n^{*b}\}$ $(b = 1, \ldots, B)$ of size n, and calculate $t^{*b} = T(d_1^{*b}, \ldots, d_n^{*b})$, where $m = [n/l]$ (if n/l is an integer) or $m = [n/l] + 1$ (otherwise).

(iii) Calculate the achieved significance level, $\widehat{\mathrm{ASL}} = \sum_{b=1}^{B} I\{t^{*b} \geq t_{obs}\}/B$, by repeating step 2 an appropriate number of times B, and reject H_0 if $\widehat{\mathrm{ASL}} \leq \alpha$, where $t_{obs} = T(d_1, \ldots, d_n)$ and α is the significance level, respectively.

We call this testing procedure "Circular Block Bootstrap (CBB) test," and abbreviate it to "CBB test" in the following sections.

3 Numerical examination

We conduct Monte Carlo simulations to investigate the size and power properties of our testing procedure proposed in Section 2. Further, we carry out Bowman and Young's (1996, p.85) test for paired data (hereafter termed "BY" for short) in order to clarify the properties of our test. In the level and power studies, the nominal level is $\alpha = 0.05$ and 0.10. All our results are based on independent 2000 simulation replications of paired two samples, $\{(Y_i(t), X_i(t))\}$, where $B = 2000$ replications of resampling are applied to every two samples in our test. Naturally, the same initial samples are used for the comparisons.

We generate initial samples according to (1) whose means are specified by $f(t) = c$ and $g(t) = 0$, where $c = 0, 0.2, 0.4, 0.6, 0.8, 1.0$. The case of $c = 0$ or $c \neq 0$ corresponds to the null hypothesis or the alternative hypothesis being true. The values, q and n, are $q = 10, 20, 30$ and $n = 10$. The size of n may be small, however it is useful in practice to examine such a behavior of similar

settings for the real data described in Section 1. As for the error terms $\varepsilon_i(t)$ and $\eta_i(t)$, if $f(t)$ and $g(t)$ explain most of the correlation structure contained in the data, then we may consider that the errors are nearly i.i.d., or very weak dependency exists in $\varepsilon_i(t)$ and $\eta_i(t)$. Thus, we choose the following Gaussian AR(1) errors: $\varepsilon_i(t) = \phi\varepsilon_i(t-1) + z_{1i}(t)$ and $\eta_i(t) = \phi\eta_i(t-1) + z_{2i}(t)$, where $z_{1i}(t) \overset{i.i.d.}{\sim} N(0, \tau^2)$, $z_{2i}(t) \overset{i.i.d.}{\sim} N(0, \tau^2)$, $\phi = 0, \pm 0.1, \pm 0.2$, $\tau^2 = (1-\phi^2)\mathrm{V}(\varepsilon_i(t)) = (1-\phi^2)\mathrm{V}(\eta_i(t))$, and $\mathrm{V}(\varepsilon_i(t)) = 1, 2, 3, 4, 5$. The computation has been carried out for all combinations of these parameters, however, to save space, we restrict ourselves to discussing the case of $\alpha = 0.05$ and $\mathrm{V}(\varepsilon_i(t)) = 1, 3$.

In CBB test, it contains a parameter l, namely block length, to be estimated. Since it is preferable that the empirical level is nearly equal to the nominal level α, our choice of l is then the length where the empirical level is close to α. If there are some candidates which have the same level errors, we make the conservative choice, *viz.*, we choose the block length such that the empirical level is less than the nominal level. Further if there are some candidates whose empirical levels are equal, we select the length where the empirical power is the highest among them. The resulting block lengths are given in Table 1.

		$\mathrm{V}(\varepsilon_i(t)) = 1$				$\mathrm{V}(\varepsilon_i(t)) = 3$			
q	ϕ	T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n
10	−0.2	8	5	4	9	8	8	4	8
	−0.1	5	2	2	8	6	2	2	6
	0	4	1	1	8	4	1	1	5
	0.1	2	1	1	7	2	1	1	4
	0.2	1	1	1	5	1	1	1	4
20	−0.2	7	6	4	9	7	7	4	9
	−0.1	6	3	2	9	6	2	2	8
	0	4	1	1	8	4	1	1	7
	0.1	2	1	2	8	2	1	2	6
	0.2	1	1	2	7	1	2	2	5
30	−0.2	8	3	3	9	8	4	4	9
	−0.1	6	3	2	9	7	3	3	9
	0	3	1	1	9	4	2	1	8
	0.1	1	1	1	9	2	1	1	7
	0.2	1	1	1	9	1	1	1	6

Table 1. Optimum block length in CBB test for $\alpha = 0.05$.

Now, we first summarize the results of the level studies. The empirical level of BY and our tests are given in Table 2. From this table, we can observe that BY test has a large level error, while CBB test has a tendency to keep the

nominal level α. In particular, as for the comparison among four statistics T_{rn} $(r = 1, 2, 3)$ and S_n in our test, the level error for T_{rn} $(r = 1, 2, 3)$ is small when $\phi \leq 0$. For $\phi > 0$, the level error for T_{1n} and S_n is small, however that for T_{2n} and T_{3n} seems to be large. We have also observed that the resulting block length in CBB test is short for $\phi \geq 0$ and long for $\phi < 0$ as is shown in Table 1, and that S_n needs longer block length than T_{rn} $(r = 1, 2, 3)$ to keep the nominal level for both cases. The resulting block lengths for S_n are greater than or equal to 4, while those for T_{1n}, T_{2n} and T_{3n} are less than or equal to 3 in most cases.

		$\mathrm{V}(\varepsilon_i(t)) = 1$					$\mathrm{V}(\varepsilon_i(t)) = 3$				
q	ϕ	T_{1n}	T_{2n}	T_{3n}	S_n	BY	T_{1n}	T_{2n}	T_{3n}	S_n	BY
10	−0.2	0.044	0.051	0.054	0.039	0.548	0.046	0.052	0.054	0.053	0.541
	−0.1	0.052	0.048	0.051	0.041	0.546	0.050	0.044	0.051	0.050	0.543
	0	0.053	0.051	0.050	0.046	0.545	0.051	0.044	0.050	0.059	0.532
	0.1	0.057	0.073	0.076	0.052	0.546	0.051	0.064	0.076	0.041	0.530
	0.2	0.059	0.098	0.118	0.053	0.537	0.054	0.087	0.111	0.052	0.529
20	−0.2	0.047	0.057	0.051	0.028	0.431	0.048	0.053	0.055	0.053	0.440
	−0.1	0.057	0.044	0.039	0.048	0.450	0.060	0.046	0.047	0.051	0.455
	0	0.053	0.044	0.056	0.039	0.447	0.046	0.049	0.054	0.049	0.444
	0.1	0.048	0.068	0.086	0.050	0.455	0.051	0.069	0.086	0.046	0.458
	0.2	0.070	0.100	0.117	0.048	0.445	0.068	0.097	0.115	0.060	0.449
30	−0.2	0.046	0.042	0.048	0.018	0.409	0.049	0.047	0.058	0.037	0.421
	−0.1	0.058	0.052	0.050	0.024	0.416	0.058	0.039	0.049	0.054	0.422
	0	0.051	0.051	0.064	0.035	0.416	0.059	0.051	0.053	0.043	0.416
	0.1	0.052	0.071	0.092	0.046	0.416	0.050	0.070	0.085	0.041	0.420
	0.2	0.066	0.107	0.127	0.063	0.414	0.059	0.094	0.121	0.049	0.416

Table 2. Empirical level for $\alpha = 0.05$.

Next, we discuss the power studies. Since we found similar tendencies among the fifteen cases of $(q, \mathrm{V}(\varepsilon_i(t)))$, we show the results for $(q, \mathrm{V}(\varepsilon_i(t))) = (10, 1), (10, 3), (30, 1), (30, 3)$ with $\phi = 0, 0.1, -0.2$. The empirical power seems to be affected by the number of subjects q as well as the variance of noise. The increase of noise variance causes the decrease of empirical power as a whole, however the power properties among the five variances given above are nearly equal to each other. Thus, we choose and discuss the case where $\mathrm{V}(\varepsilon_i(t)) = 1, 3$.

Figure 2 compares the empirical power of CBB test with that of BY test for $\phi = 0, 0.1, -0.2$ (on the left, in the middle and on the right). The dashed line with circles and other four lines correspond to BY and CBB tests; CBB

tests for T_{1n}, T_{2n}, T_{3n} and S_n are corresponding to dotted, dashed, solid and dashed-dotted lines, respectively. Since the bad behavior of BY test was observed in our level study, we omit the results of BY test and our discussion below concentrates on the properties of CBB test.

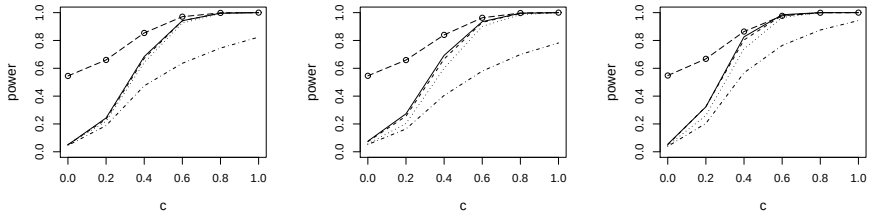


Fig. 2. Empirical power of CBB and BY for $q = 10$ and $\mathrm{V}(\varepsilon_i(t)) = 1$.

Figures 3 and 4 show the empirical powers corresponding to the four test statistics, (4) – (6), for $\phi = 0, 0.1, -0.2$ (on the left, in the middle and on the right). Each line represents the same meaning as in Figure 2. These figures show that the empirical power of T_{3n} is most powerful among them, and that the relationship among powers corresponding to the four test statistics is given by $T_{3n} \geq T_{2n} \geq T_{1n} \geq S_n$ in most cases. This indicates the numerical superiority of our test using T_{3n} in power. As the number of subjects increases, the empirical power is improved. Within the values of $0 \leq c \leq 1$, the empirical power of T_{1n} is nearly equal to that of T_{2n}, however T_{2n} is slightly higher than T_{1n}.

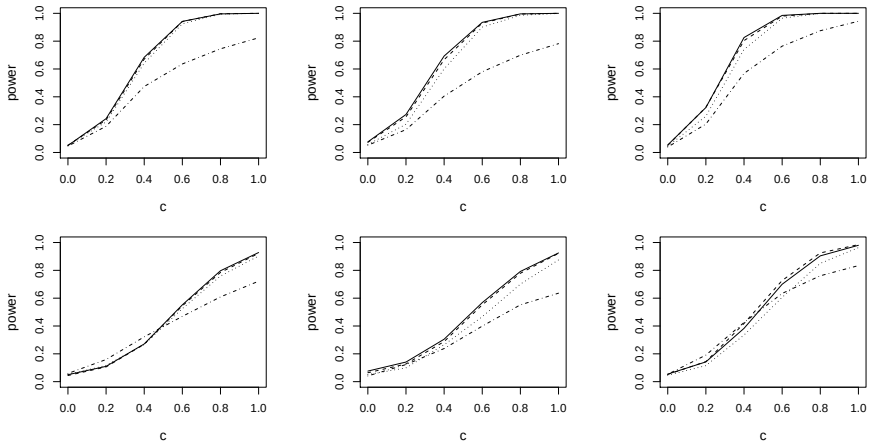


Fig. 3. Empirical power of CBB for $q = 10$ and $\mathrm{V}(\varepsilon_i(t)) = 1$ (upper), 3 (lower).

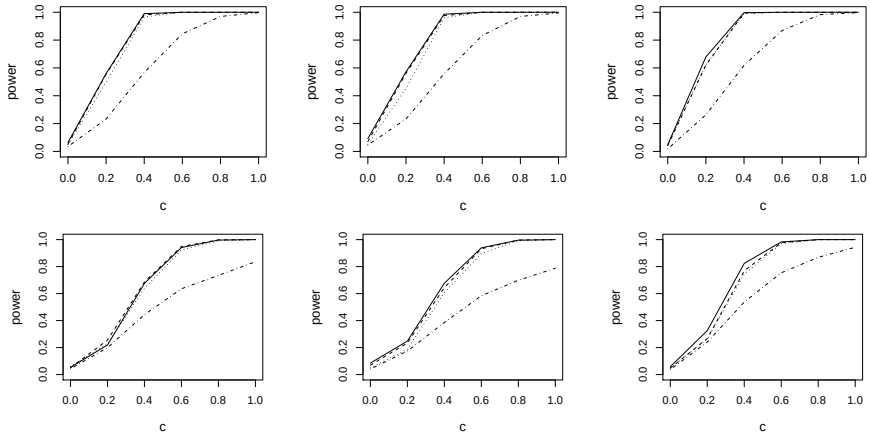


Fig. 4. Empirical power of CBB for $q = 30$ and $\mathrm{V}(\varepsilon_i(t)) = 1$ (upper), 3 (lower).

4 Concluding remarks

In this paper we have proposed a testing method for two means in paired longitudinal data based on the circular block bootstrap. Our numerical studies indicate the applicability of CBB test for weakly dependent data even when the sample size is very small. In some cases the effectiveness of application of our testing algorithm could be confirmed.

Applying our CBB with every possible block length and BY tests to the data given in Figure 1, we obtain the results that $\widehat{\mathrm{ASL}}$'s of CBB test for T_{1n} are 0.046, 0.044, 0.020, 0.008, 0.005, 0.001 for $l = 1, \ldots, 6$ and 0.000 for $l = 7, \ldots, 12$; those for T_{2n} are 0.014, 0.017, 0.009, 0.003 for $l = 1, \ldots, 4$ and 0.000 for $l = 5, \ldots, 12$; those for T_{3n} are 0.004 for $l = 1$ and 0.000 for $l = 2, \ldots, 12$; those for S_n are 0.213, 0.204, 0.209, 0.151, 0.105, 0.088, 0.085, 0.085, 0.097, 0.040, 0.038, 0.016 for $l = 1, \ldots, 12$. BY test rejects the null hypothesis in (3). Therefore, there is a possibility of the significant difference between the satellite and radar in measuring wind velocity. However, the problem on block length selection in the block resampling is very important, and the development of a fully data-driven approach to selecting block length in the CBB test will be needed for practical data analyses.

References

BOWMAN, A. and YOUNG, S. (1996): Graphical comparison of nonparametric curves. *Applied Statistics 45 (1), 83–98.*

BÜHLMANN, P. (2002): Bootstraps for time series. *Statistical Science 17 (1), 52–72.*

EFRON, B. (1979): Bootstrap methods: another look at the jackknife. *Annals of Statistics 7 (1), 1–26.*

HALL, P. and HART, J.D. (1990): Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association 85 (4), 1039–1049.*

HÄRDLE, W., HOROWITZ, J. and KREISS, J.-P. (2003): Bootstrap methods for time series. *International Statistical Review 71 (2), 435–459.*

LAHIRI, S.N. (2003): *Resampling Methods for Dependent Data.* Springer, New York.

POLITIS, D.N. and ROMANO, J.P. (1992): A circular block-resampling procedure for stationary data. In: R. LePage and L. Billard (Eds.): *Exploring the Limit of Bootstrap.* Wiley, New York, 263–270.

Part XVIII

Robustness

The Stahel-Donoho Outlyingness in a Reproducing Kernel Hilbert Space

Michiel Debruyne

Department of mathematics and computer science, Universiteit Antwerpen, Middelheimlaan 1G, 2020 Antwerpen, Belgium, *michiel.debruyne@ua.ac.be*

Abstract. The original definition of the Stahel-Donoho outlyingness in a Euclidean space is extended to a Reproducing Kernel Hilbert Space. Two robust kernel algorithms are constructed: robust Kernel Principal Component Analysis and robust Support Vector Machine classification. Both methods yield better results than their classical counterparts for data sets containing outliers. Possible outliers can be detected through appropriate diagnostic tools using the results of the robust algorithms.

Keywords: outlyingness, kernel methods, robustness, principal component analysis, support vector machine

1 Introduction

For a sample of multivariate data vectors $X = \{x_1, \ldots, x_n,\ x_i \in \mathbb{R}^d\}$ the Stahel-Donoho outlyingness $r(x_j)$ of observation x_j is defined by (Donoho, 1981; Stahel, 1982)

$$r(x_j) = \max_{a \in P} \left| \frac{a^t x_j - m(a^t X)}{s(a^t X)} \right| \tag{1}$$

with m a robust univariate estimator of location and s a univariate estimator of spread. The set P is a set of directions in $\mathbb{R}^d$. The Stahel-Donoho outlyingness measures the distance between an observation and the bulk of the data. As such it is a powerful tool to detect outlying observations, especially when the majority of data points is elliptically distributed. By downweighting or trimming observations with a large outlyingness, robust algorithms can be constructed that resist a fraction of outliers in the data. Typical applications include covariance estimation (Maronna and Yohai, 1995) and PCA (Hubert et al., 2005).

Section 2 of this paper shows that this concept of outlyingness can be extended from a simple Euclidean space $\mathbb{R}^d$ to a Reproducing Kernel Hilbert Space (RKHS). The resulting outlyingness measure can be used in many kernel algorithms. Illustrations are provided for Kernel Principal Component Analysis (KPCA, Scholköpf et al., 1998) in Section 3 and Support Vector Machine (SVM) classification (Vapnik, 1998) in Section 4.

2 Stahel-Donoho outlyingness in a RKHS

Let $\{x_1, \ldots, x_n\} \in \mathcal{X}$ be n elements in a space $\mathcal{X}$. Let K be an appropriate kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with corresponding RKHS $\mathcal{H}$ and feature map $\Phi : \mathcal{X} \to \mathcal{H}$ such that the inner product $\langle ., . \rangle$ between feature vectors in $\mathcal{H}$ can be computed by K:

$$\langle \Phi(x_i), \Phi(x_j) \rangle = K(x_i, x_j). \tag{2}$$

In practice the kernel K is chosen beforehand implicitly determining $\mathcal{H}$ and Φ. If $\mathcal{X}$ equals $\mathbb{R}^d$ and $K(u, v) = u^t v$ for instance, it is clear from (2) that Φ is the identity function and $\mathcal{H}$ is $\mathbb{R}^d$ itself. For a Radial Basis Function (RBF) kernel $K(u, v) = e^{-\|u-v\|^2/(2\sigma^2)}$ a unique infinite dimensional RKHS $\mathcal{H}$ and feature map Φ exist, but are not explicitly known. However, any algorithm only using inner products between feature vectors in $\mathcal{H}$ can still be computed through the kernel (2). An overview of kernel methodology exploiting this kernel trick can be found in the books of Scholkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004).

Denote Ω the kernel matrix containing $K(x_i, x_j)$ as entry i, j. Let a be the direction in $\mathcal{H}$ through 2 feature vectors $\Phi(x_i)$ and $\Phi(x_j)$. The projection of a feature vector $\Phi(x_l)$ onto this direction a is then

$$\langle a, \Phi(x_l) \rangle = \left\langle \frac{\Phi(x_i) - \Phi(x_j)}{\|\Phi(x_i) - \Phi(x_j)\|}, \Phi(x_l) \right\rangle.$$

Since the squared norm of an element equals the inner product of the element with itself we have that

$$\begin{aligned}
\|\Phi(x_i) - \Phi(x_j)\| &= \sqrt{\langle \Phi(x_i) - \Phi(x_j), \Phi(x_i) - \Phi(x_j) \rangle} \\
&= \sqrt{\langle \Phi(x_i), \Phi(x_i) \rangle - 2\langle \Phi(x_i), \Phi(x_j) \rangle + \langle \Phi(x_j), \Phi(x_j) \rangle} \\
&= \sqrt{K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)} \\
&= \sqrt{(\gamma^{i,j})^t \Omega \gamma^{i,j}}.
\end{aligned}$$

The vector $\gamma^{i,j} \in \mathbb{R}^n$ denotes the vector with entry i equal to 1, entry j equal to -1 and all other entries equal to 0. Then

$$\langle a, \Phi(x_l) \rangle = \left\langle \frac{\Phi(x_i) - \Phi(x_j)}{\sqrt{(\gamma^{i,j})^t \Omega \gamma^{i,j}}}, \Phi(x_l) \right\rangle = \left(\frac{\Omega \gamma^{i,j}}{\sqrt{(\gamma^{i,j})^t \Omega \gamma^{i,j}}} \right)_l.$$

Denote $v^{i,j}_{\text{proj}}$ the vector containing the projections of all feature vectors onto the direction a through feature vectors $\Phi(x_i)$ and $\Phi(x_j)$:

$$v^{i,j}_{\text{proj}} = (\langle a, \Phi(x_1) \rangle, \ldots, \langle a, \Phi(x_n) \rangle)^t = \frac{\Omega \gamma^{i,j}}{\sqrt{(\gamma^{i,j})^t \Omega \gamma^{i,j}}}.$$

From these projections the Stahel-Donoho outlyingness of a feature vector $\Phi(x_l)$ can be calculated as follows:

$$r(\Phi(x_l)) = \max_{(i,j)\in P} \left| \frac{\left(v^{i,j}_{\text{proj}}\right)_l - m(v^{i,j}_{\text{proj}})}{s(v^{i,j}_{\text{proj}})} \right|. \tag{3}$$

Again m and s are univariate robust estimators of location and scale. We take

$$m(v^{i,j}_{\text{proj}}) = \text{median}(v^{i,j}_{\text{proj}})$$
$$s(v^{i,j}_{\text{proj}}) = \text{mad}(v^{i,j}_{\text{proj}}) = \text{median}\left|v^{i,j}_{\text{proj}} - \text{median}(v^{i,j}_{\text{proj}})\right|$$

but other choices are of course possible as well, e.g. τ-estimators (Maronna and Zamar, 2002). The set P in (3) can be the entire set of all pairs of indices $\{1,\ldots,n\} \times \{1,\ldots,n\}$. Then every direction through any 2 feature vectors is considered. If n is too large P can be taken as a random subset of $\{1,\ldots,n\} \times \{1,\ldots,n\}$ of fixed size p.

3 Robust Kernel PCA

3.1 Algorithm

Classical linear PCA projects the data onto the eigenvectors of the classical sample covariance matrix. An extension of classical PCA to the kernel framework is proposed by Schölkopf et al. (1998) using the eigenvectors and eigenvalues of the kernelmatrix Ω. For the linear case it is well known that this procedure is very sensitive to outliers. Many robust alternatives have been proposed to adjust for this phenomenon (Locantore et al., 1999; Croux and Ruiz-Gazen, 2005; Hubert et al., 2005; Maronna, 2005). For other kernels outliers can be very influential as well, especially if the kernel function is unbounded, i.e. $\sup_z K(z,z) = \infty$ (Debruyne et al., 2008). Using the Stahel-Donoho outlyingness from Section 2 a robust kernel PCA algorithm can be constructed as follows.

(i) Fix $0.5 \leq \alpha \leq 1$.
(ii) Compute the Stahel-Donoho outlyingness for every observation as in (3). Retain the $h = \lfloor \alpha n \rfloor$ observations with smallest outlyingness ($\lfloor p \rfloor$ denotes the largest integer smaller than $p \in \mathbb{R}$). Denote I_h the index set of these h observations.
(iii) Perform Kernel PCA for this trimmed set of h observations to obtain the kth score function $f_k(x)$:

$$f_k(x) = \sum_{i\in I_h} \frac{\alpha_i^{(k)}}{\sqrt{\lambda_k}} \left(K(x_i,x) - \frac{1}{h}\sum_{l\in I_h} K(x_l,x) \right) \tag{4}$$

with $\alpha^{(k)}$ the unit norm eigenvector belonging to the kth largest eigenvalue λ_k of the trimmed centered $h \times h$ kernel matrix Ω^h with entry i, j equal to

$$\Omega^h_{i,j} := \Omega_{i,j} - \frac{2}{h} \sum_{k \in I_h} \Omega_{k,j} + \frac{1}{h^2} \sum_{k \in I_h} \sum_{l \in I_h} \Omega_{k,l}. \tag{5}$$

These formulas are obtained simply by applying the original formulas from Schölkopf et al. (1998) to the reduced set $\{x_i,\ i \in I_h\}$ instead of the full set of observations.

3.2 Example

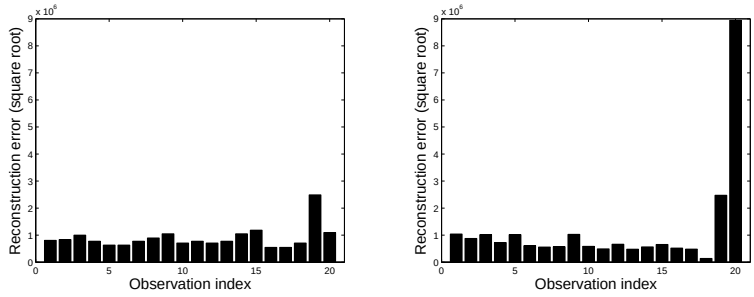
Consider a situation where the inputs are textstrings instead of numerical vectors. To analyze such data the all-subsequence kernel (Shawe-Taylor and Cristianini, 2004) can be used. Then the strings are represented by feature vectors of which each component represents a possible substring. The value of a component counts the number of times the corresponding substring occurs. However the number of possible substrings (which equals the dimension of the feature space) increases exponentially as the length of the string increases. For large strings an explicit computation of the corresponding feature vectors is thus impossible. However, with a kernel method only the inner products between any two feature vectors are needed. For the all-subsequence kernel the kernel matrix containing these inner products can be computed with fast recursive algorithms (Shawe-Taylor and Cristianini, 2004). Since our robust KPCA algorithm does not require explicit feature vectors either, but only the kernel matrix, applying the algorithm from 3.1 is straightforward.

As an example take the first 20 DNA sequences in the 'Splice-junction gene sequences' data set from the UCI data repository publicly available at http://www.ics.uci.edu/∼mlearn/MLRepository.html. This gives us 20 observations, all strings of size 60 composed out of 4 characters (A,C,G,T). As an example we contaminate observation 20 by replacing the first 25 characters by A. Figure 1 shows the square root of the reconstruction errors (=the orthogonal distances between a feature vector and its projection onto the PCA subspace in $\mathcal{H}$) of the 20 observations for (a) Classical KPCA and (b) Robust KPCA with $\alpha = 0.25$, both retaining 1 principal component. The outlier 20 clearly attracts the Classical KPCA fit: its reconstruction error is small. With robust KPCA the reconstruction error of the outlier is large, but the other observations fit better: 16 observations have a smaller error with robust KPCA than with ordinary KPCA.

4 Robust SVM classification

4.1 Algorithm

Let $(x_1, \ldots, x_n),\ x_i \in \mathcal{X}$ be the set of inputs with corresponding group labels $(y_1, \ldots, y_n),\ y_i \in \{-1, 1\}$. Let n_- be the number of inputs with label -1



(*a*) (*b*)

Fig. 1. String example: square root of the reconstruction error versus observation index for (*a*) Classical Kernel PCA, (*b*) Robust Kernel PCA, retaining 1 principal component. (Remark: the reason for taking the square root is better visibility.)

and n_+ be the number of inputs with label $+1$. The Support Vector Machine (SVM, Vapnik, 1998) is a classification technique to predict the label of a new observation $x \in \mathcal{X}$. In this case as well outliers can have a large influence on the classifying function (Christmann and Steinwart, 2004). We propose a trimmed Stahel-Donoho SVM classification algorithm (SD-SVM) as follows.

(i) Fix $0.5 \leq \alpha \leq 1$. Denote $h_- = \lfloor \alpha n_- \rfloor$ and $h_+ = \lfloor \alpha n_+ \rfloor$.
(ii) Consider only the inputs with group label -1. Compute the Stahel-Donoho outlyingnesses for every sample in this set using (3). Retain the h_- observations with smallest outlyingness. Denote the corresponding index set of size h_- as I_{h_-}. Analoguously obtain the set I_{h_+} containing the indices of the h_+ samples with group label $+1$ with smallest outlyingness.
(iii) Train a standard SVM on the inputs in $I_h = I_{h_-} \cup I_{h_+}$ by solving

$$\max_{\alpha} \left\{ \sum_{i \in I_h} \alpha_i - \sum_{i \in I_h} \sum_{j \in I_h} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right\} \text{s.t.} 0 \leq \alpha_i \leq C, \ \sum_{i \in I_h} \alpha_i y_i = 0.$$

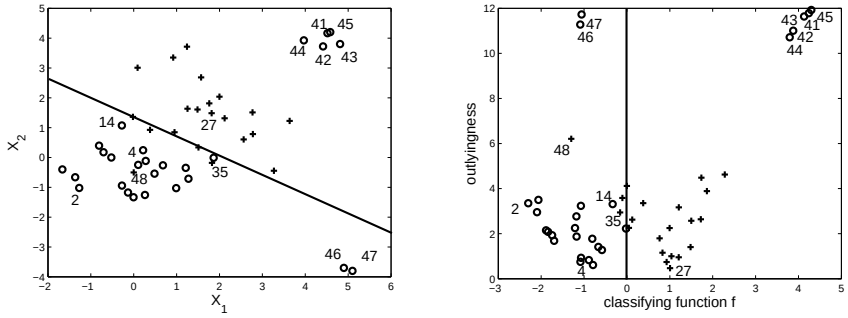
The classifying function is given by

$$f(x) = \sum_{x_i \in I_h} \alpha_i K(x_i, x) + b. \tag{6}$$

The predicted group label for a new observation x equals $\text{sign}(f(x))$.

4.2 Diagnostic plot

We propose to make a scatterplot of the pairs $(r(\Phi(x_j)), f(x_j))$ where $r(\Phi(x_j))$ is the Stahel-Donoho outlyingness of sample j computed in the trimming step of the algorithm and $f(x_j)$ can be calculated from (6). We will plot the inputs with group labels $+1$ as plusses and those with group labels -1 as circles. Finally we add a solid vertical line at horizontal coordinate 0.



(*a*) (*b*)

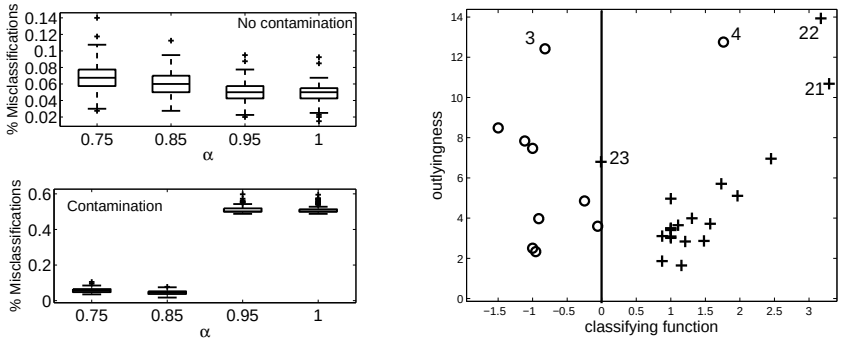
Fig. 2. (*a*) 2-dimensional classification problem. The solid line is the SD-SVM classifying line. (*b*) Diagnostic plot.

4.3 Toy example

Consider a simple example in 2 dimensions: 2 groups of 20 observations are generated from bivariate standard normal distributions with means $(0, 0)$ for the negative and $(1.5, 1.5)$ for the positive group. Eight points are added representing several types of outliers. A two-dimensional view of the data is given in Figure 2(*a*). The solid line represents the SD-SVM classification boundary with a linear kernel. Despite the 8 outliers (labeled 41-48) in the data, SD-SVM still manages to separate both groups quite nicely. Figure 2(*b*) shows the corresponding diagnostic plot as defined in Section 4.2. On the vertical axis one reads the Stahel-Donoho outlyingness. Points close to the center of their group have a small outlyingness, e.g. 4 and 27. Observations further away from the center have a larger outlyingness, e.g. 2 and 14. On the horizontal axis the value of the classifying function f as in (6) can be read. The sign of this function determines the predicted group labels. A vertical line at $f = 0$ divides the plot in two parts: every point left of line is classified into the negative group by SD-SVM; every point on the right into the positive group. Based on the diagnostic plot the outliers can be visualised and characterised. Observations $46 - 47$ are outlying with respect to the other data points in the negative group, which is clearly indicated by their large outlyingness. However, both samples still follow the classification rule. Indeed, both are lying on the left side in Figure 2(*b*). Samples $41 - 45$ on the other hand are outlying with respect to the other observations in the negative group as well as with respect to the classification line.

4.4 Simulated example

The following setup is considered: 21 data vectors are generated with 1000 i.i.d. standard Gaussian components and label -1. The positive group contains 23 observations with 1000 i.i.d. Gaussian components with unit variance



(*a*) (*b*)

Fig. 3. Experimental results. (*a*) Simulated example: misclassification percentages for different values of α. Upper: uncontaminated case. Lower: contaminated case. (*b*) Mice data: diagnostic plot.

and mean 0.2. SD-SVM is applied for different values of α. The tuning parameter C is chosen by a grid search with Leave-One-Out cross validation. The performance of the classification rule is assessed by the percentage of misclassified points in the test data, which contains 2000 newly generated test points (1000 in each group).

This is repeated 100 times. Boxplots of the resulting 100 misclassification percentages are shown in Figure 3(*a*) (upper panel) for $\alpha = 0.75,\ 0.85,\ 0.95$ and 1. We observe that the performance is quite good for every α, but nevertheless increasing as α increases. This is to be expected, since all observations come from nice normal distributions making trimming unnecessary and thus classical SVM ($\alpha = 1$) performs best. Next the same experiment was repeated adding some outliers: 3 in the negative group with Gaussian components with mean 3 and 3 outliers in the positive group with Gaussian components with mean -3. In this case classical SVM completely breaks down (see Figure 3(*a*) lower panel): it does not perform better than random guessing. This occurs for $\alpha = 0.95$ as well, since the percentage of outliers is higher than 5%. If α is chosen small enough however, SD-SVM succeeds in obtaining good classification despite the presence of the outliers.

4.5 Mice data

Finally consider a real life example to illustrate the use of the diagnostic plot. The Mice data (Wen et al., 1998) consists of NMR spectra (2050 wavelengths) measured for 30 mice with a tumor implanted, 10 of which received no treatment and 20 did. Hubert and Engelen (2004) found that the data set contains some outliers. When SD-SVM is applied with $\alpha = 0.75$ and a linear kernel the diagnostic plot in Figure 3(*b*) shows up. It confirms indeed that

some mice are deviating from the others. Mice $21 - 22$ and 3 lie far away from their group members (large outlyingness), but they are not outlying with respect to the classification rule. Mice 23 and especially 4 are outlying with respect to the classifier as well.

References

CHRISTMANN, A. and STEINWART, I. (2004): On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research 5, 1007-1034.*

CROUX, C. and RUIZ-GAZEN, A. (2005): High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis 95, 206-226.*

DEBRUYNE, M., HUBERT, M. and VAN HOREBEEK, J. (2008): Detecting influential observations in kernel PCA, submitted.

DONOHO, D.L. (1982): *Breakdown properties of multivariate location estimators.* Qualifying paper, Harvard University, Boston.

HUBERT, M. and ENGELEN, S. (2004): Robust PCA and classification in biosciences. *Bioinformatics, 20, 1728-1736.*

HUBERT, M., ROUSSEEUW, P.J. and VANDEN BRANDEN, K. (2005): A new approach to robust principal components analysis. *Technometrics 47, 64-79.*

LOCANTORE, N., MARRON, J.S., SIMPSON, D.G., TRIPOLI, N., ZHANG, J.T. and COHEN, K.L. (1999): Robust principal component analysis for functional data. *Test 8, 1-73.*

MARONNA, R.A. (2005): Principal components and orthogonal regression based on robust scales. *Technometrics 47, 264-273.*

MARONNA, R.A. and YOHAI, V.J. (1995): The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association 90, 330-341.*

MARONNA, R.A. and ZAMAR, R.H. (2002): Robust estimates of location and dispersion of high-dimensional datasets. *Technometrics 44, 307-317.*

SCHÖLKOPF, B. and SMOLA, A. (2002): *Learning with Kernels.* MIT Press, Cambridge.

SCHÖLKOPF, B., SMOLA, A. and MÜLLER, K.-R. (1998): Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation 10, 1299-1319.*

SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004): *Kernel methods for pattern analysis.* Cambridge university press, Cambridge.

STAHEL, W.A. (1981): *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen.* PhD thesis, ETH Zürich.

VAPNIK, V. (1998): *Statistical Learning Theory.* John Wiley & Sons, New-York.

WEN, X., FUHRMAN, S., MICHAELS, G.S., CARR, D.B., SMITH, S., BARKER, J.L., and SOMOGYI R. (1998): Large-scale temporal gene expression mapping of the central nervous system development. *Proc. Natl. Acad. Sci. USA, 95, 334-339.*

Estimating the Parameters of a Bivariate Extreme Value Gumbel Distribution

Alessandra Durio and Ennio Davide Isaia

Dep. of Statistics & Applied Mathematics "Diego de Castro", University of Torino
Piazza Arbarello, 8 — 10122 Torino, Italy, *durio(isaia)@econ.unito.it*

Abstract. In this paper we compare the estimates based on moments and maximum likelihood criteria with the ones obtained according to the minimum integrated square error criterion. Considering the bivariate extreme value Gumbel distribution, we measure the difference among the parameters estimates resorting to a similarity index between densities, for which a Monte Carlo significance test is introduced. Theory is outlined and main results of a simulation study are provided and commented.

Keywords: bivariate extreme value Gumbel distributions, minimum integrated square error, Monte Carlo significance test, robust estimators

1 Introduction

The extreme value Gumbel distribution is one of the most widely applied statistical distribution for problems in engineering, including, among them, meteorology analysis, network engineering, finance engineering and risk-based engineering (Kotz and Nadarajah (2000)). Usually the estimates of the parameter(s) of a Gumbel distribution are achieved by the method of the moments or resorting to the the maximum likelihood criterion.

In this paper we suggest to compare the estimates obtained from classical criteria with the ones based on the minimum integrated square error criterion (or L_2 metric). The likely discrepancy between the estimated densities is measured applying the concept of similarity between densities following from the Cauchy-Schwarz inequality and a Monte Carlo significance test of the statistical hypothesis of similarity is introduced to verify the similarity between the estimates. Whenever the hypothesis of similarity is rejected, we suggest to investigate more carefully the data structure in order to check against the presence of anomalies in the data. In this sense the L_2 criterion can be viewed as a practical diagnostic tool in building useful models. Considering the bivariate extreme value Gumbel distribution, theory is outlined and main results of a simulation study, featuring several experimental scenarios with different percentages of data contamination are provided to illustrate and corroborate the approach we propose.

2 The Gumbel's distributions and some estimators

Let (X_1, X_2) be a bivariate r.v. with marginal distributions corresponding to the univariate extreme value random variables with distribution functions $F(x_i) = \exp(-e^{-x_i})$. For such random variable we have $\mathbb{E}[X_i] = 0.577$ and $Var[X_i] = \pi^2/6$ and this for $i = 1, 2$ and $-\infty < x_i < \infty$. For our purposes we consider the type B bivariate extreme value Gumbel distribution $\mathcal{G}_1(m)$ which, for $m \geq 1$, has joint probability density function (Gumbel and Mustafi, (1967))

$$f(x_1, x_2|m) = e^{-m(x_1+x_2)} \left(e^{-m x_1} + e^{-m x_2}\right)^{\frac{1}{m}-2} \left(m - 1 + (e^{-m x_1} + e^{-m x_2})^{\frac{1}{m}}\right) e^{-(e^{-m x_1}+e^{-m x_2})^{\frac{1}{m}}}. \tag{1}$$

Usually the parametric estimate of m can be performed using a few approaches, such as the maximum likelihood (ML) and the method of the moments (MM). According to the maximum likelihood approach, the estimate $\hat{m}_{ML}$ is obtained maximizing, with respect to m, the log-likelihood function $L(m) = \sum_{i=1}^{n} \log f(X_{1i}, X_{2i}|m)$. Resorting to the method of the moments the estimate of m is given by $\hat{m}_{MM} = (1 - R)^{-1/2}$, where R is the sample correlation coefficient of (X_1, X_2).

Since many empirical situations, concerning the study of extreme values, are well described by linear transformations of the $\mathcal{G}_1(m)$ r.v. (Tawn, (1988)), we introduce, for $i = 1, 2$, the transformations $Y_i = b_i X_i + a_i$, with location parameters $a_i \in]-\infty, \infty[$ and scale parameters $b_i > 0$. The r.v. (Y_1, Y_2) has a type B bivariate extreme value Gumbel distribution $\mathcal{G}_2(m, a_1, b_1, a_2, b_2)$ with probability density function

$$f(y_1, y_2|m, a_i, b_i) = \frac{e^{-m(\frac{y_1-a_1}{b_1}+\frac{y_2-a_2}{b_2})}}{b_1 b_2} \left(e^{-\frac{m(y_1-a_1)}{b_1}} + e^{-\frac{m(y_2-a_2)}{b_2}}\right)^{\frac{1}{m}-2} \left(m - 1 + (e^{-\frac{m(y_1-a_1)}{b_1}} + e^{-\frac{m(y_2-a_2)}{b_2}})^{\frac{1}{m}}\right) e^{-\left(e^{-\frac{m(y_1-a_1)}{b_1}}+e^{-\frac{m(y_2-a_2)}{b_2}}\right)^{\frac{1}{m}}}. \tag{2}$$

In this case too, the estimate of the five parameters can be obtained either maximizing the log-likelihood function $L(m, a_1, b_1, a_2, b_2)$ or turning to the method of the moments. In this latter case, $\hat{m}_{MM} = (1 - R)^{-1/2}$, where R is the sample correlation coefficient of (Y_1, Y_2), $\hat{b}_{i_{MM}} = \frac{\sqrt{6}}{\pi} S_i$, where S_i are the standard deviations of Y_i, and $\hat{a}_{i_{MM}} = \bar{Y}_i - 0.577\, S_i$

In addition to ML and MM estimators, we consider the estimator based on the minimum integrated square error (Terrell (1990), Scott (2001)). Our choice can be motivated by the fact that in the family of the estimators based on the minimum density power divergence criterion, ML and L_2 estimators

are, respectively, the less and the more robust to outliers, even if the latter is less efficient (Basu et al., (1998)).

Given the r.v. $\mathbf{X}$ of dimension $d \geq 1$ and unknown density $f(\mathbf{X}|\boldsymbol{\theta}_0)$ for which we introduce the model $f(\mathbf{X}|\boldsymbol{\theta})$, the estimate of the vector $\boldsymbol{\theta}_0$ according to L_2 criterion is given by

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{L_2} &= \arg\min_{\boldsymbol{\theta}} \int_{\mathbb{R}^d} \left(f(\mathbf{x}|\boldsymbol{\theta}) - f(\mathbf{x}|\boldsymbol{\theta}_0)\right)^2 d\mathbf{x} = \\
&= \arg\min_{\boldsymbol{\theta}} \left[\int_{\mathbb{R}^d} f^2(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} - 2 \int_{\mathbb{R}^d} f(\mathbf{x}|\boldsymbol{\theta})\, f(\mathbf{x}|\boldsymbol{\theta}_0)\, d\mathbf{x}\right] = \\
&= \arg\min_{\boldsymbol{\theta}} \left[\int_{\mathbb{R}^d} f^2(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} - 2\, \mathbb{E}\left[f(\mathbf{x}|\boldsymbol{\theta}_0)\right]\right] = \\
&\approx \arg\min_{\boldsymbol{\theta}} \left[\int_{\mathbb{R}^d} f^2(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} - \frac{2}{n} \sum_{i=1}^{n} f(\mathbf{X}_i|\boldsymbol{\theta})\right].
\end{aligned} \tag{3}$$

With respect to the Gumbel distributions $\mathcal{G}_1(m)$ and $\mathcal{G}_2(m, a_1, b_1, a_2, b_2)$, after analytical evaluation of the integral that appears in equation (3), we have the following expressions for the L_2 estimates

$$\hat{m}_{L_2} = \arg\min_{m} \left[\frac{2\,m^2 + 1}{48\,m} - \frac{2}{n} \sum_{i=1}^{n} f(X_{1_i}, X_{2_i}|m)\right] \tag{4}$$

and, for $i = 1, 2$

$$(\hat{m}, \hat{a}_i, \hat{b}_i)_{L_2} = \arg\min_{m, a_i, b_i} \left[\frac{2\,m^2 + 1}{48\,m\,b_1\,b_2} - \frac{2}{n} \sum_{i=1}^{n} f(Y_{1_i}, Y_{2_i}|m, a_i, b_i)\right]. \tag{5}$$

Clearly, equations (4) and (5) are feasible computationally closed-form expressions so that L_2 estimates can be performed by any standard non linear optimization code, for instance the `nlm` routine of the `R` software.

3 The similarity index and the Monte Carlo test

If we apply the estimators introduced above, namely MLE, MME and L_2E, to the same sample they are likely to yield different values for the parameters. From a practical point of view, it could be useful to have an idea about the magnitude of the differences between the densities estimated by any pair of the considered estimating criteria. To this end, given the r.v. $\mathbf{X}$ of dimension $d \geq 1$, let T_0 and T_1 be two estimators and let $\hat{\boldsymbol{\theta}}_{T_0}$ and $\hat{\boldsymbol{\theta}}_{T_1}$ be the corresponding vectors of the estimated parameters of the density $f(\mathbf{x}|\boldsymbol{\theta}_0)$. In order to measure the discrepancy between the estimated densities $f(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_0})$ and $f(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_1})$, we introduce the similarity index defined as

$$sim(T_0, T_1) \stackrel{def}{=} \frac{\int_{\mathbb{R}^d} f(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_0})\, f(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_1})\, d\mathbf{x}}{\left(\int_{\mathbb{R}^d} f^2(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_0})\, d\mathbf{x} \int_{\mathbb{R}^d} f^2(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_1})\, d\mathbf{x}\right)^{\frac{1}{2}}}. \tag{6}$$

The index given by equation (6) is normalized since the concept of similarity between densities follows from the Cauchy-Schwarz inequality (Scott and Szewczyk (2001), Durio and Isaia (2007)).

If the vectors $\hat{\boldsymbol{\theta}}_{T_0}$ and $\hat{\boldsymbol{\theta}}_{T_1}$ are close to each other, then $sim(T_0, T_1)$ will be close to unity. On the other hand, if the estimated densities $f(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_0})$ and $f(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_1})$ are dissimilar, we are likely to observe a value of $sim(T_0, T_1)$ approaching zero. We propose to use the $sim(T_0, T_1)$ statistic of equation (6) to verify the following system of hypotheses

$$\begin{cases} H_0 : \boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_{T_0} \\ H_1 : \boldsymbol{\theta}_0 \neq \hat{\boldsymbol{\theta}}_{T_0}. \end{cases} \tag{7}$$

Since it is not reasonable to look for a closed-form expression for the $sim(T_0, T_1)$ distribution, to check the system of hypothesis given by (7) we resort to the Monte Carlo Significance Test (MCS test), originally suggested by Hope (1968). Denoting with $sim_{T_0T_1}$ the value of the $sim(T_0, T_1)$ statistic computed on the observed data, the MCS test consists in rejecting H_0 if $sim_{T_0T_1}$ is the $k\,\alpha$-th most extreme statistic relative to the corresponding quantities based on the random samples of the reference set, where the reference set consists in $k-1$ random samples, of size n each, generated under the null hypothesis, i.e. drawn at random from the model $f(\mathbf{x}|\hat{\boldsymbol{\theta}}_{T_0})$. In other words we generate $k-1$ random samples under H_0 and for each of them we compute $sim^*_{T_0T_1}$ and we shall reject the null hypothesis, at the α significance level, if and only if the value of the test statistic $sim_{T_0T_1}$ is greater than all the $k-1$ values of $sim^*_{T_0T_1}$. We remark that if we set $k\,\alpha = 1$ and fix $\alpha = 0.01$, we have $k - 1 = 99$, while fixing $\alpha = 0.05$ would yield $k - 1 = 19$.

The rejection of the hypothesis of similarity between the two estimated densities must be interpreted as a warning signal suggesting a more careful investigation of the data structure.

With the aim to compare the estimates based on MM and L_2 criteria with those obtained from the maximum likelihood, we set in equation (6) $T_0 = MLE$ and we perform the MCS test two times: the first one fixing $T_1 = MME$ for $sim(MLE, MME)$, the second one letting $T_1 = L_2E$, for $sim(MLE, L_2E)$.

4 Numerical example

In recent years the extreme value Gumbel distributions have been widely applied to the study of hydrological extreme events, such as storm or drought.

Yue (2001) resorts to a $\mathcal{G}_2(m, a_1, b_1, a_2, b_2)$ distribution to represent the joint distribution of correlated storm peaks and amounts of daily rainfall data from the Tokushima (Japan) meteorological station.

In this section we provide the results of a simulated example reproducing the data of storm peaks and amounts of daily rainfall from the Pino Torinese (Italy) meteorological station.

true		**(a)**			**(b)**		
values		MME	MLE	L_2E	MME	MLE	L_2E
$m = 1.8$	$\hat{m} =$	1.942	1.896	1.778	1.447	1.473	1.751
$a_1 = 47$	$\hat{a}_1 =$	45.583	45.345	45.956	42.332	42.336	46.117
$b_1 = 15$	$\hat{b}_1 =$	15.438	15.542	13.315	16.269	16.859	14.298
$a_2 = 70$	$\hat{a}_2 =$	68.536	68.598	71.612	72.406	70.445	71.836
$b_2 = 33$	$\hat{b}_2 =$	34.434	34.463	34.058	34.503	35.636	35.633

Table 1. Estimated parameters of the $\mathcal{G}_2$ Gumbel distribution.

We generate a sample of $n = 120$ points randomly drawn from the $\mathcal{G}_2(m = 1.8, a_1 = 47, b_1 = 15, a_2 = 70, b_2 = 33)$ distribution and we estimate the five parameters according to ML, MM and L_2 criteria.

If we look at the results summarized in Table 1, panel (a), we can argue that the estimates do not differ substantially among them. This idea is corroborated by the results of the two MCS tests, in fact, at the level $\alpha = 0.01$, we observe $sim_{MLE,MME} = 0.9997 > min(sim^*_{MLE,MME}) = 0.9116$ and $sim_{MLE,L2E} = 0.9892 > min(sim^*_{MLE,L2E}) = 0.9734$ and hence in both cases we accept the null hypothesis of system (7).

Table 1, panel (b) shows the estimated values of the parameters when we perturb our data set with 6 (i.e. 5%) points in the 2nd quadrant of a system with origin at the center of data mass and laying on the contour line which gives the region within the r.v. (Y_1, Y_2) falls with probability 0.9997.

In this situation both ML and MM criteria tend to underestimate the parameter m and hence the correlation coefficient. The L_2 estimates do not differ much from those obtained on the non-contaminated sample and this is due to the inherent properties of robustness of the L_2 criterion (Durio and Isaia, (2004)). Since the MCS test for $sim(MLE, MME)$ leads us to the non rejection of the null hypothesis of system (7), as $sim_{MLE,MME} = 0.9982 > min(sim^*_{MLE,MME}) = 0.9035$, we could use the ML results as good estimates for the parameters of the $\mathcal{G}_2$ distribution; doing so, we do not perceive that data have been perturbed. If we turn now to $sim(MLE, L_2E)$, the MCS test prompts us to reject the null hypothesis for $sim(MLE, L_2E)$, as $sim_{MLE,L2E} = 0.9585 < min(sim^*_{MLE,L2E}) = 0.9758$. The robustness of L_2 estimator is helpful, in this situation, since it makes us aware that either the model is not correct or we are in presence of contaminated data.

5 Main results of a simulation

At this point it is worthwhile to provide some results arising from a simulation study we carried out to check the behaviour of the three estimators

		Data infection rates					
n = 100		*4%*		*6%*		*12%*	
	MME	1.827	(0.193)	1.746	(0.177)	1.579	(0.147)
$c = 2.594$	MLE	1.799	(0.124)	1.714	(0.109)	1.519	(0.078)
	L_2E	1.977	(0.267)	1.934	(0.265)	1.798	(0.264)
	MME	1.708	(0.169)	1.605	(0.150)	1.414	(0.118)
$c = 3.406$	MLE	1.728	(0.114)	1.622	(0.036)	1.393	(0.064)
	L_2E	1.998	(0.261)	1.968	(0.257)	1.877	(0.245)
	MME	1.571	(0.143)	1.455	(0.124)	1.262	(0.092)
$c = 4.402$	MLE	1.652	(0.084)	1.526	(0.084)	1.280	(0.052)
	L_2E	2.004	(0.253)	1.977	(0.253)	1.989	(0.236)
n = 500							
	MME	1.783	(0.095)	1.705	(0.087)	1.544	(0.072)
$c = 2.594$	MLE	1.801	(0.063)	1.717	(0.056)	1.521	(0.042)
	L_2E	1.944	(0.117)	1.902	(0.117)	1.769	(0.116)
	MME	1.669	(0.083)	1.570	(0.073)	1.385	(0.057)
$c = 3.406$	MLE	1.730	(0.058)	1.623	(0.070)	1.392	(0.035)
	L_2E	1.965	(0.115)	1.936	(0.113)	1.850	(0.108)
	MME	1.537	(0.070)	1.425	(0.059)	1.239	(0.044)
$c = 4.402$	MLE	1.653	(0.053)	1.526	(0.044)	1.277	(0.030)
	L_2E	1.971	(0.113)	1.945	(0.111)	1.871	(0.104)

Table 2. Scenario C.1.b: means and standard deviations (in brackets) of the 1000 estimates of the parameter m.

introduced above. To this end, we set up some different experimental configurations according to specified data generating model corresponding to equation (1) and for each scenario we repeat 1000 times the estimates of the parameter according to ML, MM and L_2 criteria and we then compare the results given by each estimator in terms of their mean and standard deviation. Furthermore, we set in equation (6) $T_0 = MLE$ and we perform the MCS test twice: the first one letting $T_1 = MME$ for $sim(MLE, MME)$, the second one fixing $T_1 = L_2E$, for $sim(MLE, L_2E)$ and we record the number of times out of the 1000 runs that we reject the null hypothesis of system (7). In the following we provide and comment two experimental configurations.

C.1.a: we consider 1000 simulated data sets of $n = 100(500)$ points randomly drawn from a $\mathcal{G}_1(m)$ with $m = 2$ (i.e. $\rho = 0.75$).
The behaviour of the three estimators is substantially the same showing in average a good performance. In fact for the 1000 runs we obtained in mean the

		Data infection rates		
n = 100		*4%*	*6%*	*12%*
$c = 2.594$	$sim(MLE, MME)$	3	4	4
	$sim(MLE, L_2E)$	17	26	34
$c = 3.406$	$sim(MLE, MME)$	7	7	4
	$sim(MLE, L_2E)$	25	40	74
$c = 4.402$	$sim(MLE, MME)$	9	9	3
	$sim(MLE, L_2E)$	45	63	93
n = 500				
$c = 2.594$	$sim(MLE, MME)$	10	10	7
	$sim(MLE, L_2E)$	41	61	86
$c = 3.406$	$sim(MLE, MME)$	14	16	4
	$sim(MLE, L_2E)$	78	94	99
$c = 4.402$	$sim(MLE, MME)$	23	25	3
	$sim(MLE, L_2E)$	98	100	100

Table 3. Scenario C.1.b: percentages of times that the MCS test rejects the null hypothesis for $sim(MLE, MME)$ and $sim(MLE, L_2E)$.

estimates $\hat{m}_{MLE} = 2.008$, $\hat{m}_{MME} = 2.063$ and $\hat{m}_{L_2E} = 2.061$ when $n = 100$ and $\hat{m}_{MLE} = 2.011$, $\hat{m}_{MME} = 2.007$ and $\hat{m}_{L_2E} = 2.024$ when $n = 500$. For what concerns the standard deviations of the estimates, they are obviously greater when $n = 100$, while MME and L_2E show systematically a grater variance then MLE. The standard deviations we obtained for MLE, MME and L_2E are respectively, for $n = 100$, 0.163, 0.249 and 0.272, while for $n = 500$ they are 0.082, 0.123 and 0.119.
Turning to the MCS test, we rejected, when it is actually true, the null hypothesis of system (7) once for $sim(MLE, MME)$ when $n = 100$ and never when $n = 500$ while twice for $sim(MLE, L_2E)$ when $n = 100$ and only once when $n = 500$. These results let us affirm that clearly ML is the best estimator in this situation, but overall that the similarity index and the MCS test we propose have a good performance whenever the underlying model is correct.

C.1.b: we consider the same situation of case C.1.a, but we infect each sample with a proportion of 4% (6% and 12%) data points laying on the line of equation $x_2 = c + x_1$ and equally spaced in the sector delimited by the equations $x_2 = x_1^*$ and $x_2 = (x_1^* + x_2^*) - x_1$, where (x_1^*, x_2^*) are the coordinates of the mode of $\mathcal{G}_1(m)$, i.e. $x_1^* = x_2^* = (1 + m^{-1}) \log(2) - \log(\sqrt{(m-1)^2 + 4} - m + 3)$. We set $c = 2.594(3.406, 4.402)$ so that on the domain $\{(x_1, x_2) : |x_1 - x_2| - c, x_2 \in \mathbb{R}\}$ the r.v. (X_1, X_2) falls with proba-

bility $p = 0.9889(0.9978, 0.9997)$.
In all these situations again (see Table 2) the MLE shows a lowest standard deviation of its estimates, while L_2E has the widest one.
We observe that adding points as we did tends to confound a $\mathcal{G}_1(m = 2)$ with a $\mathcal{G}_1(m < 2)$. The means of the ML and MM estimates of Table 2 tend to decrease as data infection rate increases as well as c increases, while the L_2 estimates remain stable in mean around $\hat{m} = 2$. In all the sub-configurations when the data infection rate is 12%, the difference in mean between ML and MM estimates is smaller than the one we observe when the data infection rate is lower. This is the reason why we observe a low percentage of rejection of the null hypothesis for $sim(MLE, MME)$ in the last column of Table 3.
If we consider the results displayed in Table 3 as empirical powers of the MCS tests for $sim(MLE, MME)$ and $sim(MLE, L_2E)$, we obviously remark that the power of the test increases as the parameter n increases. For each sub-configuration the percentage of times we correctly reject the similarity between MLE and L_2E is greater than the corresponding percentage between MLE and MME. This is due to the robustness of L_2E with respect to MLE and to MME, as we pointed out in commenting the results of Table 2. The MCS test for $sim(MLE, MME)$ is not useful since its empirical power is very low, hence we suggest to resort to the estimators based on ML and L_2. We finally observe that the empirical power of the test for $sim(MLE, L_2E)$ increases as the data infection rate increases as well as the infecting points go far from the center of the data mass, i.e. c increases.

In conclusion, when estimating the parameter(s) of a Gumbel distribution we resort to the estimators based on the maximum likelihood and on the minimum integrated square error and we compare their results with the MCS test on the similarity index. Then, whenever a discrepancy between the two estimated densities occurs, we can perceive the presence of anomalies in the data.

References

BASU, A., HARRIS, I.R., HJORT, N.L. and JONES, M.C. (1998): Robust and Efficient Estimation by Minimizing a Density Power Divergence. *Biometrika, 85, 549-559.*

DURIO, A. and ISAIA, E.D. (2004): On robustness to outliers of parametric L_2 estimate criterion in the case of bivariate normal mixtures: a simulation study. In: M. Hubert G. Pison A.S. and Aelst S.V. (Eds.): *Theory and Applications of Recent Robust Methods*. Birkhäuser, 93–104.

DURIO, A. and ISAIA, E.D. (2007): A Quick Procedure for Model Selection in the Case of Mixture of Normal Densities. *Journal of Computational Statistics and Data Analysis, 12, 5635-5643.*

GUMBEL, E.J. and MUSTAFI, C.K. (1967): Analytical Properties of Bivariate Extremal Distributions. *Journal of the American Statistical Association, 62, 569-588.*

HOPE, A.C. (1968): A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society, B, 30, 582-598.*

KOTZ, S. and NADARAJAH, S. (2000): *Extreme Value Distributions. Theory and Applications*. Imperial College Press, London.

SCOTT, D.W. (2001): Parametric Statistical Modeling by Minimum Integrated Square Error. *Technometrics, 43, 274-285.*

SCOTT, D.W. and SZEWCZYK, W.F. (2001): From Kernel to Mixtures. *Technometrics, 43, 223-235.*

TAWN, J.A. (1988): Bivariate Extreme Value Theory: Models and Estimation. *Biometrika, 75, 397-415.*

TERRELL, G.R. (1990): Linear Density Estimates. *Proceedings of the Statistical Computing Section, American Statistical Association, 297-302.*

YUE, S. (2001): The Gumbel Logistic Model for Representing a Multivariate Storm Event. *Advances in Water Resources 24, 179-185.*

Fast Bootstrap for Robust Hotelling Tests

Ella Roelant, Stefan Van Aelst, and Gert Willems

Department of Applied Mathematics and Computer Science
Ghent University - UGent, Krijgslaan 281, S9, 9000 Gent, Belgium
Ella.Roelant@UGent.be, Stefan.VanAelst@UGent.be and Gert.Willems@UGent.be

Abstract. In this paper we consider a robust version of the one-sample and two-sample Hotelling tests. We use S- or MM-estimators of location and scatter instead of the empirical means and covariance matrices. A fast and robust bootstrap procedure is used to mimic the distribution of the test statistic. Simulations show good performance and illustrate that the bootstrap outperforms the asymptotic variance approach.

Keywords: bootstrap test, S- and MM-estimators, significance level

1 Introduction

If we need to estimate the location and scatter of a multivariate data set which may contain outliers, then the sample mean and sample covariance matrix will no longer be satisfactory as they can be extremely sensitive to outliers. Many robust alternatives for location and scatter have been discussed in the literature. We will focus on S-estimators (Davies (1987), Lopuhaä (1989)) and MM-estimators (Tatsuoka and Tyler (2000)) which are efficient, equivariant, positive breakdown estimators.

Inference for robust estimators is often based on the asymptotic distribution of these estimators. However, asymptotic estimates may be inaccurate for small sample sizes. Furthermore, when outliers are present in the data, assumptions underlying the asymptotic formulas, such as the data being elliptically distributed, are often violated.

An alternative approach to perform inference based on robust estimators is given by the bootstrap (Efron (1979)) which generally does not require distributional assumptions. The bootstrap principle is to generate a large number of samples from the original data set and to recalculate the estimate in each resample. With these bootstrapped estimates, an approximation to the estimator's true distribution can be obtained. However, the standard bootstrap procedure is non-robust, as some bootstrap samples may contain a fraction of outliers that exceeds the breakdown point of the robust estimates. Moreover, the standard procedure is also computationally demanding, due to the high computation time of robust estimators. Both these problems are solved by the fast and robust bootstrap (FRB) procedure, which was first introduced by Salibian-Barrera and Zamar (2002) and further developed by

Van Aelst and Willems (2005) and Salibian-Barrera et al. (2006, 2008). Both S- and MM-estimators can be written as the solution of a system of smooth fixed-point equations. The FRB gains a considerable amount of computation time by using the fixed-point representation of the estimator to approximate the bootstrap estimate in each resample. Because a reweighted representation of the estimator is bootstrapped, the method is also more robust as outliers that get downweighted in the original sample, also get downweighted in each resample, regardless of the fraction of outliers in that resample.

Section 2 explains the FRB principle for S- and MM-estimators. Section 3 shows how this FRB procedure can be used for robust one-sample and two-sample Hotelling tests. In Section 4 we compare the FRB to the asymptotic variance approach with a simulation study. We conclude in Section 5 with an outlook on further research.

2 Fast and robust bootstrap for S- and MM-estimators

We explain the method in general terms. Suppose that an estimator of the parameter of interest Θ can be represented by a smooth fixed-point equation $\mathbf{g}(\widehat{\Theta}_n) = \widehat{\Theta}_n$, where $\mathbf{g}$ involves the original sample. Then, using the smoothness of $\mathbf{g}$, we can calculate a Taylor expansion about Θ:

$$\widehat{\Theta}_n = \mathbf{g}(\Theta) + \nabla\mathbf{g}(\Theta)(\widehat{\Theta}_n - \Theta) + R_n$$

where R_n is a remainder term and $\nabla\mathbf{g}(.)$ is the matrix of partial derivatives. Supposing that the remainder term is small, this equation can be rewritten as

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) \approx [I - \nabla\mathbf{g}(\Theta)]^{-1}\sqrt{n}(\mathbf{g}(\Theta) - \Theta).$$

Taking bootstrap equivalents at both sides and estimating the matrix $[I - \nabla\mathbf{g}(\Theta)]^{-1}$ by $[I - \nabla\mathbf{g}(\widehat{\Theta}_n)]^{-1}$ yields

$$\sqrt{n}(\widehat{\Theta}^*_n - \widehat{\Theta}_n) \approx [I - \nabla\mathbf{g}(\widehat{\Theta}_n)]^{-1}\sqrt{n}(\mathbf{g}^*(\widehat{\Theta}_n) - \widehat{\Theta}_n). \tag{1}$$

The function $\mathbf{g}^*$ now depends on the bootstrapped data set instead of the original data set. For each bootstrap sample, we can calculate the right-hand side of this equation instead of the more expensive left-hand side. Hence, we approximate the actual estimate in each sample by computing the function $\mathbf{g}^*$ in $\widehat{\Theta}_n$ and then apply a linear correction given by $[I - \nabla\mathbf{g}(\widehat{\Theta}_n)]^{-1}$.

Multivariate MM-estimators for location, shape and scatter depend on two functions $\rho_0 : \mathbb{R} \to \mathbb{R}^+$ and $\rho_1 : \mathbb{R} \to \mathbb{R}^+$ which determine respectively the breakdown point and efficiency of the estimate. Suppose that ρ_0 and ρ_1 satisfy the following conditions:

(R1) ρ is symmetric, twice continuously differentiable and $\rho(0) = 0$,
(R2) ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$ for some finite constant c.

For $X_n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ with $n \geq p+1$, the S-estimators of location and scatter $(\tilde{\mu}_n, \widetilde{\Sigma}_n)$ minimize $\det C$ subject to $\frac{1}{n}\sum_{i=1}^n \rho_0\left(d_i(T,C)\right) = b$ among all $(T,C) \in \mathbb{R}^p \times PDS(p)$ where $d_i(T,C) = \left[(x_i - T)^t C^{-1}(x_i - T)\right]^{\frac{1}{2}}$, $i = 1, \ldots, n$. Denote $\hat{\sigma}_n := (\det \tilde{\Sigma}_n)^{1/(2p)}$. Then the multivariate MM-estimators for location and shape $(\hat{\mu}_n, \widehat{\Gamma}_n)$ minimize $\frac{1}{n}\sum_{i=1}^n \rho_1\left(d_i(T,G)/\hat{\sigma}_n\right)$ among all $(T,G) \in \mathbb{R}^p \times PDS(p)$ for which $\det G = 1$. The MM-estimator for the covariance matrix is $\widehat{\Sigma}_n = \hat{\sigma}_n^2 \widehat{\Gamma}_n$.

S-estimators are a special case of MM-estimators, obtained by choosing ρ_1 equal to ρ_0. The constant b is generally chosen to be $E_{N_p(0,I_p)}[\rho_0(\|x\|)]$ to attain consistency at the normal model. The breakdown point of the S-estimator, i.e. the smallest fraction of observations of the data set that need to be replaced by arbitrary values to carry the estimate beyond all bounds, is $\min(b/\rho_0(\infty), 1 - b/\rho_0(\infty))$. Thus, we can choose the constant c such that $b = \rho_0(\infty)/2$. The MM-estimator inherits the breakdown point of the initial S-estimator. The constant c in the ρ_1-function is chosen to obtain 95% efficiency. Hence, the MM-estimator combines a high breakdown point and a high efficiency. In this paper we use Tukey biweight functions given by $\rho_c(t) = \min(t^2/2 - t^4/(2c^2) + t^6/(6c^4), c^2/6)$ which satisfy assumptions (R1) and (R2).

The multivariate MM-estimators can be written as a system of fixed-point equations as follows

$$
\begin{aligned}
\hat{\mu}_n &= \left(\sum_{i=1}^n \frac{\rho_1'(d_i/(\det\tilde{\Sigma}_n)^{1/(2p)})}{d_i}\right)^{-1} \left(\sum_{i=1}^n \frac{\rho_1'(d_i/(\det\tilde{\Sigma}_n)^{1/(2p)})}{d_i} x_i\right) \\
\widehat{\Gamma}_n &= G\left(\sum_{i=1}^n \frac{\rho_1'(d_i/(\det\tilde{\Sigma}_n)^{1/(2p)})}{d_i}(x_i - \hat{\mu}_n)(x_i - \hat{\mu}_n)^t\right) \\
\widetilde{\Sigma}_n &= \frac{1}{nb}\left(\sum_{i=1}^n p\frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i}(x_i - \tilde{\mu}_n)(x_i - \tilde{\mu}_n)^t + \left(\sum_{i=1}^n \tilde{w}_i\right)\widetilde{\Sigma}_n\right) \\
\tilde{\mu}_n &= \left(\sum_{i=1}^n \frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i}\right)^{-1} \left(\sum_{i=1}^n \frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i} x_i\right)
\end{aligned}
$$

where we denote $G(A) = (\det A)^{-1/p} A$ for $p \times p$ matrices A, and where $d_i = [(x_i - \hat{\mu}_n)^t \widehat{\Gamma}_n^{-1}(x_i - \hat{\mu}_n)]^{1/2}$, $\tilde{d}_i = [(x_i - \tilde{\mu}_n)^t \widetilde{\Sigma}_n^{-1}(x_i - \tilde{\mu}_n)]^{1/2}$ and $\tilde{w}_i = \rho_0(\tilde{d}_i) - \rho_0'(\tilde{d}_i)\tilde{d}_i$. These equations can be used to compute fast approximations to the MM-estimates in each bootstrap sample as in (1); see also Salibian-Barrera et al. (2006). Under certain regularity conditions, the asymptotic distribution of the FRB coincides with that of the S- or MM-estimator (see Van Aelst and Willems (2005) and Salibian-Barrera et al. (2006)).

3 Robust one-sample and two-sample Hotelling test

The one-sample Hotelling test is the standard tool for inference about the center μ of a multivariate distribution. Consider $X_n = \{x_1, \ldots, x_n\}$ a sample from an underlying distribution with location μ and covariance Σ. To test the null hypothesis $H_0 : \mu = \mu_0$ Hotelling proposed the test statistic

$$T^2 := n(\bar{x} - \mu_0)^t S^{-1}(\bar{x} - \mu_0)$$

where $\bar{x}$ is the sample mean and S the sample covariance matrix. Under normality the test statistic is distributed as $[(n-1)p/(n-p)]F_{p,n-p}$.

As the empirical mean and covariance matrix can be highly influenced by outliers, this is also the case for the Hotelling T^2 test statistic. Willems et al. (2002) discuss a robust Hotelling test which uses the Minimum Covariance Determinant estimator instead of the empirical mean and covariance matrix. In Willems and Van Aelst (2004) a bootstrap method is used for the Hotelling test based on the MCD. We propose a similar idea by using S- or MM-estimates instead of the empirical estimates. These estimators can be written as fixed-points equations, hence we can use the FRB.

The distribution of the test statistic under H_0 can be determined through the FRB, assuming that this distribution does not depend on μ_0. The 5% critical value for the test is then given by the 95% quantile of the recalculated statistics

$$n(\hat{\mu}^* - \hat{\mu})(\widehat{\Sigma}^*)^{-1}(\hat{\mu}^* - \hat{\mu}),$$

where $\hat{\mu}$ is the S- or MM-estimate of the original sample and $\hat{\mu}^*$ and $\widehat{\Sigma}^*$ denote the recalculated values of the FRB.

As an alternative to bootstrap, the following asymptotic result may be used. In case of elliptically distributed data and under certain regularity conditions it follows from the asymptotic normality of the S- and MM-location estimators and the consistency of the S- and MM-scatter estimators that for large samples the S- or MM-based T^2 under H_0 should be distributed approximately as $\kappa\chi^2_p$. In case of normality, this κ equals α/β^2 with

$$\alpha = \frac{1}{p} E_{N_p(0,I_p)}[\rho'^2(\|x\|)] \text{ and } \beta = E_{N_p(0,I_p)}\left[\left(1 - \frac{1}{p}\right)\frac{\rho'(\|x\|)}{\|x\|} + \frac{1}{p}\rho''(\|x\|)\right].$$

For the S-estimator κ is calculated using ρ_0 and for the MM-estimator we use ρ_1.

The two-sample Hotelling test which tests $H_0 : \mu_1 = \mu_2$ can be applied in an analogous way. We assume that the observations $X^1_{n_1} = \{x^1_1, \ldots, x^1_{n_1}\}$ and $X^2_{n_2} = \{x^2_1, \ldots, x^2_{n_2}\}$ respectively come from underlying distributions with location μ_1 and μ_2 and common covariance matrix Σ. Classically one uses the sample means $\bar{x}_1$ and $\bar{x}_2$ and the pooled sample covariance

$$S^2_P = \frac{(n_1 - 1)S^2_1 + (n_2 - 1)S^2_2}{n_1 + n_2 - 2}$$

to construct the test statistic

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^t (S_P^2)^{-1}(\bar{x}_1 - \bar{x}_2).$$

Using the FRB, the 5% critical value for the robust version of this test is given by the 95%-quantile of the recalculated statistics

$$\frac{n_1 n_2}{n_1 + n_2}(\hat{\mu}_1^* - \hat{\mu}_2^* - (\hat{\mu}_1 - \hat{\mu}_2))^t (\widehat{\Sigma}^*)^{-1}(\hat{\mu}_1^* - \hat{\mu}_2^* - (\hat{\mu}_1 - \hat{\mu}_2))$$

where $\hat{\mu}_1$ is the S- or MM-estimate of $X_{n_1}^1$, $\hat{\mu}_2$ is the S- or MM-estimate of $X_{n_2}^2$ and $\hat{\mu}_1^*$ and $\hat{\mu}_2^*$ denote the recalculated values by the FRB. Using the recalculated $\widehat{\Sigma}_1^*$ and $\widehat{\Sigma}_2^*$, the pooled covariance $\widehat{\Sigma}^*$ is computed as:

$$\widehat{\Sigma}^* = \frac{(n_1 - 1)\widehat{\Sigma}_1^* + (n_2 - 1)\widehat{\Sigma}_2^*}{n_1 + n_2 - 2}.$$

It can be reasoned in the same way as for the one-sample Hotelling test that under H_0 this test statistic should be distributed approximately as $\kappa\chi_p^2$ for sufficiently large samples.

For the consistency of the FRB method, we can use the same argument as in Salibian-Barrera (2005). Since the asymptotic distribution of the FRB converges to that of the respective estimators, and since the test statistics are 'smooth functions' of the estimators, we obtain correct asymptotic distributions for both the one-sample and the two-sample T^2.

4 Simulations

We illustrate by simulation that the FRB generally performs better than the asymptotic variance. Simulations were performed for one (one-sample Hotelling test) or two data sets (two-sample Hotelling test) of size $n = 50$, 200 and 500 and dimensions $p = 2$, 5 and 10. We consider the following cases:

- multivariate normal $N_p(0, I_p)$
- multivariate Student T distribution with 3 degrees of freedom (T_3)
- symmetric contaminated normal, a proportion $1 - \delta$ following $N_p(0, I_p)$ and a proportion δ following $N_p(0, 9I_p)$.

For each situation we generated 1000 data sets. For each data set we performed the robust Hotelling test based on S- or MM-estimates. We considered both 25% and 50% breakdown point. The percentage of outliers δ was set to $\delta = 0.15$ for 25% breakdown point and $\delta = 0.30$ for 50% breakdown point.

Table 1 lists the observed frequency that the robust one-sample Hotelling T^2 statistic using the MM-estimator is above the 5% and 1% critical value. We considered critical values determined by the FRB (column FRB), using $B = 999$ bootstrap samples, and the asymptotic variance (column ASV).

Table 2 shows these frequencies for the one-sample Hotelling test based on the 25% breakdown S-estimator. From these tables, we see that in most cases the critical values of the test are estimated fairly accurately using the FRB, only for $n = 50$ the results are rather poor. Using the asymptotic distribution, the critical values are somewhat too low in case of normal data. On the other hand, with Student T_3 data and contaminated data and for $p = 5$ and $p = 10$ the critical values are too high, resulting in a lower Type I error.

Tables 3 and 4 show the observed frequencies above the 5% and 1% critical value for the two-sample Hotelling test based on the MM- and S-estimator respectively. In general, we notice the same trends as in Tables 1 and 2.

To summarize FRB is a fast and accurate procedure to find critical values for the S- or MM-based T^2, which requires less assumptions then the ASV approach. Compared to the robust Hotelling test based on the MCD estimator of Willems et al. (2002) we expect the power of our tests to be higher since the S- and MM-estimators are more efficient estimators.

5 Outlook

In further research, we will first investigate what will be the effect of using the S-estimator of He and Fung (2000) instead of the pooled estimator in the two-sample Hotelling test. Secondly, the robust Hotelling test will be compared to the classical test in case of asymmetric contamination. Furthermore, we will use the same idea in the MANOVA context to obtain a robust version of Wilks' lambda using S- and MM-estimators. Finally, we will investigate the power of these robust tests by simulation. Software code for the discussed tests is available at *http://users.ugent.be/∼svaelst/software*.

	level 5%						level 1%					
p	2		5		10		2		5		10	
n	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV
normal data												
50	3.4	6.4	3.5	9.2	0.9	20.3	0.7	1.4	1.0	3.3	0.0	9.3
200	5.9	7.0	5.6	6.5	5.0	7.2	1.1	1.1	2.1	2.2	1.4	1.8
500	5.1	5.1	6.5	6.7	5.3	5.8	1.0	0.8	1.4	1.3	1.2	1.2
Student T_3 data												
50	2.7	7.1	1.4	8.3	0.0	14.2	0.4	2.1	0.1	2.5	0.0	6.6
200	5.6	6.7	4.5	4.4	4.3	3.2	1.5	2.2	1.4	1.2	0.6	0.3
500	6.0	6.9	5.1	3.4	4.6	1.7	1.0	1.4	1.0	0.7	1.5	0.3
normal data with 30% outliers												
50	1.9	7.0	0.2	6.7	0.4	13.2	0.4	1.6	0.0	1.4	0.0	4.5
200	5.4	6.9	4.4	3.6	3.2	1.3	1.3	2.1	0.7	0.8	0.3	0.2
500	5.5	7.3	6.6	3.4	5.3	1.2	1.3	2.0	1.0	0.5	1.3	0.0

Table 1. One-sample Hotelling test based on the 50% breakdown MM-estimator; percentage of erroneous rejections of H_0.

	level 5%						level 1%					
p	2		5		10		2		5		10	
n	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV
normal data												
50	4.2	5.7	6.1	9.1	6.8	19.1	1.1	1.3	1.7	3.0	2.3	7.9
200	5.6	5.9	6.0	6.4	6.5	7.3	1.5	1.2	2.0	2.3	1.5	1.8
500	5.3	5.0	6.5	6.6	6.7	6.2	1.2	0.9	1.5	1.4	1.1	1.2
Student T_3 data												
50	3.6	3.0	0.9	3.1	0.0	8.5	0.4	0.4	0.0	0.6	0.0	3.7
200	5.5	2.8	5.4	1.7	3.6	2.0	1.4	0.8	1.0	0.4	0.8	0.2
500	5.3	2.4	5.0	1.2	4.7	0.8	0.9	0.1	0.8	0.3	0.8	0.3
normal data with 15% outliers												
50	3.1	2.7	0.9	3.5	0.0	8.7	0.5	0.5	0.0	0.9	0.0	3.4
200	5.9	3.0	4.8	1.7	3.6	1.3	0.8	0.1	1.1	0.1	0.5	0.3
500	4.5	2.7	4.9	1.4	4.5	0.4	1.3	0.3	1.3	0.1	0.7	0.0

Table 2. One-sample Hotelling test based on the 25% breakdown S-estimator; percentage of erroneous rejections of H_0.

	level 5%						level 1%					
p	2		5		10		2		5		10	
$n_1 = n_2$	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV
normal data												
50	4.6	7.2	3.7	7.2	1.1	13.2	0.3	1.3	0.8	2.6	0.3	4.7
200	4.5	4.6	5.0	5.5	5.4	7.2	1.3	1.8	0.7	1.2	0.9	1.6
500	5.7	5.6	5.0	4.8	5.5	5.9	1.2	1.3	1.1	1.2	1.4	1.4
Student T_3 data												
50	4.0	7.4	1.9	5.5	0.3	7.3	0.6	1.5	0.5	1.6	0.1	1.5
200	4.7	5.7	4.6	4.1	3.5	2.1	1.2	1.5	0.8	0.8	0.9	0.5
500	4.5	6.3	4.6	3.1	3.4	1.0	0.9	1.3	0.8	0.2	0.6	0.3
normal data with 30% outliers												
50	3.2	7.1	0.8	4.7	0.2	4.1	0.6	1.6	0.1	1.0	0.0	1.1
200	5.2	7.0	3.8	2.7	3.3	1.2	1.3	2.2	0.6	0.5	0.7	0.3
500	4.5	5.5	5.4	2.9	5.1	0.7	1.0	1.3	1.4	0.6	1.0	0.1

Table 3. Two-sample Hotelling test based on the 50% breakdown MM-estimator; percentage of erroneous rejections of H_0.

	level 5%						level 1%					
p	2		5		10		2		5		10	
$n_1 = n_2$	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV	FRB	ASV
normal data												
50	5.3	6.0	5.7	7.0	7.7	13.5	0.4	0.9	1.5	2.6	2.2	4.6
200	4.3	4.5	5.2	5.5	6.3	6.8	1.2	1.3	0.6	1.0	1.4	1.4
500	5.5	5.6	4.9	4.4	5.5	5.7	1.4	1.0	1.1	1.3	1.9	1.5
Student T_3 data												
50	4.2	3.1	3.1	2.7	0.5	4.0	0.5	0.4	0.8	0.9	0.0	0.6
200	4.1	2.2	4.9	1.7	3.8	1.1	1.2	0.3	0.7	0.2	0.7	0.1
500	4.5	2.0	4.6	0.7	3.6	1.0	1.1	0.2	0.6	0.0	0.9	0.2
normal data with 15% outliers												
50	4.3	2.7	2.6	2.1	0.0	3.7	0.8	0.5	0.3	0.5	0.0	0.7
200	4.6	2.3	3.6	1.0	4.0	1.0	1.1	0.4	0.5	0.1	0.7	0.0
500	5.8	2.9	5.3	1.6	4.8	0.4	1.7	0.6	1.3	0.2	1.0	0.0

Table 4. Two-sample Hotelling test based on the 25% breakdown S-estimator; percentage of erroneous rejections of H_0.

References

DAVIES, P.L. (1987): Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics 15, 1269-1292.*

EFRON, B. (1979): Bootstrap methods: another look at the jackknife. *The Annals of Statistics 7, 1-26.*

HE, X. and FUNG, W.K. (2000): High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis 72, 151-162.*

LOPUHAÄ, H.P. (1989): On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics 17, 1662-1683.*

SALIBIAN-BARRERA, M. (2005): Estimating the p-values of robusts tests for the linear model. *Journal of Statistical Planning and Inference 128, 241-257.*

SALIBIAN-BARRERA, M., VAN AELST, S. and WILLEMS, G. (2006): Principal components analysis based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association 101, 1198-1211.*

SALIBIAN-BARRERA, M., VAN AELST, S. and WILLEMS, G. (2008): Fast and robust bootstrap. *Statistical Methods and Applications 17, 41-71.*

SALIBIAN-BARRERA, M. and ZAMAR, R.H. (2002): Bootstrapping robust estimates of regression. *The Annals of Statistics 30, 556-582.*

TATSUOKA, K.S. and TYLER, D.E. (2000): On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *The Annals of Statistics 28, 1219-1243.*

VAN AELST, S. and WILLEMS, G. (2005): Multivariate regression S-estimators for robust estimation and inference. *Statistica Sinica 15, 981-1001.*

WILLEMS, G., PISON, G., ROUSSEEUW, P.J. and VAN AELST, S. (2002): A robust Hotelling test. *Metrika 55, 125-138*.

WILLEMS, G. and VAN AELST, S. (2004): A fast bootstrap method for the MCD estimator. In: J. Antoch (Ed.): *Proceedings in Computational Statistics 2004*. Springer, Physica - Verlag, Heidelberg, 1979–1986.

Estimating Partial Correlations Using the Oja Sign Covariance Matrix

Daniel Vogel and Roland Fried

Fakultät Statistik, Technische Universität Dortmund
Vogelpothsweg 87, 44225 Dortmund, Germany, *daniel.vogel@udo.edu*, *fried@statistik.uni-dortmund.de*

Abstract. We investigate the Oja sign covariance matrix (Oja SCM) for estimating partial correlations in multivariate data. The Oja SCM estimates directly a multiple of the precision matrix and is based on the concept of Oja signs, a multivariate generalisation of the univariate sign function, which obey some form of affine equivariance property. Our simulations show that the asymptotic distribution gives a good approximation of the exact finite-sample distribution already for samples of moderate size. We find it to equal the performance of the classical sample partial correlation in the normal model and outperform it in the case of heavy-tailed distributions. The high computational costs are its main disadvantage.

Keywords: elliptical distribution, graphical model, asymptotic distribution, robustness, nonparametrics

1 Introduction

Let $k \geq 3$ and $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$ with $\boldsymbol{X} = (X_1, X_2)$, $\boldsymbol{Y} = (Y_1, ..., Y_{k-2})$ be a k-dimensional random vector having a non-singular covariance matrix Σ. Let $\hat{X}_i(\boldsymbol{Y})$, $i = 1, 2$, be the projection of X_i onto the space of all affine linear functions of $\boldsymbol{Y}$. Then the *partial correlation of* X_1 *and* X_2 *given* $\boldsymbol{Y}$ is defined by

$$\varrho_{1,2\bullet Y} = \frac{\operatorname{cov}\Big(X_1 - \hat{X}_1(\boldsymbol{Y}),\, X_2 - \hat{X}_2(\boldsymbol{Y})\Big)}{\sqrt{\operatorname{var}\Big(X_1 - \hat{X}_1(\boldsymbol{Y})\Big)\operatorname{var}\Big(X_2 - \hat{X}_2(\boldsymbol{Y})\Big)}},$$

i.e. it is the correlation between the residuals $X_1 - \hat{X}_1(\boldsymbol{Y})$ and $X_2 - \hat{X}_2(\boldsymbol{Y})$. The partial correlation $\varrho_{1,2\bullet Y}$ can be computed from the covariance matrix Σ. It holds

$$\varrho_{1,2\bullet Y} = -\frac{k_{1,2}}{\sqrt{k_{1,1}k_{2,2}}}.$$

where $k_{i,j}$, $i, j = 1, ..., k$, are the elements of $K = \Sigma^{-1}$, see e.g. Whittaker (1990). The matrix K is called the *concentration matrix* (or *precision matrix*) of $\boldsymbol{Z}$.

Partial correlations play an important role for instance in graphical models, where the key notion is *conditional independence*. Roughly, a *graphical model* is a family of k-dimensional distributions of $\boldsymbol{Z} = (Z_1, ..., Z_k)$ that satisfy some given pairwise conditional independence restrictions on the components of $\boldsymbol{Z}$. One can then, based on the these pairwise conditional indepence conditions, draw inferences about conditional independencies between arbitrary disjoint subsets of $\{Z_1, ..., Z_k\}$ given some other subvector. The classical theory of graphical models for continuously distributed variables is built on the normality assumption. If $\boldsymbol{Z} = (X_1, X_2, \boldsymbol{Y})$ has a multivariate normal distribution, then X_1 and X_2 are conditionally independent given $\boldsymbol{Y}$ if and only if $\varrho_{1,2\bullet Y} = 0$, which is equivalent to $k_{1,2} = 0$. A Gaussian graphical model is thus specified by the concentration matrix K.

Now if we wish to estimate the partial correlation $\varrho_{1,2\bullet Y}$ from a sample of n independent realisations of the vector $\boldsymbol{Z} = (X_1, X_2, \boldsymbol{Y})$, then

$$\hat{\varrho}_{1,2\bullet Y} = -\frac{\hat{k}_{1,2}}{\sqrt{\hat{k}_{1,1}\hat{k}_{2,2}}}, \tag{1}$$

is a natural choice of an estimator for $\varrho_{1,2\bullet Y}$, where $\hat{K} = (\hat{k}_{i,j})_{i,j}$ is a suitable estimator of the precision matrix K. Equivalent to looking at $\hat{\varrho}_{1,2\bullet Y}$ is looking at the matrix-valued estimator

$$\hat{C} = (\hat{K}_D)^{-\frac{1}{2}}\hat{K}(\hat{K}_D)^{-\frac{1}{2}}, \tag{2}$$

where $\hat{K}_D$ denotes the diagonal matrix having the same diagonal as K. The matrix $\hat{C}$ is 1 on the diagonal and contains the negative estimated partial correlations as its off-diagonal elements, i.e. $\hat{\varrho}_{1,2\bullet Y} = -\hat{c}_{1,2}$. The estimator $\hat{K}$ can be the inverse of basically any multivariate covariance or shape estimator. We compare the partial correlation estimator based on the Oja SCM $\hat{K}^O$ to the classical normal MLE $\hat{K}^E$. Both estimators are properly defined in Section 2. Section 3 presents asymptotic distributions and influence functions under the elliptical model. Section 4 reports the findings of some finite-sample simulations on the distribution of the estimators and their sensitivity against contaminations. Section 5 is a short summary.

2 The Oja sign covariance matrix

In this section we define the Oja sign covariance matrix (Oja SCM), as it is done in Visuri et al. (2000). For an instant suppose we have a univariate data set $\mathbb{X} = (x_1, ..., x_n)$, $n \in \mathbb{N}$. We want to call $\mathrm{sgn}_{\mathbb{X}}(x) = \mathrm{sgn}(x - \mathrm{med}(\mathbb{X}))$, $x \in \mathbb{R}$, the *sign of x w.r.t. the data sample* $\mathbb{X}$. Here sgn denotes the usual univariate sign function ($\mathrm{sgn}(x) = \frac{x}{|x|}$ if $x \neq 0$ and zero otherwise) and $\mathrm{med}(\mathbb{X})$ the univariate median function applied to $\mathbb{X}$. There are several possibilities how to generalise this notion to the multivariate setting. One possibility is the Oja median and the Oja sign. Consider the k-variate data sample

$\mathbb{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, $n \in \mathbb{N}$, and let

$$Q_p = \big\{ q = \{i_1, ..., i_p\} \,\big|\, 1 \le i_1 < ... < i_p \le n \big\}, \quad 0 \le p \le n,$$

be the system of all subsets of $\{1, ..., n\}$ with p elements and $N_p = |\, Q_p| = \binom{n}{p}$. Then the *Oja median of the data sample* $\mathbb{X}$ is defined as

$$\mathbf{Omed}(\mathbb{X}) = \operatorname*{arg\,min}_{x \in \mathbb{R}^k} \sum_{Q_k} \Big| \det \begin{pmatrix} 1 & ... & 1 & 1 \\ \boldsymbol{x}_{i_1} & ... & \boldsymbol{x}_{i_k} & \boldsymbol{x} \end{pmatrix} \Big|,$$

and $\mathbf{omed}(\mathbb{X})$ as the gravity center of the set $\mathbf{Omed}(\mathbb{X})$. The *Oja sign of the point* $\boldsymbol{x} \in \mathbb{R}^k$ *w.r.t.* $\mathbb{X}$ is

$$\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}) = \frac{1}{N_{k-1}} \sum_{Q_{k-1}} \nabla_{\boldsymbol{x}} \Big| \det(\boldsymbol{y}_{i_1} \, ... \, \boldsymbol{y}_{i_{k-1}} \, \boldsymbol{y}) \Big|,$$

where $\boldsymbol{y} = \boldsymbol{x} - \mathbf{omed}(\mathbb{X})$ and $\boldsymbol{y}_i = \boldsymbol{x}_i - \mathbf{omed}(\mathbb{X})$, $i = 1, ..., n$. Note that contrary to $\operatorname{sgn}_{\mathbb{X}}$ the Oja sign $\mathbf{osgn}_{\mathbb{X}}$ does *not only* depend upon the data sample $\mathbb{X}$ through its center point $\mathbf{omed}(\mathbb{X})$. The Oja median and the Oja sign are proper multivariate generalisations of the univariate concepts, in the sense that for $k = 1$ they yield med and $\operatorname{sgn}_{\mathbb{X}}$ as defined above. If $k = 1$,

$$\mathbf{Omed}(\mathbb{X}) = \operatorname*{arg\,min}_{x \in \mathbb{R}} \sum_{1}^{n} \Big| \det \begin{pmatrix} 1 & 1 \\ x_i & x \end{pmatrix} \Big|$$

and

$$\mathbf{osgn}_{\mathbb{X}}(x) = \frac{\partial}{\partial x} |\det(x - \mathbf{omed}(\mathbb{X}))|.$$

Finally we construct a scatter estimate based on the Oja sign. The most frequently used estimate of scatter is the *empirical covariance matrix*

$$\mathrm{ECM}(\mathbb{X}) = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)^T,$$

where $\overline{\boldsymbol{x}}_n$ is the mean of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. ECM is the (biased) MLE of the covariance matrix Σ, provided the latter exists. The Oja sign covariance matrix follows the same construction principle as the ECM, with $\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n$ being replaced by $\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}_i)$. Thus we write down

$$\mathrm{OSCM}(\mathbb{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}_i) \mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}_i)^T.$$

Now we define two estimators of the *shape* of the precision matrix, i.e. the precision matrix up to scale, $\hat{K}^E = \mathrm{ECM}^{-1}$ and $\hat{K}^O = \mathrm{OSCM}$. It is on purpose that $\mathrm{OSCM}(\mathbb{X})$ is not inverted. It already estimates the precision matrix up to scale, cf. section 3. We denote the corresponding estimators for C and $\varrho_{1,2 \bullet Y}$ by $\hat{C}^E$ and $\hat{\varrho}^E_{1,2 \bullet Y}$, respectively $\hat{C}^O$ and $\hat{\varrho}^O_{1,2 \bullet Y}$. As usual $\hat{c}^E_{i,j}$ and $\hat{c}^O_{i,j}$ denote the elements of $\hat{C}^E$ and $\hat{C}^O$, respectively.

3 Some asymptotic results

A common generalisation of the multivariate normal model is the family of elliptical distributions. It is often considered in multivariate data analysis since the first and second order characteristics are an intuitive description of the actual shape of the distribution.

The density f_0 of a *spherical distribution* F_0 on $\mathbb{R}^k$ is of the form $f_0(\boldsymbol{x}) = g(\boldsymbol{x}^T\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^k$, where $g : [0,\infty) \to [0,\infty)$ is such that f_0 integrates to 1. If furthermore the covariance matrix of F_0 is the identity matrix I_k, we call F_0 a *standardized spherical distribution*. In the following we assume that $\boldsymbol{X}_0 \sim F_0$, where F_0 is a standardized spherical distribution admitting the Lebesgue-density f_0. Then, for any non-singular $A \in \mathbb{R}^{k\times k}$ and $\boldsymbol{b} \in \mathbb{R}^k$ the random variable $\boldsymbol{X} = A\boldsymbol{X}_0 + \boldsymbol{b}$ has an elliptical distribution F with mean vector $\boldsymbol{b}$, non-singular covariance matrix $\Sigma = AA^T$ and density

$$f(\boldsymbol{x}) = \det(\Sigma)^{-\frac{1}{2}} g((\boldsymbol{x}-\boldsymbol{b})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{b})).$$

Following the notation of Bilodeau und Brenner (1999) we denote the class of all elliptical distributions on $\mathbb{R}^k$ having mean $\boldsymbol{b}$ and covariance Σ by $E_k(\boldsymbol{b},\Sigma)$. By choosing the function g we model the tail behaviour of the distribution F. The most prominent member of the class of elliptical distributions is the normal distribution $N_k(\boldsymbol{b},\Sigma)$, which corresponds to $g_{N_k}(y) = (2\pi)^{-\frac{k}{2}}\exp\left(-\frac{1}{2}y\right)$. Another important subclass of the elliptical model is the family of multivariate $t_{\nu,k}$-distributions with

$$g_{t_{\nu,k}}(y) = \frac{\Gamma(\frac{\nu+k}{2})}{(\nu\pi)^{\frac{k}{2}}\Gamma(\frac{\nu}{2})}(1-\frac{y}{\nu})^{-\frac{\nu+k}{2}}.$$

Here the first subscript ν denotes the degrees of freedom. The $t_{\nu,k}(\boldsymbol{b},\Sigma)$ distribution converges to $N_k(\boldsymbol{b},\Sigma)$ as $\nu\to\infty$ and is, for small ν, a popular example of a heavy-tailed distribution. Its moments are finite only up to order $(\nu-1)$.

It is considered a shortcoming of the elliptical model that it does not include independent margins, unless the margins are normal, cf. e.g. Bilodeau and Brenner (1999), page 51. Consequently, partial uncorrelatedness (i.e. an off-diagonal zero entry in the precision matrix K) does in general not mean conditional independence. It is, however, equivalent to *conditional uncorrelatedness*, cf. Baba et al. (2004). Thus in any statistical analysis incorporating only first and second order characteristics (which is very often the case) partial correlation still provides a useful measure of conditional linear independence.

The estimator $\hat{C}$ is via (2) a function of $\hat{K}$. Hence, if the asymptotic distribution of $\hat{K}$ is known, the asymptotic distribution of $\hat{C}$ can be assessed applying the delta method. Ollila et al. (2003) state the following lemma about the Oja SCM $\hat{K}^O$.

Lemma 13. *If $F \in E_k(\boldsymbol{b}, \Sigma)$ and F_0 is its corresponding standardized spherical distribution, furthermore $\mathbb{X}_n = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$, $\boldsymbol{X}_i \sim F$ i.i.d., $i = 1, ..., n$, then*

$$\begin{aligned} &(I)\ \hat{K}^O(\mathbb{X}_n) \xrightarrow{p} \gamma_{F_0} \det(\Sigma)\Sigma^{-1}, \\ &(II)\ \sqrt{n}\Big(\hat{K}^O(\mathbb{X}_n) - \gamma_{F_0} \det(\Sigma)\Sigma^{-1}\Big) \xrightarrow{\mathcal{L}} N_k(0, \Gamma), \end{aligned}$$

where γ_{F_0} is a constant depending only on the dimension k and $\mathbb{E}||\boldsymbol{X}_0||$, $\boldsymbol{X}_0 \sim F_0$, and Γ can be written as a function of Σ, k and $\mathbb{E}||\boldsymbol{X}_0||$. Both, γ_{F_0} and Γ, are made explicit in Ollila et al. (2003).

If the true partial correlation is zero, we can also apply Slutsky's lemma to (1) and deduce the following from Lemma 13 by straightforward calculations.

Lemma 14. *If F, F_0 and $\boldsymbol{X}_0$ are as in Lemma 13 and $k_{1,2} = 0$ ($K = \Sigma^{-1}$), then*

$$\sqrt{n}\hat{\varrho}^O_{1,2\bullet Y} \xrightarrow{\mathcal{L}} N\bigg(0, \frac{k}{k+2}\Big(\frac{4k}{(\mathbb{E}||\boldsymbol{X}_0||)^2} - 3\Big)\bigg).$$

If $\boldsymbol{X_0} \sim N_k(0, I_k)$, then $\mathbb{E}||\boldsymbol{X}_0|| = \sqrt{2}\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}$. The corresponding expression for the asymptotic variance can be shown to converge to 1 as $k \to \infty$. For $k = 4$ (as in Figure 1) it equals $\frac{4^4}{3^3\pi} - 2 \approx 1.018$. Ollila et al. (2003) also report the value of $\mathbb{E}||\boldsymbol{X}_0||$ at the t-distribution. In the case of $k = 4$ and $\nu = 3$ (as in Figure 2) it results in an asymptotic variance of $\frac{2^7}{3^3} - 2 \approx 2.741$.

Lemma 14 allows to construct an asymptotic level-α-test for conditional independence, e.g. based on $n^{-1}(\hat{\varrho}^O_{1,2\bullet Y})^2$, which – appropriately standardized – will converge to a χ^2_1-distribution. It is intuitive from the results reported here that such a test – although its properties still need to be thoroughly assessed – is at the normal model asymptotically almost as efficient as the usual normal LR-test but has better robustness properties. Furthermore this test can easily be extended to an asymptotic test for conditional uncorrelatedness in the elliptical model by additionally estimating $\mathbb{E}||\boldsymbol{X}_0||$.

The asymptotics of the normal MLE $\hat{K}^E$ under normality can be found in textbooks on graphical models such as Lauritzen (1996), but a rigorous treatment under elliptical distributions is not known to us. However, since $\hat{K}^E = \hat{\Sigma}^{-1}$ is a function of the covariance estimator $\hat{\Sigma}$, one can again apply the delta method. The asymptotics of $\hat{\Sigma}$ in the elliptical model can be found in textbooks on multivariate statistics such as Bilodeau and Brenner (1999). In analogy to Lemma 14 we get

Lemma 15. *If $F = N_k(\boldsymbol{b}, \Sigma)$, Σ non-singular, then*

$$\sqrt{n}(\hat{\varrho}^E_{1,2\bullet Y} - \varrho_{1,2\bullet Y}) \xrightarrow{\mathcal{L}} N\big(0, (1 - \varrho^2_{1,2\bullet Y})^2\big).$$

In the general elliptical model a similar expression for the asymptotic variance is obtained, which in addition depends on the fourth order characteristics of F. However, if the fourth moments of F are not finite (as it is the case for $t_{3,4}$ in Figure 1) this standard approach via $\hat{\Sigma}$ is not possible.

The influence function is an important tool in robust analysis. It desribes the robustness of a statistical procedure against infinitesimal contaminations. For reasons of simplicity we consider the influence functions of our estimators at the standardized spherical distribution F_0. If we write $\boldsymbol{u} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||}$, then

$$IF(\boldsymbol{x}, \hat{C}^O, F_0) = k\Big(1 - \frac{2||\boldsymbol{x}||}{\mathbb{E}||\boldsymbol{X}_0||}\Big)\Big(\boldsymbol{u}\boldsymbol{u}^T - (\boldsymbol{u}\boldsymbol{u}^T)_D\Big)$$

whereas the influence function of $\hat{C}^E$ is

$$IF(\boldsymbol{x}, \hat{C}^E, F_0) = -||\boldsymbol{x}||^2\Big(\boldsymbol{u}\boldsymbol{u}^T - (\boldsymbol{u}\boldsymbol{u}^T)_D\Big),$$

cf. Croux and Haesbroeck (2000). The former is affine linear and the latter quadratic in $||\boldsymbol{x}||$.

4 Simulation results

We carried out a simulation study using several elliptical distributions to examine how the finite-sample performance relates to the asymptotics. In the examples that follow we fix the mean to zero and the covariance matrix to

$$\Sigma = \begin{pmatrix} 1 & -0.865 & 0.657 & -0.231 \\ -0.865 & 1 & -0.510 & 0.077 \\ 0.657 & -0.510 & 1 & -0.601 \\ -0.231 & 0.077 & -0.601 & 1 \end{pmatrix},$$

which corresponds to the following matrix of partial correlations

$$-C = \begin{pmatrix} -1 & -0.8 & 0.4 & 0 \\ -0.8 & -1 & 0 & -0.2 \\ 0.4 & 0 & -1 & -0.6 \\ 0 & -0.2 & -0.6 & -1 \end{pmatrix}.$$

Figure 1 shows the approximated densities of $-\hat{c}^O_{1,4}$ and $-\hat{c}^E_{1,4}$ (left plot) and $-\hat{c}^O_{1,3}$ and $-\hat{c}^E_{1,3}$ (right plot) calculated from 30 observations drawn from a normal distribution with covariance Σ as above. The true values to be estimated, $\varrho_{1,4\bullet 2,3} = -c_{1,4} = 0$ and $\varrho_{1,3\bullet 2,4} = -c_{1,3} = 0.4$, respectively, are indicated by the vertical lines. The density estimation is based on 2000 repititions, using the R function `density()` with a Gauss kernel and bandwidth

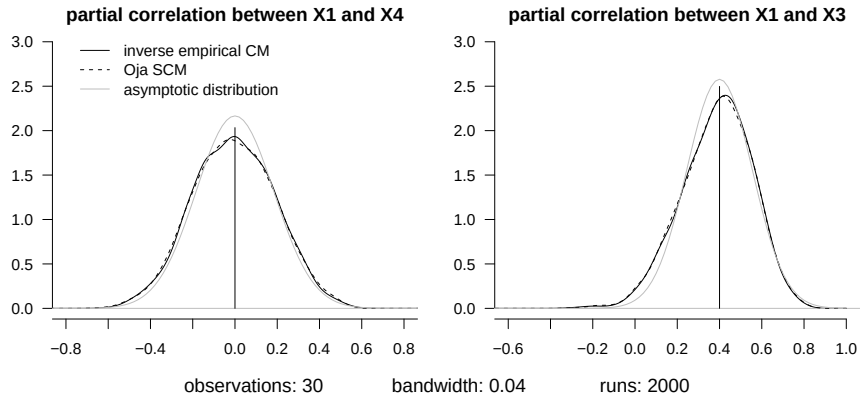


Fig. 1. Densities of two partial correlation estimators at the multivariate normal distribution.

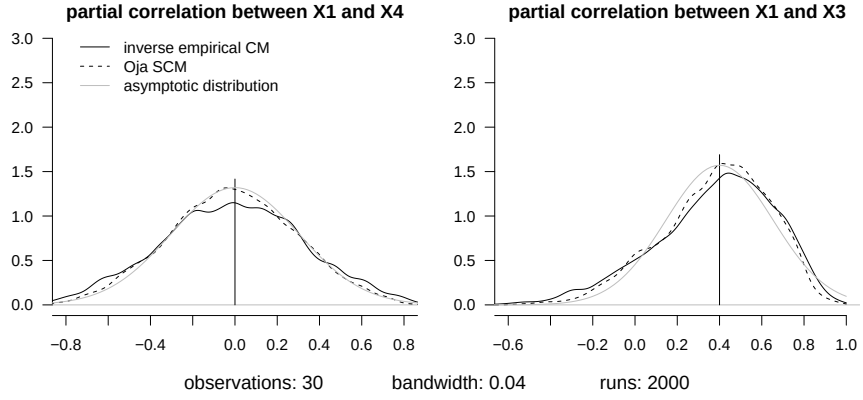


Fig. 2. Densities of two partial correlation estimators at the t_3-distribution.

.04. The solid grey lines are the asymptotic distributions of $-\hat{c}^O_{1,4}$ (left) and $-\hat{c}^O_{1,3}$, cp. Section 3. We can not detect any relevant difference between both estimators. In fact, the asymptotic relative efficiency of $\hat{c}^O_{i,j}$ at the normal model (compared to the MLE $\hat{c}^E_{i,j}$) is more than 98%.

Figure 2 shows the results of an experiment with the same parameters except that the population distribution is now $t_{3,4}(\boldsymbol{b}, \Sigma)$. We find that both estimators have a higher variability (compared to the normal model), but the Oja SCM estimator $\hat{c}^O_{i,j}$ performs substantially better than the MLE $\hat{c}^E_{i,j}$. It should be noted, though, that in the case of light tails the picture is reversed, but still both estimators are more variable than in the normal model. Again,

the solid grey lines represent the asymptotic distributions of $-\hat{c}^O_{1,4}$ and $-\hat{c}^O_{1,3}$, respectively.

In the simulation study we also examined the partial correlation estimator $\hat{C}^O$ under outlier scenarios. We found that it, though not highly robust, is less susceptible to outliers than the normal MLE $\hat{C}^E$. This is an expected behaviour considering the structure of its influence function.

5 Conclusion

The Oja SCM is well suited to the task of estimating partial correlations, in particular at heavy-tailed distributions. Note that $\sqrt{n}$-consistency of the Oja SCM only requires finite second moments. In the normal model its asymptotic and finite-sample efficiencies (almost) equal those of the MLE. The advantage is higher robustness against model misspecification. If the true distribution has heavier than Gaussian tails, the normal MLE looses strongly in efficiency whereas the Oja SCM estimator is little affected. Still, one undetected heavy outlier can make it break down.

Its major drawback remains the computational costs. Its computation necessitates the evaluation of $\binom{n}{k-1}$ $(k-1)$-dimensional hyperplanes. Using a randomized version, i.e. drawing at random a subsample of these $\binom{n}{k-1}$ hyperplanes, allows to push the limit a little bit further up to which n and k the Oja SCM is computable. In our computer experiments the approximation error turned out to be negligible compared to the estimation error up to $n = 60$ when the size of the subsample was 10%.

Finally, Visuri et al. (2000) also propose the Oja rank covariance matrix, which is based on Oja ranks instead of Oja signs. It is very similar to the Oja SCM in construction and statistical properties and exhibited the same performance in simulations.

References

BABA, K., SHIBATA, R. and SIBUYA, M. (2004): Partial correlation and conditional correlation as measures of conditional independence. *Australian and New Zealand Journal of Statistics 46 (4), 657-664.*

BILODEAU, M. and BRENNER, D. (1999): *Theory of multivariate statistics.* Springer, New York.

CROUX, C. and HAESBROECK, G. (2000): Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika 87 (3), 603-618.*

LAURITZEN, S.L. (1996): *Graphical models.* Oxford University Press, Oxford.

OJA, H. (1999): Affine invariant multivariate sign and rank tests. *Scandinavian Journal of Statistics 26 (3), 319-343.*

OLLILA, E., OJA, H. and CROUX, C. (2003): The affine equivariant sign covariance matrix: asymptotic behaviour and efficiencies. *Journal of Multivariate Analysis 87 (2), 328-355.*

VISURI, S., KOIVUNEN, V. and OJA, H. (2000): Sign and rank covariance matrices. *Journal of Statistical Planning and Inference 91, 557-575.*

WHITTAKER, J. (1990): *Graphical models in applied multivariate statistics.* John Wiley & Sons Ltd, Chichester.

Part XIX

Simulation

A Preliminary Comparison of Methods for Predicting Curves from Incomplete Data Sets - Bayesian Versus Semi-Bayesian Approaches

Carmen Ana Cabán-Mejías[1] and Toni Monleón-Getino[1]

Department of Statistics, University of Barcelona
Avda. Diagonal 645, 08028 Barcelona (Spain)
ccabanme8@alumnes.ub.edu, amonleong@ub.edu

Abstract. This study compared two different Bayesian statistical methods used to predict curves from incomplete pharmacokinetic (PK) data sets: a semi-Bayesian prediction method using the Empirical Bayes Algorithm (EBA) and a Bayesian prediction method using the Montecarlo Markov Chain (MCMC) in conjunction with Gibbs sampling. We used the diference between the predicted and the observed theophylline concentration values to calculate the mean square error (ε) and statistically validate the two methods. Our results indicate that the MCMC method offers greater precision and accuracy than EBA. Although these results are preliminary and still incomplete, they suggest that the MCMC method requires less information to better predict the kinetics record for a new case.

Keywords: Bayesian inference, MCMC, modeling, pharmacokinetic

1 Introduction

The objective of this study is to complete a Phase I clinical trial whose pharmacokinetic (PK) objective was originally modeled as a non-linear mixed model (Monleón et al. (2005)). Specifically, we compared two different Bayesian statistical methods used to estimate curves from as little information as possible.

The base-model used for this study was cited by Smith (2004) who compared different methodologies for modeling non-linear PK curves based in a study cited by Pinheiro and Bates (1995) and using their data. The original study determined the levels (mg/L) of the antiasthmatic drug theophylline in the blood serum of 12 healthy volunteers at different times (0.00, 0.25, 0.50, 1.25, 2.00, 3.75, 5.00, 7.00, 9.00, 12.00 and 24.50 hours after being administered the drug; time is approximately). Each volunteer was administered a different dose of theophylline. The doses ranged from 3.1 to 5.86 mg depending on each subjects individual characteristics. A first-order PK model was adjusted according to the levels observed in the 12 individuals at the different times.

1.1 Adjustment of the mixed-model

The parameters of the model were estimated using the PROC NLMIXED procedure in SAS. The NLMIXED procedure (Wolfinger, R.D. (1997)) fits nonlinear mixed models by numerically maximizing an approximation to the marginal likelihoodthat is, the likelihood integrated over the random effects. Different integral approximations are available, the primary one being adaptive Gaussian quadrature (Pinheiro and Bates (1995)). This approximation uses the empirical Bayes algorithm (EBA) estimates of the random effects (analogous to empirical BLUPs in linear mixed models) as the central point for the quadrature, and updates them for every iteration procedure fits nonlinear mixed models, that is, models in which both fixed and random effects are permitted to have a nonlinear relationship to the response variable. The nonlinear mixed models has been determined to be adequate for those calculations (Monleón et al. (2005)).

The PK model is as follows:

$$Cp = \frac{Dk_e k_a}{Cl(k_a - k_e)}[exp(-k_e t) - exp(-k_a t)] + e_i \tag{1}$$

Where Cp is the theophylline concentration (mg/L) in blood serum, D is the dose (mg) of theophylline administered, Cl is the clearance rate in liters/hour (= 0.04), k_a is the absorption ratio (= 1.58), k_e is the excretion ratio (= 0.086), and t is the time elapsed after administration of the theophylline in hours. Patients were monitored for 25 hours.

The author of the model states that the random parameters k_a and Cl are associated to the following parameters:

$$k_a = exp(0.47 + b_{0i})$$

$$Cl = exp(-3.22 + b_{1i})$$

where $b_{0i} \sim N(0, 0.03)$, $b_{1i} \sim N(0, 0.4)$ and $e_{ij} \sim N(0, 0.7)$.

In order to simplify the computational model, it is assumed that the random factors ((b_{0i} and b_{1i}) are independent of each other and of the residuals, which in turn are independent and identically distributed.

The conceptual model discussed above has been statistically validated (Monleón et al. (2005)). Both the model and the real values of theophylline concentration (mg/L) for the 12 individuals are represented in graphical form (Figure 1). The intervals predicted, based on the assumption of normal asymptotic distribution, for nominal 95 percent confidence based on the actual data are also shown.

Figure 1 shows the actual values of theophylline concentration at different times. They are distributed within the confidence interval of the estimated prediction for the conceptual model. The simulated PK curve (the study reported in Monleón et al. (2005)) verifies the previously estimated conceptual model statistically, as the curve of the conceptual model is consistent with

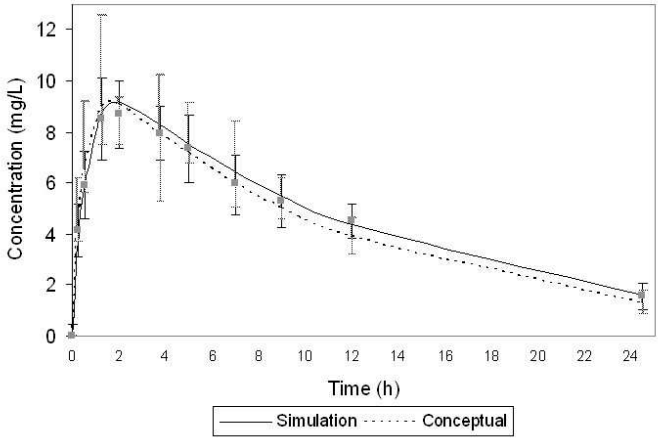


Fig. 1. Pharmacokinetic curve of theophylline according to the conceptual model and simulation. The 95 percent confidence intervals are represented.

the curve predicted by the simulation model. The simulation model is validated by the actual data since the variability represented by the confidence intervals of the simulated concentrations are within the normal range.

1.2 Adjustment of the MCMC model

The non-linear Bayesian model (MCMC) that was used in the program WinBUGS to predict the curves for the concentration of the drug is an improvement on a model originally studied in 2005 by Monleón et al. The MCMC model is as follows:

$$Cp = \frac{D * k_a}{vol * k_a - cl}\left[exp(\frac{-cl}{vol * t}) - exp(-k_a * t)\right] \tag{2}$$

where Cp is the theophylline concentration (mg/L) in blood serum, D is the dose (mg) of medication administered to the patient, Cl is the clearance rate in Liter/hour, k_a is the absortion ratio, vol is the volume of distribution, and t is the time elapsed after administration of the theophylline in hours. Although model (2) incorporated few differences was very similar to the model presented in (1); it did not cover all the kinetic parameters. Due to its complexity, the MCMC model is still in the validation phase. The a priori distributions of the model parameters are not informative.

1.3 Objectives

The objective of the present study is to improve the prediction of the trajectory of temporary curves for new individuals with missing observations,

taking into account prior available information. We aim to find an alternative to prediction using the Empirical Bayes Algorithm (EBA) known as the semi-Bayesian method, which is the usual method for prediction in mixed-models. We are particularly interested in the case where no prior probability distributions of the parameters in the model are available, whether Gaussian or previously determined, and neither are pre-established covariance structures or errors with normal distributions. The alternative method should be as simple as possible to specify in terms of parameter modeling and calculation. This will provide benefits in terms of the power to predict a kinetic curve for a new case for which less information is available, which will lead to increased applicability in clinical pharmacology.

2 Methodology

Tempelman (1998) reviewed the application of various methods for Generalized Linear Mixed Models including the MCMC method. Nevertheless, we did not find any study comparing the performance of methods to generate predicting curves from incomplete data sets based corresponding to non-linear mixed models (NLMM).

Although there are other different methods for predicting incomplete curves in NLMM such as smoothing, they were not compared in this study because it requires large effort on modelling, the corresponding statistical validation, and the subsequent implementation of the various experiments. Unfortunately those analyses are outside of the present work due to limited time and financial resources. Nevertheless, we consider that it would be necessary to compare and contrast those methods in terms of prediction errors using different public data sets.

Two methods used to predict incomplete curves (curves with data points missing) for a new individual were compared in this study. The two methods were as follows:

- A Semi-Bayesian prediction using the EBA (Morris (1983)), commonly used in mixed model prediction, using the statistical analysis software SAS version 9. PROC NLMIXED enables to use the estimated model to construct predictions of arbitrary probability functions by using the parameter estimates and the EBA estimates of the random effects. PROC NLMIXED approximates their standard errors using the first derivatives of the function that was specify (the delta method). EBA statement constructs predictions for each observation in the input data set. These predictions are linear functions of the EBA estimates of the random effects b_{0i} and b_{1i} in (1) ($b_{0i} \sim N(0, 0.03)$, $b_{1i} \sim N(0, 0.4)$).
- A Bayesian prediction method using the MCMC through Gibbs sampling, using WinBUGS software (Spiegelhalter, D. (2003)). Bayesian methods are being increasingly applied to statistical inference in sciences partly

```
#-----------------------------------------------------------------------------------------------------------
#  MCMC model definition used to predict the missing observations in the
#  experimental simulations performed in WinBugs
#-----------------------------------------------------------------------------------------------------------

model
{
   for (i in 1:n.ind)
   {
      for (j in 1:n.grid)
      {   C[i, j] ~ dnorm(model[i, j], tau)
         model[i,j] <- Dose[i]*ka[i]*(exp(-cl[i]/vol[i]*Time[i,j] ) - exp(-  ka[i]*Time[i,j] ) )
                     / ( vol[i]*ka[i] - cl[i] )
      }
      theta[i, 1:p] ~ dmnorm(mu[1:p], omega.inv[1:p, 1:p])
      ka[i]     <-  exp(theta[i,1])
      cl[i]     <-  exp(theta[i,2])
      vol[i]    <-  exp(theta[i,3])
   }
   tau ~ dgamma(0.001, 0.001)
   omega.inv[1:p, 1:p] ~ dwish(omega.inv.matrix[1:p, 1:p], omega.inv.dof)
   mu[1:p] ~ dmnorm(mu.prior.mean[1:p], mu.prior.prec[1:p, 1:p])
   omega[1:p, 1:p] <- inverse(omega.inv[1:p, 1:p])
   log.ka.mean   <- mu[1]
   log.cl.mean   <- mu[2]
   log.vol.mean  <- mu[3]
   sigma         <- 1/sqrt(tau)

   for( j in 1 : n.grid )
   { for( i in 1 : n.ind )
     { Missing[i , j] ~ dnorm( 0.0,1.0E-6) }
   }
}
```

Fig. 2. MCMC model specification for WinBugs program.

because of increased computing power and the availability of recently developed inference algorithms based on simulation. The most popular family of such algorithms is MCMC, which includes Gibbs sampling. Gibbs sampling requires knowledge about the full conditional densities (FCD) of all of the unknown parameters in the model. WinBUGS is a fully extensible modular framework for constructing and analysing Bayesian full probability models (Lunn (2000)). An introduction to WinBugs is given by Spiegelhalter (2003) and Lunn (2000) including references to Bayesian theory and its implementation using Markov chain Monte Carlo (MCMC) algorithms. Figure 2 shows the programing code for the MCMC model definition used to perform the missing data simulation experiments in the WinBugs program.

Cross validation has been used to obtain the prediction error and the mean square error was used as a statistical measure of error.

To cross validate, the drug concentration values for periods of time longer than 1 hour were removed from the data for one of the individuals in the original dataset to generate the incomplete record. The values thus removed were then predicted by the EBA and MCMC methods in order to estimate the non-linear mixed model. Thus the models always considered 11 patients in the initial matrix with complete sets of observations and 1 record corresponding to a patient with missing observations. Overall, the values of theophylline

concentration in blood serum corresponding to times over one hour after the drug administration were removed to a diferent patient in each simulation experiment. A total of 91 missing values were predicted which correspond to 12 simulation experiments to predict the missing values. Then the differences between the predicted and the real values were calculated.

2.1 Statistical validation of the prediction method

The mean square error (ε) between the predicted (calculated) and the observed (real) theophylline concentrations was employed as a measure of accuracy for each method. The ε was defined as below (3), and compared for the MCMC and EBA methods.

$$\varepsilon = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n} \tag{3}$$

In this formula, ε is the prediction mean square error, n is the number of predicted values used to calculate the error (n=91, corresponding to a total of 12 predictive curves with a different number of predicted values for each patient), $\hat{y}$ are the predicted theophylline concentrations, and y are the real values.

3 Results and discussion

The raw data for the simulations using EBA obtained by SAS and the results obtained with WinBUGS are not included is this document, although the file is available on request. The simulations by MCMC were generated with 1,000 iterations for burn-in followed by 50,000 iterations. Table 1 summarized the results of the prediction experiments.

			Range	**Median**
Method	**Mean** (ε)	**Sdv** (ε)	(min, max (ε))	(ε)
EBA	2.7627991	4.9218293	0.0000468, 27.18000	0.639538
MCMC	2.7197261	4.7150696	0.001369, 27.1441	0.840889

Table 1. Summary of results for the prediction experiments.

The ε value obtained with MCMC is slightly lower than that obtained using EBA. Likewise, the standard deviation (sdv) of ε obtained by MCMC is slightly lower than that obtained through EBA. The range of ε is wider for the calculations obtained with the EBA than the range obtained with the MCMC calculations. Nevertheless, the precision of the results is affected by

the number of iterations used in the calculation. It may therefore be the case that MCMC simulations using fewer iterations would generate results more similar to those produced using the EBA method. A t-test was performed to compare the results statistically. Overall, the results of the experiments show no statistically significant differences between those two methods (p-value = 0.95).

4 Conclusion

We aimed to find a non-addressed method; a method where no distributions are pre-determined or assumed in the model, and whose parameters are defined as simply as possible. Our results indicate that the MCMC method offers greater precision and accuracy than EBA. However, in seeking to define which of the two methods is better, it is necessary to take into consideration certain aspects such as the programming processes needed (SAS and WinBUGS) to model and predict the PK curves, as well as the time each routine takes to run the models. Overall, the SAS programming for EBA modeling is easier, while the MCMC calculations need to be worked out in more detail. On the other hand, the MCMC method requires no prior assumption of normal parameters, as EBA does, and this is an important advantage when there is no known fixed parameter structure.

Although these are preliminary and incomplete results, they seem to offer some hope that the MCMC method can predict a new kinetics record for a new case from little information, which could prove very useful in clinical pharmacology. In the near-future we will be performing further analyses to compare the results of the EBA and MCMC models against the results from others methodologies used for the predictions of time series and also considering evaluations with additional data sets.

In general, this type of prediction, modeling and simulation of clinical trials can help to improve understanding of clinical trials and thereby optimize their design. The study falls within the so-called simulation based on the effect of the drug.

References

LUNN, D.J., THOMAS, A., BEST, N. and SPIEGELHALTER, D. (2000): WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing 10, 325-337.*

MONLEÓN, T., OCAÑA, J., ARNAIZ, J.A., CARNE, X., RIBA, N. and SOY, D. (2005): Modelización, simulación y validación de un ensayo clínico de Fase I. IX *Conferencia Española de Biometría (Sociedad Española de Biometría), Oviedo.*

MORRIS C.N. (1983): Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc. 78, 47-55.*

PINHEIRO, J.C. and BATES, D.M. (1995): Approximations to the log-likelihood function in the nonlinear mixed-effect model. *Journal of Computational and Graphical Statistics, 412-435.*

SMITH MK. (2004): Software for non-linear mixed effects modelling. *RSS meeting, London, 12 May 2004. http://www.rss.org.uk/PDF/mikesmith.pdf*

SPIEGELHALTER, D., THOMAS, A. and BEST, N. (2003): WinBUGS version 1.4 user manual. *Cambridge, MA: MRC Biostatistics Unit.*

TEMPELMAN, R.J. (1998): Generalized Linear Mixed Models in Dairy Cattle Breeding. *J Dairy Sci 81, pp. 1428-1444.*

WOLFINGER, R.D. (1997): Comment: Experiences with the SAS Macro NLINMIX. *Statistics in Medicine, 16, 1258-1259.*

Outlier Detection to Hierarchical and Mixed Effects Models

Miriam Daniele and Antonella Plaia

Department of Statistics, University of Palermo
Miriam Daniele and Antonella Plaia, viale delle Scienze - ed. 13, 90128 Palermo, *mdaniele@dssm.unipa.it*, *plaia@unipa.it*

Abstract. Hierarchical and mixed effects models are models where a varying number of coefficients may be random at different levels of the hierarchy. The purpose of outlier analysis for these models is to determine whether an outlying unit at higher level is entirely outlying, or outlying due to effect of one or a few aberrant lower level units. Most works on diagnostics for these complex models have focused on the mixed model rather than on the hierarchical models, obscuring some relevant aspects of the hierarchical model. In this paper we will present an approach to influence analysis and outlier detection for mixed and hierarchical model, focusing on the special structure of nested data that these models describe.

Keywords: mixed effect models, hierarchical models, outliers, influence diagnostics

1 Introduction

Hierarchical models are useful tools in describing relationships between a response variable and covariates, in data that are grouped according to some classification factors. These models encompass two sources of variation, namely within and among individuals in the population; this naturally leads to the possibility of contaminated data at each sampling level. We refer to such data as "outliers". The term outlier has been used by different researchers with different meanings, and there is not still a very clear and univocal definition in statistical literature. Hawkins (1980) defines an outlier to be "an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism".

The purpose of a multilevel analysis of outliers is to determine whether an outlying unit at higher level is entirely outlying, or outlying due to effect of one or a few aberrant lower level units. In fact, in data structure of increasing complexity the concept of outlier becomes less clear rather than for ordinary regression. The researcher has to consider at what level(s) a particular response is outlying, and in respect of which explanatory variable(s). Furthermore, the treatment of a particular response at a level may affect its status or the status of other units at other levels in the model (Langford and Lewis (1998)). In more recent years, Zewotir and Galpin (2005) provide

routine diagnostic tools, which are computationally inexpensive. The goal of their paper is to extend the ordinary linear regression influence diagnostic approach to linear mixed models. Zewotir and Galpin (2007) proposed a unified approach to residuals, leverages and outliers in the linear mixed model. Nevertheless, all these diagnostic tools are focused on mixed model rather than on the hierarchical models, obscuring some relevant aspects of the hierarchical model. In this paper we propose an approach to outlier and high-leverage point detection, addressed both to mixed and hierarchical models, but focusing on the last ones.

2 Mixed and hierarchical linear models

Linear mixed effect models (LME) are models in which both the fixed and the random effects occur linearly in the model function. They extend linear model by incorporating random effects, which can be regarded as additional error terms, to account for correlation among observations within the same group. We refer to the formulation of LME model, described in detail by Demidenko (2004), and used by Zewotir and Galpin (2005, 2006 and 2007) for diagnostic purposes. The general form of this LME model for the j-th subject (or cluster) is

$$\mathbf{y}_j = X_j\boldsymbol{\beta} + Z_j\boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j, \qquad j = 1, \ldots, N, \tag{1}$$

where: $\mathbf{y}_j$ is a $(n_j \times 1)$ vector of responses of the j-th subject, also called individual or cluster, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown constants, the fixed effects of the model; X_j is a $(n_j \times p)$ known design matrix associated with $\boldsymbol{\beta}$, $\boldsymbol{\delta}_j$ is a $(q \times 1)$ vector of random effects from $N(\mathbf{0}, D_j)$ associated with the j-th subject or cluster, Z_j is an $(n_j \times q)$ known design matrix of the random effects, and $\boldsymbol{\varepsilon}_j$ is a $(n_j \times 1)$ vector of error term from $N(\mathbf{0}, \sigma_e^2 I)$, with $\boldsymbol{\delta}_j$ and $\boldsymbol{\varepsilon}_j$ mutually independent. Considering all the N subjects we can write (1) in a matrix form

$$\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{2}$$

where: $\mathbf{y}$ is a $(M \times 1)$ vector, with $M = \sum_{j=1}^{N} \sum_{i=1}^{n_j} n_{ij}$, X is a $(M \times p)$ known design matrix associated with the fixed effects, $\boldsymbol{\delta}$ is a $(N \times 1)$ vector of random effects, Z is an $(M \times N)$ known design matrix of the random effects, and $\boldsymbol{\varepsilon}$ is a $(M \times 1)$ vector of error terms from $N(\mathbf{0}, \sigma_e^2 I)$. The vector $\boldsymbol{\delta} \sim N(\mathbf{0}, D)$, where D in general is an unstructured matrix, usually is a diagonal block matrix

$$D_{(N\times N)} = \begin{bmatrix} D_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & D_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & D_N \end{bmatrix}, \tag{3}$$

and the response vector $\mathbf{y} \sim N(X\boldsymbol{\beta}, V)$, with

$$V_{(M\times M)} = \begin{bmatrix} \Sigma_1 + Z_1 D_1 Z_1' & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_2 + Z_2 D_2 Z_2' & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma_N + Z_N D_N Z_N' \end{bmatrix}. \tag{4}$$

The Hierarchical Linear Model (HLM) can be view as a special case of a LME model. Suppose we have M observations naturally grouped in N clusters, with n_j observations in the j-th group and $\sum_{j=1}^{N} \sum_{i=1}^{n_j} n_{ij} = M$. In the first level the model for the j-th cluster is

$$\mathbf{y}_j = X_j \gamma_j + \boldsymbol{\varepsilon}_j \tag{5}$$

where X_j has dimension $(n_j \times q)$ and $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \sigma_e^2 I)$, and $\boldsymbol{\varepsilon}_j$ and $\boldsymbol{\varepsilon}_{j'}$ are independent for $j \neq j'$. At the next level, we would like to model variation of $\boldsymbol{\gamma}_j$ from a cluster to another, that is we want a between-units model which will link the within-cluster models. Suppose we have some background of information on the j-th group, which we can summarize in a matrix W_j, so that we obtain the between group model

$$\boldsymbol{\gamma}_j = W_j \boldsymbol{\beta} + \boldsymbol{\delta}_j \tag{6}$$

where $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown constants, the fixed effects of the model, and $\boldsymbol{\delta}_j = [\boldsymbol{\delta}_{j1}, \boldsymbol{\delta}_{j2}, \dots, \boldsymbol{\delta}_{jq}]'$ with $\boldsymbol{\delta}_j \sim N(\mathbf{0}, D_j)$, and the matrix W_j has dimension $(q \times p)$. Combining the equations (5) and (6) we get the model

$$\mathbf{y}_j = X_j W_j \boldsymbol{\beta} + X_j \boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j. \tag{7}$$

This equation suggests that hierarchical models are a special case of a mixed model (1), by setting $X_j W_j = X_j$ and $X_j = Z_j$.

The fitted values of the response variable $\mathbf{y}$, in the model (2), are $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} + Z\tilde{\boldsymbol{\delta}}$, giving the residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. If the parameter of the V matrix are known then $\mathbf{e} \sim N(\mathbf{0}, R)$, and correlation between e_j and e_j' is entirely determined by the elements of R,

$$R_{(M\times M)} = V^{-1} - V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1}. \tag{8}$$

The symmetric matrix R is the projection matrix for the residual vector $\mathbf{e}$, namely $R\mathbf{y} = \mathbf{e}$. This matrix plays an important role in the analysis of residuals, especially in order to build formal and informal diagnostic tools to detect outliers and high leverage points.

3 Residuals and outliers in mixed and hierarchical linear model

In a mixed effect model there are multiple sources of error, then we may define a distinct residual for each of this source. According to Hilden-Milton (1995) there are three types of errors for model (2):

(i) case-unique or conditional error, $\boldsymbol{\varepsilon}$;
(ii) random effects, $\boldsymbol{\delta}$;
(iii) composite or marginal error, $\boldsymbol{\zeta} = Z\boldsymbol{\delta} + \boldsymbol{\varepsilon}$.

Each type of these disturbances accounts for different parts of the model, then is of diagnostic interest. As a consequence, we have three different residuals corresponding to the three types of errors:

(i) the case-unique residuals $\mathbf{e} = \mathbf{y} - X\hat{\boldsymbol{\beta}} - Z\tilde{\boldsymbol{\delta}}$;
(ii) the random effect residuals $\boldsymbol{d} = Z\tilde{\boldsymbol{\delta}}$;
(iii) the composite residuals $\boldsymbol{g} = \boldsymbol{e} + \boldsymbol{d} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$.

In order to build formal and informal procedures to detect outliers and influential observations, we have to consider these residuals and their respective projection matrices, namely that matrix which transforms the vector of observations $\mathbf{y}$ into residuals. Given the model (2) above described, it is possible to demonstrate the following results:

$$\Sigma R\mathbf{y} = P_e\mathbf{y} = \mathbf{e} \tag{9}$$
$$ZDZ'R\mathbf{y} = P_d\mathbf{y} = \mathbf{d} \tag{10}$$
$$VR\mathbf{y} = P_g\mathbf{y} = \mathbf{g}, \tag{11}$$

where R is the matrix considered in formula (8). Furthermore, known D and Σ, we can easily obtain the variances of three residuals:

$$\mathrm{Var}(\mathbf{e}) = \Sigma R\,\mathrm{Var}(\mathbf{y})\,(\Sigma R)' = \Sigma R\Sigma \tag{12}$$
$$\mathrm{Var}(\mathbf{d}) = ZDZ'R\,\mathrm{Var}(\mathbf{y})\,(ZDZ'R)' = ZDZ'RZDZ' \tag{13}$$
$$\mathrm{Var}(\mathbf{g}) = VR\,\mathrm{Var}(\mathbf{y})\,(VR)' = VRV, \tag{14}$$

and from these the Studentized residuals. Zewotir and Galpin (2005) proposed several influence measures based on R, V^{-1} and the Studentized residuals, and successively Zewotir and Galpin (2006) tested them by simulation studies. The proposal of this paper is addressed both to mixed and hierarchical models, but focuses on the last models. By different simulation studies, we noticed that only the joint treatment of all the three type of residuals and their projection matrices can highlight different and complementary aspects of the model at different levels of the hierarchy.

4 Influence measures

Influence measures aim at determining whether some observations have undue influence on the model parameter estimates. The influence measures considered in this section, suggested by Zewotir and Galpin (2005) for the only case-unique residuals, are based on the case-deletion approach, and are summarized in Table 1.

On	Name	Formula
$\hat{\boldsymbol{\beta}}$	Cook's distance	$\text{CD}_i(\boldsymbol{\beta}) = \frac{(c_{ii}-r_{ii})t^2_{e_i}}{r_{ii}p}$
	Variance Ratio	$\text{COVRATIO}_i(\boldsymbol{\beta}) = \left(\frac{M-t^2_{e_i}}{M-1}\right)^p \frac{c_{ii}}{r_{ii}}$
	Cook-Weisberg stat.	$\text{CW}_i = \frac{1}{2}\log\left(\frac{c_{ii}}{r_{ii}}\right) + \frac{p}{2}\log\left(\frac{M}{M-1} - \frac{t^2_{e_i}}{M-1}\right)$
$\tilde{\boldsymbol{\delta}}$	Cook's distance	$\text{CD}_i(\boldsymbol{\delta}) = t^2_{e_i}\,(1 - ssq(R_i))$
$\ell(\hat{\boldsymbol{\theta}})$	Likelihood distance	$\text{LD}_i = M\log\left(\frac{M-t^2_{e_i}}{M-1}\right) + \frac{[M+t^2_{e_i}(\frac{c_{ii}}{r_{ii}}-1)](M-1)}{(M-t^2_{e_i})/(M-1)} - M$

Table 1. Summary of the influence measures based on the *case-unique* residuals

In Table 1, c_{ii} and r_{ii} are the main diagonal elements of the matrices $C = V^{-1}$ and R, $t^2_{e_i}$ is the i-th case-unique Studentized residual, and ssq(R_i) is the sum of the squares of elements of R_i. The first three measures express, in different ways, the effect of the i-th observation deletion on the fixed effect parameter estimates ($\hat{\boldsymbol{\beta}}$), the others the effect of the i-th observation deletion on the random effect parameter estimates ($\tilde{\boldsymbol{\delta}}$) and on the likelihood distance respectively. We extended all these diagnostic tools to the other two types of residuals, as tables 2 and 3 show.

Setting $V_z = ZDZ'$ and indicating with pd_{ii} and pg_{ii} the main diagonal elements of the projection matrices P_d and P_g respectively, and with t_{d_i} and t_{g_i} the Studentized version of residuals, we obtain the measures in Tables 2 and 3 respectively.

5 Simulation studies and results

The various influence measures summarized in tables 1, 2, and 3 allow to evaluate the impact of single cases or group of cases on some characteristics of the fitted model. Points that stand out from the other data points with

On	Name	Formula
$\hat{\boldsymbol{\beta}}$	Cook's distance	$\text{CD}_i(\boldsymbol{\beta}) = \frac{(V_{z_{ii}} c_{ii} - p d_{ii}) t^2_{d_i}}{p d_{ii} p}$
	Variance ratio	$\text{COVRATIO}_i(\boldsymbol{\beta}) = \left(\frac{M - t^2_{d_i}}{M-1}\right)^p \frac{V_{z_{ii}} c_{ii}}{p d_{ii}}$
	Cook-Weisberg stat.	$\text{CW}_i = \frac{1}{2}\log\left(\frac{V_{z_{ii}} c_{ii}}{p d_{ii}}\right) + \frac{p}{2}\log\left(\frac{M}{M-1} - \frac{t^2_{d_i}}{M-1}\right)$
$\tilde{\boldsymbol{\delta}}$	Cook's distance	$\text{CD}_i(\boldsymbol{\delta}) = t^2_{d_i}\,(1 - ssq(Pd_i))$
$\ell(\hat{\boldsymbol{\theta}})$	Likelihood distance	$\text{LD}_i = M\log\left(\frac{M - t^2_{d_i}}{M-1}\right) + \frac{[M + t^2_{d_i}(V_{z_{ii}} c_{ii} - p d_{ii})](M-1)}{p d_{ii}(M - t^2_{d_i})/(M-1)} - M$

Table 2. Summary of the influence measures based on the *random effects* residuals

On	Name	Formula
$\hat{\boldsymbol{\beta}}$	Cook's distance	$\text{CD}_i(\boldsymbol{\beta}) = \frac{(1 - p g_{ii}) t^2_{g_i}}{p g_{ii} p}$
	Variance ratio	$\text{COVRATIO}_i(\boldsymbol{\beta}) = \left(\frac{M - t^2_{g_i}}{M-1}\right)^p \frac{1}{p g_{ii}}$
	Cook-Weisberg stat.	$\text{CW}_i = \frac{1}{2}\log\left(\frac{1}{p g_{ii}}\right) + \frac{p}{2}\log\left(\frac{M}{M-1} - \frac{t^2_{g_i}}{M-1}\right)$
$\tilde{\boldsymbol{\delta}}$	Cook's distance	$\text{CD}_i(\boldsymbol{\delta}) = t^2_{g_i}\,(1 - ssq(Pg_i))$
$\ell(\hat{\boldsymbol{\theta}})$	Likelihood distance	$\text{LD}_i = M\log\left(\frac{M - t^2_{g_i}}{M-1}\right) + \frac{[M + t^2_{g_i}(1 - p g_{ii})](M-1)}{p g_{ii}(M - t^2_{g_i})/(M-1)} - M$

Table 3. Summary of the influence measures based on the *composite* residuals

respect to their diagnostic measures should be further examined. The simulation study that we carried out consists in generating "clean" data sets, then in introducing some aberrant cases or groups of cases, finally in seeing if the influence measures detect them.

In particular, we generated $m = 1000$ clean data sets from model (2), considering a model with one fixed effect and random intercept and slope, $y_{ij} = \beta_0 + \beta_1 x_{ij} + \delta_{0j} + \delta_{1j} z_{ij} + \varepsilon_{ij}$. Each data set has $M = 200$ cases, nested in $N = 20$ balanced groups, namely $n_j = 10$ $(j = 1, \ldots, N)$. The main steps of the performed simulation study can be summarized as follows:

(i) we generated $\mathbf{x} \sim U(0,1)$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, I)$, and $\boldsymbol{\delta} \sim N(\mathbf{0}, I)$,

(ii) we computed $\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\delta} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\beta} = \mathbf{1}$;
(iii) we estimated the parameters of the model;
(iv) we dirtied the generated data sets by introducing abnormal cases or groups of cases;
(v) we evaluated the sensitivity of the proposed influence measures on these dirty data sets for each type of residual.

For each data set generated, we compared graphically the influence measure computed on the clean data set with the same measure on the dirty data set, for the three residuals simultaneously, in order to underline analogies and differences. To evaluate the sensitivity of the proposed measures to highlight the anomalies introduced in the simulation studies, we decided to express their performance as percentage of successes, where successes are the number of times that the specific measure detects the introduced anomaly.

Nevertheless, there are not cutoffs for these influence measures, to establish when a case is outlier or is an high-leverage point. Then, we thought to use as term of comparison, some summary values of the influence measures computed on all the 1000 clean data sets. We have to distinguish between the introduction of single outliers in X, in Z or in $\mathbf{y}$ separately (z_{11}, y_{11} and x_{11}), or simultaneously (z_{11}, y_{1010} and x_{1020}), and the introduction of a cluster of outliers. Table 1 shows the results relative to the Cook's distance for the fixed parameter estimates, but the simulation considered all the influence measures in Tables 1, 2, and 3.

In the above Table, column $CD_e(\boldsymbol{\beta})$ refers to the influence measure already known in literature, while $CD_d(\boldsymbol{\beta})$ and $CD_g(\boldsymbol{\beta})$ refer to the influence measures for the random effect and the composite residuals introduced in this paper. The results obtained on the 1000 data sets have been summarized according to:

- "single-case outlier": we compute the third quartile (3Qu) on the 1000 clean data sets for each influence measure, and we compare it with the value of the influence measure corresponding the dirty data set. For example, introducing an anomaly in X (x_{11}) we consider:

$$Q = (CD_e(\boldsymbol{\beta})_{11}^{dirty} - 3\mathrm{Qu}(CD_e(\boldsymbol{\beta})_1^{clean}))/CD_e(\boldsymbol{\beta})_{11}^{dirty}. \qquad (15)$$

  Q assumes values in the range $[0, 1]$, where values close to 1 indicate that the influence measure detects the presence of an outlier. Actually, as we compare $CD_e(\boldsymbol{\beta})_{11}^{dirty}$ with the third quartile (and not a mean or a median), values of Q greater than 0.5 (the value considered to compute the Table 4) can be considered high values.
- "single-cluster outlier": we compare, by the same relative difference, the mean of the group means on 1000 clean data sets with the mean of $CD_e(\boldsymbol{\beta})$ for the anomalous cluster in the dirty data set.

Results in Table 4 show the importance of computing the three measures, as they highlight different types of anomalies in the data. It is well known

Table 4. Percentage of the successes for the *Cook's distance* on the fixed effect parameter estimates

Anomaly in	Cases or groups	$CD_e(\boldsymbol{\beta})$	$CD_d(\boldsymbol{\beta})$	$CD_g(\boldsymbol{\beta})$
X	x_{11}	79.2	47.8	82.3
$\mathbf{y}$	y_{11}	83.9	51.9	99
Z	z_{11}	3.2	78.7	67.4
X; Z	x_{1020}; z_{11}	28.5	32.5	19.5
	x_{1020}	77.7	50.9	81.2
	z_{11}	36.8	63.7	22.8
X; Z; $\mathbf{y}$	x_{1020}; z_{11}; $\mathbf{y}_{1010}$	1	2.9	58.3
	x_{1020}	79.2	17.9	82.7
	z_{11}	12.2	65.2	70.9
	y_{1010}	40.7	17.9	99
X	$\mathbf{x}_1$	90	32.5	77.3
Z	$\mathbf{z}_1$	51.7	76.4	46.3
$\mathbf{y}$	$\mathbf{y}_1$	0.8	98.8	14.9
X;Z; $\mathbf{y}$	$\mathbf{x}_{20}$; $\mathbf{z}_1$; $\mathbf{y}_{11}$	1.6	18.1	5.6
	$\mathbf{x}_{20}$	98.5	26.9	81.1
	$\mathbf{z}_1$	26.8	73.1	12.1
	$\mathbf{y}_{11}$	4	96.4	65

in literature that the simultaneous introduction of outliers in X, in Z, and in $\mathbf{y}$ may often cause the *masking effect*. The case-unique residual diagnostic measures ($CD_e(\boldsymbol{\beta})$) show this problem, resulting in very low percentages of success. Moreover, the Cook's distance computed on the random effect residuals ($CD_d(\boldsymbol{\beta})$) and on the composite residuals ($CD_g(\boldsymbol{\beta})$) perform better than $CD_e(\boldsymbol{\beta})$, when we introduce anomalous cases in Z, and when we introduce cluster of outliers, as random effect residuals can be considered as Level-2 residuals.

Concluding, the proposed diagnostic tools, used simultaneously, allow to perform both a Level-1 and a Level-2 influence analysis, and are particularly appropriate to hierarchical models. The joint evaluation of the measures for all the three types of residuals allow to reduce the problem of masking effect.

References

DEMIDENKO, E. (2004): *Mixed Models: Theory and Applications*. Wiley.
HAWKINS, D.M. (1980): *Identification of Outliers*. Chapman and Hall, London.

HILDEN-MINTON, J.A. (1995): Multilevel Diagnostics for Mixed and Hierarchical Linear models. University of California.

LANGFORD, I.H. and LEWIS, T. (1998): Outliers in multilevel data. *J. Roy. Statist. Soc. Ser. A* 161, 121–160.

ZEWOTIR, T. and GALPIN, J.S. (2005): Influence Diagnostics for Linear Mixed Models. *Journal of Data Science* 3, 153–177.

ZEWOTIR, T. and GALPIN, J.S. (2006): Evaluation of Linear Mixed Model Case Deletion Diagnostic Tools by Monte Carlo Simulation. *Comm. Statist. Simulation Comput.* 35, 645–682.

ZEWOTIR, T. and GALPIN, J.S. (2007): A unified approach on Residuals, Leverages and Outliers in the Linear Mixed Model. *Comm. Statist. Simulation Comput.* 16, 58–75.

On the Multivariate Goodness-of-Fit Test

Grzegorz Konczak

Karol Adamiecki University of Economics
Dr Grzegorz Konczak, 40-226 Katowice, Bogucicka 14, Poland,
koncz@ae.katowice.pl

Abstract. In the paper the proposal of the nonparametric test to verify the hypothesis on the form of the distribution of the multivariate random variable is presented. The main idea of the proposed test is based on the well known empty cells test. In the empty cells test the area of variability of the random variable is divided into some fixed cells. In the proposed modification there is one cell only but the cell is moving over the whole area of variability of the random variable. The area of empty cells test is determined.

The analysis of testing the hypothesis that the random variable has a multivariate uniform distribution is presented. The table with critical values of the test statistic for the case of two-dimensional random variable is presented.

Keywords: multivariate test, empty cells test, Monte Carlo

1 Introduction

One of the goodness-of-fit tests is the empty cells test (David F.N. 1950, Hellwig Z. 1965). This test can be used to test the hypothesis of the distribution of random variable. The area of variability of random variable is divided into m cells and the number of elements from the sample in each cell is counted. Then the number of empty cells is determined. This number of empty cells is compared to the critical value. The interpolated critical values for empty cells statistic were presented by Domanski Cz. and Pruska K. (2000). Konczak G. (2005) has proposed the modification of the empty cells test which takes into account the number of cells with k elements ($k = 0, 1, ..., m - 1$).

The proposal of the multivariate goodness-of-fit test is presented in the paper. The hypothesis about the form of distribution of multivariate random variable $\mathbf{X}$ will be tested. The idea of this proposal is based on the empty cells test. Let S denotes the area of variability of the random variable $\mathbf{X}$. In the proposed multivariate test there is one cell only which is moving over the whole area S. The analysis in the case of verifying the hypothesis that random variable has a multivariate uniform distribution is presented. The table with critical values of the test statistic for two-dimensional case is presented. The results of the computer simulation of the power of the proposed test were presented.

2 Basic notations and assumptions

Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ be the continuous vector random variable and let $F(\mathbf{x})$ be the cumulative distribution function of this random variable. Let us assume that $X_1, X_2, \ldots, X_k$ are independent random variables with cumulative distributions $F_1(x), F_2(x), \ldots, F_k(x)$. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ for $i = 1, 2, \ldots, n$, be an n-element simple sample. We will test the hypothesis H_0 that the sample is taken from the $F_0(\mathbf{x})$ distribution.

The hypothesis can be written as follows

$$\begin{aligned} H_0 &: F(\mathbf{x}) = F_0(\mathbf{x}) \\ H_1 &: F(\mathbf{x}) \neq F_0(\mathbf{x}) \end{aligned} \tag{1}$$

Let $q_{i,\alpha}$ denotes a quantil of order α of the random variable X_i, for $i = 1, 2, \ldots, k$.

For each $\mathbf{x} = (x_1, x_2, ..., x_k) \in S^*$ where

$$S^* = \left[q_{1,\frac{\pi}{2}}; q_{1,1-\frac{\pi}{2}}\right] \times \left[q_{2,\frac{\pi}{2}}; q_{2,1-\frac{\pi}{2}}\right] \times \cdots \times \left[q_{k,\frac{\pi}{2}}; q_{k,1-\frac{\pi}{2}}\right]$$

the cell $S_{\mathbf{x}}$ can be written

$$S_{\mathbf{x}} = \left[q_{1,\beta_1-\frac{\pi}{2}}; q_{1,\beta_1+\frac{\pi}{2}}\right] \times \left[q_{2,\beta_2-\frac{\pi}{2}}; q_{2,\beta_2+\frac{\pi}{2}}\right] \times \cdots \times \left[q_{k,\beta_k-\frac{\pi}{2}}; q_{k,\beta_k+\frac{\pi}{2}}\right] \tag{2}$$

where $\beta_i (\frac{\pi}{2} \leq \beta_i \leq 1 - \frac{\pi}{2})$ is given as follows

$$\beta_i = \begin{cases} \frac{\pi}{2} & x_i < F_i^{-1}(\frac{\pi}{2}) \\ F_i(x_i) & x_i \in \left[F_i^{-1}(\frac{\pi}{2}); F_i^{-1}(1-\frac{\pi}{2})\right] \\ 1 - \frac{\pi}{2} & x_i > F_i^{-1}(1-\frac{\pi}{2}) \end{cases} \tag{3}$$

for $i = 1, 2, \ldots, n$.

The probability p that $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$ is in the cell $S_{\mathbf{x}}$ under H_0 is constant and can be written as follows

$$P(\mathbf{x}_i \in S_{\mathbf{x}}) = \pi \cdot \pi \cdot ... \cdot \pi = \pi^k$$

It can be rewritten as follows

$$P(\mathbf{x}_i \in S_{\mathbf{x}}) = \sqrt[k]{p} \cdot \sqrt[k]{p} \cdot ... \cdot \sqrt[k]{p} = p$$

In this paper it will be considered the case where $p = \frac{1}{n}$.

The idea of the construction of the cell $S_{\mathbf{x}}$ in the two-dimensional uniform random variable case is presented in Figure 1. There is one cell $S_{\mathbf{x}}$ which moves over the whole area of $[a, b] \times [c, d]$. The point $\mathbf{x} = (x_1, x_2)$ is the mid-point of the cell $S\mathbf{x}$.

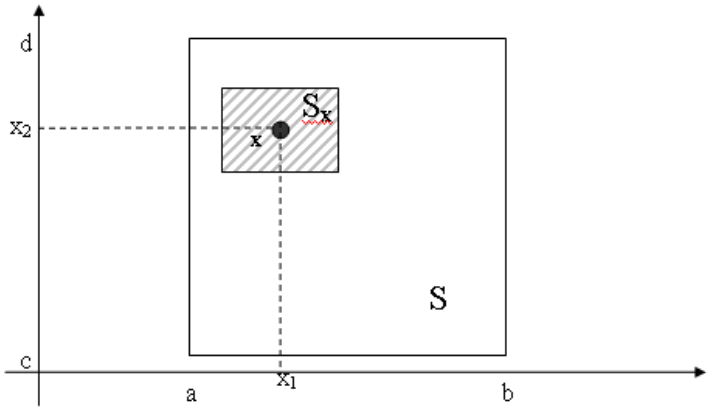


Fig. 1. The idea of the proposed test (the uniform random variable case).

3 Construction of the test statistics

Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ be a k-dimensional random variable vector. Let the random variables $X_1, X_2, \ldots, X_k$ are independent (in general case the random variables $X_1, X_2, \ldots, X_k$ can be dependent but the construction of the cells $S_{\mathbf{x}}$ in this case is much more difficult). Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be an i.i.d. sample. The hypothesis (1) that the sample is taken from F_0 distribution will be tested. For each

$$\mathbf{x} \in \left[q_{1,\frac{\pi}{2}}; q_{1,1-\frac{\pi}{2}}\right] \times \left[q_{2,\frac{\pi}{2}}; q_{2,1-\frac{\pi}{2}}\right] \times \cdots \times \left[q_{k,\frac{\pi}{2}}; q_{k,1-\frac{\pi}{2}}\right]$$

under H_0 we have

$$P(\mathbf{x}_i \in S_{\mathbf{x}}) = p = const$$

for $i = 1, 2, \ldots, n$.

Let us consider the number of elements from n element sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ which are in the cell $S_{\mathbf{x}}$. The number of elements in the cell $S_{\mathbf{x}}$ can be written as follows

$$card\{S_{\mathbf{x}}\} = card\{j : \mathbf{x}_j \in S_{\mathbf{x}}, j = 1, 2, \ldots, n\} \tag{4}$$

The probability that the cell $S_{\mathbf{x}}$ is empty can be written as follows:

$$P(card\{S_{\mathbf{x}}\} = 0) = P((\mathbf{x}_1 \notin S_{\mathbf{x}}) \wedge (\mathbf{x}_2 \notin S_{\mathbf{x}}) \wedge \cdots \wedge (\mathbf{x}_k \notin S_{\mathbf{x}})) \tag{5}$$

Under the assumption that $X_1, X_2, \ldots, X_n$ are independent it can be written as follows

$$\begin{gathered} P(card\{S_{\mathbf{x}}\} = 0) = P(\mathbf{x}_1 \notin S_{\mathbf{x}}) \cdot P(\mathbf{x}_2 \notin S_{\mathbf{x}}) \cdot \ldots \cdot P(\mathbf{x}_n \notin S_{\mathbf{x}}) = \\ = (1-p) \cdot (1-p) \cdot \ldots \cdot (1-p) = (1-p)^n \end{gathered} \tag{6}$$

Let us consider the function

$$h : \left[q_{1,\frac{\pi}{2}}; q_{1,1-\frac{\pi}{2}}\right] \times \left[q_{2,\frac{\pi}{2}}; q_{2,1-\frac{\pi}{2}}\right] \times \cdots \times \left[q_{k,\frac{\pi}{2}}; q_{k,1-\frac{\pi}{2}}\right] \to \{0,1\}$$

given as follows

$$h(\mathbf{x}) = \begin{cases} 0 & \text{if } card\{S_{\mathbf{x}}\} > 0 \\ 1 & \text{if } card\{S_{\mathbf{x}}\} = 0 \end{cases} \tag{7}$$

The formula (7) can be written equally as follows

$$h(\mathbf{x}) = \begin{cases} 0 & \text{if } \exists_i \quad x_i \in S_{\mathbf{x}} \\ 1 & \text{if } \forall_i \quad x_i \notin S_{\mathbf{x}} \end{cases} \tag{8}$$

The function takes the value 1 if and only if none of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is in the cell $S_{\mathbf{x}}$ area, that's means that if the cell $S_{\mathbf{x}}$ is empty. The formula (8) can be written

$$h(\mathbf{x}) = \begin{cases} 0 & \text{if } \exists_i \quad \mathbf{x}_i \in \left[q_{1,\frac{\pi}{2}}; q_{1,1-\frac{\pi}{2}}\right] \times \left[q_{2,\frac{\pi}{2}}; q_{2,1-\frac{\pi}{2}}\right] \times \cdots \times \left[q_{k,\frac{\pi}{2}}; q_{k,1-\frac{\pi}{2}}\right] \\ 1 & \text{if } \forall_i \quad \mathbf{x}_i \notin \left[q_{1,\frac{\pi}{2}}; q_{1,1-\frac{\pi}{2}}\right] \times \left[q_{2,\frac{\pi}{2}}; q_{2,1-\frac{\pi}{2}}\right] \times \cdots \times \left[q_{k,\frac{\pi}{2}}; q_{k,1-\frac{\pi}{2}}\right] \end{cases} \tag{9}$$

That's mean that the value $h(\mathbf{x})$ is equal to 1 if and only if the cell $S_{\mathbf{x}}$ is empty. Then we have

$$h(\mathbf{x}) = 1 \Leftrightarrow card\{S_{\mathbf{x}}\} = 0$$

It can be written as follows

$$P(h(\mathbf{x}) = 1) = P(card\{S_{\mathbf{x}}\} = 0) = (1-p)^n$$

for each $\mathbf{x} \in S^*$.

The idea of the function $h(\mathbf{x})$ in the two-dimensional case is presented in the Figure 2. The $n = 4$ element sample is taken. The value of the function $h(\mathbf{x})$ is equal to 1 (the empty cells area) if and only if $\mathbf{x}_i \notin S_{\mathbf{x}} (i = 1, 2, 3, 4)$ and $h(\mathbf{x}) = 0$ (the non-empty cells area) if and only if there exist i such $\mathbf{x}_i \in S_{\mathbf{x}}$ (see Figure 2).

To test the hypothesis (1) it can be used following statistic

$$T = \frac{1}{\prod_{i=1}^{k} \left(q_{i,1-\frac{\pi}{2}} - q_{i,\frac{\pi}{2}}\right)} \int_{\frac{\pi}{2}}^{1-\frac{\pi}{2}} \int_{\frac{\pi}{2}}^{1-\frac{\pi}{2}} \ldots \int_{\frac{\pi}{2}}^{1-\frac{\pi}{2}} h(\mathbf{x}) dx_k \ldots dx_2 dx_1 \tag{10}$$

The statistic T measures the relational area (in general case the relational volume) of the empty cells area. We reject the hypothesis H_0 if the value of this statistic is greater or equal to the critical value $(T \geq T_\alpha)$. It is difficult to find quantiles of the statistic T in general case. Tehe estimated values of these quantiles can be found using Monte Carlo study.

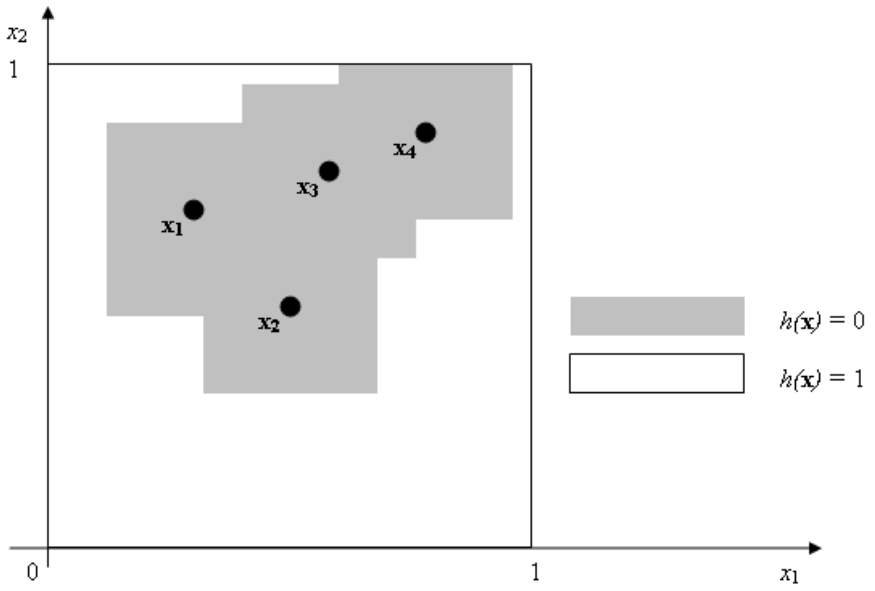


Fig. 2. Function $h(\mathbf{x})$ - the idea of the construction (two dimensional case).

4 Monte Carlo study - the case of two-dimensional uniform distribution

Let the random variable $\mathbf{X} = (X_1, X_2)$ has two-dimensional uniform distribution on the set $[0,1] \times [0,1]$. The density function $f(\mathbf{x})$ of this random variable can be written as follows:

$$f(\mathbf{x}) = \begin{cases} 0 & \text{if } x_1 \notin [0,1] \text{ or } x_2 \notin [0,1] \\ 1 & \text{if } x_1 \in [0,1] \text{ and } x_2 \in [0,1] \end{cases} \tag{11}$$

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be the n element i.i.d. sample from uniform distribution $U[0,1]^2$. Let $\pi = \frac{1}{\sqrt{n}}$ then in the two dimensional case we have $p = \frac{1}{n}$.

The set S^* of mid-points of the set cells $S_{\mathbf{x}}$ can be written as follows:

$$S^* = \left[q_{1,\frac{\sqrt{p}}{2}}; q_{1,1-\frac{\sqrt{p}}{2}}\right] \times \left[q_{2,\frac{\sqrt{p}}{2}}; q_{2,1-\frac{\sqrt{p}}{2}}\right]$$

To obtain the critical values for test the hypothesis that random variable $\mathbf{X}$ has two-dimensional uniform distribution the Monte Carlo simulation were made. For $k = 2$ and sample sizes of $n = 5, 6, \ldots, 20$ there were found quantiles of the statistic T. They were found for the significance levels $\alpha = 0.10, 0.05$ and 0.01. There were found the empirical quantiles of the statistic T and then they were accepted as the estimated values of the quantiles of the statistic T. There are four following steps in the Monte Carlo simulations:

(i) The vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n (n = 5, 6, \ldots, 20)$ were generated from two-dimensional uniform distribution U[0,1] x [0,1].
(ii) For each sample the value of the T statistic was calculated.
(iii) The steps 1-2 were repeated 10 000 times.
(iv) The empirical quantiles 0.90, 0.95 and 0.99 were accepted as estimates of quantiles of the statistic T.

Table 1. The estimates quantiles of the test statistic $T(k = 2)$.

Sample size n	significance level 0.10	0.05	0.01
5	0.7917	0.9714	—
6	0.7082	0.9142	—
7	0.6217	0.8043	—
8	0.6115	0.7762	0.9762
9	0.5471	0.6733	0.9842
10	0.5333	0.7025	0.9660
11	0.5251	0.6531	0.8941
12	0.4591	0.5965	0.8390
13	0.4661	0.5881	0.8180
14	0.4254	0.5334	0.7681
15	0.4083	0.5161	0.7027
16	0.3722	0.4652	0.6951
17	0.3620	0.4471	0.6751
18	0.3493	0.4312	0.6029
19	0.3254	0.4042	0.5821
20	0.3302	0.4241	0.6120

Source: Monte Carlo study

The results of Monte Carlo study are presented in Table 1. For sample size from 5 to 20 there are presented estimates quantiles of the statistic T. These quantiles can be used as a critical values for the proposed goodness-of-fit test in the two-dimensional case.

The quatiles can be used as critical values in the case of verification that the sample is taken from uniform distribution. In general case the critical values can be obtain by the same way.

In general case there is difficult to determine the exact value of the statistic T. Let us assume that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is the n element i.i.d. sample. We test

the hypothesis that the sample is from $F_0(\mathbf{x})$ distribution. The estimate value of the statistic T we can obtain with following six steps:

(i) We generate a vector $\mathbf{x}$ from $F_0(x)$ distribution.
(ii) We construct the cell $S_{\mathbf{x}}$ with mid-point $\mathbf{x}$.
(iii) We observe if the cell $S_{\mathbf{x}}$ is empty.
(iv) The steps 1-2 should be repeat 10 000 times.
(v) We calculate the fraction of cells which are empty.
(vi) The fraction we accept as an estimated value of statistic T.

The determined value of the statistic T is should be compared to the critical value from Table 1.

The power of the proposed test was analyzed using Monte Carlo study. Samples of size $n = 20$ were taken from two dimensional uniform distribution with density function

$$f(\mathbf{x}) = \begin{cases} 0 & \text{if } x_1 \notin [\mu_1 - 0.5, \mu_1 + 0.5] \text{ or } x_2 \notin [\mu_2 - 0.5, \mu_2 + 0.5] \\ 1 & \text{if } x_1 \in [\mu_1 - 0.5, \mu_1 + 0.5] \text{ and } x_2 \in [\mu_2 - 0.5, \mu_2 + 0.5] \end{cases} \tag{12}$$

where $\mu_1 \in [0, 1]$, $\mu_2 \in [0, 1]$ and $\mathbf{x} = (x_1, x_2)$.

The significance level $\alpha = 0.05$ was assumed and the critical value $T_{n,\alpha}$ was taken from Table 1 ($T_{20,0.05} = 0.4241$).

The estimated probabilities of rejection H_0 for various (μ_1, μ_2) are presented in Figure 3.

5 Concluding remarks

The proposed modification of the empty cells test can be used to test the hypothesis on the form of distribution of the continuous vector random variables. This test can be used in quality control procedures. It can be especially used in process monitoring using Shewhart's control chart to test the hypothesis of normality distribution in small sample cases.

The Monte Carlo study have been made. The critical values of the proposed statistic have been derived. The method of determining the value of the statistic T were described. The simulation analysis of the power of the proposed test in the two-dimensional case were done. The proposed multivariate modification of the empty cells is natural enhancement of the classical form of this test and is easy to use.

References

DAVID, H.F. (1950): *Order Statistics*. J. Wiley & Sons Inc. New York.

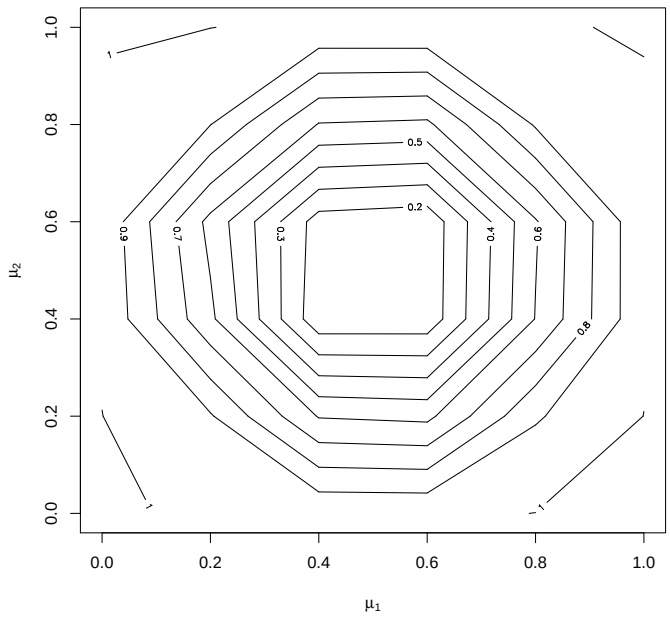


Fig. 3. The estimated probabilities of rejection H_0.

DOMANSKI, Cz. and PRUSKA, K.(2000): *Nieklasyczne metody statystyczne*. PWE Warszawa.

HELLWIG, Z. (1965): Test zgodnosci dla malej proby. *Przeglad Statystyczny 12, 99-112*.

KONCZAK, G. (2005): On the Modification of David-Hellwig Test In: Daniel Baier and Klaus-Dieter Wernecke (eds.): *Innovations in Classification, Data Science, and Information Systems. Proc. 27th Annual GfKl Conference*, Springer-Verlag, Heidelberg-Berlin, 138–145.

A Robustified MCMC Sampler – Metropolis Hastings Simulator with Trimming

Veit Köppen[1] and Hans-J. Lenz[1]

Institute of Production, Information Systems and OR, Freie Universität Berlin
Garystr. 21, 14195 Berlin, Germany, {*koeppen;hjlenz*}*@wiwiss.fu-berlin.de*

Abstract. One facet of data quality is the integrity of data. Most main business and economic indicators suffer from statistical discrepancies. Such indicators are modeled as random variables and related by a non-linear stochastic system of equations. In order to check integrity of data with respect to a fully specified model consisting of balance equations or equations due to definitions we need the joint distribution of the right hand side of each single equation, and the distribution of the related left hand side. As the Gaussian distribution is not closed under all four arithmetic operations, we need MCMC simulation to determine the probability distributions. In this paper we use the Metropolis-Hastings (MH) method. Various distributions and moments of indicators are simulated. Using the MH method in a classical way imprecise estimates may be caused by large measurement errors of the variables. Consequently, robust estimation becomes mandatory.

Keywords: particle filter theory, Monte Carlo simulation, trimming, Metropolis-Hastings algorithm

1 Introduction

Business indicators are part of many business reports. The same is true for main economic indicators like Gross Domestic Product, inflation rate or rate of unemployment as collected by the national account group of UNO. A vital question is whether such indicators contradict given balance equations or simple definitions. Business processes are related to services and goods, and deliver indicators which are measured. Of course, some integrated and aggregated indicators may be corrupted by errors or must be estimated because of being not directly observable. The same is true for economic indicators which are characterized by an higher aggregation level. Therefore business and economic indicators can be modeled as random variables. Of course, a special case are crisp data where all variances are zero. The system of equations we consider is a non-linear system with arithmetic operators connecting the variables in each equation. A classical system of business indicator is the DuPont-System, which will be investigated here for the sake of simplicity. Other systems may differ in the equations, but can be handled as well. Markov Chain Monte Carlo (MCMC) simulation is a helpful tool to investigate random variables if it is not possible to analytically determine the proper

distribution functions. The Metropolis-Hastings (MH) algorithm can be used to generate the probability distribution of random variables. This method can be easily implemented for instance in R as we did, and also proves to have reasonable computational performance.

2 The DuPont-system of business indicators

In 1919 the chemical company "DuPont" developed a system of business indicators. The equation system is:

- profit = sales - cost
- profit margin = profit / sales
- return on investment (ROI) = profit / capital
- capital turnover = sales / capital.

The two types of variables are: (1) endogenous (explained or left hand) variables: profit, ROI, profit margin, and capital turnover; (2) exogenous (explaining or right hand) variables: sales, cost, and capital. In some applications some of these variables may have missing values, or can only be estimated or measured with large imprecision. Other variables have a restricted range due to a share holder policy. We assume that all equations considered are mathematical separable. Now, as there exist various ways to compute an indicator, the question arises whether the single equation estimates are "model consistent", and how to compute a combined estimate in the sense of a full information procedure. It is evident that the same problem carries over to the main economic indicators of the national accounts system, i.e. UNO-SNA 2008, which have hundreds of variables.

3 Simulation

Computation of the corresponding joint probability function or marginal distributions of a simultaneous non-linear equation system is usually not a trivial task. We use MCMC methods for the sake of generality. Therefore, the very restrictive assumption of a Gaussian distribution family can be relaxed. Because of special features of the MH algorithm any density function can be used, cf. Hastings (1970), Chib (2004). This is specially true for mixed, skewed and heavy tail distributions.

3.1 Extending the Metropolis Hastings algorithm

In the first phase of the MH algorithm the probability functions of the given variables are determined. Furthermore, a proposal distribution is chosen for each exogenous variable. To reduce the sampling cost the shape of proposal should be as close as possible to the desired probability function. If this is

not a priori possible, the sampling size has to be extended caused by the so called burn-in phase. The size of the sample depends upon two parameters: the number of particles and the number of repetitions. The second parameter describes how many simulated means of particles per run will be used to estimate the distribution function of the exogenous variables. In the second phase all exogenous variables are simulated using MH algorithm. In the third phase distributions of the endogenous variables are estimated using only those equations where the corresponding exogenous variables are sampled. The number of iterations for these two phases is kept as a parametric constant. At the end of a MCMC simulation experiment a sample for all variables is generated. Therefore estimates of moments or densities can be derived, tests can be performed and given indicators can be checked.

Extended MH algorithm:

Experimental set-up: Fix the repetition size and the number of particles.

1. Initialize the exogenous variables with proposal distributions and target probability functions.

2. Draw samples from exogenous variables using MH algorithm.

3. Derive distribution of endogenous variables from equation system.

4. Compute the means and variances of all variables.

5. If the repetition size is not reached, go to 2.

3.2 Evaluation of the algorithm

An evaluation of the extended MH algorithm requires to analyze whether or not the quality of the simulation depends upon artifacts. We consider this step as a kind of "calibration". This primarily concerns non linearity due to products and quotients as well as non-normality. The following estimators are used for the mean and the variance of the simulated data, where T describes the sample size of the simulated data:

$$\hat{\mu} = 1/T \sum_T X_i \qquad \hat{\sigma}^2 = 1/(T-1) \sum_T (X_i - \hat{\mu})^2$$

As estimators for the triangular distribution parameters we use:

$$\hat{l} = 1/T \sum_T min(X_i) \qquad \hat{p} = 1/T \sum_T \hat{\mu} \qquad \hat{u} = 1/T \sum_T max(X_i)$$

In all of our simulation experiments the simulation uses 1000 particles (sampled values) per experiment and each experiment is repeated 5000 times.

3.3 Single linear equation with Gaussian and non-Gaussian distributed variables

We use the single linear equation $sales = profit + cost$. The exogenous random variables are profit and cost, and the endogenous random variable is

sales. Firstly, we consider the joint Gaussian distribution mainly for comparison purposes. Computation of theoretical mean is done by $E[X_1 \pm X_2] = E[X_1] \pm E[X_2]$ and variance by $Var[X_1 \pm X_2] = Var[X_1] + Var[X_2]$, of course, under the independence assumption. The results for the theoretical and simulated estimates are given in Table 1.

distribution	μ	σ^2	$\hat{\mu}$	$\hat{\sigma}^2$	l	p	u	$\hat{l}$	$\hat{p}$	$\hat{u}$
(i)	100	17	100	17						
(ii)	100	61.85	100	67.40	95	100	105	95.12	100	104.81
(iii)	100	67.45	99.62	67.42						

Table 1. Results of endogenous variable in linear equation case.

(i) Gaussian distribution

We assume Gaussian probability functions and specify the parameters as follows: $profit \sim N(20, 1^2)$ and $cost \sim N(80, 4^2)$. The proposal distribution is the Gaussian distribution with the same parameters, too. By running the algorithm estimated means and standard deviations of the exogenous variables do not differ from the desired levels. As the variables are Gaussian distributed, the sum of two Gaussian random variables is also a Gaussian distribution. Thus a Kolmogorov-Smirnov (KS) goodness of fit test can be used based on the simulated data of *sales*. The H_0-hypothesis of all tests is that the simulated data is corresponding to a Gaussian distribution with mean 100 and variance 17. Mean of p-values from simulations delivers $\overline{p} \approx 50.4\%$. We infer that the simulation at least for uncorrelated Gaussian distributed variables in the linear case is a good choice.

(ii) Triangular distribution

The next example uses the same equation with a triangular probability function. $Profit$ is distributed in the interval [19, 21] with peak (mode) at 20. *Cost* are distributed in the interval [76, 84] with peak at 80. The proposal distributions are now a uniform distribution of $profit$ in the interval [19, 21] and of *cost* in the interval [76, 84]. Testing of the simulated exogenous variables against the theoretical distribution the p-value of a two-sided KS test has a value of about 0.

Fig. 1 left upper part makes clear that simulations are insufficient near to the upper and lower bounds. The simulated densities and distribution functions are compared with their theoretical functions in Fig. 1.

(iii) Contaminated Gaussian distributions

In our third example $profit$ and *cost* are now corresponding to an ϵ-contaminated Gaussian distribution. This probability function is described by:

$$(1 - \epsilon) \cdot N(\mu_1, \sigma_1^2) + \epsilon \cdot N(\mu_2, \sigma_2^2) \text{ with } \epsilon \in [0, 1]. \quad (1)$$

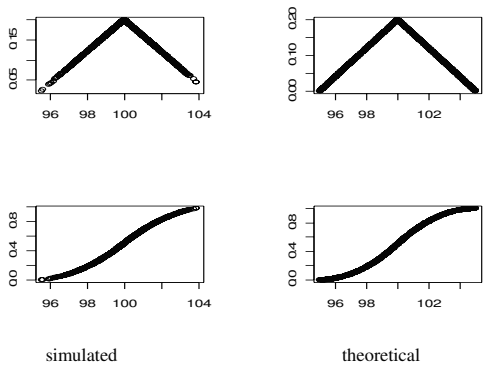


Fig. 1. Simulated and theoretical triangular distributions.

profit: $\mu_1 = 21\ \sigma_1^2 = 2\ \mu_2 = 16\ \sigma_2^2 = 1\ \epsilon = 0.1$
cost: $\mu_1 = 75\ \sigma_1^2 = 4\ \mu_2 = 90\ \sigma_2^2 = 3\ \epsilon = 0.3$

Our proposal distribution is a Gaussian distribution and the starting values for the MH algorithm are uniformly distributed in the interval [10, 30] for profit and [70, 100] for *cost*. The theoretical value of sales is well supported by the simulated variance as easily seen in Tab. 1. The KS test can not be applied, because the distribution function of the sum is not known.

As a first conclusion we note that simulation of a linear equation system using the MH algorithm leads to good results not only for endogenous variables, but also for exogenous variables. Some problems may arise if the variance is very large, since it affects the simulated data. A proposal of tackling this problem by robustification is given in the last section.

3.4 Single nonlinear equation with Gaussian and non-Gaussian distributed variables

As an example of a non-linear equation we take from the DuPont system: $profit = ROI \cdot capital$. The exogenous variables are ROI and $capital$. $Profit$ is here the endogenous variable. The theoretical estimates for the endogenous variable can be computed under independence assumption as (Mood (1973)):

$$E[X_1 \cdot X_2] = E[X_1] \cdot E[X_2],$$
$$Var[X_1 \cdot X_2] \approx E^2[X_1] \cdot Var[X_2] + E^2[X_2] \cdot Var[X_1] + Var[X_1] \cdot Var[X_2]$$

The theoretical and simulated results for the endogenous variable $profit$ are shown in Table 2.

(i) Gaussian distribution

In the case of a Gaussian probability function the variables are distributed as: $ROI \sim N(0.25, 0.025^2)$ and $capital \sim N(80, 0.4^2)$. To reduce computation

distribution	μ	σ^2	$\hat{\mu}$	$\hat{\sigma}^2$	l	p	u	$\hat{l}$	$\hat{p}$	$\hat{u}$
(i)	20	5.01	19.99	4.99						
(ii)	20	0.275	20	0.215	18.24	20	21.84	18.29	20	21.77
(iii)	20	6.19	20.03	6.78						

Table 2. Results for endogenous variable in case of a non-linear equation.

effort the proposal distributions are Gaussian distributions with same parameters. A KS test for all simulated particles per run has a mean of $\overline{p} = 0$. The underlying distribution is a normal distribution with the theoretical mean and variance. Because of the zero values of these two-sided tests, the H_0-hypothesis that the simulated data have a Gaussian distribution must be rejected.

(ii) Triangular distribution

In case of a triangular distribution, ROI is symmetrically distributed in the interval [0.24, 0.26] and capital is also symmetrically distributed, however, in the interval [76, 84]. The proposal distribution is a uniform distribution. As before the simulated data differs clearly from the theoretical distribution at the boundaries. If the exogenous variables are described by a triangular probability function, the simulated data for the endogenous variable should follow a triangular distribution. The two-sided KS test has a mean in the p-values of 0. Thus the Hypothesis, that $profit$ is described by a symmetrically triangular distribution in the interval [18.24, 21.84] is rejected.

(iii) Contaminated Gaussian distribution

In the next example variables ROI and $capital$ are distributed as a two-peak Gaussian which is described by equation 1:

ROI: $\mu_1 = 0.21\ \sigma_1^2 = 0.05^2\ \mu_2 = 0.31\ \sigma_2^2 = 0.06^2\ \epsilon = 0.4$

capital: $\mu_1 = 71\ \sigma_1^2 = 10^2\ \mu_2 = 86\ \sigma_2^2 = 8^2\ \epsilon = 0.6$

The proposal distribution for each variable is a Gaussian distribution. Due to the unknown distribution function of the product a KS test can not be applied.

4 Variance-reduction techniques by trimming

The problem of too large deviations caused by MCMC simulation can be reduced by a robust estimation procedure. To achieve robustness one possible solution is to eliminate the extreme values at both ends from the sample. A solution is the γ-trim mean and γ-trim variance, for instance cf. Büning (1991). This implies to drop the $\gamma \cdot R$ upper and lower sampled values, where $0 \leq \gamma < 0.5$. The estimators change to:

$$\hat{\mu}_\gamma = \frac{1}{R(1-2\cdot\gamma)} \sum_{i=\gamma\cdot R+1}^{R(1-\gamma)} X_i \qquad \hat{\sigma}^2 = \frac{1}{R(1-2\cdot\gamma)-1} \sum_{i=\gamma\cdot R+1}^{R(1-\gamma)} (X_i - \hat{\mu}_\gamma).$$

Because the sample size is reduced, sampling should be extended. The amount of post sampled particles depends on the target sampled size and the γ-trim. The relationship can be described by: sampling particles = demanded particles / $(1 - 2 \cdot \gamma)$. The γ parameter is dependent on the variance of the variable. If the variance is high, γ should be increased. On the other hand, the parameter should be set to 0 for a low variance. This will reduce the sampling effort and, consequently, speed-up the algorithm.

5 Simulation of a non-linear stochastic system of equations

The following example illustrates, that not only a single equation but a full system of equations can be simulated by our algorithm with trimming. The exogenous variables are capital, cost and sales and they are distributed as follows: $capital \sim N(80, 4^2)$, $cost \sim N(80, 4^2)$ and $sales \sim N(100, 5^2)$. To keep the simulation size small, the proposal distribution for the endogenous variables is also a Gaussian distribution. For simulation the DuPont system is used. Fig. 2 shows the distributions of mean and standard deviation of all variables. The sampling is done with γ equals to 0 (no trimming). Fig. 3 shows the results for γ-trimmed estimation. The value of γ is set to 0.3. The shrinkage of the spread of all distributions becomes evident, cf. Fig. 2 and 3.

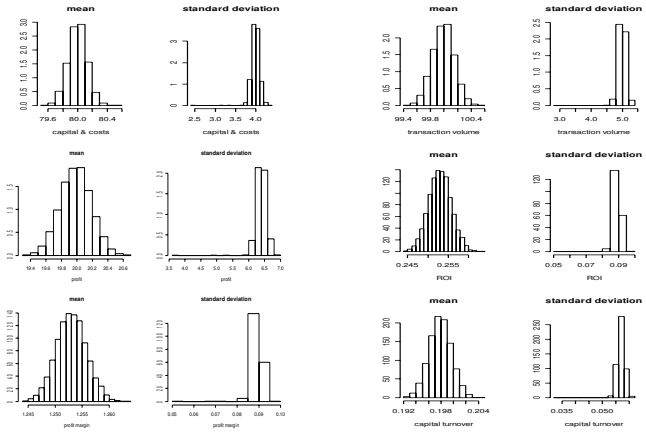


Fig. 2. Simulated means and standard deviations of the DuPont system ($\gamma = 0$).

6 Results and future work

We showed that simulation based upon the MH algorithm is a sound method to simulate simultaneous equation systems. The inputs are the system of

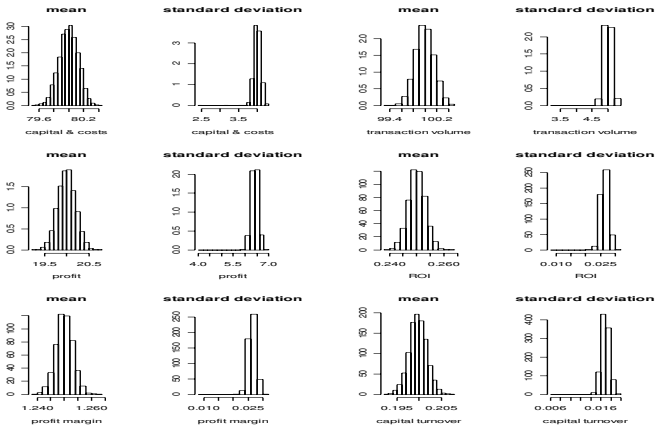


Fig. 3. Simulated means and standard deviations of γ-trimmed DuPont System ($\gamma = 0.3$.)

equations and the probability functions of the exogenous variables. All other variables can be simulated. If the sampling size is adequate, the sampling distributions are similar to the theoretical distributions. Most of those theoretical distributions are not analytically derivable, thus sampling is the only way to solve such systems. The improvement of using random variables instead of crisp quantities for main business and economic indicator systems is obvious. The often published "statistical discrepancies" related to variables become crucial if measurement errors are effective.

The results of our simulation study of non-linear equation systems are: (1) MH algorithm is a very flexible method to sample from any joint probability function. (2) A critical drawback is the sampling from a probability density function defined on a finite domain. (3) For large variances of variables the classic MH algorithm should be modified to a MH with trimming. The simulation size must accordingly be adapted. The pay-offs for increased simulation are improved estimators. However, further problems exist like missing values as well as stochastic dependencies between variables.

References

BÜNING, H. (1991): *Robuste und adaptive Tests*. Walter de Gruyter, Berlin.

CHIB, S. (2004): Markov Chain Monte Carlo Technology. In: J.E. Gentle, W. Härdle and Y. Mori (Eds.): *Handbook of Computational Statistics, Concepts and Methods*. Springer, Berlin, 71–102.

HASTINGS, W.K. (1970): Monte Carlo sampling method using Markov Chains and their applications. *Biometrica 57 (1), 97–109*.

KÖPPEN, V., HAUSMANN, A., and LENZ, H.-J. (2005): Simulation - A Support for Controllers Decision Process. In: *Proceedings of ICTM 2005: "Challenges and Prospects"*. Multimedia University, Melakka, 1155–1170.

MOOD, A.M., GRAYBILL, F.A., and BOES, D.C. (1973): *Introduction to the theory of statistics*. 3rd edition. McGraw-Hill, Tokyo.

Numerical Comparisons of Power of Omnibus Tests for Normality

Shigekazu Nakagawa[1], Naoto Niki[2], and Hiroki Hashiguchi[3]

[1] Kurashiki University of Science and the Arts
Kurashiki, 712-8505, Japan, *nakagawa@cs.kusa.ac.jp*
[2] Tokyo University of Science
Tokyo, 162-8601, Japan, *niki@ms.kagu.tus.ac.jp*
[3] Saitama University
Saitama, 338-8570, Japan, *hiro@ms.ics.saitama-u.ac.jp*

Abstract. In this paper, we propose a new omnibus test statistic for normality and give its normalizing transformation. We shows this statistic is more useful than the other omnibus test statistics by illustrating powers with some alternative hypotheses.

Keywords: omnibus test, percentiles, powers, simulation

1 Introduction

Normality is one of the most common assumptions made in the development and use of statistical procedures (see D'Agostino and Stephens(1986) and Thode(2002)). The focus of this article is related to the problem of testing whether a sample of observations comes from a normal distribution. In particular, the omnibus test that combines with the sample skewness and the sample kurtosis is treated.

An omnibus test statistic for normality, which is based on sample moments, first appeared in D'Agostino and Pearson(1973). Jarque and Bera(1987) have pointed out that it is the Lagrange multiplier test when the sample is drawn from the Pearson distributions. Urzúa(1996) has derived a better-behaved omnibus test for normality that is a natural extension of the Jarque-Bera test. Nakagawa et al.(2007) have suggested a new omnibus test statistic and have determined that a simple transformation of it to normality based on the Wilson-Hilferty transformation. Recently, the omnibus test has attracted a great deal of attention in economics. See, for example, Urzúa(1996) and Poitras(2006).

In this paper, we consider a distribution of the statistic suggested by Nakagawa et al.(2007) and we give the numerical evidence that shows this statistic is more appropriate than the other omnibus test statistics. We start with a summary explanation of the omnibus tests in Section 2. In Section 3, we provide the upper percentiles. Finally, we show powers with comparison to other omnibus statistics.

2 Jarque-Bera type omnibus tests

There are many Jarque-Bera type omnibus tests, e.g., Jarque and Bera(1987), Urzúa(1996), Nakagawa et al.(2007). The omnibus test introduced by Jarque and Bera (1987) is as follows:

$$JB = n\left[\frac{\left(\sqrt{b_1}\right)^2}{6} + \frac{(b_2-3)^2}{24}\right], \tag{1}$$

where n is the sample size, $\sqrt{b_1} = m_3/m_2^{3/2}, b_2 = m_4/m_2^2$ and the central moments are defined as $m_r = (1/n)\sum_{i=1}^{n}(X_i - m_1)^r$ and $m_1 = (1/n)\sum_{i=1}^{n} X_i$ is the sample mean. We remark here that $\sqrt{b_1}$ and b_2 are not mutually independent but they are uncorrelated and nearly independent (See D'Agostino and Stephens(1986)). We consider the JB is asymptotically distributed as the χ^2 with two degrees of freedom, because $\sqrt{b_1}$ and b_2 are asymptotically distributed as $N(0, 6/n)$ and $N(3, 24/n)$, respectively.

A natural extension of the Jarque-Bera test suggested by Urzúa(1996) is as follows:

$$EJB = \frac{\left(\sqrt{b_1}\right)^2}{\mathrm{var}(\sqrt{b_1})} + \frac{(b_2 - E(b_2))^2}{\mathrm{var}(b_2)}, \tag{2}$$

where $E(\sqrt{b_1}) = 0$, $\mathrm{var}(\sqrt{b_1}) = 6(n-2)/\{(n+1)(n+3)\}$, $E(b_2) = 3(n-1)/(n+1)$, and $\mathrm{var}(b_2) = 24n(n-2)(n-3)/\left\{(n+1)^2(n+3)(n+5)\right\}$. The EJB adjusts for the sampling distribution of the mean and variance.

The omnibus test statistic introduced by Nakagawa et al.(2007) is

$$JB' = \frac{\left(\sqrt{b_1}\right)^2}{6} + \frac{b_2^2}{24}, \tag{3}$$

and the first three central moments of the distribution of JB' are given as follows (Even though, they have obtained the first four central moments):

$$\mu_1(JB') = (3\,n-5)\left(n^2+12\,n+7\right)/\left\{8\,(n+5)\,(n+3)\,(n+1)\right\}, \tag{4}$$

$$\begin{aligned}\mu_2(JB') = {} & (n-2)\left(3\,n^7+239\,n^6+6819\,n^5+37283\,n^4-8775\,n^3\right.\\ & \left.-329451\,n^2-327711\,n-99175\right)n/\left\{2\,(n+9)\,(n+13)\right.\\ & \left.(n+7)\,(n+11)\,(n+5)^2\,(n+3)^2\,(n+1)^2\right\},\end{aligned} \tag{5}$$

$$\begin{aligned}\mu_3(JB') = {} & (n-2)\,(n-3)\left(36\,n^{12}+7621\,n^{11}+695016\,n^{10}\right.\\ & +30383269\,n^9+394400346\,n^8+1869091018\,n^7+233869944\,n^6\\ & -24473922254\,n^5-63720419472\,n^4-33172995119\,n^3\\ & \left.+20403993840\,n^2+23868462025\,n+5853965250\right)n/\\ & \left\{(n+11)\,(n+21)\,(n+9)\,(n+19)\,(n+7)\,(n+17)\right.\\ & \left.(n+15)\,(n+13)\,(n+5)^3\,(n+3)^3\,(n+1)^3\right\}.\end{aligned} \tag{6}$$

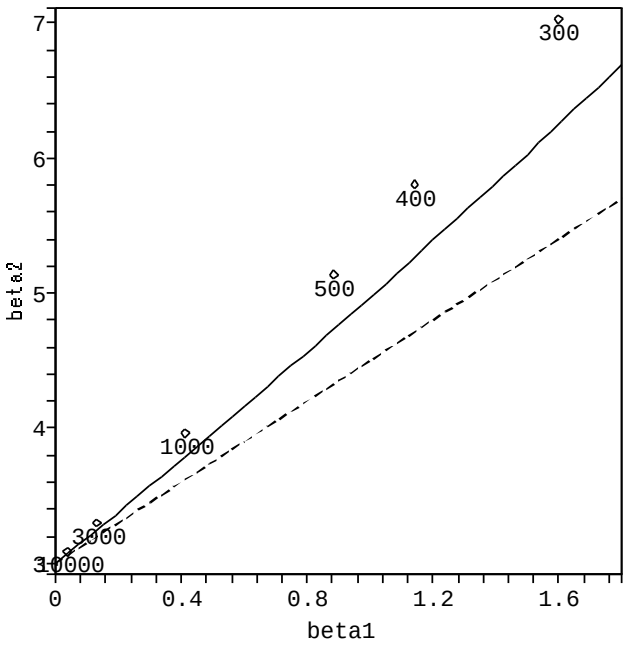


Fig. 1. $(\beta_1(JB'), \beta_2(JB'))$ chart.

From the above expressions, the Cornish-Fisher assumptions for approximate cumulants of the distribution of the standardized variate $\sqrt{n}(JB' - \mu_1(JB'))/\sqrt{\mu_2(JB')}$ are satisfied. This is the reason why they have derived the statistic JB' instead of (1).

The standardized skewness $\sqrt{\beta_1(JB')} = \mu_3(JB')/(\mu_2(JB'))^{3/2}$ and the standardized kurtosis $\beta_2(JB') = \mu_4(JB')/(\mu_2(JB'))^2$ are obtained immediately. For $n = 300, 400, 500, 1000, 3000, 10000$, points are plotted on the $(\beta_1(JB'), \beta_2(JB'))$ chart in Fig. 1. The upper solid curve refers to the Pearson type III distributions, or χ^2 distributions, and the lower dashed curve to the Pearson type V distributions, i.e. reciprocal of χ^2 distributions (Stuart and Ord (1994)). Fig. 1 shows us that the points are closed to the Pearson type V distributions when the sample of size n is large. This fact suggests to us the Wilson-Hilferty transformation is available for the reciprocal of a linear function of JB'.

In order to show a normalizing transformation, let $\sqrt{\beta_1} = \sqrt{\beta_1(JB')}$ and

$$A = 6 + \frac{8}{\sqrt{\beta_1}} \left[\frac{2}{\sqrt{\beta_1}} + \sqrt{1 + \frac{4}{\beta_1}} \right], \tag{7}$$

Table 1. Probability points of JB' for $n = 20, 35, 50, 100, 200, 500$ (approximation by using (8)).

n	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
20	0.6590	0.8426	1.0555	1.3957
35	0.6186	0.7707	0.9468	1.2280
50	0.5875	0.7156	0.8621	1.0928
100	0.5336	0.6194	0.7134	0.8538
200	0.4909	0.5454	0.6016	0.6804
500	0.4489	0.4782	0.5066	0.5435

then (if the sample of size n is large)

$$\left(\left(1-\frac{2}{9A}\right)-\left(\frac{1-(2/A)}{1+Z\sqrt{2/(A-4)}}\right)^{1/3}\right)/\sqrt{2/9A}\sim N(0,1), \qquad (8)$$

where Z is the standardized variate

$$Z = (JB' - \mu_1(JB')/\sqrt{\mu_2(JB')}.$$

The expression (8) is the normal approximation based on the Wilson-Hilferty transformation. Choosing (7) appropriately, JB' is distributed as a χ^2 with A degrees of freedom (In detail, see Nakagawa et al.(2007)).

3 Probability points

From the normal approximation (8), we have probability points u_α such that

$$\Pr(JB' > u_\alpha) = \alpha$$

for any n, immediately. For $\alpha = 0.1, 0.05, 0.025, 0.01$, Table 1 shows u_α for a range of sample sizes.

4 Powers

This section compares the power of JB, EJB and JB' when these are used as tests for normality. Monte Carlo simulation procedures are performed with four alternatives. These are as follows (These are following Urzúa (1996)): Student's t distribution with five degrees of freedom; χ^2 distribution with two degrees of freedom; Laplace distribution with mean zero and variance 25; log-normal distribution with mean zero and variance 25. The number of

replications in each simulation is 10^4 and the results are presented in Table 2, where the significance level is $\alpha = 0.1$.

In Table 2, W is the Shapiro Wilk test statistic (Shapiro and Wilk(1965)) and is used to do a traditional omnibus test for normality. It is defined as follows:

$$W = \frac{\left(\sum_{i=1}^{[n/2]} a_i \left(X_{(n+i-1)} - X_{(i)}\right)\right)^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2},$$

where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the order statistics. For $n \leq 50$, the coefficients a_i are found in many textbooks (See eg, D'Agostino and Stephens(1986) and Thode(2002)).

Table 2 shows that the power of JB', for fixed n, is the most highest except the case that alternative distribution is χ^2 with 2 degrees of freedom. For $n = 20, 35, 50$, the power of W is superior than the others.

In order to explain the relatively poor performance of their test against the chi-square alternative, we show the population skewness and the population kurtosis in Table 3. Comparing Table 2 and Table 3, we observe Jarque-Bera type omnibus test statistics are suitable for detecting kurtosis, while Shapiro Wilk test statistic seems to be suitable for detecting skewness.

References

BOWMAN, K.O. and SHENTON, L.R. (1975): Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika, 62(2), 243–250.*

D'AGOSTINO, R.B. and PEARSON, E.S.(1973): Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika, 60, 613–622.*

D'AGOSTINO, R.B. and STEPHENS, M.A.(1986): *Goodness-Of-Fit Techniques (Statistics, a Series of Textbooks and Monographs).* Marcel Dekker Inc., New York.

JARQUE, C.M. and BERA, A.K.(1987): A test for normality of observations and regression residuals. *Internat. Statist. Rev., 55(2), 163–172.*

NAKAGAWA, S., NIKI, N. and HASHIGUCHI, H.(2007): An omnibus test for normality. *Proceedings of the ninth Japan-China symposium on statistics,* 191-194.

POITRAS, G.(2006): More on the correct use of omnibus tests for normality. *Econom. Lett., 90(3), 304–309.*

SHAPIRO, S.S. and WILK, M.B.(1965): An analysis of variance test for normality: Complete samples. *Biometrika, 52, 591–611.*

STURT, A and ORD, J.K.(1994): *Kendall's Advanced Theory of Statistics, vol. I, 6th ed.*, Edward Arnold.

THODE, H.C.(2002): *Testing for normality (Statistics, a Series of Textbooks and Monographs).* Marcel Dekker Inc., New York.

URZÚA, C.M.(1996): On the correct use of omnibus tests for normality. *Econom. Lett., 90(3), 304–309.*

Table 2. Power with 10^4 replications for $n = 20, 35, 50, 100$ and $\alpha = 0.1$.

n		t_5	χ^2_2	Laplace	log-normal
20	JB	0.314	0.790	0.397	1.000
	EJB	0.329	0.704	0.422	1.000
	JB'	0.339	0.604	0.438	0.999
	W	0.254	0.904	0.351	1.000
35	JB	0.430	0.970	0.536	1.000
	EJB	0.441	0.943	0.562	1.000
	JB'	0.457	0.792	0.595	1.000
	W	0.312	0.996	0.434	1.000
50	JB	0.530	0.998	0.654	1.000
	EJB	0.545	0.994	0.679	1.000
	JB'	0.569	0.889	0.723	1.000
	W	0.355	1.000	0.502	1.000
100	JB	0.721	1.000	0.865	1.000
	EJB	0.735	1.000	0.879	1.000
	JB'	0.780	0.989	0.919	1.000

Table 3. $\sqrt{\beta_1}, \beta_2$ for each alternatives.

	t_5	χ^2_2	Laplace
$\sqrt{\beta_1}$	0	2	0
β_2	9	6	6

Parameter Estimation for Events in the Divided Observation Periods in a Poisson Process

Michio Sera, Hideyuki Imai, and Yoshiharu Sato

Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 060-0814 Hokkaido, Japan,
rgbalpha@main.ist.hokudai.ac.jp
imai@ist.hokudai.ac.jp
ysato@main.ist.hokudai.ac.jp

Abstract. In a nonhomogeneous Poisson process, we assume that the observation terms are divided into some intervals, and the aggregated numbers of events in each interval are available. In the situations, we show the asymptotic relative efficiency relative to the full data. Moreover, we proposed the EM-type algorithm for the interval data. Numerical experiments suggest that the algorithm works well.

Keywords: nonhomogeneous Poisson process , asymptotic relative efficiency, EM algorithm

1 Introduction

We consider that situations where individuals are received some treatments, and their events are monitored. The number of events are assumed to be a nonhomogeneous Poisson process with the intensity function

$$\lambda(t) = \rho(t;\alpha)\exp(\boldsymbol{x}'\boldsymbol{\beta}),$$

where α and $\boldsymbol{\beta}$ are parameters to be estimated.

When the precise event times are available, estimation of the parameter are considered in Lawless (1987). This kind of data is called "the full data" (Fig 1, top). When only the aggregated counts over the observation periods are available, estimation of the parameter and efficiency comparison relative to the full data are considered in Dean and Balshaw (1997). In this paper, this kind of date is called "the aggregated data" (Fig 1, middle). When analyzing the aggregated data, instead of the full data, some loss of efficiency may entailed. In Dean and Balshaw (1997), it is measured by the asymptotic relative efficiency (ARE) of the estimators derived from the aggregated data relative to those derived from the full data. ARE of the parameter is the ratio of the asymptotic variance of the parameter derived from the aggregated data to that of the parameter derived from the full data. The asymptotic variances of the maximum likelihood estimators are derived from the information matrix.

Moreover, when the intensity function of the Poisson process may change in the observation terms, ARE are shown in Nishijima and Kamakura (2006).

In this article, we consider situations where the observation terms are divided into some intervals, and the aggregated numbers of events in each interval are available. In this article, this data set is called the interval data. (Fig 1, bottom) It is an extension of the aggregated data.

We show the parameter estimation in the situation, and ARE related to the full data. The interval data is an incomplete data set because exact event times are not recorded. EM algorithm provides us a broadly applicable approach for maximum likelihood estimation. Thus, we propose the EM type algorithm for the interval data to improve ARE. Numerical experiments suggests that the proposed method is effective in parameter estimation for the interval data.

We consider a comparison of k treatments, where m_j individuals are given treatment j, and let $M = \sum_{j=1}^{k} m_j$ be the total sample size. Also, let G_j be the index set of individuals given treatment j. Events of each individual are monitored. We assume that the numbers of events for individuals in the term $(0, t]$ are the counting process. The counting process of the i-th individual is denoted by $X_i(t)$.

We also assume that the i-th individual is observed up to time T_i, and that T_i is independent of the counting process $X_i(t)$. In this article, we consider the three types of observation, namely (1) the full data, (2) the aggregated data, and (3) the interval data.

The full data means that all event times are precisely recorded. The event time of the i-th individual is denoted by $\omega_{is}, i = 1, \ldots, M$, and $s = 1, \ldots, n_i$ where $n_i = X_i(T_i)$. We can use $X_i(t)$ for all $i = 1, \ldots, M$, and $t \in (0, T_i]$. The aggregated data means that only the total aggregated count over observation periods $n_i = X_i(T_i)$ are available.

The interval data means that the observation term is divided into some intervals, and the aggregated numbers of events in each interval are available. The observation intervals of the i-th individual $(0, T_i]$ are denoted by $(\varphi_{i,0}, \varphi_{i,1}], (\varphi_{i,1}, \varphi_{i,2}], \ldots, (\varphi_{i,h_i-1}, \varphi_{i,h_i}]$, where h_i is the number of intervals of the i-th individual, $\varphi_{i,0} = 0$, and $\varphi_{i,h_i} = T_i$. We obtain the aggregated counts of events in each interval. The aggregated counts in the v-th interval of the i-th individual is denoted by $c_{i,v}$. These types of data are illustrated in Figure 1.

The model used in the article is a nonhomogeneous Poisson process. The conditions for a nonhomogeneous Poisson process with the intensity function $\lambda(t)$ are that as $h \to 0$

$$P(X(t+h) - X(t) = 0) = 1 - \lambda(t)h + o(h),$$
$$P(X(t+h) - X(t) = 1) = \lambda(t)h + o(h),$$

and that the random variable $X(t+h) - X(t)$ is statistically independent of the number and position of events in $(0, t)$ (Box and Lewis (1966)). We

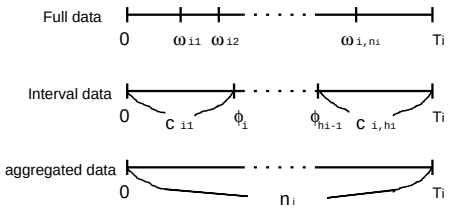


Fig. 1. The full data (top): ω_{is} is the the s-th event time of the i-th individual, the interval data (middle): c_{iv} is the aggregated count of events over the v-th interval, and the aggregated data (bottom): n_i is the aggregated count over the period $(0, T_i)$.

assume that

$$\lambda_i(t; \alpha, \boldsymbol{\beta}) = \rho(t; \alpha) \exp(\mathbf{x}_i' \boldsymbol{\beta}),$$

where $\rho(t; \alpha)$ is a twice differentiable function, called baseline intensity, depending on the parameter α, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ is the treatment indicator vector for i-th individual such that

$$x_{i1} = 1, i = 1, \ldots, M$$
$$x_{ij} = \begin{cases} 1, & \text{individual } i \text{ received treatment } j, \\ 0, & \text{otherwise,} \end{cases}$$

and $\boldsymbol{\beta}$ is a unknown parameter. This means that the $rho(t; \alpha)$ reflects the shape of the intensity function with parameter α, and $\boldsymbol{\beta}$ is a vector of regression parameters. In the interval data, we write

$$r_{i,v}(\alpha) = \int_{\varphi_{i,v-1}}^{\varphi_{i,v}} \rho(t; \alpha) dt, \quad v = 1, \ldots, h_i,$$

and in the aggregation data, $R_i(\alpha) = \int_0^{T_i} \rho(t; \alpha) dt$.

2 Likelihood function, maximum likelihood estimator and asymptotic relative efficiency

Likelihood function of a nonhomogeneous Poisson process is

$$L_{full}(\alpha, \boldsymbol{\beta}) = \exp\left(\sum_{i=1}^{M} n_i \mathbf{x}_i' \boldsymbol{\beta}\right) \left\{\prod_{i=1}^{M} \prod_{s=1}^{n_i} \rho(\omega_{is}; \alpha)\right\} \times \exp\left(-\sum_{i=1}^{M} \mu_i\right), \quad (1)$$

where $\mu_i = E(n_i) = R_i(T_i) \exp(\mathbf{x}_i' \boldsymbol{\beta})$.

Thus, the likelihood function of the full data is given by $L_{full}(\alpha, \boldsymbol{\beta})$. Similarly, the likelihood function of the interval data is

$$L_{\text{int}}(\alpha, \boldsymbol{\beta}) = \exp\left(\sum_{i=1}^{M} n_i \mathbf{x}_i' \boldsymbol{\beta}\right) \left(\prod_{i=1}^{M} \prod_{v=1}^{h_i} (r_{i,v}(\alpha))^{c_{i,v}}\right) \times \exp\left(-\sum_{i=1}^{M} \mu_i\right), \quad (2)$$

and that of aggregated data is

$$L_{ag}(\alpha,\boldsymbol{\beta}) = \exp\left(\sum_{i=1}^{M} n_i \mathbf{x}_i'\boldsymbol{\beta}\right)\left(\prod_{i=1}^{M}(R_i(\alpha))^{n_i}\right) \times \exp\left(-\sum_{i=1}^{M}\mu_i\right).$$

The maximum likelihood estimators of the parameters of these three cases satisfy

$$\left[\frac{\partial \log L_*(\alpha,\boldsymbol{\beta})}{\partial\alpha}, \frac{\partial \log L_*(\alpha,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}\right]' = \mathbf{0}$$

where $*$ is either full, int, or ag for the full data, the interval data and the aggregated data, and

$$\begin{aligned}
\frac{\partial \log L_{full}(\alpha,\boldsymbol{\beta})}{\partial\alpha} &= \sum_{i=1}^{M}\left(\sum_{s=1}^{n_i}\frac{\partial \log \rho(\omega_{is};\alpha)}{\partial\alpha} - \mu_i\frac{\partial \log R_i(\alpha)}{\partial\alpha}\right),\\
\frac{\partial \log L_{\text{int}}(\alpha,\boldsymbol{\beta})}{\partial\alpha} &= \sum_{i=1}^{M}\left(\sum_{v=1}^{h_i} c_{iv}\frac{\partial \log r_{iv}(\alpha)}{\partial\alpha} - \mu_i\frac{\partial \log R_i(\alpha)}{\partial\alpha}\right),\\
\frac{\partial \log L_{ag}(\alpha,\boldsymbol{\beta})}{\partial\alpha} &= \sum_{i=1}^{M}\left(n_i\frac{\partial \log R_i(\alpha)}{\partial\alpha} - \mu_i\frac{\partial \log R_i(\alpha)}{\partial\alpha}\right),\\
\frac{\partial \log L_*(\alpha,\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{M}(n_i\mathbf{x}_i - \mu_i\mathbf{x}_i).
\end{aligned}$$

The maximum likelihood estimators of the parameters of these three types of data are denoted by $(\hat{\alpha}_{full}, \hat{\boldsymbol{\beta}}_{full}), (\hat{\alpha}_{int}, \hat{\boldsymbol{\beta}}_{int})$ and $(\hat{\alpha}_{ag}, \hat{\boldsymbol{\beta}}_{ag})$, respectively. It is shown in Dean and Balshaw (1997) that the information matrices based on the full data and on the aggregated data are

$$\mathbf{I}_{full}(\alpha,\boldsymbol{\beta}) = \begin{bmatrix}\mathbf{X}'\mathbf{V}\mathbf{X} & \mathbf{X}'\mathbf{V}\mathbf{Z}\\ \mathbf{Z}'\mathbf{V}\mathbf{X} & \mathbf{Z}'\mathbf{V}\mathbf{Z}+\mathbf{H}_{full}\end{bmatrix}, \text{ and } \mathbf{I}_{ag}(\alpha,\boldsymbol{\beta}) = \begin{bmatrix}\mathbf{X}'\mathbf{V}\mathbf{X} & \mathbf{X}'\mathbf{V}\mathbf{Z}\\ \mathbf{Z}'\mathbf{V}\mathbf{X} & \mathbf{Z}'\mathbf{V}\mathbf{Z}\end{bmatrix},$$

where

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k}\\ \vdots & \ddots & \vdots\\ x_{M1} & \cdots & x_{Mk}\end{bmatrix}, \quad \mathbf{Z} = \left[\frac{\partial \log R_1(\alpha)}{\partial\alpha},\ldots,\frac{\partial \log R_M(\alpha)}{\partial\alpha}\right]'$$

$$\mathbf{V} = \begin{bmatrix}\mu_1 & & \\ & \ddots & \\ & & \mu_M\end{bmatrix}, \quad \mathbf{H}_{full} = -\sum_{i=1}^{M} E\left\{\sum_{s=1}^{n_i}\frac{\partial^2 \log[\rho(\omega_{is};\alpha)/R_i(T_i;\alpha)]}{\partial\alpha^2}\right\}.$$

The information matrix based on the interval data is

$$\mathbf{I}_{int}(\alpha,\boldsymbol{\beta}) = \begin{bmatrix}\mathbf{X}'\mathbf{V}\mathbf{X} & \mathbf{X}'\mathbf{V}\mathbf{Z}\\ \mathbf{Z}'\mathbf{V}\mathbf{X} & \mathbf{Z}'\mathbf{V}\mathbf{Z}+\mathbf{H}_{int}\end{bmatrix} \tag{3}$$

where

$$\mathbf{H}_{int} = -\sum_{i=1}^{M} E\left\{\sum_{s=1}^{h_i} \frac{\partial^2 \log r_{i,s}(\alpha)}{\partial \alpha^2}\right\}.$$

ARE of the parameter is the ratio of the asymptotic variance of the parameter derived from the aggregated data to that of the parameter derived from the full data. Since the information matrix is the asymptotic variance of the maximum likelihood estimator, we obtain the ARE from the information matrix.

ARE of $(\hat{\alpha}_{ag}, \hat{\boldsymbol{\beta}}_{ag})$ relative to $(\hat{\alpha}_{full}, \hat{\boldsymbol{\beta}}_{full})$ is shown in Dean and Balshaw (1997) such that

$$ARE(\hat{\alpha}_{ag}) = 1 - \frac{\mathbf{H}_{full}}{\mathbf{E} + \mathbf{H}_{full}}, \tag{4}$$

$$ARE(\hat{\beta}_{ag,1}) = 1 - \left\{\frac{l_1}{(\sum_{i\in G_1} \mu_i)^{-1}\mathbf{E} + l_1}\right\}\left(1 - \frac{\mathbf{H}_{full}}{\mathbf{E} + \mathbf{H}_{full}}\right),$$

$$ARE(\hat{\beta}_{ag,j}) = 1 - \left\{\frac{l_j}{((\sum_{i\in G_1} \mu_i)^{-1} + (\sum_{i\in G_j} \mu_i)^{-1})\mathbf{E} + l_j}\right\}\left(1 - \frac{\mathbf{H}_{full}}{\mathbf{E} + \mathbf{H}_{full}}\right),$$

where

$$\mathbf{E} = \sum_{j=1}^{k} \sum_{i\in G_j} \mu_i \left(\frac{\partial \log R_i(\alpha)}{\partial \alpha} - \frac{\sum_{i\in G_j} R_i(\alpha) \frac{\partial \log R_i(\alpha)}{\partial \alpha}}{\sum_{i\in G_j} R_i(\alpha)}\right)^2,$$

and

$$l_j = \left(\frac{\sum_{i\in G_j} \frac{\partial R_i(\alpha)}{\partial \alpha}}{\sum_{i\in G_j} R_i(\alpha)} - \frac{\sum_{i\in G_1} \frac{\partial R_i(\alpha)}{\partial \alpha}}{\sum_{i\in G_1} \frac{\partial R_i(\alpha)}{\partial \alpha}}\right)^2.$$

As a similar fashion, we obtain the ARE of $(\hat{\alpha}_{int}, \hat{\boldsymbol{\beta}}_{int})$ relative to $(\hat{\alpha}_{full}, \hat{\boldsymbol{\beta}}_{full})$. From (3), we obtain

$$ARE(\hat{\alpha}_{int}) = 1 - \frac{\mathbf{H}_{full} - \mathbf{H}_{int}}{\mathbf{E} + \mathbf{H}_{full}},$$

$$ARE(\hat{\beta}_{int,1}) = 1 - \left\{\frac{l_1}{(\sum_{i\in G_1} \mu_i)^{-1}(\mathbf{E} + \mathbf{H}_{int}) + l_1}\right\}\left(1 - \frac{\mathbf{H}_{full} - \mathbf{H}_{int}}{\mathbf{E} + \mathbf{H}_{full}}\right),$$

$$ARE(\hat{\beta}_{int,j}) = 1 - \left\{\frac{l_j}{((\sum_{i\in G_1} \mu_i)^{-1} + (\sum_{i\in G_j} \mu_i)^{-1})(\mathbf{E} + \mathbf{H}_{int}) + l_j}\right\} \times \left(1 - \frac{\mathbf{H}_{full} - \mathbf{H}_{int}}{\mathbf{E} + \mathbf{H}_{full}}\right).$$

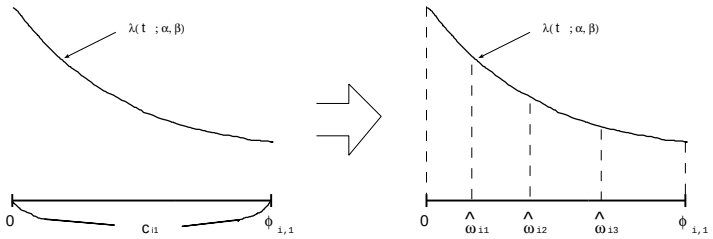


Fig. 2. Estimators of $\omega_{i,n}$.

From (4) and (5), we obtain the following inequalities:

$$ARE(\hat{\alpha}_{ag}) \leq ARE(\hat{\alpha}_{\text{int}}) \leq 1,$$
$$ARE(\hat{\beta}_{ag,j}) \leq ARE(\hat{\beta}_{\text{int},j}) \leq 1.$$

These show that the asymptotic relative efficiency of the maximum likelihood parameters derived from the interval data is greater than those derived from the aggregated data.

3 The EM type algorithm for the interval data

The aggregated data and the interval data are both incomplete data sets because exact event times are not recorded. EM algorithm provides us a broadly applicable approach for maximum likelihood estimation Mclachlan and Krishnan (1997).

We propose the EM type algorithm for the interval data. The proposed algorithm is as follows:

STEP 0: Compute $(\hat{\alpha}_{int}, \hat{\boldsymbol{\beta}}_{int})$, based on the interval data, which maximize the likelihood function $L_{int}(\alpha, \boldsymbol{\beta})$ in (2). Put $k \to 1$ and $(\hat{\alpha}_{int}^{(0)}, \hat{\boldsymbol{\beta}}_{int}^{(0)}) = (\hat{\alpha}_{int}, \hat{\boldsymbol{\beta}}_{int})$.
STEP 1: Estimate event times. In v-th interval of i-th individual, the estimator $\hat{\omega}_{in}^{(k)}$, $u = n_{i,v-1} + 1, \ldots, n_{i,v}$ satisfies that all

$$\int_{\hat{\omega}_{i,\mu-1}^{(k)}}^{\hat{\omega}_{i,\mu}^{(k)}} \lambda(t; \alpha^{(k)}, \beta^{(k)}) dt,$$

are the same value for $\hat{\omega}_{in}^{(k)}$, $u = n_{i,v-1} + 1, \ldots, n_{i,v} + 1$ where $n_{i,v} = \sum_{j=1}^{v} c_{i,j}$, $\hat{\omega}_{i,\mu_i,v-1} = \varphi_{i,v-1}$ and $\hat{\omega}_{i,n_i,v+1} = \varphi_{i,v}$.
STEP 2: Compute $(\hat{\alpha}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})$ based on the full data $\{\hat{\omega}_{iu}^{(k)},\ i = 1, \ldots, M,\ u = 1, \ldots, n_{i,h_i}\}$, which maximizes the likelihood function $L_{full}(\alpha, \boldsymbol{\beta})$ in (1).

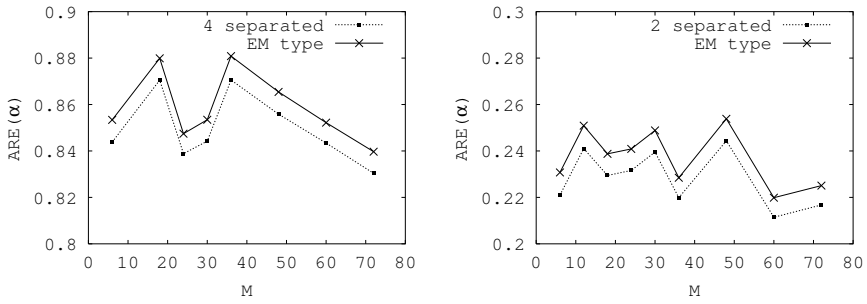


Fig. 3. Asymptotic relative efficiency of α.
Case 1(left): 4 intervals $(0, T_i/4], (T_i/4, 2T_i/4], (2T_i/4, 3T_i/4], (3T_i/4, T_i]$.
Case 2(right): 2 intervals $(0, T_i/20], (T_i/20, T_i]$

STEP 3: If $\left|\hat{\alpha}^{(k)} - \hat{\alpha}^{(k-1)}\right| + \left\|\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}}^{(k-1)}\right\| < \varepsilon$, then goto STEP 4, otherwise put $k \leftarrow k+1$ and go to STEP 2.
STEP 4: Output $(\hat{\alpha}^{(k)}, \hat{\boldsymbol{\beta}}^{(k)})$ as the estimator.

The STEP 1 is the maximization step, and the STEP 2 is the estimation step. In Step 1, we estimate the event time by the predicted values under the condition that the parameters α and $\boldsymbol{\beta}$ are fixed. Figure 2 illustrates the case for $v = 1$ and $c_{i1} = 3$. The expectation values in four periods $(\hat{\omega}i, s-1, \hat{\omega}_{i,s}), s = 0, 1, 2, 3$ are the same. In Step 2, the maximum likelihood estimators based on the estimated event time $\{\hat{\omega}_{iu}^{(k)},\ i = 1, \ldots, M,\ u = 1, \ldots, n_{i,h_i}\}$ are calculated.

4 Numerical experiments

We perform numerical experiments to confirm the performance of the proposed method. A Weibull baseline intensity function $\rho(t; \alpha) = \alpha t^{\alpha-1}$ was used, because the distribution is often suitable where the conditions of "strict randomness" of the exponential distribution are not satisfies.

The true parameters are $\alpha = 1.5, \boldsymbol{\beta} = (1.0, -1.0, 1.0)'$, and the end time T_i is distributed according to the uniform distribution on $[20, 25]$. For the interval data, observation terms are divided as follows:

Case 1 : 4 intervals $(0, T_i/4]$, $(T_i/4, 2T_i/4]$, $(2T_i/4, 3T_i/4]$, $(3T_i/4, T_i]$

Case 2 : 2 intervals $(0, T_i/20]$, $(T_i/20, T_i]$

The number of individuals M varies from 5 to 75 Figure 3 illustrates ARE of α, and Figure 4 illustrates ARE of β_1.

In all figures, solid lines show ARE of the interval data based on EM type algorithm, and dotted lines show ARE of the estimator based on ML. These figures show that the proposed method improves the ARE of the estimators.

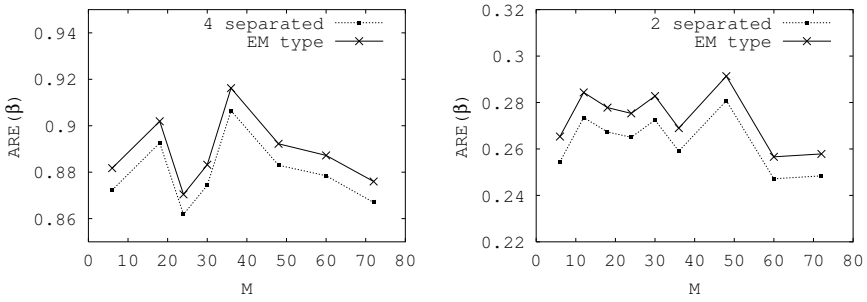


Fig. 4. Asymptotic relative efficiency of β_1.
Case 1 (left) : 4 intervals $(0, T_i/4]$, $(T_i/4, 2T_i/4]$, $(2T_i/4, 3T_i/4]$, $(3T_i/4, T_i]$.
Case 2 (right) : 2 intervals $(0, T_i/20]$, $(T_i/20, T_i]$.

5 Conclusion

In this article, we have considered the parameter estimation when number of event of individual are assumed to be a nonhomogeneous Poisson process. Three types of observations have been considered, the full data, the interval data, and the aggregated data. We have derived the information matrix based on the interval data. Consequently, we have obtained the asymptotic variances of the parameter based on the interval data. We also have shown the asymptotic relative efficiency for the interval data based on the asymptotic variances.

We have also proposed the EM type algorithm both for the aggregated data and the interval data because exact event times are not recorded. By estimating event times by the algorithm, we have shown that the asymptotic relative efficiency is improved. Moreover, numerical experiments have performed to confirm the efficacy of the proposed method.

References

BOX, D. R. and LEWIS, P. A. W (1966): *The Statistical Analysis of Series of Events.* Methuen, London.

DEAN, C. B. and BALSHAW, R. (1997): Efficiency lost by analyzing counts rather than event times in Poisson and overdispersed Poisson regression models. *Journal of the American Statistical Society 92*, 1387–1398.

LAWLESS, J. F. (1987): Regression methods for Poisson Data. *Journal of the American Statistical Society 82*, 808–815.

MCLACHLAN, G. J. and KRISHNAN, T. (1997): *The EM Algorithm and Extension.* Wiley, New York.

NISHIJIMA, K. and KAMAKURA, T. (2006): Precision of parameter estimation in statistical models for recurrent events with time-dependent covariates. *Bulletin of the Computational Statistics of Japan 19(2)*, 103–126 (in Japanese).

Part XX

Spatial Statistics

Detection of Space-Time Hotspots for Korean Earthquake Data Using Echelon Analysis

Sanghoon Han[1] , Fumio Ishioka[2] and Koji Kurihara[3]

[1] Graduate School of Environmental Science, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan, *shhan@ems.okayama-u.ac.jp*
[2] Graduate School of Environmental Science, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan, *fishioka@ems.okayama-u.ac.jp*
[3] Graduate School of Environmental Science, Okayama University, 3-1-1 Tsushima-naka Okayama 700-8530, Japan, *kurihara@ems.okayama-u.ac.jp*

Abstract. The purpose of this paper is to detect hotspot areas for spatial data using echelons. We perform echelon analysis on Korean earthquake data using ESRI's ArcGIS a piece of geographical information system (GIS) software which meshes areas in order to acquire contiguity information about them. With this contiguity information on the meshed areas, we can detect hotspots using echelon analysis and spatial scan statistics. In addition, we can detect space-time hotspots for earthquakes of magnitude greater than 3 that people can feel for the 1st period (1978-1987), 2nd period (1988-1997), and 3rd period (1998-2007) from within the Korean earthquake data.

Keywords: hotspot, echelon analysis, spatial scan statistics, spatial-temporal data

1 Introduction

Recently, the detection of areas which differ significantly from other areas, as in the incidence of diseases in certain areas, is an important subject for study. The use of spatial scan statistics (Kulldorff, 1997) is a method for the detection and inference of zones with significantly high or low rates based on the likelihood ratio. These zones are called hotspots. Kulldorff detected hotspots using a method which scans the area based on circular form. This method is excellent for finding circular hotspots, but it is not appropriate for the detection of hotspots consisting of lines or other complex forms.

To overcome this problem, we applied echelon analysis (Myers et al., 1997), which is an analytical method for investigating the phase-structure of spatial data systematically and objectively, based on neighboring information between each cell. Kurihara (2004) studied the classification of geospatial lattice data and their graphical representation. And Kurihara et al. (2006) expanded the multivariate spatial data to include data with a time-series structure and analyzed it on the time space. In addition, Ishioka et al. (2007) applied echelon analysis to 3-dimensional spatial data.

The contemporary seismicity of the Korean peninsula is found to be low compared to the past, and to neighbors Japan and China. Although the Korean Peninsula is usually believed to have little seismicity, this is not accurate when considering its historical seismicity (Kim, 1980). In addition, the seismic intensity of earthquakes is strengthening more and more in South Korea.

In this paper, we will attempt to find hotspots based on echelon analysis of South Korean earthquake data. First, we will gather contiguity information from data obtained using ArcGIS. We will find candidate hotspots using echelon analysis of Korean earthquake data from 1978 to 2007. In addition, we detect the space-time hotspots for the earthquakes of magnitude greater than 3 that people can feel for the 1st period (1978-1987), 2nd period (1988-1997), and 3rd period (1998-2007) from within the Korean earthquake data.

2 Detection of hotspots

2.1 Echelon analysis

The echelon analysis is based on the areas of relative high and low values of response variables for spatial data. The echelon approach aggregates the areas in which the values have the same topological structure and makes hierarchically related structure of these areas. The echelon dendrogram is a graph which represents the surface topology of cellular data and hierarchical structure of these data. For one horizontal dimension case, the hypothetical set of hillforms is divided to the same structured areas shown in the Figure 1.

The horizontal line shows the position (x) of spatial data, and the vertical line shows the value (h) of response value for specified horizontal position. Thus, the data are given by set of (x, h). The structure of these hillforms is given by the following echelon dendrogram shown in the Figure 2. The echelon approach is able to apply every spatial data which have connective relation (neighbor) information and response value.

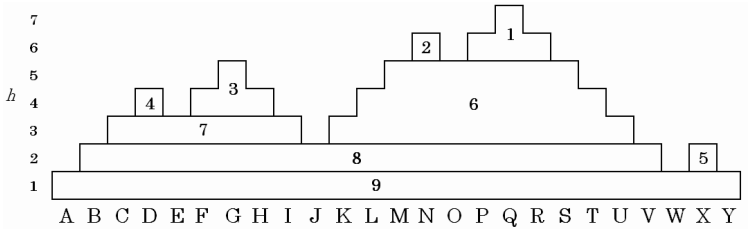


Fig. 1. The hypothetical set of hillforms in one horizontal dimension.

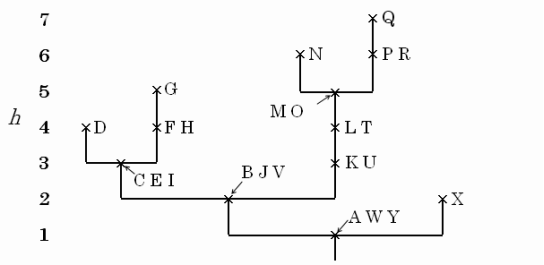


Fig. 2. The echelon dendrogram of Figure 1.

2.2 Spatial scan statistics

Spatial scan statistics are used to detect the areas with significantly high or low significant rates and to find features of the data. These areas are called hotspots. Suppose hotspot candidate area Z is within the whole area G. Each individual within the area Z has population probability p_1 for the attribute, while the population probability for individuals outside of the area Z is p_2. The probability for any individual is independent of the probabilities for all others. The null hypothesis is $H_0 : p_1 = p_2$. The alternative hypothesis for detecting high rates is $H_1 : p_1 > p_2$. Let $n(G)$ be the total population in the whole area G, and $n(Z)$ be the population within area Z. $c(G)$ is the total number of individuals with the attribute in the whole area G and $c(Z)$ is the number of individuals with the attribute within the area Z. Here, we consider a model based on the Poisson distribution. We can hence write the likelihood function as

$$L(Z, p_1, p_2) = \frac{\exp[-p_1 n(Z) - p_2(n(G) - n(Z))]}{c(G)!} p_1^{c(Z)} p_2^{c(G)-c(Z)} \prod_{x_i}^{n} n(x_i).$$

In order to maximize the likelihood function, we calculate the maximum likelihood function conditioned for area Z. The maximum likelihood estimators $\hat{p_1} = c(Z)/n(Z)$ and $\hat{p_2} = (c(G) - c(Z))/(n(G) - n(Z))$ are substituted.

$$L(Z) = \frac{\exp[-c(G)]}{c(G)!} (\frac{c(Z)}{n(Z)})^{c(Z)} (\frac{c(G) - c(Z)}{n(G) - n(Z)})^{c(G)-c(Z)} \prod_{x_i}^{n} n(x_i)$$

The likelihood ratio λ is maximized over all subsets of the whole area in order to detect the hotspots.

$$\lambda = \frac{\max_z L(Z)}{L_0} = \frac{(\frac{c(Z)}{n(Z)})^{c(Z)} (\frac{c(G) - c(Z)}{n(G) - n(Z)})^{c(G)-c(Z)}}{(\frac{c(G)}{n(G)})^{c(G)}}$$

Here, L_0 is the following likelihood function under the null hypothesis.

$$L_0 = \sup_p \frac{\exp[-pn(G)]}{c(G)!} p^{c(G)} \prod_{x_i}^{n} n(x_i) = \frac{\exp[-c(G)]}{c(G)!} (\frac{c(G)}{n(G)})^{c(G)} \prod_{x_i}^{n} n(x_i)$$

The test statistic λ is also represented by

$$\lambda = (\frac{c(Z)}{e(Z)})^{c(Z)} (\frac{c(G) - c(Z)}{e(G) - e(Z)})^{c(G)-c(Z)}$$

where $e(Z)$ is the expected value of the attribute within area Z, and $e(G) = c(G)$.

3 Korean earthquake data

3.1 Description of data and kriging

For analysis, we used Korean earthquake data acquired from 1978 to 2007. For the 24 aggregated zones, a total of 755 cases were reported. From these data, we used only the data on South Korea (613, 81%), excluding North Korea and the missing value (23).

Kriging is a method of interpolation named after a South African mining engineer named D. G. Krige who developed the technique in an attempt to more accurately predict ore reserves. Over the past several decades kriging has become a fundamental tool in the field of geostatistics. The types of kriging are simple, ordinary, indicator and so on (Cressie (1991)). The candidate of hotspot with statistical significance can not be found in such kriging (Figure 3).

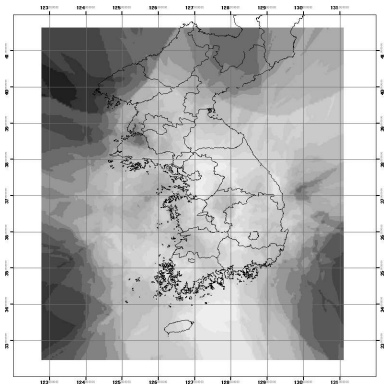


Fig. 3. Indicator kriging of Korean earthquake data.

3.2 Detection of hotspots for spatial data

Many methods have been proposed to detect the candidate of hotspot, such as Moran's I statistic, Tango's test statistic, GAM (Geographical Analysis Machine), Kulldorff's scan, Flexible scan and so on. Echelon scan is based on statistical map (Right of Figure 4). Statistical map with shading is used to show quantitative information and it varies geographically. But we can only find contiguous hotspots in this map with low accuracy of the visual decoding.

Echelon dendrogram (Left of Figure 4) has the hierarchical structure of the data. In addition, the features of each area (mesh) are also shown as the abbreviated labels in dendrogram. The candidate area of hotspot is scanned from the top of dendrogram based on hierarchical spatial structure (Kurihara, 2004). The analytical procedure was as follows.

Step 1) Plot the epicenter of an earthquake on the map using latitude and longitude

Step 2) Create a mesh (22×20) **including all points (nearly** $\mathbf{1600}km^2$**)**

Usually the 1st mesh and the 2nd mesh are used as each nearly $6400km^2$ and $100km^2$ (Area of each mesh (lattice)). In this paper, we used these but most areas became the hotspot. So, we use the 1.5 mesh. The 1.5 mesh is nearly $1600km^2$.

Step 3) Count the frequency of points in each mesh

Step 4) Perform echelon analysis, calculate spatial scan statistics and find the candidate of hotspots

Using contiguity information for each mesh, we performed echelon analysis. The candidates of hotspots will be located in the top echelon of the dendrogram.

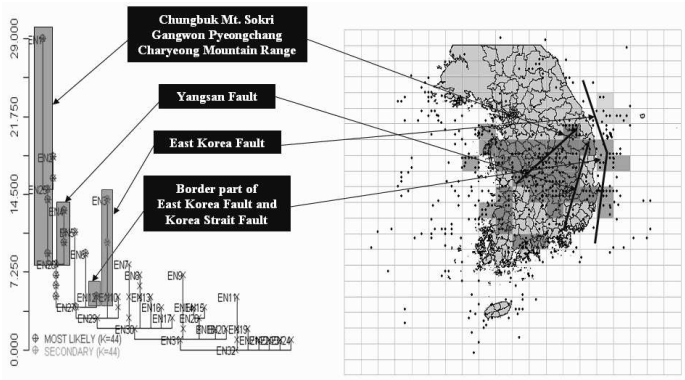


Fig. 4. Echelon dendrogram and hotspots on the map (Area of each mesh (lattice) is nearly $1600km^2$).

Figure 4 is drawn on the basis of the hotspots detected in the echelon dendrogram and is drawn separately for the comprehension. From Figure 4, we can see that the hotspots include the middle and southern part of South Korea. That is, hotspot areas are Chungnam, Chungbuk and Gyeongbuk ($\log\lambda$=354.9508, p-value=0.001). This hotspot area includes faults (the East Korea Fault and the Korea Strait Fault), a capable fault (the Yangsan Fault) and a mountain range (the Charyeong Mountain Range). The hotspot that likelihood ratio is high such as this result is only found in Flexible scan.

3.3 Detection of hotspots for spatial-temporal data

For hotspot detection, we handled only space data provided by observation results from a certain point in time until the present. However, the space data have in many cases shown time-series orders. Therefore, it is very important that we analyze changes in the time-series orders of the hotspots. Spatial-temporal data is considered to be continuous data consisting of multiple spatial data. Figure 5 shows examples of space-time hotspots. In the left part of Figure 5, the horizontal axis represents the usual spatial area (or space), and the vertical axis represents the passage of time. The trajectory (sequence of time slices) of the hotspot areas is shown in the right part of Figure 5. By way of the space-time hotspots, we can see the time changes in hotspot areas, such as expansion, reduction, movement, division and so on. Kulldorff et al. (2005) used overlapping cylinders to scan the spatial-temporal area and detect space-time hotspots. Here, we perform the echelon technique for spatial-temporal data. Echelon scan is only technique which detect any types of space-time hotspot shown in Figure 5. We define the neighbor information for spatial-temporal data and detect the space-time hotspots.

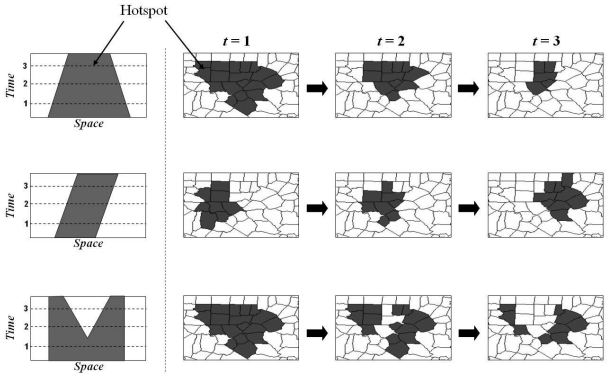


Fig. 5. The three different hotspot types over space and time (left). The trajectory (sequence of time slices) of hotspot areas (right).

The data on the space-time of the local model are provided as a certain division $D_i, i = 1, 2, \ldots, k$ in $t, t = 1, 2, \ldots, T$ at a point in time, and the data are given in $h(D_{t,i})$. Then we define contiguity information $NB(D_{t,i})$ between each pair of domains as follows.

$$\begin{aligned} NB(D_{t,i}) = \{&D_{t,j} | \text{regions } i \text{ and } j \text{ are connected}\} \\ &\cap D_{t+1,i} \\ &\cap D_{t-1,i} \end{aligned}$$

From the data value $NB(D_{t,i})$ and contiguity information $h(D_{t,i})$ for each domain, the making of an echelon dendrogram for the data over space and time is enabled.

As an example, we detected the space-time hotspots for earthquakes of magnitude greater than 3 that people can feel consisting for the 1st period (1978-1987), 2nd period (1988-1997) and 3rd period (1998-2007) from the Korean earthquake data. We analyzed the data based on the interval α of the period t, α=1, 2, 3, 5. However, we was not able to get a clear, good result. So, we used α=10 and analyzed it. We define neighbors as the meshes which border each other within the same period. For different periods, we define the same meshes as neighbors to each other. Under these conditions, we were able to make an echelon dendrogram. To detect the hotspot areas maximizing the log likelihood ratio (LLR), we scanned the areas from the upper echelon to the bottom, based on the hierarchical structure. Here, we decided on the maximum number of hotspot areas in order to avoid allowing the hotspot regions to become too wide. We thought that a suitable size for a hotspot area would be less than one-fourth or one-fifth the size of the whole

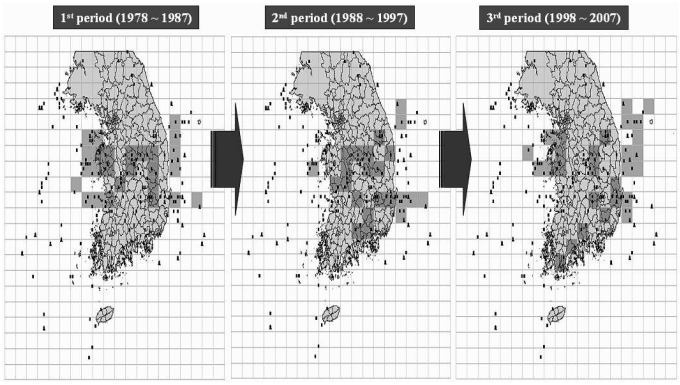


Fig. 6. Hotspots in the Korean earthquake data consisting of the 1st, 2nd and 3rd periods.

area. In addition, we calculated the p-value using the Monte Carlo hypothesis test on the null hypothesis.

Space-time hotspot areas ($\log\lambda$=186.2112, p-value=0.001) are shown on the geographical map in Figure 6. In this case, the hotspot areas for earthquakes of magnitude greater than 3 within the Korean earthquake data show a slight reduction from the 1st period to the 2nd period and then a slight increase again into the 3rd period. Additionally, Figure 6 shows that the hotspot areas divide into two regions as time passes. That is, the hotspots were separated between the West and Southeast.

4 Conclusions

Echelon analysis enables the expression of the phase-structure of space data. Hotspots are given as the upper echelon in the dendrogram. In this paper, we detected the hotspot areas for Korean earthquake data using echelon and spatial scan statistics. Additionally, we detected hotspots for earthquakes of magnitude greater than 3 for the 1st period (1978-1987), and then expanded to the spatial data with a time-series structure, to which the data from the 2nd period (1988-1997) and 3rd period (1998-2007) have been added. According to the results of the analysis we can say that the hotspots have changed as time has passed. By considering the spatial data on the time space, we can verify the expansion or reduction of the hotspots.

References

CRESSIE, N. (1991): *Statistics for Spatial data.* Willey, New York.

ISHIOKA, F., KURIHARA, K., SUITO, H., HORIKAWA, Y. and ONO, Y. (2007): Detection of hotspots for three-dimensional spatial data and its application to environmental pollution data. *Journal of Environmental Science for Sustainable Society. 1, 15-24.*

KIM, S.G. (1980): Seismicity of the Korean Peninsula and Its Vicinity. *Jour. Kor. Inst. Mining Geol. 13 (1), 51-63.*

KULLDORFF, M. (1997): A spatial scan statistics. *Communications in Statistics, Theory and Methods. 26, 1481-1496.*

KULLDORFF, M., HEFFERNAN, R., HARTMAN, J., ASSUNCAO, R.M. and MOSTASHARI, F. (2005): A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine, 2, 216-224.*

KURIHARA, K. (2004): Classification of geospatial lattice data and their graphical representation. *Classification, Clustering, and Data Mining Applications.* Springer, Canada, 251-258.

KURIHARA, K. ISHIOKA, F. and MOON, S.H. (2006): Detection of Hotspots on Spatial Data Using Principal Component Analysis. *Journal of the Korean Data Analysis Society. 8 (2), 447-458.*

MYERS, W.L., PATIL, G.P. and JOLY, K. (1997): Echelon approach to areas of concern in synoptic regional monitoring. *Environmental and Ecological Statistics. 4, 131-152.*

Using Geometric Anisotropy in Variogram Modeling

Takafumi Kubota[1] and Tomoyuki Tarumi[2]

[1] Okayama University, Graduate school of humanities and social sciences Tsushimanaka 3-1-1 Okayama, Japan, *kubota@law.okayama-u.ac.jp*
[2] Okayama University, Admission Centre Tsushimanaka 3-1-1 Okayama, Japan

Abstract. We propose to apply geometric anisotropy to geostatistical data and fit to the ellipse model by least square method. The range values of variogram in several directions are calculated and the values optimized using the ellipse model to obtain the parameters of the ellipse. To ensure the validity of this method we employ the water pollution data of Okayama prefecture in Japan, and show the prediction error of the geometric anisotropic model. We find that the parameters calculated by our method are smaller than that determined by the isotropic model.

Keywords: kriging, variogram, anisotropy

1 Introduction

Environmental data, such as data on weather and water pollution, have several parameters which provide not only their observed characteristic values but also geometric information. These are called spatial data or geostatistical data. One purpose for geostatistical data analyses is to predict the characteristic values of unobserved points. The most famous prediction method is Kriging, and for this method variogram which measures variance of data is required.

When we construct a variogram model and calculate its parameters, we need to incorporate directional behavior of the parameters. A variogram exhibiting the same behavior in different directions is called an isotropic variogram, while an anisotropic variogram displays different behavior in different directions. Anisotropy expressed as variograms with different range values in different directions is known as geometric anisotropy.

There are many studies which discuss geometric anisotropy. For example, Ecker and Gelfand (1997) adopts a Bayesian perspective to study models involving both trend surface and range anisotropic covariance parameters, and Ecker and Gelfand (1999) proposes methodology which simultaneously estimates the linear transformation and the other semivariogram parameters which allow full inference for any characteristic of the geometrically anisotropic model. Furthermore there are many software systems using geometric anisotropy. *geoR* is a free and open-source package for geostatistical

analysis which interactively models the variogram anisotropy. A web application performing geostatistical analysis, especially estimation of the variogram model with anisotropy, has been developed by Kubota and Tarumi (2007).

To obtain the parameters of geometric anisotropy we propose to fit to the ellipse model by least square method. We calculated the range values of the variogram in several directions and optimized these values using the ellipse model. To ensure the validity of our method we employed Kriging using geometric anisotropic variogram and calculated the prediction error. Employing the method of cross validation, we compared the resulting sum of squared errors with those of an isotropic variogram. In our study we used the water pollution data of Okayama prefecture in Japan, and found the prediction error of the geometric anisotropic model. The sum of squared errors calculated by our method were smaller than those obtained from the isotropic model.

We describe estimation of variogram parameters with geometric anisotropy in section 2, apply this to practical data in section 3, and provide concluding remarks in section 4.

2 Estimation of variogram parameters with geometric anisotropy

We first of all set the number of directions (*ndir*) and the tolerance values (*tol*) for each direction. We assume the same tolerance value in every direction. Figure 1 shows the four directions (0, 45, 90, and 135 degree), clockwise rotations from the north direction (N, NE, E, and SE) with 22.5 degree of tolerance which includes both side of each direction.

To obtain the variogram cloud, we measured the difference between pairs of characteristic values $z(\mathbf{x}_\alpha)$ and $z(\mathbf{x}_\beta)$ located at points $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$ within each direction. The difference is

$$\gamma(\mathbf{h})^* = \frac{1}{2}(z(\mathbf{x}_\alpha + \mathbf{h}) - z(\mathbf{x}_\alpha))^2 \tag{1}$$

where $\mathbf{h} = \mathbf{x}_\alpha - \mathbf{x}_\beta$. We have to consider not only the distance but also the direction for each pair of observed points. For exploratory data analysis it is conventional to fit the theoretical variogram to define several classes and calculate the mean value of the distances and differences between the characteristic values. But we directly fitted the theoretical variogram to a variogram cloud. In calculating a variogram cloud, the cutoff value, which limits the distance between observation pairs, is very significant. We adopt as the cutoff value half of the maximum distance between every pair, as proposed in Kubota et al. (2005). As the model for theoretical variogram we

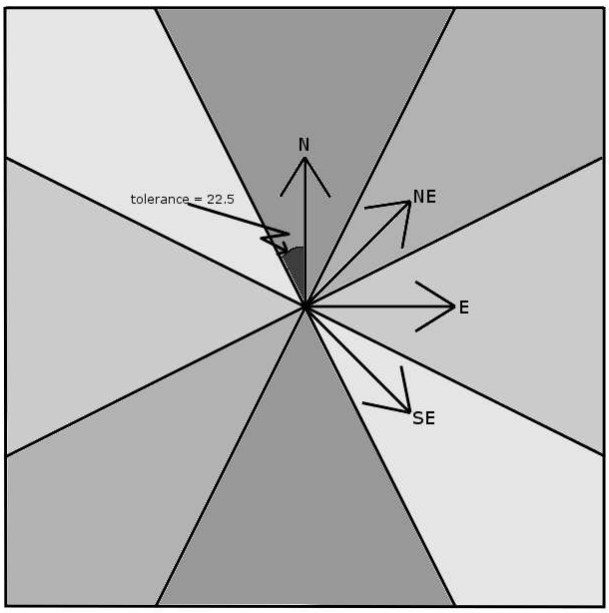


Fig. 1. The four directions (0, 45, 90, and 135 degree), clockwise rotations from the North direction (N, NE, E, and SE).

use the spherical model.

$$\gamma(\mathbf{h};\xi) = \begin{cases} \xi_0 + \xi_1(\frac{3}{2}\frac{||\mathbf{h}||}{\xi_2} - \frac{1}{2}[\frac{||\mathbf{h}||}{\xi_2}]^3), & 0 < ||\mathbf{h}|| \leq \xi_2 \\ \xi_0 + \xi_1, & ||\mathbf{h}|| > \xi_2 \\ 0, & ||\mathbf{h}|| = 0 \end{cases} \tag{2}$$

where ξ_0 is nugget effect value, ξ_1 is sill value, and ξ_2 is range value. Figure 2 shows graph of spherical model which parameters are ξ_0=1, ξ_1=5, and ξ_2=10.

We then calculate the coordinate of point $\mathbf{P}_{0j}$ in jth direction. It is expressed by direction and the corresponding range value which are parameters of the fitted variogram model.

$$\mathbf{P}_{0j} = \begin{pmatrix} d_j \cos\theta_{0j} \\ d_j \sin\theta_{0j} \end{pmatrix} \tag{3}$$

$\mathbf{P}_{1j}$, a point on the ellipse line corresponding to $\mathbf{P}_{0j}$, is defined as follows.

$$\mathbf{p}_{1j} = \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} \begin{pmatrix} a\cos\theta_{1j} \\ b\sin\theta_{1j} \end{pmatrix} \tag{4}$$

with a semimajor axis a, semiminor axis b, and ψ which is degree of the angle of a semimajor axis. Where $\tan\theta_{1j} = \frac{a}{b}\tan\theta_{0j}$. We calculate the weighted

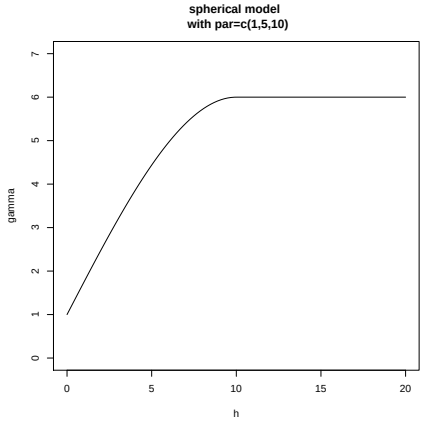


Fig. 2. Spherical model which parameters are ξ_0=1, ξ_1=5, and ξ_2=10.

least squares criterion, weighted regression sum of squared error of ellipse (WRSSE), by using $\mathbf{OP}_{0j}$ which is distance between $\mathbf{O}$ and $\mathbf{P}_{0j}$ and $\mathbf{OP}_{1j}$ which is distance between $\mathbf{O}$ and $\mathbf{P}_{1j}$

$$WRSSE = \sum_{j=1}^{ndir} \frac{np_j}{N} (\mathbf{OP}_{0j} - \mathbf{OP}_{1j})^2 \tag{5}$$

where N is the number of all pairs and np_j is the number of pairs in jth direction. We fit the parameters a, b and ψ optimizing WRSSE under conditions $a > 0$ and $b > 0$. Figure 3 shows four directional variograms and the theoretical variogram. And Figure 4 shows these corresponding to four directions (0, 45, 90, and 135 degree), as well as the graph of $\mathbf{P}_{0j}$ and the fitted ellipse.

The predicted value of the unobserved point is given by using the fitted variogram parameters, the ratio of ellipse parameters between a and b, and the rotating angle parameter of ψ.

Using this estimation method, we can obtain the ellipse parameters by parametric method, neither empirical nor interactive method which is adopted in *geoR*.

3 Applying to practical data

To ensure its validity we applied our method to water pollution data (155 points pH) of Okayama Prefecture in 2005. We used R environments and its library *geoR* and *gstat* to calculate and perform geostatistical analysis. In optimizing parameters of the ellipse we used R functions *optim* and *constrOptim*. Using cross validation we calculated the variogram parameters

and ellipse parameters from ith removed observed points of the pH data by the method described in section 2, and predicted ith removed points by Kriging. This prediction was performed for all observed points, and the predicted error calculated.

$$\sum_{i=1}^{155}(x(\mathbf{x}_i) - x(\hat{\mathbf{x}}_i))^2 \tag{6}$$

Cross validation was performed by changing the parameters' number of directions and tolerance values as follows.

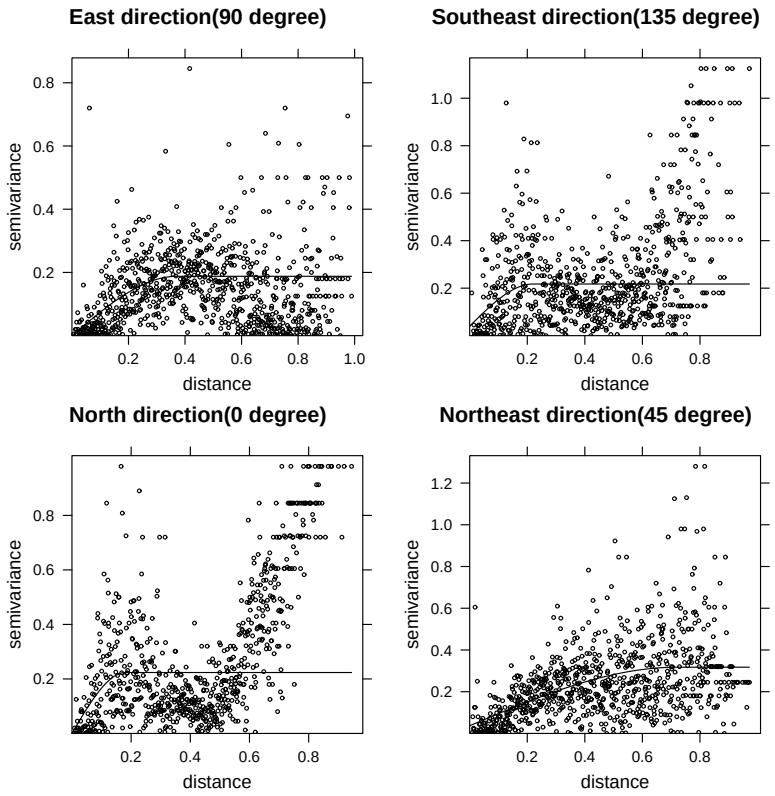


Fig. 3. Four directional variograms.

- *ndir*
  - 2
  - 4
  - 8
  - 16
- *tol*

– $90/ndir$
– $45/ndir$

Table 1 shows the result values equation (6) of cross validation.

Comparing the predicted error in Table 1 with the predicted error from the isotropic model which value of equation (6) was 7.33, we found that the model using geometric anisotropy had a smaller error than that of the isotropic model, not considering the results of the four directions.

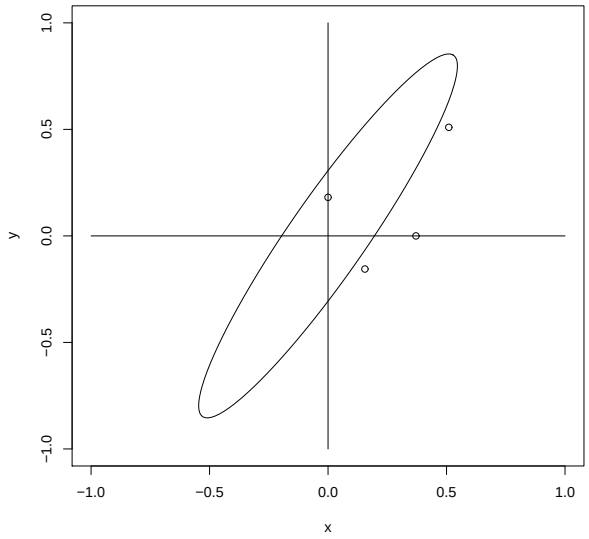


Fig. 4. The fitted ellipse.

ndir	2	4	8	16
$90/ndir$	7.08	7.40	6.72	6.68
$45/ndir$	6.66	7.68	6.73	6.73

Table 1. The result of cross validation.

We also found that when the number of direction was two and tolerance value was 22.5 degrees the lowest error occured in cross validation. We furthermore performed other cross validations to find better parameters (number of directions and tolerance values) by the following methods.

ndir	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$90/ndir$	7.08	6.93	7.40	7.44	6.95	6.76	6.72	6.59	6.68	6.65	6.77	6.79	6.83	6.77	6.68

Table 2. Fix *tol* to $90/ndir$ and change *ndir* from 2 to 16.

A. Fix *tol* to $90/ndir$ and change *ndir* from 2 to 16
B. Fix *ndir* to 2 and change *tol* from $90/ndir$ to $10/ndir$

Table 2 shows the result of A, and table 3 shows the result of B.

ndir	2
$90/ndir$	7.08
$80/ndir$	6.89
$70/ndir$	6.81
$60/ndir$	6.78
$50/ndir$	6.69
$40/ndir$	6.64
$30/ndir$	6.73
$20/ndir$	6.49
$10/ndir$	6.94

Table 3. Fix *ndir* to 2 and change *tol* from $90/ndir$ to $10/ndir$.

The result of A shows 9 directions had smallest error with fixed *tol* of $90/ndir$, and the result of B shows $20/ndir$(=10 degree) had smallest error with fixed *ndir* of 2. Figure 5 displays the map of predicted values of geometric anisotropic variogram with smallest error and that of isotropic variogram, and corresponding these errors.

4 Concluding remarks

We proposed to apply geometric anisotropy to geostatistical data and fitted to the ellipse model by the least square method. To ensure its validity we applied our method to water pollution data (155 points pH) of Okayama prefecture in 2005. Not considering the results of the four directions, we found the model using geometric anisotropy had a smaller error than the isotropic model. When *tol* was kept constant and *ndir* was changed the result of the case *ndir*=9 had smallest error, whereas if *ndir* was kept constant and tolerance was changed the result of the case tol=$20/ndir$ had the smallest error. These

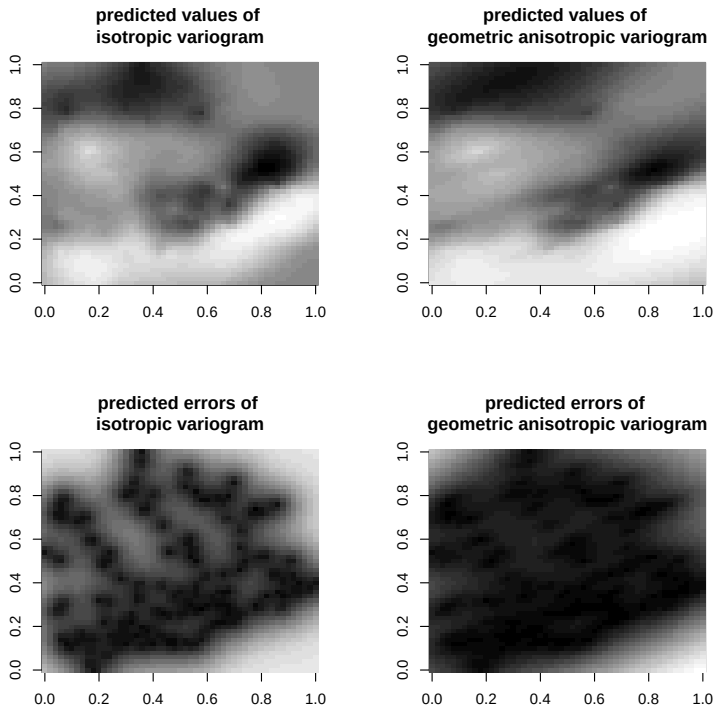


Fig. 5. The map of predicted values of geometric anisotropic variogram with smallest error and that of isotropic variogram, and corresponding these errors.

results represent that if *ndir* is too small or *tol* is too large it is difficult to identify the directional characteristics. In contrast, if *ndir* is too large or degree of *tol* is too small there are few pairs which correspond to particular directions, so fitted parameters become unstable.

The other type of anisotropy is zonal anisotropy which contains different sill values in different directional variograms. To model the zonal anisotropic variogram linear coupling of multiple isotropic variograms and/or geometric anisotropic variograms is proposed by Deraisme (2002). For further studies, we propose to apply zonal anisotropy to geostatistical data in fitting the ellipse model by the least square method in every geometric anisotropic variogram.

References

DERAISME, J.(2002): *Zonal Anisotropy: how to model the variogram?*. Newsletter #14 - February 2002 - Environment

ECKER, D.M. and GELFAND, E.A.(1997): *Spatial Modeling and Prediction Under Range Anisotropy*. Environmental and Ecological Statistics. 10, 165–178.

ECKER, D.M. and GELFAND, E.A.(1999): *Modeling and Inference for Geometrically Anisotropic Spatial Data*. Mathematical Geology 31, 67–83.

KUBOTA, T. IIZUKA, M. FUEDA, K. and TARUMI, T.(2005): *The Selection of the Cutoff in Estimating Variogram Model*. The 5th IASC Asian Conference on Statistical Computing, 97–100.

KUBOTA, T. and TARUMI, T. (2007): *Estimation of variogram model with anisotropy*. Bulletin of the International Statistical Institute 56th Session Proceedings

Part XXI

Statistical Software and Development Projects

Use' Evaluation of the TIC in Statistic Subjects Studied by the Students of Social Sciences

Miguel Á. Montero Alonso[1] and José A. Roldán Nofuentes[2]

[1] School of Social Sciences, Campus of Melilla, University of Granada
Avd. Alfonso XIII s/n, 52006 Melilla, Spain, *mmontero@ugr.es*
[2] Biostatistics, School of Medicine, University of Granada, Spain, *jaroldan@ugr.es*

Abstract. In this communication we want to evaluate the work done and the use of the new technologies (TIC) to involve students in a statistical teaching and to prepare professors as students to an agreed future towards which we are approaching with the adaptation to the European Space of Higher Education (EEES). In this work the created materials are evaluated as much as their use in diverse projects without finalizing, with the objective to improve the obtained results.

Keywords: TIC, *e*learning, teaching-learning process

1 Introduction

The creation of the European Space of Higher Education and the establishment of the system of credits ECTS have shown the necessity to introduce innovative changes and the necessity of an improvement in the *teaching-learning process*, which goes guided towards an independent performance of the students in which every day the initiative of work and the active incorporation of the students in their process of learning acquire bigger value (López,(2004)).

The academic course 2005/2006, the University of Granada (Ugr) in its Convocation of Innovation Educational Projects financed the project "APPLICATION OF NEW TECHNOLOGIES FOR THE ELABORATION OF MATERIALS TO SUPPORT THE UNIVERSITY STUDENTS OF SOCIAL SCIENCES" that a year before allowed to elaborate materials and create a virtual platform of teaching, a Learning Management System (LMS). With that call we tried to be a first passage in the adaptation of the virtualized subjects to the European Space of Superior Education, since the tools and methodologies of virtual teaching are directly applicable in many of the necessary tasks to teach a subject in the new European model's mark.

In the following course, and within the *Plan of Virtualizacin de Subjects for Academic Course* 2006/2007 of the Virtual Training Center of the University of Granada (CEVUG) - Secretaryship of Technologies for Support to

Teaching (STAD), a subject of Statistic in way *b-learning* [1] was taught with a pursued task, to improve the teaching quality by means of an semi-presential sessions in which another ways of teaching are distributed in the classroom, according to the traditional model, whereas another is based on denominated education online, virtual teaching or teleformation, in which the student carries out a process of the autolearning under professor's supervision. For it, they have didactic materials of support generated by the professor and accessible via Internet or by Learning Content Management System (LCMS). This way, the student can establish his own learning rhythm, dedicating in each moment the necessary time to assimilate each concept (Montero (2007)). Working into this plan, we teach the subject about Virtual Statistics and Tourism.

With both projects the process of changes is tried to harness the use of the TIC that is generating the processes of convergence to the EEES, change imposed by European supranational institutions that forces the states and their institutions to change structures, organizations and policies to adapt to Common Space. In the case of the TIC the change is obvious since it makes no sense to use the TIC to do just like it was already made without them, for that reason the TIC takes associated the change of opportunities, the opportunity to a better way of making things by different forms and when being associated both scenes, EEES and TIC, offer the opportunity to rethink the form in which the university students make their more genuine functions (Montero-Quesada (2007)).

A second aspect is that the change required from the reform of the EEES and the TIC has common elements: to facilitate the construction of the knowledge, to take the responsibility of the own learning and to have a greater control on the contents and activities, a possibility of collaborative work as much for the students as for the professor, (Batanero-Díaz (2007)) and (Davies et al.(2007)). We will always consider the advantages and disadvantages of virtual teaching, as for the professor as for the students, (Montero-Roldan, (2008)).

We want to give our vision and our experience about a new approached, showing some of the tools and made activities because until now not many studies have existed about how the professor sees the process of convergence and the incorporation of the TIC in it.

2 Contents

With this project we are persecuting to generate didactic materials to help the students of Social Sciences like complementary and/or alternative tools to the environment web already created for these students, thus continuing in this

[1] Blending Learning, educational modality in which combines the present formation and the on-line formation.

way to reinforce present teaching and to foment the active and autonomous learning of the students.

To carry out our project the tool used for the creation of the courses is **WebCT** of the University British Columbia. WebCT has a huge number of communication tools, contents, evaluations and studies, and in the same way, a flexible educational environment where the students can, besides learning, share experiences and knowledge with virtual communities composed by users of the system.

The contents that have been elaborated have been divided in the following big blocks:

- Traditional educational material: subjects of theory, problem relations and scripts of practices. All of them can be found by the students in web pages and pdf file, that can be lowered and be worked like and when they want. The theory notes have been published in manuals that are used and demanded by the students, without any costs for them.
- Material directed to the autolearning: tutorial videos on practices of computer and problems type made in Power Point.
- Material directed to the autoevaluation: autoevaluation question classified by topics and/or level of difficulty.

The students can find all elaborated material for Statistic subject in $http://cevug.ugr.es/asignaturas\_online.php$ or the web of the investigation group, $http://www.ugr.es/local/eues/webgrupo/index.html$, where necessary information for the student can be found, like dates of examinations, schedules of tutorship, etc.

This web page has been created with the objective to continue including more professors and therefore more subjects, more materials to improve the process of teaching-learning of the students and where it can act in an autonomous way acquiring bigger work initiative.

3 Evaluation of the project

We understand that the first phase of the developed innovation project concluded in May, end of the past academic course, the evaluation and valuation of the tool corresponds the present to the students in course, although some of them participated in an initial evaluation which we will comment next, but, in general, we believe it had a good welcome.

Anyway, the professors implied in this educational experience fixed the following main objectives with the purpose of evaluating the degree of advantage or use of the same one:

- To value the degree of improvement in the derived learning from the existence and use of the materials of created support to teaching.

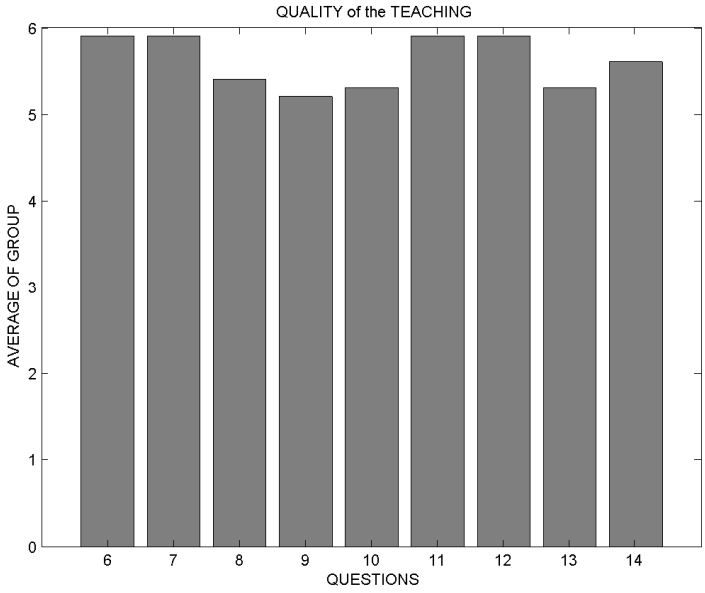


Fig. 1. Evaluation of the Quality of Education.

- To evaluate the perception that the students have about their didactic utility, as well as of the different curricular materials including.
- To estimate the degree of satisfaction of the students to improve and elaborate new materials.

The degree of attainment of these objectives will be valued from the results of a poll between the student-users, which reflects on these issues, wonders what utility and view them deserves and they put on in evidence the deficiencies observed for possible improvements in the future.

Likewise, we try to objectively assess the influence of the innovative teaching methodology introduced this academic year in the obtained results, considering the total number of approved, the average rating of the groups, the relative dispersion of qualifications, etc. These measures will allow us, among others, comparing results between groups of students from previous academic courses involved or not in this type of educational experience and to compare results to extract the opportune conclusions.

If we see Figures 1 and 2, obtained from an initial assessment carried out to the students, it can be seen that the valuation can be considered good, but we have also collected a number of issues that the students have proposed and, of course, we believe that an improvement is needed.

In each figure we can appreciate some questions asked to the students and the evaluation of the group regarding to each represented question. The

questions in Figure 1 were about the quality of the teaching and they were the following ones:

Q1. The professors / tutors dominate the subject?
Q2. Have the answers emitted by the professor in the resolution of doubts been quick and clear?
Q3. The professor has made an appropriate use of illustrations and examples.
Q4. The students have implied themselves and they have participated in the course (construction of contents, debates, elaboration of materials, etc.)
Q5. Has the motivation, on the part of the students, toward the course been high?
Q6. Have the tutors been available to solve the students' doubts?
Q7. Has the interaction between profesor/tutor-student been flowed?
Q8. Has the individual learning been instigated as much the collaboration in learning groups?
Q9. Has the student's active rol been incited in order to be a participant in the course?

Questions of the Figure 2 were related to the activities and the materials elaborated. Those questions were the following ones:

Q15. Have the activities and resources used in the course been helpful to reach the objectives?
Q16. Among all the activities, Has the study of real cases which are close and related to the student's professional field been included?
Q17. Have the techniques and the evaluation procedures used been in consonance with the objectives of the course?

We still have a lot of work to do, we are in an early stage, but we believe that the results and reception has been good, although we consider that we must continue and improve the undertaken work line.

4 Academic results

The new technologies of communication open new representation windows and transmission of the mathematical-statistical knowledge. Foregone, the advances in computer technology and communication are following one another are going to suppose, are supposing already, the possibility of creating new virtual spaces that make possible a new didactic utility.

Professors in the virtual environment are no longer direct instructors and they become facilitators, providing tools and tips to students to help them develop their own learning process, while addressing their concerns and needs. The tutor becomes the person in charge of having a direct contact with the student, predisposing him and advising him in following up the study materials, and therefore it is the figure that must avoid the non motivation and the student's abandonment in his auto learning process.

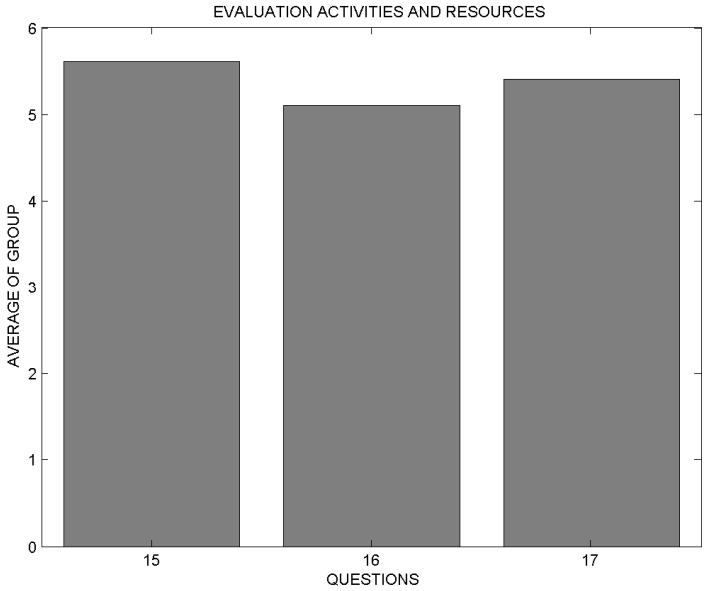


Fig. 2. Evaluation of the Activities and Used Resources.

We have obtained important profits that we will assist and motivate us to continue and improve this education model. These results, in our view, quite important we highlight next:

- It has significantly increased the number of registered students, like you can see in the Figure 3. In the last course has been 13 registered students compared to 3 or 4 of previous years. In fact, it has been well received from registered students in degree programs outside the E.U. Social Sciences, with 30% of them[2]. In the present course are 15 students, two more than last course and there are even *Erasmus* students.
- In the last course, the rate of abandonment of the subject has improved considerably, since only 2 people have not connected themselves neither made the tasks assiduously, that is to say, 85% of the registered students has worked the subject, has presented/displayed its weekly exercises, they have made its examination, and therefore they have surpassed the subject in the June convocation. All this thanks to the discussion forums, chat, support, put in common with the partners, and thanks to the good atmosphere that has arisen in the virtual class.

To emphasize, from our point of view, the student not only learned statistical knowledge, but they also, as we think have had a tool which has served

[2] In this project the University of Granada allows a maximum of 50 students, but it is recommended that this is around 25 students.

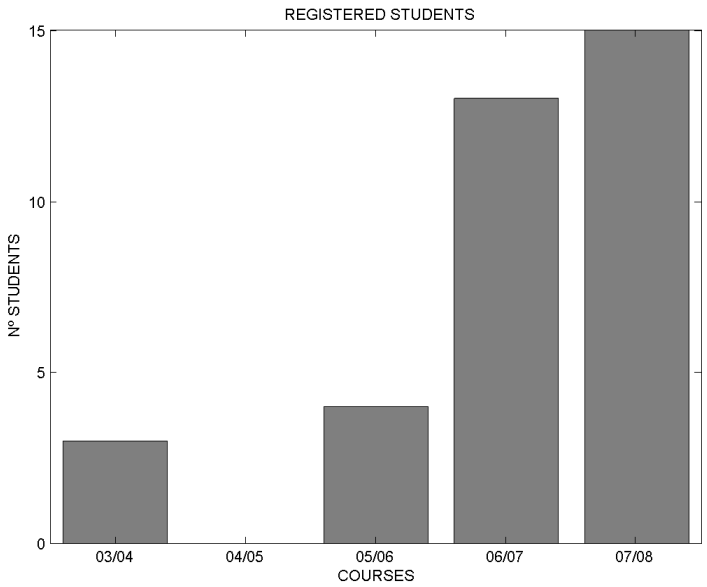


Fig. 3. Evolution of the Number of Registered Students.

as guide in the way of the learning, enriching the process through the use of the technologies, fomenting a renovated process and renovating of teaching-learning, where have been developed competitions that will allow to form individuals for a learning throughout life.

On the other hand, we could not finish without valuing the opinion of students, who really are the main link. For it a poll on the part of the CEVUG of the University of Granada was made during the first week of June of the last year, where diverse questions on the subject, contents, presentation or the work of the tutor became, among others. This poll can be found in $http : //cevug.ugr.es/gestion/encuestas\_cevug/muestra\_estadisticas.php?curso = 201\&submit = Estadisticas$. We emphasize that in a scale from 1 to 6, the global valuation of the course has been of 5.3, which makes us think that we are doing it well, even so, given that everything can be improved, it will be necessary to polish those small deficiencies that we have found in this first course.

Therefore, we conclude that the incorporation of TIC in university teaching is one of the strategic priorities in the methodological environment according to the Proposals for the Renovation of the Educational Methodologies in the Spanish University published recently by the Ministry of Education (MECD, 2006). Our present objective should be to update us in this methodology that uses the available resources to foment the autonomous learning of the students, establishing a continuous guide, at the same time that to cover

necessities of those cannot habitually attend class (for work reasons, for coincidence time or any other eventuality).

In the on-line teaching, after the initial moments of contact with the virtual environment, the greater or smaller degree of the student's participation won't depend so much on the technological aspects as on the methodological, supposing the student's initial interest for the matter object of the virtual course. This one will be, the main starting point for a new educational methodology centered in the student, in front of the traditional model of masterful classes in which the center of the teaching process is the professor.

References

BATANERO, C., DÍAZ, C. (2007): Training future statisticians to teach Statistics. In: M.Ivette Gomes, D. Pestana, P. Siva, (Eds.), *Proceedings in ISI 2007*, Lisbon, Portugal.

CRAMER, E., CRAMER, K., KAMPS, U. (2002): e-stat: A web-based learning environment in applied statistics. In: W. Hrdle, B. Rnz (Eds.) *COMPSTAT 2002 Proceedings in Computational Statistics*, Physica, Heidelberg 309-314.

DAVIES, N., MARRIOTT, J., GIBSON, L. (2007): Solving the problem of teaching statistics?. In: M.Ivette Gomes, D. Pestana, P. Siva, (Eds.), *Proceedings in ISI 2007*, Lisbon, Portugal.

LÓPEZ, A.J. (2004): El papel del e-Learning en el Espacio Europeo de Educacin Superior, In: *Congreso Online Educa*, Madrid.

MECD (2006). *Propuestas para la Renovación de las Metodologías Educativas en la Universidad*. Madrid: SGT.

MONTERO, M.A.(2007): Enseanza de Estadstica en un entorno virtual. *Revista de Informtica Educativa y Medios Audiovisuales 4, (4), 1-6.*

MONTERO, M.A., QUESADA, I. (2007): Ensear Estadística bajo un entorno Virtual. In: *XXX National Congress of Statistic and O.R.*, Valladolid, Spain.

MONTERO, M.A., ROLDÁN, J.A. (2008a): Statistic *e*-Learning. In: Monica Kovácová (Eds.): *Aplimat - Journal of Applied Mathematics*, Bratislava, Slovakia.

MONTERO, M.A., ROLDÁN, J.A. (2008b): New tendencies in the Teaching and the Learning in Statistic. In: *8th International Conference on Operations Research*, Havana, Cuba.

SYMANZIK, J., VEKANUSOVIC, N. (2002): Teaching Statistics with electronic textbooks. In: W. Hrdle, B. Rnz (Eds.) *COMPSTAT 2002 Proceedings in Computational Statistics*, Physica, Heidelberg 79-90.

SYMANZIK, J., VEKANUSOVIC, N. (2006): Teaching an introductory Statistics course with CyberStats, an electronic textbook. *Journal of Statistics Education*, Volume 14, Number 1, www.amstat.org/publications/jse/v14n1/symanzik.html

Fast Text Mining Using Kernels in R

Ingo Feinerer[1] and Alexandros Karatzoglou[2]

[1] Theory and Logic Group
Institute of Computer Languages
Vienna University of Technology
Austria, *feinerer@logic.at*
[2] LITIS, INSA de Rouen
Avenue de Universite
76801 Saint-Etienne du Rouvray
France, *alexis@ci.tuwien.ac.at*

Abstract. Recent advances in the field of kernel-based machine learning methods enable the fast processing of text using string kernels which are built with the use of suffix arrays. `kernlab` provides both kernel methods infrastructure and a large collection of already implemented algorithms and includes an implementation of suffix array based string kernels. Along with the use of `tm` these packages provide `R` with functionality in processing, visualizing and grouping large collections of text data using kernel methods. We focus on the performance of various types of string kernels at these tasks.

Keywords: string kernels, text mining, clustering, `R`

1 Introduction

Kernel-based methods are state of the art machine learning methods who have gained prominence mainly through the successful application of Support Vector Machines (SVMs) in a large domain of applications. SVMs have also been applied in classification of text documents using typically the inner product between two vector representations of text documents.

String kernels (Watkins (2000), Herbrich (2002)) have proven to be a promising alternative providing excellent results in text classification using SVMs and clustering using kernel based clustering methods like e.g. spectral clustering. One of the main drawbacks in the application of string kernels in large document collections is that until recently most algorithms and implementations where both slow and scaled usually in $O((n+m)^2)$ where n and m are the number of characters in the two documents. Vishwanathan and Smola (2004) introduced string kernels based on suffix trees which scaled $O(n+m)$ but had still drawbacks since the construction and use of a suffix tree requires a large amount of memory (typically $40n$) and has poor locality. Suffix trees based string kernels scale in theory in $O(n)$ but have been proven to be relatively slow for large scale text processing due to this inherent weakness. The use of suffix arrays (Teo and Vishwanathan (2006)) resolved most

of these issues and has transformed string kernels into a very useful option for large scale text mining using kernel-based machine learning.

In this paper we will demonstrate the viability of kernel-based machine learning for text mining using suffix array based string kernel implementations in the `kernlab` (Karatzoglou et al. (2007)) R (R Development Core Team (2008)) package and we will benchmark these string kernels. We will also briefly present the `tm` (Feinerer (2007)) R package which provides text mining functionality for R.

2 String kernels using suffix arrays

String kernels work by calculating a weighted sum over the common substring between two strings, as shown in Equation 1. Different types of kernels arise by the use of different weighting schemas. The generic form of string kernels between two sets of characters x and x' is given by the equation

$$k(x, x') = \sum_{s \sqsubseteq x, s' \sqsubseteq x'} \lambda_s \delta_{s,s'} = \sum_{s \in A^*} \mathrm{num}_s(x) \mathrm{num}_s(x') \lambda_s \ , \tag{1}$$

where A^* represents the set of all non empty strings and λ_s is a weight or decay factor which can be chosen to be fixed for all substrings or can be set to a different value for each substring. This generic representation includes a large number of special cases, e.g. setting $\lambda_s \neq 0$ only for substrings that start and end with a whitespace character gives the "bag of words" kernel. In this paper we consider four different types of string kernels:

- Constant (constant): All common substrings are matched and weighted equally.
- Exponential decay (exponential): All common substrings are matched but the substring weight decays as the matching substring gets shorter.
- k-spectrum (spectrum): This kernel considers only matching substrings of exactly length n, i.e. $\lambda_s = 1$ for all $|s| = n$.
- Bounded range (boundrange) kernel where $\lambda_s = 0$ for all $|s| > n$ that is comparing all substrings of length less or equal to a given length n.

String kernels can be computed by building the suffix tree of a string x and computing the matching statistics of a string x' by traversing string x' through the suffix tree of x. Given a suffix tree $S(x)$ it can be proven that the occurrence of a certain substring y can be calculated by the number of nodes at the end of the path of y in the suffix tree. Auxiliary suffix links, linking identical suffixes in the tree are utilized to speed up the computations. Two main suffix tree operations are required to compute string kernels, a top down traversal for annotation and a suffix link traversal for computing matching statistics, both operations can be performed more efficiently on a suffix array.

The enhanced suffix array (Abouelhoda et al. (2004)) of a string x, is an array of integers corresponding to the lexicographically sorted suffixes of x

with additional information stored to allow for the reproduction of almost all operations available on a suffix tree. Suffix arrays bring the advantage of better memory use and locality thus most operations can be performed faster than on the original suffix trees.

3 R infrastructure

R provides a unique environment for text mining but has until recently lacked tools that would provide the necessary infrastructure in order to handle text and compute basic text related operations, e.g. the computation of a term matrix. Package tm provides this functionality.

3.1 tm

The tm package provides a framework for text mining applications in R. It offers functionality for managing text documents, abstracts the process of document manipulation and eases the usage of heterogeneous text formats in R. The package has integrated database backend support to minimize memory demands. An advanced metadata management is implemented for collections of text documents to alleviate the usage of large and with metadata enriched document sets. Its data structures and algorithms can be extended to fit custom demands, since the package is designed in a modular way to enable easy integration of new file formats, readers, transformations and filter operations. tm provides easy access to preprocessing and manipulation mechanisms such as whitespace removal, stemming, or conversion between file formats. Further a generic filter architecture is available in order to filter documents for certain criteria, or perform full text search.

3.2 kernlab

kernlab is an extensible package for kernel-based machine learning methods in R. The package contains implementations of most popular kernels, and also includes a range of kernel methods for classification, regression (Support Vector Machine, Relevance Vector Machine), clustering (kernel k-means, Spectral Clustering), ranking, and Principal Component Analysis (PCA).

We shortly describe two of the lesser known kernel methods we are using for our experiments.

Kernel k-means The k-means clustering algorithm has the drawback that it cannot separate clusters that are not linearly separable in input space. One technique for dealing with this problem is mapping the data into a high-dimensional non-linear feature space with the use of a kernel. Denoting

clusters by π_j and a partitioning of points as ${\pi_j}_{j=1}^k$ and if Φ is the mapping function then the k-means objective function using Euclidean distances becomes

$$\mathcal{D}({\pi_j}_{j=1}^k) = \sum_{j=1}^{k} \sum_{a \in \pi_j} \|\Phi(a) - m_j\|^2 \ , \tag{2}$$

where $m_j = \frac{1}{\|\pi_j\|} \sum_{a \in \pi_j} \Phi(a)$ and in the expansion of the square norm only inner products of the form $\langle \Phi(a), \Phi(b) \rangle$ appear which are computed by the kernel function $k(a, b)$.

Spectral clustering Spectral clustering (Ng et al. (2001), Shi and Malik (2000)) works by embedding the data points of the partitioning problem into the subspace of the k largest eigenvectors of a normalized kernel matrix. The data is then embedded into the subspace of the largest eigenvectors of the normalized kernel matrix. This embedding leads to more straightforward clustering problems since points tend to form tight clusters in the eigenvector subspace.

4 Experiments

4.1 Data

Our first dataset is a subset of the Reuters-21578 dataset (Lewis (1997)) containing stories collected by the Reuters news agency. The dataset is publicly available and has been widely used in text mining research within the last decade. Our subset contains 800 documents in the category "acq" (articles on acquisitions and mergers) and 400 in the category "crude" (stories in the context of crude oil). In case of clustering this is the full dataset, for classification this is the training set. The classification test set are further 200 acquisition and 100 crude oil stories.

The second dataset is a subset of the SpamAssassin public mail corpus (`http://spamassassin.apache.org/publiccorpus/`). It is freely available and offers authentic e-mail communication with classifications into normal (ham) and unsolicited (spam) mail. For the experiment we use 800 ham documents and 400 spam documents for clustering and as training set for our classification tasks. For the latter the test set consists of additional 100 ham and 100 spam documents.

4.2 Experiment setup

The experiment setup is twofold. First, we performed clustering methods on the two datasets Reuters and SpamAssassin. We use kernel k-means and spectral clustering algorithms implemented in the `kernlab` package. We manually set the number of desired clusters to two (i.e., $k = 2$) for both datasets since

in both cases we know the true labels for each document (for the Reuters set the categories acquisitions and crude oil, for the SpamAssassin set ham and spam, respectively). Secondly, we perform a "C-SVC" classification on both datasets. Again we use the SVMs in package `kernlab`.

We used four types of recent string kernels: exponential, constant, spectrum, and boundrange. We conducted runs with various values of the string length parameter used by the spectrum and bounded range kernel. We evaluate the quality of the computed cluster and classification results via the cross-agreements between the cluster or classification labels and the true labels of each dataset. In addition we measured the consumed computation time for each step. In case of clustering these are the timings for creation and computation of the string kernel matrix (denoted as Kernel Time) and the actual clustering process (denoted as Cluster Time). For classification we logged the time for string kernel matrix creation (Kernel Time), the amount of time necessary for training the SVM (SVM Training Time), and the spent time for predicting the test documents (Predict Time).

4.3 Results

Type	Length	Agree	Cluster Time	Agree	Cluster Time	Kernel Time
Algorithm		k-means		Spectral		
exponential		0.6863	1.323	0.6891	3.507	558.400
constant		0.5141	0.426	0.7958	3.604	514.080
spectrum	4	0.7766	1.346	0.8408	3.563	522.073
spectrum	6	0.8641	1.597	0.8575	3.374	519.446
spectrum	8	**0.8641**	2.365	0.8566	3.618	521.971
spectrum	10	0.6833	2.807	**0.8683**	3.669	519.317
boundrang	4	0.6900	1.472	0.6208	3.643	520.817
boundrang	6	0.6491	1.852	0.6975	3.665	520.981
boundrang	8	0.6991	1.495	0.7700	3.532	522.097
boundrang	10	0.7283	2.903	0.7816	3.486	519.738
fullstring	8			0.8266	3.851	2195.795
string	8			0.8558	3.652	4521.459

Table 1. Cross-agreement and timing results for kernel k-means and spectral clustering on the Reuters data for various types of string kernels set and different values for string length parameter. The string and fullstring kernels are dynamic programming based implementations of kernels similar to the spectrum and boundrange kernel. They are slower by a factor of 4 and 8 respectively.

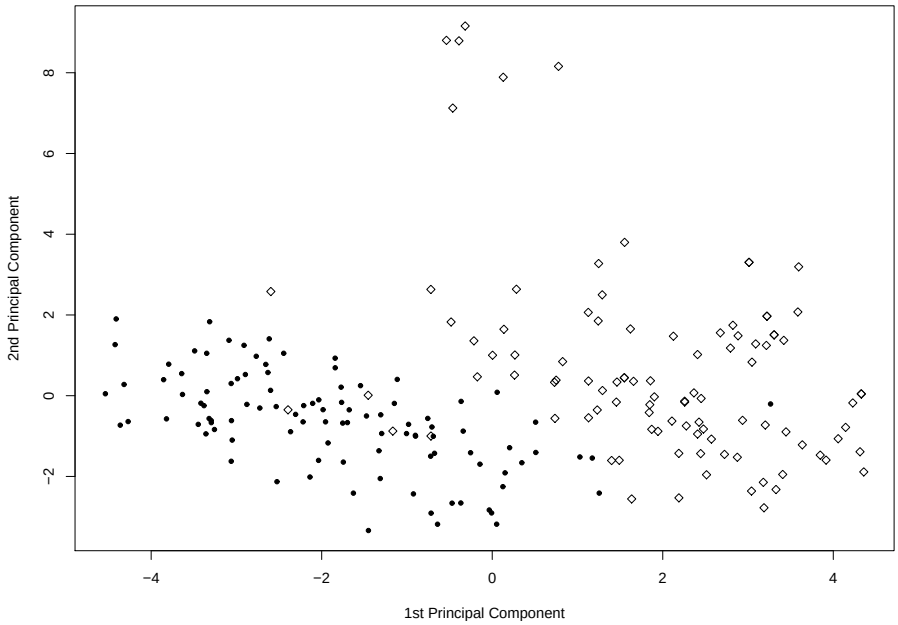


Fig. 1. The projection on two principal components of 200 Reuters text documents (100 crude in dots and 100 acquisition in diamonds) using a spectrum kernel and kernel PCA.

Type	Length	Agree	Cluster Time	Agree	Cluster Time	Kernel Time
Algorithm		k-means		Spectral		
exponential		0.7358	2.372	0.6183	3.393	1765.602
constant		0.6170	2.535	0.5991	3.635	1674.338
spectrum	4	0.6525	1.901	**0.7491**	3.351	1675.842
spectrum	6	0.6366	1.034	0.6691	3.170	1673.387
spectrum	8	0.6366	1.688	0.6691	3.147	1677.165
spectrum	10	0.5891	1.160	0.6675	3.170	1678.230
boundrang	4	**0.7750**	1.154	0.7200	3.343	1670.671
boundrang	6	0.7583	1.314	0.7300	3.353	1672.437
boundrang	8	0.7541	1.280	0.6175	3.359	1670.454
boundrang	10	0.7450	1.987	0.6158	3.378	1679.011

Table 2. Cross-agreement and timing results for kernel k-means and spectral clustering on the SpamAssassin dataset under different string kernel parameters.

Type	Length	Agree	Predict Time	Kernel Time	SVM Training Time
exponential		0.9633	0.007	562.769	0.115
constant		0.8333	0.010	516.054	0.595
spectrum	4	**0.9900**	0.005	525.716	0.102
spectrum	6	0.9800	0.006	525.044	0.128
spectrum	8	0.9766	0.007	527.880	0.141
spectrum	10	0.9566	0.008	523.306	0.168
boundrang	4	0.9600	0.007	534.470	0.118
boundrang	6	0.9666	0.006	520.825	0.117
boundrang	8	0.9633	0.008	524.199	0.108
boundrang	10	0.9633	0.007	519.974	0.114

Table 3. Cross-agreement and timing results for SVM classification on the Reuters dataset under different string kernel parameters.

Type	Length	Agree	Predict Time	Kernel Time	SVM Training Time
exponential		0.9400	0.005	1740.865	0.097
constant		0.8850	0.006	1644.471	0.199
spectrum	4	**0.9700**	0.004	1667.624	0.112
spectrum	6	0.9500	0.005	1667.321	0.127

Table 4. Cross-agreement and timing results for SVM classification on the SpamAssassin dataset under different string kernel parameters.

From the results of our experiments on the Reuters dataset (see Table 1) we can conclude that the spectrum kernel could be a generally good choice for processing text documents. It performs well overall compared to the alternative kernels. This can be attributed to the fact that it considers matches of exact length k while the other kernels match also all substrings of length smaller than k thus introducing additional information into the kernel matrix that might not be of benefit in the case of rather lengthy text documents where the probability of matches will be high and thus additional noise is introduced.

The results of the experiments on the SpamAssassin dataset (see Table 2) show that it is important to also consider the nature of the text that is being processed. In the case of the SpamAssassin dataset alternative kernels types seem to perform on the same level if not better than the spectrum kernel. This could be attributed to the particular composition of spam messages where short words occur more frequently and arbitrary text and string sequences are added into the messages, to bypass filters. Kernels such as the

boundrange kernel will capture more of this information. The spectral clustering algorithm performs in a slightly more consistent way across different kernel and parameter configurations than the kernel k-means although the later seems to outperform the former for some configurations.

The results in Table 3 and Table 4 confirm the excellent performance of SVMs in text classification when using string kernels. Note that until recently string kernels have been computationally expensive and this prohibited applications in many real world datasets. This constraint has been lifted as the comparison of the performance of the suffix based kernels with a typical dynamic programming implementation shows. The new version of the kernels outperforms the older on the Reuters dataset by a factor of 4 to 8 depending on the kernel type.

Figure 1 illustrates the projection of 200 text documents from the Reuters dataset on two principal components using kernel PCA. This type of visualization can be used as an explorative tool in order to e.g. test for the existence of clusters in a set of documents.

4.4 Conclusion

We presented a fast implementation of string kernels with excellent performance in clustering and classification using spectral clustering and support vector machines. The availability of the kernels in `kernlab` along with the text processing infrastructure in the `tm` package provides R with advanced algorithms for the processing of text documents for the practitioner. In our experiments we demonstrated the excellent performance of the kernel both in term of computational cost and results and showed that the spectrum kernel is a good choice when working with normal text data while this might change when using non-text related string data like e.g. DNA sequences.

References

ABOUELHODA, M.I., KURTZ, S., OHLEBUSCH, E. (2004): Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, 2, 53–86.

FEINERER, I. (2007): tm: Text Mining Package: *CRAN tm version 0.3*.

HERBRICH, R. (2002): *Learning Kernel Classifiers Theory and Algorithms* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

KARATZOGLOU, A., SMOLA, A. and HORNIK, K. (2007): kernlab: Kernel Methods Lab *CRAN kernlab version 0.9-5*.

LEWIS, D. (1997): Reuters-21578 text categorization test collection.

NG, A., JORDAN, M. and WEISS, Y. (2001): On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems, 14*.

R Development Core Team (2008): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

SHI, J. and MALIK, J. (2000): Normalized cuts and image segmentation. *Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888–905*.

TEO, C.H. and VISHWANATHAN, S.V.N. (2006): Fast and space efficient string kernels using suffix arrays. *Proceedings of the 23rd International Conference on Machine learning (ICML)* ACM Press, Pittsburgh, Pennsylvania, 929–936.

VISHWANATHAN, S.V.N. and SMOLA, A.J. (2004): Fast Kernels for String and Tree Matching. In B. Schölkopf and K. Tsuda and J. P. Vert (Eds.): *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, 113–130.

WATKINS, C. (2000): Dynamic Alignment Kernels. *Advances in Large Margin Classifiers*. A.J. Smola and P. L. Bartlett and B. Schölkopfand D. Schuurmans. MIT Press, Cambridge, MA, 39–50.

e-status: a Problem-based Learning Web Tool Powered by R

José A. González, Lluís Marco, Lourdes Rodero, and Josep A. Sánchez

Universitat Politècnica de Catalunya
Statistics and Operations Research Dept., 08034 Barcelona, Spain,
{*jose.a.gonzalez — lluis.marco — lourdes.rodero — josep.a.sanchez*}*@upc.edu*

Abstract. This paper describes some features of the e-status tool, devoted to help the students by providing them with exercises that are corrected automatically. e-status is available via web to all the students of the course, and it uses internally the R environment for the needed computations. Unlike many other web resources, e-status has been conceived as a versatile and dynamic tool: as a consequence, the students making use of it can benefit of valuable feedback and can improve their knowledge of the subject.

Keywords: statistical education, learning-software, R, web-based tools

1 Introduction

e-status is a web-based platform designed to help the students in their courses, specifically when practical exercises are a substantial component for the students' learning. This framework can fit well into statistics subjects, especially in applied studies like engineering or medicine, strongly based in techniques of problem solving and case studies. Its current address on the internet is http://ka.upc.es/.

The tool is a computer environment able to provide exercises to the students and correct their answers automatically. The exercises normally come with randomly generated data; an analytical model built by the problem creator (usually a teacher) allows assessment of correct answers, and even providing information related to predictable errors. It is important to keep in mind that each exercise solved is logged in a database.

In order to accurately explain our goals, the term *problem* should be redefined. *Problem* is meant as the whole structure being composed mainly of:

- the description of a supposed situation;
- the model, written with instructions of the statistical language R;
- the questions, with additional information used to assess the answers.

Thus, the exercises seen by the students are just particular cases of some general problem. This can be useful to allow repetition of the topic, or to avoid students cheating, for instance.

As said above, e-status keeps all the work done in a database. Therefore, both the students and teachers have precise information and feedback about the work carried out. e-status shows the user statistics related to students performance in a variety of ways: summarized or detailed, tabulated or graphically, relative to class groups, problems or students, etc.

In this paper we are not presenting a concrete, complete, online course of statistics. The list of internet-based learning tools would be lengthy, but e-status does not exactly match the general approach. Our aim is to briefly present the features of the tool, designed to be: (1) flexible and versatile, allowing the creation of absolutely diverse problems; and (2) able to provide data and feedback that could improve the learning process of the students.

2 Development

The e-status project began in 2002, announcing the broad outline of its features by González and Muñoz (2002). Later, didactical background was described in detail to emphasize the importance of the methodology, see González and Muñoz (2006). The application was put in practice in 2003 with students from the Barcelona School of Informatics (FIB), although this phase revealed the need for some improvements, so that the tool would became really operational and useful to the students. For instance, it was obvious that the students had to identify themselves to enter the application, preferably with their current password. This could be done for FIB students, but users from other schools had to receive a randomly created password by e-mail.

At the beginning, e-status included Java code, because the problems were not modeled in R, but in a simple language especially designed for the tool, similar to R though evidently limited in its functions. Problems in the early versions of e-status were edited with a Java stand-alone application that was eventually integrated into the web tool, using servlets (based on Java as well). However, Java technologies were given up later, evolving to a simpler software environment supported only by the PHP language.

The most important innovation that has been applied is the inclusion of R (2006) as the language used for computations. The necessity of modeling problems with more capabilities and less effort required consideration of another language to replace the original one. R stood out as the best candidate: we took into account an open source environment, widely known in the statistics community for its computational and graphical capabilities, a powerful and robust language, and an extensive collection of packages freely developed by a large number of contributors.

In the Summer of 2007 a new version including significant changes was installed in the server, and it has been used since September by several hundreds of students without any noteworthy incident. It is remarkable that the early method to assess an answer (each question has only one solution) has been replaced by the inclusion of procedural criteria, written with R instruc-

tions. This change greatly extends the power of the tool, since it allows the teachers to process the student answer, considering anything that could be expressed with a programming language.

3 Architecture

Figure 1 shows the main elements of the computer architecture of e-status. They are:

(i) The web server (ka.upc.es), where the application is hosted as well as the R software and Rserve, though they could be running on another server;
(ii) The LDAP server, to authenticate the users;
(iii) The database server, where the data is saved;
(iv) Finally, the internet may be considered as another element, since it is the bridge communicating the user with the application.

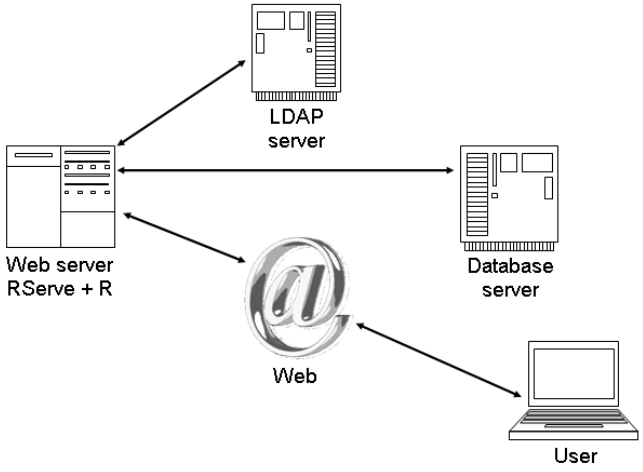


Fig. 1. Some elements of the architecture of e-status.

e-status is conceived as a web service, that is, users can access the tool by means of an internet client, such as Netscape, IExplorer or Mozilla Firefox, and an internet connection, not necessarily with large bandwidth. Several information technologies have been used during its life (for instance: at the beginning it ran on a machine with a Microsoft Operating System). At the present, e-status is hosted on a Linux server with Apache, and it is entirely written in the PHP language, more precisely CakePHP, a structured framework that aims to help PHP users to rapidly develop robust web applications.

The communication between e-status and R is made by means of the Rserve software (Urbanek, 2003), a package which allows binary requests to

be sent to R. The client-side implementation for the language PHP was developed for the e-status project. This way, each user has a separate workspace without needing to initialize R.

The first versions of e-status used SQLServer as a database management system (DBMS) but at present it has been replaced by MySQL, a popular open source DBMS, closely tied to the success of PHP and Apache, two of the main components of e-status.

Finally, we added to the 2007 update the capability of connecting to LDAP servers (Koutsonikola and Vakali, 2004), allowing identification of users present in the directory, by means of their own password.

At present and from now on, e-status is entirely based on open source software, so it is the authors' wish to make the platform easily accessible to the academic community. We are working to find a means for simple and independent installation on any system. Parenthetically, any researcher can try the tool on our server by requesting a teacher's account from the authors.

4 Functionalities

Two main profiles can interact with the application: teacher and student. In broad terms, teachers build problems that are solved by the students, and teachers monitor their results also. In fact, the division is not so sharp: teachers can solve a problem as a student, and students can access some non-revealing information from their colleagues. However, the distinction between the teacher and student profile is useful for explanation purposes.

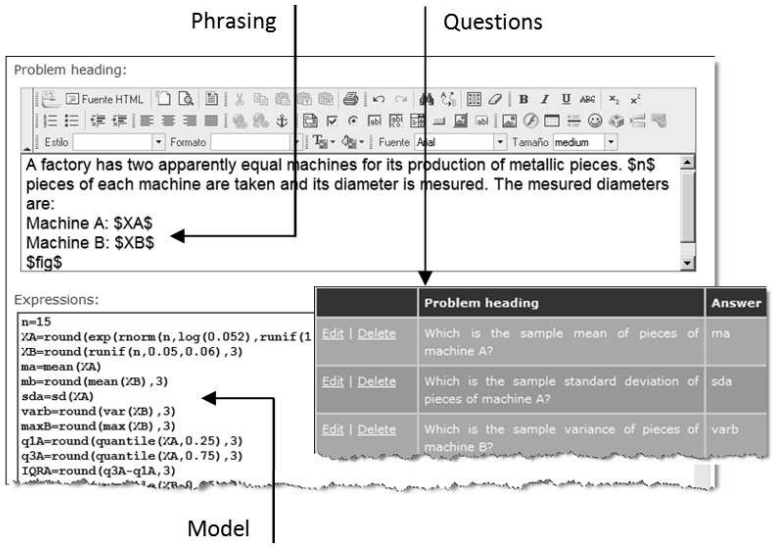


Fig. 2. The three elements of a problem: Phrasing, Model and Questions.

From the teacher's point of view, the application allows the creation of problems for the students. Problems have three parts: Phrasing, Background Model and Questions (Figure 2), as well as secondary elements like *description*, *difficulty*, *author*, etc. Statements for the exercises and questions are written using text processor utilities already embedded in the existing text boxes. Any variable or graphic from the statistical model can be included in these statements, writing them between two dollar signs (see an example in Figure 2).

The statistical model is created using R instructions. Including random generation of data among the R code guarantees that the dataset for the exercise will change every time the problem is requested. All the power of R is available for defining the model behind the problem, and any needed package can be installed and used. The teacher chooses the variable that contains the answer for each question and the amount of feedback to the student (for example: returning some hints for another try depending on the student answer, or showing the correct answer when every try failed).

On the other hand, students can solve exercises repeatedly with different data every time. It is not necessary for the student to know the R code which works in the background, generating the numerical part of the exercise. The existence of R behind the platform is thus transparent to the student. The sequence followed by students when solving an e-status problem is described next (Figure 3):

(i) Phrasing and questions are shown with data generated by the model;
(ii) The student writes the answers in the text boxes and sends the problem for being revised;

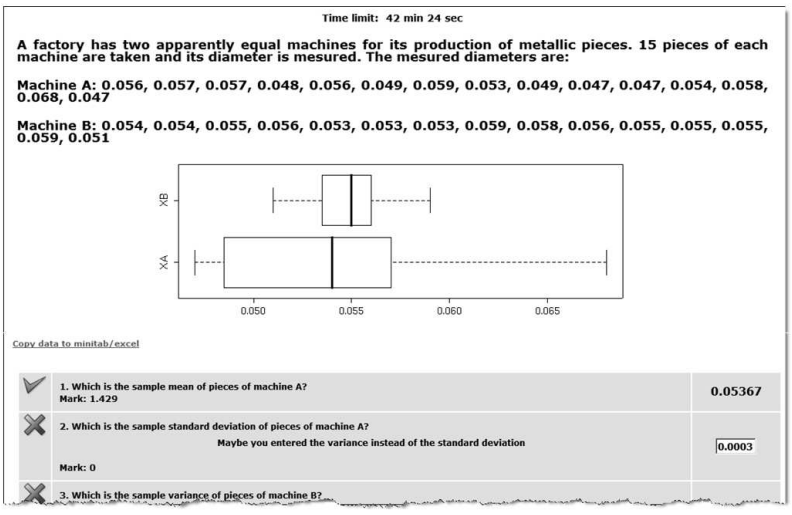


Fig. 3. The problem view for a student.

(iii) Answers are automatically revised, comparing the correct answer calculated by the R code with the answers given by the student; more generally, the student answer is contrasted with some procedure to assess its validity, since perhaps the question does not have a uniquely correct answer;

(iv) If chosen by the teacher, hints are returned for wrong answers and more tries are allowed.

Each execution of a problem is registered so that both teacher and student can have a registry of the work done. Teachers know the intensity of use for each exercise, the score and even each response for every exercise solved; all of them are indicators of possible weaknesses in the students' learning process. Students can schedule their activities and view their performance compared with other students (Figure 4).

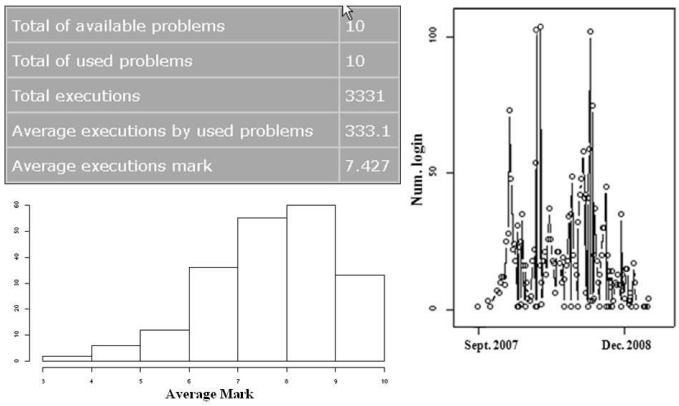


Fig. 4. Information about student activity.

The e-status platform has some extra functionality useful for an interactive student experience, such as polls and statistical calculators with some common functions. They can evaluate the usefulness of each problem on a scale from 1 to 5 as well. They can even choose a nickname and an avatar (a small image) which appears in rating lists instead of the actual name.

Note that the R programming language is central to e-status. It is used for generating data and correct answers for each problem, for assessing the answers given by the students, and for numerically and graphically analyzing registry data from the students' database. R can even be used to customize the analysis of results, as teachers can write and execute their own procedures.

5 Experiences

e-status has been successfully used on many different courses (Table 1 shows only the latest). We have checked that there is a positive correlation between e-status marks and final grades, although this could happen because of a confusion effect: the best students use e-status more and obtain better marks. However, the perception from both teachers and students is that e-status makes students more skilful in resolving common problems. The authors are planning an experimental design to confirm the efficacy of e-status as a learning resource, which will be carried out in the first semester of 2008.

Until now the platform has been used only in statistics subjects, but it could be used in many other areas of knowledge: the only condition is that a problem should be able to be expressed through R instructions.

Teachers are in charge of filling the repository of problems using the e-status application. We have experienced that the platform is exciting not only for students, but also for teachers. The creation of a problem that is educational and motivating depends fully on the teacher's ability. Teachers learn to create problems with e-status (and collaterally improve their skill with the R programming language). The common life cycle of a problem usually begins as a simple exercise, being enriched in subsequent years.

e-status is useful in distance learning courses and for self-study because of its internet nature. However, we are extensively using it in "on-site" classes. Problems can be structured in blocks available to the students while they are studying a topic. This is aligned with the syllabus of the course and it can likely help students in their schedule of subject preparation.

6 Conclusion

Although e-status could be implemented without R, undoubtedly its inclusion has meant a significant change. The R language is easy to use, and it can be applied to model any problem, basic or complex. Nice-looking graphics are

Course	Studies	Number students	Number problems	Number solutions
1 EST-FIB	Computer Science	144	30	1815
2 MEEI	Engineering	266	10	3327
3 Bio	Statistics	14	17	495
4 Bio-Odon	Dentistry	128	12	1699
5 Best	Postgraduate course	29	14	483

Table 1. Some courses using e-status at the end of the year 2007, with number of students enrolled, number of problems available and number of exercises solved.

simple to produce in R, and they are a good complement to many problems, if not necessary. Besides, e-status benefits from R to compute most of the statistics and graphics we can use to assess student performance.

Compared to other platforms (Cramer *et al.*, 2003), the tool described here is quite modest, but it has the advantage of being both powerful and simple at the same time. Moreover, we are deeply committed to promoting the use of information and feedback from the work done with the tool, a valuable resource for getting the students motivated and to awaken their interest in statistics.

The students of our content-overloaded subjects mostly focus on problem-solving techniques whenever possible, which is usually the case for many statistics courses. This was the core idea that gave birth to e-status, with a satisfactory degree of utilization suggesting that it has been well accepted. Students, and teachers also, may be *caught* by the operation, perhaps fascinated by the never-equal repetitions process. The added value of the Information Technologies in teaching should be remarked on, when they are applied, as in our case. We teach courses in a classroom setting with additional online activities (exercises in e-status), that can be supervised by the instructors, which is not possible for many web resources. According to the compromises for consistency of quality assurance across the European Higher Education Area, e-status is clearly an active learning method and a utensil to allow measurement of student effort too.

7 Acknowledgments

The authors wish to acknowledge the Institut de Ciències de l'Educació (UPC) for supporting this research under grant PMD 2006/07.

References

CRAMER, E., HÄRDLE, W., KAMPS, U. and WITZEL, R. (2003): e-stat: Views, Methods, Applications. INT STATISTICAL INSTITUTE: *Proceedings of the ISI 54th*, Berlin (Germany).

GONZÁLEZ, J.A. and MUÑOZ, P. (2002): e-status: a web tool for Learning by Doing Exercises. In: John Wiley & Sons Inc.: *Proceedings of the 2nd ICTM*, July 2002. Crete (Greece).

GONZÁLEZ, J.A. and MUÑOZ, P. (2006): e-status: an Automatic Web-Based Problem generator-Applications to Statistics. *Computer Applications in Engineering Education, 14 (2) 151-159.*

KOUTSONIKOLA, V. and VAKALI, A. (2004): LDAP: Framework, Practices, and Trends. *IEEE Internet Computing, 8 (5) 66-72.*

R Development Core Team (2006): R: A language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

URBANEK, S. (2003): Rserve – A fast way to provide R functionality to applications. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*.

MASTINO: Learning Bayesian Networks Using R

Massimiliano Mascherini[1], Fabio Frascati[2], and Federico M. Stefanini[2]

[1] European Commission, Joint Research Centre
Via E. Fermi 2479, 21027 Ispra(Va), Italy, *massimiliano.mascherini@jrc.it*
[2] Dipartimento di Statistica 'G.Parenti', University of Florence
V.le Morgagni 59, 50100 Florence, Italy, *stefanini@ds.unifi.it*
fabiofrascati@yahoo.it

Abstract. Bayesian Networks are increasingly used to represent conditional independence relations among variables and causal information in problem domains in which decisions are based on probabilistic reasoning. Structural learning is NP-hard therefore the database of observed cases must be often supplemented with search heuristics based on prior information. In this paper we present a software package for R, called MASTINO, that extends the existing DEAL package by providing new tools for learning Bayesian Networks and Conditional Gaussian networks in a score-and-search framework, such as the score function *P-metric* and the *M-GA* genetic algorithm. MASTINO is freely available under the terms of the GNU General Public License Version 2, and it has been recently submitted to be part of the CRAN repository. Meanwhile it can be downloaded from the website: `http://statind.jrc.it/mastino`.

Keywords: Bayesian networks, structural learning, R

1 Introduction

Bayesian Networks (BNs), Cowell et al. (1999), are a widespread tool in many areas of artificial intelligence and statistics because of efficient algorithms which make probabilistic inference effective in highly structured problem domains. BNs are suited to represent conditional independence relationships but they have been extended to represent causal information, Spirtes et al. (2000), and utility of decisions, so that probabilistic expert systems are increasingly developed in areas ranging from technology to medical problem domains.

Inference about the structure of a BN, also called structural learning, has been proved to be a NP-hard problem, Chickering (1995). Structural learning is typically performed by combining expert's priori knowledge with the information contained in a database of cases. Several heuristics have been shown to work in practice and it seems that specialized problem domains take benefits from problem-dependent tuning. A software package for `R`, (R Development Core Team, 2008), suited to quickly implement hypothesized heuristics and

to test them against standard benchmarks in structural learning is therefore welcome.

Structural learning of Bayesian Networks has been first implemented in R by Bøttcher and Dethlefsen (2003), who wrote the DEAL package in which the BDe metric was implemented for learning Conditional Gaussian (CG) networks, Bøttcher (2005). A Greedy Search algorithm with random restart is the search engine optimizing the BDe score.

In this paper we present MASTINO, a R package that provides new tools for learning BNs and CG networks. MASTINO extends DEAL in several ways. In particular, the P-metric score function is implemented to evaluate CG newtorks under strong but partial prior information on the structure, Mascherini and Stefanini (2005b, 2007). The *M-GA* genetic algorithm is based on a new population-based heuristic aimed at a robust search for the best CG network, Mascherini and Stefanini (2005). The suite of functions includes some utilities to help with the manipulation of BNs and CG networks. MASTINO is freely available under the terms of the GNU General Public License Version 2, and it can be downloaded from the website: `http://statind.jrc.it/mastino`. The package works under a R version >= 2.4.1 and it has been recently submitted to be part of the CRAN repository, therefore it should be soon available for download among contributed packages.

This paper starts with Section 2 in which Bayesian Networks are shortly described and structural learning of Bayesian Networks is introduced in the score-and-search approach. New methods implemented in MASTINO are explained. Then, Sections 3 and 3 describe two simple examples together with the corresponding R code and conclusions and issues to be addressed by further research concludes the paper. For the discussion of more complex real data examples we address the reader to Mascherini and Stefanini (2007).

2 Learning Bayesian Networks

A BN is a graph-based representation of random variables encoding their joint probability distribution in a compact way. For terminology and theoretical aspects on BNs, we refer to Cowell et al. (1999). In this paper discrete BNs are shortly defined as a directed acyclic graph (DAG) $D = (V, E)$ where V is a finite set of vertices and E is a finite set of directed edges between vertices. The DAG D encodes the structure of the Bayesian Networks. To each vertex $v \in V$ in the graph corresponds a discrete random variable X_v. The set of variables associated with the graph D is $X = (X_v)_{v \in V}$. For shortness, sometimes the label v also indicates the correspondent random variable X_v. In addition, for each vertex v a set of parents, $pa(v)$, is defined. A conditional probability table (CPT) is attached to every pair $(v, pa(v))$. Thus the set P of all local probability distributions $p(x_v \mid x_{pa(v)})$ is obtained. The joint

probability distribution is defined through the factorization:

$$p(x) = \prod_{v \in V} p(x_v \mid x_{pa(v)})$$

In order to completely specify a Bayesian Network for X, we must therefore specify a DAG D and a set P of local probability distributions, operatively CPTs. Note that nodes at the tail of directed edges reaching node v denote the random variables in the conditional distribution of X_v, thus a conditional independence assertion is associated to the lack of one or more directed edges. Further conditional independence relations may be read from the graph by exploiting separation theorems. CG-BNs are probabilistic Networks in which both continuous and discrete random variables are present. To ensure exact local computation, discrete random variables are not allowed in CG-BNs to have continuous parents. A method to perform parameter and structural learning in CG-BNs has been recently described in Bøttcher (2005), where a comprehensive discussion is performed.

The set of directed edges E on V defines a DAG, the structure of a BN. Structural learning of BNs may be performed following two main different approaches. In the first approach, learning follows the original PC schema, Spirtes et al. (2000), which performs statistical tests to produce a list of conditional independency relations. In the second approach, called "score-and-search", algorithms compare candidate DAG structures according to a given score, also called metric, which is used as objective function during optimization. Widely used scores include BIC, AIC and MDL. The most important score from the Bayesian standpoint is the BDe metric, Heckerman et al. (1995), in which a candidate DAG structure B_s on a fixed set of nodes V is supported by observed data if the conditional posterior distribution of B_s given observed data is large.

The above metrics have been all successfully used in actual learning tasks. Despite the recognized possibility of improving the learning process by exploiting prior information, attempts to elicit prior beliefs on networks structure in a quantitative way are still quite limited in the literature.

Mascherini and Stefanini (2005, 2007) proposed a new score function, the P-metric, which mixes prior beliefs and experimental information following the BDe assumptions, Heckerman et al. (1994). In particular, the P-metric encodes the a-priori belief on the structure of a candidate network B_s by a score function $S_{prior}(B_s)$ which captures some local and some global network features. Local features are defined by score component $S_p^{\delta}(B_s)$ that describe beliefs on the presence of oriented edges, each one marginally considered. Partial prior beliefs on network topology is encoded by score component $S_p^{\tau}(B_s)$, which takes the form of an expected degree of connectivity, for example the expected number of parents for one child, and it is scaled according to the Kullback-Leibler distance, (Kullback et al., 1951).

The proposed score function $S_{prior}(B_s)$ combines the two components $S_p^{\delta}(B_s)$ and $S_p^{\tau}(B_s)$ on the logarithmic scale:

$$S_{prior}(B_s) = \log\left[\left(\frac{P(B_s)}{P(\{\emptyset\})}\right)^{\alpha} \cdot e^{(1-\alpha)(-KL(\mathcal{P}_{pa}\|\mathcal{Q}_{pa}))}\right] \tag{1}$$

where the role of α, $0 \leq \alpha \leq 1$, is to balance the relative strength of components due to edge orientation and to network topology. A value $\alpha = 1$ is suited to perform learning without accounting for the network topology component. According to the prior belief, the most plausible structure is obtained by maximizing the score $S_{prior}(B_s)$ with respect to B_s.

Our Bayesian-inspired metric, called *P-metric*, mixes the elicited prior information and experimental information in a way close to Heckerman et al. (1994). The Bayesian Dirichlet with Equivalence metric, (BDe), assigns the same likelihood value to structures which are likelihood equivalent, i.e. DAGs encoding the same assertions on conditional independence relations. The equivalence is obtained by choosing BN parameters through a prior procedure in which Dirichlet hyperparameters are defined using the notion of equivalent sample size. Then, the *P*-metric defining the score of a candidate structure B_s given a complete database of cases $\mathcal{D}$ is, on the log scale:

$$log\left(S_{P\text{-}metric}(B_s)\right) = \beta_z \cdot \log(S_{prior}(B_s)) + ll_{BDe}(D \mid B_s, \theta) \tag{2}$$

The role of the parameter β_z is to calibrate the strength of the prior score with respect to the likelihood function. The value of β_z depends on the size of the problem domain and on the sample size of cases as well as on the elicited belief. Clearly for $\beta_z = 0$ the *P*-metric is equal to the BDe metric, if a uniform prior distribution over structures is chosen in the BDe. In MASTINO the P-metric is implemented with the `Pmetric()` function. The best network using the P-metric can be found by two heuristic strategies: the greedy search, `Pmetric.search()` and the perturbated hill-climb, `p.hill.climb()`, that are an extension of the algorithms already implemented in DEAL.

Focusing on the heuristic strategies, in order to search the best Conditional Gaussian Bayesian Networks using the BDe metric, in MASTINO a genetic algorithm, named *M-GA*, (Mascherini et al. (2005)), is implemented with the `MGA()` function.

Genetic algorithms (GAs) have been first used by Larrañaga et al. (1996) to search for optimal discrete BN structures. The GA implemented in MASTINO is a modification of the method proposed by Larrañaga et al. (1996) which also works with CG networks. In the *M-GA* the single crossover point of Larrañaga et al. (1996) is extended and the two parents of a new individual equally contribute to offsprings on a gene by gene basis, taking part bit by bit in the creation of the new individual string, to maintain or increase the genetic variability of the population of candidates. Furthermore, a fixed number of randomly generated networks is added at each generation as immigrants.

3 Examples

The ASIA network is a small fictitious and well-known Bayesian network, Lauritzen et al. (1988), for calculation of the probability of a patient having tuberculosis, lung cancer or bronchitis given values taken by some other variables, like visit-to-Asia which is one if the patient recently visited Asia. The problem domain is here quite rich, for example shortness-of-breath, called dyspnoea (D), may be due to different factors, like tuberculosis (T), lung cancer (L), and bronchitis (B). Then a recent visit to Asia, (A), increases the risk of tuberculosis, while smoking, (S), is known to be a risk factor for both lung cancer and bronchitis. Results of a single chest X-ray, (X), do not discriminate between lung cancer and tuberculosis, (E), as neither does the presence or absence of dyspnoea. All the 8 variables of the model are binary and the dataset included in MASTINO contains 1500 cases.

In MASTINO, the initialization of a network precedes the score-and-search step of structural learning. Being based on the package DEAL, the library MASTINO can be used to learn both discrete and CG networks. In particular MASTINO exploits the representation of a Bayesian Network as an object of class network defined in DEAL. Networks are generated from a dataframe, and discrete variables must be specified factors. In Bøttcher et al. (2003) a complete description of the DEAL functions is provided. Network building is performed using DEAL resources:

```
> library(MASTINO); data(asia); df = asia; net=network(df)
```

Parameter learning follows the Bayesian approach. Parameters of the joint distribution of variables in the network are determined by the function `newprior()`, that is based on the function `jointprior()` defined in DEAL. To improve the learning process fully discrete and CG networks are treated as different objects by setting automatically the parameters of the master prior function, Bøttcher (2003):

```
> prior=newprior(net); net.2=getnetwork(learn(net,df,prior))
```

After defining the prior distribution the P-metric may be used to learn the structure using expert's belief. We assume that the available prior information was partially quantified by experts concerning three pairs of nodes: (A, T), (S, L) and (L, T). The expert states that the node "Tuberculosis" (T) is not reputed to have any effect on "Visiting Asia" (A), so the probability of the event $A \leftarrow T$ was set to be equal to 0.01, and the remaining probabilities are set to capture a slight effect of the event "Visiting to Asia" on "Tuberculosis"; then, "Smoking" (S) was believed to have an effect on "Lung Cancer" (L) but "Lung Cancer" did not have any effect on "Smoking". The probability of those events was set to $P(S \rightarrow L) = 0.6$ and $P(S \leftarrow L) = 0.01$ respectively. Finally, no effects between "Lung Cancer" and "Tuberculosis" (T) was believed to exist, so the probability of the event $L \not\leftrightarrow T$ was set equal to 0.8. For simplicity, the prior information elicited above was then encoded in MASTINO using vectors, where the first value is equal to the prior

probability, the second is the identificative number (ID) of the parent node and the last is the ID of the child node:

```
> bel1=c(0.01,5,3); bel2=c(0.55,3,5); bel3=c(0.6,4,6)
> bel4=c(0.01,6,4); bel5=c(0.1,6,5); bel6=c(0.1,5,6)
```

The six vectors are then merged into a matrix and they are included in MASTINO using the dedicated function:

```
> belief=rbind(bel1,bel2,bel3,bel4, bel5, bel6)
> PV=includeBelief(belief, net)
```

Prior information on network topology was defined by requiring that 80% of network nodes has at most one parent, `q_x=c(0.8,0.2); cl=1`. The set of input parameters in the P-metric function must be specified before structural learning. In particular the strength of the prior information, β and the importance of the local features, α, must be specified. Mascherini and Stefanini (2007) numerically explored the effect of several different pairs of parameter values on the overall score. The P-metric seemed not highly sensitive to the precise numerical choice of two parameters. Based on these results, here we set $\beta = 0.5$, i.e the strength of the prior information is reduced of 50%, and α=0.75, because due to the small size of the network the local features are considered more important than the global features. Structural learning takes place by invoking `Pmetric.search` or `P.hill.climb`, which respectively extend the greedy search algorithm and the perturbated hill climbing implemented in DEAL:

```
> alpha= 0.75; beta=0.5; best.gs=
Pmetric.search(net.2, df, prior, beta, alpha, PV, cl, q_x)
> best.hc=p.hill.climb(net.2, df, prior, beta, alpha, PV, cl,
q_x)
```

DEAL provides functions to plot learned networks. In this case study the two algorithms converged to the same network presented in Figure (1, a). The comparison of these two networks is performed by means of function `compareBN`, which gives a summary of the similarities of the two networks in terms of correct/wrong arcs:

```
> plot(best.gs[[1]]); plot(best.hc[[1]])
> compareBN(best.gs[[1]],best.hc[[1]])
```

The comparison of the learned network with the original ASIA network shows that a total of 7 arcs (correct + reverse oriented) out of 8 are successfully identified by MASTINO. Although in the network learned using the P-metric one arc is missing, these findings suggest the overall effectiveness of our algorithms. In particular, the P-metric outperforms other algorithms as the BDE metric implemented in DEAL and PC-NPC algorithms implemented in the commercial software HUGIN, Andreassen et al. (1989). In fact, the P-metric correctly identified a total of 7 arcs in the best case against a total of 5 arcs discovered by the PC and NPC algorithms. The comparison of computer runs performed with the BDe metric implemented in DEAL shows an unexpected feature of the DEAL implementation: if prior parameters are automatically set by DEAL then the learning algorithm discover a total of 7 arcs out of 8 but it also adds other 17 incorrect arcs. On other hands, if prior

parameters are manually set then a total of 5 arcs are correctly identified by the BDe metric implemented in DEAL (Figure 1, b).

RATS is a network created by simulated data available in DEAL. The dataset is structured in 24 rats (12 females, 12 males) receiving a randomized assignment of one drug among three products for loosing weight. The weight loss for each rat is noted after one and two weeks. The variable included in the dataset are: Sex (discrete, binary), Drug (discrete, trinomial), W1 (numeric, weight loss, one week), W2 (numeric, weight loss, second week). The aim is to assess the effects of Drugs on the rats' weight loss. The network is initialized as described above:

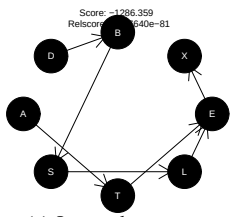
```
> data(rats); df = rats; net=network(df); prior=newprior(net)
> net.2=getnetwork(learn(net,df,prior))
```

After network initialization, structural learning of a CG network is performed by invoking the `MGA()` function, a population-based algorithm performing stochastic search in the space of CG networks. According to the literature, we set the following parameter values: immigration rate = 0.05; mutation rate = 0.01; crossover = 0.5; population size = 10; number of generations = 10. Besides setting different values for parameters of the M-GA algorithm, the user must specify at least a dataframe of observations: `>best.MGA = MGA(df,0.05,0.01,0.5,10,10)`. The 10 best structures from one run are contained into the list `best.MGA`. The graphical representation of the best structure is obtained by executing the function `plot`. The best network found by M-GA is equal to the original network and also to the network learned using the autosearch function implemented in DEAL. Values of the BDe score for top scored structures along generations are easily obtained and plotted as the output of the commands below shows:

```
>plot(best.MGA[[1]]); best.DEAL=autosearch(net.2, df, prior)
>compareBN(best.MGA[[1]], best.DEAL[[1]])
```

4 Conclusion

In this paper we presented MASTINO, a `R` package to learn Bayesian Networks following a score-and-search approach, which extends the DEAL pack-



(a) Learned structure.

Algorithm	Correct and Reverse Oriented	Missing Arcs	Incorrect Added
PC	5	3	0
NPC	5	3	1
BDe^{*}_{DEAL}	7	1	17
BDe^{**}_{DEAL}	5	3	2
P-metric	7	1	0

(b) *prior parameters automatically set; **prior parameters manually specified.

Fig. 1. Structural learning of the ASIA network in MASTINO [a]. Performances of some learning algorithms are compared at a sample size equal to 1500 [b].

age. The score metric, called *P-metric*, is implemented to evaluate structures using an informed score and the *M-GA* genetic algorithm is coded to perform a robust search in the space of CG networks. Although some computational constraints of `R` limit the use of MASTINO to problem domains with few variables, the package represents an original implementation of a set of recently proposed tools. Further work could be directed towards making MASTINO suited to learn large sized networks. A preliminary inquiry seems to suggest that a low-level recoding will both make MASTINO independent on DEAL and increase its speed up to handle reasonably large problems domains.

References

ANDREASSEN, S.K., OLESEN, K.G., JENSEN, F.V. and JENSEN, F. (1989): HUGIN: a shell for building bayesian belief universes for expert systems. *Proceedings of the 11th International Joint Conference on Artificial Intelligence.*

BØTTCHER, S.G. and DETHLEFSEN, C. (2003): DEAL: A package for learning bayesian networks. *Journal of Statistical Software 8(20), 1–40.*

CHICKERING, D.M. (1995): Learning bayesian networks is NP-complete. *Proceedings on Artificial Intelligence and Statistics, 121–130..*

COWELL, R.G., DAWID, P.A., LAURITZEN, S.L. and SPIEGELHALTER, D.J. (1999): *Probabilistic Networks and Expert Systems.* New York: Springer-Verlag.

HECKERMAN, D., GEIGER, D., and CHICKERING, D.M. (1994): Learning Bayesian Network: A combination of knowledge and statistical data. *Proceedings of 10th Conf. Uncertainty in Artificial Intelligence, 293–301.*

KULLBACK, S. AND LEIBLER, R.A. (1951): On information and sufficiency. *Annals of Mathematical Statistics 22, 79-86.*

LARRAÑAGA, P. and POZA, M. (1996): Structure Learning of Bayesian Networks by Genetic Algorithms: A Performance Analysis of Control Parameters. *IEEE Journal on Pattern Analysis and Machine Intelligence 18(9), 912-926.*

LAURITZEN, S. and SPIEGEHALTER, D. (1988): Local computation with probabilities on graphical structures and their application to expert system. *Journal of the Royal Statistical Society - B Series 50(2), 157–192.*

MASCHERINI, M and STEFANINI, F.M. (2005): M-GA: A genetic algorithm to learn Conditional Gaussian Bayesian Networks. *Proceedings of the IEEE International Conference on Computational Intelligence for Modelling, Control and Automation.*

MASCHERINI, M and STEFANINI, F.M. (2005b): Encode prior information to learn Bayesian networks. *WP n.13 of the Department of Statistics*, Florence University Press.

MASCHERINI, M. and STEFANINI, F.M. (2007): Using Weak Prior Information on Structures to Learn Bayesian Networks. *Lecture Notes in Artificial Intelligence. 4692(1), 413–420.*

R DEVELOPMENT CORE TEAM (2008): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051- 07-0, URL http://www.R-project.org.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000): *Causation, Prediction, and Search, 2nd ed.* New York, N.Y.: MIT Press.

Interactive Software for Optimal Designs in Longitudinal Cohort Studies

Frans E. S. Tan, Fetene B. Tekle, and Martijn P. F. Berger

University of Maastricht, Dept. of Methodology and Statistics,
P.O.Box 616, 6200 MD, Maastricht, The Netherlands,
frans.tan@stat.unimaas.nl

Abstract. A new user-interface computer-program for the optimization of designs for longitudinal cohort studies is presented. Users can specify the random effects model and the cost parameters for the data analysis and the software will produce the optimal design results. The software helps users to identify the optimal cohort design, optimal number of repeated measurements per subject, optimal allocations of time points within a given study period. Further, users can compute the loss in relative efficiencies of any other alternative design compared to the optimal one.

Keywords: optimal design, cohort designs, longitudinal study, relative efficiency, software

1 Introduction

A user-interface computer-program for the optimization of designs for longitudinal cohort studies is presented. Many large scale longitudinal studies have been set-up in the past to describe health status changes over time and to identify the associated risk factors. Some examples of such studies are the Framingham heart study (Dawber, 1980), the national cancer prevention study (Calle et al. 2002), and the Survey of Medical Information and Life style study in Eindhoven- SMILE (Bruijn et al. 2005). Researchers of such large scale studies often face the problem of how to allocate the repeated measurements over time. Many of these studies may have not been benefited from the optimal design theory. The number of repeated measurements and the selection of the time points at which the measurements are taken in most large scale studies is done on more or less ad-hoc basis.

The problems of optimal designs for longitudinal cohort studies have been discussed in the literature by Berger (1986), Tan and Berger (1999), Ouwens et al. (2002), Berger and Tan (2004), and Tekle et al. (2007) among others. However, no generally available software for the optimization of longitudinal studies has been developed so far. In this paper we introduce and describe a new interactive software program.

The paper is organized as follows. In section 2, design problems for longitudinal cohort studies are briefly discussed. The generalized mixed effects models and optimal designs will be described in section 3. Relative efficiency

is defined as a measure of comparison of two designs in the same section. In section 4, the software program to implement the optimal designs for longitudinal studies will be described.

2 Design problems for longitudinal cohort studies

Longitudinal cohort studies sample a cohort, defined as a group experiencing some event in a selected time period, and measure them on selected outcome through a time interval or study period T. Depending on the number of cohorts, the cohort design may be purely longitudinal, cross-sectional or mixed longitudinal. In purely longitudinal design a group of subjects followed over the study period. Alternatively, in a cross-sectional cohort design different cohorts of subjects are selected and measured at different points of time within the study period, i.e. the number of cohorts is the same as the number of time points at which the samples are considered. Sometimes, a follow up period is considered too long to complete data collection for the purely longitudinal cohort design. As the follow up period gets longer , there is high chance for dropouts and non compliances. If the longitudinal information for subjects is still required, a mixed longitudinal design with several cohorts measured over a shorter time interval for each cohort can be considered.

Planners of cohort studies often face the problem of choosing a purely longitudinal, cross-sectional or mixed longitudinal design. Tekle et al. (2007) have considered a numerical study to compare these cohort designs under fixed costs of a study and found that the purely longitudinal cohort design is the most efficient one in minimizing the variance of the parameter estimators of the linear mixed-effects models. Tan and Berger (1999) have suggested that the optimal number of repeated measurements for purely longitudinal cohort design should be the same as the number of regression parameters in the model. Tekle et al. (2007) extended the result for other cohort designs and concluded that the optimal number of repeated measurement is equal to the sum of the number of regression parameters in the model and the number of cohorts in the study minus one.

The software that we will describe helps users to identify the optimal cohort design, optimal number of repeated measurements for a given cohort design, optimal allocations of time points. Furthermore, the software also computes the loss in relative efficiency of any other alternative design compared to the optimal one. The user will specify the random-effects model and the costs parameters to obtain the optimal design results.

3 The model and optimal cohort designs

For the analysis of a longitudinal trend of a continuous response variable over time, a linear mixed effects model is discussed by Diggle et al. (1994), and Verbeke and Molenberghs (2000), among others. For a cohort design of N

subjects with the number of cohorts equal to S, the model is given by:

$$y_{si} = X_s\beta + Z_s b_{si} + \varepsilon_{si}, \tag{1}$$

where the $q_s \times 1$ vector y_{si} is a vector of repeated measurements on a continuous response variable for person i, $i = 1, \cdots n_s$, at q_s time points $t_s = (t_{s1}, \cdots, t_{sq_s})'$ in cohort s, $s = 1, \cdots, S$ such that $\sum_{s=1}^{S} n_s = N$ and the total number of time points $q = \sum_{s=1}^{S} q_s$, X_s is the $q_s \times p$ matrix of explanatory variables including polynomials of the measurement times and Z_s is the $q_s \times r$ submatrix of X_s; p and r are numbers of fixed and subject-specific regression parameters, respectively. The $p \times 1$ vector β is a vector of fixed regression coefficients. The $r \times 1$ vector b_{si} is the vector of subject-specific coefficients with mean zero and $r \times r$ variance-covariance matrix G. The $q_s \times 1$ vector ε_{si} is the vector of random errors for subject i within cohort s, with mean zero and variance-covariance matrix $\sigma^2 R_s$, where the $q_s \times q_s$ matrix R_s is the correlation matrix of the error vector and σ^2 is a common variance for error components. We have considered a first order autoregressive correlation structure (AR1).

The regression parameters can be estimated using maximum likelihood method. The interest in optimal design theory is to optimize a certain function of the variance-covariance matrix of the parameter estimators.

The cohort designs $\xi \in \Xi_{Sq}$ are defined as follows:

$$\xi = \begin{Bmatrix} t_1 & t_2 & \cdots & t_S \\ w_1 & w_2 & \cdots & w_S \end{Bmatrix}, \tag{2}$$

where Ξ_{Sq} is the design space of all cohort designs with S cohorts such that the total number of time points within a study period is equal to q, $t_s$is a vector of time points $t_s = (t_{s1}, \cdots, t_{sq_s})'$ in cohort s, $s = 1, \cdots, S$, and w_s is the relative size of cohort s such that $\sum_{s=1}^{S} w_s = 1$. The interest is now to identify an optimal design among elements of Ξ_{Sq} that minimizes a certain function of the variance-covariance matrix of the parameter estimators. We restrict in this paper to optimal designs using the D-optimality criterion. The popular D-optimal design minimizes the determinant of the variance-covariance matrix of the parameter estimators.

The number of measurements per subject usually needs to be restricted because of monetary cost, subject fatigue, and other logistical reasons. Therefore, it is reasonable to take the cost of measuring into account when a cohort design is optimized. Fedorov et al.(2002) and Tekle et al. (2007) have considered a variance-covariance matrix of the parameter estimates that is normalized for the cost parameters.

The normalized variance-covariance matrix under a design ξ that takes

into account the cost parameters is:

$$var(\hat{\beta}_\xi) = \left[\sum_{s=1}^{S} w_s \left(\frac{X_s' V_s^{-1} X_s}{c_1 + c_2 q_s)}\right)\right]^{-1} = \left[\sum_{s=1}^{S} w_s \left(\frac{X_s' V_s^{-1} X_s}{k + q_s}\right)\right]^{-1}, \quad (3)$$

with $w_s = \frac{n_s(k+q_s)}{C-c_0}$, where V_s is the variance-covariance matrix of the response variable in cohort s , C is the total budget of a study, c_0 is the initial setup costs of a study and may also include the costs of maintaining the research staff during the study period, c_1 is the costs of recruiting a new subject, c_2 is the costs of a measurement at each time point for each subject, k is the ratio of the costs of recruiting a new subject c_1 and the costs of each measurement c_2, i.e. $k = c_1/c_2$. As it can be seen in (2.1), the variance-covariance matrix of the parameter estimators depends only on k among the cost parameters (see Tekle et al. (2007) for details on the cost functions and cost parameters).

If the parameters in the variance-covariance matrix V_s and the ratio of the costs c_1 and c_2 are known, then the generalized variance of $var(\hat{\beta}_\xi)$ in equation (2.1) can be minimized over the design space Ξ_{Sq} by choosing a D-optimal cohort design ξ^* such that, for each design $\xi \in \Xi_{Sq}$,

$$\det\left\{var(\hat{\beta}_{\xi^*})\right\} \leq \det\left\{var(\hat{\beta}_\xi)\right\}, \quad (4)$$

where $\hat{\beta}_{\xi^*}$ and $\hat{\beta}_\xi$ are estimators of β under ξ^* and ξ, respectively. That is, an optimal design is a design that has the smallest generalized variance of model parameter estimators for a given total cost.

The optimal design problem for a cohort study is to identify the optimal time locations of the vector t_s, the associated optimal weights w_s's, and the number of repeated measurements in each cohort, q_s for a given total cost of a study and cost parameters.

Relative efficiency (RE) can be used to compare the efficiencies of two designs. The RE of a design ξ_2 compared to a reference design ξ_1 is the ratio of the generalized variances to the power of the incerse of the number of parameters (Atkinson and Donev 1992). If the ratio is less than 1, then ξ_2 is less efficient than ξ_1 and design ξ_1 is to be preferred compared to ξ_2. The reciprocal of the RE is equal to the relative amount of extra observations that must be taken under ξ_2 to obtain the same efficiency as the design ξ_1 (Atkinson and Donev, 1992).

4 POLS program

Program for Optimal design of Longitudinal Studies (POLS) is an interactive program that helps to obtain optimal cohort designs. The program runs in a MATLAB software environment. POLS for linear mixed-effects models will be

described in details in this section while the software and the corresponding program for logistic mixed effects models are available upon request from the authors. The program requires the user to choose the type of the polynomial of the model and give input values on the cost parameters, the variance-covariance parameters for random-effects and error terms. Then, the user can request and obtain the optimal design solutions via the main menu of the program. The program provides results on Optimal allocations of time points, optimal weights of cohorts, relative efficiency of optimal designs with different number of time points at fixed costs parameters, relative efficiency of equi-distance designs with different number of time points at fixed costs parameters, relative efficiency of an optimal design with a specific number of time points with different values of cost ratios, and relative efficiency of cohort designs with different number of cohorts S.

4.1 Description and use of the program

The software is written in MATLAB code. Type *mainmenu* in the command window and press the Enter key to begin with the program. Then, the menu shown in Figure 1 will appear. The user should give the necessary input values by filling the corresponding boxes and pressing the corresponding buttons. The results can be obtained by clicking on the buttons shown on the right side of the main menu. The user has to provide input values for certain parameters and make choices. Note that the correlation structure of the R_s matrix in each cohort s is fixed to AR(1) structure. The parameters that require input values are:

- *Choice of the polynomial*: selecting among linear, quadratic and cubic models. The default polynomial of the mixed effects model is linear as shown in Figure 1, but the user can change to either quadratic or cubic from the list in the pop-up menu.
- *Values of the parameters in the variance-covariance matrix* G: these are the variance-covariance parameter values for the random-effects parameters. If a fixed-effects model is assumed, then all the values required in G should be set to zero. In general, the parameter in G can be set to values greater or equal to zero. Previous numerical studies have shown that the choice of the parameter values has little or no effect on the results of optimal designs (Tan and Berger (1999), Tekle et al. (2007)). The user will find a sub-menu to give the values for the variance components in G matrix and the variance of the error terms after clicking on the 'variance parameter (G)' button.
- *Value of variance parameter for the error terms,* σ^2: The value can be set based on the results from previous related studies or other sources and it should be greater than zero.
- *Values for costs parameters*: the costs parameters are the total budget of the study C, the cost per new subject c_1, the cost of measurement (new

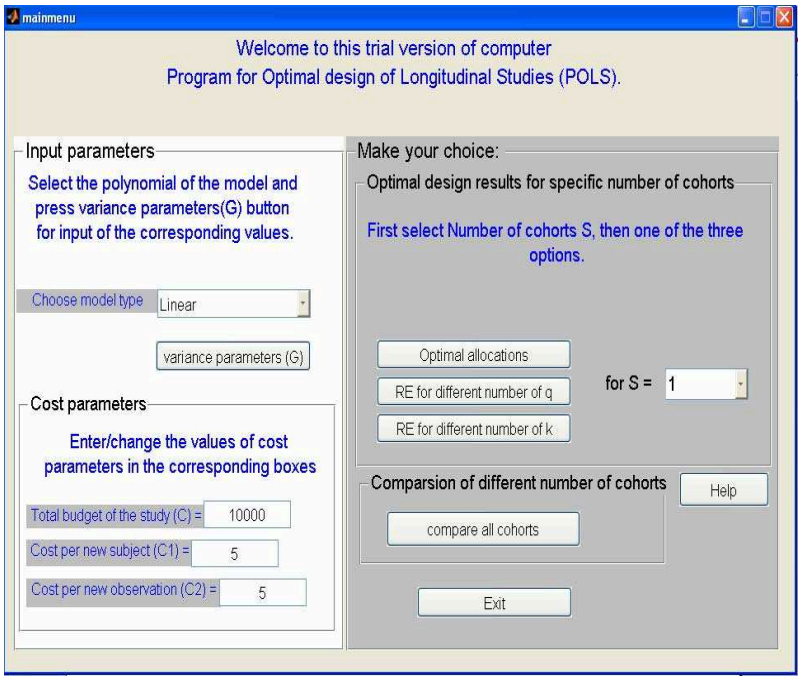


Fig. 1. The main menu with the default input values for POLS.

observation) c_2. The initial setup costs parameter c_0 is set in the program to a constant for all cohort designs. The value of the cost parameter C can be any value greater than zero. Even if, as discussed in section 3, the value of this parameter does not affect the objective function of the optimization and will not affect the results, the budget of a cohort study is usually large in practice. The value of c_1 must be greater than or equal to the value of c_2. The cost parameter values can be changed from the default values given in the corresponding boxes shown at the left bottom in the main menu (Figure 1).

As shown in Figure 1, the default value for the number of cohorts S is 1. The user can change the number of cohorts S by choosing the values from the pop-up menu. The user can choose, by clicking on buttons, between the following options:

- '*Optimal allocations*': Dialog boxes will appear to fill the specific number of time points and the starting and end period for the study time interval (T).

- '*RE for different number of q*': It gives the relative efficiency of designs with the number of time points $q > p$ compared to an optimal design with the number of time points $q = p$ for fixed values of cost parameters.
- '*RE for different number of k*': Relative efficiencies of designs with different values of the cost ratio k (c_1/c_2) can be compared by selecting this option. To compare designs with different values of the cost ratio, one has to fix the number of time points q. The user will find an input dialog box to fill the specific number of time points q for the comparison of the values of cost ratios.
- '*compare all cohorts*': This option allows to obtain results for the comparison of cohort designs with different number of cohorts. The user will be asked to choose the number of time points for each cohort. Values with $q \geq p$ and $q \geq S$ can be filled in the boxes of a sub-menu. The computer time increases exponentially as the number of cohorts and time points increases.

All the results can be displayed in an editable matlab figure and if required also in an excel table. Upon request, the program opens an excel file to print the results. Users can change input values or obtain results by pressing the corresponding buttons on the main menu shown in Figure 1 as many times as they wish. A 'Help' button is also available to guide users. The button 'Exit' will help to quite the program.

4.2 The optimization algorithm

The core element of the program is optimization of the generalized variance of the parameter estimates. This has been done using the fmincon function from MATLAB optimization toolbox. Fmincon is a function that finds a constrained minimum of a function of several variables by using a sequential quadratic programming (SQP) method. That is, the function solves a quadratic programming (QP) sub problem at each iteration. The starting designs in our case are equi-distance time-point designs with equal weight for each cohort. An estimate of the Hessian of the Lagrangian is updated at each iteration using the BFGS (Broyden, Fletcher, Goldfarb and Shanno) formula (Matlab, 2004).

4.3 Limitations of the software

Users need to have a matlab package on their computer to run the current version of the software. However, the authors are arranging to obtain a license on matlab compiler so that the program will be converted to stand alone software such that the software can run without a prerequisite of having a matlab package.

We have noticed that the program takes longer time to give results when the number of cohorts is increasing. The fact that the optimal solution is

searched among all possible combinations of number of time points for each cohort, causes the program to take longer time to obtain the results. Thus, the number of cohorts S is restricted to 4 in the program.

References

ATKINSON, A.C., Donev, A.N. (1992): *Optimum Experimental Designs*. Clarendon Press: Oxford.

BERGER, M.P.F. and TAN, F.E.S. (2004): Robust designs for linear mixed effects models. *Journal of the Royal Statistical Society*, series C 53(4): 569-581.

BERGER, M.P.F. (1986): A comparison of efficiencies of longitudinal, mixed longitudinal and cross-sectional designs. *Journal of Educational Statistics* 11(3):171-181.

BRUIJN, G.J., KREMERS, S.P.J., MECHELEN, W.V., BRUG, J. (2005): Is personality related to fruit and vegetable intake and physical activity in adolescents? *Health Education Research* 20(6): 635-644.

CALLE, E.E., RODRIGUEZ, C., JACOBS, E.J., ALMON, M.L., CHAO, A., MCCULLOUGH, M.L., FEIGELSON, H.S., THUN, M.J. (2002): The American Cancer Society Cancer Prevention Study II Nutrition Cohort. *Cancer* 94(2): 500-511.

DAWBER, T.R. (1980): *The Framingham study: the epidemiology of atherosclerotic disease*. Harvard University Press: Harvard.

DIGGLE, P.J., LIANG, K-Y., ZEGER, S.L. (1994): *Analysis of Longitudinal Data.* Oxford University Press: Oxford.

FEDOROV, V.V., GAGNON, R.C., LEONOV, S.L. (2002): Design of experiments with unknown parameters in variance. *Applied Stochastic Models in Business and Industry* 18: 207-218.

MATLAB (2004): *Handbook* Version 7.0.1. (R14), MathWorks Inc.

OUWENS, M.J.N.M., TAN, F.E.S., and BERGER, M.P.F. (2002): Maximin D-optimal design for longitudinal mixed effects models, *Biometrics* 58: 735-741.

TAN, F.E.S., and BERGER, M.P.F. (1999). Optimal allocation of time points for the random-effects model. *Communication in Statistics, Simula* 2: 517-540.

TEKLE, F.B., TAN, F.E.S., and BERGER, M.P.F. (2007): D-optimal cohort designs for linear mixed effects models. *Statistics in Medicine*, 999999, DOI: 10.1002/sim.3045.

VERBEKE, G. and MOLENBERGHS, G. (2000): *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Part XXII

Time Series

An Efficient Estimation of the GLMM with Correlated Random Effects

Moudud Alam

Dept. of Economics and Society, Dalarna University and ESI, Örebro University. Dalarna University, SE 78188 Borlänge, Sweden, *maa@du.se*.

Abstract. This paper presents a two-step pseudo likelihood estimation technique for the generalized linear mixed models (GLMM) with random effects being correlated (possibly between subjects). Due to the use of the two-step estimation technique the proposed algorithm outperforms the conventional pseudo likelihood algorithms, *e.g.* Wolfinger and O'Connell (1993), in terms of computational time. Moreover, it does not require any reparametarisation of the model such as Lindstrom and Bates (1989). Multivariate Taylor's approximation has been used to approximate the intractable integrals in the likelihood function of the GLMM. Based on the analytical expression for the estimator of the covariance matrix of the random effects, a condition has been presented as to when such a covariance matrix can be estimated through the estimates of the random effects. An application of the estimation technique with a binary response variable is presented using a real data set on credit defaults.

Keywords: Laplace approximation, PQL, defaults correlation

1 Introduction

The literature on estimation of the generalized linear mixed models (GLMM) is abundant. The justification of yet another paper is given from the fact that none of the estimation methods currently available can provide an exact maximum likelihood estimator except for the case of Gaussian family with identity link. Most of them are computationally too extensive and all needs some restrictive assumptions about the random effects, namely they being independent. This paper presents a general (in that it works for independent and correlated random effects) algorithm to estimate the GLMM parameters using a two-step estimation procedure. The estimation procedure is derived in the lines of the Pseudo (or Penalized Quasi) Likelihood (PL or PQL) approach (Breslow and Clayton (1993); Wolfinger and O'Connell (1993)) with allowing for the correlated random effects. This two-step estimation procedure makes the mathematical derivation simple and computational implementation fast in comparison to the conventional PQL approaches.

The paper has been organized in the following way. Section two outlines on the GLMM and its estimation techniques while section three presents the proposed estimation technique. The detailed mathematical derivations

have been omitted in this version of the paper however, those details can be obtained from the author. Section four gives an application of the estimation technique to a real data set and section five ends the paper with a concluding discussion.

2 Estimation of generalized linear mixed models

An extension of the generalized linear models (GLM) with the random effects terms is called the generalized linear mixed models (McCulloch and Searle (2001)). Under the standard assumptions and with a canonical link, the conditional log-likelihood of a GLMM given the realization of the random effects is expressed in matrix notation as

$$\log(\beta, \mathbf{D}|\mathbf{Y}, \mathbf{u}) = l = \frac{\mathbf{Y}^T(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\beta + \mathbf{Z}\mathbf{u})}{a(\phi)} + \mathbf{1}^T c(\mathbf{Y}, \phi) \quad (1)$$

where, $\mathbf{Y} = \{y_{ij}\}$, $(i = 1, 2, .., n_j$ and $j = 1, 2, .., k)$ is the vector of response variable, $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is the vector of fixed effects parameters, $\mathbf{u} = (u_1,\ u_2, ..., u_k)^T$ is the vector of random effects $\eta_{ij} = \mathbf{X}_{ij}\beta + \mathbf{Z}_{ij}\mathbf{u}$ is called the linear predictor, $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices associated with fixed and random effects respectively, $b(.)$ is called the cumulant function, $a(\phi)$ is called the dispersion parameter and the conditional expectation, $E(\mathbf{Y}|\mathbf{u}) = \mu$, satisfies $g(\mu) = \eta$ for some link function, $g(.)$. Generally u_j's are assumed to be *i.i.d.* normal however, this paper considers $\mathbf{u}$ as a multi normal variate with mean vector, $\mathbf{0}$, and an unstructured covariance matrix, $\mathbf{D}$, *i.e.* $f(\mathbf{u}) = (2\pi)^{-\frac{k}{2}} |\mathbf{D}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right]$. An explanation of such correlated random effects can be given from the genetic relation's view point (Searle *et al.* (2001)) or from the defaults correlation's perspective (see section 4). It is worth noting that a method derived for the unstructured $\mathbf{D}$ contains *i.i.d.* u_j's as a special case where $\mathbf{D}$ is a diagonal matrix. Since $\mathbf{u}$ is not observable, the joint marginal likelihood of β and $\mathbf{D}$ is given as

$$L(\beta,\ \mathbf{D}|\mathbf{Y}) = \int (2\pi)^{-\frac{k}{2}} |\mathbf{D}|^{-\frac{1}{2}} \exp\left[l - \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right] d\mathbf{u} \quad (2)$$

where, the integral is taken over the dimensions of $\mathbf{u}$. The integral in equation (2) has the form, $I = \int_R C \exp[h(\mathbf{u})]\, d\mathbf{u}$, where, $h(\mathbf{u}) = l - \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}$. Such a functional form permits us to use the Laplace approximation to evaluate the integral in (2). There are several other techniques to evaluate the integral such as Gauss-Hermite quadrature and Markov-Chain Monte-Carlo (MCMC) methods (McCulloch and Searle (2001)) however, the Laplace approximation is chosen because of its computational speed. Applying Laplace approximation in (2), based on the second order Taylor's expansion of $h(\mathbf{u})$, we obtain

$$l_{PQL} = \frac{\mathbf{Y}^T(\mathbf{X}\beta + \mathbf{Z}\widetilde{\mathbf{u}}) - \mathbf{1}^T b(\mathbf{X}\beta + \mathbf{Z}\widetilde{\mathbf{u}})}{a(\phi)} + \mathbf{1}^T c(\mathbf{Y}, \phi) - \frac{1}{2}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\widetilde{\mathbf{u}} \quad (3)$$

where, l_{PQL} is the log-likelihood function of the GLMM under PQL approach and $\widetilde{\mathbf{u}}$ is the value of $\mathbf{u}$ evaluated at the maximum of $h(\mathbf{u})$.

The PQL estimation with unstructured $\mathbf{D}$ is complicated and some reparametarisation of $\mathbf{D}$ such as the Cholesky decomposition is suggested in those cases (Lam and Lee (2004); Lindstrom and Bates (1998)). However, the following section presents a way to estimate $\mathbf{D}$ where no such reparametarisation is required (see section 3) while it allows the random effects to be correlated within or between subjects.

3 Two-step pseudo likelihood estimation

From equation (2), the marginal likelihood of β and $\mathbf{D}$ can be interpreted as an expectation, $E\left(\exp\left[l\right]\right)$, with respect to the multivariate normal distribution of $\mathbf{u}$. Using the multivariate version of Taylor's expansion of the function $m(\mathbf{u}) = \exp\left[l\right]$ around the marginal mean of $\mathbf{u}$, which is 0, we have

$$L(\beta,\ \mathbf{D}|\mathbf{Y}) = E\left(m(\mathbf{u})\right) \approx m\left(\mathbf{0}\right) + \mathbf{0} + \frac{1}{2}E\left\{\mathbf{u}^T m^{(2)}\left(\mathbf{u}\right)_{\mathbf{u}=\mathbf{0}}\mathbf{u}\right\} + ct \qquad (4)$$

where, $m^{(2)}\left(\mathbf{u}\right) = \frac{\partial^2 m(\mathbf{u})}{\partial \mathbf{u}\partial \mathbf{u}^T}$ and ct represents the correction terms.

Regarding the estimation of the fixed effects parameters, β, already the second order term in equation (4) is flat and commonly ignored in the PQL methods. After ignoring the second and higher order term, the likelihood for β becomes a likelihood of the GLM having no random effect left in it. Thus, it suggests estimating the β parameters through a simple GLM procedure. This is not surprising since Maddala (1987) concluded that the fixed effect approach can produce a consistent estimate of the β parameters even if there is an autocorrelation in the model due to the random effects, while Alam and Carling (2007) supported Maddala (1987)'s claim by empirical evidence.

In order to make the scenario more clear we plot the log-likelihood for β of a simple logistic mixed model. Assume, a binary response variable, y_{ijt}, presents the defaults status of the $i^{th}\,(i = 1, 2, \cdots, n_{jt})$ loan in industry $j\,(j = 1, 2, \cdots, k)$ at time $t\,(t = 1, 2, \cdots, T)$ and is modeled through a single covariate, $\{x_{ijt}\}$, *i.e.* $\eta_{ijt} = x_{ijt}\beta + Z_{ijt}\mathbf{u_t}$ where, $\mathbf{u_t} \sim iid\ N_k\left(\mathbf{0}, \mathbf{D}\right)$. In simulation, we use $\mathbf{D} = \mathbf{I}$, $x_{ijt} \sim N(0, 1)$, $\beta = 0.5$, $k = 3$, $T = 20$ and 40, $n_{jt} = 200$. Figure 1 present the plots of the conditional log-likelihoods for the single fixed effect parameter, β, evaluated with $T = 20$ (Fig. 1 left panel) and 40 (Fig. 1, right panel), at the true value of the random effects, *i.e.* $\mathbf{u} = \mathbf{u}_{true}$, (solid line) and at its marginal mean *i.e.* $\mathbf{u} = \mathbf{0}$ (broken line). Fig. 1 reveals that both the conditional log-likelihoods have their maximum at the same value for β hence they should provide the same estimate for it.

The estimation of the covariance parameters and the random effects are yet to be done. Assume, $h\left(\mathbf{u}\right)$ has a single maxima at $\mathbf{u} = \widetilde{\mathbf{u}}$. Then, applying

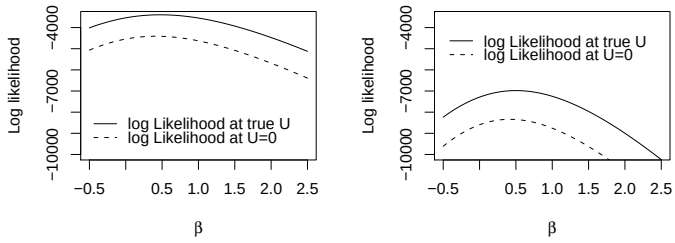


Fig. 1. Conditional log-likelihood for β of a logistic GLMM with $T = 20$ (left panel) and $T = 40$ (right panel)

the multivariate version of the Laplace approximation (Evans and Swartz (1995)) in equation (2) we have

$$log\left(L(\beta,\ \mathbf{D}|\mathbf{Y})\right) = -\frac{1}{2}log\left(|\mathbf{D}|\right) - \frac{1}{2}log\left\{-|H_h\left(\widetilde{\mathbf{u}}\right)|\right\} + h\left(\widetilde{\mathbf{u}}\right) \tag{5}$$

where,

$$H_h\left(\widetilde{\mathbf{u}}\right) = -\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/\mathbf{a}\left(\phi\right) - \mathbf{D}^{-1} \tag{6}$$

is the Hessian of $h\left(\mathbf{u}\right)$ and $\widetilde{\mathbf{W}}$ is the diagonal weight matrix (McCullagh and Nelder (1989)) evaluated at $\mathbf{u} = \widetilde{\mathbf{u}}$. Since $\widetilde{\mathbf{u}}$ is the maxima of $h(\mathbf{u})$, it can be obtained by solving $\frac{\partial h(\mathbf{u})}{\partial \mathbf{u}}|_{\mathbf{u}=\widetilde{\mathbf{u}}} = 0$ for which a Newton-Raphson algorithm leads us to solve the following equation iteratively.

$$\widetilde{\mathbf{u}}_{r+1} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}_r}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}^T\widetilde{\mathbf{W}}_r\left(\mathbf{Y}^* - \mathbf{X}\beta\right) \tag{7}$$

where, $\mathbf{Y}^*$ is the linearized version of the response variable which is given as: $\mathbf{Y}^* = \widetilde{\mathbf{W}}^{-1}\left(\mathbf{Y} - \widetilde{\mu}\right) + \widetilde{\eta}$ and the "r"'s in the subscript indicate that the matrix/vector is evaluated at the r^{th} iteration when $\mathbf{u} = \widetilde{\mathbf{u}}_r$, $(r = 0, 1, ...)$. This $\widetilde{\mathbf{u}}$ happens to produces the predicted values of the random effects vector, $\mathbf{u}$, given the data.

The covariance matrix of the random effects, $\mathbf{D}$, can be estimated by maximizing equation (5) given a particular value of $\widetilde{\mathbf{u}}$ and $\widehat{\beta}$. To simplify the calculation we consider $a\left(\phi\right) = 1$ which is true for binomial and Poisson distributions. After taking matrix differentiation of (5), equating it to zero, and simplifying, using Magnus and Neudecker (1999), it can be shown that

$$\widehat{\mathbf{D}} = \left(-\mathbf{H}_{h(\mathbf{u})}\right)^{-1} + \widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T \tag{8}$$

Equations (7) and (8) lead us to conclude that, the covariance parameters of the random effects can be estimated consistently along with the random effects through a joint iterative procedure. Since the proposed estimation

procedure estimated fixed effects parameters independently of $\mathbf{u}$ and $\mathbf{D}$, let us call it a two-step pseudo likelihood (2PL) approach. The 2PL reduces the computational effort of conventional PQL since it uses independent estimation for fixed effects while in PQL has to check the convergence of all the model parameters the same time. Moreover, an initial estimate of $\mathbf{D}$ through the fixed effects approach as proposed in Alam and Carling (2007) might reduce the computational burden to some extent.

It should be noted here again that the procedure for the models with a free dispersion parameter, $a(\phi)$, is not discussed in this paper. With non-canonical link, the calculations presented in this section become more complex and are avoided.

The right hand side of equation (8) can be explained as $E(V(\mathbf{u}|\mathbf{Y})) + V(E(\mathbf{u}|\mathbf{Y}))$ since $\widetilde{\mathbf{u}} = E(\mathbf{u}|\mathbf{Y})$ and $E(\widetilde{\mathbf{u}}) = 0$. From such a structure the consistency arguments for $\widehat{\mathbf{D}}$ is very straightforward. Furthermore, a relation of $E(V(\mathbf{u}|\mathbf{Y}))$ to the Hessian of $h(\mathbf{u})$ reveals that $E(V(\mathbf{u}|\mathbf{Y})) \to 0$ as $n_{ij} \to \infty$. This means that given a large data set we can estimate the covariance matrix, $\mathbf{D}$, by using $\widetilde{\mathbf{u}}$ only. This feature justifies the fixed effect approach suggested in Alam and Carling (2007) for estimating $\mathbf{D}$. In doing so equation (6) can be used to check the applicability of the fixed effect approach.

4 Application with credit defaults data

This section illustrates some merits of the 2PL algorithm through an application with a real data set on credit defaults. The data were collected from two major Swedish banks and were analyzed, for the first time, by Carling, *et al.* (2004). The data set contains quarterly information, between the 2^{nd} quarter of 1994 and the 2^{nd} quarter of 2000, on the borrowing companies' financial status, bank data on loan types, credit bureau data, two macro economic variables and an indicator variable stating whether a loan is default by a certain quarter. The research interest involved with the data analysis was to derive a credit risk model by incorporating industry specific defaults correlation. In Carling *et al.* (2004), only the within industry defaults correlation was considered while this paper aims at investigating the possibility of both within and between industry correlations.

In order to be consistent with SNI industry definition[1], this paper redefines the industries by merging the SNI codes of the companies, at the first two digits level, in the way that mimics Carling *et al.*(2004)'s industry definition. Furthermore, the number of industries is reduced from 7 (in Carling *et al.*(2004)) to 6 to avoid non-convergence problem.

Following Carling *et al.*(2004) and considering a between industry correlation a logistic GLMM is proposed where $y_{ijt}|u_{jt} \sim iid\ Bin(1, p_{ijt})$, $\log\left(\frac{p_{ijt}}{1-p_{ijt}}\right) = \mathbf{X}_{ijt}\beta + \mathbf{Z}_{ijt}\mathbf{u}_t$, $\mathbf{u}_t = (u_{1t}, u_{2t}, \cdots, u_{kt})^T$, $\mathbf{u}_t \sim iid\ \mathbf{N}_k(\mathbf{0}, \mathbf{D})$,

[1] www.scb.se.

$i = 1, 2, ..., n_{jt},\ j = 1, 2, ..., k = 6$ (number of industries), $t = 1, 2, ..., T = 25$ (number of quarters). The list of covariates are given in Table 1. To insure comparability, Carling *et al.* (2004)'s model is reanalyzed with logistic link and with the new 6 industries. The results are presented in Tables 1 and 2. We denote the model with a diagonal $\mathbf{D}$ as PQLD which resembles the model used in Carling *et al.* (2004). The same model with an unstructured $\mathbf{D}$ is denoted as PQLU. The above model with cluster effects as fixed effects is also estimated and is denoted as FE.

Table 1 shows that most of the fixed effects parameter estimates are very close between the three models. Except for the coefficient associated with "Bank A", the big differences in estimates are found only for the insignificant coefficients (see rows 2-6 in Table 1). The above feature indicates that the fixed effects parameter estimates are not much sensitive to these three types of model specifications.

The covariance matrix, $\mathbf{D}$, estimated through PQLD is a diagonal matrix with the estimates of the diagonal elements being (0.3577, 0.2592, 0.2514, 0.1097, 0.2312, 0.4203). However, the $\mathbf{D}$ matrix estimated by PQLU is an unstructured matrix (see Table 2). The covariance matrix, $\mathbf{D}$, is also estimated as

$$\widehat{\mathbf{D}} = \frac{1}{T}\sum_{i=1}^{T} \widetilde{\mathbf{u}}_\mathbf{t}\widetilde{\mathbf{u}}_\mathbf{t}^\mathbf{T} \tag{9}$$

where, $\widetilde{\mathbf{u}}_t$'s are the realizations of the random effects obtained from PQLD. The estimates of the covariance matrix of the random effects, $\mathbf{D}$, obtained from PQLD, computed with equation (9), and from PQLU, computed with 2PL method, are given in Table 2.

Table 2 shows that the covariance parameters estimates are not much different between the two models except for those parameters regarding 4^{th} industry. The close similarity in the above covariance matrix is not surprising in this case since the Hessian matrix for $\mathbf{u}$ was point wise very closed to zero. The non zero diagonal elements of $\mathbf{D}$ estimated by the both approaches indicate the existance of both within and between industry defaults correlation.

The computational time is often a matter of great concern especially for the large data cases. The estimation of PQLD implemented in SAS 9.1 using %GLIMMIX macro required 44 minutes (appr.) on a Pentium 4 PC (3.19 GHz processor, 0.99 GB RAM). On the contrary, the estimation of the PQLU, implemented partially in SAS and partially in R, took 29 minutes (appr.). While comparing the computation time it should be noted that PQLD estimated 6 covariance parameters while PQLU estimated 21 covariance parameters.

5 Concluding discussion

The proposed 2PL estimation is fast, in comparison to standard PQL approaches, and is applicable for the GLMMs with random effects being corre-

Covariates	Models					
	PQLD		PQLU		FE	
	Estimate	Std. Er.	Estimate	Std. Er.	Estimate	Std. Er.
Intercept	-4.44	0.214	-4.34	0.175	-3.90	0.378
Credit Survived 1 Yr.	-0.18	0.137	-0.30	0.132	-0.18^{ns}	0.136
Credit Survived 2 Yr.	0.03^{ns}	0.137	0.05^{ns}	0.132	0.02^{ns}	0.137
Credit Survived 3 Yr.	0.13^{ns}	0.138	-0.16^{ns}	0.133	0.17^{ns}	0.138
Credit Survived 4 Yr.	0.28^{ns}	0.136	0.22^{ns}	0.132	0.28^{ns}	0.135
Credit Survived 5+ Yr.	0.30^{ns}	0.148	0.10^{ns}	0.141	0.31^{ns}	0.148
Short term credit	0.53	0.039	0.50	0.039	0.54	0.039
Long term credit	-0.32	0.051	-0.30	0.050	-0.31	0.051
Mixed credit	0.00	NA	0.00	NA	0.00	NA
Account. data complete	-2.60	0.090	-2.53	0.088	-2.61	0.090
,, Reported previously	0.73	0.087	0.74	0.084	0.73	0.087
,, Reported afterward	-3.55	0.242	-3.44	0.240	-3.56	0.241
,, Missing	0.00	NA	0.00	NA	0.00	NA
Bank A	-0.09	0.040	-0.18	0.035	-0.08^{ns}	0.041
Remarks 8,11,16,25	1.09	0.055	1.08	0.054	1.09	0.055
Remark 25	1.11	0.077	1.11	0.075	1.12	0.076
Sales data missing	0.85	0.120	0.85	0.115	0.86	0.120
Sales ($\log_e$)	-0.04	0.005	-0.04	0.005	-0.04	0.005
Earnings/Sales	-0.25	0.038	-0.24	0.038	-0.25	0.038
Inventory/Sales	0.54	0.106	0.53	0.102	0.53	0.106
Loan/Asset	1.02	0.041	1.04	0.040	1.01	0.041
Output gap	-0.18	0.024	-0.19	0.010	NA	NA
Slope of yield curve	-0.23	0.060	-0.26	0.021	NA	NA

Note 1. ^{ns} stands for not significant at 2.5% level. No. of observations = 950693.

Table 1. Fixed effects parameter estimates.

lated between subjects. The derivation of 2PL provides an analytical expression (6), as a by-product, to check the applicability of the fixed effects approach (Alam and Carling (2007)) for estimating the parameters of a GLMM.

Approximate likelihood methods, *e.g.* PQL and h-likelihood (Lee and Nelder (1994)), are widely criticized especially for binary response models (see *e.g.* Engel (1998)). However, we should note that the Laplace approximation used in PQL approaches is an asymptotic approximation (Evans and Swartz (1995)) therefore they should be judged on the basis of their large sample performances which the critics of PQL have hardly considered. Moreover,

D matrix estimated through PQLU						**D** matrix estimated through PQLD					
0.28	0.16	0.18	0.16	0.16	0.26	0.31	0.16	0.18	0.09	0.15	0.21
0.16	0.25	0.25	0.17	0.23	0.15	0.16	0.24	0.22	0.09	0.18	0.13
0.18	0.25	0.26	0.18	0.22	0.16	0.18	0.22	0.23	0.09	0.17	0.13
0.16	0.17	0.18	0.14	0.15	0.13	0.09	0.09	0.09	0.05	0.06	0.06
0.16	0.23	0.22	0.15	0.22	0.16	0.15	0.18	0.17	0.06	0.19	0.13
0.26	0.15	0.16	0.13	0.16	0.36	0.21	0.13	0.13	0.06	0.13	0.31

Table 2. Estimated covariance matrix of the random effects.

for large sample cases, simulation based computations, *e.g.* MCMC methods, for GLMM are awfully time consuming. High speed computers are becoming available but still there is a need for some approximate method which can handle large data sets within a reasonable time limit.

References

ALAM, M.M. and CARLING, K. (2007): Computationally feasible estimation of the covariance structure of the generalized linear mixed models (GLMM). *Journal of Statistical Computation and Simulation* (to appear).

BRESLOW, N.E. and CLAYTON, D.G. (1993): Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88, 9-25.*

CARLING, K., RÖNNEGÅRD, L. and ROSCZBACH, K. (2004): Is firm interdependence within industries important for portfolio credit risk? Sverige Riksbank Working Paper Series No. 168.

ENGEL, B. (1998): A simple illustration of the failure of the PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal 40(2), 141-154.*

EVANS, M. and SWARTZ, T. (1995): Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science 10(3), 254-272.*

LAM, K.F. and LEE, Y.W. (2004): Merits of modelling multivariate survival data using random effect proportional odds model, *Biometrical Journal 46(1), 331-342.*

LEE, Y. and NELDER, J.A. (1996): Hierarchical generalized linear models. *Journal of the Royal Statistical Society (B) 58, 619-678.*

LINDSTROM, M.J. and BATES, D.M. (1988): Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association 43 (404), 1014-1022.*

MADDALA, G.S. (1987): Limited dependent variable models using panel data. *Journal of Human Resources 22(3), 307-338.*

MAGNUS, J.R. and NEUDECKER, H. (1999): *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Wiley, New York.

McCULLAGH, P. and NELDER, J.A. (1989): *Generalized Linear Models*. Chapman and Hall, London.

McCULLOCH, C.E. and SEARLE, S.R. (2001): *Generalized Linear and Mixed Models*. Wiley: New York.

SEARLE, S.R., CASELLA, G. and McCULLOCH, C.E. (1992): *Variance Components*. Wiley, New York.

WOLFINGER, R. and O'CONNELL, M. (1993): Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation 48, 233-243*.

The SVM Approach for Box-Jenkins Models

Saeid Amiri[1], Dietrich von Rosen[2], and Silvelyn Zwanzig[3]

[1] Center of Biostochastics, Swedish University of Agriculture Sciences P.O.Box 7032, SE 750 07 Uppsala, Sweden, *saeid.amiri@et.slu.se*
[2] Center of Biostochastics, Swedish University of Agriculture Sciences P.O.Box 7032, SE 750 07 Uppsala, Sweden, *Dietrich.von.Rosen@et.slu.se*
[3] Uppsala University, Department of Mathematics, Box 480, SE-751 06 Uppsala, Sweden , *zwanzig@math.uu.se*

Abstract. Support Vector Machine (SVM) is known in classification and regression modeling . It has been receiving attention in the application of nonlinear functions. The aim of this article is to motivate the use of the SVM approach to the analysis of the time series models. We discuss the efficiency of SVM in comparison with ARMA model. The applicability of this approach for a unit root situation is also considered, which violates the stationarity.

Keywords: support vector machine, time series analysis, unit root

1 Introduction

A time series analysis is the study of observations made sequentially in time. It is a complicated field in statistics because of direct and indirect effects of time on the variables in the model. The essential difference between the modeling via time series and ordinary method is that data points taken over time may have an internal relation that should be accounted for. It can be a correlation structure, a trend, seasonality and so on.
Time series can be studied in the time domain and in the time frequency domain. The time domain is more known among researchers in sciences whereas the frequency domain has many applications in engineering. The time domain is modeled by two main approaches. The traditional approach that was advocated by Box and Jenkins (1970) in their influential book, includes a systematic class of models called autoregressive integrated moving average (ARIMA). Shumway (2000), explains this model in detail. A defining feature of these models is that they are multiplicative models, meaning that observed data are assumed to result from products of factors involving differential or difference equation operators responding to a white noise input. The other approach uses additive models or structural models. In this approach, it is assumed that observations are from sum of components, each with a specified time series structure.
None of them has inferential tools such as the Box-Jenkins model, for example model selection, parameter estimation and model validation. Therefore

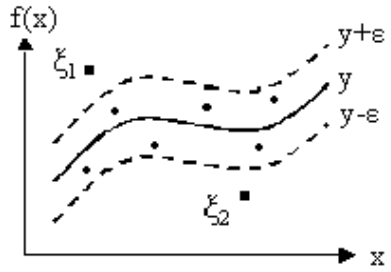


Fig. 1. SVM regression with insensitive tube, slack variables ξ_1, ξ_2 and observations.

when using a new method in time series one should compare the results with this classical model. Support Vector Machine is one of the new methods in modeling that has good performance in classification and regression analysis, unfortunately a few papers have tried to use the special case of time series, see Müller (1997) and Murkharejee (1997). These two paper consider dynamic models e.g., the Mackey class equation is used to show the efficiency of SVM. In this paper, the concept of SVM regression is reviewed. Then the approach of time series modeling is used to show that it can be written as a SVM model. At the end, the performance of SVM for time series data are studied by using real and simulated data.

2 Support Vector Machine

During the last decades many people have been working on it in a variety of fields. SVM has impact on improving the statistical learning method and has been used to solve problems in classification. Concerning SVM has literature on modeling, especially for nonlinear models. The review of Burges (1998), Cristianini and Shaw-Taylor (2000) and Bishop (2006) help to understand the basic concept of SVM. For more details see Vapnik (1995) and Vapnik (1998). In this section, we briefly present SVM regression.

In statistics, the aim of modeling is often to find a function $f(x)$ which predicts y in a model $y = f(x) + error$. It is not easy to find $f(x)$. Using mathematical methods, we can interpolate and approximate by using statistical methods. Via some statistical criteria like sum square or ML, the model can be exploited. To evaluate our procedure, we need a criterion or loss function. Here it is defined as "ignoring observation which error is less than ϵ",

so

$$L(x, y, f) = |y - f(x)|_\epsilon = max(0, |y - f(x)| - \epsilon).$$

It is called "ϵ-insensitive error function". Another loss function is Huber's loss function which is the squared distance between the observations and the function, see Cristianini and Shaw-Taylor (2000) and Hasti et al. (2001). In Figure 1. the points outside the tube around the function are called slack variables which clarify as ξ_{1i} and ξ_{2j} for above and below the tube, respectively. The value of points inside the tube is zero and outside the tube is nonzero. To find ξ_{1i} and ξ_{2j}, one should estimate parameters by the error function as below,

$$\begin{aligned} &\text{minimize} \quad \textstyle\sum_{i=1}^{N}(\xi_{1i} + \xi_{2i}) + \frac{\lambda}{2} \parallel W \parallel^2, \\ &\text{subject to } y_i \leq f + \epsilon + \xi_{1i}, \\ &\qquad\qquad y_i \geq f - \epsilon - \xi_{2i}, \\ &\qquad\qquad \xi_{1i}, \xi_{2j} \geq 0. \end{aligned}$$

By using the Lagrange multiplier to find parameters and optimize by the Karush-Kuhn-Tucker condition, $f(x)$ can be shown as below:

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i), \tag{1}$$

where α_i are support vectors, i.e. those points that contribute in the prediction. All points within the tube have $\alpha_i = 0$ and a few of α_i are nonzero. In (1), $k(x, x_i)$ is the kernel function, which is inner product of variables i.e.,

$$k(x, x_i) = < \phi(x), \phi(x_i) > \tag{2}$$

The followings are some kernels:

Linear kernel	$k(x, x') = < x, x' >$,
Polynomial kernel	$k(x, x') = (a < x, x\prime > +k)^d$,
Radial Basis Function kernel (RBF)	$k(x, x') = exp(-\sigma \|x - x'\|^2)$,
Laplacian kernel	$k(x, x') = < x, x' > exp(-\sigma \|x - x'\|)$.

Other kernels are the hyperbolic tangent kernel, the spline kernel, the Bassel and the ANOVA RBF kernel. The number of kernels is unlimited and one can define new kernels by combining them. Discussions about them can be found in Burges (1998), Shaw-Taylor (2000) and Karatzoglou et al. (2007). Some advantages and disadvantages are the following. SVM is based on kernels so finding suitable kernels is most important. However in practice one needs to study only a few kernel functions, Burges (1998). The key in SVM is the transformation of a nonlinear problem to a higher dimensional linear space using the kernel function. SVM is not based on any distribution assumption.

3 Time series analysis

Following the Box-Jenkins approach, it involves identifying an appropriate ARIMA process by a mathematical model for forecasting. The model is a combination of AR and MA models. $AR(p)$ is defined as bellow,

$$x_{t+1} = \sum_{j=1}^{p} \phi_j x_{t+1-j} + \epsilon_{t+1}. \tag{3}$$

Many real data are not linear. Hence it is better to find an appropriate function of the observations which is not linear, because it will be closer to reality. The nonlinear model can be written as:

$$x_{t+1} = \sum_{j=1}^{p} \phi_j h_j(x_{t+1-j}) + \epsilon_{t+1}, \qquad \epsilon_{t+1} \sim N(0, \sigma^2), \tag{4}$$

$$x_{t+1} = (h_1(x_t), \ldots, h_p(x_{t+1-p})) \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}, \tag{5}$$

$$x = H\phi. \tag{6}$$

If H is known, the parameters can be estimated. To simplify assume $\underline{\mathrm{x}}_t = (x_t, x_{t-1}, \ldots, x_{t+1-p})$, $p < t$. The parameters of the model can be estimated by the conditional ML:

$$\begin{aligned} L(\phi, \sigma|\underline{x}_p) &= f(x_{p+1}|\underline{\mathrm{x}}_p) f(x_{p+2}|\underline{\mathrm{x}}_{p+1}) \ldots f(x_t|\underline{\mathrm{x}}_{t-1}) = \prod_{i=p}^{t-1} f(x_{i+1}|\underline{\mathrm{x}}_i) \\ &= \prod_{i=p}^{t-1} \frac{1}{\sqrt{2\pi}\sigma} \quad exp - \frac{(x_{i+1} - \sum_{j=1}^{p} \phi_j h(x_{i+1-j}))^2}{2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{(t-p)/2} exp - \sum_{i=p}^{t-1} \frac{(x_{i+1} - \sum_{j=1}^{p} \phi_j h(x_{i+1-j}))^2}{2\sigma^2} \end{aligned} \tag{7}$$

Thus one needs to minimize,

$$SS = \sum_{i=p}^{t-1} (x_{i+1} - \sum_{j=1}^{p} \phi_j h_j(x_{i+1-j}))^2 = \sum_{i=p}^{t-1} (x_{i+1} - H_i\phi)^2 \tag{8}$$

To improve estimation, one can use penalty:

$$SS2 = \sum_{i=p}^{t-1} (x_{i+1} - H_i\phi)^2 + \lambda\|\phi\| = (x - H\phi)^T (x - H\phi) + \lambda\|\phi\|, \tag{9}$$

$$\frac{\partial SS2}{\partial \phi} = 0 \Rightarrow -H^T(x - H\phi) + \lambda\phi = 0,$$

which implies that

$$H\phi = (HH^T + \lambda I)^{-1} HH^T y, \tag{10}$$

where HH^T is matrix of inner product of observation. Thus (6) can be written as inner product. By considering (2), we can rewrite the nonlinear equation as kernel function,

$$x_{t+1} = f(\underline{x}_t) + e_{t+1} = \sum_{i=1}^{p} \phi_i h_i(x_{t+1-i}) + e_{t+1} = \sum_{i=1}^{t} \alpha_i k(\underline{x}_t, \underline{x}_i) + e_{t+1}. \tag{11}$$

Another formula that can be considered is using time directly as independent in model, it is logical because the time series data are collected during time,

$$x_t = \sum_{i=1}^{t} \alpha_i k(\underline{x}_t, i). \tag{12}$$

MA(q): This model is as below,

$$x_t = \sum_{j=0}^{q} \theta_j w_{t-j}, \qquad w_t \sim N(0, \sigma^2). \tag{13}$$

We can follow previous method and use nonlinear of this model,

$$x_t = \sum_{j=0}^{q} \theta_j h(w_{t-j}).$$

Unfortunately, it is difficult to decide about the distribution of $h(.)$ in advance. If we assume $h(w_{t-j}) \sim N(\mu_n, \sigma_n^2)$, there is no improvement for modeling. However if this model is invertible, we can write MA as AR and follow the previous model. Hence we have two problems. The distribution of $h(.)$ and invertibility of it that make unclear the behavior of MA for using kernel. Also about $ARMA(p, q)$, there are two viewpoints, first the ignorance of MA and consideration it as AR or if $ARMA(p, q)$ is invertible, $ARMA$ can be written as AR. Hence interpretation of AR can be used.

Unit root is as follow,

$$x_t = \mu + x_{t-1} + w_t = \mu + \mu + x_{t-2} + w_{t-1} + w_t = t\mu + x_0 + \sum_{i=0}^{t} w_i. \tag{14}$$

This is a problem for the Box-Jenkins approach because it violates the stationarity condition. It can not be formulated by the Box-Jenkins model, see Brockwell and Davis (1991). The modeling of the unit root has been discussed a lot in econometrics. There exists some statistical tests for diagnosis

and also modeling in the special conditions. The equation (14) explains that the unit root has a regression form of time but because of dependency between observations, the common regression can not be used for it but SVM is not based on the distribution hence the dependency does not affect on it therefore by using the previous discussion and rewriting it as kernel formula, SVM can be used to analyze it. However if $\mu = 0$, then this model has big problem and its behavior is completely random.

4 Applications

In this section, some models are studied by using the Residual Sum Square (RSS) and Akaike Information Criterion (AIC). AIC is calculated based on $ln\widehat{\sigma}_k^2 + \frac{2k}{n}$, where $\widehat{\sigma}_k^2 = \frac{RSS}{n}$. The k and n are number of parameters and observations, respectively. These results admit the discussion of previous section by using examples from $AR(2)$, $MA(1)$ and $ARMA(2,1)$.

4.1 *AR*

The example is chosen from Brockwell and Davis (1991), example 9.2.1. It includes 200 observations. Table 1 shows RSS and AIC of $AR(2)$ and SVM with different kernel. Here SVM is calculated by equation (11). In this table, the result of a few kernels are presented because the RSS of many kernels is more larger than $AR(2)$. It shows the efficiency of laplacian kernel in comparison with the Box-Jenkins modeling. Also RBF with $\sigma = 50$ fitted fairly good.

Table 1. RSS and AIC of AR(2) and SVM.

Model	RSS	AIC
AR(2)	176.99	-0.102
RBF[1]	171.73	-0.136
RBF[2]	144.33	-0.368
Bessel[1]	161.16	-0.176
Bessel[2]	194.46	0.009
Laplacian[1]	100.83	-0.664
Laplacian[2]	202.68	0.33
linear	177.75	-0.102
poly[3]	176.43	-0.085

[1]Fitted by $\sigma = 10$. [2]Fitted by $\sigma = 50$. [3]With 2 degree.

The calculations in Table 2 are based on equation (12). This model uses the time as independent variable. The table shows how much fitting is improved. By using this model, the Laplacian kernel and Bessel kernel have

smaller RSS than AR but other kernels are larger than AR. These values show Bessel kernel is fitted approximately to our example but its change is very large. The changes of Laplacian kernel is small in comparison with the Bessel kernel, so it seems more trustable. The Laplacian kernel for this model is better than the previous model.

Table 2. Modeling directly based on time.

Model	RSS	AIC
Laplacian[1]	56.6	-1.252
Laplacian[2]	21.55	-2.217
Bessel[1]	29.5	-1.83
Bessel[2]	980.17	1.619

[1]Fitted by $\sigma = 10$. [2]Fitted by $\sigma = 50$.

Table 3. Percent and rank of model in simulation.

	model based on x_t		model based on t	
Model	percent	rank	percent	rank
AR(2)	0.020	6.930	0.006	2.93
RBF[1]	0.283	3.67	0.00	9.18
RBF[2]	0.000	4.43	0.00	6.00
Bessel[1]	0.023	3.77	0.00	7.90
Bessel[2]	0.000	5.85	0.994	1.006
tanhdot	0.000	12.51	0.000	12.636
tangant1	0.000	12.49	0.000	12.536
splinedot	0.000	14.51	0.000	14.423
spline1	0.0000	14.48	0.000	14.36
Laplacian[1]	0.540	2.17	0.000	3.923
Laplacian[2]	0.003	6.27	0.000	2.143
linear	0.020	6.36	0.000	10.210
poly[3]	0.110	5.52	0.000	10.220
ANOVA[1]	0.000	10.98	0.000	7.526
ANOVA[2]	0.000	10.013	0.000	4.99

[1]Fitted by $\sigma = 10$. [2]Fitted by $\sigma = 50$. [3]With 2 degree.

Moreover, consider $AR(2)$ with $x_t = x_{t-1} - 0.9x_{t-2} + \omega_t$. This model is stationary. The Box-Jenkins model fits very well. To compare the Box-Jenkins

model with SVM, the simulation of this model are performed 1000 times with 100 observations. The results for the Box-Jenkins model and different kernels are shown in Table 3. The first two columns include the results of using (11) and the second two columns include the results of using (12). The column of rank shows the mean of rank of model in all of the simulations and the percent shows how many times the model has the smallest RSS in all of the simulations. Table 3 shows that by using x_t, the Laplacian kernel in 54% time has minimum but by using time as explanatory variable, Bessel kernel has minimum RSS. The results of table 3 is the same as tables 1. Therefore the Bessel and Laplacian kernel are suitable for AR. Also Table 2 shows model fits better by using the time as the explanatory variable instead of x_t.

4.2 *MA*

The example 10.4.2 of Brockwell and Davis (1991) is $MA(1)$ with 160 observations, which is used to study the modeling of SVM. The results of this modeling are presented in Table 4.

Table 4. RSS and AIC of $MA(1)$ and SVM.

Model	RSS	AIC
$MA(1)$	147	-0.072
Bessel[1]	227.373	0.388
Bessel[2]	198.415	0.252
Laplacian[1]	178.720	0.123
Laplacian[2]	79.282	-0.689

[1]Fitted by $\sigma = 10$. [2]Fitted by $\sigma = 50$.

The result shows the Laplacian kernel with large sigma fitted well to $MA(1)$ and also the RSS with Bessel kernel is near $MA(1)$ but other kernels have very large RSS. As it is expected from the discussion, SVM should be fitted to the AR model better than the MA model. For AR, the Laplacian kernel with small σ has smaller RSS but for MA, the Laplacian kernel with larger σ has smaller RSS. To complete it look at Table 5. It is the result of the simulation $y_t = \omega_t + 0.5\omega_{t-1}$ with 100 observations. This includes the rank and the percent of different models in comparison with the Box-Jenkins model. This illustrates the previous results which the Laplacian kernel with large σ in % 88 has fitted better than other with MA behavior.

4.3 *ARMA*

Here ARMA(2,1) with 200 observations is considered from Brockwell and Davis (1991), example 9.2.3. The Table 6 includes RSS and AIC of the ARMA

Table 5. Percent and rank of model in simulation of MA.

Model	percent	rank
MA(1)	0.000	8.08
RBF[1]	0.000	8.54
RBF[2]	0.000	5.002
Bessel[1]	0.000	6.512
Bessel[2]	0.112	2.90
Tanhdot	0.000	12.594
Tangant1	0.000	12.442
Splinedot	0.000	14.502
Spline1	0.000	14.462
Laplacian[1]	0.000	3
Laplacian[2]	0.888	1.11
linear	0.000	10.68
poly[3]	0.000	13.68
ANOVA[1]	0.000	6.89
ANOVA[2]	0.000	4

[1]Fitted by $\sigma = 10$. [2]Fitted by $\sigma = 50$. [3]With 2 degree.

(2,1) and different kernels. The first two columns include the results of using formula (11) and the second two columns include the results of using formula (12). It admits the efficiency of Laplacian kernel for the ARMA. The Laplacian kernel has the smallest RSS in both formula.

Table 6. RSS and AIC of ARMA and SVM.

	model based on x_t		model based on t	
Model	RSS	AIC	RSS	AIC
ARMA(2,1)	197.169	0.0157		
RBF[1]	244.168	0.209	1536.55	2.048
RBF[2]	176.267	-0.008	1216.10	1.815
Bessel[1]	201.50	0.037	1460.53	2.018
Bessel[2]	195.39	0.006	56.82	-1.228
Laplacian[1]	116.96	-0.526	350	0.569
Laplacian[2]	200.143	0.0107	46.76	-1.443

[1]Fitted by $\sigma = 10$. [2]Fitted by $\sigma = 50$. [3]With 2 degree.

To simulate ARMA(2,1), consider $x_t = 0.4x_{t-1} + 0.5x_{t-2} + \omega_t + 0.2\omega_{t-1}$. The simulation is done with 100 observations in 1000 times. The results of ARMA(2,1) by using the Box-Jenkins and different kernel are presented in Table 7. These results agree with the Table 6, which is based on one time series data. It depicts that in both modeling, equation (11) and (12), the Laplacian kernel has better results than others. By using x_t and time as explanatory variables, Laplacian kernel with $\sigma = 10$ has smallest RSS in % 923 and %66 times of simulations, respectively.

Table 7. Percent and rank of model in simulation of ARMA(2,1).

	model based on x_t		model based on t	
Model	rank	percent	rank	percent
ARMA(2,1)	0.000	8.803	0.000	9.00
RBF[1]	0.020	5.143	0.000	7.99
RBF[2]	0.000	3.33	0.000	4.93
Bessel[1]	0.000	4.23	0.000	6.47
Bessel[2]	0.002	3.86	0.044	2.33
Tanhant[1]	0.000	12.56	0.000	12.606
Tangant[2]	0.000	12.46	0.000	12.39
Spline [1]	0.000	14.56	0.000	14.473
Spline[1]	0.000	14.4	0.000	14.526
Laplacian[1]	0.923	1.14	0.660	1.483
Laplacian[2]	0.045	3.81	0.296	2.39
Linear	0.000	9.51	0.000	10.876
Poly[3]	0.000	8.676	0.000	10.123
ANOVA[1]	0.000	9.643	0.000	6.486
ANOVA[2]	0.000	7.83	0.000	3.90

[1]Fitted by $\sigma = 10$. [2]Fitted by $\sigma = 50$. [3]With 2 degree.

4.4 Unit root

To study the SVM for the unit root, the model $x_t = x_{t-1} + \omega_t$ is simulated 1000 times with 100 observations. The results of using SVM to this data are presented in Table 8. It includes the rank of model in comparison with the others and the percent that shows how many times the model has the smallest RSS in comparison with others. For these data, the study of $ARMA$ is impossible because of the non stationarity of data. This table depicts the Laplacian kernel is fitted very well to these data.

Table 8. Percent and rank of model in simulation of unit root.

Model	rank	percent
RBF[1]	0.00	7.68
RBF[2]	0.019	4.08
Bessel[1]	0.000	6.23
Bessel[2]	0.0367	2.77
tangant[1]	0.00	11.63
tangant[2]	0.000	11.36
spline[1]	0.00	13.57
spline[2]	0.000	13.42
Laplacian[1]	0.915	1.14
Laplacian[2]	0.002	5.68
linear	0.00	9.87
poly[3]	0.00	9.08
ANOVA[1]	0.0025	5.75
ANOVA[2]	0.0240	2.85

[1]Fitted by $\sigma = 10$. [2] Fitted by $\sigma = 50$. [3]With 2 degree.

5 Conclusion

Although the Box-Jenkins model is still one of the most applied model in time series but there is some problems. For examples, the Box-Jenkins models are based on stationarity but this is not often sufficient for example, consider a $AR(1)$ model:

$$x_t = \phi x_{t+1} + w_t.$$

It can be shown as $x_t = -\sum_{j=1}^{\infty} \phi^{-j} w_{t+j}$. Hence for $\|\phi\| > 1$ it is stationary but this is based on the future and not logical modeling process. Also the modeling unit root by $ARMA$ is impossible. The result of this study shows that SVM has good results in comparison with the Box-Jenkins modeling. Especially by using the Laplacian kernel and Bessel kernel. As the discussion shown in section 3, the use of SVM is logical for the ARMA model. Moreover the usage of time directly as explanatory variable in model fits very well, also it is logical because the time series data are collected based on the time. One of the outstanding result is about unit root, where SVM by using the Laplacian kernel is fitted well.

Acknowledgment

The authors gratefully acknowledges Dr. Mats Gustafsson for his encouragement and useful comments that led to the improvement of this paper.

References

BISHOP, C.M. (2006): *Pattern Recognition and Machine Learning*. Springer, New York.

BOX, G. and JENKINS, G. (1970): *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

BROCKWELL, P.J. and DAVIS, R.A. (1991): *Time Series: Theory and Methods*, 2nd ed. Springer, New York.

BURGES, C.J.C. (1998): A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery, 2, 121-167.*

CRISTIANINI, N. and SHAW-TAYLOR, J. (2000): *An introduction to Support Vector Machine*. Cambridge University Press, New York.

HASSTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *Elements of Statistical learning: Data Mining, Inference and Prediction*. Springer, New York.

KARATZOGLOU, A. et al. (2007). The kernel lab package. *http://cran.r-project.org/src/contrib/Descriptions/kernlab.html*

MÜLLER, K.R. et al. (1997): Predicting Time Series with Support Vector Machines. *ICANN'97, page 999-1004*. Berlin.

MURKHAREJEE, S. et al. (1997): Nonlinear Prediction of Chaotic Time Series using Support Vector Machines. *IEEE workshop on Neural Network for Signal Processing.*

SHUMWAY, R.H. and STOFFER, D.S. (2000): *Time Series Analysis and Its Applications*. Springer, New York.

VAPNIK, V.N. (1995): *The Nature of Statistical Learning Theory*. Springer, New York.

VAPNIK, V.N. (1998): *Statistical Learning Theory*. Wiley, New York.

Comparison of Financial Time Series Using a TARCH-Based Distance

Jorge Caiado[1, 2] and Nuno Crato[2]

[1] Department of Economics and Management, School of Business Administration, Polytechnic Institute of Setbal, *jcaiado@esce.ips.pt*
[2] Cemapre, Instituto Superior de Economia e Gestão
Rua Miguel Lupi 20, 1200 Lisboa, Portugal, *ncrato@iseg.utl.pt*

Abstract. This paper proposes an asymmetric-volatility based method for cluster analysis of stock returns. Using the information about the estimated parameters in the TARCH equation, we compute a distance matrix for the stock returns. Clusters are formed by looking to the hierarchical structure tree (or dendrogram) and the computed principal coordinates. We employ these techniques to investigate the similarities and dissimilarities between the "blue-chip" stocks used to compute the Dow Jones Industrial Average (DJIA) index.

Keywords: asymmetric effects; cluster analysis; DJIA stock returns; threshold ARCH model; volatility

1 Introduction

Cluster analysis of financial time series plays an important role in several areas of application. In stock markets, the examination of mean and variance correlations between asset returns can be useful for portfolio diversification and risk management purposes. In international equity markets, we may be interested in identifying similarities in index returns and volatilities for grouping countries. The existence of asymmetric cross-correlations and dependences in asset returns is also of interest for many financial researchers.

Many existing statistical methods for analysis of multiple asset returns use multivariate volatility models imposing conditions on the covariance matrix that are hard to apply. To avoid these problems, three types of multivariate statistical techniques have been used for analysing the structure of asset returns comovements. One is the principal component analysis (PCA) that is concerned with the covariance structure of asset returns and can be used in dimension reduction. The second is the factor model for asset returns that uses multiple time series to describe the common factors of returns (see Zivot and Wang, 2003 and Tsay, 2005 for further discussion). The third is the identification of similarities in asset return volatilities using cluster analysis (see, for instance, Bonanno, Caldarelli, Lillo, Micciché, Vandewalle and Mantegna, 2004).

A fundamental problem in clustering of financial time series is the choice of a relevant metric. Mantegna (1999), Bonanno, Lillo and Mantegna (2001), among others, used the Pearson correlation coefficient as similarity measure of a pair of stock returns. Although this metric can be useful to ascertain the structure of stock returns movements, it does not take into account the stochastic volatility dependence of the processes and cannot be used for comparison and grouping stocks with unequal sample sizes. The latter is a common problem of most existing nonparametric-based metrics for cluster analysis of economic and financial time series.

In this paper, we introduce a distance measure between the threshold autoregressive conditionally heteroskedastic (TARCH) parameters of the return series. In order to summarize and better interpret the results, we suggest using a hierarchical clustering tree and a multidimensional scaling map to explore the existence of clusters.We apply these steps to investigate the similarities and dissimilarities among the "blue-chip" stocks of the Dow Jones Industrial Average (DJIA) index.

The remaining sections are organized as follows. Section 2 provides the asymmetric-volatility based method for clustering asset returns. Section 3 describes the data. Section 4 presents the empirical findings on the analyzed data. Section 5 summarizes and concludes.

2 Asymmetric-volatility based distance

Many time-varying volatility models have been proposed to capture the asymmetric volatility effects in asset returns. These include the common univariate asymmetric models of Nelson (1991), Engle and Ng (1993), Glosten, Jagannathan and Runkle (1993) and Zakoian (1994), the multivariate generalized autoregressive conditionally heteroskedasticity (GARCH) models of Engle and Kroner (1995) and Kroner and Ng (1998), and the asymmetric dynamic autoregressive conditional correlation model of Capiello, Engle and Sheppard (2006).

Glosten, Jagannathan and Runkle (1993) and Zakoian (1994) introduced independently the Threshold ARCH model to allow for asymmetric shocks to volatility. The simple TARCH(1,1) model assumes the form

$$\varepsilon_t = z_t \sigma_t, \tag{1}$$

$$\sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \alpha \varepsilon_{t-1}^2 + \gamma \varepsilon_{t-1}^2 d_{t-1}, \tag{2}$$

where $\{z_t\}$ is a sequence of independent and identically distributed random variables with zero mean and unit variance, $d_t = 1$ if ε_t is negative, and $d_t = 0$ otherwise. In this model, volatility tends to rise with "bad news" ($\varepsilon_{t-1} < 0$) and to fall with "good news" ($\varepsilon_{t-1} > 0$). Good news has an impact of α while bad news has an impact of $\alpha + \gamma$. If $\gamma > 0$ then the leverage effect exists. If $\gamma \neq 0$, the shock is asymmetric, and if $\gamma = 0$, the shock is symmetric.

The persistence of shocks to volatility is given by $\alpha + \beta + \gamma/2$. Nelson (1991) proposed also an heteroskedasticity model to incorporate the asymmetric effects between positive and negative stock returns, called the exponential GARCH (or EGARCH) model, in which the leverage effect is exponential rather than quadratic. To capture all the skewness and excess kurtosis in the volatility processes with asymmetric distributions, Nelson (1991) suggested a "fat-tailed" distribution, the generalized error distribution (GED), with density function given by

$$f(z) = \frac{v \exp\left[-0.5\left|z/\lambda\right|^{v}\right]}{\lambda 2^{(1+1/v)}\Gamma(1/v)}, 0 < v \leq \infty, -\infty < z < +\infty \tag{3}$$

where v is the tail-tickness parameter, $\Gamma(\cdot)$ is the gamma function, and

$$\lambda = \left[\frac{2^{(-2/v)}\Gamma(1/v)}{\Gamma(3/v)}\right]^{0.5}. \tag{4}$$

When $v = 2$, $\{z_t\}$ is normally distributed, and is fat-tailed distributed if $v < 2$. For $v > 2$, it has thinner tails distribution (for example, for $v = +\infty$, it has a uniform distribution on the interval $[-\sqrt{3}, \sqrt{3}]$).

We now introduce a distance measure for clustering time series with similar asymmetric volatility effects. Let $r_{x,t} = \log P_{x,t} - \log P_{x,t-1}$ denote the continuously compounded return of an asset x from time $t-1$ to t ($r_{y,t}$ is similarly defined for asset y). Suppose we fit a common TARCH(1,1) model to both time series by the method of maximum likelihoods assuming GED innovations. Let $T_x^G = (\widehat{\alpha}_x, \widehat{\beta}_x, \widehat{\gamma}_x, \widehat{v}_x)'$ and $T_y^G = (\widehat{\alpha}_y, \widehat{\beta}_y, \widehat{\gamma}_y, \widehat{g}_y)'$ be the vectors of the estimated ARCH, GARCH, leverage effect and tail-tickness parameters, respectively, with the estimated covariance matrices given by V_x^G and V_y^G, respectively. A Mahalanobis-like distance between the asymmetric features of the volatilities (TARCH-based distance) of the return series $r_{x,t}$ and $r_{y,t}$ can be defined by

$$d_{TARCH}(x,y) = \sqrt{(T_x^G - T_y^G)'\Omega^{-1}(T_x^G - T_y^G)}, \tag{5}$$

where $\Omega = V_x^G + V_y^G$. This measure takes into account the information about the asymmetric structure of the time series volatilities and solves the problem of unequal lengths. The distance measure (5) fulfills the usual properties of a metric (except the triangle inequality): (i) $d(x,y)$ is asymptotically zero for independent time series generated by the same DGP; (ii) $d(x,y) \geq 0$; and (iii) $d(x,y) = d(y,x)$.

3 Data description

We consider data of the 30 "blue-chip" US daily stocks used to compute the Dow Jones Industrial Average (DJIA) index for the period from June

1990, 11 to September 2006, 12 (4100 daily observations), as shown in Table 1. This data was obtained from Yahoo Finance (http://finance.yahoo.com) and correspond to closing prices adjusted for dividends and splits. In Table

Table 1. Stocks used to compute the Dow Jones Industrial Average (DJIA) Index

Stock	Code	Sector	Stock	Code	Sector
Alcoa Inc.	AA	Basic mat.	Johnson & Johns.	JNJ	Healthcare
American Int. Gr.	AIG	Financial	JP Morgan Chase	JPM	Financial
American Express	AXP	Financial	Coca-Cola	KO	Cons. goods
Boeing Co.	BA	Ind. goods	McDonalds	MCD	Services
Caterpillar Inc.	CAT	Financial	3M Co.	MMM	Conglomerates
Citigroup Inc.	CIT	Ind. goods	Altria Group	MO	Cons. goods
El Dupont	DD	Basic mats.	Merck & Co.	MRK	Healthcare
Walt Disney	DIS	Services	Microsoft Corp.	MSFT	Technology
General Electric	GE	Ind. goods	Pfizer Inc.	PFE	Healthcare
General Motors	GM	Cons. goods	Procter & Gamble	PG	Cons. goods
Home Depot	HD	Services	AT&T Inc.	T	Technology
Honeywell	HON	Ind. goods	United Techs.	UTX	Conglomerates
Hewlett-Packard	HPQ	Technology	Verizon Comm.	VZ	Technology
Int. Bus. Machin.	IBM	Technology	Walt-Mart Stores	WMT	Services
Inter-tel Inc.	INTC	Technology	Exxon Mobile CP	XOM	Basic mats.

2 we present the estimation results of TARCH(1,1) models for DJIA stock returns with GED innovations, including diagnostic tests for residual and squared residuals. The estimated coefficients are statistically significant for all stocks except the ARCH estimates for CAT, DIS, GE and MRK, and the leverage-effect for INTC and MMM, which are not significant at conventional levels. The distribution of the innovation series is fat-tailed for all stocks. As expected, the persistent estimates for all the asymmetric models are very close to one. This extreme persistence in the conditional variance is very common in many empirical application using high frequency data (see Bollerselev, Chou and Kroner, 1992, and Kroner and Ng, 1998).

The Lagrange multiplier test statistic shows evidence of no serial correlation in the squared residuals up to order 20 for all stocks except CAT, MCD and VZ. In terms of the mean equation, the Ljung-Box test statistic does not reject the null hypothesis of no serial correlation in the residuals for all stocks except AIG, JNJ, PFE, UTX, VZ, and XOM.

Table 2. Estimated TARCH(1,1) models with conditional GED innovations for DJIA stock returns

Stock	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{v}$	Persistence	$Q(20)$	$Q^2(20)$	$LM(20)$
AA	0.02403*	0.95053*	0.03220*	1.482*	0.9907	26.4	19.3	18.9
AIG	0.04141*	0.91677*	0.05873*	1.417*	0.9874	35.0**	15.6	16.3
AXP	0.01958*	0.94808*	0.06949*	1.343*	1.0024	24.2	3.2	3.2
BA	0.03346*	0.93562*	0.03709*	1.317*	0.9876	15.5	21.8	21.0
CAT	0.00340	0.98055*	0.02344*	1.320*	0.9957	21.9	36.2**	16.3
CIT	0.02722*	0.95570*	0.03781*	1.405*	1.0018	21.1	17.0	16.9
DD	0.01787*	0.96790*	0.02372*	1.466*	0.9976	15.1	16.2	16.4
DIS	0.00494	0.97643*	0.03166*	1.344*	0.9972	17.5	10.7	10.4
GE	0.00816	0.96498*	0.05153*	1.598*	0.9989	17.6	21.1	21.2
GM	0.02065*	0.94330*	0.04757*	1.380*	0.9877	23.0	13.5	13.2
HD	0.01317*	0.95588*	0.05286*	1.397*	0.9955	29.8	7.7	7.9
HON	0.04347*	0.87160*	0.11698*	1.247*	0.9736	17.7	16.5	16.3
HPQ	0.01362*	0.97216*	0.01908*	1.224*	0.9953	19.6	9.0	8.9
IBM	0.02417*	0.95046*	0.04493*	1.259*	0.9971	14.2	12.1	11.8
INTC	0.02642*	0.96920*	0.00817	0.969*	0.9997	25.7	11.2	11.0
JNJ	0.03090*	0.93535*	0.06490*	1.450*	0.9999	35.5**	26.1	26.5
JPM	0.02044*	0.95543*	0.06946*	1.418*	1.0006	27.2	15.0	14.9
KO	0.02089*	0.95719*	0.04040*	1.416*	0.9983	22.8	22.6	22.7
MCD	0.01897*	0.95870*	0.02784*	1.405*	0.9916	13.9	44.6*	45.5*
MMM	0.01216*	0.98754*	-0.00219	1.186*	0.9986	21.9	17.1	16.6
MO	0.06040*	0.88601*	0.05836*	1.098*	0.9756	16.3	3.7	4.0
MRK	0.01701	0.90773*	0.06365*	1.186*	0.9566	28.8	0.9	0.9
MSFT	0.05052*	0.92676*	0.04293*	1.316*	0.9988	10.8	6.2	6.4
PFE	0.04057*	0.93469*	0.02592**	1.468*	0.9882	31.9**	11.6	11.2
PG	0.03159*	0.94220*	0.04236*	1.336*	0.9950	26.9	2.6	2.8
T	0.03919*	0.93948*	0.03402*	1.450*	0.9957	22.1	22.4	22.7
UTX	0.02540*	0.90959*	0.10784*	1.324*	0.9889	32.2**	4.4	4.4
VZ	0.02877*	0.94453*	0.04853*	1.520*	0.9976	33.6**	41.2*	37.8*
WMT	0.02549*	0.95718*	0.03206*	1.543*	0.9987	30.2	18.9	18.2
XOM	0.03407*	0.93796*	0.03420*	1.610*	0.9891	45.8*	26.1	26.4

* (**) Significant at the 1% (5%) level; $Q(20)$ is the Ljung-Box statistic for serial correlation in the residuals up to order 20; $Q^2(20)$ is the Ljung-Box statistic for serial correlation in the squared residuals up to order 20 (McLeod and Li, 1983); $LM(20)$ is the Lagrange multiplier test statistic for ARCH effects (Engle, 1982) in the residuals up to order 20.

4 Cluster analysis

Cluster analysis of time series attempts to determine groups (or clusters) of objects in a multivariate data set. The most commonly used partition clustering method is based in hierarchical classifications of the objects. In hierarchical cluster analysis, we begin with each time series being considered as a separate cluster (k clusters). In the second stage, the closest two groups are linked to form $k-1$ clusters. This process continues until the last stage in which all the time series are in the same cluster (see Everitt, Landau and Leese, 2001 for further discussion).

Figure 1 shows the cluster analysis of DJIA stock returns using a hierarchical clustering tree (or dendrogram) by complete linkage (see, e.g., Johnson and Wichern, 2002). For this purpose we used the TARCH-based distance measure defined in (5).

Figure 2 shows the multidimensional scaling map of distances constructed with the same distance measure. The multidimensional scaling is a multivariate statistical method closely related to principal coordinates analysis, and uses the information about the similarities (or dissimilarities) between the time series to construct a configuration of k points in the r-dimensional space (in this case, two dimensions). For details, see Morrison (2005). The plot can also help to identify the clusters.

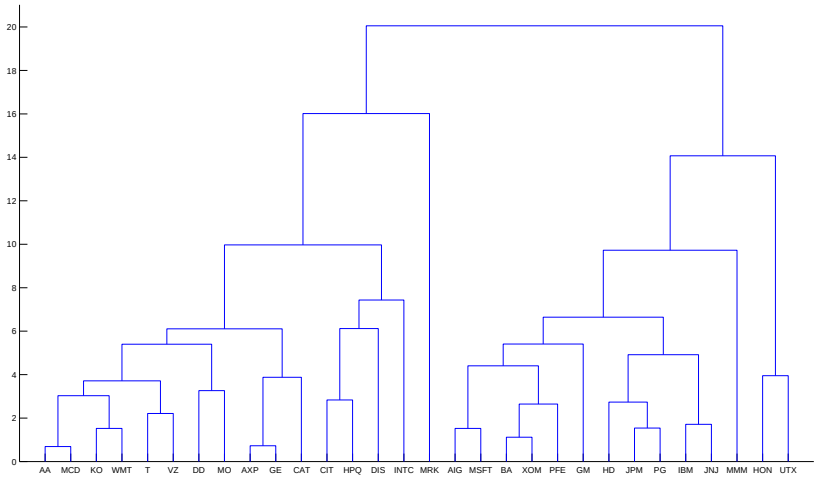


Fig. 1. Dendrogram of DJIA stock returns using the TARCH-based distance

The dendrogram associated with the stochastic features of returns series suggests two clear clusters. One is formed by Merck, consumer goods (Coca-Cola and Altria), financial (American Express and Caterpillar), technology (Hewlett-Packard, Inter-tel, Verizon and AT&T), basic materials (Alcoa and El Dupont), industrial goods (Citigroup, General Electric), and services cor-

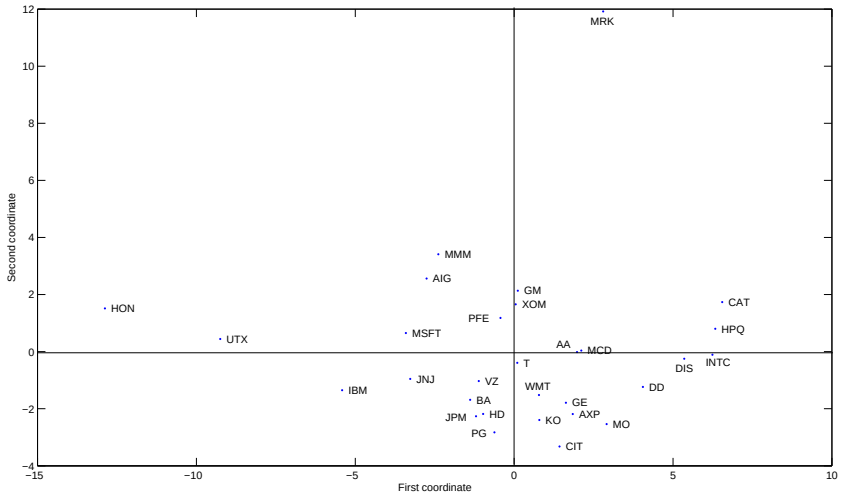


Fig. 2. Multidimensional scaling of DJIA stock returns using the TARCH-based distance

porations (Walt Disney, McDonalds and Walt-Mart Stores). The second is formed by Healthcare (Johnson & Johnson and Pfizer), conglomerates (3M and United Technologies), technology (Microsoft and IBM), financial (JP Morgan and American Int. Group), consumer goods (General Motors and Procter & Gamble), industrial goods (Boeing and Honywell), and miscellaneous sector corporations (Exxon and Home Depot).

Looking at the map of distances across the stocks, most technology companies appear close together, as most services and basic materials companies tend to cluster together, and most consumer goods companies are close to each other and close to the industrial goods companies (with exception of HON at the first coordinate). MRK company is a clear outlier.

5 Conclusions

In this paper, we introduced an asymmetric-volatility based metric for clustering financial time series. Using the information about the simple TARCH model estimates of the squared returns, we investigated the similarities among the stocks of the Dow Jones Industrial Average (DJIA) index. From this study, we found homogenous clusters of stocks with respect to the conglomerates, services and technology sectors, and we found heteregoneous clusters of stocks with respect to the financial, consumer goods and industrial goods sectors.

Acknowledgment: This research was supported by a grant from the Fundação para a Ciência e a Tecnologia (FEDER/POCI 2010).

References

BONANNO G., LILLO F., and MANTEGNA, R. (2001): "High-frequency cross-correlation in a set of stocks", *Quantitative Finance*, 1, 96-104.

BONANNO, G., CALDARELLI, G., LILLO, F., MICCIECHÉ, S., Vandewalle N. and Mantegna, R. (2004): "Networks of equities in financial markets", *European Physical Journal B*, 36, 363-371.

CAPIELLO, L., ENGLE, R. and SHEPPARD, K. (2006): "Asymmetric dynamics in the correlation of global equity and bond returns", *Journal of Financial Econometrics*, 4, 537-572.

ENGLE, R. and NG, V. (1993): "Measuring and testing the impact of news on volatility", *Journal of Finance*, 48, 1022-1082.

ENGLE, R. and KRONER, K. (1995): "Multivariate simultaneous generalized ARCH", *Econometric Theory*, 11, 122-150.

EVERITT, B., LANDAU, S. and LEESE, M. (2001): *Cluster Analysis*, 4th ed., Edward Arnold, London.

GLOSTEN, L. JAGANNATHAN, R. and RUNKLE, D. (1993): "On the relation between the expected value and the volatility of the nominal excess return on stocks", *The Journal of Finance*, 48, 1779-1801.

JOHNSON, R. and WICHERN, D. (2002): *Applied Multivariate Statistical Analysis*. 5th Ed., Prentice-Hall.

KRONER, K. and NG, V. (1998): "Modeling asymmetric comovements of asset returns", *Review of Financial Studies*, 11, 817-844.

MANTEGNA, R.N. (1999): "Hierarchical structure in financial markets", *The European Physical Journal B* 11, 193-197.

MCLEOD, A. and LI, W. (1983): "Diagnostic checking ARMA time series models using squared-residual autocorrelations", *Journal of Time Series Analysis*, 4, 269-273.

MORRISON, D. (2005): *Multivariate Statistical Methods*, 4th ed., Duxbury, Brooks/Cole Thomson Learning, Belmont.

NELSON, D. (1991): "Conditional heteroskedasticity in asset returns: a new approach", *Econometrica*, 59, 347-370.

TSAY, R. (2005), *Analysis of Financial Time Series*, 2nd ed., Wiley, New Jersey.

ZAKOIAN, J. (1994): "Threshold heteroskedasticity models", *Journal of Economic Dynamics and Control*, 18, 931-944.

ZIVOT, E. and WANG, J. (2003): *Modeling Financial Time Series with S-Plus*. Springer-Verlag, New York.

Analyzing Regime Changes in Time Series with Regression Trees

Carmela Cappelli and Francesca Di Iorio

Dipartimento di Scienze Statistiche, Università di Napoli Federico II
Via L. Rodinò n.22, 80138 Napoli, Italy,
{*carmela.cappelli, fdiorio*}*@unina.it*

Abstract. The detection of regime changes in time series is nowadays a popular subject of research both in the econometric and statistical literature. The most challenging task is to identify multiple breaks occurring at unknown date, in this context most contributions have addressed the case of level shifts. In particular, Cappelli and Reale (2005) have proposed a method that employs regression trees to detect multiple mean breaks. In this paper we propose an extension of their procedure meant to deal with regime changes due to instability in model parameters. In order to evaluate the performance of the proposed approach a simulation study is carried on. An application to the US labor productivity is also presented and discussed.

Keywords: time series, regime changes, regression trees, Chow Test

1 Introduction

In the last two decades a large amount of research both in the econometric and statistics literature has been concerned with the detection of regime changes in time series (for a review see Hansen, 2001).
Indeed, it's not uncommon that economic and financial phenomenon show instability over time and this instability needs to be investigated for several purposes: first, in the context of forecasting it is sensible to base the forecasts on a model estimated on a recent segment of the series, instead of using the entire series, and this is especially true for time series covering extended periods for which the specification of a single model can be inadequate. Second, the identification of regime changes might isolate short intervals between longer ones revealing the presence of outliers and thus suggesting the need for adjusting the data. Third, the presence of structural breaks reveals a behavior of the time series that could otherwise be misunderstood and modelled inadequately. In particular such a presence may lead to an erroneous identification of a long memory process (Granger and Hyung, 2004).
The most challenging task is to identify multiple breaks occurring at unknown dates. Most contributions have addressed the case of level shifts, in this context Cappelli *et* al (2005, 2008) have proposed a computational efficient procedure that employs regression trees to identify breaks and their

locations.
In this streamline this paper focuses on a different problem: regime changes due to instability in model parameters. To investigate these types of regime changes, we propose an extension of the procedure that uses in the tree growing stage the residuals of parametric models fitted to contiguous subseries obtained by splitting the original series.
The remainder of the paper is organized as follows. In section 2 we remind the basics of regime changes analysis and we introduce regression trees showing how they can be employed for regime changes detection. Section 3 reports the results of a simulation study meant to evaluate the performance of the proposed approach. Final remarks follows in section 4.

2 Regression trees for regime changes detection

Let y_t be a time series characterized by G regimes and $G-1$ breaks so that $t = T_{g-1}+1, \ldots, T_g$ and $g = 1, \ldots, G$ (we adopt the common convention that $T_0 = 0$ and $T_G = T$ where T is the length of the series).
Regardless of the nature of the breaks, the problem resorts to estimate the set of break dates $\{1, \ldots, g, \ldots, G-1\}$ that define a partition of the series $P(T, G) = \{I_1, \ldots, I_g, \ldots, I_G\}$ into homogeneous segments.
In case of mean breaks Cappelli and Reale (2005) have proposed a procedure that employs Regression Trees (so forth denoted RT). Briefly, in a standard RT procedure a node h is split into its left and right descendants h_l and h_r to reduce the deviance of the response variable y_t fitting to each node the mean of the y's values falling into the node (for details see Breiman *et* al., 1984). Thus, the algorithm selects the split such that

$$SSR(h) - [SSR(h_l) + SSR(h_r)] = max \tag{1}$$

where $SSR(h) = \sum_{y_t \in h}(y_t - \mu(\hat{h}))^2$, is the residual sum of squares for node h, $\hat{\mu}(h)$ is the mean of the y_t values in node h and $SSR(h_l)$ and $SSR(h_r)$ are the corresponding quantity computed for the left and right descendants, respectively.
For the purpose of locating the breaks-dates, let k be an arbitrary ascending (or descending) sequence of completely ordered numbers, for sake of simplicity take $k = 1, 2, \ldots, i, \ldots, T$.
Tree regressing y_t on k yields to a partition of the series into homogeneous subperiods such that $\hat{\mu}_g \neq \hat{\mu}_{g+1}$. The split points in the tree identify the break-dates whereas the different regimes are given by the y_t values falling into the tree terminal nodes.
As shown in Cappelli *et.* al (2008) this approach, that mimics the well known break estimation method of Bai and Perron (2003), is computationally fast and thus it is suitable for workers who routinely analyze large numbers of time series or those dealing with long series which are impractical to analyze with current methods due to the prohibitive computing time.

Regime changes can be due to model instability i.e. to breaks in the parameters of a given model. To deal with these types of changes we propose a modification of the above described regression tree approach.
To focus our discussion let us consider the following $AR(p)$ model:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t,\ \epsilon_t \sim WN(0, \sigma_\epsilon^2).$$

The parameters are $(\phi_i, \sigma_\epsilon^2)$. Here, we consider only changes in the autoregressive parameters ϕ_i that reflect changes in the structure of the serial correlation of y_t. At the aim to identify the breaks we estimate for each value of k with $min_{obs} \leq k \leq T - min_{obs}$ [1] two separate $AR(p)$ models computing the corresponding residuals. The best split is selected according to the expression 1 employing in the computations the residuals of the $AR(p)$ model, thus, in this case for a given node h it is $SSR(h) = \sum_{y_t \in h}(y_t - \hat{y}_t)^2$ where $\hat{y}_t = \sum \hat{\phi}_i y_{t-i}$.
As in any tree procedure a criterion to select the relevant splits, i.e. the sets of final break dates and corresponding regimes, is needed. Usually retrospective pruning is employed, in the present case it seems more appropriate the use of an hypothesis testing based stopping rule. In particular we consider the Chow test (1960) that represents a milestone in the field. This test checks the presence of a single structural break at a known date by splitting the series into two subperiods and testing the equality of the two sets of parameters using a classic F statistic. In the case of our tree procedure, upon the choice of a significance level, splitting stops at a given node h if

$$\frac{SSR(h) - [SSR(h_l) + SSR(h_r)]/p}{[SSR(h_l) + SSR(h_r)]/(T(h) - 2p)} < F_{(p,T(h)-2p)}$$

where $T(h)$ is the length of the subseries in node h and p is the number of estimated parameters.
Eventually, since our approach requires the *a priori* specification of a model, to circumvent the problem of model mispecification, in our programme various models can be assumed growing candidate trees (i.e., sets of breaks). For the purpose of choosing among the competing models the preferable one, model selection criteria can be employed (for the computation of these criteria in regression trees see Su *et* al., 2004).

3 Simulation study

In this section we present the results of a simulation study carried out to evaluate the ability of the proposed approach in detecting regime changes

[1] min_{obs} denotes a minimum number of observations needed to estimate the model; conditions such as this and/or on the reduction in the residuals are easily handled within the framework of tree procedures.

due to instability in the model parameters. At this aim we considered two simulation settings for each we generated 1000 series of length 450 attaching subseries drawn from an $AR(2)$ and a $MA(2)$, respectively. The parameters values are reported in Table 1.

Table 1. Parameter values for the two simulation settings.

	Setting 1		Setting 2	
	$AR(2)$		$MA(2)$	
	ϕ_1	ϕ_2	θ_1	θ_2
Y_{t_1}	0.6	-0.6	0.5	0.3
Y_{t_2}	1.2	-0.6	-1.0	-0.6
Y_{t_3}	-0.4	-0.2	1.6	-0.2

where $t_1 = 1, \ldots, 150$; $t_2 = 151, \ldots, 300$ $t_3 = 301, \ldots, 450$, thus in both cases there are two breaks at obs. 150 and 300 respectively. For illustrative purposes, we have depicted one of the $AR(2)$ simulated series and the corresponding autocorrelation functions in Figure 1.

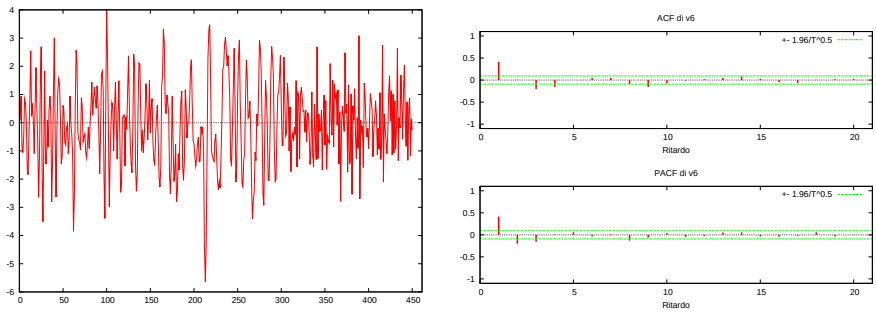


Fig. 1. An illustrative simulated series with the corresponding autocorrelation and partial autocorrelation functions.

As it can be seen there's no graphical evidence of the presence of breaks. In particular the second regime that is characterized by a more marked sinusoidal behavior it's not distinguishable from the other regimes in particular from the first one.
On the simulated series we have searched for breaks considering as reference models for the $AR(2)$ an $AR(p)$ ($p = \{1, 2\}$) and for the $MA(2)$ a $MA(q)$ ($q = \{1, 2\}$); we fixed a significance level of 0.01 in the Chow test and set $min_{obs} = 30$. Table 1 reports the estimated mean number of breaks nb, the

percentages of correct identifications $\%ci$ (number of breaks correctly identified within a short interval from the real date) and the percentage of correct selections in the sense of AIC ($\%cs$) associated with the two reference models for the two simulation settings.

Table 2. Main simulation findings averaged over 1000 replications.

	nb	$\%ci$ (± 8 obs.)	$\%ci$ (± 4 obs.)	$\%ci$ (± 2 obs.)	$\%cs$
$AR(1)$	2.3	80	70	60	0
$AR(2)$	2.6	100	95	85	100
$MA(1)$	3.3	70	45	30	20
$MA(2)$	2.3	85	70	50	80

As we can see the results are quite different for the AR and the MA models. In the case of the AR models the percentage of correct identifications are very high even in case of incorrect model of lower order and the mean number of breaks very close (slightly higher) to the actual number; also, the preferable model according to the AIC is always the correct one. On the contrary, for the MA models the correct identifications are lower dropping to unsatisfactory levels when the $MA(1)$ is used as reference model. Also, in this latter case the estimated mean number of breaks is higher than the actual number and in the 20% of cases the preferable model would be the incorrect one. Indeed, these two findings are related because the AIC is computed on the tree residuals, and a tree overfitted i.e., with a higher number of splits (breaks) and terminal nodes (regimes) tends to produce smaller residuals and consequently a better AIC. It is also worth noticing that, in general, a final number of breaks higher than the actual one, does not surprise because, this number depends on the stopping rule and it's well known (see Hansen, 2001) that when the break date is estimated with the data at hand, the Chow test might falsely indicate a break. Thus, alternative stopping rules will be considered in the future, in particular it's our belief that the testing procedure based on the AR metric (Piccolo, 1989) can be useful in this context.

4 Regime changes analysis of labor productivity in US

Based on the encouraging results of the simulation study we have employed our approach for the regime changes analysis of the labor productivity in the US manufacturing sector. In particular we have considered the seasonally adjusted quarterly index series that measures output per hour of all persons for the manufacturing sector (the series is freely available at *www.bls.gov*). The sample period is 1987:1-2007:3 yielding to 83 observations. Since the issue

of primary interest is the growth rate of labor productivity, we have considered the log-first differences $\nabla \log y_t$. The original series and the log-first differences are displayed in Figure 2 whereas Figure 3 depicts the estimated autocorrelation functions of the log-first differences.
The graphical inspection of Figure 3 suggests that the series may be well represented by an $AR(1)$ model. We argue that this series might present regime changes and that they might be due to both changes in the constant term (that accounts for the mean through the relation $\mu = \frac{\phi_0}{1-\phi_1}$) and in the AR parameter. Thus, we have considered a model with the intercept:

$$\nabla \log y_t = \phi_0 + \phi_1 \nabla \log y_{t-1} + \epsilon_t$$

.

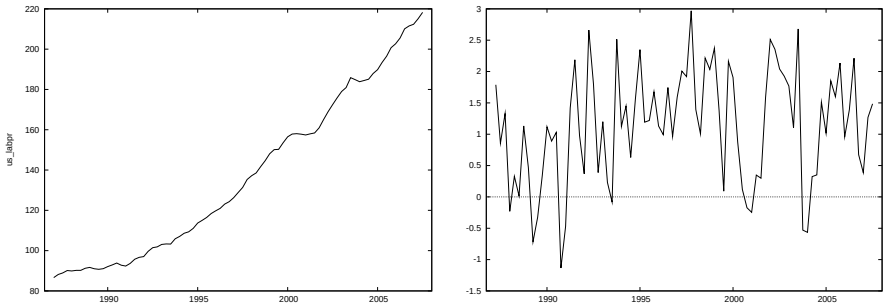


Fig. 2. The index series of output per hour of all persons in the manufacturing sector and the corresponding log-first differences.

We have applied the tree based procedure described in section 2 setting a significance level for the Chow test of 5% and $min_{obs} = 20$) and we found evidence of two regime changes in 1993:3, that corresponds to a well documented resurgence in productivity (see for example Hansen, 2001), and in 2000:1. These findings are in accord with the visual inspection of the series plotted in Figure 2.
Table 3 reports the estimated models for the entire series and for the regimes identified by the breaks.

We see that the AR coefficient that is significant for the entire series, remains significant only on the third regime. i.e., the subperiod running form 2000:2 On the first and second regime, only the constant is significant. Indeed, the constant term, always significant, varies considerably across the regimes. Since the series analyzed is the labor productivity growth rate, changes in the constant represent changing trend. In particular, the reduction of the constant in the third regime shows a change in the slope of the underlying original series, i.e., a reduction in the growth rate of the labor productivity in recent years

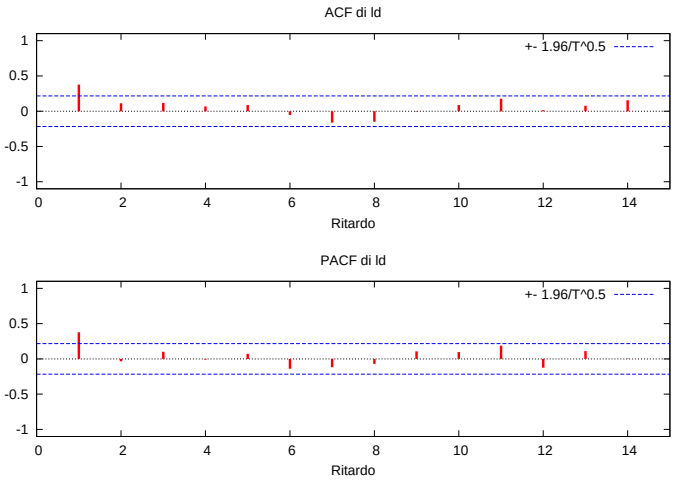


Fig. 3. The autocorrelation and partial autocorrelation functions of the log-first differences for the labor productivity index.

Table 3. Estimates of the $AR(1)$ model for the entire series and for the subseries identified by the breaks. Standard errors are given in brackets, crosses indicate a significance.

	n.obs	ϕ_0	ϕ_1
		Entire series	
1987:2-2008:3	82	$0.70^{\dagger}(0.15)$	$0.38^{\dagger}(0.10)$
		Regimes	
1987:1-1993:3	26	$0.45^{\dagger}(0.25)$	$0.26(0.26)$
1993:4-2000:1	26	$1.59^{\dagger}(0.47)$	$-0.02(0.29)$
2000:2-2007:3	30	$0.59^{\dagger}(0.28)$	$0.47^{\dagger}(0.28)$

It's worth noticing that, with such a short time series both model identification and parameter estimation might not be reliable, thus, the results must be interpreted with caution. Nevertheless, the case study has shown that our procedure is able to reveal relevant features of the series and thus it represents a useful aid to data understanding and model specification.

5 Concluding remarks

In this paper we have proposed a tree based procedure designed to identify regime changes due to instability in model parameters. The results of

the simulation study are encouraging: the proposed approach seems rather promising in detecting model parameter changes over time and, as shown by the application to the US labor productivity index series, it provides useful insights into the data.
Further simulations are the subject of ongoing research, in particular the unsatisfactory results associated with the MA models require further investigations. Future work will also address the use of alternative stopping rules for the selection of the relevant breaks and the issue of detecting breaks in the differencing parameter in long memory processes.

Acknowledgements

Authors acknowledge financial support from Dipartimento di Scienze Statistiche University of Naples Federico II. Although this is a joint work, C. Cappelli wrote sections 2 and 4, and F. Di Iorio sections 1 and 3.

References

BAI, J. and PERRON, P. (2003): Computation and analysis of multiple structural change models, *J*ournal of Applied Econometrics, 18, 1-22.

BREIMAN, L., FREIDMAN, J.H., OLSHEN R.A. and STONE C.J. (1984): *C*lassification and Regression Trees, Wadsworth & Brooks, Monterey (CA).

CAPPELLI, C., PENNY, R.N., REA, W. and REALE, M. (2008): Detecting multiple mean breaks at unknown points with regression trees, *Mathematics and Computers in Simulation*, forthcoming.

CAPPELLI, C. and REALE, M. (2005): Dating multiple structural breaks occurring at unknown dates via Atheoretical Regression Trees, in: C. Provasi (Ed): *S.Co. 2005: Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*, Provasi C. (Ed), Cleup, Padova, 479-484.

CHOW, G.C. (1960): Tests of euquality between sets of coefficients in two linear regressions, *Econometrica 28(3), 591-605* .

GRANGER, C.W.J. and HYUNG, N. (2004): Occasional structural breaks and long memory with an application to the *S&P 500* absolute stock returns, *Journal of Empirical Finance, 11, 399-421.*

HANSEN, B. (2001): The new econometrics of structural change: dating breaks in U.S. labor productivity, *Journal of Economic Perspectives, 15, 117-128.*

PICCOLO, D. (1990): A distance measure for classifying ARIMA models, *Journal of Time Series, 11, 153-164.*

SU, X.G., WANG M. and FAN, J.J. (2004): Maximum Likelihood Regression Trees. *Journal of Computational and Graphical Statistics, 13, 586-598.*

Bootstrap and Exponential Smoothing Working Together in Forecasting Time Series

Clara Cordeiro[1] and M. Manuela Neves[2]

[1] Faculdade de Ciências e Tecnologia, Universidade do Algarve
Campus de Gambelas, 8005-139 Faro, Portugal, *ccordei@ualg.pt*
[2] Departamento de Matemática, Technical University of Lisbon (TULisbon)
Instituto Superior de Agronomia, Tapada da Ajuda, 1349-017 Lisboa, Portugal,
manela@isa.utl.pt

Abstract. In these article we propose an automatic procedure to forecast time series. First, the procedure selects the best exponential smoothing method among a set of methods proposed, by comparing the mean squared errors (MSE). Second, a resampling scheme is used over the residuals obtained after an autoregressive adjustment selected by AIC criterium. The time series is then reconstructed and forecasts are obtained with the selected model. The procedure developed is compared with the best prediction methods selected using the 3003 time series of M3 competition.

All these computational work is performed with the software R2.6.2 (R Development Core Team (2008)).

Keywords: accuracy measures, bootstrap, exponential smoothing, forecasting, M3 competition

1 Introduction

In our days it is well known the importance of time series studies. These studies provide indicators about a country economy, the unemployment rate, the export and import product rates, etc. In this article an automatic procedure to analyze time series and to obtain forecast estimates is proposed. In order to evaluate its performance the 3003 time series of the M3 competition are used. The Makridakis and Hibon (2000) article lists 24 methods of forecasting used in the M3 competition. Several accuracy measures were used to analyze the performance of the various methods. In this study we compare in-sample performance of the methods using MSE (based on the one-step-ahead forecasts) and the out-of-sample performance based on the forecasts for a given period. Further, results of our method are compared with the best six methods in the M3 competition.

2 Exponential smoothing methods

Exponential smoothing (EXPOS) refers to a set of forecasting methods, several of which are widely used. The EXPOS is a procedure that continually

updates a forecast emphasizing the most recent experience, that is, recent observations are given more weight than the older observations. Single exponential smoothing, Holt's linear trend, HoltWinters seasonal smoothing with either additive or multiplicative seasonality are some examples of EXPOS methods, see DeLurgio(1998) for more details. The forecasting performance of exponential smoothing methods has been addressed by several authors. A very good reviewing of the past 25 years of time series forecasting is given by De Gooijer and Hyndman (2006). These methods are relatively simple but robust approaches to forecasting and accurate in model identification. The Box-Jenkins ARIMA models require the user to identify an appropriate model and to use at least 50 observations to have a good chance of success (Chatfield, 1978).

The four EXPOS methods addressed here, where data with or without trend and/or with or without seasonal components are considered, are in Table 1.

Table 1. The EXPOS methods considered

Classification	Method
1	Single exponential smoothing
2	Holt's linear trend
3	HoltWinters seasonal smoothing with additive seasonality
4	HoltWinters seasonal smoothing with multiplicative seasonality

Using R software (R Development Core Team (2007)) and some of their packages, such as **forecast** and **tseries**, some functions are constructed. For the EXPOS methods the R function **HoltWinters**() is used in the model selection.

The *additive* Holt-Winters has the following recursive equations to estimate the trend and the seasonal factor at time t

$$T_t = \alpha(X_t - S_{t-s}) + (1-\alpha)(T_{t-1} + b_{t-1})$$
$$b_t = \beta(T_t - T_{t-1}) + (1-\beta)b_{t-1}$$
$$S_t = \gamma(X_t - T_t) + (1-\gamma)S_{t-s}$$

with α, β, $\gamma \in [0,1]$ and

T_t smoothed value at end of period t after adjusting for seasonality
X_t value of actual demand at end of period t
S_{t-s} smoothed seasonal index, s periods ago
b_t smoothed value of trend through period t
α smoothing constant used for T_t
β smoothing constant used to calculate the trend (b_t)
γ smoothing constant used for calculate the seasonal index in period t.

The forecast function used is **predict**() with the equation

$$\hat{X}_t(h) = T_t + h \times b_t + S_{t+h-rs} \tag{1}$$

where $h = 1, 2, 3, \cdots$ is the forecast horizon and $r = 1$ if $1 \leq h \leq s$, $r = 2$ if $s < h \leq 2s$, etc.

The *multiplicative* Holt-Winters has as recursive equations

$$T_t = \alpha(X_t/S_{t-s}) + (1-\alpha)(T_{t-1} + b_{t-1})$$
$$b_t = \beta(T_t - T_{t-1}) + (1-\beta)b_{t-1}$$
$$S_t = \gamma(X_t/T_t) + (1-\gamma)S_{t-s}$$

and prediction equation

$$\hat{X}_t(h) = (T_t + h \times b_t) \times S_{t+h-rs}, \tag{2}$$

where the parameters are defined above.

A new function **best.EXPOS**() is defined for choosing the model that better fits a time series according to Table 1. It is extremely important to choose the model that better describes the behavior of the series in study because it can be the reason for a *good* or a *bad* forecast. Another function **boot.EXPOS**() described in the next section is derived to obtain future values.

3 Bootstrapping EXPOS methods

Efron's bootstrap classical procedure, Efron 1979, does not contemplate structures of dependent data, such as time series, where the dependence data arrangement should be kept during the resampling scheme. A great development in the area of resampling methods for dependent data has been observed (Lahiri (2003)). Several authors have proposed bootstrap methodologies for time series. Bootstrap estimates can incorporate the variability due to parameter estimation without assuming any specific distribution for the errors. The majority of those methods suggests the use of blocks, in order to keep the dependence structure. In Cordeiro and Neves (2006) several bootstrap methodologies for dependent data were compared in constructing forecast intervals and the sieve bootstrap (Bühlmann (1997)) showed the best results. Inspired by this approach and by Alonso et al. (2002) a different bootstrap approach is implemented to obtain forecast estimates: **boot.EXPOS**(). This approach was better explained in Cordeiro and Neves (2007a, b). The overall procedure can be shortly described as:

(i) Use **best.EXPOS**() to choose the method, where the components (if they exist) are removed and the residuals obtained.
(ii) Use AIC for selecting the autoregressive model that better fits the residuals.

(iii) For B replicates do
   (a) Resample the residuals centered and use the model in *step 2* to simulate a new residuals series.
   (b) Obtain a bootstraped time series by adding the components (if they exist) estimated in *step 1*.
   (c) Forecast future values according to the EXPOS model in *step 1*.

(iv) Determine for each h step-ahead an optimal point forecast. Here the mean or the median of the B forecasts will be considered.

4 Forecast accuracy measures

To evaluate the performance of **boot.EXPOS**() procedure some accuracy measures are used. In M3 competition some of them are calculated. Concerning this subject it is also interesting to read the article proposed by Hyndman and Koehler (2006).

Let X_t denote the observation at time t and $\hat{X}_t$ the forecast of X_t. The forecast error is defined by $e_t = X_t - \hat{X}_t$. The forecasts are computed for a hold-out period. Thus the out-of-sample forecasts $\hat{X}_n(1), \cdots, \hat{X}_n(h)$ are computed based on the data from time $t = 1, \cdots, n$. Accuracy measures are then computed in order to compare our results with those presented in Makridakis and Hibon (2000) in http://www.forecastingprinciples.com/m3-competition.html. The following accuracy measures are here considered:

Table 2. Accuracy measures

Acronyms	Definition	Formula
sMAPE	Symmetric Mean Absolute Percentage Error	$mean(200\frac{\lvert e_t \rvert}{X_t+\hat{X}_t})$
RMSE	Root Mean Squared Error	$\sqrt{(mean(e_t^2)}$

5 Application to M3 competition data

As said before, the data sample M3 competition (R package:Mcomp) were used to obtain point forecasts based on our procedure. The 3003 series include various types of time series data: micro, demographic, finance, etc, and different categories according to the time intervals between successive observations: yearly, quarterly, monthly and "other", see Figure 1 for example. All time series considered are strictly positive. For each category, different forecasting periods are considered: six for yearly, eight for quarterly and "other", eighteen for monthly. Each M3 competition series was classified according to Table 1, using function **best.EXPOS**(). This function select the model

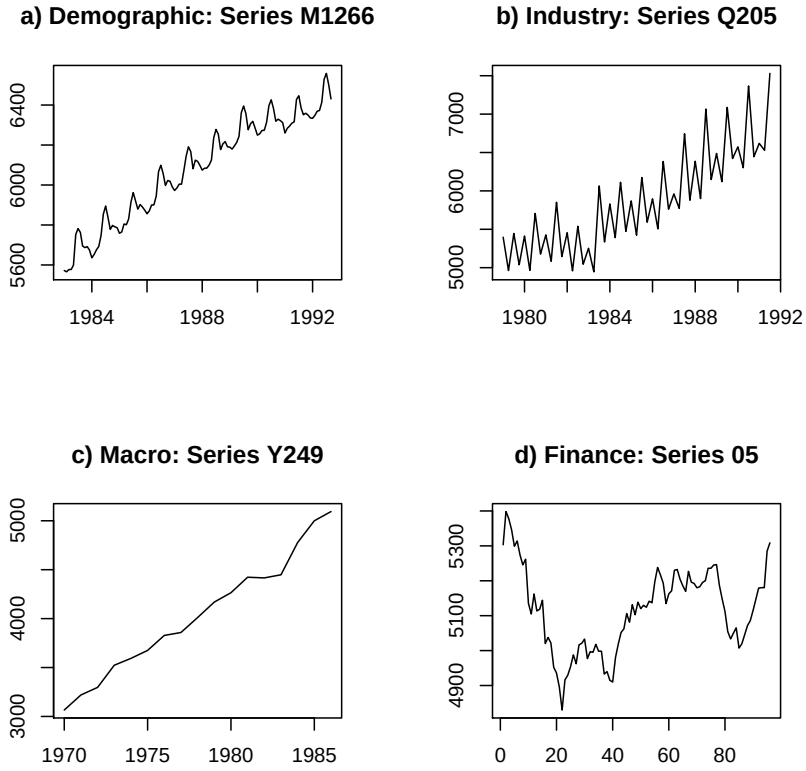


Fig. 1. M3 competition category: monthly (a), quarterly (b), yearly (c) and "other" (d).

Table 3. Number of series in each category and method

	Method				
Categories	1	2	3	4	*TOTAL*
Monthly	305	274	367	482	*1428*
Quarterly	119	204	227	206	*756*
Yearly	216	429			*645*
Other	54	120			*174*
TOTAL	*694*	*1027*	*594*	*688*	3003

with the potential to produce the best forecasts. Table 3 shows the number of series in each category and method.

Then the function **boot.EXPOS**() is used. One thousand replications (B=1000) are carried out and forecasts are obtained for each replication. Two forecast estimates are obtained using the mean (method1) and median

(method2) of the B forecasted values. The mean calculation is based in the 95% central forecast simulations. The methods Naive2, Box-Jenkins automatic, ForecastPro, THETA, RBH and ForecastX have the best performance in M competition (Makridakis et al. (1982)) and M3 competition and they are described in Makridakis and Hibon (2000). These well behaved methods are used in our accuracy comparisons. Results for sMAPE using the 3003 time series are given in Table 4 for method1 and method2.

Table 4. Average symmetric MAPE: all data

Method	Forecasting horizon										Average of forecasting horizon					
	1	2	3	4	5	6	8	12	15	18	1-4	1-6	1-8	1-12	1-15	1-18
Naive2	10.5	11.3	13.6	15.1	15.1	15.8	14.5	16.0	19.3	20.7	12.62	13.55	13.74	14.22	14.8	15.46
B-J automatic	9.2	10.4	12.2	13.9	14.0	14.6	13.0	14.1	17.8	19.3	11.42	12.39	12.52	12.78	13.33	13.99
ForecastPRO	8.6	9.6	11.4	12.9	13.3	14.2	12.6	13.2	16.4	18.3	10.64	11.67	11.84	12.12	12.58	13.18
THETA	8.4	9.6	11.3	12.5	13.2	13.9	12.0	13.2	16.2	18.2	10.44	11.47	11.61	11.94	12.41	13.00
RBF	9.9	10.5	12.4	13.4	13.2	14.1	12.8	14.1	17.3	17.8	11.56	12.26	12.4	12.76	13.24	13.74
ForecastX	8.7	9.8	11.6	13.1	13.2	13.8	12.6	13.9	17.8	18.7	10.82	11.72	11.88	12.21	12.80	13.48
Our method1	8.0	9.8	11.5	12.8	14.2	15.1	12.8	13.8	17.0	18.8	10.49	11.87	12.03	12.63	13.35	14.19
Our method2	7.9	9.8	11.3	12.8	14.2	15.1	12.8	13.8	16.9	18.8	10.43	11.82	11.98	12.60	13.32	14.15

The performance of the various methods depends upon the length of the forecasting horizon. Our method is among the selected methods in what refers to sMAPE, in particular it has a very good performance for one step-ahead. This is an interesting fact that gave us new ideas for investigation.

When, only the data in a category is considered for the study, as it is shown in Table 5, our method revealed a good overall performance for short periods. In Table 6 RMSE is calculated for the monthly category and compared with the other methods. Here we have the same conclusions presented before.

Table 5. Average symmetric MAPE: 1428 monthly series

Method	Forecasting horizon										Average of forecasting horizon					
	1	2	3	4	5	6	8	12	15	18	1-4	1-6	1-8	1-12	1-15	1-18
Naive2	15.0	13.5	15.7	17.0	14.9	14.4	15.6	16.0	19.3	20.7	15.30	15.08	15.26	15.55	16.16	16.89
B-J automatic	12.3	11.7	12.8	14.3	12.7	12.3	13.0	14.1	17.8	19.3	12.78	12.70	12.86	13.19	13.95	14.80
ForecastPRO	11.5	10.7	11.7	12.9	11.8	12.0	12.6	13.2	16.4	18.3	11.72	11.78	12.02	12.43	13.07	13.85
THETA	11.2	10.7	11.8	12.4	12.2	12.2	12.7	13.2	16.2	18.2	11.54	11.75	12.09	12.48	13.09	13.83
RBF	13.7	12.3	13.7	14.3	12.3	12.5	13.5	14.1	17.3	17.8	13.49	13.14	13.36	13.64	14.19	14.76
ForecastX	11.6	11.2	12.6	14.0	12.4	12.0	12.8	13.9	17.8	18.7	12.32	12.28	12.44	12.81	13.58	14.44
Our method1	11.4	11.8	12.1	13.1	13.4	13.4	13.6	13.8	17.0	18.8	12.10	12.53	12.95	13.25	13.84	14.61
Our method2	11.4	11.8	12.1	13.1	13.4	13.4	13.6	13.8	16.9	18.8	12.10	13.53	12.94	13.24	13.83	14.58

As in Makridakis and Hibon (2000) the relative performance of the methods compared to a benchmark: Naive2 and Dampen is also observed. Description about the Naive2 benchmark is in Appendix A in that article and description about Dampen is in Gardner and McKenzie (1985). Table 7 and Table 8 list the difference sMAPE(benchmark)-sMAPE(select method) using

Table 6. Root mean squared error: 1428 monthly series

Method	Forecasting horizon										Average
	1	2	3	4	5	6	8	12	15	18	1-18
Naive2	1144	1367	1466	1643	1363	1201	1453	1329	1766	1673	1448
B-J automatic	864	942	934	1061	1006	1100	1107	1208	1454	1563	1185
ForecastPRO	812	905	913	1068	1032	990	1157	1135	1411	1463	1146
THETA	810	936	1067	1181	1130	979	1170	1138	1445	1487	1168
RBF	984	1636	1468	1850	1503	1000	1355	1197	1764	1651	1459
ForecastX	794	977	920	1087	1008	966	1175	1169	1457	1510	1163
Our method1	837	952	928	1148	1131	1071	1202	1157	1418	1500	1175
Our method2	838	956	929	1137	1127	1058	1183	1157	1414	1483	1168

the results in Table 4. It is clear that all the methods are more accurate than Naive2. The comparisons in Table 8 are similar, but in this case the negative values occur whenever the method listed is worse than Dampen method.

Table 7. Comparison (in %) of various methods with Naïve2 as the benchmark

	Forecasting horizon(s)				
	1	Average	Average	Average	Average
		1-4	1-6	1-12	1-18
B-J automatic	1.3%	1.2%	1.2%	1.4%	1.5%
ForecastPRO	1.9%	2.0%	1.9%	2.1%	2.3%
THETA	2.1%	2.2%	2.1%	2.3%	2.5%
RBF	0.6%	1.1%	1.3%	1.5%	1.7%
ForecastX	1.8%	1.8%	1.8%	2.0%	2.0%
Our method1	2.5%	2.1%	1.7%	1.6%	1.3%
Our method2	2.6%	2.2%	1.7%	1.6%	1.7%

Table 8. Comparison (in %) of various methods with Dampen as the benchmark

	Forecasting horizon(s)				
	1	Average 1-4	Average 1-6	Average 1-12	Average 1-18
Naive2	-1.7%	-1.6%	-1.5%	-1.8%	-1.8%
B-J automatic	-0.4%	-0.4%	-0.4%	-0.4%	-0.4%
ForecastPRO	0.2%	0.4%	0.4%	0.3%	0.4%
THETA	0.4%	0.6%	0.5%	0.5%	0.6%
RBF	-1.1%	-0.5%	-0.2%	-0.3%	-0.1%
ForecastX	0.1%	0.2%	0.3%	0.2%	0.1%
Our method1	0.8%	0.6%	0.2%	-0.2%	-0.6%
Our method2	0.9%	0.6%	0.2%	-0.2%	-0.5%

6 Remarks

In this work an automatic procedure based on EXPOS and bootstrap in order to obtain forecasts is presented. Our methodology is applied to the 3003 time series of the M3 competition. Some accuracy measures such as sMAPE, RMSE and comparison with a benchmark method (Naive2 and Dampen) are evaluated as in Makridakis and Hibon (2000). In a first analyze our method presents a good performance- it is an automatic procedure that stands among the best methods in M3 competition data.

Some ideas are in progress for improving the results obtained, as well for considering more methods for the initial adjustment.

References

ALONSO, A.M., PEÑA, D. and ROMO, J. (2002): Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference 100, 1-11.*

BÜHLMANN, P. (1997): Sieve Bootstrap for Time series. *Bernoulli 3, 123-148.*

CHATFIELD, C. (1978): The Holt-Winters forecasting procedure. *Applied Statistics 27, 3, 264-279.*

CORDEIRO, C. and NEVES, M.M. (2006): The Bootstrap methodology in time series forecasting. In: Proceedings of CompStat2006, 17th Conference of IASC-ERS, Springer Verlag, 1067–1073.

CORDEIRO, C. and NEVES, M.M. (2007a): Bootstrap prediction intervals: a case-study. In: Joan del Castillo, Anna Espinal and Pere Puig (Eds.): Proceedings of the 22nd International Workshop on Statistical Modelling (IWSM2007), pag. 191-194.

CORDEIRO, C. and NEVES, M M. (2007b): Resampling techniques in time series prediction: a look at accuracy measures. In: Gomes, M.I., Pestana, D. and

Silva, P.(eds): Proceedings of the 56th Session of the International Statistical Institute(ISI 2007), pag. 353. *Extended Abstrat* in CD-ROM.

DE GOOIJER, J.G. and HYNDMAN, R.J. (2006): 25 years of time series forecasting. *International Journal of Forecasting 22, 443-473*.

DELURGIO, S.A. (1998): *Forescasting Principles and Applications*. McGraw-Hill International Editions.

EFRON, B. (1979): Bootstrap methods: another look at the Jackknife. *The Annals of Statistics 7, 1-26*.

GARDNER, E.S. and MCKENZIE, E. (1985): Forecasting trends in time series. *Management Science 31, 1237-1246*.

HYNDMAN, R.J. and KOEHLER,A.B. (2006): Another look at measures of forecast accuracy. *International Journal of Forecasting 22, 679-688*.

LAHIRI, S.N. (2003): *Resampling Methods for Dependente Data*. Springer Verlag Inc.

MAKRIDAKIS, S., ANDERSON, A., CARBONE, R., FILDES, R., HIBON, M., LEWANDOWSKI, R., NEWTON, J., PARZEN, E. and WINKLER,R. (1982): The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition. *Journal of Forecasting 1, 111-153*.

MAKRIDAKIS, S. and HIBON, M. (2000): The M3-Competition: results, conclusions and implications. *International Journal of Forecasting 16, 451-476*.

R DEVELOPMENT CORE TEAM (2008): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

A Note on the Estimation of Long-Run Relationships in Dependent Cointegrated Panels*

Francesca Di Iorio[1] and Stefano Fachin[2]

[1] Department of Statistical Sciences, University of Naples Federico II
v. L. Rodinó, 80138 Naples, Italy, *fdiiorio@unina.it*
[2] Faculty of Statistics, University of Rome "La Sapienza"
p.le A. Moro 5, 00138 Rome, Italy, *s.fachin@caspur.it*

Abstract. We address the issue of estimation and inference in dependent non-stationary panels of small cross-section dimensions. The main conclusion is that the best results are obtained applying bootstrap inference to single-equation estimators. SUR estimators perform badly, or are even unfeasible, when the time dimension is not very large compared to the cross-section dimension.

Keywords: panel cointegration, FM-OLS, FM-SUR

1 Introduction

The estimation of cointegrating relationships in heterogenous, dependent panels can be considered a still largely unsettled problem. In presence of cross-units linkages system estimation methods should be more efficient than single equation ones; however, in practice there are many difficulties. Full information methods (Groen and Kleibergen, 2003) are feasible only when the number of time observations (T) is much larger than that of cross-section observations (N). Very much the same holds for Mark, Ogaki and Sul's (2005) DSUR, the extension of DOLS to seemingly unrelated systems. Further, since the estimation requires inverting the long-run covariance matrix of the system, only short-run dependence is allowed; cointegration across units is ruled out. This assumption is shared by the analogous system extension of FM-OLS, FM-SUR (Moon, 1999), which is however somehow more parsimonious than DSUR. Summing up, system estimation of cointegrated panels is feasible only in particular conditions: large datasets or no cointegration across units. This prompts two main questions. First, how large are the efficiency gains actually delivered by SUR estimators with respect to single-equation ones with empirically relevant sample sizes? The second question requires

* Financial support from the Department of Statistics of the University of Naples Federico II, University of Rome "La Sapienza" and MIUR is gratefully acknowledged. Correspondence to: s.fachin@caspur.it, fdiiorio@unina.it

taking a completely different perspective. Efficiency improvements are desired in order to have more accurate interval estimation and more reliable tests. Can we reach these targets applying some alternative inference procedure, such has the bootstrap, to standard single-equation estimators? The good simulation results on bootstrap inference in FM-OLS (Psaradakis, 2001, Fachin, 2004) and in panel unit root and panel cointegration tests (see *inter alia*, respectively Chang, 2004, and Fachin, 2007) suggest this point is worth investigating. The first goal of our paper is thus to compare the estimation performances of single-equation and system estimators in panels with short-run dependence across units. Since FM-SUR, contrary to DSUR, is feasible in systems of the dimension typically encountered in practice, it is natural to concentrate on the two FM estimators. Second, we will compare simulation and asymptotic inference performance of the FM-OLS estimator with those delivered by asymptotic inference on the FM-SUR estimator. We shall now first discuss the bootstrap inference procedures (section 2), then present the design and results of our Monte Carlo experiment (section 3), while some conclusions are drawn in section 4.

2 Bootstrap procedures

Consider for the sake of simplicity the case of two I(1) variables, Y and X, observed on a panel of N units and T time observations. We assume a linear long-run equilibrium relationship, with possibly heterogenous coefficients, holds in all units. Formally:

$$y_{it} = \theta_i + \beta_i x_{1it} + u^y_{it} \tag{1}$$

where $x_{it} = x_{it-1} + u^x_{it}$. In both cases $i = 1, \ldots, N$, and $t = 1, \ldots, T$. Bootstrap inference involves two key steps: first, constructing the pseudo-datasets; second, defining the test statistics or confidence intervals to be used. Let us examine them in turn.

When constructing pseudo-data sets from non-stationary dependent panels the key point is to reproduce the presence of dependence both in the time series and in the cross-section dimensions. The former aspect has been the subject of the vast debate, whose details are beyond the scope of this paper (for a review, see Politis, 2003). Essentially, we can either follow a model-based (parametric) approach or a non-parametric one. In the former case in a first step the data are filtered through AR models, so to obtain white-noise residuals to be resampled. In the latter blocks of observations of length proportional to the memory of the series and random starting point are drawn with replacement from the dependent series. As in Di Iorio and Fachin (2007), we will follow the latter approach. The bootstrap noises are thus obtained resampling the residuals of the FM regressions with a block bootstrap algorithm, the Stationary bootstrap (SB; Politis and Romano, 1994). To preserve the cross-unit dependence structure we simply need to resample the entire

$T \times N$ matrix of residuals. The systematic part of the bootstrap Data Generating Process (DGP) will depend on the purpose of the exercise: in the case of hypothesis testing it is given by the null hypothesis to be tested, while for interval estimation by the results of unconstrained estimation. Summing up, when the aim is testing the hypothesis $H_{0:}\ \beta_i = \beta_i^{(0)}$ the bootstrap DGP is:

$$y_{it}^* = \widehat{\theta}_i + \beta_i^{(0)} x_{1it} + u_{it}^{*y} \tag{2}$$

while for interval estimation we use

$$y_{it}^* = \widehat{\theta}_i + \widehat{\beta}_i x_{1it} + u_{it}^{*y} \tag{3}$$

where $\widehat{\alpha}_i, \widehat{\beta}_i$ are the unconstrained FM-OLS estimates. In both cases $\widehat{u}_{it}^{*y}$ is obtained applying a stationary bootstrap algorithm to the unconstrained residuals $\widehat{u}_{it}^{y} = y_{it} - \widehat{\theta}_i - \widehat{\beta}_i x_{1it}$. A thorough discussion of the choice of block length, not allowed here by space constraints, is included in Paparoditis and Politis (2003). As usual, in the cases of two-tailed tests the bootstrap estimate of the p-value will be $p^* = prop(|t_b^*| > t)$, with $t = s_{\beta_i}^{-1}(\widehat{\beta}_i - \beta_i^{(0)})$, $t_b^* = s_{\beta_i}^{*-1}(\widehat{\beta}_{ib}^* - \widehat{\beta}_i)$, $\widehat{\beta}_{ib}^*$ the FM-OLS estimate of β_i computed on the $b-th$ pseudo-dataset $(b = 1, \ldots, B)$, s_{β_i} and $s_{\beta_i}^*$ the estimated standard errors of the estimators. One simple way to compute confidence intervals is to take the desired quantiles of the distribution of the $\widehat{\beta}_{ib}^{*\prime}$s. An α-level confidence interval for β_i is then simply given by $[Q_{\alpha/2}(\widehat{\beta}_i^*), Q_{1-\alpha/2}(\widehat{\beta}_i^*)]$, where $\widehat{\beta}_i^* = \left[\widehat{\beta}_{i1}^* \ldots \widehat{\beta}_{iB}^*\right]$. In principle basing the interval on a pivotal quantity should deliver better results. Psaradakis (2001) suggests the percentile-t interval $[\widehat{\beta}_i - Q_{1-\alpha/2}(\mathbf{t}_b^*)s_{\beta_i}, \widehat{\beta}_i - Q_{\alpha/2}(\mathbf{t}_b^*)s_{\beta_i}]$, where the Gaussian quantiles used in asymptotic inference are replaced by those of the bootstrap distribution (empirical estimate of the unknown small sample distribution of the studentized statistic). The superiority of the second type of interval depends entirely upon the quality of the estimates of the standard errors (see *e.g.*, Kilian, 1999). Hence, in our study we shall compute both type of intervals.

3 Monte Carlo experiment

3.1 Design

The natural starting point of our simulation study is Moon and Perron (2004), who carried out a simulation study in a traditional seemingly unrelated equations set-up of small system size. These authors concluded that system estimators were overall superior to single-equation ones. Our contribution will be considering larger systems, able to mimic the datasets used in non-stationary panel analysis, and bootstrap as well as traditional Gaussian inference. As we will see, we will reach conclusions rather different from Moon and Perron's. Our DGP can be thus summarised as follows. In each unit of the panel two

right-hand side $I(1)$ variables (X_1, X_2) and a left-hand side variable (Y) are linked by a linear long-run equilibrium relationship:

$$y_{it} = \theta_i + \beta_{1i} x_{1it} + \beta_{2i} x_{2it} + u^y_{it}, i = 1, \ldots, N; \tag{4}$$

$$x_{kit} = x_{kit-1} + u^x_{kit}, k = 1, 2;\ i = 1, \ldots, N. \tag{5}$$

Generalising the design used in Pedroni (2000), where homogeneity is assumed, we generate the regression coefficients as Uniform variates, respectively $\theta_i \sim Uniform(2, 4)$ and $\beta_{ki} \sim Uniform(1, 3)$, where $k = 1, 2$. The errors of equations (5) and (4) are drawn from a Multivariate Normal distribution with non-diagonal covariance matrix, so that there is feedback across equations and units. More precisely, letting $\mathbf{u}^x_t = [\mathbf{u}^{x\prime}_{1t} \mathbf{u}^{x\prime}_{2t} \ldots \mathbf{u}^{x\prime}_{Nt}]'$, where $\mathbf{u}^{x\prime}_{it} = [u^x_{1it} u^x_{2it}]'$ and $\mathbf{u}^y_t = [u^y_{1t} u^y_{2t} \ldots u^y_{Nt}]'$, we have

$$\underset{(N+2N)\times 1}{\begin{bmatrix} \mathbf{u}^y_t \\ \mathbf{u}^x_t \end{bmatrix}} = MN\left(\underset{(N+2N)\times(N+2N)}{\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} R & \Delta \\ \Delta' & \Phi \end{bmatrix}} \right),$$

where R is a full $N \times N$ matrix governing the dependence across units in the $u^{y\prime}_{it}$s, Δ is a $N \times 2N$ matrix governing the dependence between the u^x and u^y noises, and finally Φ is a $2N \times 2N$ matrix governing the dependence in the $u^{x\prime}s$ within and across units. Since Moon and Perron report both the performances of both FM-OLS and FM-SUR estimators to be negatively affected by the degree of endogeneity of the X's we decided to control accurately δ, imposing the homogeneity assumption and running two sets of experiments with $\delta = 0.2$ and $\delta = 0.4$. In both cases the Δ matrix has a block form ensuring that there is constant correlation between the noise of any X and that of the relevant Y equation, and no correlation across units:

$$\underset{N\times 2N}{\Delta} = \begin{bmatrix} \delta & \delta & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \delta & \delta & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \delta & \delta \end{bmatrix}.$$

We instead allow heterogeneity across units in the dependence parameters, generating them as $Uniform(0.3, 0.4)$ random variates; without loss of generality we assume different X's in the same unit to be orthogonal. Letting $\phi^{(ij)}_{lk} = cov(u^x_{li}, u^x_{kj})$, the covariance between the noise of X_l in the i^{th} unit

and X_k in the j^{th} unit, we then have:

$$\underset{2N\times 2N}{\boldsymbol{\Phi}} = \begin{bmatrix} 1 & 0 & \phi_{11}^{(12)} & \phi_{12}^{(12)} & \dots & \phi_{11}^{(1N)} & \phi_{12}^{(1N)} \\ 0 & 1 & \phi_{21}^{(12)} & \phi_{22}^{(12)} & \dots & \phi_{21}^{(1N)} & \phi_{22}^{(1N)} \\ \phi_{11}^{(21)} & \phi_{12}^{(21)} & 1 & 0 & \dots & \phi_{11}^{(2N)} & \phi_{12}^{(2N)} \\ \phi_{21}^{(21)} & \phi_{22}^{(21)} & 0 & 1 & \dots & \phi_{21}^{(2N)} & \phi_{22}^{(2N)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi_{11}^{(N1)} & \phi_{12}^{(N1)} & \phi_{11}^{(N2)} & \phi_{12}^{(N2)} & \dots & 1 & 0 \\ \phi_{21}^{(N1)} & \phi_{22}^{(N1)} & \phi_{21}^{(N2)} & \phi_{22}^{(N2)} & \dots & 0 & 1 \end{bmatrix}$$

The choice of the time and cross-section sample sizes have been fixed trying to strike a balance between empirical relevance (with most macroeconomic panels with N around 10 or 20 and T around 30) and the requirements of the SUR estimator, which is feasible only with a rather large T/N ratio. We thus fixed $N = 5, 10$ and $T = 50, 100$. Finally, we set the number of bootstrap redrawings (B) and Monte Carlo simulations (M) to 1000.

3.2 Results

In Tables 1 and 2 we report summary statistics of the performances of respectively FM-OLS and FM-SUR estimators, further averaging the usual Monte Carlo means over units and variables.

Point estimation performance is evaluated by the average absolute relative bias $100 \times (2N)^{-1} \sum_k^2 \sum_i^N \mid M^{-1} \sum_m^M (\widehat{\beta}_{kim} - \beta_{ki})\beta_{ki}^{-1} \mid$ while the dispersion by the relative Monte Carlo standard error. The first remark in order is that the fully modified estimators are indeed affected by the degree of endogeneity present in the system; the performance of both FM-OLS and FM-SUR estimators are slightly, but clearly, inferior with a higher δ. The second is that the SUR procedure turned out to be practically unfeasible for $T = 50$ and $N = 10$. The covariance matrix, although not exactly singular, was always so ill-conditioned that the estimators turned out highly numerically unstable. Hence, we do not report them here. Since this (T, N) combination can be considered rather representative of the sample sizes used in applied work on non-stationary panels (with indeed the time sample often actually smaller than this one) this is an important finding. In the other cases both estimators are essentially unbiased even with the smaller time sample, although the SUR estimator is always somehow more biased than the OLS one. For instance, for $T = 50$, $N = 5$ and $\delta = 0.2$ the average relative bias is 0.40% for the former and 0.34% for the latter. The Monte Carlo standard errors are very similar, with again the single-equation estimator always slightly superior to the SUR one (for the same case, respectively 3.98% and 3.41%). Coverage and Type I errors of Gaussian inference on FM-OLS are both disappointing, with severe overrejection and undercoverage. For the same parameters

combination quoted above the Type I error of a 5% test is 14.63% and the coverage, as a consequence, 85.37%. Both problems are partially solved using the bootstrap, although coverage is inferior to nominal for both the basic and the studentized intervals (respectively, 89.38% and 90.11%) and the test underrejects (Type I error 2.36%). On the other hand, the performance of asymptotic inference on the SUR estimator is simply disastrous, with Type I errors close to 50% when T is not large with respect to N (that is, always for $T = 50$ and when $N = 10$ for $T = 100$) and around 20% even in the more favorable case of $T = 100$, $N = 5$. The reason for this extremely poor performance, not obvious from the bias and Monte Carlo variability statistics, is found in Table 3: the detailed results for the case $T = 50$, $N = 5$, $\delta = 0.4$[1] show that the standard formulas for the variance of the SUR estimator grossly underestimate its actual variance.

Table 1

FM-OLS: Estimation and Inference Performance.

				Coverage			Type I Err	
T	N	$\overline{bias}$	$\overline{s.e.}$	*Asy*	*boot*	*boot-t*	*Asy*	*boot*
				$\delta = 0.2$				
50	5	0.34	3.41	85.37	89.38	90.11	14.63	2.36
	10	0.35	4.53	86.13	89.91	90.25	13.88	3.51
100	5	0.15	1.87	90.76	91.89	92.23	9.24	0.88
	10	0.17	1.89	90.57	91.54	93.52	9.43	3.81
				$\delta = 0.4$				
50	5	0.64	3.40	84.99	89.71	90.08	15.01	2.78
	10	0.57	4.49	86.00	90.23	90.27	14.00	3.36
100	5	0.25	1.86	90.40	91.99	92.00	9.60	0.78
	10	0.20	1.88	90.55	91.93	93.34	9.45	3.51

$\overline{bias}$: $100 \times (2N)^{-1} \sum_k^2 \sum_i^N \mid M^{-1} \sum_m^M (\widehat{\beta}_{kim} - \beta_{ki}) \beta_{ki}^{-1} \mid$

$\overline{s.e.}$: $(2N)^{-1} \sum_i^N \sum_k^2 \left[\left(\sqrt{M^{-1} \sum_m^M (\widehat{\beta}_{kim} - \overline{\widehat{\beta}}_{ki})^2} \right) \beta_{ki}^{-1} \right] \times 100;$

Coverage: proportion of 5% confidence intervals including the true value of the coefficient of interest;

[1] Detailed results for the other cases do not provide any additional insights and thus are not reported here, but, as customary, available on request.

Table 2

FM-SUR: Estimation and Inference Performance.

				Coverage	Type I Err
T	N	$\overline{bias}$	$\overline{s.e.}$	Asy	Asy
				$\delta = 0.2$	
50	5	0.40	3.98	54.62	45.38
	10	-	-	-	
100	5	0.18	1.92	80.96	19.04
	10	0.27	2.20	57.18	42.82
				$\delta = 0.4$	
50	5	0.79	4.04	52.56	47.44
	10	-	-	-	
100	5	0.35	1.94	79.72	20.28
	10	0.51	2.22	55.48	45.52

-: not available (numerical overflow);

all symbols and abbreviations: see Table 1

Table 3

Bias and Variability of FM-OLS and FM-SUR estimators.

$T = 50$, $N = 5, \delta = 0.4$

		$bias$		$MC\ s.e.$		$\overline{\widehat{\sigma}}$		$\overline{\widehat{\sigma}} - MC\ s.e$	
Unit		OLS	SUR	OLS	SUR	OLS	SUR	OLS	SUR
1	β_1	-0.13	0.44	5.1	5.3	3.8	2.2	1.2	3.1
	β_2	0.25	0.57	6.4	7.9	5.2	3.0	1.2	4.9
2	β_1	-0.33	0.21	3.0	3.2	2.3	1.4	0.7	1.9
	β_2	1.09	1.27	6.0	6.6	4.7	2.7	1.4	3.9
3	β_1	0.65	0.60	10.7	13.6	8.1	4.3	2.6	9.3
	β_2	-0.38	0.35	5.7	6.8	4.3	2.3	1.4	4.5
4	β_1	0.52	2.34	9.8	12.7	8.2	4.5	1.6	8.2
	β_2	-0.38	0.01	3.8	4.4	3.1	1.8	0.8	2.7
5	β_1	2.35	0.96	11.1	13.5	8.5	4.6	2.6	9.0
	β_2	0.23	1.10	4.7	5.3	3.6	2.1	1.1	3.2

MC $s.e.$: Monte Carlo s.e.; $\overline{\widehat{\sigma}}$: average estimated standard error $\times 100$;

other symbols and details: see Table 1

4 Conclusions

Our main conclusion is very simple: on the basis of our simulation exercise the best option in non-stationary panel analysis seems to be given by single-equation estimators with bootstrap inference. The potential efficiency gains of SUR-type estimators remain such even in the restrictive case of no long-run relationships across units. In fact, when the time dimension is not very large relatively to the cross-section dimension the covariance matrix is likely to be so ill-conditioned to make the resulting estimates essentially meaningless. Further, even when some meaningful point estimates can be obtained, their variance is likely to be grossly understimated by standard formulas, with disastrous effects on inference. These conclusions are in stark contrast to Moon and Perron's (2004). However, this should not come as a surprise. The properties of SUR estimators depend critically upon the quality of the estimate of the covariance matrix. This task may be easy in small systems, such as those examined by Moon and Perron, but is typically difficult in even slightly larger systems, such those considered in our study.

References

BREITUNG, J. (2005): A Parametric Approach to the Estimation of Cointegration Vectors in Panel Data *Econometric Reviews, 24, 151-174.*

CHANG, Y. (2004): Bootstrap Unit Root Tests in Panels with Cross-Sectional Dependency. *Journal of Econometrics 120, 263-293.*

DI IORIO, F. and SFACHIN, S. (2007): Testing for breaks in cointegrated panels. *Economics - The Open-Access, Open-Assessment E-Journal 2007-14.*

FACHIN, S. (2004): Bootstrap inference on Fully Modified Estimates of Cointegrating Coefficients: A Comment. *Economics Bulletin 3, 1-8.*

FACHIN, S. (2007): Long-Run Trends in Internal Migrations in Italy: a Study in Panel Cointegration with Dependent Units. *Journal of Applied Econometrics 22, 401-428.* DOI:10.1002/jae.907 Further information in IDEAS/RePEc

GROEN, J.J.J. and KLEIBERGEN, F. (2003): Likelihood-based cointegration analysis in panels of vector error-correction models. *Journal of Business and Economic Statistics 21, 295-318.*

KILIAN, L. (1999): Finite-Sample Properties of Percentile and Percentile-t Bootstrap Confidence Intervals for Impulse Responses. *The Review of Economics and Statistics 81, 652-660.*

MARK, N.C., OGAKI, M. and SUL, D. (2005): Dynamic Seemingly Unrelated Cointegrating Regressions. *Review of Economic Studies 72, 797-820.*

MOON, H.R. (1999): A note on fully-modified estimation of seemingly unrelated regressions models with integrated regressors. *Economics Letters 65,25-31.*

MOON, H.R., PERRON, B. (2004): Efficient Estimation of the Seemingly Unrelated Regression Cointegration Model and Testing for Purchasing Power Parity. *Econometric Reviews 23, 293-323.*

PAPARODITIS, E., POLITIS, D.N., (2003): Residual-based block bootstrap for unit root testing. *Econometrica 71, 813-855.*

PEDRONI, P. (2000): Fully Modified OLS for heterogenous cointegrated panels. In *Advances in Econometrics, 15, 93-130.*

PESARAN, M.H., SHIN, Y. and SMITH, R.P.(1999): Pooled Mean Group Estimation of Dynamic Heterogeneous Panels. *Journal of the American Statistical Association 94, 621-624.*

POLITIS, D.N., ROMANO, J.P. (1994): The Stationary Bootstrap. *Journal of the American Statistical Association 89,1303-1313.*

POLITIS, D. (2003): The Impact of Bootstrap Methods on Time Series Analysis. *Statistical Science 18,219-230.*

PSARADAKIS, Z. (2001): On bootstrap inference in cointegrating regressions. *Economics Letters 72,1-10.*

The Exact Likelihood Function of a Vector Autoregressive-Moving Average Process

José L. Gallego

Departamento de Economía, Universidad de Cantabria
Avda. de los Castros, 39005 Santander, Spain, *jose.gallego@unican.es*

Abstract. Several algorithms proposed to evaluate the exact likelihood function of a vector autoregressive moving average model are derived following a common and very simple approach. The transparency in the new derivation reveals clearly the relative merits of these procedures and eases the development of statistical software to estimate this class of models. A collection of C++ routines to build multivariate transfer function models is available from the author.

Keywords: estimation, multiple time series, vector autoregressive moving average process

1 Introduction

Let $\mathbf{w}_t = (w_{1t}, \ldots, w_{kt})'$ be a stationary k-dimensional vector autoregressive moving average (VARMA) process of order (p, q) defined by the equation

$$\boldsymbol{\Phi}(B)\mathbf{w}_t = \boldsymbol{\Theta}(B)\mathbf{a}_t, \quad t = 0, \pm 1, \pm 2, \ldots \tag{1}$$

where $\boldsymbol{\Phi}(B) = \mathbf{I}_k - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p$ and $\boldsymbol{\Theta}(B) = \mathbf{I}_k - \boldsymbol{\Theta}_1 B - \cdots - \boldsymbol{\Theta}_q B^q$ are matrix polynomials in the backshift operator B, such that $B^j \mathbf{w}_t = \mathbf{w}_{t-j}$ and $B^j \mathbf{a}_t = \mathbf{a}_{t-j}$, $\mathbf{I}_k$ is the identity matrix of order k, $\boldsymbol{\Phi}_i$ $(i = 1, \ldots, p)$ and $\boldsymbol{\Theta}_j$ $(j = 1, \ldots, q)$ are $k \times k$ unknown parameter matrices, and $\mathbf{a}_t = (a_{1t}, \ldots, a_{kt})'$ is a k-dimensional white noise process with zero mean vector, $E(\mathbf{a}_t) = \mathbf{0}$, and positive definite covariance matrix, $E(\mathbf{a}_t \mathbf{a}_t') = \boldsymbol{\Omega}_a$. It has been assumed, for convenience, that $E(\mathbf{w}_t) = \mathbf{0}$. The process $\{\mathbf{w}_t\}$ is said to be stationary and invertible if the zeros of the determinantal equations $|\boldsymbol{\Phi}(B)| = 0$ and $|\boldsymbol{\Theta}(B)| = 0$ lie outside the unit circle, respectively.

Assuming Gaussians errors, the $\mathbf{w}_t$'s have a normal distribution and the likelihood function for the parameters $\boldsymbol{\beta} = (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q, \boldsymbol{\Omega}_a)$ is given by

$$L(\boldsymbol{\beta}|\mathbf{w}) = (2\pi)^{-kn/2} |\boldsymbol{\Omega}_w|^{-1/2} \exp(-\frac{1}{2}\mathbf{w}' \boldsymbol{\Omega}_w^{-1} \mathbf{w}) \tag{2}$$

where n is the number of observations, $\mathbf{w} = (\mathbf{w}_1', \ldots, \mathbf{w}_n')'$ is a sample vector time series of length kn from (1) and $\boldsymbol{\Omega}_w$ is the $kn \times kn$ covariance matrix $\boldsymbol{\Omega}_w = E(\mathbf{w}\mathbf{w}')$. The $(i, j)th$ block of $\boldsymbol{\Omega}_w$ is the $k \times k$ covariance matrix

$\boldsymbol{\Gamma}_{i-j} = E(\mathbf{w}_{t-i}\mathbf{w}'_{t-j})$ $(i, j = 1, \ldots, n)$, which can be expressed in terms of $\boldsymbol{\beta}$ (e.g., see Nicholss and Hall (1979) and Mittnik (1990)).

The direct evaluation of the exact likelihood function (2) requires computing the covariance matrices $\boldsymbol{\Gamma}_i$ $(i = 0, \ldots, n-1)$ and the determinant and inverse of the covariance matrix $\boldsymbol{\Omega}_w$. In the univariate case $(k = 1)$, three approaches have been proposed for significantly reducing the computational burden and the time processing when working with moderate size samples. Newbold (1974) and Ljung and Box (1979) gave various closed forms for the determinant and the inverse by applying the generalized least squares theory, Ansley (1979) found that the time series filtered by the AR polynomial has a band-diagonal covariance matrix whose Cholesky factor is easily computable, and Harvey and Phillips (1979) and Pearlman (1980) used the Kalman filter. Multivariate extensions of these algorithms have been proposed by, among others, Hilmer and Tiao (1979), Nicholls and Hall (1979), Mauricio (1995) and Ma (1997); Mauricio (2002); Ansley and Kohn (1983) and Shea (1987).

Mauricio (1995) found a computationally efficient way of evaluating the explicit expressions for the determinant and inverse given by Nicholls and Hall (1979). Besides computational superiority, it is claimed that the algorithm gathers all the advantages attributed to the competing procedures (for example, the automatic detection of nonstationary and noninvertible models, which is very useful in the subsequent maximization) and does not suffer any of their drawbacks. However, a possible disadvantage of this algorithm is that its theoretical derivation based on tricky matrix identities appears obscure and difficult to follow. On the contrary, Reinsel (1995, chap. 5) describes a very intuitive closed form algorithm, whose derivation only requires to familiarize oneself with the matrix form of the VARMA model and apply a well-known matrix inversion lemma. To date, the connection between the Mauricio and Reinsel algorithms was not noticed and, consequently, the advantages of efficiency and simplicity provided by both have not been yet combined.

The purpose of this note is to show that the Mauricio algorithm can be readily obtained as a variant of a modified version of the Reinsel algorithm. The new derivation described in the next section reveals clearly that the Mauricio algorithm is computationally superior to the Reinsel algorithm only for mixed VARMA processes, being both equally efficient for pure autoregressive or moving average processes. For the sake of completeness, some guidelines useful for the efficient implementation of a computational algorithm are also given in Section 3.

2 The exact likelihood function

The common starting point in the derivation of a closed-form algorithm is the model (1) written in matrix form (e.g., see Reinsel (1995) p. 122).

$$\boldsymbol{\Phi}\mathbf{w} = \boldsymbol{\Theta}\mathbf{a} + \mathbf{F}\mathbf{u} \tag{3}$$

where $\mathbf{a} = (\mathbf{a}_1', \ldots, \mathbf{a}_n')'$ is an $kn \times 1$ vector, $\mathbf{u} = (\mathbf{w}_{-p+1}, \ldots, \mathbf{w}_0, \mathbf{a}_{-q+1}, \ldots, \mathbf{a}_0)$ is an $k(p+q) \times 1$ vector of presample values, $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are $kn \times kn$ parameter matrices, and $\mathbf{F}$ is an $kn \times k(p+q)$ parameter matrix. The block matrix $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_{ij}]$ $(i = 1, \ldots, n;\ j = 1, \ldots, n)$ is defined as $\boldsymbol{\Phi}_{ij} = -\boldsymbol{\Phi}_{i-j}$ for $0 \leq i - j \leq p$ and $\boldsymbol{\Phi}_{ij} = \mathbf{0}$ otherwise, where the $k \times k$ matrix $\boldsymbol{\Phi}_{ij}$ is the typical block of $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}_0 = -\mathbf{I}_k$. Similarly, the block matrix $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_{ij}]$ $(i = 1, \ldots, n;\ j = 1, \ldots, n)$ is such that $\boldsymbol{\Theta}_{ij} = -\boldsymbol{\Theta}_{i-j}$ for $0 \leq i - j \leq q$ and $\boldsymbol{\Theta}_{ij} = \mathbf{0}$ otherwise, where $\boldsymbol{\Theta}_0 = -\mathbf{I}_k$; and the block matrix $\mathbf{F} = [\mathbf{F}_{ij}]$ $(i = 1, \ldots, n;\ j = 1, \ldots, p+q)$ is such that $\mathbf{F}_{ij} = \boldsymbol{\Phi}_{p-(j-i)}$ for $0 \leq j - i < p$, $\mathbf{F}_{ij} = -\boldsymbol{\Theta}_{p+q-(j-i)}$ for $p \leq j - i < p+q$, and $\mathbf{F}_{ij} = \mathbf{0}$ otherwise. Although the above notation can seem confusing at first, it is very useful in the computer implementation of the algorithm.

Since $\mathbf{u}$ is independent of $\mathbf{a}$, the covariance matrix of $\mathbf{w}$ is

$$\boldsymbol{\Omega}_w = \boldsymbol{\Phi}^{-1}[\boldsymbol{\Theta}(\mathbf{I}_n \otimes \boldsymbol{\Omega}_\mathbf{a})\boldsymbol{\Theta}' + \mathbf{F}\boldsymbol{\Omega}_\mathbf{u}\mathbf{F}']\boldsymbol{\Phi}^{-1'}$$

where the $k(p+q) \times k(p+q)$ covariance matrix $\boldsymbol{\Omega}_u = [\boldsymbol{\Omega}_{ij}]$ $(i = 1, \ldots, p+q;\ j = 1, \ldots, p+q)$ is defined as $\boldsymbol{\Omega}_{ij} = \boldsymbol{\Gamma}_{i-j}$ for $1 \leq i, j \leq p$, $\boldsymbol{\Omega}_{ij} = \boldsymbol{\Psi}_{q+i-j}\boldsymbol{\Omega}_\mathbf{a}$ for $1 \leq i \leq p$ and $p < j \leq p+q$, $\boldsymbol{\Omega}_{ij} = \boldsymbol{\Omega}_\mathbf{a}$ for $p < i \leq p+q$ and $i = j$, $\boldsymbol{\Omega}_{ji} = \boldsymbol{\Omega}_{ij}'$ and $\boldsymbol{\Omega}_{ij} = \mathbf{0}$ otherwise. The matrices $\boldsymbol{\Psi}_i$ $(i = 0, 1, \ldots, q-1)$ are the weights of the infinite matrix polynomial $\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}(B)^{-1}\boldsymbol{\Theta}(B)$ (e.g., see Reinsel (1995) p. 123).

In order to express the likelihood function (2) in a computationally efficient form, the covariance matrix of $\mathbf{w}$ is written as

$$\boldsymbol{\Omega}_w = \mathbf{A}(\mathbf{I}_{kn} + \mathbf{B}\mathbf{B}')\mathbf{A}'$$

where $\mathbf{A} = \boldsymbol{\Phi}^{-1}\boldsymbol{\Theta}(\mathbf{I}_n \otimes \mathbf{P}_a)$ is an $kn \times kn$ lower triangular block matrix, $\mathbf{B} = (\mathbf{I}_n \otimes \mathbf{P}_a^{-1})\boldsymbol{\Theta}^{-1}\mathbf{F}\mathbf{P}_u$ is an $kn \times k(p+q)$, and $\mathbf{P}_a$ and $\mathbf{P}_u$ are the Cholesky factors of $\boldsymbol{\Omega}_a$ and $\boldsymbol{\Omega}_u$, respectively. Using standard matrix results, we find that

$$\boldsymbol{\Omega}_w^{-1} = \mathbf{A}'^{-1}[\mathbf{I}_{kn} - \mathbf{B}(\mathbf{I}_{k(p+q)} + \mathbf{B}'\mathbf{B})^{-1}\mathbf{B}']\mathbf{A}^{-1}$$

and

$$|\boldsymbol{\Omega}_w| = |\mathbf{A}||\mathbf{I}_{kn} + \mathbf{B}\mathbf{B}'||\mathbf{A}'| = |\mathbf{I}_{k(p+q)} + \mathbf{B}'\mathbf{B}||\mathbf{P}_a|^{2n}$$

Thus, the quadratic form in the exponent of the likelihood function (2) can be evaluated as

$$\mathbf{w}'\mathbf{A}'^{-1}[\mathbf{I}_{kn} - \mathbf{B}(\mathbf{I}_{k(p+q)} + \mathbf{B}'\mathbf{B})^{-1}\mathbf{B}']\mathbf{A}^{-1}\mathbf{w} \tag{4}$$

This expression is equivalent to the equation (5.27) of Resinsel (1995), and shows that in order to evaluate the likelihood function (2) it is never necessary to invert matrices larger than $k(p+q) \times k(p+q)$, which reduces considerably the computational burden compared to the direct evaluation. Furthermore, it should be noticed that, due to their triangular block structure, the inversion of $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ is not required to compute the matrices $\mathbf{A}$ and $\mathbf{B}$. More on this later.

Note now that the matrices $\mathbf{F}$ and $\mathbf{F}\boldsymbol{\Omega}_u\mathbf{F}'$ can be partitioned as

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{F}\boldsymbol{\Omega}_u\mathbf{F}' = \begin{pmatrix} \mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

where $\mathbf{F}_1$ and $\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1'$ are $kr \times k(p+q)$ and $kr \times kr$ matrices, respectively, with $r = \max(p, q)$. Therefore, there exists a $kn \times kr$ matrix $\mathbf{R} = [\mathbf{R}_1' \; \mathbf{0}']'$ such that $\mathbf{F}\boldsymbol{\Omega}_u\mathbf{F}' = \mathbf{R}\mathbf{R}'$, where $\mathbf{R}_1$ is the Cholesky factor of $\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1'$.

Hence, a variant of the above outlined approach, which is computationally more efficient when both p and q are positive, can be obtained by writing the covariance matrix $\boldsymbol{\Omega}_w$ as

$$\boldsymbol{\Omega}_w = \mathbf{A}(\mathbf{I}_{kn} + \mathbf{C}\mathbf{C}')\mathbf{A}'$$

where $\mathbf{C} = (\mathbf{I}_n \otimes \mathbf{P}_a^{-1})\boldsymbol{\Theta}^{-1}\mathbf{R}$ is an $kn \times kr$ matrix. Thus, the quadratic form in the exponent of the likelihood function

$$\mathbf{w}'\mathbf{A}'^{-1}[\mathbf{I}_{kn} - \mathbf{C}(\mathbf{I}_{kr} + \mathbf{C}'\mathbf{C})^{-1}\mathbf{C}']\mathbf{A}^{-1}\mathbf{w} \tag{5}$$

and the determinant

$$|\boldsymbol{\Omega}_w| = |\mathbf{I}_{kr} + \mathbf{C}'\mathbf{C}||\mathbf{P}_a|^{2n} \tag{6}$$

are now equivalent to equations (15) and (16) of Mauricio (1995), but the new derivation is considerably simpler.

Comparing (4) with (5), we can see that the main difference among the algorithms of Reinsel and Mauricio lies in the treatment of the matrix $\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1'$ associated to the presample values. While that the former defines $\mathbf{B}$ in terms of the $k(p+q) \times k(p+q)$ Cholesky factor of $\boldsymbol{\Omega}_u$, the latter defines $\mathbf{C}$ in terms of the $kr \times kr$ Cholesky factor of the entire matrix $\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1'$. In this way, the dimension reduction in the larger matrix to be inverted $\mathbf{I}_{kr} + \mathbf{C}'\mathbf{C}$ with respect to $\mathbf{I}_{k(p+q)} + \mathbf{B}'\mathbf{B}$ is given by $\min(kp, kq)$. Note also that the inverse of the covariance matrix used in (5) is equivalent to the equation (3.2) of Ma (1997), but the latter requires the inversion of two $kr \times kr$ matrices due to the partition the vector $\mathbf{u} = (\mathbf{w}_{-p+1}, \ldots, \mathbf{w}_0|\mathbf{a}_{-q+1}, \ldots, \mathbf{a}_0)$. Finally, defining $\mathbf{L}' = [\mathbf{0}|\mathbf{A}'^{-1}]$ and $\mathbf{X}' = [\mathbf{I}_{kr}| - \mathbf{C}']$, (5) can be expressed as the residual sum of squares of the regression of $\mathbf{L}\mathbf{w}$ on $\mathbf{X}$

$$\mathbf{w}'\mathbf{L}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{L}\mathbf{w} \tag{7}$$

which is the quadratic form suggested by Nicholls and Hall (1979), but that has been here derived without using concentrated maximum likelihood or generalized least squares procedures.

3 Computational considerations

From equations (5) and (6) it should be clear that the problem of evaluating the exact likelihood function of a VARMA model is reduced to that of computing the matrices $\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1'$, $\mathbf{C}$ and $\mathbf{A}^{-1}\mathbf{w}$.

The most difficult task is to create $\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1'$, which can be indirectly computed by using the matrix identity

$$\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1' = \boldsymbol{\Phi}_{11}\boldsymbol{\Omega}_{w,11}\boldsymbol{\Phi}_{11}' - \boldsymbol{\Theta}_{11}(\mathbf{I}_r \otimes \boldsymbol{\Omega}_a)\boldsymbol{\Theta}_{11}'$$

obtained from the r first observations in the partition of the model (3) coherent with the partitioned matrix $\mathbf{F} = (\mathbf{F}_1', \mathbf{0}')'$. Note that the $kr \times kr$ submatrices $\boldsymbol{\Phi}_{11}$, $\boldsymbol{\Theta}_{11}$ and $\boldsymbol{\Omega}_{w,11}$ contain the first kr rows and columns of $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}_w$, respectively. Additionally, to create the submatrix $\boldsymbol{\Omega}_{w,11}$ the covariance matrices $\boldsymbol{\Gamma}_i$ $(i = 0, \ldots, r-1)$ must be determined in terms of $\boldsymbol{\beta}$.

Once the matrix $\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1'$ has been calculated, its Cholesky factor $\mathbf{R}_1 = [\mathbf{R}_{ij}]$ $(i, j = 1, \ldots, r)$ is used to compute recursively the blocks $\mathbf{C}_{ij}$ $(i = 1, \ldots, n;\ j = 1, \ldots, r)$ of the matrix $\mathbf{C}$ from

$$\mathbf{C}_{ij} = \boldsymbol{\Theta}_1^*\mathbf{C}_{i-1,j} + \boldsymbol{\Theta}_2^*\mathbf{C}_{i-2,j} + \cdots + \boldsymbol{\Theta}_q^*\mathbf{C}_{i-q,j} + \mathbf{P}_a^{-1}\mathbf{R}_{ij}$$

where $\boldsymbol{\Theta}_i^* = \mathbf{P}_a^{-1}\boldsymbol{\Theta}_i\mathbf{P}_a$, $\mathbf{C}_{ij} = 0$ for $i < 1$ and $\mathbf{R}_{ij} = \mathbf{0}$ for $i > r$.

Similarly, the tth element of the block vector $\mathbf{A}^{-1}\mathbf{w} = [\mathbf{P}_a^{-1}\hat{\mathbf{a}}_t]$ is obtained from the recursion

$$\hat{\mathbf{a}}_t = \mathbf{w}_t - \boldsymbol{\Phi}_1\mathbf{w}_{t-1} - \cdots - \boldsymbol{\Phi}_p\mathbf{w}_{t-p} + \boldsymbol{\Theta}_1\hat{\mathbf{a}}_{t-1} + \cdots + \boldsymbol{\Theta}_q\hat{\mathbf{a}}_{t-q}$$

where the presample values are fixed at zero and the $\hat{\mathbf{a}}_t$'s are the so-called conditional residuals.

We can now follow similar steps to those suggested by Mauricio (1995) for evaluating the likelihood at a given value of $\boldsymbol{\beta}$.

(i) Compute the Cholesky factor $\mathbf{P}_a$, its inverse $\mathbf{P}_a^{-1}$ and determinant $|\mathbf{P}_a|$.
(ii) Calculate the theoretical covariance matrices $\boldsymbol{\Gamma}_k$ for $k = 0, \ldots, r-1$.
(iii) Compute $\mathbf{F}_1\boldsymbol{\Omega}_u\mathbf{F}_1'$ and its Cholesky factor $\mathbf{R}_1$.
(iv) Generate recursively $\mathbf{C} = [\mathbf{C}_{ij}]$.
(v) Compute $\mathbf{I}_{kr} + \mathbf{C}'\mathbf{C}$, its Cholesky factor $\mathbf{T} = [\mathbf{T}_{ij}]$ $(i, j = 1, \ldots, r)$ and determinant $|\mathbf{T}|$.
(vi) Generate recursively the conditional residuals $\hat{\mathbf{a}}_t$.
(vii) Transform the conditional residuals $\hat{\mathbf{a}}_t^* = \mathbf{P}_a^{-1}\hat{\mathbf{a}}_t$.
(viii) Obtain the $k \times 1$ vectors $\hat{\mathbf{b}}_j$ $(j = 1, \ldots, r)$

$$\hat{\mathbf{b}}_j = \sum_{i=1}^{n} \mathbf{C}_{ij}'\hat{\mathbf{a}}_i^*$$

(ix) Transform $\hat{\mathbf{b}}_j$ into

$$\hat{\mathbf{b}}_j^* = \sum_{i=1}^{j} \mathbf{T}_{ij}\mathbf{e}_j$$

(x) Evaluate the exact likelihood function as

$$(2\pi)^{-kn/2}|\mathbf{T}|^{-1}|\mathbf{P}_a|^{-n}\exp[-\frac{1}{2}(\sum_{t=1}^{n}\hat{\mathbf{a}}_{*t}'\hat{\mathbf{a}}_{*t} - \sum_{j=1}^{r}\hat{\mathbf{b}}_{*j}'\hat{\mathbf{b}}_{*j})]$$

Several computational considerations are of interest. Firstly, with regards to the storage requirements, steps (4)-(8) can be taken storing only $k(q+1)$ rows of the matrix $\mathbf{C}$. Therefore, in a computer implementation it is not necessary to allocate memory depending on the sample size, apart from the one used by the data. Secondly, the computing time to estimate an VARMA model using the above outlined algorithm obviously depends on factors such as the sample size n, the dimension k, the orders p and q, and the goodness preestimates $\hat{\boldsymbol{\beta}}$. Using a C++ implementation of this algorithm along with the UMINCK optimization algorithm of Dennis and Schnabel (1983), the estimation of the numerical examples considered by Reinsel (1995, pp. 142-153) usually takes less than a second or a few seconds in a conventional notebook. An experimental version of this statistical software can be freely obtained from the author. Thirdly, the algorithm can be easily adapted to evaluate the concentrated log likelihood with respect to one of the variances of $\boldsymbol{\Omega}_a$ or with respect to the entire matrix $\boldsymbol{\Omega}_a$. Finally, the so-called exact residuals $\hat{\mathbf{e}} = E(\mathbf{a}|\mathbf{w})$ can be obtained from (3) as

$$\hat{\mathbf{e}} = \boldsymbol{\Theta}^{-1}\boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\Theta}^{-1}\mathbf{F}\hat{\mathbf{u}} = \boldsymbol{\Theta}^{-1}\boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\Theta}^{-1}\mathbf{R}(\mathbf{I}_{kr} + \mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{A}^{-1}\mathbf{w}$$

where $\hat{\mathbf{u}} = E(\mathbf{u}|\mathbf{w})$ is the vector of estimated presample values.

4 Conclusions

Despite the variety of algorithms proposed to evaluate the exact likelihood function of VARMA processes, this class of models has not got the popularity of the univariate time series models. While the list of statistical software with ARIMA capabilities is quite long, the packages providing VARIMA capabilities can be counted with the fingers on one hand. One possible explanation is that the theoretical derivation of the existing algorithms appears obscure and difficult to implement. Therefore, it seems convenient to develop algorithms more accessible to non-time series specialists. The new derivation of both the Mauricio's algorithm and Hall and Nicholls' algorithm presented here, along the lines of Reinsel's algorithm, has the advantages of simplicity, computational efficiency and ease of implementation, and can easily be extended to handle explanatory variables, intervention effects and missing data.

References

ANSLEY, C. (1979): An algorithm for the exact likelihood of a mixed autoregressive-moving average model. *Biometrika 66 (1), 59-65.*

ANSLEY, C. and KOHN, R. (1983): Exact likelihood of vector autoregressive-moving average process with missing or aggregated data. *Biometrika 70 (1), 275-8.*

DENNIS, J.E. and SCHNABEL, R.B. (1983): *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, Englewood Cliffs.

HARVEY, A.C. and PHILLIPS, G.D.A. (1979): Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika 66 (1), 49-58.*

HILLMER, S.C. and TIAO, G.C. (1979): Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association 74 (367), 652-660.*

LJUNG, G. and BOX, G. (1979): The likelihood function of stationary autoregressive- moving average models. *Biometrika 66 (2), 265-70.*

MA, C. (1997): On the exact likelihood function of a multivariate autoregressive-moving average model. *Biometrika 84 (4), 957-964.*

MAURICIO, J.A. (1995): Exact maximum likelihood estimation of stationary vector arma models. *Journal of the American Statistical Association 90 (429), 282-291.*

MAURICIO, J.A. (2002): An algorithm for the exact likelihood of a stationary vector autoregressive-moving average model. *Journal of Time Series Analysis 23 (4), 473-86.*

MITTNIK, S. (1990): Computation of theoretical autocovariance matrices of multivariate autoregressive moving average time series. *Journal of the Royal Statistical Society. Series B (Methodological) 52 (1), 151-155.*

NEWBOLD, P. (1974): The exact likelihood function for a mixed autoregressive-moving average process. *Biometrika 61 (3), 423-426.*

NICHOLLS, D. and HALL, A. (1979): The exact likelihood function of multivariate autoregressive-moving average models. *Biometrika 66 (2), 259-64.*

PEARLMAN, J. G. (1980): An algorithm for the exact likelihood of a high- order autoregressive-moving average process. *Biometrika 67 (1), 232-233.*

REINSEL, G. (1995): *Elements of multivariate time series analysis*, 2nd Edition. Springer-Verlag, New Jersey.

SHEA, B. (1987): Estimation of multivariate time series. *Journal of Time Series Analysis 8, 95-109.*

A Test for Seasonal Fractional Integration

Uwe Hassler[1], Paulo M.M. Rodrigues[2], and Antonio Rubia[3]

[1] Goethe University Frankfurt, Germany *hassler@wiwi.uni-frankfurt.de*
[2] University of Algarve, Portugal *prodrig@ualg.pt*
[3] University of Alicante, Spain *antonio.rubia@ua.es*

Abstract. In this paper we propose regression based test statistics to detect seasonal fractional integration. Our setting extends Robinson's (1994) approach to the time domain and generalizes the procedures in Agiakloglou and Newbold (1994), Tanaka (1999) and Breitung and Hassler (2002) by allowing for fractional integration at the zero and seasonal frequencies. Our testing procedure can be easily implemented in practical settings and is flexible enough to account for a broad family of long- and short-memory specifications, including ARMA-type and/or GARCH-type dynamics, among others. Furthermore, it has power against different types of alternative hypotheses and inference is conducted under critical values drawn from a standard chi-squared distribution, independently of the long-memory parameters.

Keywords: nonstationarity, fractional integration, seasonality

1 Introduction

Seasonality is an important characteristic of many economic time series. In this paper, we discuss a statistical procedure wich is able to test the suitability of a given order of fractional integration at the seasonal frequencies. For seasonal data, several studies have investigated long-memory dynamics and its properties, and several tests have been proposed; see, among others, Hassler (1994), Robinson (1994), Ooms (1995), and Gil-Alana and Robinson (2001). The procedure presented in this paper generalizes the regression based approach of Breitung and Hassler (2002) and Hassler, Rodrigues and Rubia (2008) to the seasonal context. Furthermore, we allow for different types of errors in the data generating process (DGP) which include martingale differences sequences (MDS) and weakly correlated errors, thus allowing for ARMA and/or time varying volatility patterns. As in the frequency-domain case, the tests do not require formal knowledge of the true values of the fractionally-integrated coefficients. These are mainly intended for formally pretesting hypotheses about the extent of persistence, and to construct confidence sets that include the true values of the long-memory coefficients with a certain asymptotic coverage level. This is valuable for descriptive inference and, furthermore, provides reliable values for initiating optimization routines upon which several estimation procedures, such as (quasi) maximum likelihood procedures, build on.

The remaining of the paper is organized as follows. Section 2 introduces the general setting, discusses the set of sufficient conditions for the regression based tests and presents the regression-based tests, the relevant test statistics, and their asymptotic distributions. Section 3 analyzes the finite-sample performance of the tests by means of Monte Carlo experimentation. Section 4 summarizes the main conclusions.

2 The seasonal fractionally integrated model

Consider the pure seasonal model,

$$\Delta_\gamma (L;\delta)\, x_t = \varepsilon_t, \qquad t = 1, ..., T \tag{1}$$

where $x_t = 0$ for $t \leq 0$, $n = [S/2] + 1$, $[.]$ denotes the integer part of its argument and for even S the corresponding filter becomes

$$\Delta_\gamma (L;\delta) = (1-L)^{\delta_1} \left[\prod_{i=2}^{S/2} (1 - 2cos(\gamma_i)L + L^2)^{\delta_i} \right] (1+L)^{\delta_n}$$

with dimension $n = S/2 + 1$ and $\gamma_i = 2\pi(i-1)/S$, $i = 2, \ldots, [S/2]$. When S is odd, the component $(1+L)^{\delta_n}$, which corresponds to a cycle of two periods, is omitted. The properties of ε_t will be discussed below. A special case is when $\delta_1 = ... = \delta_n = 1$, originating a seasonal integrated process, i.e., $\left(1 - L^S\right) x_t = \varepsilon_t$. By allowing non-integer values in $\delta = (\delta_1, ..., \delta_n)'$, x_t is said to be generated by a seasonal fractionally integrated process of order δ; see, among others, Hassler (1994) and references therein.

For empirical purposes, the main interest lies in testing whether $\delta = \mathbf{d}$, with $\mathbf{d} \in R^n$ being specified *a priori*, against the alternative for which the order of integration is $\mathbf{d} + \theta$, $\theta \neq \mathbf{0}$. Thus, the hypothesis of interest is generally stated as

$$\mathrm{H}_0 : \delta = \mathbf{d}, \text{ or } \mathrm{H}_0 : \theta = \mathbf{0}, \tag{2}$$

against the alternative hypothesis that H_0 is false, *i.e.*, $\mathrm{H}_1 : \delta \neq \mathbf{d}$ or $\mathrm{H}_1 : \theta \neq \mathbf{0}$.

2.1 Preliminaries

We start our theoretical analysis by introducing and discussing the initial set of assumptions, general notational issues and definitions which are necessary for the regression-based tests.

Assumption $\mathcal{A}$:

i) The observable process $\{x_t,\ t = 1, ..., T\}$ *is generated by* $\Delta_\gamma(L; \mathbf{d})\, x_t = \varepsilon_t$, *with* $x_t = 0$ for $t \leq 0$, $\Delta_\gamma(L; \mathbf{d})$ *defined in* (1) *and* $\mathbf{d}$ *being a possibly non-integer vector in* R^n, $n \geq 1$.

(ii) The innovation process satisfies $a(L)\varepsilon_t = v_t$, *where* $a(L) = 1 - \sum_j^p a_j L^j$, $p \geq 0$, *such that* $a(z)$ *has all its roots outside the unit circle.*

(iii) The innovation process $\{v_t, \mathcal{F}_t\}$, $\mathcal{F}_t = \sigma(v_j : j \leq t)$, *is a stationary and ergodic MDS,* $E(v_t^2) = \sigma^2$, and $\{v_t\}$ *obeys one of the following restrictions:*

iii.a) $\{v_t\}$ *is independent and identically distributed and* $E(|v_t^4|^{1+r})$ *absolutely bounded for some* $r > 0$.

iii.b) $\{v_t\}$ *is strictly stationary and ergodic with*

$$\sum_{l_1=-\infty}^{\infty} \sum_{l_2=-\infty}^{\infty} ... \sum_{l_7=-\infty}^{\infty} |\kappa_v(0, l_1, ..., l_7)| < \infty,$$

where $\kappa_v(0, l_1, ..., l_7)$ *is the eigth-order joint cumulant of* $\{v_t\}$.

Condition *i)* is general enough to cover both types (type I and type II) of fractional processes considered in the literature. This assumption has become standard in the fractional unit root literature, because it may permit the observable processes to be well-defined in the mean-square sense regardless of the values of $\mathbf{d}$. In the context of the present paper, however, it does not play a major role and the relevant results hold both if we consider $\{\varepsilon_t\}_{t\geq 1}$ or $\{\varepsilon_t\}_{-\infty}^{\infty}$. Condition *ii)* allows for short-run dynamics. Condition *iii.a)* can be weakened by requiring that, conditional on the σ-field of events $\mathcal{G}_t$, moments up to the fourth-order (and suitable cross-products of elements of ε_t) equal the corresponding unconditional moments, so that essentially $\{\varepsilon_t\}$ is only required to behave as an i.i.d process up to the fourth-order moment. The main purpose of *iii.b)* is to allow for time-varying conditional volatility patterns in $\{\varepsilon_t\}$. This requires additional restrictions limiting the extent of temporal dependence, which are provided by restricting the absolute summability of the eight-order joint cumulants; see also Gonçalves and Kilian (2007) and Demetrescu, Kuzin and Hassler (2008). Finally, normality is not required to derive the asymptotic theory, however efficiency in Gaussian-score based procedures will only be attainable under that restriction.

Before presenting the regression-based test statistics, consider the following relevant definitions:

Definition 2.1. *For all* $j \geq 1$ *and* $\gamma \in [0, \pi]$, *define the non-stochastic weighting process* $\omega_j(\gamma)$ *as follows,*

$$\omega_j(\gamma) = \begin{cases} 1/j, & if\ \gamma = 0 \\ 2j^{-1}\cos(j\gamma), & if\ \gamma \in (0, \pi) \\ (-1)^j/j, & if\ \gamma = \pi \end{cases}. \tag{3}$$

Similarly, for $\gamma = (\gamma_1, ..., \gamma_n)'$ *such that* $\gamma_s = 2\pi(s-1)/S$, $s = 1, ..., n$, *define*

$$\omega_j(\gamma) = (\omega_j(\gamma_1), ..., \omega_j(\gamma_n))'. \tag{4}$$

Definition 2.2. *Given the real-valued stochastic process* $\{x_t, t \geq 1\}$ *and a vector* $\delta \in R^n$, *define the filtered series*

$$\varepsilon_{\delta,t} = \Delta_\gamma(L;\delta)\, x_t, \tag{5}$$

where, if $\delta = \mathbf{d}$, *then* $\Delta_\gamma(L;\mathbf{d})\, x_t = \varepsilon_t$ *and* $\varepsilon_{\mathbf{d},t} = \varepsilon_t$. *For any frequency* $\gamma_s \in [0, \pi]$, *define the following (truncated and non-truncated) stochastic processes which are constructed by linearly filtering* $\varepsilon_{\delta,t}$ *with the weighting processes given in Definition 2.1, i.e.,*

$$\varepsilon^*_{\gamma_s,t-1} = \sum_{j=1}^{t-1} \omega_j(\gamma_s)\, \varepsilon_{\delta,t-j}, \tag{6}$$

$$\varepsilon^{**}_{\gamma_s,t-1} = \sum_{j=1}^{\infty} \omega_j(\gamma_s)\, \varepsilon_{\delta,t-j}. \tag{7}$$

Definition 2.3. *Given* $\gamma = (\gamma_1, ..., \gamma_n)'$, *define the n-dimensional vectors*

$$\varepsilon^*_{\gamma,t-1} = \left(\varepsilon^*_{\gamma_1,t-1}, ..., \varepsilon^*_{\gamma_n,t-1}\right)' = \sum_{j=1}^{t-1} \omega_j(\gamma)\, \varepsilon_{\delta,t-j};$$

$$\varepsilon^{**}_{\gamma,t-1} = \left(\varepsilon^{**}_{\gamma_1,t-1}, ..., \varepsilon^{**}_{\gamma_n,t-1}\right)' = \sum_{j=1}^{\infty} \omega_j(\gamma)\, \varepsilon_{\delta,t-j}. \tag{8}$$

2.2 Regression-based tests for fractional integration

The regression based approach was pioneered by Agiakloglou and Newbold (1994) for the context of fractional unit roots at the zero frequency, and further developed in Breitung and Hassler (2002), Hassler and Breitung (2006), and Demetrescu *et al.*, (2008) in the same context. Regression-based tests are particularly advantageous for the empirically relevant case in which the data exhibit weak correlation. We extend this procedure to the seasonal case testing principle and present the relevant asymptotic distribution.

Conditions ii) and iii) allow for stationary AR(p) dynamics in the generating process, which may appear jointly with time-varying volatility patterns, such as GARCH or Stochastic Volatility errors, under the same set of restrictions as those in condition ii). The results are formally discussed for the case in which p is known. For practical purposes, the short-run dynamics of the underlying process may be characterized by a stationary and invertible linear

process $\varepsilon_t = \sum_{j=0}^{\infty} b_j v_{t-j}$ such that the AR(p) model, for some large enough $p < \infty$, approaches the underlying AR representation reasonably well. The effects on the finite-sample properties of the regression-based tests when the underlying correlation structure in the short-run dynamics is unknown shall be discussed in the Monte Carlo section.

The following proposition states the general testing strategy for generalized fractional integration in the regression framework:

Proposition 2.1 *Consider the basic auxiliary regression augmented with p lags of the dependent variable, i.e.,*

$$\varepsilon_{\mathbf{d},t} = \sum_{l=1}^{n} \phi_l \varepsilon^*_{\gamma l,t-1} + \left(\sum_{i=1}^{p} \zeta_i \varepsilon_{\mathbf{d},t-i} \right) + e_{tp}, \quad t = p+1, ..., T. \tag{9}$$

Then, the null hypothesis $H_0 : \theta = \mathbf{0}$ can be tested by addressing the joint significance of the estimated $\phi_l, l = 1, ..., p$, in (9).

Augmentation is standard in many testing procedures having the null of (fractional) integration; see for instance Dolado, Gonzalo and Mayoral (2002) and Breitung and Hassler (2002). Essentially, augmenting the auxiliary regression with lags of the dependent variable seeks to whiten the correlation structure of the regression residuals so that they can behave asymptotically as a MDS.

The following theorems present the asymptotic properties of the regression based test statistic for general fractional integration.

Theorem 2.1. *Let β_T be the (n+p) estimated vector of parameters in the pth order augmented auxiliary regression $\varepsilon_{\mathbf{d},t} = \beta' \mathbf{X}^*_{tp} + e_{tp}$, with $\mathbf{X}^*_{tp} = \left(\varepsilon'^*_{\gamma,t-1}, \varepsilon_{\mathbf{d},t-1}, ..., \varepsilon_{\mathbf{d},t-p}\right)'$, and let the (n+p) vector $\mu_0 = (0, ..., 0, a_1, ..., a_p)'$, with the a_i parameters corresponding to the autoregressive coefficients in $\left(1 - \sum_{i=1}^{p} a_i L^i\right) \varepsilon_t = v_t$. Then, under Assumption $\mathcal{A}$, the null hypothesis, and as $T \to \infty$,*

$$\sqrt{T}\left(\beta_T - \mu_0\right) \Rightarrow \mathcal{N}\left(\mathbf{0},\ \left(\mathbf{\Omega}^{**}_p\right)^{-1} \mathbf{\Lambda}_p \left(\mathbf{\Omega}^{**}_p\right)^{-1}\right) \tag{10}$$

where $\mathbf{\Omega}^{**}_p \equiv E\left(\mathbf{X}^{**}_{tp} \mathbf{X}'^{**}_{tp}\right)$, $\mathbf{\Lambda}_p \equiv E\left(v_t^2 \mathbf{X}^{**}_{tp} \mathbf{X}'^{**}_{tp}\right)$, and $\mathbf{X}^{**}_{tp} = \left(\varepsilon'^{**}_{\gamma,t-1}, \varepsilon_{\mathbf{d},t-1}, ..., \varepsilon_{\mathbf{d},t-p}\right)'$.

Proof. *See Hassler, Rodrigues and Rubia (2008).*

Theorem 2.2. *Let $\mathbf{R}$ be an $n \times (n+p)$ matrix such* that $[\mathbf{R}]_{ij} = 1$ *for all $i = j$ and zero otherwise. Consider the Wald-type test statistic on the estimates of the augmented auxiliary regression, i.e.,*

$$\Upsilon^{(n)}_{Wp} = [\mathbf{R}\beta_T]' \left[\frac{1}{T} \mathbf{R} \widehat{\mathbf{V}}_{\mathbf{T}} \mathbf{R}'\right]^{-1} [\mathbf{R}\beta_T] \tag{11}$$

with $\widehat{\mathbf{V}}_{\mathbf{T}}$ *being the sample estimation of the covariance matrix of* β_T *such that*

$$\widehat{\mathbf{V}}_{\mathbf{T}}\ /T = \left(\sum_{t=p+1}^{T} \mathbf{X}_{tp}^{*}\mathbf{X}_{tp}^{*\prime}\right)^{-1} \left(\sum_{t=p+1}^{T} \widehat{e}_{tp}^{2}\mathbf{X}_{tp}^{*}\mathbf{X}_{tp}^{*\prime}\right) \left(\sum_{t=p+1}^{T} \mathbf{X}_{tp}^{*}\mathbf{X}_{tp}^{*\prime}\right)^{-1},$$

where $\widehat{e}_{tp}$ *denotes the estimated residuals. Under the same conditions of Theorem 2.1,* $\Upsilon_{Wp}^{(n)} \Rightarrow \chi_{(n)}^{2}$.

Proof. *See Hassler, Rodrigues and Rubia (2008) for details.*

Remark: Throughout this paper, we have focused on the model $\Delta_{\gamma}\left(L;\mathbf{d}\right)\left(x_t-\mu_t\right)=\varepsilon_t\mathbb{I}_{(t>0)}$, restricting $\mu_t=0$. As commented in Breitung and Hassler (2002), the simplest way to deal with non-zero deterministic patterns, $\mu_t\neq 0$, is to demean/detrend x_t prior to computing the relevant tests statistics. This does not affect the limit distribution of the relevant statistics; see the discussion in Robinson (1994). Regarding augmentation of the test regression, following Demetrescu *et al.* (2008), we suggest the the rule of thumb proposed by Schwert (1989) which shows relatively good performance in finite-samples in this context. This rule sets $p=\left[c(T/100)^{1/4}\right]$, where c is a positive constant and $[\cdot]$ denotes the integer value of the argument.

3 Finite-sample analysis

In this section we address the empirical properties of the regression-based test statistic in finite samples. In our simulations we consider a sample of 400 observations and use 5000 replications in the computation of the size and power of the test statistics. The data generation process used is,

$$(1-L)(1+L)(1+L^2)x_t=(1-L)^{\theta_1}(1+L)^{\theta_2}(1+L^2)^{\theta_3}\varepsilon_t$$

where $\varepsilon_t\sim nid(0,\sigma^2)$.

The rejection frequencies of the $\Upsilon_{Wp}^{(n)}$ test for a nominal significance level of 5% and sample size of $T=400$ are shown in Table 1. The null hypothesis considered is seasonal integration and the alternatives used in the power evaluation allow for fractional integration at the zero, nyquist and harmonic frequencies. A test regression as in (9) was used to compute the tests and we considered p=0, 1 and 4 lags.

We observe from Table 1 that the size of the test ($\theta_1=\theta_2=\theta_3=0$), independently of the augmentation considered in the analysis, is well-behaved, i.e. empirical size is very close to the nominal size considered (5%). Table 1 also shows very good power performance of the proposed test procedures

(particularly for $p = 0$). The inclusion of unnecessary regressors (the augmentation used in the present context) has negative effects over the procedure. A power reduction is clearly noticeable as the number of the lags used in the augmentation increases. Moreover, in simulations not reported in this paper, but available from the authors, the power performance of the procedures clearly improves as the sample size increases, as expected. Notice that this is an important point given that these procedures are typically applied on relatively large samples.

We have also experimented with different combinations of $\theta_i \neq 0$, $i = 1, 2, 3$, but the results were qualitatively similar to those of Table 1 and have therefore been omitted.

Table 1: Frequency of rejection of the $\Upsilon_{Wp}^{(n)}$ test. The sample size considered is T=400.

θ_1	θ_2	θ_3	*p=0*	*p=1*	*p=4*
0.00	0.00	0.00	.051	.048	.049
-0.30	0.00	0.00	1.00	.906	.404
-0.20	0.00	0.00	.990	.669	.238
-0.10	0.00	0.00	.564	.216	.091
0.10	0.00	0.00	.502	.216	.097
0.20	0.00	0.00	.987	.726	.264
0.30	0.00	0.00	1.00	.986	.588
0.00	-0.30	0.00	1.00	.992	.246
0.00	-0.20	0.00	.966	.663	.076
0.00	-0.10	0.00	.420	.180	.061
0.00	0.10	0.00	.397	.100	.073
0.00	0.20	0.00	.947	.264	.141
0.00	0.30	0.00	1.00	.427	.162
0.00	0.00	-0.30	1.00	.998	.742
0.00	0.00	-0.20	.999	.979	.737
0.00	0.00	-0.10	.871	.719	.477
0.00	0.00	0.10	.247	.125	.072
0.00	0.00	0.20	.292	.221	.302
0.00	0.00	0.30	.388	.380	.519

Note: θ_i, i=1,2,3, refers to the value considered to generate fractional noise at a particular frequency, and p indicates the order of augmentation used in the test regression.

4 Conclusion

This paper presents regression-based test statistics in the time-domain that allow testing for fractionally-integrated patterns against fractional or integer integration in seasonal models. The tests can be computed from simple least-squares regressions, and are asymptotically equivalent to the frequency-domain tests of Robinson (1994) and the likelihood-based tests in Nielsen (2004), for which the relevant critical values are obtained from a χ^2 distribution with as many degrees of freedom as the restrictions being tested, regardless of the order of integration. Augmented versions of these tests are asymptotically robust against weakly-dependent errors following unknown patterns under quite general conditions, and exhibit good statistical performance in samples of moderate size. This makes the general regression-based testing strategy discussed in this paper a valuable tool for addressing preliminary data analysis in which parsimonious yet potentially restrictive hypothesis related to the order of integration of the data may be formally validated or refuted.

Acknowledgments

Financial support from POCTI/ FEDER (grant ref. PTDC/ECO/64595/2006) and the SEJ2005-09372/ECON project is gratefully acknowledged.

References

AGIAKLOGLOU, C. and NEWBOLD, P. (1994): Lagrange multiplier tests for fractional difference. *Journal of Time Series Analysis* 14, 253–262.

BREITUNG, J. and HASSLER, U. (2002): Inference on the cointegration rank in fractionally integrated processes. *Journal of Econometrics* 110(2), 167-185.

DEMETRESCU, M., KUZIN, V., HASSLER, U. (2008): Long Memory Testing in the Time Domain. *Econometric Theory*, 24(1), 176-215.

DOLADO, J.J., GONZALO, J. and MAYORAL, L. (2002): A Fractional Dickey-Fuller Test for Unit Roots, *Econometrica* 70(5), 1963-2006.

GIL-ALANA, L.A. and ROBINSON, P.M. (2001): Testing of Seasonal Fractional Integration in UK and Japanese Consumption and Income. *Journal of Applied Econometrics* 16, 95-114.

GONÇALVES, S. and KILIAN, L. (2007): Asymptotic and Bootstrap Inference for AR(∞) Processes with Conditional Heteroskedasticity. forthcoming *Econometric Reviews*.

HASSLER, U., BREITUNG, J. (2006): A Residual-Based LM Type Test Against Fractional Cointegration. *Econometric Theory* 22, 1091-1111.

HASSLER, U. (1994): (Mis)specification of Long Memory in Seasonal Time Series. *Journal of Time Series Analysis* 15, 19–30.

HASSLER, U., RODRIGUES, P.M.M. and RUBIA, A. (2008): Testing for the General Fractional Unit Root Hypothesis in the Time Domain, Funcas Working Paper, n 380.

NIELSEN, M.Ø. (2004): Efficient likelihood inference in nonstationary univariate models. *Econometric Theory* 20, 116-146.

OOMS, M. (1995): Flexible seasonal long memory and economic time series, Working Paper, Erasmus University Rotterdam.

ROBINSON, P.M. (1994): Efficient tests of nonstationary hypotheses. *Journal of the American Statistical Association* 89(428), 1420-1437.

SCHWERT, G.W. (1989): Tests for Unit Roots: A Monte Carlo Investigation. *Journal of Business and Economic Statistics* 7(2), 147-59.

TANAKA, K. (1999): The Nonstationary Fractional Unit Root. *Econometric Theory* 15, 549 - 582.

Time Series Analysis Using Local Standard Fractal Dimension -Application to Fluctuations in Seawater Temperature-

Kenichi Kamijo[1] and Akiko Yamanouchi[2]

[1] Graduate School of Life Sciences, Toyo University,
1-1-1 Izumino, Itakura, Gunma, 374-0193, Japan, *kamijo@toyonet.toyo.ac.jp*
[2] Izu Oceanics Research Institute,
3-12-23 Nishiochiai, Shinjuku, Tokyo, 161-0031, Japan,
gx0400018@toyonet.toyo.ac.jp

Abstract. The local standard fractal dimension (LSFD) and the local standard contribution ratio (LSCR) as new indices in discrete time series have been proposed in this paper. That is, the standardization of the local fractal dimension and local contribution ratio has been evaluated by the standard deviation as the absolute amount in a given short time series. Applied to the difference time series between altitudinal seawater temperatures, it can be seen that moving LSFD and LSCR progress at almost the same rates. As applications to random processes, the various interesting features can be observed, and the possibility of the classification for general time series is suggested using these results.

Keywords: local standard fractal dimension, local standard contribution ratio, standard deviation, difference time series between altitudinal seawater temperatures

1 Introduction

We have already proposed the use of local fractal dimension (LFD) and local contribution ratio (LCR) in discrete dynamical systems, and several examples of applications of these have been presented (Kamijo and Kato (2004), Kamijo and Yamanouchi(2005), Kamijo and Yamanouchi(2007)). Based on our experience of having applied these indices to various time series, when interpreting these indices as being prognostic indices for predicting the occurrence of a certain physical phenomenon, we consider that it is necessary to not only redefine the change patterns of these time series, but also the concept of absolute amount. Specifically, when a certain physical quantity is to be observed continuously with a view to predicting sudden changes or phase transition phenomena of that physical quantity, abnormalities in LFD and LCR must be detected early; in addition, the absolute amount of physical

quantities must be considered at the same time. In cases when there are extremely small changes in the physical quantity, and, by contrast, when there are large fluctuations, if the change patterns are the same, LFD and LCR will have exactly the same values. Consequently, they are no longer able to reflect the difference in the size of the absolute amount of change, which is unreasonable. For this reason, we decided to adopt a type of standardization for LFD and LCR in this study that we previously proposed to solve this kind of problem. These indices are denoted by LSFD and LSCR, and they can standardize the degree of complexity in time series fluctuations using the standard deviation as an absolute amount. As an example of this, we applied LSFD and LSCR to the difference time series between altitudinal seawater temperatures in the area around the Izu Peninsula, Japan, and found that moving LSFD and moving LSCR progress at almost the same rates. Different characteristics were observed in simulations based on other random number series. It has been suggested that discrete dynamical systems can be categorized if these analytical results are used positively. This will open up applications in process control and statistical quality control in the future.

2 Local standard fractal dimension and local standard contribution ratio

2.1 LFD

For a discrete time series, which can be considered as an objective vector,

$$\boldsymbol{x}_k = \{x_k, x_{k+1}, \cdots, x_{k+L-1}\}, \tag{1}$$

the accumulated change $N(r, k, L)$ can be defined as

$$N(r,k,L) = \frac{1}{r}\sum_{i=1}^{r}\sum_{j=0}^{[\frac{L}{r}]-2}|x_{k+jr+r+i-1} - x_{k+jr+i-1}| \tag{2}$$

where
L: length of the discrete time series, r: sampling interval.
The accumulated change $N(r, k, L)$ can also be redefined as

$$N(r,k,L) = Ar^{-D_k} \tag{3}$$

where
D_k: k-th local fractal dimension, A: proportion constant.

Therefore,

$$\log N(r,k,L) = -D_k \log r + \log A \tag{4}$$

Here $N(r,k,L)$ can be replaced by Y, and also $\log r$ by X, then we have

$$Y = -D_k X + \log A \tag{5}$$

Then, D_k can be obtained as the local fractal dimension based on a regression analysis, and in this paper we will refer to LFD_k instead of D_k, as the k-th local fractal dimension.

That is, D_k is the so-called regression coefficient, and can be obtained by the following equation.

$$D_k = -\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} \equiv LFD_k \tag{6}$$

A method for obtaining the LFD from the information on six plots has already been proposed as the Six-Point Evaluation Method (Kamijo and Kato(2004), Kamijo and Yamanouchi(2005), Kamijo and Yamanouchi(2007)), where $N(r,k,L)$ can be obtained varying r from 1 to 6 in fixed condition of L=30.
In case of r=6, L=30, we have 6 pairs for (X,Y) on the so-called X-Y plane.

2.2 LCR

In this paper, the "degree-of-freedom-adjusted contribution ratio" that indicates how well the regression line fits, is called the local contribution ratio (LCR). This LCR is related to the fractality associated with the "manner of change" of the said time series, and indicates the extent of fractal strength. In the case of a perfect mathematical fractal, the LCR is 1, while, alternately, the LCR becomes a value close to 0 when there is absolutely no fractality. The LCR is defined as follows using each variance (mean square; MS) in the so-called variance analysis table for the regression analysis.

$$LCR = 1 - \frac{V_e}{V_T} \tag{7}$$

where
V_e: error variance, V_T: total variance.

Then we obtain

$$0 \le LCR \le 1 \tag{8}$$

2.3 Moving measurement

The finite time series $\{x_k, x_{k+1}, \cdots, x_{k+L-1}\}$ can be extracted from the infinite time series $\{x_0, x_1, x_2, \cdots, x_k, x_{k+1}, \cdots, x_{k+L-1}, \cdots\}$ for calculation with the length L, and the k-th standard deviation $STDEV_k$ can be obtained by the following equation.

$$STDEV_k = \sqrt{\frac{\sum_{i=k}^{k+L-1} (x_i - \overline{x}_k)^2}{L-1}} \tag{9}$$

where

$$\overline{x}_k = \frac{\sum_{i=k}^{k+L-1} x_i}{L} \tag{10}$$

Next, k-th local standard fractal dimension $LSFD_k$ and k-th local standard contribution ratio $LSCR_k$ can be defined as follows:

$$LSFD_k = \frac{LFD_k}{STDEV_k} \tag{11}$$

$$LSCR_k = \frac{LCR_k}{STDEV_k} \tag{12}$$

Then we can have the discrete time series for moving $LSFD$ and moving $LSCR$ by increasing the suffix k one by one.

2.4 Explanation

In discrete dynamical systems, observed values are plotted in time sequences. Accordingly, this time sequence is most important. Generally, when observing physical quantities, the values observed at a certain time are considered to be influenced in some manner by the values of an earlier time, and often an apparently similar fluctuation pattern is demonstrated as the fluctuation pattern of that dynamical system even if the time scale or sampling interval is changed. In other words, appropriate irregular fluctuations are generally considered as having a self-similar or fractal structure. It is known that in a given finite time series, although the value of its standard deviation (STDEV) does not vary even when its time sequence is changed, the values of LFD and LCR, by contrast, are strongly dependent on its fluctuation pattern and complexity.

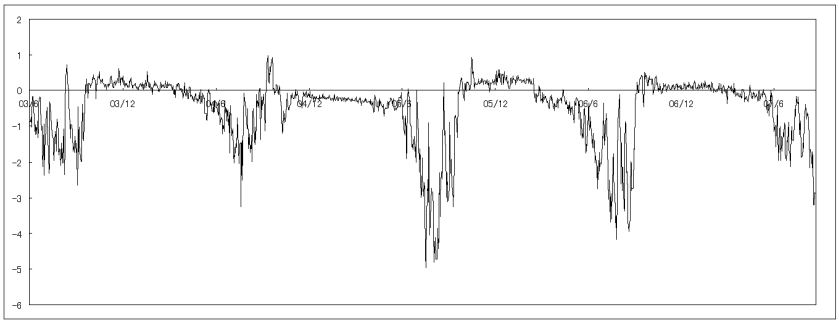


Fig. 1. Difference time series between altitudinal seawater temperatures(15-5m; 2003/6/5-2007/8/24; Usami, Ito, Japan).

3 Application to difference time series between altitudinal seawater temperatures

The meteorological characteristics of the near-surface layer were obtained at a point off the Izu Peninsula (Usami, Ito, Japan) by placing temperature data loggers at three depths: 5 m (upper layer), 10 m (middle layer) and 15 m (lower layer). The loggers were set to continuously measure seawater temperatures at 1-hour intervals. The discrete time series data was averaged on a daily basis and the differences in time series data for seawater temperatures between water depths 15 m and 5 m were compiled. We used seawater temperature difference time series as an example of a discrete dynamical system. Fig. 1 shows the time series used in this analysis. From this figure, negative values are observed in summer and they fluctuate greatly, while in winter, positive values are seen, and they exhibit relatively stable change.

Concerning the time series in Fig. 1, Fig. 2 shows the results of calculating LSFD and LSCR using the method proposed in this paper. Fig. 2 demonstrates that LSFD and LSCR exhibit almost the same kind of change pattern in the case of a discrete dynamical orbit.

Nevertheless, as shown in Fig. 3, the changes in LFD, LCR and STDEV, from which LSFD and LSCR are calculated, are completely different to those shown in Fig. 2. It is very interesting that this phenomenon is observed in spite of LFD and LCR being divided by the corresponding STDEV.

Next, Fig. 4 shows the results of having applied the above-mentioned method to a uniform random number series as an example of a stochastic process. In the case of a random time series, the same kinds of changes are observed when LSFD and LSCR are viewed in a macro manner when a large scale is used. However, considerable deviation is observed when the scale is reduced. As shown in Fig. 5, changes in LFD, LCR and STDEV, from which LSFD and LSCR are calculated, appear completely different from those in Fig. 4. This is an interesting phenomenon, resembling that shown in Fig. 3

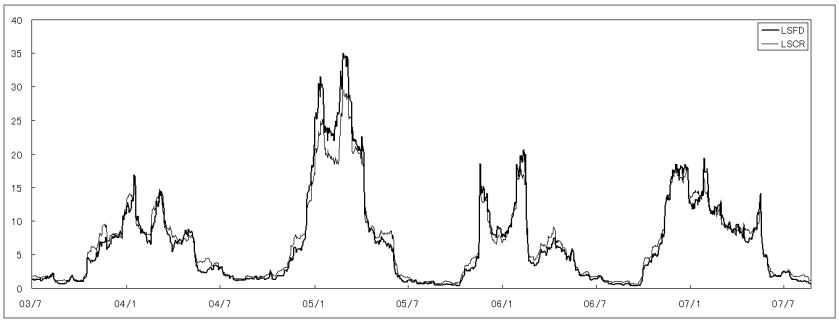


Fig. 2. LSFD and LSCR changes in case of difference time series between altitudinal seawater temperatures; L=30.

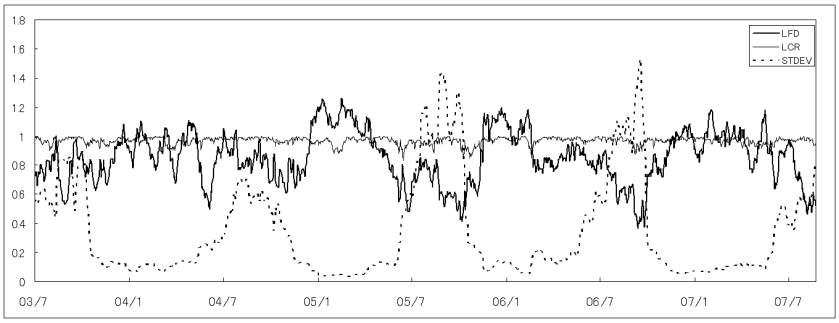


Fig. 3. LFD, LCR and STDEV changes in case of difference time series between altitudinal seawater temperatures; L=30.

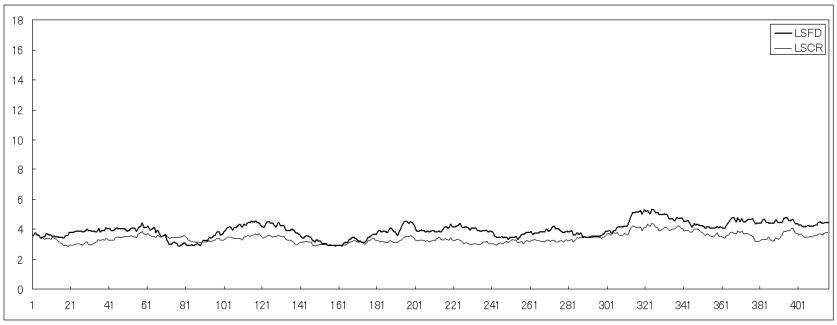


Fig. 4. LSFD and LSCR changes in case of random process; L=30.

that shows in the case for fluctuations in the difference of altitudinal seawater temperature.

Furthermore, when the original time series is completely linear, it is known that LFD is close to 0 and that LCR decreases to about 80% even if the initial gradient is changed. However, when the gradient of the line is increased,

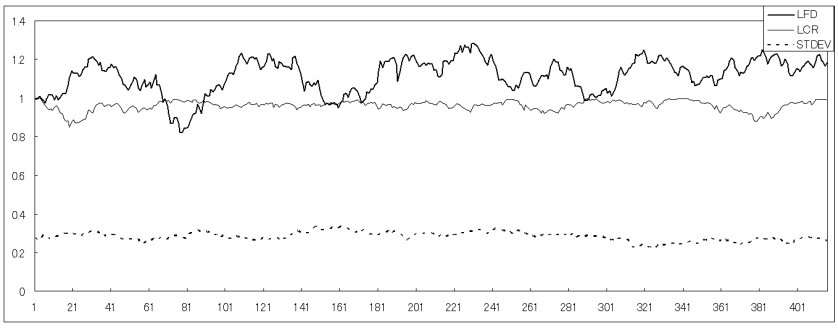


Fig. 5. LFD, LCR and STDEV changes in case of random process; L=30.

STDEV naturally increases, so that LSFD and LSCR decrease with this increase, and then progress as fixed values respectively. It is considered that the LSFD and LSCR have the same values if the unit of the graph scale is getting larger.

4 Considerations

When LSFD and LSCR are applied to fluctuations in the difference in altitudinal seawater temperatures, LSFD and LSCR have almost the same values and their change patterns are almost the same. Upon further investigation, the correlation coefficient between LSFD and LSCR is relatively high. In contrast to this, a strange phenomenon of the correlation coefficient between LFD and LCR, from which the LSFD and LSCR are calculated, being unexpectedly low is observed.

Table 1 shows a correlation coefficiet matrices in the seawater temperature difference time series and the uniform random number series, which are related to LSFD and LSCR. The correlation coefficient between LFD and LCR in both time series is small. While the values of the correlation coefficient between LSFD and LSCR are high in both time series, the coefficient value of the seawater temperature difference time series is much closer to 1. LSFD and LSCR are considered to be changing while taking on almost the same values, and the indices in the uniform random number series indicate a smaller difference in the deviation in the graphs. Although the values of LSFD and LSCR appear to be closer to each other in the seawater temperature difference time series, it is considered that it appears same way if the values of LSFD and LSCR are compared using the same maximum graph scale.

Furthermore, when the seawater temperature difference time series is viewed on a macro scale, we can determine that the LSFD is high in winter and in contrast to this, low in summer.

SeaWater	LFD	LCR	STDEV	LSFD	LSCR
LFD	1				
LCR	0.1240	1			
STDEV	-0.6554	-0.0626	1		
LSFD	0.7536	0.0152	-0.6412	1	
LSCR	0.7131	0.0488	-0.7004	0.9867	1

Random	LFD	LCR	STDEV	LSFD	LSCR
LFD	1				
LCR	-0.2709	1			
STDEV	-0.2525	-0.0183	1		
LSFD	0.7635	-0.1333	-0.8096	1	
LSCR	0.1618	0.3278	-0.9449	0.7337	1

Table 1. Correlation coefficient matrices in the seawater temperature difference time series and the uniform random number series.

5 Conclusions and future outlook

As we have discussed above, in this paper we have considered one kind of standardization for LFD and LCR. Namely, when a time series with an appropriate length is given, LFD and LCR serve as indicators as to how many times the standard deviation should be multiplied; this is based on the similar approach as the SN ratio. If these indices are denoted by LSFD and LSCR, then these indices are considered as being able to standardize the degree of complexity in general time series fluctuations by the absolute amount. As one example of a discrete dynamical system, this approach was applied to a difference time series between altitudinal seawater temperatures. In this case, it was found that LSFD and LSCR change by almost the same amounts. In simulation tests based on uniform random number series as a stochastic process, various characteristics that differ from those of dynamical systems were observed. The possibility that these analytical results are useful for categorizing the characteristics of time series has been suggested if these phenomena are positively used, and will open up applications in process control and statistical quality control in the future.

References

KAMIJO, K. and KATO, T. (2004): Sliding Measurement for Local Fractal Dimension in Discrete Time Series and its Applications, *2004 Hawaii International Conference on Statistics, Mathematics and Related Fields*, Honolulu, U.S.A.

KAMIJO, K. and YAMANOUCHI, A. (2005): Sliding Measurement for Signal Processing Using Local Fractal Dimension -An Example of Fluctuations in Sea-

water Temperature as Discrete Dynamical Systems-, *Fractals in Engineering 2005*, Tours, France.
KAMIJO, K. and YAMANOUCHI, A. (2007): Signal Processing Using Fuzzy Fractal Dimension and Grade of Fractality -Application to Fluctuations in Seawater Temperature-, *2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing*, Honolulu, U.S.A.

Spectral Homogeneity for a Set of Time Series

Inmaculada Luengo[1], Pedro Saavedra[2], and Carmen N. Hernández[2]

[1] Departamento de Informática y Sistemas
Universidad de Las Palmas de Gran Canaria
35017 Las Palmas, Spain, *mluengo@dis.ulpgc.es*
[2] Departamento de Matemática Aplicada
Universidad de Las Palmas de Gran Canaria
35017 Las Palmas, Spain, *saavedra@dma.ulpgc.es*, *cflores@dma.ulpgc.es*

Abstract. In this paper we propose a test of spectral homogeneity to decide whether a collection of time series are generated from the same pattern. In particular, we test if the variance of the log of the spectral density is zero. We propose a test statistic whose probability distribution under the null hypothesis is approximated by bootstrap. The consitency of the overall procedure is analyzed through Mallows metrics. We illustrate our results through a simulation (2000 repetitions) of a MA(2) with random coefficients.

Keywords: replicated time series, spectral analysis, homogeneity test, bootstrap, Mallows metric

1 Introduction

In recent years various papers have been written on the spectral analysis of replicated time series from a biomedical point of view. Diggle and Al-Wasel (1993) introduce a population pattern known as the population spectrum inspired in an asymptotic representation of the periodogram of Gaussian linear processes. They consider a set of time series evaluated at the same times, where each one is the realization of a specific stationary process with an absolutely continuous spectral distribution. In order to analyse these data they consider a random effects model which assumes that the spectral density function of each individual process is the trajectory of a stochastic process, whose average function is known as the population spectrum. Hernández et al. (1999) estimate this parameter in more general conditions and they approximate the probability distribution of the estimate using bootstrap. Saavedra et al (2000) develop the theory of doubly stochastic linear stationary processes and estimate the population spectrum by means of smoothing of the average periodogram. Luengo et al. (2006) propose an additive model for the logperiodogram of replicated time series, with each one corresponding to a specific object of the population.

In this paper we test the spectral homogeneity of a set of time series. To do this, we consider the additive random effects model and the estimates for

both the population component and the individual components, propossed by Luengo-Merino et al. (2006), also based on the asymptotic representation of the general linear processes. The distribution of the proposed statistical test under the null hypothesis is approximated using the bootstrap thechnique. A simulation study is carried out and the power of the test for a specific set of alternatives is evaluated.

2 Model for the set of time series

2.1 Definition of the model

We will use the exact same model defined by Luengo-Merino et al. (2006). Let $(B, \mathbf{B}, P_B)$ be a probabilistic space so that each object $b \in B$ has a general linear process $\{X(b,t) : t \in \mathbf{Z}\}$ associated. Let be

$$\{X_i(t) : i = 1, \cdots, r;\ t = 1, \cdots, N\} \tag{1}$$

a set of time series evaluated on a random sample of r objects $b_1, \ldots, b_r$ from B and observed at the same N times. The periodogram of each series at the j-th Fourier frequency $\omega_j = 2\pi j/N$, for $j = 1, \cdots, \nu = [N/2] - 1$ is defined as:

$$I_i(\omega_j) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} X_i(t)\, e^{-i\omega_j t} \right|^2 \tag{2}$$

Let $C_0 = -0.5771$ be the Euler constant and $y_{ij} = \log I_i(\omega_j) - C_0$.

We suppose that, for any objects $b_1, \ldots, b_r \in B$ and for any times $t = 1, \ldots, N$, y_{ij} can be written:

$$y_{ij} = \mu(\omega_j) + Z_i(\omega_j) + e_{ij} \tag{3}$$

for $i = 1, \cdots, r$; $j = 1, \cdots, \nu = [N/2] - 1$, where:

(i) $\mu(\omega)$ for $|\omega| \leq \pi$ represents a underlying pattern in the population.
(ii) $\{Z_i(\omega); |\omega| \leq \pi\}$, $i = 1, \cdots, r$ involves the specificity of the i-th object in the population and are r independent trajectories of a stochastic process $\{Z(\omega)\}$ such that $E[Z(\omega_j)] = 0$ for $j = 1, \ldots, \nu$. We will write $\operatorname{cov}[Z(\omega), Z(\theta)] = \Psi_Z(\omega, \theta)$, for $|\omega|, |\theta| \leq \pi$.
(iii) $\{e_{ij}\}$ are independent and identically distributed on i random variables, with $E[e_{ij}] = 0$ and $E\left[e_{ij}^2\right] = \sigma_e^2 < \infty$ for $j = 1, \ldots, \nu$.

An apparently restrictive aspect of the model is the use of the Euler constant C_0 in the transformation of the periodograms. Nevertheless, as we will see in a subsequent section, the statistical test, the aim of this paper, is independent of the value of C_0.

Remark. The proposed model is approximately satisfied by sets of time series such as:

i) For each $i = 1, \ldots, r$, $\{X_i(t), t \in \mathbf{Z}\}$, is a trajectory of a general linear process with absolutely continuous spectral distribution, being the corresponding spectral density $\{Q_i(\omega), |\omega| \leq \pi\}$.
ii) The process $\{X_i(t), t \in \mathbf{Z}\}$ verify the conditions of theorem 6.2.2 in Priestley (1981, pgs 424-425)

Under these conditions the periodogram of each trajectory $\{X_i(t)\}$ satisfies:

$$I_i^{(N)}(\omega_j) = Q_i(\omega_j) U_{ij}^{(N)} + R_{i,j}^{(N)} \tag{4}$$

where for each $i = 1, \cdots, r$, $R_{i,j}^{(N)}$ denotes a term that is asymptotically negligible, and $U_{ij}^{(N)}$ are asymptotically independent random variables having the standard exponential distribution (this distribution is exact if the processes are Gaussian). Thus, we can consider $E\left[\log U_{ij}^{(N)}\right] \approx C_0$, the above mentioned Euler constant. Neglecting the term $R_{i,N}(\omega_j)$ (as in Franke and Härdle (1992)) and making the necessary transformations in (4), we obtain:

$$Y_{ij} = \mu(\omega_j) + Z_i(\omega_j) + e_{ij} \text{ for } i = 1, \cdots, r;\ j = 1, \cdots, \nu$$

where

$$\mu(\omega_j) = E[\log Q_i(\omega_j)]$$
$$Z_i(\omega_j) = \log Q_i(\omega_j) - E[\log Q_i(\omega_j)]$$
$$e_{ij} = \log U_{ij}(\omega_j) - C_0$$

It is obvious that $E[Z_i(\omega_j)] = 0$, for all i and $|\omega| \leq \pi$ and that the random variables e_{ij} are independent on i and approximately identically distributed on j and $E[e_{ij}] = 0$.

2.2 Estimation of the parameters

In this section we remember the kernel estimates for the population parameter $\mu(\omega)$ and for each of the individual trajectories $Z_i(\omega)$ defined in equation (3) proposed by Luengo-Merino et al. (2006).
Consider a kernel function $K(\theta)$ having the following properties:

K1) $K(\theta)$ is a symmetric, nonnegative function on the real line, with compact support $[-\kappa, \kappa]$, uniformly Lipschitz with constant L_κ.

K2) $\frac{1}{2\pi}\int_{-\infty}^{\infty} K(\theta)\, d\theta = 1$ and $\frac{1}{2\pi}\int_{-\infty}^{\infty} \theta^2 K(\theta)\, d\theta = 1$.

So the kernel estimate of $\mu(\omega)$ is defined as a smoothing of the averages $y_{\cdot j}^{(N)} = (1/r)\sum_{i=1}^{r} y_{ij}^{(N)}$ in the form:

$$\widehat{\mu}(\omega) = \frac{1}{Nh} \sum_{j=-\nu}^{\nu} K\left(\frac{\omega - \omega_j^{(N)}}{h}\right) y_{\cdot j}^{(N)} \tag{5}$$

being h the bandwidth.

Once this parameter has been estimated, we smooth the residuals $y_{ij}^{(N)} - \widehat{\mu}\left(\omega_j^{(N)}\right)$ to define the estimate of each individual trajectory $Z_i(\omega)$, for $i = 1, \ldots, r$ so

$$\widehat{Z}_i(\omega) = \frac{1}{N\lambda_i} \sum_{j=-\nu}^{\nu} K\left(\frac{\omega - \omega_j^{(N)}}{\lambda_i}\right)\left(y_{ij}^{(N)} - \widehat{\mu}\left(\omega_j^{(N)}\right)\right) \tag{6}$$

being λ_i the corresponding bandwidth.

From here on, we assume following conditions (*M1*) and (*M2*) for which suitable properties for $\widehat{\mu}(\omega)$ and $\widehat{Z}_i(\omega)$ hold (Luengo-Merino et al. (2006)).

(*M1*) The function $\mu(\omega)$ is twice continuously differentiable on $[-\pi, \pi]$.
(*M2*) The function $\Psi_Z(\omega, \theta) = \operatorname{cov}[Z_i(\omega), Z_i(\theta)]$ is twice continuously differentiable on the square $[-\pi, \pi] \times [-\pi, \pi]$.

3 Statistical test for homogeneity testing

3.1 The statistical test

The natural spectral homogeneity hypothesis should be

$$H_0 : \operatorname{var}[Z(\omega)] = 0, \text{ for } |\omega| \leq \pi$$

However, since the processes are only observed at a finite number of times, we will work on a finite number of equidistant frequencies in $[0, \pi]$ (the times series are real valued, and thus the spectrums are symmetric).
More concretely, we choose a fixed value $a \in \mathbf{Z}^+$ and consider the set of frequencies $\varphi_j = \pi j/a$, $j = 1, \ldots, (a-1)$. Then, the homogeneity null hypothesis for the set of objects B is formulated as

$$H_0 : \operatorname{var}[Z(\varphi_j)] = 0, \text{ for } j = 1, \ldots, (a-1) \tag{7}$$

From here on, we suppose that the number of observations by series is $N = 2aM$. So we have $\varphi_j = (2\pi jM)/N = \omega_{jM}$, i. e. the selected fixed frequencies φ_j always coincide with the (jM)-th Fourier frequency, for $j = 1, \ldots, (a-1)$. In this context, the statistics test proposed for testing H_0 is:

$$D_{r,N}^2 = \frac{1}{\sqrt{(a-1)(r-1)}} \sum_{i=1}^{r} \sum_{j=1}^{a-1} \left(\frac{y_{i,jM}^{(N)} - y_{\cdot jM}^{(N)}}{\sigma_e}\right)^2 - \frac{1}{\sqrt{(a-1)(r-1)}} \tag{8}$$

The test consists of rejecting H_0 when $D_{r,N} > d_0$, for a certain critical value d_0. Note that $y_{i,jM}^{(N)}$ is the observation corresponding to b_i at the fixed frequency φ_j. We can immediately prove that:

$$E\left[D_{r,N}^2\right] = \sqrt{\frac{r-1}{a-1}} \sum_{j=1}^{a-1} \frac{\operatorname{var} Z_i\left(\varphi_j^{(a)}\right)}{\sigma_e^2} \tag{9}$$

which allows us to evaluate the suitability of the test proposed for the testing of H_0.

It is interesting to point out that $D_{r,N}$ is independent of the value of C and that under H_0 has the form:

$$D_{r,N}^2 = \frac{1}{\sqrt{(a-1)(r-1)}} \sum_{i=1}^{r} \sum_{j=1}^{a-1} \left(\frac{e_{i,jM}^{(N)} - e_{\cdot jM}^{(N)}}{\sigma_e} \right)^2 - \frac{1}{\sqrt{(a-1)(r-1)}} \quad (10)$$

where $e_{\cdot jM}^{(N)} = (1/r) \sum_{i=1}^{r} e_{i,jM}^{(N)}$.

3.2 Distribution of the statistical test

In order to calculate p-values, we must obtain the distribution of the statistical test $D_{r,N}^2$ under H_0. We propose a bootstrap approximation given by the following algorithm:

Step 1. Obtain by means of (5) the estimate $\widehat{\mu}(\omega)$ of the population parameter .

Step 2. Obtain by means of (6) the estimation $\widehat{Z}_i(\omega)$ of each $Z_i(\omega)$ for $i = 1, \ldots, r$.

Step 3. Determine the $(a-1)$-dimensional r vectors of the residuals $\widehat{\mathbf{e}}_i = \left(\widehat{e}_{i,M}, \ldots, \widehat{e}_{i,(a-1)M}\right)$ where $\widehat{e}_{i,jM} = y_{i,jM}^{(N)} - \widehat{\mu}(\varphi_j) - \widehat{Z}_i(\varphi_j)$, for each $i = 1 \ldots, r$; $j = 1, \ldots,(a-1)$

Step 4. Calculate the $(a-1)$-dimensional r vectors of centred residuals $\widetilde{\mathbf{e}}_i = \left(\widetilde{e}_{i,M}, \ldots, \widetilde{e}_{i,(a-1)M}\right)$ where $\widetilde{e}_{i,jM} = \widehat{e}_{i,jM} - (1/r) \sum_i \widehat{e}_{i,jM}$, for $i = 1 \ldots, r.$ and $j = 1, \ldots,(a-1)$

Step 5. Choose L random samples $\mathbf{e}_1^*, \ldots, \mathbf{e}_r^*$, each of size r, of the empirical distribution corresponding to the vectors $\widetilde{\mathbf{e}}_1, \ldots, \widetilde{\mathbf{e}}_r$.

Step 6. Calculate the L values corresponding to the bootstrap statistic

$$D_{r,N}^{*2} = \frac{1}{\sqrt{(a-1)(r-1)}} \sum_{i=1}^{r} \sum_{j=1}^{a-1} \left(\frac{e_{i,jM}^* - e_{\cdot jM}^*}{\sigma_e} \right)^2 - \frac{1}{\sqrt{(a-1)(r-1)}} \quad (11)$$

where $e_{\cdot jM}^* = (1/r) \sum_i e_{i,jM}^*$, which provide an approximation of the probability distribution of $D_{r,N}^2$, under the null hypothesis.

4 The bootstrap principle holds

To prove the theoretical validity of the bootstrap principle, according to Bickel and Freedman (1981), we use the Mallows distance.

For $p \geq 1$, let $\Gamma_p\left(\mathbf{R}^k\right)$be the set of probability distributions F on $\mathbf{R}^k$ such that $\int \|x\| \, dF(x) < \infty$.

For probability distribution functions F and G in $\Gamma_p\left(\mathbf{R}^k\right)$, the Mallows distance $d_p^{(k)}(F,G)$ is defined as the inferior of $E\left[\|X-Y\|^p\right]^{1/p}$ over the pairs of random vectors X and Y, such that X has law F and Y has law G. If X and Y are k-dimensional random vectors with probability distributions F and G respectively, we understand $d_p^{(k)}(X,Y)$ for $d_p^{(k)}(F,G)$.

We will prove the consistency of the method in the sense of the Mallows metric with a fixed and $r, N \to \infty$.

Throughout this paper, $P_N^{(a-1)}$denotes the $(a-1)$-dimensional probability distribution of the random vectors $\mathbf{e}_i^{(N)} = \left(e_{i,M}^{(N)}, \ldots, e_{i,(a-1)M}^{(N)}\right)$, $i = 1, \ldots, r$ corresponding to model (3). For the random sample $b_1, \ldots, b_r$, $P_{r,N}^{(a-1)}$denotes the $(a-1)$-dimensional empirical distribution corresponding to the random errors $\mathbf{e}_i^{(N)}$, $i = 1, \ldots, r$; $\widehat{P}_{r,N}^{(a-1)}$denotes the empirical distribution of the $(a-1)$-dimensional residuals $\widehat{\mathbf{e}}_i, i = 1, \ldots, r$ defined in step 3 of the bootstrap algorithm. Finally, $\widetilde{P}_{r,N}^{(a-1)}$denotes the $(a-1)$-dimensional empirical distribution corresponding to the random sample of centred residuals $\widetilde{\mathbf{e}}_i, i = 1, \ldots, r$ defined in step 4.

The following theorems assure the closeness of these distributions according to the Mallows metric of order 2.

According to lemma (8.4) of Bickel and Freedman (1981), for each N fixed, $d_2^{(a-1)}\left(P_N^{(a-1)}, P_{r,N}^{(a-1)}\right) \to 0$, almost sure for $r \to \infty$. However, this could not be true in general for $r, N \to \infty$. The next theorem assures the convergence of this distance to zero, whatever.

Theorem 1. *Under the conditions of the model, and* $N = 2aM$

$$d_2^{(a-1)}\left(P_N^{(a-1)}, P_{r,N}^{(a-1)}\right) \to 0, \quad \textit{for} \quad r, N \to \infty$$

Theorem 2. *Let our model be such that (M1) and (M2) are true for* $\mu(\omega)$ *and* $\Psi_Z(\omega, \theta)$, *respectively. Let* $\widehat{\mu}(\omega)$ *and* $\widehat{Z_i}(\omega)$ *be defined as in (5) and (6), respectively, and let* $K(\theta)$ *satisfy the assumptions K1) and K2). Hence, for* $h, \lambda \to 0$ *and* $r, Nh, N\lambda \to \infty$, *N=2aM, we obtain:*

(i) $d_2^{(a-1)}\left(P_{r,N}^{(a-1)}, \widehat{P}_{r,N}^{(a-1)}\right)^2 \to 0$ *in probability*

(ii) $d_2^{(a-1)}\left(\widehat{P}_{r,N}^{(a-1)}, \widetilde{P}_{r,N}^{(a-1)}\right)^2 \to 0$ *in probability*

Finally, we evaluate the consistency of the bootstrap approach in the form given for the next theorem.

Theorem 3. *With the same hypothesis as in theorems 1 and 2, and being moreover* $d_2^{(a-1)}\left(P_N^{(a-1)}, P_{r,N}^{(a-1)}\right)^2 = o_p\left(r^{1/2}\right)$, *and* $r^{1/2}h^2, r^{1/2}\lambda^2$, $r^{1/2}(Nh)^{-1}, r^{1/2}(N\lambda)^{-1} \to 0$, *then*

$$d_2^{(1)}\left(\Delta_{r,N}, \Delta_{r,N}^*\right)^2 \to 0 \quad \textit{in probability}$$

where $\Delta_{r,N} = \left\{ \frac{1}{\sqrt{(a-1)(r-1)}} \sum_{j=1}^{a-1} \sum_{i=1}^{r} \left(e_{i,jM}^{(N)} - e_{\cdot,jM}^{(N)} \right)^2 \right\}^{1/2}$

and $\Delta_{r,N}^{*} = \left\{ \frac{1}{\sqrt{(a-1)(r-1)}} \sum_{j=1}^{a-1} \sum_{i=1}^{r} \left(e_{i,jM}^{*} - e_{\cdot,jM}^{*} \right)^2 \right\}^{1/2}$

5 Simulation study

We illustrate the bootstrap procedure proposed, by means of a simulation study in which the set of time series is generated from the following process of moving averages of order 2 and random coefficients:

$$X(\mathbf{b}, t) = \varepsilon(\mathbf{b}, t) + b_1 \varepsilon(\mathbf{b}, t-1) + b_2 \varepsilon(\mathbf{b}, t-2)$$

where $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ with law $N_2 \left(\begin{pmatrix} 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 2\gamma & \gamma \\ \gamma & 2\gamma \end{pmatrix} \right)$, $\gamma \geq 0$, and for any $\mathbf{b} \in B$, $\{\varepsilon(\mathbf{b}, t) : t \in \mathbf{Z}\}$ is a white noise process with probability distribution uniform $\left[-\sqrt{3}, \sqrt{3}\right]$

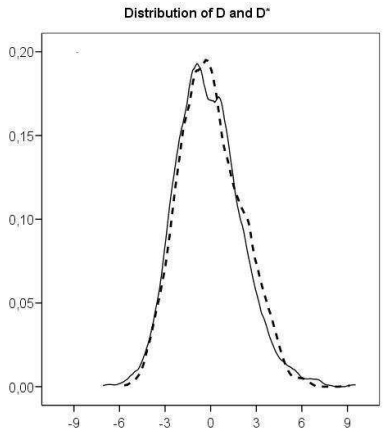


Fig. 1. Comparison of the density functions of D and D^*.

Obviously, if $\gamma = 0$, the null hypothesis of homogeneity is true. We have considered $a = 100$, $N = 200$ and $r = 60$ and have obtained an approximation of the theoretical distribution of $D_{r,N}^2$ under H_0, making 2000 simulations.

To obtain the bootstrap distribution, both the parameter and the trajectories were estimated using the kernel function $K(x) = \frac{3\pi}{2\sqrt{5}} \left(1 - \frac{x^2}{5} \right)$ if $|x| \leq \sqrt{5}$ with bandwidths estimated by a cross validation method. The number of bootstrap repetitions was $L = 2000$.

Figure 1 shows the density distributions of $D_{r,N}$ and $D^*_{r,N}$ obtained smoothing by means of kernel estimators the 2000 respective values.

To study the power of the test we have considered alternative hypotheses in terms of the trace of the covariance matrix of the coefficients in the process. In Figure 2 we show the trace-power graph.

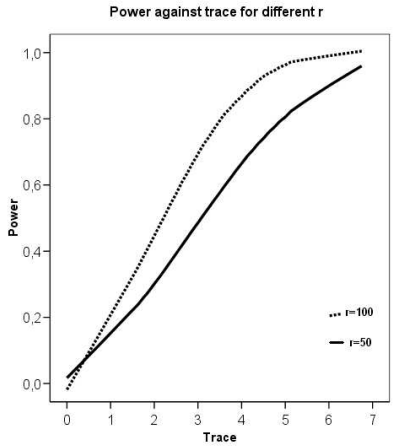


Fig. 2. Comparison of trace-power lines with $r = 50$ and $r = 100$.

References

BICKEL, P. and FREEDMAN, D. (1981): Some Asymptotic Theory for the Bootstrap. *Annals of Statistics, 9, 1196-1217.*

DIGGLE, P.J. and AL-WASEL, I. (1993): On Periodogram-Based Spectral Estimation for Replicated Time Series. In: Subba Rao (EdS.): *Developments in Time Series Analysis*. Chapman and Hall, Great Britain, 341–354.

FRANKE, J. and HÄRDLE, W. (1992): On Bootstraping Kernel Spectral Estimates. *Annals of Statistics, 20, 121-145.*

HERNÁNDEZ, C.N., ARTILES, J. and SAAVEDRA, P. (1999): Estimation of the Population Spectrum with Replicated Time Series. *Computational Statistics and Data Analysis, 30, 271-280.*

LUENGO, I., HERNÁNDEZ, C.N., and SAAVEDRA, P. (2006): Test to Compare Two Population Logspectra. *Computational Statistics, 21 (1), 91-101.*

LUENGO–MERINO, I., HERNÁNDEZ–FLORES C.N., and SAAVEDRA–SANTANA, P. (2006): Spectral Estimation in a Random Effect Model. *Proceedings in Computational Statistics 2006, 1217-1224.*

PRIESTLEY, M.B. (1981): *Spectral Analysis and Time Series*. Wiley, New York.

SAAVEDRA, P., HERNÁNDEZ, C.N. and ARTILES, J. (2000): Spectral Analysis with Replicated Time Series. *Communications in Statistics Theory and Methods, 29, 2343-2362.*

Unit Root Tests Using the ADF-Sieve Bootstrap and the Rank Based DF Test Statistic: Empirical Evidence

Valderio A. Reisen[1] and Maria Eduarda Silva[2]

[1] Departamento de Estatística, CCE, UFES, Vitoria, Brazil, *valderio@cce.ufes.br*
[2] Faculdade de Economia Universidade do Porto & UIMA, Portugal, *mesilva@fep.up.pt*

Abstract. In this paper we consider unit root tests under general time series models, including long-memory processes. The first test is a modified version of the ADF (Augmented Dickey-Fuller) sieve bootstrap test given in Chang and Park (2003) to test unit root process. The test is based on an approximation of the ARFIMA (fractionally differenced autoregressive moving average) model, where the fractional parameter is estimated by semi-parametric approaches.The second test under study is a rank based Dickey-Fuller test (RDF). The empirical power of the tests are investigated through Monte Carlo simulations under general time series models, including long-memory processes. The slight modification of the unit root sieve bootstrap test proposed here gives, in general, higher power than the ADF test. The rank based DF test for a unit root when the data generation process is a long memory process presents, in general, higher power than the usual ADF and sieve bootstrap unit root tests.

Keywords: unit root, long memory, sieve bootstrap, rank test

1 Introduction

The order of integration of a time series is a crucial determinant of the properties exhibited by the series. Recently, the theory and practice of unit root tests have produced many works in the literature of economics and related areas. One of the most popular unit root tests is the Augmented Dickey-Fuller test, given by Said and Dickey (1984). However, in the last decade a considerable number of published works has been calling attention to the low power of this test in finite sample sizes. In this context, two issues have received special attention. The first is concerned with the low power of unit root tests under the presence of MA errors and, also, the near unit root case and is, now, well documented, see for example Maddala and Kim (1998) and references therein. The second is concerned with the effect of the use of the standard unit root tests when the underlying process presents long memory behaviour.

A common assumption in time series analysis is that observations separated by a long-time span are independent or nearly so. Thus the most widely

applied models for stationary time series have spectral densities which are bounded at the frequency $\omega = 0$ and autocorrelation functions decaying exponentially to zero. However in many time series, particularly those arising in economics and hydrology, the dependence between distant observations, though small, is not negligible. These series are called long memory: they exhibit cycles and changes of level of all orders of magnitude, suggesting non-stationarity and their spectral densities increase indefinitely as the frequency decreases to zero, see Beran (1994) and references cited there. The ARFIMA(p, d, q) (fractionally differenced ARMA) model has been used to model time series with long-memory behaviour. The parameter d, usually called fractional parameter, describes the memory of the process and has been recently considered to test for unit root, Santander et al (2003) and references therein.

Bootstrap approaches to estimation and testing of general classes of time series have been the focus of many works recently, see for example Franco and Reisen (2004), Chang and Park (2003) and references therein. One special bootstrap technique that has recently received attention is the sieve bootstrap Buhlmann (1997). This bootstrap technique is based on the approximation of the data generation process by a sequence of autoregressive processes of order p which depends on the sample size n. This sieve bootstrap enjoys a nonparametric property, being model free within a class of linear process. The use of the sieve bootstrap for testing unit root process has been considered in Bisaglia and Procidano (2002) and Chang and Park (2003). This last reference gives an overview of the recent developments on the bootstrap methodology applied to time series analysis. We follow the effort made by these two works to propose a modified version of the unit root sieve bootstrap test. The modification is based on the AR representation of the fractional long-memory process. Our Monte Carlo empirical investigation indicates that the proposed methodology to test unit root processes against a general class of time series models presents, in general, higher power than the usual ADF and sieve bootstrap unit root tests.

Recently, non-parametric procedures and in particular rank tests have been object of increased interest in time series analyis, due to their distribution free, exact, easy to compute and robust behaviour. Granger and Hallman (1991), among others proposed a rank Dickey-Fuller test, where the ranks of the data are used instead of the data themselves and whose asymptotic properties were derived by Fotopoulos and Ahn (2003), when the alternative is a stationary AR(1) process. Here we investigate the use of the rank DF (RDF) test when the data generation process is an fractionally integrated, ARFIMA process. Our empirical Monte Carlo investigation indicates that the rank based DF test to test unit root processes against a general class of time series models presents, in general, higher power than the usual ADF and sieve bootstrap unit root tests. In the following section, we define the fractionally integrated long-memory model, estimation methods, the unit root, the sieve

bootstrap and the rank unit root tests. In section 3 an empirical comparison of the tests is presented. Section 4 concludes the paper.

2 Fractional model and the unit root tests

Let $\{X_t\}_{t\in\mathbb{Z}}$ be a discrete zero-mean time series satisfying the equation

$$X_t = \alpha X_{t-1} + U_t \tag{1}$$

where $\{U_t\}_{t\in\mathbb{Z}}$ is a time series generated by

$$U_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k} \tag{2}$$

and $\{\epsilon_t\}$ is a sequence of independent and identically distributed random variables with zero-mean and constant variance σ_ϵ^2. Here, the process $\{U_t\}$ satisfies a general fractionally integrated ARMA(ARFIMA) $(I(\delta))$ representation as follows:

$$\phi_p(B)U_t = (1-B)^{-\delta}\theta_q(B)\epsilon_t, \tag{3}$$

where B is the backshift operator, $\phi_p(B)$ and $\theta_q(B)$ are polynomials of order p and q, respectively, and all of their roots are outside the unit circle. $\delta \in R$ is the long memory parameter and $(1-B)^{-\delta}$ is defined as a binomial expansion. When $|\delta| < \frac{1}{2}$, U_t is a stationary and invertible process.

The test of the uni t root null hypothesis $\alpha = 1$ will be considered for model (1) against fractional alternatives. Under this null hypothesis, model (1) becomes a $I(d)$ process, with $d = 1 + \delta$.

The ADF tests and the estimation methods are given in the sequence of this section.

The fractional estimation methods

The parameter δ of U_t may be estimated by semiparametric methods. One of the most popular methods was proposed by Geweke and Porter-Hudak (1983) (GPH) and it takes advantage of the form of the spectral density function of the ARFIMA process, $f(w)$, given by:

$$f(w) = \frac{\sigma_\varepsilon^2}{2\pi}\frac{\theta_q(e^{-iw})|^2}{\phi_p(e^{-iw})|^2}\left(2\sin\left(\frac{w}{2}\right)\right)^{-2\delta}, \quad w \in [-\pi, \pi].$$

Let $\{u_t\}$, $t = 1, ..., n$ a set of observations of U_t and $I(w)$ denote the periodogram at frequency w, $I(w) = n^{-1}|\sum_{t=1}^{n} u_t e^{iwt}|^2$. The GPH method uses $\ln I(w)$ as an estimate of the $\ln(f(w))$ to build a regression equation where the linear parameter to be estimated by OLS is $-\delta$. The number of observations in the regression equation is a function of n, that is, $g(n) = n^\tau$, $0 < \tau < 1$. Hurvich et al (1998) proved that, under some regularity conditions on the choice of the bandwidth, the GPH-estimator is consistent for the memory

parameter and asymptotically normal for Gaussian time series process. They also established that the optimal bandwidth is the order $o(n^{4/5})$. Phillips (2007) shows the consistency and derives the asymptotic distribution of the GPH-estimator for unit root process. Modified versions of GPH estimator have been proposed in the literature. In particular Reisen (1994) proposed the use of the smoothed periodogram which is a consistent estimator for the spectral density function. The smoothing function is based on the Parzen lag-window where the truncation point is $m = n^{\beta}$, $0 < \beta < 1$ and in the resulting regression equation, function $g(n)$ is chosen as above. This estimation method is, hereafter, denoted by Sp.

The ADF tests

To test the unit root hypothesis of model (1), i.e., $H_0 : \alpha = 1$ ($\delta = 0$), the ADF test is based on the regression equation

$$X_t - X_{t-1} = (\alpha - 1)X_{t-1} + \sum_{k=1}^{\nu} \alpha_k \triangle X_{t-k} + \epsilon_{\nu,t}. \tag{4}$$

As previously noted, under the null hypothesis of $\alpha = 1$, $\triangle X_t = U_t$. The ADF statistic test is given by

$$ADF = \frac{\hat{\alpha} - 1}{s(\hat{\alpha})} \tag{5}$$

where $\hat{\alpha}$ is the OLS estimate of α and $s(\hat{\alpha})$ is the estimated standard error for $\hat{\alpha}$. See for example, Said and Dickey (1984) or Chang and Park (2003) for more details. The ADF test requires a number of lags to be considered in the regression equation (4). There are many works that report studies related to this. Here, the number of lags was set equal to $n^{0.25}$, Diebold and Nerlove (1990).

ADF-Sieve test

To obtain the ADF-Sieve test, the bootstrap samples are generated from the fitted residuals of the AR representation of equation (4). The choice of the order ν is based on the AIC criterion. Here, we fixed $\nu(max) = 10log_{10}(n)$. The estimates for the AR coefficients are obtained from the Yule-Walker equations. Since the sieve bootstrap procedure is well described in Buhlmann (1997) and Chang and Park (2003), to save space we do not give the full details of this method here.

ADF-Sievegph and ADF-SieveSp tests

The stationary and invertible ARFIMA process has infinite MA representation as given in equation (2) where, as $k \longrightarrow \infty$, $\psi_k \sim C^* k^{\delta - 1}$ for some constant C^* which depends on the parameters of the process (see Hosking, 1981). Then, $\sum_{k=0}^{\infty} k^r \psi_k < \infty$ for $r = -[v + \delta]$, for any $v > 0$ and $[\cdot]$ denotes integer part. Hence, for the model above the assumptions A1 and A2 given in Buhlmann (1997) are satisfied. These assumptions are equivalent to assumptions (1) and (2) in Chang and Park (2003). Our sieve bootstrap

ADF test follows the same procedure as the ADF-Sieve. However, in the equation (4) the AR(ν) representation is replaced by a truncated binomial expression of $(1-B)^{\delta}$, where the parameter δ is estimated by GPH and Sp methods. Hence, the resulting ADF-Sieve tests are called ADF-Sievegph and ADF-SieveSp for GPH and Sp methods, respectively. In the simulation investigation discussed in the next section, we set the truncation point of the binomial expression at the value when the coefficient is smaller than 10^{-5}. We have also considered the truncation point at the sample size value, but the results did not change significantly.

The simulation results presented in the next section are very encouraging to consider the slight modification of the ADF-Sieve unit root test here proposed as a alternative method to test unit root hypothesis.

The Rank DF (RDF) test

Let $R_{n,t}$ denote the rank of X_t among $X_1, X_2, \ldots, X_n$. To test the null hypothesis of unit root, $H_0 : \alpha = 1$, $\hat{\alpha}$ and $s(\hat{\alpha})$ in the test statistic defined in (5) are replaced by their rank counterparts, $\hat{\alpha}^{(r)}$ and $s(\hat{\alpha}^{(r)})$, given by

$$\hat{\alpha}^{(r)} - 1 = \frac{\sum_{t=2}^{n} R_{n,t-1} \triangle R_{n,t}}{\sum_{t=2}^{n} R_{n,t-1}^2} \tag{6}$$

and

$$s(\hat{\alpha}^{(r)}) = \frac{S^{(r)}}{(\sum_{t=2}^{n} R_{n,t-1}^2)^{1/2}} \tag{7}$$

where

$$S^{(r)2} = \frac{1}{n-2} \sum_{t=2}^{n} (R_{n,t} - \hat{\alpha}^{(r)} R_{n,t-1})^2$$

yielding the following test statistic

$$RDF = \frac{\hat{\alpha}^{(r)} - 1}{s(\hat{\alpha}^{(r)})} = \frac{\hat{\alpha}^{(r)} - 1}{S^{(r)}} (\sum_{t=2}^{n} R_{n,t-1}^2)^{1/2} \tag{8}$$

This is the same test statistic as that proposed by Granger and Hallman (1991), however the authors dealt with ARMA(p,q) processes in the alternative of $H_0 : \alpha = 1$. Fotopoulos and Ahn (2003) proved weak convergence of RDF and obtained critical values by Monte Carlo simulations. In this work we consider RDF for unit root tests and obtain empirical size and powers when the underlying model is an ARFIMA(0,d,0), $0 \leq d \leq 1$.

3 Monte Carlo Simulation study

We simulated realizations of $n = 100$ observations from the ARFIMA$(0, d, 1)$, model, $X_t = (1-B)^{-d}(1-\theta B)\epsilon_t$, $t = 1, ..., n$, for different parameter values of d and θ and tested the null hypothesis of unit root process ($H_0 : d = 1$). The

processes were simulated recursively, Hosking(1981) and setting ϵ_t a Gaussian white noise process with zero-mean and unit variance. To avoid any effect related to the initialization of the recursions, the initial 600 observations were discarded. The computations were carried out using the FORTRAN program. To obtain the fractional estimates to be used in the ADF-Sievegph and ADF-SieveSp tests, the bandwidth in the regression equation was $g(n) = n^{0.8}$. In the Sp method, the Parzen window was set equal to $n^{0.9}$ (see Reisen (1994) for details). The reported results are based on 3000 replications using 3000 bootstrap repetitions. As previously described, in the ADF-Sieve we use the AIC criterion to select the order of the approximated AR representation of (4). The 5% critical points for the ADF are given in the Tables 1 and 2 (see Hamilton, 1991). The 5% critical point for the RDF was obtained by Monte Carlo simulation. In our practical experiment, to generate the bootstrap samples we followed all suggestions addressed in Chang (2003), pages 390-391.

The empirical rejection rates of the ADF, sieve ADF and RDF tests when the data generation process (DGP) is an ARFIMA$(0, d, 0)$ are given in Table 1. Table 2 presents the empirical rejection rates for the ADF and sieve ADF tests when de DGP is an ARFIMA$(0, d, 1)$ with $\theta_1 = -0.4$.

Tests	Pc	d					
		1.0	0.9	0.8	0.7	0.5	0.3
ADF	-1.95	0.046	0.053	0.083	0.152	0.542	0.957
ADF-Sieve	-	0.061	0.065	0.106	0.191	0.611	0.974
ADF-Sievegph	-	0.062	0.074	0.107	0.198	0.604	0.969
ADF-SieveSp	-	0.059	0.070	0.101	0.193	0.616	0.973
RDF	-1.75	0.048	0.073	0.184	0.375	0.898	0.994

Table 1. Rejection rates (power). True model ARFIMA$(0, d, 0)$.

The sieve bootstrap procedures seem to improve the ADF test (Table 1). The empirical size of all tests are very close to the nominal value. In general, all sieve tests give superior power compared to the usual ADF test, although the proposed sieve tests using the AR representation of the ARFIMA process seem to be superior(higher power) than the others. The overall conclusions about the empirical power results do not change when the MA coefficient is included in the model (Table 2). Regarding the DF rank based test, RDF, in this limited study we obtained a size which is comparable to the size of the ADF test and a slightly higher power for large values of d. These results are promising and need further investigation.

Tests	Pc	d				
		1.0	0.9	0.7	0.5	0.3
ADF	-1.95	0.047	0.050	0.151	0.555	0.971
ADF-Sieve	-	0.056	0.057	0.175	0.585	0.964
ADF-Sievegph	-	0.068	0.066	0.175	0.567	0.955
ADF-SieveSp	-	0.062	0.063	0.163	0.572	0.961
RDF	-1.75	0.046	0.062	0.379	0.937	1.0

Table 2. Rejection rates (power). True model ARFIMA$(0, d, 1)$, $\theta = -0.4$.

4 Conclusions

This paper proposes the use of the sieve bootstrap to test unit root processes against fractional alternatives. An alternative ADF-Sieve test is also given which is based on the AR representation of the ARFIMA$(0, d, 0)$ model. To estimate the fractional parameter, two semi-parametric estimation methods are considered. The sieve bootstrap unit root tests here proposed give, in general, higher power to reject the null hypothesis of unit root process when the underling process is not a unit root. The research related to this topic is still under-way by the authors for different sample sizes and models. It is well-known the effect of bandwidth parameter estimate in the semi-parametric methods. Here, this was fixed although different values will be also considered in our future research related to the ADF-Sieve test. Also, the use of parametric estimation methods may be considered in this context. Moreover, the performance of a rank based Dickey-Fuller test is investigated when we are testing unit root against long-memory. This study indicates the need for further investigation into the use of this statistic.

Acknowledgements

V.A. Reisen gratefully acknowledge partial financial support from CNPq,Brazil. The present paper is a part of the results that were obtained when the first author was a visiting researcher at UMIST-Manchester, UK, and he thanks Prof. T. Subba Rao for the invitation. The authors also Marcelo Bourguignon Pereira, undergraduate student under the first author supervision, to provide some empirical results presented in the paper. This research was partly supported by *Fundação para a Ciência e Tecnologia* (Portugal) through *Unidade de Investigação Matemática e Aplicações* of Universidade de Aveiro and Projecto *POCI/MAT/61096/2004*.

References

BERAN, J. (1994): *Statistics for long–memory*. Chapman and Hall, New–York.

BISAGLIA, L. and PROCIDANO, I. (2002): On the power of the Augmented Dickey–Fuller test against fractional alternatives using bootstrap. *Economic Letters 77, 343-347*.

BUHLMANN, P. (1997): Sieve bootstrap for time series. *Bernoulli 3, 123–148*.

CHANG, Y. and PARK, J.Y. (2003): A sieve bootstrap for the test of a unit root. *Journal of Time Series Analysis 24, 370-400*.

DIEBOLD, F.X. and NERLOVE, M. (1990): Unit roots in Economic time series: a selective survey. In G.F. Rhodes and T.B. Fomby (Eds.): *Advances in Econometrics: A Research Annual 8*. JAI Press, 3–69.

FOTOPOULOS, S.B. and AHN, S.K. (2003): Rank based Dickey-Fuller test statistics. *Journal of Time Series Analysis 24, 647-662*.

FRANCO, G. and REISEN, V.A. (2004): Bootstrap techniques in semiparamentric estimation methods for ARFIMA models: a comparison study. *Computational Statistics 19, 243-259*.

GEWEKE, J. and PORTER–HUDAK, S. (1983): The estimation and application of long memory time series models. *Journal of Time Series Analysis 4, 221-238*.

GRANGER, C.W.J. and HALLMAN, J. (1991): Nonlinear transformations of integrated time series. *Journal of Time Series Analysis 12, 207-224*.

HAMILTON, J.D. (1991): *Time series analysis*. Princeton University, New Jersey.

HOSKING, J. (1981): Fractional differencing. *Biometrika 68, 165-176*.

HURVICH, C.M., DEO, R. & BROKSKY, J. (1998): The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series. *Journal of Time Series Analysis 19, 19-46*.

MADDALA, G.S. and KIM, I. (1998): *Unit Roots, Cointegration and Structural Change*. Cambridge University Press.

PHILIPPS, P.C.B. (2007): Unit root log periodogram regression. *Journal of Econometrics 138, 104-124*.

REISEN, V.A. (1994): Estimation of the fractional difference parameter in the ARIMA(p, d, q) model using the smoothed periodogram. *Journal of Time Series Analysis 15, 335-350*.

ROBINSON, P.M. (1995): Log–periodogram regression of time series with long range dependence. *Annals of Statistics 23, 1048-1072*.

SAID, S.E. and DICKEY, D.A. (1984): Testing for unit root autoregressive–moving average models of unknown order. *Biometrika 71, 599-608*.

SANTANDER, L., REISEN, V.A. and ABRAHAM, B. (2003): Non–cointegration tests and a fractional ARFIMA process. *Statistical Methods 5, 1-22*.

SOWELL, F. (1992): Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *Journal of Econometrics 53, 165-188*.

Monitoring Calibration of the Singular Spectrum Analysis Method

Paulo Canas Rodrigues and Miguel de Carvalho

CMA and Faculty of Sciences and Technology, Universidade Nova de Lisboa
2829-516 Caparica, Portugal, *paulocanas@fct.unl.pt* and *mb.carvalho@fct.unl.pt*

Abstract. Singular Spectrum Analysis (SSA) is an increasingly popular extension of Principal Component Analysis, suited for data sets on which the dependence constraint is not fulfilled. Since the method was introduced in the seminal work of Broomhead and King (1986), a broad number of applications have been developed (for a reference, see Golyandina et al. (2001)). SSA leaves two decisions to the analyst, namely the choices of the values of the window length and the number of leading eigentriples for conducting the reconstruction. In this work we provide a comparative analysis of several well known measures of forecast accuracy, for different values of the window length and ways of grouping the eigentriples. For such purpose we used data on the monthly US unemployment rate, ranging from 1948 to 2007. The results evidence the need for an extreme precaution in the choice of the above mentioned parameters, as they can compromise the forecast accuracy.

Keywords: eigentriples, out-of-sample forecast, recurrent forecast algorithm, singular spectrum analysis, US unemployment rate, window length

1 Introduction

Principal Component Analysis (PCA) is one of the main standard tools in multivariate data analysis (c.f., Jolliffe (2002)). However, given the original context under which PCA was developed, it was not suited in order to account for dependence, and so as a consequence not adequate for most time-series problems. One increasingly popular extension of PCA for frameworks wherein the requirement of independence is not fulfilled, is given by Singular Spectrum Analysis (SSA, from hereafter). This method was introduced in the seminal work of Broomhead and King (1986). Barely speaking the core idea of SSA relies in the decomposition of the time series within several components that usually can be identified as trends, oscillatory components or noise components. SSA leaves two decisions to the analyst, namely the choices of the window length and the number of leading eigentriples for conducting the reconstruction. One particular peculiarity of the method is that it relies on no statistical assumption. In fact, quoting Golyandina et al. (2001), SSA is more a "technique of multivariate geometry than of statistics". Furthermore, unlike other time-series techniques which can be used for the very same purposes (such as the AutoRegressive Integrated Moving Average - ARIMA - model), there is no need to make the data stationary before carrying out the analysis.

Since its introduction, the method has been successfully applied in a broad number of fields of research with a main emphasis in climatology and meteorology (see Paegle et al. (2000) and Allen and Smith (1996)). Applications in Econometrics are still at their early days. Unfortunately, as referred by Golyandina et al. (2001) this technique is not yet well known among Econometricians. A remarkable exception is provided by Carvalho and Rodrigues (2007), who have applied SSA and the Recurrent Forecast algorithm in order to produce out-of-sample forecasts for the unemployment rate in the Euro Area. Their results evidenced a larger forecast accuracy of SSA relatively to the results produced by the European Central Bank Survey of Professional Forecasters.

In this paper, we contribute by assessing how "costly" (from the forecast accuracy standpoint) it can be a less suited calibration of the technique. This work is largely motivated by the recent discussion originated by Chatfield (2001), concerning the calibration of the SSA technique. We used the US monthly unemployment rate (1948(1) - 2007(12)), in order to conduct the experiment. To the produced out-of-sample were then applied some well known measures of forecast accuracy, namely: the Root Mean Squared Error (RMSE); the Average Absolute Error (AAE); the Average Directional Accuracy (ADA) and the Mean Absolute Percentage Error (MAPE). The results emphasize the need for a careful choice of the windows length value, as well as the number leading eigentriples for conducting the reconstruction. In fact as shown below, an inadequate choice of the aforementioned parameters can become quite expensive from the forecasting accuracy standpoint.

This article is organized as follows: in the next section we introduce SSA. This section assumes some familiarity with the concepts of hankelization and eigentriple. In section 3 we summarize and comment the achieved results. Finally in section 4, we present some final remarks.

2 The forecast method

2.1 Singular spectrum analysis

In this section, we closely follow the exposition on Golyandina et al. (2001). Given a time series of length n, the basic SSA method relies on four steps. At first step (called the *embedding step*), the one-dimensional series is represented as a multidimensional series whose dimension is called the window length, L. The multidimensional time series (which is a sequence of vectors) forms the trajectory matrix (1).

If we consider $\mathbf{U} = [u_1 : u_2 : \cdots : u_n]$ the time series of length n, and L, $1 < L < n$, the window length, we obtain $K = n - L + 1$ lagged vectors of length L, $\mathbf{X}_j = [u_j : u_{j+1} : \cdots : u_{j+L-1}]'$, $j = 1, 2, \ldots, K$ and the trajectory matrix

$$\mathbf{X} = [\mathbf{X}_1 : \cdots : \mathbf{X}_K]' = \begin{bmatrix} u_1 & u_2 & \cdots & u_L \\ u_2 & u_3 & \cdots & u_{L+1} \\ \vdots & \vdots & \ddots & \vdots \\ u_K & u_{K+1} & \cdots & u_n \end{bmatrix} \tag{1}$$

The second step, SVD step, is the Singular Value Decomposition of the trajectory matrix (1) into a sum of rank-one bi-dimensional matrices as:

$$\mathbf{X} = \sum_{i=1}^{d} \mathbf{X}_i = \sum_{i=1}^{d} \sqrt{\lambda_i} W_i V_i', \tag{2}$$

where λ_i, W_i and V_i, $i = 1, \ldots, L$, are the eigenvalues, the left and right singular vectors, respectively and $d = rank(\mathbf{X}) \leq L$. The collection $(\sqrt{\lambda_i}, W_i, V_i)$ is called the ith *eigentriple* of the SVD of matrix $\mathbf{S} = \mathbf{X}'\mathbf{X}$. The first two steps together are considered as the *decomposition stage* of Basic SSA.

The next two steps form the *reconstruction stage*. The point of this stage is the reconstruction of the original time series as the sum of principal components using the *diagonal averaging* procedure.

The *grouping step* corresponds to splitting the matrices, computed at the SVD step, into several groups (m - the intended number of principal components) and summing the matrices within each group. The result of the step is a representation of the trajectory matrix as a sum of several *resultant matrices*.

The last step transfers each resultant matrix into a time series, which is an additive component of the initial series. The corresponding operation is called *diagonal averaging*. It is a linear operation and maps the trajectory matrix of the initial series into the initial series itself. In this way we obtain a decomposition of the initial series into several addictive components.

Let $\mathbf{X}_{I_k}$, where $I_k = \{i_{k1}, \ldots, i_{km}\}$ is a set of indexes, $k = 1, \ldots, d$, is the number of nonzero singular values of $\mathbf{X}$ and m is the number of eigentriples used in the reconstruction step. If we apply the diagonal averaging procedure to $\mathbf{X}_{I_k}$, we obtain the series $\widetilde{\mathbf{U}}^{(k)} = [\widetilde{u}_1^{(k)} : \cdots : \widetilde{u}_n^{(k)}]$, which decomposes the initial series, $\mathbf{U}$, into the sum of m series:

$$\mathbf{U} \approx \sum_{k=1}^{m} \widetilde{\mathbf{U}}^{(k)}, \qquad k = 1, \ldots, K. \tag{3}$$

These m series represent the first m principal components. Note that the equality in (3) only happens if $m = L$, i.e., if we sum all series/principal components.

Further, we use the following notation: let $\mathbf{Y} \in \mathbb{R}^J$, then by $\mathbf{Y}^{\nabla}$ we denote the first $J - 1$ components of vector $\mathbf{Y}$.

Carvalho and Rodrigues (2007) provide an illustration of the course of the process followed by SSA, referring to the quarter unemployment rate in the Euro Area.

2.2 Recurrent forecast algorithm

In order to define the recurrent forecast algorithm, we first need to define the space on which the algorithm acts. Consider a linear space $\mathcal{L} \subset \mathbb{R}^L$ such that $dim(\mathcal{L}) = r$. Further, let $P_1, ..., P_r$ denote an orthogonal basis in $\mathcal{L}$. The recurrent forecast algorithm is based on the following recurrently defined variable,

$$\widehat{u}_t = \widetilde{u}_t \mathbb{I}(1, ..., n) + \sum_{j=1}^{L-1} \alpha_j \widehat{u}_{t-j} \mathbb{I}(n+1, ..., n+m), \tag{4}$$

where $\mathbb{I}(.)$ denotes the indicator function, and where the weights α_j are defined as follows:

$$\alpha_j \equiv [\mathcal{R}]_{L-j+1}, \quad j = 1, ..., L-1, \tag{5}$$

where

$$\mathcal{R} = \left[1 - \left(\sum_{j=1}^{r} \pi_i^2 \right)^2 \right]^{-1} \sum_{j=1}^{r} \pi_i P_i^{\nabla} \tag{6}$$

and π_i is such that, $P_i = \left[P_i^{\nabla} \quad \pi_i \right]'$.

The forecasted values for the m periods following n are then given by

$$\widehat{\mathbf{U}} = [\widehat{u}_{n+1} : \cdots : \widehat{u}_{n+m}]'. \tag{7}$$

Further details regarding the above mentioned algorithm can be found in Golyandina et al. (2001).

3 Results

In order to conduct the propose experiment we considered the monthly US unemployment rate, ranging from January of 1948 to December of 2007. The pre-forecast period ranges from January of 1948 to December of 1978. The database used was collected using the Labor Force Statistics from the Current Population Survey of the Bureau of Labor Statistics.

The main decisions that the analyst has to make are the choice of the window length, L, and the number of eigentriples to perform the reconstruction step, m. Golyandina et al. (2001) suggest that L must be "small", whereas K should be very "large" (formally, $K \to \infty$). To the best of our knowledge there is no rule of thumb for choosing the number of eigentriples used in the reconstruction step, m.

3.1 Adjustment

Our application of the SSA method was calibrated considering different values for the window length. The considered values were $L = 12$, $L = 24$ and $L = 360$, for limit values, and plus seven values of L: 72; 120; 168; 216; 264; 312 (separated values between 24 and 360 by 48 observations - 4 years). These values (multiples of 12) were chosen in order to obtain trends and seasonal effects (Vautard et al. (1992)). Table 1 presents the share of variance explained by the first three, and remaining principal components for the aforementioned values of window length. The results provide statistical evidence supporting an overall domination of the first principal component, even when shifting the window length.

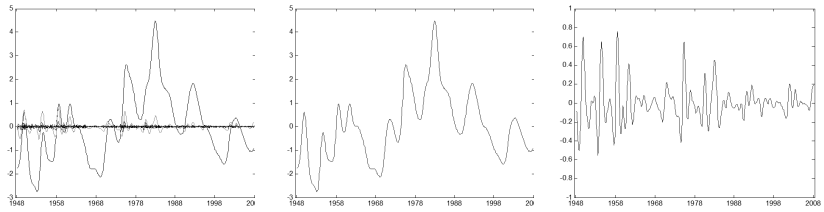
	Proportion of Explained Variance (%)			
L	PC1	PC2	PC3	Remainder
12	92.75	5.95	0.66	0.64
24	82.33	12.73	3.32	1.62
72	57.29	19.05	9.83	12.83
120	44.17	20.46	10.43	24.94
168	35.99	19.71	11.70	32.60
216	29.59	21.02	9.96	39.43
264	24.56	22.03	8.68	44.73
312	22.82	20.81	8.47	47.90
360	24.07	18.27	8.08	49.58

Table 1. A cross table evidencing the percentage of variance explained by the principal components, over the different values for the window length.

If we intend to explain a larger proportion of variance using a small number of principal components we confirm the suggestion of Golyandina et al. (2001), i.e., L should be "small" and K must be "large" (Table 1).

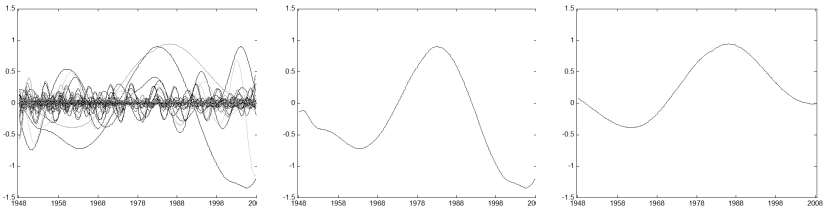
Without loss of generality, from hereafter we rely our focus considering a window length of 12 and 312. The principal components for the referred L values, are depicted in Figure 1 and Figure 2, respectively.

As it can be observed with a window length of 312 we obtain an overall smoother behavior of the principal components. Figure 1 shows a first principal component with proportion of variance explained of 92.75% (see Table 1) which represents a trend of the original time series. The second and following principal components for a window length of 12 can be seen as noise. The first two principal components of Figure 2 represent the trend of the US unemployment rate and the remaining the noise. No seasonal effects of any order



(a) (b) (c)

Fig. 1. All the extracted principal components are represented in plot (a). The remaining plots representing the first and second depicted in (b) and (c), respectively. These plots were obtained using a window length of 12.



(a) (b) (c)

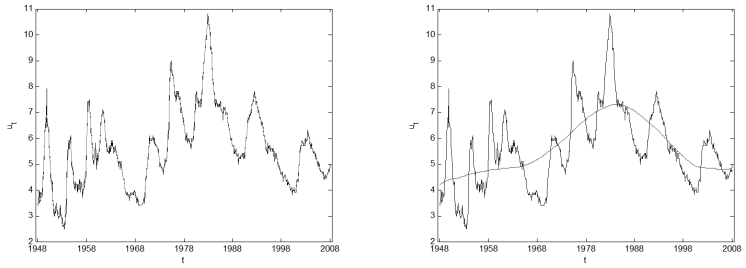
Fig. 2. All the extracted principal components are represented in plot (a). The remaining plots representing the first and second principal components are depicted in (b) and (c), respectively. These plots were obtained using a window length of 312.

seem to be observed. If we superimposing both first principal components in the same graph it can be seen the same trend with an overall (too) smoother behavior of the component obtained using a window length of 255.

The comparison between the observed and the forecasted values of the US unemployment rate, for a window length of 12 and 312, is depicted in Figure 3. As a consequence of the relatively "small" values of the RMSE, AAE and MAPE, the first plot ($L = 12$) of Figure 3 presents the observed and forecasted index almost entirely overlapped. From the inspection of the second plot ($L = 312$), it is visible an overall larger deviation due to "larger" values of RMSE, AAE and spatially of MAPE. The values of ADA were almost the same.

3.2 Forecasting

In what concerns to the production of forecasts, we used the aforementioned recurrent forecast algorithm. The yielded values are produced on a month-by-month basis, departing from a pre-forecast period ranging from January of 1948 to December of 1978. The CPU time of this procedure increases with the window length and with the number of leading eigentriples considered



(a) (b)

Fig. 3. The comparison between the observed and forecasted values of the US unemployment rate, considering a windows length of 12 and 312, respectively represented in plots (a) and (b).

on reconstruction. The predictive power of the method is illustrated over an analysis of the results from the considered values of the window length. The forecast accuracy of the method was assessed using three different criteria, namely: the RMSE, the AAE, the ADA and the MAPE. The results are summarized in Table 2 for one-step-ahead.

	Window Length (L)								
	12	24	72	120	168	216	264	312	360
RMSE	0.284	0.565	1.260	1.330	1.328	1.365	1.398	1.410	1.193
AAE	0.207	0.388	1.100	1.162	1.129	1.131	1.176	1.212	1.036
ADA (%)	70.12	69.25	62.83	60.92	60.63	60.35	62.07	62.93	62.22
MAPE (%)	3.24	5.87	17.73	19.24	18.83	18.80	20.30	21.08	17.34

Table 2. Evaluating the predictive power of the method one-step-ahead.

After the analysis of the values of Table 2, we decided to choose several small values for window length and check which one was the best for prediction, i.e., the one which produces smaller errors and larger accuracy. The considered values for L were taken from 3 to 24 and the errors are depicted in Figure 4. Table 3 summarizes the better values for the window length according to the considered errors for one, three and six steps ahead.

We also tried to optimize the number of leading eigentriples for reconstruction. Figure 5 depicts the values of RMSE, AAE, ADA and MAPE for

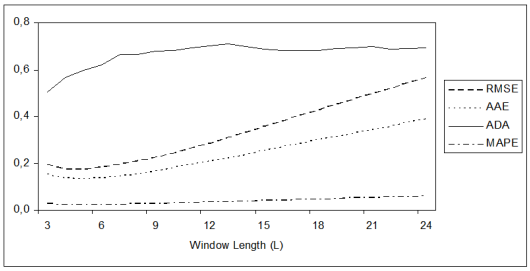


Fig. 4. RMSE, AAE, ADA and MAPE for the values of window length between 3 and 24 for one-step-ahead.

	Number of steps ahead		
	one-step-ahead	three-step-ahead	six-step-ahead
RMSE	$4 \rightarrow 0.173$	$6 \rightarrow 0.547$	$6 \rightarrow 0.890$
AAE	$5 \rightarrow 0.131$	$6 \rightarrow 0.255$	$6 \rightarrow 0.465$
ADA (%)	$13 \rightarrow 70.69$	$20 \rightarrow 68.97$	$23 \rightarrow 66.09$
MAPE (%)	$5 \rightarrow 2.17$	$6 \rightarrow 4.14$	$6 \rightarrow 7.41$

Table 3. Window length values with better fit according to the considered measures of forecast accuracy. $a \rightarrow b$ denotes the forecast accuracy, b, given the window length, a.

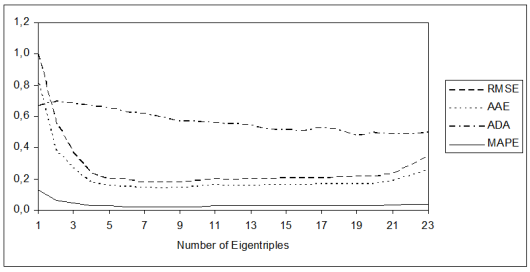


Fig. 5. RMSE, AAE, ADA and MAPE for the values of window length 24 and according to the number of leading eigentriples for reconstruction.

a window length of 24 and according to the number of leading eigentriples used for reconstruction (one-step-ahead).

In the overall, the results with better forecast accuracy for each considered window length were achieved taking the number of leading eigentriples for reconstruction close to $\frac{L}{3}$. To the best of our knowledge, this feature is not

known in the literature, and it is our intention to devote some attention to this topic in future work.

4 Summary and conclusions

The Singular Spectrum Analysis Method is combined with Recurrent Forecast Algorithm in order to produce out-of-sample forecasts to the monthly US unemployment rate. The main interest here was to assess how the forecasting accuracy can shift over different calibrations for the method. As mentioned earlier, this work is strongly motivated by the recent discussion originated with some remarkable comments of Chatfield (2001), concerning the calibration of the SSA techniques used in Golyandina et al. (2001). Our results evidence the need for a precautionary choice of the above mentioned parameters, as that decision can compromise the forecasting accuracy of the method.

One approach to suitably calibrate the method would be given by solving the following optimization problem,

$$L^* = \arg\min_{L \in \mathbb{N}} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(u_t - \hat{u}_t(L)\right)^2}, \tag{8}$$

for a fixed sample size.

Even though this naive approach can be in most circumstances computationally prohibitive, one can apply to it some stochastic search algorithms in order to achieve a reasonable approximation to the optimal MSE calibration. A similar optimization should be taken for the number of leading eigentriples for reconstruction. Even though the suggested objective function minimizes the MSE, other criteria would be equally legitimate, depending on the problem of interest. We postpone this problem for future work.

References

ALLEN, M.R. and SMITH, L.A. (1996): Monte Carlo SSA: Detecting irregular oscillations in the presence of colored noise. *Journal of Climate 9, 3373-3404.*

BROOMHEAD, D.S. and KING, G.P. (1986): Extracting Qualitative Dynamics from Experimental Data. *Physica D 20, 217-236.*

CHATFIELD, C. (2001): Reviewed Work: Analysis of Time Series Structure: SSA and Related Techniques by N. Golyandina; V. Nekrutkin; A. Zhigljavsky. *Biometrics 57(4), 1272-1273*

DE CARVALHO, M. and RODRIGUES, P.C. (2007): Forecasting the Unemployment Rate in the Euro Area. *Mimeo.*

GOLYANDINA, N., NEKRUTKIN, V. and ZHIGLJAVSKY, A. (2001): *Analysis of Time Séries Structure: SSA and Related Thechniques.* Chapman & Hall. New York.

JOLLIFFE, I.T. (2002): *Principal Component Analysis.* Springer Verlag. New York, Inc.

PAEGLE, J.N., BYERLE, L.A. and MO, K.C. (2000): Intraseasonal Modulation of South American Summer Precipitation. *Monthly Weather Review 128, 837-850.*

VAUTARD, R., YIOU, P. and GHIL, M. (1992): Singular Spectrum Analysis: A toolkit for Short Noisy Chaotic Signals. *Physica D 58, 95-126.*

Parameter Estimation for INAR Processes Based on High-Order Statistics

Isabel Silva[1] and M. Eduarda Silva[2]

[1] Faculdade de Engenharia & CEC, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal, *ims@fe.up.pt*
[2] Faculdade de Economia, Universidade do Porto & UIMA
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal, *mesilva@fep.up.pt*

Abstract. The high-order statistics (moments and cumulants of order higher than two) have been widely applied in several fields, specially in problems where it is conjectured a lack of Gaussianity and/or non-linearity. Since the INteger-valued AutoRegressive, INAR, models are non-Gaussian, the high-order statistics can provide additional information that allows a better characterization of these processes. Thus, an estimation method for the parameters of an INAR model, based on Least Squares applied on third-order moments is proposed. The results of a Monte Carlo study, to investigate the performance of the estimator, are presented and the method is applied to a set of real data.

Keywords: INAR process, estimation, high-order statistics

1 Introduction

In the recent past, the high-order statistics (HOS) have been widely applied in several fields, specially in problems where is conjectured a lack of Gaussianity and/or non-linearity. By HOS it is meant the moments and cumulants of order higher than two, in the time domain, and the corresponding multidimensional Fourier transform (polyspectrum), in the frequency domain. In this work, the time domain approach is considered.

Let $\{X_t\}$ be a kth-order stationary stochastic process. The kth-order joint moment of $X_t, X_{t+s_1}, \ldots, X_{t+s_{k-1}}$, for $s_1, \ldots, s_{k-1} \in \mathbb{R}$, is a function of $k-1$ variables defined by $\mu_X(s_1, \ldots, s_{k-1}) = \mathrm{E}[X_t X_{t+s_1} \ldots X_{t+s_{k-1}}]$, with $\mu_X = \mathrm{E}[X_t]$.

Recently, the integer-valued autoregressive process has been proposed in the literature to model time series of counts. The pth-order integer-valued autoregressive, INAR(p), process is defined as a discrete time non-negative integer-valued stochastic process, $\{X_t\}$, that satisfies the following equation (Latour (1998)):

$$X_t = \alpha_1 * X_{t-1} + \alpha_2 * X_{t-2} + \cdots + \alpha_p * X_{t-p} + e_t, \tag{1}$$

where

(i) $\{e_t\}$, designated the innovation process, is a sequence of independent and identically distributed (i.i.d.) non-negative integer-valued random variables with $\mathrm{E}[e_t] = \mu_e$, $\mathrm{Var}[e_t] = \sigma_e^2$ and $\mathrm{E}[e_t^3] = \gamma_e$;

(ii) the symbol $*$ represents the thinning operation (Steutel and Van Harn (1979), Gauthier and Latour (1994)), defined by

$$\alpha_i * X_{t-i} = \sum_{j=1}^{X_{t-i}} Y_{i,j}, \text{ for } i = 1, \ldots, p,$$

where $\{Y_{i,j}\}$, designated the counting series, is a set of i.i.d. non-negative integer-valued random variables such that $\mathrm{E}[Y_{i,j}] = \alpha_i$, $\mathrm{Var}[Y_{i,j}] = \sigma_i^2$ and $\mathrm{E}[Y_{i,j}^3] = \gamma_i$. All the counting series are assumed independent of $\{e_t\}$;

(iii) $0 \leq \alpha_i < 1$, $i = 1, \ldots, p-1$, and $0 < \alpha_p < 1$. Note that the stationarity condition for the INAR(p) process is that $\sum_{k=1}^{p} \alpha_k < 1$.

A special case is the Poisson INAR process with binomial thinning operation, where $\{e_t\}$ has a Poisson distribution with parameter λ and the counting series, $\{Y_j^{(i)}\}$, are a set of Bernoulli random variables with $P(Y_j^{(i)} = 1) = 1 - P(Y_j^{(i)} = 0) = \alpha_i$.

Since the INAR models are non-Gaussian, the HOS can provide additional information in the characterization of these processes. Thus, an estimation method for the parameters of an INAR model that uses HOS is proposed in this work. This method applies the Least Squares estimation method to minimize the errors between the third-order moment of the observations and of the fitted model.

This work is organized as follows: in Section 2 the third-order characterization of INAR(p) models is provided and the proposed Least Squares Estimation method using HOS is described. In Section 3 the results of a simulation study to assess the small sample properties of the proposed estimator are given and the method is applied to a set of observations concerning the number of plants within the industrial sector. Finally, some remarks are presented in Section 4.

2 Least squares estimation using HOS

The third-order characterization, in terms of moments and cumulants, of INAR models has been obtained by Silva and Oliveira (2004, 2005) and Silva (2005). In particular, the third-order moments of an INAR(p) process, defined by (1), satisfy a set of Yule-Walker type equations similar to those satisfied

by the bilinear process, that can be written as:

$$\begin{aligned}\mu_X(0,0) = &\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\alpha_i\alpha_j\alpha_k\,\mu_X(i-j,i-k)\\ &+3\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_j{\sigma_i}^2\mu_X(i-j)+3\mu_X(\sigma_e^2+{\mu_e}^2)\sum_{i=1}^{p}\alpha_i\\ &+3\mu_e\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_i\alpha_j\mu_X(i-j)+3\mu_X\mu_e\sum_{i=1}^{p}{\sigma_i}^2\\ &+\mu_X\textstyle\sum_{i=1}^{p}(\gamma_i-3\alpha_i{\sigma_i}^2-\alpha_i^3)+\gamma_e,\end{aligned} \tag{2}$$

$$\mu_X(0,k) = \textstyle\sum_{i=1}^{p}\alpha_i\mu_X(0,k-i)+\mu_e\mu_X(0),\quad k>0, \tag{3}$$

$$\begin{aligned}\mu_X(k,k) = &\textstyle\sum_{i=1}^{p}\sum_{j=1}^{p}\alpha_i\alpha_j\,\mu_X(k-i,k-j)+\sum_{i=1}^{p}{\sigma_i}^2\mu_X(k-i)\\ &+2\mu_e\mu_X(k)-\mu_X({\mu_e}^2-{\sigma_e}^2),\quad k>0,\end{aligned} \tag{4}$$

$$\mu_X(k,m) = \textstyle\sum_{i=1}^{p}\alpha_i\mu_X(k,m-i)+\mu_e\mu_X(k),\quad m>k>0, \tag{5}$$

where $\mu_X(0)=\sum_{i=1}^{p}\alpha_i\mu_X(i)+\mu_e\mu_X+V_p$, is the second-order moment of $\{X_t\}$, with $V_p={\sigma_e}^2+\mu_X\sum_{i=1}^{p}{\sigma_i}^2$, which represents the variance of the one-step-ahead prediction error (Silva (2005)).

These equations indicate that the INAR processes have a non-linear structure, therefore the first- and second-order moments are not sufficient to describe the dependence structure of the process.

Let $\{x_1,x_2,\ldots,x_n\}$ be a realization of a non-negative integer-valued stationary stochastic process with third-order moments $\mu(0,k)$, $k>0$. The approximating model considered is an INAR(p) process (order known) with parameters $\alpha_1,\ldots,\alpha_p,\mu_e,\sigma_e^2$ and third-order moments $\mu_X(0,k)$, $k>0$, satisfying (3), which can be represented in the following matrix form

$$\boldsymbol{\mu}_{3,X} = \mathbf{M}_{3,X}\boldsymbol{\alpha}+\mu_e\mu_X(0)\mathbf{1}_p, \tag{6}$$

where $\boldsymbol{\mu}_{3,X}$ is defined as

$$\boldsymbol{\mu}_{3,X} = [\,\mu_X(0,1)\;\cdots\;\mu_X(0,p)\,]^T,$$

$\mathbf{M}_{3,X}$ is the $p\times p$ non-symmetric Toeplitz matrix of the third-order moments of the INAR(p) process

$$\mathbf{M}_{3,X} = \begin{bmatrix} \mu_X(0,0) & \mu_X(1,1) & \ldots\ \mu_X(p-1,p-1)\\ \mu_X(0,1) & \mu_X(0,0) & \ldots\ \mu_X(p-2,p-2)\\ \vdots & \vdots & \ddots & \vdots\\ \mu_X(0,p-1)\ \mu_X(0,p-2) & \ldots & \mu_X(0,0)\end{bmatrix},$$

with $\mu_X(\cdot,\cdot)$ given in (2) to (5), $\boldsymbol{\alpha}=[\,\alpha_1\;\cdots\;\alpha_p\,]^T$ is the vector of coefficients, $\mu_X(0)$ is the second-order moment of the INAR(p) process and $\mathbf{1}_p$ is a $p\times 1$ vector of ones.

Defining

$$\mathbf{H} = [\mathbf{M}_{3,X} \qquad \mu_X(0)\mathbf{1}_p] \qquad \text{and} \qquad \boldsymbol{\theta} = [\,\alpha_1 \cdots \alpha_p \; \mu_e\,]^T,$$

equation (6) can be rewritten as

$$\boldsymbol{\mu}_{3,X} = \mathbf{H}\boldsymbol{\theta},$$

suggesting that $\boldsymbol{\theta}$ may be estimated by least squares, i.e., minimizing the squared error between the third-order moments of the fitted INAR(p) model, $\boldsymbol{\mu}_{3,X}$, and the third-order moments of the data,

$$\boldsymbol{\mu}_3 = [\,\mu(0,1) \cdots \mu(0,p)\,]^T.$$

Thus, $\hat{\boldsymbol{\theta}}$, the Least Squares estimator of $\boldsymbol{\theta}$ based on HOS (LS_HOS) satisfies

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \{L^*(\boldsymbol{\theta})\}$$

where

$$L^*(\boldsymbol{\theta}) = (\boldsymbol{\mu}_3 - \mathbf{H}\boldsymbol{\theta})^T(\boldsymbol{\mu}_3 - \mathbf{H}\boldsymbol{\theta}).$$

In practice, the estimator is calculated by substituting the moments in $\boldsymbol{\mu}_3$ and $\mathbf{H}$ by their sample counterparts.

Thus,

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \{\hat{L}^*(\boldsymbol{\theta})\} = \min_{\boldsymbol{\theta}} \{(\hat{\boldsymbol{\mu}}_3 - \hat{\mathbf{H}}\boldsymbol{\theta})^T(\hat{\boldsymbol{\mu}}_3 - \hat{\mathbf{H}}\boldsymbol{\theta})\}.$$

Note that an estimator for σ_e^2 can be obtained by $\hat{\sigma}_e^2 = \hat{V}_p - \overline{X}\sum_{i=1}^p \hat{\sigma}_i^2$, where $\overline{X}$ is the sample mean of the observations, $\hat{\sigma}_i^2$ is an estimator of the counting series variance for the i-th thinning operation, $\alpha_i * X_{t-i}, i = 1, \ldots, p$, and $\hat{V}_p = \hat{R}(0) - \sum_{i=1}^p \hat{\alpha}_i \hat{R}(i)$, with $\hat{R}(i) = \frac{1}{N}\sum_{t=1}^{N-i}(X_t - \overline{X})(X_{t+i} - \overline{X})$, representing the sample autocovariance function. The estimation of $\hat{\sigma}_i^2$ depends on the distribution of the counting series, for instance, in the case of the binomial thinning operation (when the counting series are Bernoulli distributed), $\hat{\sigma}_i^2 = \hat{\alpha_i}(1 - \hat{\alpha_i})$, for $i = 1, \ldots, p$.

3 Monte Carlo results and application to real data

The aim of the simulation study presented in this section is twofold: to examine the small sample properties of the estimator previously described and compare its performance with other estimation methods for the parameters of an INAR process.

Thus, 1000 realizations of Poisson INAR(p) processes ($e_t \sim \mathcal{P}o(\lambda)$), with binomial thinning operation, are generated, for $p = 0, \ldots, 3$. The sample sizes used are $N = 50, 200, 500$ and 1000 and parameter values considered

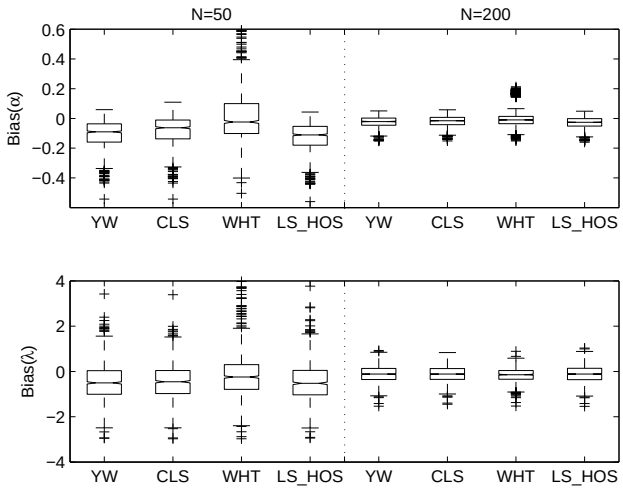


Fig. 1. Boxplots of the sample bias for the estimates obtained in 1000 realizations of 50 and 200 observations of the INAR(1) model: $X_t = 0.9 * X_{t-1} + e_t$, where $e_t \sim \mathcal{P}o(1)$.

are: $\lambda \in \{1.0, 3.0\}$, for $p = 1$, $\alpha_1 \in \{0.1, 0.4, 0.6, 0.9\}$, for $p = 2$, $(\alpha_1, \alpha_2) \in \{(0.1, 0.6), (0.6, 0.1), (0.3, 0.4), (0.4, 0.3), (0.1, 0.1), (0.4, 0.4)\}$, and for $p = 3$, $(\alpha_1, \alpha_2, \alpha_3) \in \{(0.1, 0.1, 0.4), (0.1, 0.4, 0.1), (0.4, 0.1, 0.1), (0.3, 0.3, 0.3)\}$.

For each realization, the estimation methods used to obtain $\hat{\boldsymbol{\theta}} = [\hat{\alpha}_1, \ldots, \hat{\alpha}_p, \hat{\mu}_e]^T$ are Yule-Walker (YW), Conditional Least Squares (CLS), Whittle (WHT) and Least Squares using HOS (LS_HOS). For a detailed description of the YW, CLS and WHT estimation methods see Silva (2005). The minimizations necessary in the methods CLS, WHT and LS_HOS are performed through the MATLAB function *fminunc*, which finds a minimum of a scalar unconstrained multivariable function by using the BFGS Quasi-Newton method with a mixed quadratic and cubic line search procedure (MathWorks (2004)). The initial values of the iterative methods (CLS, WHT and LS_HOS) are the YW estimates. For each case, the mean bias, variance and mean square error are evaluated.

With respect to the small sample properties of the LS_HOS estimator, the following conclusions can be drawn from the analysis of all the simulations. In general, the sample bias, variance and mean square error decrease as the sample size increases, indicating that the distribution of the estimators is consistent and symmetric. However, for a small sample size there is evidence of departure from symmetry in the marginal distributions, specially for values of the parameter near the non-stationary region.

When the several estimation methods are compared it is found that the LS_HOS provides similar results, in terms of the smallest values of sample

bias, variance and mean square error, to the other methods. It is also verified that, in general, the proportion of non-admissible estimates of the methods is less for LS_HOS, followed by WHT and CLS. In order to illustrate some of these conclusions, Figure 1 shows the boxplots of the sample bias for the estimates obtained from 50 and 200 observations of the INAR(1) process with parameter values $(\alpha_1, \lambda) = (0.9, 1.0)$. Note that the value of α is near the non-stationary region, however, even for $N = 50$ observations, the LS_HOS estimates presents the best results.

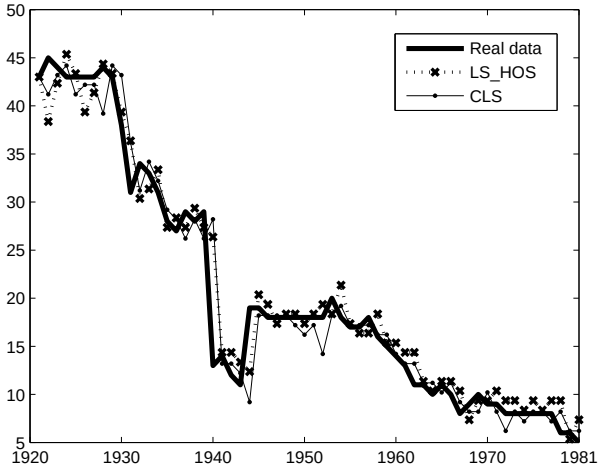


Fig. 2. The number of Swedish mechanical paper and pulp mills, from 1921 to 1981, and the fitted values considering the LS_HOS and CLS estimates.

Figure 2 presents the number of Swedish mechanical paper and pulp mills, from 1921 to 1981, used by Brännäs (1995) and Brännäs and Hellström (2001). These authors fitted an INAR(1) process to this dataset using some explanatory variables. Here, an INAR(1) process where the innovations are i.i.d. with mean μ_e and variance σ_e^2 is considered. Since the mean of the data is 20.40 and its variance is 155.16, a Poisson innovation process is not assumed but then the method does not require that or anyother assumption on the distribution of the innovations. Table 1 presents the parameter estimates obtained by CLS and LS_HOS methods. The fit of both models, based on LS_HOS and CLS estimates, are also shown in Figure 2. The mean square errors (MSE) between the observations and the fitted values are exhibited in Table 1. It can be seen that the MSE is slightly smaller for the LS_HOS fit than for CLS fit. The last two columns of the table present the mean and

variance of the estimated models:

$$\hat{\mu}_x = \frac{\hat{\mu}_e}{1-\hat{\alpha}} \quad \text{and} \quad \hat{\sigma}_x^2 = \frac{(1-\hat{\alpha})(\hat{\mu}_e\hat{\alpha}+\hat{\sigma}_e^2)}{(1-\hat{\alpha})^2(1+\hat{\alpha})}.$$

It is noticeable that the model estimated by LS_HOS presents mean and variance closer to the sample values. The residuals from both fitted models are uncorrelated.

Method	$\hat{\alpha}$	$\hat{\mu}_e$	$\hat{\sigma}_e^2$	MSE	$\hat{\mu}_x$	$\hat{\sigma}_x^2$
CLS	0.9591	0.2017	15.2268	8.5494	4.9315	192.2764
LS_HOS	0.9269	1.3635	19.2253	7.4465	18.6525	145.4513

Table 1. The parameter estimates of the number of Swedish mechanical paper and pulp mills, from 1921 to 1981.

4 Final remarks

The principal advantage of HOS is the capability to detect and characterize the deviations from Gaussianity and non-linearity of the processes. Thus in this work a new estimation method for the parameters of INAR processes based on HOS is proposed. This method uses the Least Squares estimation to minimize the errors between the third-order moment of the observations and of the fitted model. A Monte Carlo study indicates that this estimation method provides good results in small samples, in terms of sample bias, variance and mean square error. Moreover, when used in the context of a non-Poisson real dataset the LS_HOS estimates provide a model with mean, variance and autocorrelations closer to the sample values.

Acknowledgments

For the first author, this work reports research developed under financial support provided by FCT - Fundação para a Ciência e Tecnologia, Portugal.

References

BRÄNNÄS, K. (1995): Explanatory variables in the AR(1) count data model. *Umeå Economic Studies 381.*

BRÄNNÄS, K. and HELLSTRÖM, J. (2001): Generalized integer-valued autoregression. *Econometric Reviews 20 (4), 425-443.*

GAUTHIER, G. and LATOUR, A. (1994): Convergence forte des estimateurs des paramtres dun processus GENAR(p). *Annales des Sciences Mathématiques du Québec 18, 49-71.*

LATOUR, A. (1998): Existence and stochastic structure of a non-negative integer-valued autoregressive process. *Journal of Time Series Analysis 19, 439-455.*

MATHWORKS (2004): Optimization toolbox user's guide for MATLAB. Available from `http://www.mathworks.com/access/helpdesk/help/pdf_doc/optim/optim_tb.pdf`

SILVA, I. (2005): *Contributions to the analysis of discrete-valued time series.* PhD Thesis. Universidade do Porto, Portugal.

SILVA, M.E. and OLIVEIRA, V.L. (2004): Difference equations for the higher-order moments and cumulants of the INAR(1) model. *Journal of Time Series Analysis 25, 317-333.*

SILVA, M.E. and OLIVEIRA, V.L. (2005): Difference equations for the higher-order moments and cumulants of the INAR(p) model. *Journal of Time Series Analysis 26, 17-36.*

STEUTEL, F.W. and VAN HARN, K. (1979): Discrete analogues of self-decomposability and stability. *The Annals of Probability 7, 893-899.*

Forecasting in INAR(1) Model

Nélia Silva[1], Isabel Pereira[1] and M. Eduarda Silva[2]

[1] Departamento de Matemática, Universidade de Aveiro & UIMA
Campus de Santiago, 3810-193 Aveiro, Portugal,
neliasilva@ua.pt, isabel.pereira@ua.pt
[2] Faculdade de Economia, Universidade do Porto, & UIMA
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal *mesilva@fep.up.pt*

Abstract. In this work we consider the problem of forecasting integer-valued time series modelled by the INAR(1) process introduced by McKenzie (1985) and Al-Osh and Alzaid (1987). The theoretical properties and practical applications of INAR and related processes have been discussed extensively in the literature but there is still some discussion on the problem of producing coherent, ie, integer predictions. Here Bayesian methodology is used to obtain point predictions as well as confidence intervals for future values of the process. The predictions thus obtained are compared with their classic counterparts. The proposed approaches are illustrated with a simulation study and a real example.

Keywords: INAR models, Bayesian prediction, integer prediction, Markov Chain Monte Carlo algorithm

1 Introduction

In this work we are interested in a special class of observation-driven models, the so-called integer-valued autoregressive (INAR) process introduced by McKenzie (1985) and Al-Osh and Alzaid (1987). The theoretical properties and practical applications of INAR and related processes have been discussed extensively in the literature. Silva *et al.* (2005) consider independent replications of count time series modelled by INAR(1) and proposed several estimation methods using the classical and Bayesian approaches in time and frequency domains. Usually the forecast values are obtained from the conditional expectations, which have the optimal property but rarely generate integer values. In order to produce coherent forecasts Freeland and McCabe (2004) use the median of the k-step-ahead conditional distribution to emphasize the intention of preserving the integer structure of the data in generating the forecasts. McCabe and Martin (2005) develop a general methodology for producing coherent predictions of low count data.

The main purpose of this paper is to obtain coherent forecasts for the Poisson INAR(1) process. Bayesian methodology is used to obtain point predictions as well as credibility intervals for future values of the process which are compared with their classic counterpart. Section 2 provides the theoretical results in order to obtain the point forecasts. Section 3 presents methods

for producing confidence intervals or highest posterior predictive density intervals for forecasts. In Section 4 we conduct a simulation study to compare the performance of the classical and Bayesian approaches, considering point and interval predictions. In Section 5 the proposed methodology is applied to a data set and compared with classical inference and forecasting procedures of Freeland (1998).

2 Point prediction

Consider a non negative integer-valued random variable X and $\alpha \in [0,1]$. The generalized thinning operation, hereafter denoted by '$\circ$', is defined as $\alpha \circ X = \sum_{j=1}^{X} Y_j$, where $\{Y_j\}$, $j = 1, \ldots, X$, is a sequence of independent and identically distributed non-negative integer-valued random variables, independent of X, with finite mean α and variance σ^2. The well-known INAR(1) process $\{X_t; t = 0, \pm 1, \pm 2, \ldots\}$ is defined on the discrete support $\mathbb{N}_0$ by the equation

$$X_t = \alpha \circ X_{t-1} + \epsilon_t,$$

where $0 < \alpha < 1$, $\{\epsilon_t\}$ is a sequence of independent and identically distributed integer-valued random variables, with $E[\epsilon_t] = \mu_\epsilon$ and $Var[\epsilon_t] = \sigma_\epsilon^2$.
In this paper we consider only Poisson INAR(1) process, denoted henceforth as PoINAR(1), where $\{\epsilon_t\}$ is a sequence of independent Poisson distributed variables with parameter λ, independent of the counting series $\{Y_j\}$. Note that, assuming $\epsilon_t \frown Po(\lambda)$ it is straightforward to show that $X_t \frown Po(\lambda/(1-\alpha))$. Given an observed series up to time n, $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$, the most common procedure for constructing predictions for time series is to use conditional expectations. Now, the probability function of $X_{n+h}|X_n$ is given by (Freeland, 1998, pp. 30)

$$\begin{aligned} f(x_{n+h}|x,\alpha,\lambda) = \exp\left\{-\lambda\tfrac{1-\alpha^h}{1-\alpha}\right\} \sum_{i=0}^{M_h} \tfrac{1}{(x_{n+h}-i)!} \times \\ \times \left(\lambda\tfrac{1-\alpha^h}{1-\alpha}\right)^{x_{n+h}-i} \binom{x_n}{i} (\alpha^h)^i (1-\alpha^h)^{x_n-i}, \quad x_{n+h} = 0, 1, \ldots, \end{aligned} \tag{1}$$

where $M_h = \min(X_{n+h}, X_n)$. Consequently, $E[X_{n+h}|X_n, \alpha, \lambda] = \alpha^h \left[X_n - \frac{\lambda}{1-\alpha}\right] + \frac{\lambda}{1-\alpha}$, $h = 1, 2, 3, \ldots$ and as $h \longrightarrow +\infty$, $X_{n+h}|X_n, \alpha, \lambda$ is Poisson distributed with parameter $\lambda/(1-\alpha)$.

2.1 Classical methodology

The h-step-ahead predictor based on the conditional expectation of INAR(1),

$$\hat{X}_{n+h}|\mathbf{x}_n = E[X_{n+h}|X_n] = \alpha^h \left[X_n - \frac{\lambda}{1-\alpha}\right] + \frac{\lambda}{1-\alpha},\ h = 1, 2, 3, \ldots \tag{2}$$

was obtained by Freeland and McCabe (2004). In order to obtain coherent predictions for X_{n+h}, Freeland and McCabe (2004) suggest minimizing the expected absolute error given the sample, $E[|X_{n+h} - \hat{X}_{n+h}| \Big| X_n]$, concluding that $\hat{X}_{n+h} = \hat{m}_{n+h}$ is the median of the h-step-ahead conditional distribution $f(x_{n+h}|x_n)$.

2.2 Bayesian methodology

The Bayesian predictive probability function is very simple to understand: it is based on the assumption that both, the future observation, X_{n+h} and the vector of unknown parameters $\boldsymbol{\theta} = (\alpha, \lambda)$ are random.

Definition 1 *The h-step-ahead Bayesian posterior predictive distribution is given by*

$$f(x_{n+h}|\mathbf{x}_n) = \int_{\Theta} f(x_{n+h}|\mathbf{x}_n; \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x}_n)d\boldsymbol{\theta}, \tag{3}$$

where $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$ is the posterior probability function of $\boldsymbol{\theta}$ and $f(x_{n+h}|\mathbf{x}_n; \boldsymbol{\theta})$ is the predictive distribution (classical) given by (1).

The Bayesian h-step-ahead predictor is given by the expected valued, the median or the mode of X_{n+h} given $\mathbf{x}_n$, computed from $f(x_{n+h}|\mathbf{x}_n)$. To compute the posterior probability function of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$, we use the beta and gamma as prior distributions of the parameters to INAR(1) model, $\alpha \frown Beta(a, b)$, $a, b > 0$ and $\lambda \frown Gamma(c, d)$, $c, d > 0$ since these are conjugate of binomial and Poisson distributions, respectively, and also consider α and λ independent. If $M_t = \min(X_{t-1}, X_t)$, the Bayesian predictive function of X_{n+h} given $\mathbf{x}_n$ is given by,

$$\begin{aligned} f(x_{n+h}|\mathbf{x}_n) \propto & \int_{\alpha}\int_{\lambda} \sum_{i=0}^{M_h} \binom{x_n}{i} (\alpha^h)^i (1-\alpha^h)^{x_n - i} \frac{1}{(x_{n+h}-i)!} \times \\ & \times \exp\left(-\lambda \tfrac{1-\alpha^h}{1-\alpha}\right) \left(\lambda \tfrac{1-\alpha^h}{1-\alpha}\right)^{x_{n+h}-i} \exp\left[-(d+n)\lambda\right] \lambda^{c-1} \\ & \times \alpha^{a-1}(1-\alpha)^{b-1} \prod_{t=2}^{n} \sum_{i=0}^{M_t} \frac{\lambda^{x_t - i}}{(x_t - i)!} \binom{x_{t-1}}{i} \alpha^i (1-\alpha)^{x_{t-1}-i} d\alpha d\lambda \,. \end{aligned} \tag{4}$$

In order to estimate X_{n+h}, we can adapt the Tanner composition method (Tanner, 1996) to the integer case . After sampling $X_{n+h,1}, \ldots, X_{n+h,m}$, the h-step-ahead predictor of X_{n+h}, can be calculated from sample mean ($\hat{X}_{n+h}$), median ($\hat{m}_{n+h}$) or mode ($\hat{mo}_{n+h}$).

Another approach is to calculate $E(X_{n+h}|\mathbf{x}_n)$ using $E(X_{n+h}|\mathbf{x}_n) = E[E(X_{n+h}|\boldsymbol{\theta}, \mathbf{x}_n)|\mathbf{x}_n] = X_n E\big[\alpha^h|\mathbf{x}_n\big] + E\left[\frac{1-\alpha^h}{1-\alpha}\lambda \Big| \mathbf{x}_n\right]$. These expected values can be estimated through Metropolis algorithm in conjunction with Adaptive Rejection Sampling Method; thus the predictor is

$$\tilde{X}_{n+h} = X_n \left(\frac{1}{m}\sum_{i=1}^{m} \alpha_i^h\right) + \left(\frac{1}{m}\sum_{i=1}^{m} \frac{1-\alpha_i^h}{1-\alpha_i}\lambda_i\right). \tag{5}$$

3 Interval prediction

3.1 Classical methodology

A confidence interval for the predictor $\hat{X}_{n+h}$ can be calculated through the probability function of the h-step-ahead prediction error, $e_{n+h}|\mathbf{x}_n = X_{n+h}|\mathbf{x}_n - \hat{X}_{n+h}|\mathbf{x}_n$, which is given by

$$P(e_{n+h} = k - \alpha^h x_n - \lambda\tfrac{1-\alpha^h}{1-\alpha}|\mathbf{x}_n) = P(X_{n+h} = k|X_n = x_n) = \exp\left\{-\lambda\frac{1-\alpha^h}{1-\alpha}\right\}\sum_{i=0}^{M_h}\frac{1}{(k-i)!}\left(\lambda\frac{1-\alpha^h}{1-\alpha}\right)^{k-i}\binom{x_n}{i}(\alpha^h)^i(1-\alpha^h)^{x_n-i}. \tag{6}$$

From the expression (6) we obtain a γ level confidence interval for X_{n+h}

$$(\hat{X}_{n+h} + e_{t_1}, \hat{X}_{n+h} + e_{t_2}), \tag{7}$$

where $\hat{X}_{n+h}$ is given by (2), e_{t_1} is the largest value of e_{n+h} such as $\mathrm{P}(e_{n+h} \leq e_{t_1}) \leq (1-\gamma)/2$ and e_{t_2} is the smallest value of e_{n+h} such as $\mathrm{P}(e_{n+h} \leq e_{t_2}) \geq (1+\gamma)/2$.

3.2 Bayesian methodology

We propose an adaptive generalization of the method used to obtain HPD (*Highest Posterior Density*) intervals of the model parameters, in which we consider the predictive distribution instead of the posterior.

Definition 2 *$R(\gamma) = (X_L, X_R)$ is a $100\gamma\%$ HPD interval for X_{n+h} if*

$$P(X_L \leq X_{n+h} \leq X_R) = \sum_{x_{n+h}=X_L}^{X_R} f(x_{n+h}|\mathbf{x}_n) \geq K_\gamma, \tag{8}$$

where K_γ is the largest constant such that $P[X_{n+h} \in R(\gamma)] \geq \gamma$. Due to complexity of the predictive probability function given by (4) it is not possible to calculate the exact HPD interval for X_{n+h}; we can give an approximation for $R(\gamma)$ by using the Chen and Shao algorithm, (see Chen *et al.*,2000).

4 Simulation study

For the simulation study we consider samples of size $n = 40, 90, 190$, generated by INAR(1) models with the parameters values $\alpha = 0.2, 0.5, 0.8$ and $\lambda = 1, 3$ and considering the hyperparameters $a = b = c = d = 10^{-4}$ (a vague prior distribution).

4.1 Point prediction

From the various simulated samples we conclude that large values of α and λ are related with high dispersion values. Consequently the increase in α and λ provides large values of $|x_{n+h} - x_n|, h > 1$. Independently of prediction methodology used, the forecasts performance depends on two basic aspects: one is the difference between x_n and $x_{n+h}, h > 1$; the other is the approximation between x_n and $\hat{\lambda}/(1-\hat{\alpha})$, in particular with the increase in h (note that $\hat{E}(X_{n+h}|X_n) \to \hat{\lambda}/(1-\hat{\alpha}), h \to \infty$). These findings are illustrated in Table 1 where point predictions for 10 steps ahead for a particular model are given. The table includes the h-step ahead simulated and predicted values and the absolute deviations between x_{190} and $x_{190+h}, h = 1, \ldots, 10$. The last line contains the classical limiting distribution.

To confront classical and Bayesian methodologies we use the mean square error (MSE) to compare means, the mean absolute deviation (MAD) to compare medians and the "everything or nothing" lost function (FPTN), given by $1/n \sum I(x_{n+h})$ where $I(x_{n+h}) = \begin{cases} 1 \text{ if } |\hat{x}_{n+h} - x_{n+h}| > \delta \\ 0 \text{ if } |\hat{x}_{n+h} - x_{n+h}| \leq \delta \end{cases}$ to compare modes. We consider $\delta = 1$ since we have integer values.

Table 2 shows the MSE, MAD and FPTN values from 10 one-step-ahead predictions. Values of MSE($\tilde{X}_{n+h}$) are obtained considering the Bayesian predictors given by (5) (values of MSE($\hat{X}_{n+h}$ are similar). Values of MAD and FPTN were calculated, respectively, through medians and modes. The indices "C" or "B" indicate which methodology is used (classical or Bayesian, respectively). As we can see, when $\alpha = 0.8$ Bayesian methodology provides smaller values than classical methodology, so the Bayesian predictions seems to have a better performance than classical predictions.

In order to study and compare the estimates given by the sample mean, sample median and sample mode we used the mean absolute percentual error (MAPE), given by $1/H \sum_{h=1}^{H} |\hat{X}_{n+h} - X_{n+h}|/X_{n+h}$, where H represents the number of predictions realized. This criteria does not benefit any measure (mean, median or mode) in particular. The results are illustrated in Table 3 for three samples with sizes 40, 90 and 190 of the model $x_t = \alpha \circ x_{t-1} + \epsilon_t, \epsilon_t \frown P(3)$. MAPE minimum is always obtained with Bayesian approach. Similar results are obtained for $\lambda = 1$.

4.2 Interval prediction

The performance of the intervals (with 95 % of confidence or credibility) obtained by each approach is measured by the amplitudes and coverage probabilities. The coverage probability is estimated through the correspondent frequency of the future observation $x_{n+i} \in R_i(\gamma)$, $i = 1, 2, \ldots, h$, considering 100 replicates. Prediction intervals for future observations were calculated using expression (7) for classical methodology and Chen and Shao algorithm

for Bayesian methodology. The simulation results for the case $\lambda = 3, \gamma = 0.95$ and n=100, are presented in Table 4. Table 4 indicates that when $\alpha = 0.2$ the bayesian intervals have large coverage probability; it can be noted that, when $h > 2$ and for small values of α and λ the classic intervals have small amplitudes and they are coincident with those by considering the asymptotic distribution. Moreover, it is worthwhile to mention that when $\alpha \geq 0.5$ the mean amplitudes of the prediction intervals obtained by bayesian methodology are smaller than their classic counterparts.

5 Analysis of burn claims data

We apply the proposed methodology to a data set analysed by Freeland (1998) comprising 120 monthly counts of workers collecting Wage Loss Benefits (WLB) for burn injuries. All the descriptive details of the data set can be found in Freeland (1998) who concludes that PoINAR(1) is a plausible choice for modelling the data. In order to evaluate and compare the different prediction methodologies, h-step ahead forecasts ($h = 1, 2, 3, 4, 5, 6$) are produced for the time period from July to December 1994, for which we know the observed values. The point forecasts based on the mean, median and mode and the observed values are presented in Table 5. In general, it can be noted that MAPE values of classic point predictions are smaller than those of Bayesian predictions. This result is expected in view of the simulation results presented in the last section since the estimated value for alpha is 0.4.

Interval predictions for the period July to December 1994 are obtained using the two approaches proposed given by (7). The intervals obtained, presented in Table 5, are analogous, although the Bayesian have smaller width.

6 Final remarks

Forecasting low integer values of time series of counts remains an open problem. Although conditional means do not preserve coherently the integer nature of the data, it seems there is no advantage in using median or mode values of the predictive distribution. Simulations indicate that the performance of the different approaches depend on the parameters of the model and that Bayesian methodology provides the best results when MAPE statistic is used.

References

AL-OSH, M.A. and ALZAID, A.A.(1987): First-Order Integer-Valued Autoregressive (INAR(1)) process. *Journal of Time Series Analysis, 8, 261-275.*

CHEN, M-H., SHAO, Q-M and IBRAHIM, J.G. (2000): *Monte Carlo Methods in Bayesian Computation.* Springer Series in Statistics.

FREELAND, R.K. (1998): *Statistical Analysis of Discrete Time Series With Application to the Analysis of Workers Compensation Claims Data*. Ph.D. thesis. The University of British Columbia, Canada.

FREELAND, R.K. and MCCABE, B.P.M. (2004): Forecasting Discrete Valued Low Count Series. *International Journal of Forecasting, 20, 427-434*.

GILKS, W.R. and BEST, N.G. (1987): Adaptive Rejection Metropolis Sampling Within Gibbs Sampling. *Applied Statistics, 44, 455-472*.

MCCABE, B.P.M. and MARTIN, G.M. (2005): Bayesian Predictions Of Low Count Time Series. *International Journal of Forecasting , 21, 315-330*.

MCKENZIE, E. (1985): Some ARMA Simple for Discrete Variate Time Series. *Water Resources Bulletin, 21, 645-650*.

SILVA, I., SILVA, M.E., PEREIRA, I. and SILVA, N. (2005): Replicated INAR(1) Process. *Methodology and Computing in Applied Probability, 7, 517-542*.

TANNER, M.A. (1996): *Tools for Statistical Inference*. 3rd. Springer Verlag, New York.

Table 1. Point predictions considering a sample of size n=190 with parameters $(\lambda = 3, \alpha = 0.8, x_{190} = 16)$.

$(\lambda = 3, \alpha = 0.8; x_{190} = 16))$						
			classical approach		Bayesian approach	
h	x_{190+h}	jump	$\hat{x}_{190+h}$	$\lvert x_{190+h} - \hat{x}_{190+h}\rvert$	$\hat{x}_{190+h}$	$\lvert x_{190+h} - \hat{x}_{190+h}\rvert$
1	16	0	15.477	0.523	15.530	0.470
2	16	0	15.084	0.916	15.010	0.990
3	16	0	14.787	1.213	14.678	1.322
4	20	4	14.564	5.436	14.574	5.426
5	19	3	14.396	4.604	14.408	4.592
6	17	1	14.270	2.730	14.516	2.484
7	18	2	14.175	3.825	14.128	3.872
8	19	3	14.103	4.827	14.182	4.818
9	20	4	14.049	5.951	14.066	5.934
10	17	1	14.008	2.993	13.986	3.014
∞			13.884			

Table 2. Values of MSE, MADE and FPTN considering 10 one-step-ahead predictions for the model $x_t = \alpha \circ x_{t-1} + \epsilon_t,\ \epsilon_t \frown \mathrm{P}(3)$ and sample sizes 40, 90 and 190.

	α	0.2			0.8		
	n	40	90	190	40	90	190
MSE	$\hat{X}_{n+h,C}$	6.68	2.01	**5.56**	12.98	3.82	16.17
	$\hat{X}_{n+h,B}$	**6.46**	**1.93**	6.06	**5.22**	**3.05**	**3.57**
MAD	$\hat{m}_{n+h,C}$	2.27	1.18	**2.00**	3.18	1.45	3.81
	$\tilde{m}_{n+h,B}$	**2.00**	1.18	2.05	**1.82**	**1.27**	**1.68**
FPTN	$\hat{mo}_{n+h,C}$	**0.45**	0.45	**0.45**	0.73	0.55	1.00
	$\tilde{mo}_{n+h,B}$	0.55	**0.36**	0.64	**0.36**	**0.36**	**0.55**

Table 3. Values of MAPE considering 10 one-step-ahead predictions for the model $x_t = \alpha \circ x_{t-1} + \epsilon_t,\ \epsilon_t \frown \mathrm{P}(3)$ and sample sizes 40, 90 and 190.

α	0.2			0.8		
n	40	90	190	40	90	190
$\hat{X}_{n+h,C}$	0.714	0.573	0.870	0.707	0.564	0.919
$\tilde{X}_{n+h,B}$	**0.178**	0.137	**0.260**	0.110	0.109	**0.081**
$\hat{m}_{n+h,C}$	0.652	0.588	**0.631**	**0.606**	0.561	**0.831**
$\tilde{m}_{n+h,B}$	0.187	**0.120**	0.209	0.115	**0.103**	0.091
$\hat{mo}_{n+h,C}$	**0.619**	**0.464**	0.929	0.625	**0.506**	**0.831**
$\tilde{mo}_{n+h,B}$	0.187	0.127	0.231	**0.086**	0.125	0.098

Table 4. Coverage probability estimates and mean amplitudes of the intervals for the h-step-ahead future values, in INAR(1) model with $n = 100$ and $\lambda = 3$.

	$\alpha = 0.2$				$\alpha = 0.8$			
	cov.prob. estimates		mean ampl.		cov.prob. estimates		mean ampl.	
h	Classical	Bayes	Classical	Bayes	Classical	Bayes	Classical	Bayes
1	0.68	0.96	6.34	7.34	0.97	0.96	10.18	9.53
2	0.78	0.99	7.16	7.79	0.96	0.94	12.63	12.00
3	0.98	0.99	8.00	7.88	0.95	0.93	13.89	13.29
4	0.99	0.94	8.00	7.77	0.95	0.89	14.81	14.31
5	0.98	0.98	8.00	7.78	0.95	0.87	15.36	14.67
6	0.96	1.00	8.00	7.86	0.96	0.92	15.58	14.92
7	0.96	0.99	8.00	7.79	0.96	0.91	15.78	14.97
8	0.95	0.99	8.00	7.79	0.92	0.94	15.83	15.30
9	0.96	0.96	8.00	7.77	0.93	0.93	15.95	15.47
10	0.93	0.98	8.00	7.84	0.96	0.91	15.97	15.54

Table 5. h-step ahead predictions of monthly claims from July to December 1994.

			point pred.						interval pred.	
			class.			Bayes.			class.	Bayes.
h	year/month	claims of	$\hat{x}$	$\hat{m}$	$\hat{m}_0$	$\hat{x}$	$\hat{m}$	$\hat{m}_0$		
1	94/07	11	7.89	8	7	7.67	7	7	(2.1,13.0)	(3.0,13.1)
2	94/08	12	8.24	8	8	8.01	8	7	(2.0,14.0)	(3.0,14.0)
3	94/09	11	8.38	8	8	8.14	8	7	(2.0,14.0)	(3.0,14.0)
4	94/10	12	8.44	8	8	8.21	8	7	(2.0,15.0)	(3.0,14.0)
5	94/11	7	8.46	8	8	8.22	8	7	(2.0,15.0)	(3.0,14.0)
6	94/12	11	8.47	8	8	8.22	8	7	(2.0,15.0)	(3.0,14.0)
		MAPE	0.26	0.27	0.27	0.27	0.27	0.32		

A Fuzzy Trend Model for Analyzing Trend of Time Series

Norio Watanabe[1] and Masami Kuwabara[2]

[1] Department of Industrial and Systems Engineering, Chuo University
Kasuga 1-13-27, Bunkyo-Ku, Tokyo 112-8551, Japan,
watanabe@indsys.chuo-u.ac.jp
[2] Graduate School of Chuo University
Kasuga 1-13-27, Bunkyo-Ku, Tokyo 112-8551, Japan

Abstract. A fuzzy trend model is proposed for analyzing trend of time series. This model is a simple time series model based on a fuzzy system composed of fuzzy if-then rules. Applicability of the model is shown by some examples and simulation studies. The fuzzy trend model provides a new framework for trend analysis.

Keywords: time series analysis, fuzzy system, fuzzy if-then rule

1 Introduction

A fuzzy trend model was proposed by Watanabe and Kuwabara (2005) and related topics were discussed by Kuwabara and Watanabe (2006). Their model is specified for a long-term financial time series such as the returns of stock price indices. In this paper we extend a fuzzy trend model to a general time series model for analyzing trend of time series. The fuzzy trend model is a simple model based on fuzzy system composed of fuzzy if-then rules. The usage of fuzzy sets is only for describing a system and the fuzzy trend model is a usual stochastic model.

Standard methods for trend analysis are the regression analysis and moving average method. Our model can be represented by a form of a regression model but is not a usual regression model. And our model has a mechanism of the weighted sum similar to moving average. In our model, however, weighted sum of unobserved series is considered. Fuzzy sets are used for setting weights.

Many smoothing methods including the locally weighted regression and smoothing scatterplots by Cleveland (1979) can be applied for trend analysis also. Our approach is model based differently from usual smoothing methods.

There are some stochastic trend models including ARIMA and state space models. Our model is simpler but much flexible than those models for capturing the trend.

In this paper we use the term "trend" as the mean value function of the time series. We assume that the time series can be decomposed into the trend and the zero mean stationary process. The purpose is the inference

on the trend. When the seasonal component exists, we regard the seasonal component as a part of the trend. In our model the trend can be either deterministic or stochastic. When the trend is deterministic, it is shown that a fuzzy trend model can provide a piecewise linear function which approximates the trend.

The aim of this paper is to propose the fuzzy trend model as a general tool for trend analysis. In Section 2 we introduce the model. A basic property of our model is discussed briefly in Section 3. We show applicability of the model by examples and simulation studies in Sections 4 and 5.

2 Fuzzy trend model

Let $\{y_n|n = 1, 2, ..., N\}$ denote the observed time series. The fuzzy trend model for $\{y_n|n = 1, 2, ..., N\}$ is given by

$$y_n = \mu_n + x_n \tag{1}$$

$$\mu_n = \sum_{k=1}^{K} \nu_k(n)\mu_n(k), \tag{2}$$

$$R_k \ : \text{ If } n \text{ is } A_k, \text{ then } \mu_n(k) = u(k), \qquad (k = 1, ..., K) \tag{3}$$

under the assumptions:

(A1) $\{x_n\}$ is a stationary stochastic process whose mean is zero and variance is σ_x^2,
(A2) $\{u(k)\}$ is the unobserved deterministic or stochastic process, and
(A3) the joint distribution of $\{x_n|n = 1, ..., N\}$ is the same as the conditional distribution of $\{x_n|n = 1, ..., N\}$ for given $\{u(k)|k = 1, ..., K\}$,

where R_k is the fuzzy if-then rule and ν_k is the membership function of the fuzzy set A_k. On fuzzy set A_k we assume that

(A4) the entire set of A_k is the set $\{t|0 \le t \le N_{\max}\}$, where $N \le N_{\max}$, and the membership functions ν_1,..., ν_K satisfy the equation:

$$\sum_{k=1}^{K} \nu_k(t) = 1 \quad \text{for} \quad 0 \le \ \forall t \le N_{\max}. \tag{4}$$

We do not refer to the fuzzy set theory (see Negoita and Ralescu (1975), for example), since the model given by (1)–(3) can be represented by a regression form $y = Bu + x$ (see Eq. (6)), where B is determined by ν_k. However, B is unknown differently from the usual regression model (see Grenander and Rosenblatt (1957) and Vogelsang (1998), for example).

In this paper we use the membership function ν_k which has the form of the left graph in Fig. 1, where the width parameter d is an integer larger than one, $a_k = (k - 1)d$ and the shape parameter c satisfies $0 \le c \le 1$. Typical

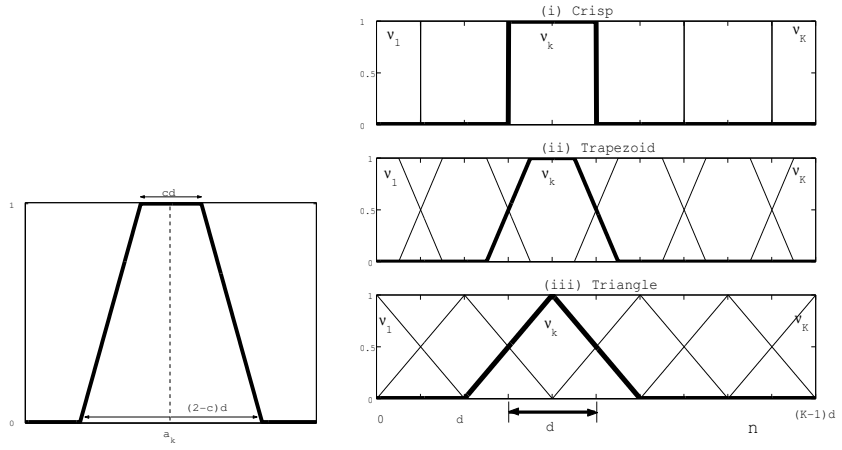


Fig. 1. Membership function ν_k.

examples of $\{\nu_1, .., \nu_K\}$ are shown by the right graph in Fig. 1, where (i) $c = 1.0$, (ii) $c = 0.5$ and (iii) $c = 0.0$.

In the case of above membership functions the number of rules or length of the process $\{u(k)\}$ is determined from N as follows:

$$K = \left[\frac{N + d - 1}{d}\right] + 1, \tag{5}$$

where $[t]$ means the minimum integer that is greater than or equal to t, for $2 \leq d \leq N$. This means that K is determined by d. The integer d is unknown and then K is also unknown. For $d > N$ we set $K = 1$, which implies that μ_n is constant.

The input-output system given by (2) and (3) is a special case of the Takagi-Sugeno's fuzzy system (Takagi and Sugeno (1985)), which is a well-known model in the fuzzy control theory. The output is the trend $\{\mu_n\}$ and the input is the unobserved process $\{u(k)\}$. The output $\{\mu_n\}$ is given by a weighted sum of the input and weights are determined by membership values.

For given $\{u(k)\}$ the first term $\{\mu_n\}$ is the conditional mean function of $\{y_n\}$. Note that $\{\mu_n\}$ and $\{x_n\}$ are independent from the condition (A3), when $\{u(k)\}$ is stochastic. Thus μ_n can be regarded as a trend. In Section 4 we show examples of trend models.

The feature of the fuzzy trend model is the existence of the latent process $\{u(k)|k = 1, ..., K\}$. The analysis on the trend can be achieved by investigating the latent process $\{u(k)\}$.

For this purpose the most important problem is how to identify the model. The first step of the identification is to determine d and c from the observed series $\{y_1, ..., y_N\}$. For fixed d and c the latent process $\{u(k)\}$ can be estimated by the least squares method, as follows.

Let $\hat{u}$ denote the least squares estimator of $u = (u(1), ..., u(K))'$ for fixed d, that is, fixed K. We can represent the model (1)–(3) as

$$y = Bu + x, \tag{6}$$

where $y = (y_1, ..., y_N)'$, $x = (x_1, ..., x_N)'$ and (n, k)-element of the $N \times K$ matrix B is $\nu_k(n)$. Then we have $\hat{u} = (\hat{u}(1), ..., \hat{u}(K))' = (B'B)^{-1}B'y$. Note that B is unknown and its size is also unknown. If d and c is given, then we can calculate B.

If the joint distribution of $\{x_1, ..., x_N\}$ is available, the maximum likelihood method can be applied also. And if we can assume that numbers of possible values of d and c are finite respectively, an information criterion such as AIC or BIC can be defined. When the trend structure is unknown, however, the structure of $\{x_n\}$ is also unknown generally. In this case the quasi AIC can be defined by assuming $\{x_n\}$ is the Gaussian white noise and using the least squares estimate of $\{u(k)\}$. At the first stage of time series analysis such an approach might be efficient.

3 Approximation by fuzzy trend model

We show a general ability of the fuzzy trend model with the triangle membership functions given by (iii) $c = 0$ in Fig. 1. We have the following result.

Theorem 1. Let $g(t)$ be the continuous piecewise linear function given by connecting K points $(0, u_1)$, $(d, u_2), ..., ((K-1)d, u_K)$ sequentially, where $u_1, ..., u_K$ are any real numbers for $K \geq 2$. Then $g(n)$ is equal to the n-th element of Bu for any positive integer $n \leq N$, where $u = (u_1, u_2, ..., u_K)'$ and B is determined from the triangular membership functions.

The proof is easy and we omit it.

Suppose that the trend is deterministic. Then we can regard the smooth curve given by connecting N points $(1, \mu_1), (2, \mu_2), ..., (N, \mu_N)$ as the trend curve. The above theorem implies that fitting the fuzzy trend model is equal to the approximation of the trend curve by a piecewise linear function.

Let $\mu = (\mu_1, ..., \mu_N)'$ denote the trend and put $\epsilon_\mu = \mu - Bu^*$, where

$$u^* = \arg\min_u (\mu - Bu)'(\mu - Bu). \tag{7}$$

Then we obtain

$$B\hat{u} = \mu - \epsilon_\mu + B(B'B)^{-1}B'(x + \epsilon_\mu), \tag{8}$$

where $\hat{u}$ is the least squares estimator, for fixed d. When N is large comparing to K, the last term is close to zero generally. If $||\epsilon_\mu||$ is sufficiently small, the above equation shows that the true trend μ can be approximated by $B\hat{u}$.

As a special case of approximation we can show that $K = 2$ realizes the linear trend. Note that $K = 1$ corresponds to the constant trend. We show examples in the next section.

Usually we have to determine d or K from time series data. An appropriate d or K will be selected by an information criterion as shown in the following section.

4 Examples and simulation

In this section we use the triangular membership functions ($c = 0.0$) and the Akaike's information criterion (AIC). And we assume that $\{x_n\}$ is the Gaussian white noise. Then AIC can be calculated by using the least squares estimator $\hat{u}$.

4.1 Deterministic trend

We fit the fuzzy trend model to the artificial data generated from the models:

$$y_n = 0.005n + x_n, \qquad x_n \sim \mathrm{NID}(0, 1), \tag{9}$$

and

$$y_n = 3 + 0.005n + \sin(\frac{n}{5\pi}) + x_n, \qquad x_n \sim \mathrm{NID}(0, 1), \tag{10}$$

where $N = 1000$. Examples of time series are shown by the upper left graphs in Figs. 2 and 3 We select d or K by the AIC. Note that $d > N$ and $d = N$

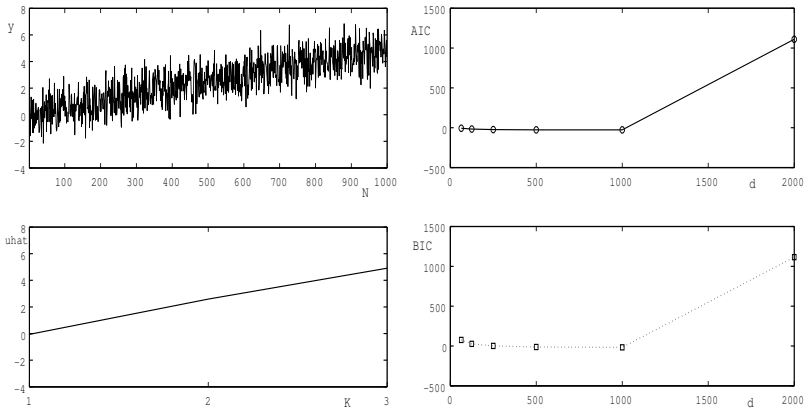


Fig. 2. Linear trend model (9).

correspond to $K = 1$ and $K = 2$ respectively. The upper right graphs show the values of AIC. (We also calculate the values of BIC, for reference.) The estimates $\hat{u}$ for the selected d are shown by the lower left graphs. In these cases $K = 2$ and $K = 51$ are selected respectively. It is found that fuzzy trend models are well identified.

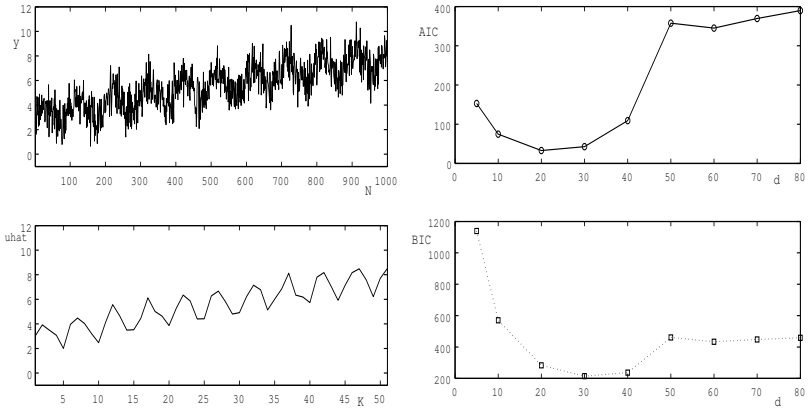


Fig. 3. Linear plus Sin curve model (10).

4.2 Stochastic trend

The typical example of stochastic trend is the random walk. The Fig. 4 shows a result for the model:

$$y = Bu + x, \qquad x_n \sim \mathrm{NID}(0, 0.8^2), \tag{11}$$

where $\{u(k)\}$ is the random walk such that $u(k) - u(k-1) \sim \mathrm{NID}(0, 1)$ and $N = 1000,\ d = 20,\ K = 51$. The middle left graph is the estimated trend and

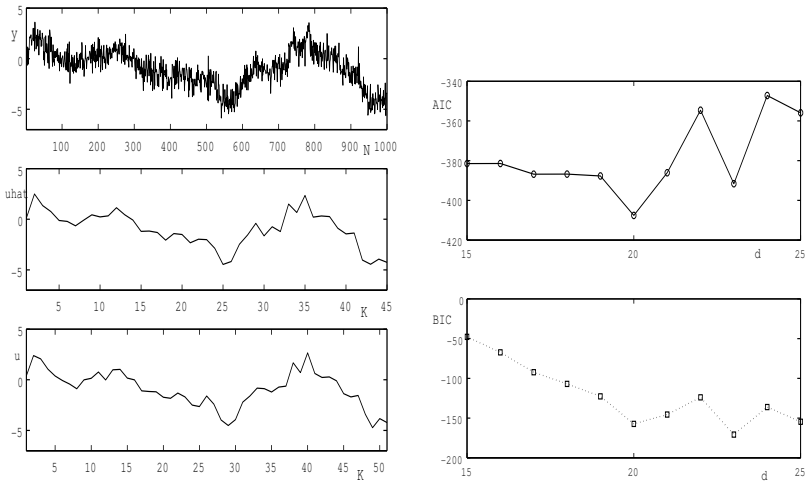


Fig. 4. Random walk plus white noise model (11).

the lower is the true trend. It is found that the trend is estimated adequately.

When the trend is stochastic, statistical inference on the latent process can be considered. In Watanabe and Kuwabara (2005) and Kuwabara and Watanabe (2006) the stationary AR model is assumed for $\{u(k)\}$, and they consider the estimation of the autocorrelation function of $\{u(k)\}$. Note that their model is a quite special case of our model.

5 Trend with change point

In this section we show that the fuzzy trend model can be applied to change point detection problems on trend. We assume that $\{u(k)\}$ has a sudden structural change. When $c = 1.0$, y_n has a sudden structural change also. When $c < 1.0$, the structural change of y_n is not sudden but gradually.

In Figs. 5 and 6 two examples of trends with one change point and two change points are illustrated. Fig. 7 shows the estimated (upper graphs) and true (lower graphs) trends. The left graphs are trends of the model with one change point and the right graphs are for two change points model. These results show that the fuzzy trend model detect structural changes successfully.

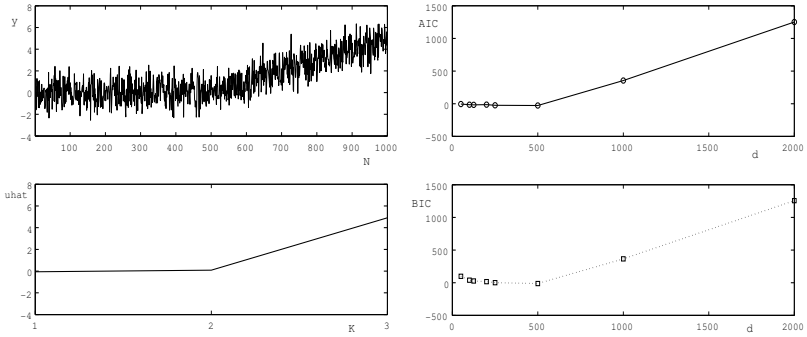


Fig. 5. Model with a change point.

6 Conclusion

In this paper we propose the fuzzy trend model for trend analysis. Simulation studies show the applicability of the model. The fuzzy trend model provides an alternative approach for trend analysis. A feature of the model is the existence of the latent process, which can be either deterministic or stochastic.

However simulation studies are limited. Further studies are required for various cases. Moreover statistical inference on the model should be studied in detail.

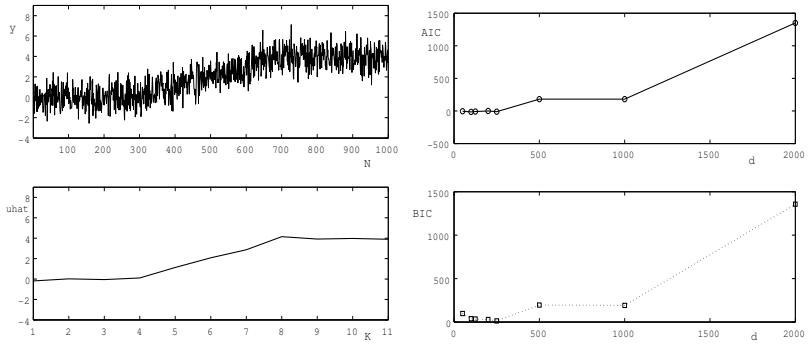


Fig. 6. Model with two change points.

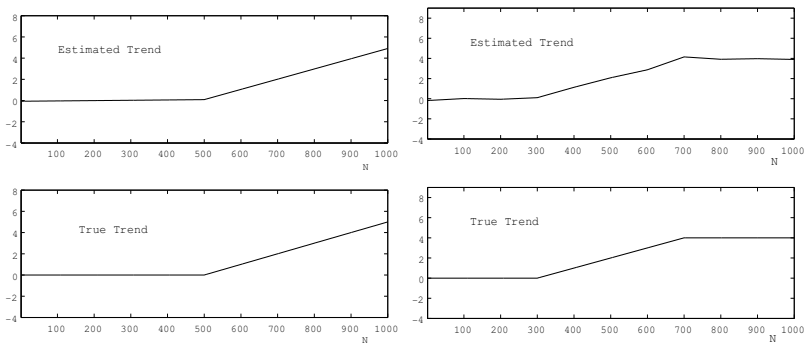


Fig. 7. Estimated and true trends (Left: one, Right: two).

References

CLEVELAND, W.S. (1979): Robust locally weighted regression and smoothing scatterplots. *Journal. of the American Statistical Association, 74, 829–836.*

GRENANDER, U. and ROSENBLATT, M. (1957): *Statistical Analysis of Stationary Time Series*, Chelsea Publishing Company.

KUWABARA, M. and WATANABE, N. (2006): A fuzzy trend model for long-term financial time series and its identification. *Proceedings of 2006 Annual Meeting of the North American Fuzzy Information Processing Society*, Montreal, Canada, 6 pages.

NEGOITA, C.V. and RALESCU, D.A. (1975): *Application of Fuzzy Sets to Systems Analysis*, Birkhäuser Verlag.

TAKAGI, T. and SUGENO, M. (1985): Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. on Systems, Man, and Cybernetics, 15(1), 116–132.*

VOGELSANG, T.J. (1998): Trend function hypothesis testing in the presence of serial correlation. *Econometrica, 66(1), 123–148.*

WATANABE, N. and KUWABARA, M. (2005): A fuzzy model for long-term financial time series. *Proceedings of the Ninth IASTED International Conference on Artificial Intelligence and Soft Computing*, Benidorm, Spain, 76–81.

Part XXIII

Index

Index

p

p

p

p

p