P. Brito (Editor)

Proceedings of COMPSTAT'2008

International Conference on Computational Statistics

Porto - Portugal, August 24th-29th 2008

Keynote and Invited 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 the Book of Proceedings hence correspond to the invited sessions, as follows:

Keynote
Advances on Statistical Computing Environments
Classification and Clustering of Complex Data
Computation for Graphical Models and Bayes Nets
Computational Econometrics
Computational Statistics and Data Mining Methods for Alcohol Studies (Interface session)
Finance and Insurance (ARS session)
Information Retrieval for Text and Images
Knowledge Extraction by Models
Model Selection Algorithms
Models for Latent Class Detection (IFCS session)
Multiple Testing Procedures
Random Search Algorithms
Robust Statistics
Signal Extraction and Filtering

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

Keynote

Nonparametric Methods for Estimating Periodic Functions, with Applications in Astronomy

Peter Hall

Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3130, Australia, *p.hall@ms.unimelb.edu.au*

Abstract. If the intensity of light radiating from a star varies in a periodic fashion over time, then there are significant opportunities for accessing information about the star's origins, age and structure. For example, if two stars have similar periodicity and light curves, and if we can gain information about the structure of one of them (perhaps because it is relatively close to Earth, and therefore amenable to direct observation), then we can make deductions about the structure of the other. Therefore period lengths, and light-curve shapes, are of significant interest. In this paper we briefly outline the history and current status of the study of periodic variable stars, and review some of the statistical methods used for their analysis.

Keywords: astronomy, curve estimation, light curve, local-linear methods, Nadaraya-Watson estimator, nonparametric regression, periodogram, stars

1 Introduction

1.1 Periodic variation arising in astronomy

Stars for which brightness changes over time are referred to, unsurprisingly, as variable stars. Some 31,000 such stars are known to exist, and at least another known 15,000 light sources are likely candidates. For many (although not all) such stars, brightness varies in a periodic, or approximately periodic, way. Moreover, stars of this type can often be observed with relatively unsophisticated equipment, for example with small telescopes, binoculars and even with the naked eye. The first variable stars were discovered by direct, unaided observation.

The pulsating star Mira, Latin for "the wonderful," was the first-discovered periodic variable star. It was recorded by David Fabricius, a German minister of religion, in 1596. At first he did not give it much of his attention, but when he noticed the star brighten during 1609 he realised that he had found a new type of light source.

The periodicity of Mira was established by Jan Fokkens Holwarda, a Dutch astronomer, who during 1638 and 1639 estimated the period to be about 11 months. Today we know that the length of the cycle is close to 331

days. For much of its cycle, Mira can be seen unaided. Its brightness varies from about magnitude 2 or 3 up to about 10, and then back again. (On the "magnitude" scale of star brightness, stars of higher magnitude are dimmer, or more difficult to see. Stars of magnitude 8 or larger are not visible to the naked eye.) The relative brightness of Mira, at least for much of its period, would have made it visible to astronomers in classical times.

Variable stars are catalogued into two broad classes — Intrinsic, for which the sources of variability lie within the star itself, and Extrinsic, where the variability comes, in effect, from the star's surface or from outside the star. About 65% of Intrinsic variable stars are "pulsating," and in those cases the brightness varies on account of cyclic expansions and contractions. Mira is of this type; it is a Long-Period Variable star, and stars in this category have periods of between a few days and several years.

Extrinsic variable stars are either Eclipsing Binaries or Rotating Variables. These sources of variation are perhaps the simplest for non-astronomers to understand. In the case of Eclipsing Binaries, one star rotates around the other, and when that star gets between its partner and the observer, the total amount of recorded light is reduced. When the two stars are well separated, as seen by the observer, the total amount of recorded light is maximised. The light emitted by a Rotating Variable star changes through the rotation of material on the star's surface.

This brief account of the nature of variable stars, and more specifically of periodic-variable stars, indicates that we often have only sketchy knowledge of the mechanisms that cause brightness to fluctuate. Even in the case of eclipsing binary stars, for which the nature of the mechanism is relatively clear, the extent of interaction between the two stars may be unknown. For example, mass can be transferred from one star to the other in an eclipsing binary system, although the scale of the transfer may be unclear.

Having a graph of star brightness, as a function of phase during the cycle, can give insight into the nature of these mechanisms within the star, or within the star system. Sometimes an understanding of the mechanisms can be gained for stars that are relatively close, and by comparing their brightness curves with those of distant stars we have an opportunity to gain information about the latter. It is therefore advantageous to have nonparametric estimators of brightness curves, which do not impose mathematical models that dictate the shape of the curve estimates.

1.2 Related literature in astronomy and statistics

Astronomers typically refer to a plot of the mean brightness of a periodic variable star, representing a function of phase during the time duration of a period, as the star's "light curve." Distinctions between the notion of a theoretical light curve, on which we have only noisy data, and an estimate of that curve based on the data, are generally not made. Likewise, the difference between the function on which the true light curve or its estimate are based,

and a graph of that function, is generally not remarked on. These issues should be borne in mind when reading the astronomy literature, and also when interpreting the discussion below.

Ways of explaining the mechanisms that lead to periodic variation in brightness are continuously under development; see Prigara (2007), for instance. Likewise, estimates and interpretations of the curves that represent this variation are constantly becoming available. For example, Norton et al. (2008) present and discuss the light curves of 428 such stars, of which only 68 of had previously been recognised as being of this type. Eyer and Cuypers (2000) predict that the GAIA space mission, expected to be launched by the European Space Agency in 2011, will be able to detect some 18 million variable sources, among them five million classic periodic variable stars and two to three million eclipsing binary systems. Thus, the potential scope of the research problems discussed in this paper is likely to expand rapidly.

Book-length accounts of of variable stars, their properties and their light curves, include those given by Hoffmeister et al. (1985), Sterken and Jaschek (1996), Good (2003), North (2005), Warner (2006) and Percy (2007). The MACHO project, where the acronym stands for MAssive Compact Halo Objects, includes a very large catalogue of light curves for variable stars. See Axelrod et al. (1994) for discussion of statistical issues in connection with the MACHO data.

The astronomy literature on periodic variable stars is sophisticated from a quantitative viewpoint. For example, it includes methodology for discovering light curves that are "outliers" in a catalogue of such curves; see e.g. Protopapas et al. (2006). And it involves automated methodology for identifying periodic variable stars among millions of light sources in the night sky; see e.g. Derue et al. (2002) and Kabath et al. (2007).

There is a large literature on modelling curves in terms of trigonometric series. In statistics and related fields it includes work of Pisarenko (1973), Hannan (1974), Frost (1976), Quinn and Fernandes (1991), Quinn and Thompson (1991), Quinn (1999) and Quinn and Hannan (2001). Many other contributions can be found in the engineering literature. If the number of components is taken large then the methodology essentially amounts to nonparametric curve estimation, and is closely related to approaches discussed below in section 3. Computational and statistical-efficiency issues connected with the estimation of periodic functions are addressed by McDonald (1986) and Bickel et al. (1993, p. 107), respectively.

Early work in astronomy on nonparametric methods for analysing data on periodic variable stars includes contributions from Lafler and Kinman (1965) and Renson (1978). The method most favoured by astronomers for estimating light curves is the periodogram, which was used by statisticians more than a century ago to assess periodicity. Work on formal testing for periodicity includes that of Fisher (1929), Whittle (1954) and Chiu (1989). The theory of periodogram estimation owes much to Walker (1971, 1973) and Hannan

(1973). The periodgram was introduced to astronomy largely through the work of Barning (1963), Deeming (1975), Lomb (1976), Ferraz-Mello (1981) and Scargle (1982). See also Vityazev (1997). For examples of analyses undertaken using this approach, see Waelkens et al. (1998), de Cat and Aerts (2002), DePoy et al. (2004), Lanza et al. (2004), Aerts and Kolenberg (2005), Maffei et al. (2005), Hall and Li (2006) and Shkedy et al. (2004). Bayesian methods were proposed by Shkedy et al. (2007). Alternative techniques include those of Reimann (1994), Hall et al. (2000) and Hall and Yin (2003).

For some variable stars, the fluctuation of brightness is explained well by a model where period and/or amplitude are also functions of time. See, for example, work of Eyer and Genton (1999), Koen (2005), Rodler and Guggenberger (2005), Sterken (2005), Hart, Koen and Lombard (2007) and Genton and Hall (2007).

1.3 Summary

Section 2 provides an account of least-squares methods for inference in the simplest case, where the light curve can reasonably be modelled in terms of a single periodic function. Periodogram-based methods, and inference when the curve is more plausibly a superposition of p different periodic functions, are treated together in section 3. The case of evolving periodic models is addressed in section 4. Our treatment follows lines given in greater detail by Hall et al. (2000), Hall and Yin (2003), Hall and Li (2006) and Genton and Hall (2007).

2 Models and methodology in the case of periodicity based on least squares

2.1 Models for brightness and observation times

Let $g(x)$ denote the "true" value of brightness of the star at time x. A graph of g, as a function of phase, would be called by astronomers the true "light curve" of the star. We make observations Y_i at respective times X_i, where $0 < X_1 \le \cdots \le X_n$, and obtain the data pairs (X_i, Y_i) for $1 \le i \le n$. The model is superficially one of standard nonparametric regression:

$$Y_i = g(X_i) + \epsilon_i \,, \tag{1}$$

where the ϵ_i's, describing experimental error, are independent and identically distributed random variables with zero mean and finite variance. We take g to be a periodic function with period θ; its restriction to a single period represents the light curve. From the data (X_i, Y_i) we wish to estimate both θ and g, making only periodic-smoothness assumptions about the latter.

A range of generalisations is possible for the model (1). For example, we might replace the errors ϵ_i by $\sigma(X_i)\,\epsilon_i$, where the standard deviation $\sigma(X_i)$

is either known, as is sometimes the case with data on star brightness, or accurately estimable. Then, appropriate weights should be incorporated into the series at used to estimate θ; see (3) below. To reflect many instances of real data, the time points X_i should remain separated as n increases, and in particular the standard "infill asymptotics" regime of nonparametric regression is inappropriate here.

Neither should the X_i's be modelled as equally spaced quantities. Indeed, it is straightforward to see that in this case, and for many values of θ (in particular where θ is a rational multiple of the spacing), consistent estimation is not possible.

Realistic mathematical models for the spacings between successive X_j's include the case where they are approximately stochastically independent. One such model is

$$X_j = \sum_{i=1}^{j} V_i\,, \quad 1 \le j \le n\,, \tag{2}$$

where V_1, V_2, ... are independent and identically distributed nonnegative random variables. Clearly there are limitations, however, to the generality of the distribution allowable for V, representing a generic V_i. In particular, if the distribution is defined on an integer lattice, and if θ is a rational number, then identifiability difficulties continue to cause problems.

These problems vanish if we assume that the X_j's are generated by (2), where the distribution of $V > 0$ is absolutely continuous with an integrable characteristic function and that all moments of V are finite. Call this model $(\mathrm{M}_{X,1})$. The fact that the characteristic function should be integrable excludes the case where the X_i's are points of a homogeneous Poisson process, but that context is readily treated separately.

Another class of processes $\mathcal{X}$ is the sequence $X_j = X_j(n) \equiv nY_{nj}$, where $Y_{n1} < \cdots < Y_{nn}$ are the order statistics of a random sample $Y_1, \ldots, Y_n$ from a Uniform distribution on the interval $[0, y]$, say. Call this model $(\mathrm{M}_{X,2})$. Models $(\mathrm{M}_{X,1})$ and $(\mathrm{M}_{X,2})$ are similar, particularly if V has an exponential distribution. There, if $\mathcal{X}(n+1) = \{X_1, \cdots, X_{n+1}\}$ is a sequence of observations generated under $(\mathrm{M}_{X,1})$, if $\mathcal{X}'(n) = \{X'_1, \ldots, X'_n\}$ is generated under $(\mathrm{M}_{X,2})$ with $y = 1$, and if we define $X_{\mathrm{tot}} = \sum_{i \le n+1} X_i$, then $\{X_1/X_{\mathrm{tot}}, \ldots, X_n/X_{\mathrm{tot}}\}$ has the same distribution as $\mathcal{X}'(n)$.

A third class of processes $\mathcal{X}$ is the jittered grid of Akaike (1960), Beutler (1970) and Reimann (1994), where $X_j = j + U_j$, for $j \ge 1$, and the variables U_j are independent and Uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$. Call this model $(\mathrm{M}_{X,3})$. Each of $(\mathrm{M}_{X,1})$, $(\mathrm{M}_{X,2})$ and $(\mathrm{M}_{X,3})$ has the property that the spacings $X_j - X_{j-1}$ are identically distributed and weakly dependent.

2.2 Least-squares estimation of g and θ

In this section we give an overview of methods for inference. The first step is to construct a nonparametric estimator $\hat{g}(\cdot \,|\, \theta)$ of g on $(0, \theta]$, under the

assumption that the period of g is θ. Next we extend $\hat{g}$ to the real line by periodicity, and, using a squared-error criterion,

$$S(\theta) = \sum_{i=1}^{n} \{Y_i - \hat{g}(X_i \,|\, \theta)\}^2 \,, \tag{3}$$

take our estimator $\hat{\theta}$ of θ to be the minimiser of $S(\theta)$. (We could use a leave-one-out construction of $S(\theta)$, omitting the pair (X_i, Y_i) from the data. While this would give slightly different numerical results, it would not influence first-order asymptotic properties of the method.) Finally, for an appropriate estimator $\tilde{g}(\cdot \,|\, \theta)$ of g under the assumption of period θ, we employ $\hat{g}_0 \equiv \tilde{g}(\cdot \,|\, \hat{\theta})$ to estimate g.

Even if $\hat{g}$ and $\tilde{g}$ are of the same type, for example both local-linear estimators, it is usually not a good idea to take them to be identical. In particular, to ensure approximately optimal estimation of θ, the version $\hat{g}(\cdot \,|\, \theta)$ that we use to define $S(\theta)$ at (3) should be smoothed substantially less than would be appropriate for point estimation of g. In general the function S has multiple local minima, not least because any integer multiple of θ can be considered to be a period of g.

Next we discuss candidates for $\hat{g}$. Under the assumption that the true period of g is θ, the design points X_i may be interpreted modulo θ as $X_i(\theta) = X_i - \theta \lfloor X_i/\theta \rfloor$, for $1 \le i \le n$, where $\lfloor x \rfloor$ denotes the largest integer strictly less than x. Then, the design points of the set of data pairs $\mathcal{Y}(\theta) \equiv \{(X_i(\theta), Y_i), 1 \le i \le n\}$ all lie in the interval $(0, \theta]$. We suggest repeating *ad infinitum* the scatterplot represented by $\mathcal{Y}(\theta)$, so that the design points lie in each interval $((j-1)\theta, j\theta]$ for $-\infty < j < \infty$; and computing $\hat{g}(\cdot \,|\, \theta)$, restricted to $(0, \theta]$, from the data, using a standard second-order kernel method such as a Nadaraya-Watson estimator or a local-linear estimator. In practice we would usually need to repeat the design only in each of $(-\theta, 0]$, $(0, \theta]$ and $(\theta, 2\theta]$, since the effective bandwidth would be less than θ. We define $\hat{g}(\cdot \,|\, \theta)$ on the real line by $\hat{g}(x \,|\, \theta) = \hat{g}(x - \theta \lfloor x/\theta \rfloor \,|\, \theta)$.

In view of the periodicity of g it is not necessary to use a function estimation method, such as local linear, which accommodates boundary effects. Indeed, our decision to repeat design points in blocks of width θ means that we do not rely on the boundary-respecting properties of such techniques. The Nadaraya-Watson estimator, which suffers notoriously from boundary problems but is relatively robust against data sparseness, is therefore a good choice here. The resulting estimator of g is

$$\hat{g}(x \,|\, \theta) = \frac{\sum_i Y_i \, K_i(x \,|\, \theta)}{\sum_i K_i(x \,|\, \theta)} \,, \quad 0 \le x \le \theta \,, \tag{4}$$

where $K_i(x \,|\, \theta) = K[\{x - X_i(\theta)\}/h]$, K is a kernel, h is a bandwidth, and the two series on the right-hand side of (4) are computed using repeated blocks of the data $\mathcal{Y}(\theta)$.

Alternative estimators of g, of slightly lower statistical efficiency than that defined in (4), can be based on the periodogram. This approach tends to be favoured by astronomers, not least because it is readily extended to the case of multiperiodic functions; see section 3.

2.3 Properties of estimators

If g has r bounded derivatives; if the estimator $\hat{g}$ is of rth order, meaning that its asymptotic bias is of size h^r and its variance is of size $(nh)^{-1}$; and if $h = h(n)$ has the property that for some $\eta > 0$, $n^{-(1/2)+\eta} \le h = o(n^{-1/(2r)})$; then $\hat{\theta} = \operatorname{argmin} S(\theta)$ is consistent for θ and, under regularity conditions,

$$n^{3/2}\,(\hat{\theta} - \theta) \to \mathrm{N}\big(0, \tau^2\big) \tag{5}$$

in distribution, where $0 < \tau^2 < \infty$. When $\hat{g}$ is a Nadaraya-Watson or local-linear estimator,

$$\tau^2 = 12\,\sigma^2\,\theta^3\,\mu^{-2} \left\{ \int_0^\theta g'(u)^2\,du \right\}^{-1}, \tag{6}$$

where $\sigma^2 = \operatorname{var}(\epsilon_i)$ and $\mu = \lim_{j\to\infty} E(X_j - X_{j-1})$, assumed to be finite and nonzero. Formula (5) implies that $\hat{\theta}$ converges to θ at a parametric rate. In Quinn and Thompson's (1991) parametric analysis of a closely related problem they obtained the same limit theorem for $\hat{\theta}$, albeit with a different value of τ^2.

Formula (6) implies that estimators of period have lower variance when the function g is 'less flat', i.e. when g has larger mean-square average derivative. This accords with intuition, since a perfectly flat function g does not have well-defined period, and more generally, the flatter g is, the more difficult it is to visually determine its period.

If $h \sim Cn^{-1/(2r)}$ for a constant $C > 0$, and $\hat{g}$ is an r'th order regression estimator, then $n^{3/2}\,(\hat{\theta} - \theta)$ remains asymptotically Normally distributed but its asymptotic bias is no longer zero. In the r'th order case, $h = O(n^{-1/(2r)})$ is the largest order of bandwidth that is consistent with the parametric convergence rate, $\hat{\theta} = \theta + O_p(n^{-3/2})$.

This high degree of accuracy for estimating θ means that, if $\tilde{g}(\cdot\,|\,\theta)$ is a conventional estimator of g under the assumption that the period equals θ, then first-order asymptotic properties of $\hat{g}_0 \equiv \tilde{g}(\cdot\,|\,\hat{\theta})$ are identical to those of $\tilde{g}(\cdot\,|\,\theta)$. That is, from an asymptotic viewpoint the final estimator $\hat{g}_0$ behaves as though the true period were known. These results follow by Taylor expansion. For example, if $\tilde{g}(\cdot\,|\,\theta)$ is the Nadaraya-Watson estimator defined at (2.4), but with a different bandwidth h_0 say, satisfying $h_0 \ge n^{-(1/2)+\xi}$ for some $\xi > 0$, then a Taylor-expansion argument shows that for all $\eta > 0$,

$$\tilde{g}(\cdot\,|\,\hat{\theta}) = \tilde{g}(\cdot\,|\,\theta) + o_p\big\{(nh_0)^{-1/2}\big\}\,. \tag{7}$$

The remainder $o_p\big\{(nh_0)^{-1/2}\big\}$ here is of smaller order than the error of $\tilde{g}(\cdot\,|\,\theta)$ about its mean.

3 The case of multiperiodic functions

3.1 Model for g, and issues of identifiability

In some cases the radiation from a star can reasonably be modelled as a superposition of more than one periodic function. To avoid problems of non-identifiability we take g to be representable as

$$g(x) = \mu + \sum_{j=1}^{p} g_j(x)\,, \quad -\infty < x < \infty\,, \tag{8}$$

where μ denotes a constant, g_j is a smooth, nonvanishing, real-valued periodic function with minimal period θ_j, $0 < \theta_1 < \ldots < \theta_p < \infty$, and each g_j is centred by the condition

$$\int_0^{\theta_j} g_j(x)\,dx = 0\,. \tag{9}$$

Therefore, the constant term in any orthogonal expansion of g_j on $[0, \theta_j]$, with respect to an orthonormal system where one of the orthonormal functions is constant, is absorbed into μ at (8). This property will motivate our estimators of $g_1, \ldots, g_p$; see section 3.3 below.

We assume p is known, and address the problem of estimating $\theta_1, \ldots, \theta_p$ and $g_1, \ldots, g_p$ without making parametric assumptions about the latter. Of course, by conducting inference for different values of p one can obtain significant information about its "true" value, but we do not have a satisfactory approach to formally estimating p.

By saying that θ_j is the minimal period of g_j we mean that if g_j is also periodic with period θ' then $\theta_j \le \theta'$. This does not render either the θ_j's or the representation at (8) uniquely defined, however. Indeed, the representation is unique if and only if the periods are "relatively irrational", meaning that θ_i/θ_j is irrational for each $1 \le i < j \le p$. We shall say that the periods are "relatively rational" if each value of θ_i/θ_j is a rational number.

At first sight this suggests an awkward singularity in the statistical problem of conducting inference about g_j and θ_j, as follows. Since each irrational number is approximable arbitrarily closely by rational ones, then so too each statistically identifiable problem can be approximated arbitrarily closely by a non-identifiable one, by slightly altering the periods θ_j and leaving the shape of each g_j essentially unchanged. And since the periods in the approximating problem can be chosen to be relatively rational, then new and quite different representations may be constructed there, involving finite mixtures of periodic functions that are different from those in the relatively irrational form of the problem. This implies that, even if the original mean function g uniquely enjoys the representation at (8), there is an infinity of alternative mean functions that, while being themselves very close to g, have representations, as mixtures of periodic functions, that differ significantly from the unique representation of g.

While this is correct, it does not often hinder statistical analysis of real or simulated data, since the alternative representations involve functions g_j that are either very rough or have very long periods. In such cases the g_j's are often not practically recognisable as periodic functions, and in particular they lead to solutions that usually appear as pathological.

3.2 Period estimators based on the periodogram

Assume that the data pairs (X_i, Y_i) are generated as at (1), but that g is now a multiperiodic function. Least-squares methods can be used to construct estimators of g and of the periods θ_j, but they are awkward to use in practice, at least without appropriate "starting estimators," since the analogue of $S(\theta)$, at (3), has many local extrema. On the other hand, methods based on the periodogram are relatively easy to implement; we describe them below.

Let cs denote either the cosine or the sine function. For any real number ω, define the squared periodogram by

$$A(\omega)^2 \equiv A_{\cos}(\omega)^2 + A_{\sin}(\omega)^2 \,,$$

where $A_{\rm cs}(\omega) = n^{-1} \sum_i Y_i \,{\rm cs}(\omega X_i)$. If $p = 1$, in which case there is a unique period θ, say, then the quantity $\widehat{\omega}$ which produces a local maximum of $A(\omega)$ achieves a local maximum in the vicinity of each value $\omega^{(k)} = 2k\pi/\theta$, where k is any nonzero integer. This property is readily used to estimate θ.

More generally, in the multiperiodic case the periodogram A has its large peaks near points $2k\pi/\theta_j$, for arbitrary integers k and for $j = 1, \ldots, r$. By sorting peak locations into r disjoint sets, for each of which adjacent values are approximately equal distances apart, the values of θ_j may be estimated as before. In either case the estimators converge at the rate $n^{-3/2}$ discussed for the least-squares methods introduced in section 2.

3.3 Estimators of g

Having constructed $\hat\theta_1, \ldots, \hat\theta_p$ we use orthogonal series methods to develop estimators $\hat g_1, \ldots, \hat g_p$, as follows. Let $\{\psi_0, \psi_1, \ldots\}$ denote a complete orthonormal sequence of functions on the interval $[0, 1]$, with $\psi_0 \equiv 1$. Extend each function to the real line by periodicity. Given an integer $m \geq 1$, which will play the role of a smoothing parameter; given generalised Fourier coefficients a_{jk} for $1 \leq j \leq p$ and $1 \leq k \leq m$; and given a constant μ; put

$$\tilde g(x \,|\, a, \mu) = \mu + \sum_{j=1}^{p} \sum_{k=1}^{m} a_{jk} \, \psi_k(x/\hat\theta_j) \,, -\infty < x < \infty \,, \tag{10}$$

where a denotes the parameter vector of length $q = mp$ made up of all values of a_{jk}. Of course, the functions $\psi_k(\cdot/\hat\theta_j)$ used in this construction are periodic

with period $\hat{\theta}_j$. The estimator (10) reflects the model (8) and the constraint (9), the latter imposed to help ensure identifiability.

Take $(\hat{a}, \widehat{\mu})$ to be the minimiser of

$$T(a, \mu) = \sum_{i=1}^{n} \{Y_i - \tilde{g}(X_i|a, \mu)\}^2 .$$

In this notation our estimator of g_j is

$$\hat{g}_j(x) = \sum_{k=1}^{m} \hat{a}_{jk}\, \psi_k(x/\hat{\theta}_j) .$$

In practice we recommend taking $\{\psi_j\}$ to be the full trigonometric series: $\psi_0 \equiv 1$ and, for $j \geq 1$,

$$\psi_{2j}(x) = 2^{1/2} \cos(2j\pi x) \quad \text{and} \quad \psi_{2j-1}(x) = 2^{1/2} \sin(2j\pi x) .$$

4 Evolving periodic functions

4.1 Introduction

The notion that star brightness is given by a fixed periodic function, unchanging over time, is of course a simplification. The very mechanisms that produce periodicity are themselves the subject of other mechanisms, which affect their properties and so influence the period and amplitude of the supposedly periodic function. Thus, while the model at (1) might be reasonable in many circumstances, in some instances we should allow for the fact that the characteristics of g will alter over time.

In the sections below we develop models for functions with evolving amplitude and period, and then we combine these to produce a model for g. Finally we use that model to motivate estimators.

4.2 The notions of evolving period and amplitude

Write g_0 for a periodic function with unit period, and let t denote a continuously differentiable, strictly increasing function. Represent time by x, and put $t_x = t(x)$ and $t'_x = t'(x) > 0$. We shall consider the function t to provide a change of time, from x to t_x.

Assume that a function g can be represented as

$$g(x) = g_0(t_x) . \tag{11}$$

We think of g as having period $1/t'_x$ at time x, and in fact for small $u > 0$,

$$g(x + u) = g_0\{t_x + t'_x\, u + o(u)\} .$$

Since the function $d(u) = g_0(t_x + t'_x\, u)$ has period $1/t'_x$, then, if the time-transformation t_{x+u} were to be applied in a linear way for $u > 0$, g would have period $1/t'_x$ at all future times $x+u$. More generally, without the linearity assumption, the function g given by (11) can be considered to have a period $1/t'_x$ that evolves as time, x, increases.

Amplitude can also evolve. If $a > 0$ is a smooth function, representing amplitude; and if we write a_x for $a(x)$; then we might generalise (11) to:

$$g(x) = a_x\, g_0(t_x)\,. \tag{12}$$

Here we could consider g to have period $1/t'_x$, and amplitude $a_x\, g_0(t_x)$, at x.

The concept of evolving amplitude has to be treated cautiously, however. While altering time can change only the distances between successive peaks and troughs in the function g_0, altering both amplitude and time can produce a function which is very different. Any smooth, strictly positive function g can be constructed non-uniquely as at (12), with $a > 0$ representing a smooth amplitude change, $t_x \equiv x$ being the identity transformation, and g_0 denoting any strictly positive, smooth function, periodic or otherwise.

One conclusion to be drawn from this discussion is that, unless amplitude is determined by a relatively simple parametric model; and unless it changes only very slowly over time, relative to the lengths of periods; it can interact too greatly with period to be interpretable independently of period.

It is possible for non-identifiability of g_0 to occur even when $a \equiv 1$ and the function t has a simple parametric form. For example, suppose that, in the particular case $p = 1$, $t_{x+kp} = t_x + k$ for each $x \in [0, 1]$ and each integer $k \geq 1$. Then, since g_0 has period 1, it follows that $g_0(t_{x+k}) = g_0(t_x)$ for each x and each integer k. Therefore, the periodic function $g \equiv g_0(t)$ is representable as either a time-changed version of the function g_0 with unit period, or more directly as the non-time changed function $g_1 \equiv g_0(t)$ with unit period. If we consider this particular time-change function t, and also the identity time-change, to be members of a larger parametric class, $\mathcal{T}$ say, of time-change functions, then there is ambiguity in determining the member s of $\mathcal{T}$ that enables us to represent $g \equiv g_0(t)$ as $g = g_2(s)$ where g_2 has period 1.

4.3 Models for period

We shall interpret (12) as a model for a regression mean, g, where the functions a and t are determined parametrically and g_0 is viewed nonparametrically. In order for (12) to be interpretable in astronomical terms, it is helpful for the models for t to be quite simple. For example, taking $t_x = \theta_2^{-1}\, \log(\theta_1 + \theta_2 x) + \theta_3$, for constants $\theta_1 > 0$, θ_2 and θ_3, implies that $1/t'_x = \theta_1 + \theta_2 x$. In this case the initial period is θ_1, and the period changes linearly with time, with slope θ_2. If we start measuring time at zero when $x = 0$ then we require $\theta_3 = -\theta_2^{-1} \log \theta_1$, and then the model becomes:

$$t_x = \theta_2^{-1}\, \log\left(1 + \theta_1^{-1}\theta_2 x\right). \tag{13}$$

We might refer to (13) as a "linear model," since it results from a linear model for period. Analogously we could describe the model

$$t_x = (\theta_1\theta_2)^{-1}\left(1 - e^{-\theta_2 x}\right), \tag{14}$$

for which $1/t'_x = \theta_1\, e^{\theta_2 x}$, as an "exponential model." It is an attractive alternative to the linear model in certain cases, since its period is unequivocally positive.

A time-change function such as

$$t_x = \int_0^x \left(\theta_1 + \theta_2\, u + \ldots + \theta_k\, u^{k-1}\right)^{-1} du \tag{15}$$

produces a period the evolution of which, in time, is described exactly by a polynomial of degree $k-1$, and represents a generalisation of the linear model.

It should be appreciated that in models (13)–(15), and in a setting where data are assumed to be observed at an approximately constant rate over a time interval $[0, n]$ of increasing length n, usually only the parameter θ_1, representing period at time $x = 0$, would be kept fixed as n increased. The parameters $\theta_2, \ldots, \theta_k$ would typically decrease to zero as n increased, and in fact would usually decrease at such a rate that $n^{j-1}\,|\theta_j|$ was at least bounded, if not decreasing to zero, for $2 \le j \le k$. This prevents period from changing by an order of magnitude over the observation time-interval. Moreover, if $\theta_1 > 0$ is fixed and $\sup_{1\le j\le k}\, n^{j-1}\,|\theta_j| \to 0$ as $n \to \infty$, then for all sufficiently large values of n, t_x is strictly monotone increasing on $[0, n]$. In such cases, (15) is asymptotically equivalent to the simpler model,

$$t_x = \theta_1^{-1}\, x + \theta_2\, x^2 + \ldots + \theta_k\, x^k\,, \quad 0 \le x \le n\,, \tag{16}$$

modulo a reparametrisation. An exponentiated version of (16) is also possible.

4.4 Models for amplitude

Models for the function a_x can be constructed similarly to those for t_x. However, in order to avoid identifiability problems we should insist that $a_x = 1$ at the initial time, so that initial amplitude is incorporated into the function g_0. Bearing this in mind, and taking the initial time to be $x = 0$, potential models include

$$a_x = 1 + \omega_1\, x + \ldots + \omega_\ell\, x^\ell\,, \quad 0 \le x \le n\,,$$

and its exponentiated form, $a_x = \exp(\omega_1\, x + \ldots + \omega_\ell\, x^\ell)$.

4.5 Model for data generation

Assume that data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated by the model

$$Y_i = a(X_i \,|\, \omega^0)\, g_0\{t(X_i \,|\, \theta^0)\} + \epsilon_i\,,$$

where $a_x = a(x \,|\, \omega)$ and $t_x = t(x \,|\, \theta)$ denote smooth, positive functions determined by finite vectors ω and θ of unknown parameters, ω^0 and θ^0 are the true values of the respective parameters, $t(\,\cdot\,|\,\theta)$ is strictly increasing, $a(\,\cdot\,|\,\omega)$ is bounded away from zero and infinity, and the experimental errors, ϵ_i, have zero mean. For example, $t(\,\cdot\,|\,\theta)$ and $a(\,\cdot\,|\,\omega)$ could be any one of the models introduced in sections 4.3 and 4.4, respectively.

As in section 4.2, g_0 is assumed to be a smooth, periodic function with unit period. Therefore, even if the regression mean, $g(x) = a(x \,|\, \omega)\, g_0\{t(x \,|\, \theta)\}$, were a conventional periodic function, without any amplitude or time change, the period, p say, would be inherited from the time-change function $t(x \,|\, \theta)$, which here would be linear: $t(x \,|\, \theta) = x/p$ and $\theta = p$, a scalar. We shall take $a(0 \,|\, \omega) = 1$ if $x = 0$ is the earliest time-point on our scale, so that amplitude is inherited from g_0.

Similar results, and in particular identical convergence rates of estimators, are obtained for a variety of processes X_i that are weakly stationary and weakly independent. They include cases where the X_i's are (a) points of a homogeneous Poisson process with intensity μ^{-1} on the positive real line; or (b) the values of $[n/\mu]$ (integer part of n/μ) independent random variables, each uniformly distributed on the interval $[0, n]$; or (c) the values within $[0, n]$ of the "jittered grid" data $j\mu^{-1} + V_j$, where the variables V_j are independent and identically distributed on a finite interval. See section 2.1 for discussion of models such as (a), (b) and (c). In each of these cases the average spacing between adjacent data is asymptotic to μ as $n \to \infty$.

4.6 Estimators

To estimate g_0, ω and θ, put

$$\hat{g}_0\big\{t(x \,|\, \theta) \,\big|\, \theta, \omega\big\} = \frac{\sum_i a(X_i \,|\, \omega)^{-1}\, Y_i\, K_i(x \,|\, \theta)}{\sum_i K_i(x \,|\, \theta)}\,,$$

$$S(\theta, \omega) = \sum_i \Big[Y_i - a(X_i \,|\, \omega)\, \hat{g}_0\big\{t(X_i \,|\, \theta) \,\big|\, \theta, \omega\big\}\Big]^2\,,$$

where $K_i(x \,|\, \theta) = K[\{x(\theta) - X_i(\theta)\}/h]$, K is a kernel function, h is a bandwidth,

$$x(\theta) = t(x \,|\, \theta) - \lfloor t(x \,|\, \theta) \rfloor\,, \quad X_i(\theta) = t(X_i \,|\, \theta) - \lfloor t(X_i \,|\, \theta) \rfloor\,,$$

and $\lfloor u \rfloor$ denotes the largest integer strictly less than u.

Let $(\theta, \omega) = (\hat{\theta}, \widehat{\omega})$ be the minimiser of $S(\theta, \omega)$. Then, potentially using, to construct $\hat{g}_0$, a bandwidth different from the one employed earlier, our estimator of g_0 is $\hat{g}_0(\,\cdot\,|\,\hat{\theta},)$. Estimators of the time-change function $t_x = t(x \,|\, \theta^0)$ and amplitude function $a_x = a(x \,|\, \omega^0)$ are given by $\hat{t}_x = t(x \,|\, \hat{\theta})$ and $\hat{a}_x = a(x \,|\, \widehat{\omega})$, respectively. We estimate g, defined at (12), as $\hat{g}(x) = \hat{a}_x\, \hat{g}_0(\hat{t}_x)$.

References

AKAIKE, H. (1960): Effect of timing error on the power spectres of sampled-data. *Ann. Inst. Statist. Math. 11, 145-165.*

AXELROD, T.S. and 18 OTHER AUTHORS (1994): Statistical issues in the MACHO Project. In: G. J. Babu and E. D. Feigelson (Eds.) *Statistical Challenges in Modern Astronomy II*, Springer, New York, 209-224.

AERTS, C. and KOLENBERG, K. (2005): HD 121190: A cool multiperiodic slowly pulsating B star with moderate rotation. *Astronom. Astrophys. 431, 614-622.*

BARNING, F.J.M. (1963): The numerical analysis of the light-curve of 12 Lacertae. *Bull. Astronom. Institutes of the Netherlands 17, 22-28.*

BEUTLER, F.J. (1970): Alias-free randomly timed sampling of stochastic processes. *IEEE Trans. Inform. Theor. IT-16, 147-152.*

BICKEL, P.J., KLAASSEN, C.A.J., RITOV, Y. and WELLNER, J.A. (1993): *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

CHIU, S. (1989): Detecting periodic components in a white Gaussian time series. *J. Roy. Statist. Soc. Ser. B 5, 249-259.*

DE CAT, P. and AERTS, C. (2002): A study of bright southern slowly pulsating B stars, II. The intrinsic frequencies. *Astronom. Astrophys. 393, 965-982.*

DEEMING, T.J. (1975): Fourier analysis with unequally-spaced data. *Astrophys. Space Sci. 36, 137-158.*

DEPOY, D.L., PEPPER, J., POGGE, R.W., STUTZ, A., PINSONNEAULT, M. and SELLGREN, K. (2004): The nature of the variable galactic center source IRS 16SW. *Astrophys. J. 617, 1127-1130.*

DERUE, F. and 49 OTHER AUTHORS (2002): Observation of periodic variable stars towards the Galactic spiral arms by EROS II. *Astronom. Astrophys. 389, 149-161.*

EYER, L. and CUYPERS, J. (2000): Predictions on the number of variable stars for the GAIA space mission and for surveys as the ground-based International Liquid Mirror Telescope. In: L. Szabados and D. W. Kurtz (Eds.): *The Impact of Large Scale Surveys on Pulsating Star Research.* ASP Conference Series, vol. 203. Astronomical Society of the Pacific, San Francisco, 71-72.

EYER, L., GENTON, M.G. (1999): Characterization of variable stars by robust wave variograms: an application to Hipparcos mission. *Astron. Astrophys, Supp. Ser. 136, 421-428.*

FERRAZ-MELLO, S. (1981): Estimation of periods from unequally spaced observations. *Astronom. J. 86, 619-624.*

FISHER, R.A. (1929): Tests of significance in harmonic analysis. *Proc. Roy. Soc. London Ser. A 125, 54-59.*

FROST, O.L. (1976): Power-spectrum estimation. In: G. Tacconi (Ed.): *Aspects of Signal Processing with Emphasis on Underwater Acoustics*, Part I. Reidel, Dordrecht, 12-162.

GENTON, M.G. and HALL, P. (2007): Statistical inference for evolving periodic functions. *Roy. Statist. Soc. Ser. B 69, 643-657.*

GOOD, G.A. (2003): *Observing Variable Stars.* Springer, New York.

HALL, P. and LI, M. (2006): Using the periodogram to estimate period in nonparametric regression. *Biometrika 93, 411-424.*

HALL, P., REIMANN, J. and RICE, J. (2000): Nonparametric estimation of a periodic function. *Biometrika 87, 545-557.*

HALL, P. and YIN, J. (2003): Nonparametric methods for deconvolving multiperiodic functions. *J. Roy. Stat. Soc. Ser. B 65, 869-886.*

HANNAN, E.J. (1973): The estimation of frequency. *J. Appl. Prob. 10, 510-519.*

HANNAN, E.J. (1974): Time-series analysis. System identification and time-series analysis. *IEEE Trans. Automatic Control AC-19, 706-715.*

HART, J.D., KOEN, C. and LOMBARD, F. (2004): An analysis of pulsation periods of long-period variable stars. *J. Roy. Statist. Soc. Ser. C 56, 587-606.*

HOFFMEISTER, C., RICHTER, G. and WENZEL, W. (1985): *Variable Stars.* Springer, Berlin.

KABATH, P., EIGMÜLLER, P., ERIKSON, A, HEDELT, P., RAUER, H., TITZ, R. and WIESE, T. (2007): Characterization of COROT Target Fields with BEST: Identification of Periodic Variable Stars in the IR01 Field. *Astronom. J. 134, 15601569.*

KOEN, C. (2005): Statistics of O-C diagrams and period changes. In: C. Sterken (Ed.): *The Light-Time Effect in Astrophysics.* ASP Conference Series vol. 335. Astronomical Society of the Pacific, San Francisco, 25-36.

LAFLER, J. and KINMAN, T.D. (1965): An RR Lyrae survey with the Lick 20-inch astrograph II. The calculation of RR Lyrae period by electronic computer. *Astrophys. J. Suppl. Ser. 11, 216-222.*

LANZA, A.F., RODONÒ, M. and PAGANO, I. (2004): Multiband modelling of the Sun as a variable star from VIRGO/SoHO data. *Astronom. Astrophys. 425, 707-717.*

LOMB, N.R. (1976): Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci. 39, 447-462.*

MAFFEI, P., CIPRINI, S. and TOSTI, G. (2005): Blue and infrared light curves of the mysterious pre-main-sequence star V582 Mon (KH 15D) from 1955 to 1970. *Monthly Not. Roy. Astronom. Soc. 357, 1059-1067.*

MCDONALD, J.A. (1986): Periodic smoothing of time series. *SIAM J. Sci. Statist. Comput. 7, 665-688.*

NORTH, G. (2005): *Observing Variable Stars, Novae, and Supernovae.* Cambridge University Press, Cambridge, UK.

NORTON, A.J. and 19 OTHER AUTHORS (2008): New periodic variable stars coincident with ROSAT sources discovered using SuperWASP. *Astronom. Astrophys.*, to appear.

PERCY, J.R. (2007): *Understanding Variable Stars.* Cambridge University Press, Cambridge, UK.

PISARENKO, V. (1973): The retrieval of harmonics from a covariance function. *Geophys. J. Roy. Astronom. Soc. 33, 347-366.*

PRIGARA, F.V. (2007): Radial solitary waves in periodic variable stars. Manuscript.

PROTOPAPAS, P., GIAMMARCO, J.M., FACCIOLI, L., STRUBLE, M.F., DAVE, R. and ALCOCK, C. (2006): Finding outlier light curves in catalogues of periodic variable stars. *Monthly Not. Roy. Astronom. Soc. 369, 677-696.*

QUINN, B.G. (1999): A fast efficient technique for the estimation of frequency: interpretation and generalisation. *Biometrika 86, 213-220.*

QUINN, B.G. and FERNANDES, J.M. (1991): A fast efficient technique for the estimation of frequency. *Biometrika 78, 489-497.*

QUINN, B.G. and HANNAN, E.J. (2001): *The Estimation and Tracking of Frequency.* Cambridge University Press, Cambridge, UK.

QUINN, B.G. and THOMPSON, P.J. (1991): Estimating the frequency of a periodic function. *Biometrika 78, 65-74.*

REIMANN, J.D. (1994): *Frequency Estimation Using Unequally-Spaced Astronomical Data.* Unpublished PhD thesis, University of California, Berkeley.

RENSON, R. (1978): Méthode de recherche des périodes des étoiles variables. *Astron. Astrophys. 63, 125-129.*

RODLER, F. and GUGGENBERGER, E. (2005): Spurious period shifts and changes among variable stars. In: C. Sterken (Ed.): *The Light-Time Effect in Astrophysics.* ASP Conference Series vol. 335. Astronomical Society of the Pacific, San Francisco, 115-118.

SCARGLE, J.D. (1982): Studies in astronomical time-series analysis II: Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J. 263, 835-853.*

SHKEDY, Z., DECIN, L., MOLENBERGHS, L. and AERTS, C. (2007): Estimating stellar parameters from spectra using a hierarchical Bayesian approach. *Monthly Not. Roy. Astronom. Soc. 377, 120-132.*

SHKEDY, Z., DECIN, L., MOLENBERGHS, L., AERTS, M. and AERTS, C. (2004): Estimating stellar parameters from spectra. I. Goodness-of-fit parameters and lack-of-fit test. *Astronom. Astrophys. 421, 281-294.*

STERKEN, C. (2005): The O-C diagram: basic procedures. In: C. Sterken (Ed.): *The Light-Time Effect in Astrophysics.* ASP Conference Series vol. 335. Astronomical Society of the Pacific, San Francisco, 3-22.

STERKEN, C. and JASCHEK, C. (1996): *Light Curves of Variable Stars: A Pictorial Atlas.* Cambridge University Press, Cambridge, UK.

VITYAZEV, V.V. (1997): Time series analysis of unequally spaced data: Intercomparsion between estimators of the power spectrum. In: G. Hunt and H. E. Payne (Eds.) *Astronomical Data Analysis Software and Systems VI.* ASP Conference Series vol. 125. Astronomical Society of the Pacific, San Francisco, 166-169.

WAELKENS, C., AERTS, C., KESTENS, E., GRENON, M. and EYER, L. (1998): Study of an unbiased sample of B stars observed with Hipparcos: the discovery of a large amount of new slowly pulsating B stars. *Astronom. Astrophys. 330, 215-221.*

WALKER, A.M. (1971): On the estimation of a harmonic component in a time series with stationary independent residuals. *Biometrika 58, 21-36.*

WALKER, A.M. (1973): On the estimation of a harmonic component in a time series with stationary dependent residuals. *Adv. Appl. Probab. 5, 217-241.*

WARNER, B.D. (2006): *A Practical Guide to Lightcurve Photometry and Analysis.* Springer, New York.

WHITTLE, P. (1954): The statistical analysis of a seiche record. *J. Marine Res. 13, 78-100.*

Part II

Advances on Statistical Computing Environments

Back to the Future: Lisp as a Base for a Statistical Computing System

Ross Ihaka[1] and Duncan Temple Lang[2]

[1] University of Auckland, New Zealand, *ihaka@stat.auckland.ac.nz*
[2] University of California, Davis, USA

Abstract. The application of cutting-edge statistical methodology is limited by the capabilities of the systems in which it is implemented. In particular, the limitations of R mean that applications developed there do not scale to the larger problems of interest in practice. We identify some of the limitations of the computational model of the R language that reduces its effectiveness for dealing with large data efficiently in the modern era.

We propose developing an R-like language on top of a Lisp-based engine for statistical computing that provides a paradigm for modern challenges and which leverages the work of a wider community. At its simplest, this provides a convenient, high-level language with support for compiling code to machine instructions for very significant improvements in computational performance. But we also propose to provide a framework which supports more computationally intensive approaches for dealing with large datasets and position ourselves for dealing with future directions in high-performance computing.

We discuss some of the trade-offs and describe our efforts to realizing this approach. More abstractly, we feel that it is important that our community explore more ambitious, experimental and risky research to explore computational innovation for modern data analyses.

Keywords: Lisp, optional typing, performance

1 Background

The growth in popularity of R over the last decade has been impressive and has had a significant impact on the practice and research of statistics. While the technical achievements have been significant, the fostering of a community which has continued the development of R and the myriad of packages that provide cutting-edge statistical methodology is perhaps the most significant achievement of the R project.

R is not unlike the S language that was developed at Bell Labs over the last 3 decades of the last century. At that time, S was revolutionary in concept and enabled a different approach to data analysis that continues today. A similar change in the way we do data analysis and statistical computing is needed again. This is no small part due to the changing nature of scientific computing (parallel and distributed computing, Web-based data access and computing,

massive data sets, computationally intensive methods). But also, we need to be undertaking bold research that involves experimenting with these new technologies and guiding statisticians to new computational paradigms rather than focusing mostly on making the existing, familiar facilities easier to use and implementing ideas available in numerous other programming languages.

It is important that the statistical community recognize the impact that R has had and not assume that it is sufficient for the long-term or that new developments will simply happen. Rather, they must encourage, support and participate in the development of new ideas and infrastructure.

In part, due to the success and popularity of R, it is no longer a research vehicle for more ambitious experiments. The focus of R development has changed gradually to be one of adding important usability features found in other languages, e.g. graphical user interfaces, support for Unicode and internationalisation, and improving portability and ease of use. People wanting to pursue more experimental research projects have been faced with the "nobody will use it" issue as there is a single, "official" R. Simply put, the phrase "the good is the enemy of the better" expresses well the sentiment that R has proven to be good enough for our needs and that an incremental, more localized mindset has developed and has made development of R somewhat conservative. This has inhibited significant changes in direction and has encouraged the more incremental, short term developments rather than a big picture research oriented view of statistical computing. Unfortunately, this has become dominant within the statistics community and journals, and we are now focused more on implementations of existing algorithms than novel new paradigms. To encourage and retain good minds in this field, we need to provide a more significant innovative and exciting research environment where concepts not code are the topics discussed and we are working on large problems, not just details of smaller issues.

2 Issues with R

Before commenting on any of R's deficiencies, we should note that R has been and continues to very effective and successful and there have been numerous significant developments within its history. However, modern data analysis and statistical and scientific computing are continuing to change at a dramatic rate and the essential computational model underlying R is tied to that of the early S systems from 20 to 30 years ago. We outline some of the issues below and note that they refer to efficiency of code execution and support for better programming practices with type specification.

Copying: R uses a pass-by-value semantic for function calls. This means that when a function modifies the contents of one of its arguments, it is a local copy of the value which is changed, not the original value. This has many desirable properties, including aiding reasoning about and debugging code,

and ensuring precious data is not corrupted. However, it is very expensive as many more computations need to be done to copy the data, and many computations require excessive memory due to the large number of copies needed to guarantee these semantics.

Whole-Object and Vectorized Computations versus Scalar Operations: There is a significant benefit to using vectorized functions that perform operations on the whole object rather than writing code that processes elements individually. However, many operations are hard to vectorise and operations that need to "unbox" individual values are extremely expensive. (See section 4 for timing results.)

Compiled Native Code: To obtain efficient code, it is quite common to move important code to C and access that from R. While this is not very difficult, it does pose a challenge to many users and requires knowledge of an additional programming language. Further, it make the resulting software less amenable to extensions by others, and involves significantly more work by the author to bridge the interface between the two languages, and especially debugging the two separate pieces of code.

Software and Type checking: Like many high-level programming languages, R does not require or support declarations and type specification for variables. This is very useful for rapid, interactive programming and prototyping. However, when developing larger systems or software for others to use, being able to annotate code with type information and have the system enforce it is an important productivity gain and produces more robust and reflective software.

These are issues with the language, not the implementation. They reflect sensible decisions that we need to reevaluate in the face of significant changes to computing and data analysis over the last and next decade.

3 Common Lisp

The R engine began life as a very simple Lisp interpreter. The similarities between S and Lisp made it easy to impose an S-like syntax on the interpreter and produce a result which looked very much like S. The fact that this approach has succeeded once raises the question of whether it might be possible to do even better by building a statistical language over a more robust, high-performance Lisp. There are both pluses and minus to taking this approach. On the minus side, "we" will no longer own the implementation details of all aspects of our computing environment. This reduces our independence to enact our own modifications. On the plus side, we gain the experience and effort of an entirely different and broader community in the implementation of an engine. This means that there is no necessity to add

features like namespaces, conditions/exceptions and an object system to the language. They are already present. One of the most important benefits of the approach is that we can use a version of Lisp that compiles to machine code and get significantly improved performance for general code via optional specification of the data types. This raises the possibility of greatly improving the performance of our statistical systems.

Common Lisp is a natural choice of Lisp for building large software systems. It is a formally standardized specification with many implementations – both open source and commercial. The implementations of interest to us provide various different types of small and very large number types (including rationals); language macros; lexical scoping/closures; dynamic scoping; optional type declarations; machine-code compilation; name spaces; a basic package mechanism; extensible I/O types via connections/streams; foreign function interface (FFI); thread support; reflection and programming on the language; additional user-level data structures (e.g. general hash tables, linked lists); unicode; reference semantics; destructive in-situ operations; error handling system (conditions and exceptions); profiling and debugging tools; interfaces for Emacs; IDEs for commercial versions of Lisp and an object/class system similar but richer than the S4 system in R. It is one of the few languages that is both high-level (interactive) and low-level & efficient (compiled to machine code) and offers features similar to those that have proven effective in statistics, with more idealized semantics for statistical computing. The syntax is quirky, but a thin layer on top of this that provides a more familiar form (see section 6) makes Lisp an extremely attractive candidate for moving forward in statistical computing practice and research.

Using Common Lisp provides significant advantages. It would free up the limited and valuable resources that the statistical computing community invests in maintaining, extending and innovating its own language and interpreter when a better one is already available to us. We do lose some control over aspects of the software environment, but the similarities of R and Lisp are such that this does not seem of any consequence. We can contribute changes to Open Source Lisp implementations (e.g. SBCL) or even fork development if we truly need such autonomy. But with the resources not tied to porting new features already existing in Lisp – both now and in the future – we can focus on innovations in statistical computing rather than computer science and information technology.

If we are willing to embark on building a new statistical computing environment, we need to consider all possible languages that might serve as a good base, and not just Lisp. We discuss other candidates in section 8.

4 Speed, compilation & timings

As mentioned previously, many R packages use compiled C/FORTRAN code in order to gain efficiency. As a result, R is a good prototyping environment but requires low-level programming for computationally intensive methods. And this has led people to disregard it for use in large-scale, high performance computing tasks. We want to reduce the gap between programming in the high-level language (R) and making things efficient in the system-level language (C), and also to allow methodology developed and implemented by statisticians to be used in real, industrial-strength applications. We believe that the optional type declaration and machine-code compiler provided by implementations of Lisp achieves this.

Let's consider a basic and overly simple example in which we implement the sum function directly within a high-level language. The following are obvious implementations of this in both R and Python[1], also a dynamic, interpreted language without type specification but with byte-code compilation.

R

```
Sum =
function(x) {
  ans = 0
  for(e in x)
    ans = ans + e
  ans
}
```

Python

```
def Sum(x):
    ans = 0.0
    for i in x:
        ans = ans + i
    return ans
```

We are ignoring issues such as missing values (NAs), and of course, both systems provide built-in, compiled versions of the sum function. However, we are interested in using this elementary example that focuses on scalar computations to compare the performance of our implementations written in the language with a similar implementation in Lisp.

We used a vector of length $100,000$ and computed its sum $10,000$ times to compare the relative performances of our two implementations above with the built-in ones and also a similar implementation Lisp. We also implemented and measured equivalent code in Java and C and explored different ways to compute the result in both Lisp and Python, i.e. using the general reduce function in both systems. The example is sufficiently small and we want to compare the naïve, obvious implementations, so we did not spend much time optimizing the code. The results are given in Table 1.

Python's built-in `sum` is much slower than R's built-in function because, while both are written in C, the Python code accepts generic, extensible Python sequences and must use generic dispatch (at the C-level or perhaps

[1] A potential point of confusion is that the compiler module within CMU Common Lisp is called Python and predates the programming language Python.

Implementation	*Time*	*Performance factor relative to slowest*
R interpreted	945.71	1
Python interpreted	385.19	2.50
Python reduce() function	122.10	7.75
Lisp no type declarations	65.99	14.33
Python built-in sum()	49.26	19.20
R built-in sum()	11.2	84.40
Lisp with type declarations*	2.49	379.80
Java	1.66	569.70
C	1.66	569.70

Table 1. Execution time (in seconds) of the summation of a vector of size 100, 000 repeated 10, 000 times. In all but the case of the call to R's built-in sum() function, there is no test for `NA`s. These measurements were taken on a Linux machine with a 2.4Ghz AMD 64 bit chip and 32 GB of RAM. *We also performed the experiments on an Intel Mac (2.33Ghz, 3Gb RAM) and the results were similar, but the actual values were quite different for some situations. The built-in R sum() took only 2.68 seconds and so is much more similar to Lisp which took 1.85 seconds on that machine. The Java code was 3 times slower than the C code.

to a Python function) to fetch the next element of the sequence and then similarly for adding the number to the total. While it is reasonable to point out that both the Python and R built-in functions are more general than the compiled lisp function in that they can handle arbitrary sequences and numeric and integer vectors respectively, this objection has one serious flaw. While the Lisp function has been limited to vectors of double-float elements, Lisp allows us to declare these limitations; R and Python do not. We can easily create a collection of specialized, fast sum functions for other data types in Lisp, but we cannot in R and Python. This optimization is not available to us in R and Python.

The timings show that the simple implementation entirely within Lisp is essentially as fast as R's C routine, taking into account that the latter tests for `NA`s. What is also informative is the factor of 35 between the Lisp code that has just two type declarations and the version that has none; optional type declarations are effective. But the important comparison is between the type-declared Lisp version and the equivalent version written entirely in both R and Python. Here we see that the Lisp version is 380 times faster than R and 150 times faster than Python.

Over the last several years, Luke Tierney has been making progress on byte-code compilation of R code. His results indicate an improvement of a factor between 2 and 5 (Tierney (2001)). Luke Tierney has also been experimenting with using multiple processors within the internal numerical computations done by R. This has the potential to speed up the code, but will yield, at best, a factor given by the number of available processors. Further,

this work would need to be done manually for all functions and would not directly apply to user-level code.

The timing results illustrate that the optimized Lisp code runs about 30% slower than optimized C code. Clearly, if the majority of the computations in the high-level language amount to calling primitives written efficiently in C, then this 30% slow-down will lead to an overall slow-down and the resulting system will be a potential step-backwards. However, we can of course implement such primitives ourselves in C and use them from within Lisp. But more importantly, we do not believe that these primitives form the majority of the operations, and further that copying objects is a large contributor to performance issues in R. While vectorized operations are fundamental, they are not relevant to the many common computations which cannot be readily vectorized. And the primary message from this section is that when we implement an algorithm that deals with individual elements of a vector in the high-level language, the gains in the Lisp approach are immense. Some Lisp implementations are not slow and the language is viable for high-performance computing. Furthermore, the gains in speed are available incrementally along a continuum ranging from an initial version that is subsequently annotated with increasing amount of information about the types of the data/variables. So the improvement in run-time will also be frequently accompanied by gains in development time as we don't have to switch to another language (e.g. C) to obtain the necessary performance improvements.

5 Actual examples

5.1 Reinforced random walk

Motivated by a research problem of a colleague, we simulated a simple discrete two dimensional reinforced random walk. This is a random walk in which the transition probabilities of moving North, South, East or West from the current position are a function of the number of times the walk has previously visited the current spot. In our simulation, the probability of moving East if this was the second or greater time we had visited the current location is $(1+\beta)/4$ and the probability of moving West is $(1-\beta)/4$; all other probabilities are $1/4$.

This is a potentially expensive simulation as we must keep a record of how often each location has been visited, and further we need to be able to quickly determine the number of times we have visited the a particular position. The choice of data structure and algorithm for computing this is important for the efficiency of this algorithm. We use a hash table with the location as a key (in Lisp the object can be used directly, but in R, we must create a string from the x, y pair). Furthermore, since this is a Markov process, it is not readily vectorized.

We implemented the algorithm in both R and Lisp using the same algorithm. With $\beta = .5$, we ran $100,000$ steps of the random walk on several

different machines. The execution times for 3 different machines are given below. (the times are in seconds).

Lisp	R	Machine characteristics
0.215	6.572	2.33Ghz/3GB Intel, Mac OS X
0.279	7.513	2.4Ghz/32GB AMD Opteron, Linux
0.488	8.304	1Ghz/2GB AMD Athlon, Linux

So we see a significant benefit from using Lisp, with a speedup of a factor ranging from 17 to 30.

The person interested in doing these simulations proposed looking at 50 different values of β and performing $10,000$ random walks, each of length $10,000$. The goal is to look at the distributions of both the drift and the standard deviation of the walk. On the Intel Mac laptop, the R version takes .75 seconds for $10,000$ iterations. 50 replications of this takes 39.912 seconds, and 100 takes 80.213 seconds. So this is close to linear and $10,000$ replications of $10,000$ iterations would take at least 133 minutes. And to do this for 50 values of beta would take at least $4\frac{1}{2}$ days! This assumes that the computation will complete and not run out of memory.

The Lisp version takes 212.9 seconds for $10,000$ iterations of $10,000$ steps for a given β. So for 50 values of β, the expected completion time is 3 hours in total.

5.2 Biham-Middleton-Levine traffic model

We also implemented Biham-Middleton-Levine traffic model in both R, with computationally intensive parts written in C code that are called from R and in pure, type declared Lisp code. The results again indicate that the Lisp code out-performed the combination of R and C code. While both R implementation could be further optimized, a reasonable amount was done using profiling in R and then recoding the bottlenecks in C.

6 Syntax

Lisp is a powerful computing language which provides a rich set of resources for programmers. Despite this, many programmers have difficulty with it because of its syntax. The S expression

```
sum(x)/length(x)
```

is represented in Lisp by the "s-expression"

```
(/ (sum x) (length x))
```

It is our intent to provide a thin layer of syntax over Lisp to provide a comfortable environment for carrying out data analysis. Although we intend

to change the appearance of Lisp, it is important that the layer which does this be as thin as possible. This would make it possible for users to work in Lisp, should they choose to do so. This would make the applications developed in the framework useful to the Lisp community as well as to statisticians.

There are a number of ways in which the syntax layer could be implemented. A standard LALR parser generator is available and this could be used to translate an S-like syntax into Lisp. As an alternative, we (together with Brendan McArdle of the University of Auckland) are examining the use of a PEG (parsing expression grammar) based parser. Such parsers provide the ability to extend the grammar at run-time which is useful for experimentation.

The syntax of the language is not yet finalised, but we would expect that a simple function definition such as the one below on the left would be translated to a Lisp form given on the right.

```
defun sum(x)
{
  local s = 0
  do i = 1, n {
    s = s + x[i]
  }
  s
}
```

```
(defun sum (x)
  (let ((s 0))
     (doloop (i 1 (length x))
       (setf s (+ s (elt x i))))
     s))
```

Here, `doloop` and `elt` are Lisp macros which implement a Fortran-style do-loop and 1-based element access for vectors.

Adding declarations to the original code would simply add corresponding declarations to the Lisp code. The annotated version of sum function above is given below on the left and the Lisp translation on the right.

```
defun sum(double[*] x)
{
  local double s = 0
  do i = 1, n {
    s = s + x[i]
  }
  s
}
```

```
(defun sum (x)
  (declare
    (type (simple-array double (*))
        x))
  (let ((s 0))
       (declare (type double s))
    (doloop (i 1 (length x))
      (setf s (+ s (elt x i))))
    s))
```

In fact, we will probably use macros to provide specialized versions for the different data types from a single “template”.

7 Other issues

Memory consumption & copying: As we have mentioned, the pass-by-value semantics of R impose a significant performance penalty. Moving to a pass-by-reference approach would avoid this but involve a very different style of programming. For common, interactive use this may not be desirable, but also would not be a significant issue. For more computationally intensive tasks and "production analyses", the approach may be very beneficial. So too would be a computational model that facilitated working on data sets record at a time or in blocks. This approach has been used very effectively in SAS, for example. We plan on making streaming data and out-of-memory computations a significant part of the fundamental framework. By combining pass-by-reference with a flexible, fast programming language and data delivery mechanism for streaming data, we expect that statisticians and others can use the same tools for interactive, exploratory data analysis and intensive, production-level data processing and mining tasks.

Parallel computing: Parallel computing using multiple cores executing code concurrently with shared memory is becoming increasingly important. Many statistical methods are "embarrassingly parallel" and will benefit greatly from such facilities. Thus, we want to be able use a high-level language to express parallel algorithms. Progress on this front has been slow in R for various reasons. By adopting another community's engine, i.e. SBCL or Allegro Lisp, we inherit much of the work that is already done to provide user-level parallel facilities which are close to completion for the different platforms. Additionally, some of the commercial vendors of Lisp platforms have rich thread support. Further, we expect that there will be advances in compiler technology in general, and implemented in Lisp systems, for identifying and automating aspects of parallelism that we are unlikely to achieve within the statistical community alone.

Backward compatibility?: The R community is already large and growing. There are over 1000 contributed R packages on CRAN (`www.r-project.org`), 150 from BioConductor (`www.bioconductor.org`) and 40 from Omegahat (`www.omegahat.org`). It is a not a trivial decision to embark on building a new system and losing access to this code. So backward-compatibility is an important early decision. We could attempt to re-implement R on a Lisp foundation and this would likely lead to improvements in performance. However, we feel that it is better to move to a new computational model. But, we might still implement a Lisp-based R interpreter that can run concurrently within the Lisp session and can interpret R code. Alternatively, we can develop a translator that converts R code to a Lisp equivalent. And an additional approach is to embed R within Lisp so that we can call R functions directly from within Lisp or our new language. rsbcl (Harmon (2007)) already

provides this interface and allows us access to arbitrary R functionality. We are exploring these different approaches to reusing R code.

Extensibility: The R interpreter is written in C and there is sharp divide between R-language code, interpreted code and the system itself. The fundamental internal data structures are compiled and fixed. With a system written using Lisp, however, we are working in a language that performs run-time compilation to machine code. There is no divide between the "interpreter" and the user-level language. This means that users can introduce new "core" data types within their code and they can be used in the same manner as the core data types provided by our "new" environment. This extensibility allows others outside of the language developers to perform new experiments on the system itself and to disseminate them to others without needing to alter the system. This gives us a great deal of flexibility to handle new tasks and explore alternative approaches to computing. This is also important if we are to foster research on the topic of statistical computing environments themselves, which is necessary if statistical computing is to continue to evolve.

8 Alternative systems

If we are prepared to build a new system, an obvious question is why choose Lisp as the underlying language/environment. Python is becoming increasingly widely used and supported. There is a great deal of advanced design in the upcoming Perl 6/Parrot environment. And each of Perl, Python and Java have extensive add-on modules that are of interest to the scientific and statistical communities.

Each of these systems is a worthy choice on which to build a new system. All are compiled languages in the sense of creating byte-code that is executed on a virtual machine. Java has just-in-time compilation (JIT) which gives it performance comparable to code compiled to machine-instructions. But this is what Lisp provides transparently. And Lisp provides optional type checking, whereas Java *requires* type specification and Python and Perl do no permit type specification (in the standard language). While Java is potentially very fast, its focus on secure code execution and hence array-bound checking introduces a significant overhead for scientific/numerical computing.

The timing results in section 4 indicate that a good Lisp implementation outperforms each of these other higher-level languages. While most of these are more popular than Lisp, we think it is important to engage in ambitious work with greater potential to improve statistical computing and its availability for, and impact on, scientific computing.

We could use a low-level language such as C++ and this would provide us with a potentially better foundation than we currently have in the C-based code underlying R. However, we would still be in the situation of owning our own interpreter and so be responsible for every detail, both now and

in the future. We could build on top of projects such as Root (Brun and Rademakers (1997)), which provides an interactive C++-like language that provides direct access to C++ libraries but we believe that there is a greater benefit to using a rich high-level language which is compiled to machine code rather than an interactive, interpreted C-based language.

Having chosen Lisp, we could elect to use one of the existing statistical systems based on Lisp, i.e. XLisp-Stat, Quail. The choice of Common Lisp will allow us to run the code on any of the standard Common Lisp implementations. For us, the main attraction is the presence of a good, high-performance machine-code compiler. If the code for these systems can be deployed on such a Common Lisp implementation and is not tied to a particular implementation of Lisp, then we will use it (license permitting). Otherwise, it is opportune to design the environment anew with fundamental support for more modern advanced data analysis, e.g. streaming data with out-of-memory algorithms.

9 Conclusion

The statistics community needs to engage in developing computing infrastructure for the modern and future challenges in computationally intensive, data rich analyses. The combination of run- and development-time speed and memory usage is important, and a language that supports optional/incremental type specification helps in both compilation and good programming, while enabling interactive use. We are pursuing Common Lisp, with its several high-performance implementations to develop a framework on which to implement a new statistical computing environment. And in this new development, we are seeking to build in at a fundamental level different, modern computing paradigms (e.g. streaming data and out-of-memory/record-at-a-time algorithms).

Starting the development of a new computing environment using Lisp is not a guaranteed success. Lisp is not a widely used language within the statistics community. And to a large extent, many people are content with their existing environments. This is a long-term project and we are also hoping to engage new and different additional communities and to benefit from their knowledge and activity.

By putting an R-like syntax on Lisp, we feel that the obvious benefits of Lisp can become accessible to a community in need of them, and allow software developed by statisticians to be used in real, high-performance applications.

References

BRUN, R. and RADEMAKERS, F. (1997): ROOT - An Object Oriented Data Analysis Framework, *Nucl. Inst. & Meth. in Phys. Res. A*, 389, 81–86. (Proceedings AIHENP '96 Workshop.)

HARMON, C. (2007): rsbcl - An Interface Between R and Steel Bank Common Lisp. Personal communication.

TIERNEY, L. (2001): Compiling R: A Preliminary Report, *DSC 2001 Proceedings of the 2nd International Workshop on Distributed Statistical Computing.*

Computable Statistical Research and Practice

Anthony Rossini[1,2]

[1] Novartis Pharma AG, Basel, Switzerland, *anthony.rossini@novartis.com*
[2] University of Washington, Seattle WA, USA, *rossini@u.washington.edu*

Abstract. The work of research and applied statisticians is driven by both manual and electronic computation. This computation, which supports our work, can be crudely described by 3 stages: scoping, to assess what can and needs to be done; analysis, where this is accomplished; and reporting, which communicates the results to others. Barriers to the reuse of computations can be found in the translational needs driving the transition between sub-activities; for example, scoping activities are seldom directly reusable during analysis, and there is a limited amount of direct reuse going from analysis to reporting. There is an additional high barrier for translating statistical theory and methodology to practical activities, with both sides (applied and theoretical statisticians) pointing the blame at the other for not using appropriate tools or addressing appropriate needs. In this sense, much statistical research is not really computable, but rather, translatable. This work describes some proposals for exploring novel information technology support to address the translational challenges during transition between stages of statistical practice. These are illustrated in the on-going design of CommonLisp Stat, a platform for investigating the interaction between statistical computing environments supporting research and/or practice.

Keywords: statistical computing, Lisp, expressible research

1 Introduction

Statistical Computing concerns itself with technologies, systems, and methodologies which support both the on-going development as well as proper practice of statistics. The term *statistical practice* as used in this paper is intended to describe the range of activities performed by a statistician, from practical applications focusing on the analysis of data to the development of theory which can support the selection of strategies for such analyses. *Statistical Computing* drives efficiencies in *statistical practice*, and in an essential area of research to drive better ways to express, communicate, and enable the practice of statistics.

Process improvement first requires assessment of current work habits. The processes surrounding statistical practices are no different, and the identification of common activity patterns provide a baseline and context from which to improve. Though there are many possibly characterisations, we will select a 3 step process to describe the core components of the activity; these steps

will be referred to as scoping, analysis/verification, and reporting. Other characterisations can be used, and the methodology in the paper appropriately applied.

The names given to the 3 steps have meanings in other related contexts, so we briefly clarify. Scoping refers to the refinement of the problem to tackle, whether methodological or applied, and describes the preparation work to establish the feasibility of potential approaches. Analysis/verification refers to the critical work for the activity, whether actually working through required proofs or the planned data analysis, both computations, though not necessarily in the sense of using an electronic computer. Reporting refers to generation of materials such as technical reports, journal articles, and presentations based on the computations. The critical point we want to make is that the transistion between the stages can result in losing material which otherwise might be available.

1.1 Biased history

The mid-90s, which predated the popularity of open source software, was a golden time of statistical computing systems. Many experiments and features were implemented, and there were a number of interesting experimental and production systems in use. To a certain extent, many of the implementations have died out or were marginalized. Part of this is due to the success of the R system for statistical computing, which provides many features which provide adequate support for practical statistical applications and research. An evolutionary process has resulted in selective pressure for R and eliminated a number of interesting alternatives.

This is not to say that the field is dead. There are some very interesting and more subtle improvements and experiments found in some of the R packages. These user interfaces provide interesting and more direct experiments focused on enhancing particular practices. One example is the `Rcmdr` package which provides workflow support for basic statistical analysis. This builds on the usual introductory statistical package by mixing GUI and script work. Another set of examples can be found in the tools supporting graphical models. The packages `deal` and `dynamicGraph` contain user interface support for the analysis of data through bayesian networks (directed graphs) and other graphical model methodologies.

However, current statistical computing research and experiments tend towards the applied and target deliverables over exploration of new ideas and approaches.

1.2 Considerations for design

The goal of the research program described in this paper is to construct a system for experimenting with different data analysis workflows. Rather than focusing on a limited or single workflow which a focused system such as R (R

Core, 2008) or Mondrian (Theus, 2002) provide, this system works to focus on workflow construction, though at the initial sacrifice of efficiency.

The key consideration we have is on work-flow support. In particular, we focus on the identification of information surrounding the development of theorems or data analyses which can be captured and reused later. Support is required both for the experimentation and development of theorems as well as for the application of these to data analysis.

The capabilities which such a platform should support include features for flexible model specification and manipulation, numerical support for linear algebra and optimization, reproducibility methods such as Literate Statistical Practice (Rossini, 2001), reporting capabilities such as visualization and description of models and data, and algorithm description and characterization support. One enabling feature would be the means to precisely describe a model and optimization features which might provide a characterization through the description ("source code") of the algorithm of the theoretical and practical features of a particular statistical procedure.

Programming language features and libraries can be used to implement many of the Computational needs, but there needs to be a separation between the realizations and instances of these features and the resulting capabilities. From a systems design specification, the features form the user specifications and functional requirements of such a system, while the implementation approach selects the tools that can be used to implement this. At the basic level, any reasonable programming environment can be used to support statistical computing innovations, but that is not the point. Through the description of features which enhance support for statistical data analysis, we can select a programming platform primarily driving the breadth of focus of the goals of a particular research program, or the depth of exploring or using the capabilities of a specific computational aspect.

While the general overview presented just now provides a high-level view, many of the specific details are perhaps poorly conveyed. We will consider the following use-cases to drive requirements:

1. The consulting statistician who requires a wide toolbox of procedures as well as clear reporting and summarization tools
2. A research statistician whose primary interest is the methodology being developed rather than the application
3. The generalist who is interested in technology-transfer, studying comparative procedures and equally interested in determining which procedures in their own toolbox are acceptable as well as in getting the right level of an answer to client and collaborator questions delivered at the right time.

Wickham and Rossini (2008, in preparation) describe Statistical Design Patterns, a structural approach to describing practical statistical activities which can be additionally used to characterize statistical computing system features. Once the features have been characterized in a value-free manner,

opinions and beliefs can be placed on the structure to ascribe qualities and rankings to it. We can use those concepts to both gauge the capabilities of the current available systems as well plan future goals. Values can then be laid on using context.

The next section of the paper describes some of the capabilities that we aspire to in a statistical computing platform. We follow that by illustrating how these capabilities are useful in supporting statistical activities. CommonLisp Stat design, status, and goals is described next, and we close by describing tactics to achieve our long-term goals.

2 Capabilities

Statistical modeling activities strive towards the pinnacle of creating knowledge, i.e. information and including the uncertainty that we comprehend within the information. Moving along the knowledge pathway from data, which could come in a numerical form from experiments or literature, or in the form of quantitative opinions from interested parties or experts, to information, where this is organized in some form, to knowledge, where this is then translated into informed uncertainty, is the process that we would like to support. There is a range of functionality that can be automated through computing systems to support this pathway, and we will cover a few of these aspects.

Statistical Design Patterns (Wickham and Rossini, 2008, in preparation) describes a framework for patterns, procedures, and activities which a statistician would engage in. It describes an ontological structure, i.e. a framework for constructing a vocabulary, which is intended for use in describing activities which can then have values assigned to them. This is useful for retroactive quantitative decision analysis regarding both existing support systems such as R as well as activities such as statistical consulting, explaining the differences between pragmatic approaches to statistical consulting, as well as proactive specification and valuing of desired features in a new system, which is the application desired here.

2.1 Numerics

Numerical capabilities is perhaps the most obvious capability required, and by this we refer to precision of calculation and infrastructure such as optimization and numerical linear algebra facilities to provide building blocks for constructing numerically suitable algorithmic implementations.

Historically this has served as a reachable and obvious value proposition, dating back to World War 2 and earlier when calculators, both mechanical and human were used to take care of calculations which were required for accurate warfare. More recent work has used modern computers to make

models and predictions of large scale phenomena ranging from the human body in the context of health to financial and weather systems.

Many books and articles describe the importance of this capability; in fact, it should be quite obvious. Traditionally, statistical computing has been concerned with proper numerical computation, a focused attribute, rather than the general problem we are striving to address, which is the communication of the algorithm being followed for computation along with the corresponding knowledge (estimates and uncertainty) that should be conveyed.

2.2 Reproducibility and reporting

Literate Statistical Practice (Rossini 2001; Rossini and Leisch; 2003) embodies the reporting and reproducibility aspects. This has also been covered in later work by Gentleman and Temple-Lang, as well as in domain specific work, for example targeting Econometrics and Epidemiology.

There are two key pieces: the first is computational reproducibility, also coined "reproducible research", and this has been covered to a great extent in the literature. The approach has been applied to a myriad of uses in both teaching data analysis as well as describing the practical issues faced in applied statistical data analysis.

However, the second key concept, that of reporting, is not well covered. While the reporting of substantive results has leveraged such literate practice, the reporting of computational methods has seen much less use of this approach, with the precise communication of the computational implementation and details of the practical challenges often glossed over or left out. While many research statisticians tend to use mathematical tricks to simplify proofs to the bare essentials, honest statistical computation often requires just the opposite, that is the accommodation of the full range of possibilities.

2.3 Algorithm expressibility

The description and characterization of algorithms is essential for better communication of knowledge about statistical practice. It is very common to use the equivalent of advertisements or "bullet-point arguments" to describe statistical research rather than precise characterizations of the implementation and practical ramifications.

The implementation of expressibility and characterization can be done through the use of computer science programming language theory such as through the use of appropriately designed object systems, method and function selection and dispatch, the appropriate use of selective and non-selective context, and understanding the limitations of modularity.

2.4 Data management

Computable, statistically oriented data structures are critical to simplifying the analysis of data. While traditionally considered to be the realm of infor-

matics, this is a key prerequisite to enabling proper statistical analysis. The quality of this activity can simplify the communication of computable data analysis. Practically, this falls into the practice of data storage, integration, annotation, and related management activities.

3 Lisp and statistical computing

There has been a strange interest in Lisp-based systems by a number of statistical computing researchers in the last few years. While none of these has become a functioning, widely used product, there are some experiments which we hope will trickle down into production systems such as R and SAS.

Common Lisp is a particular Lisp-based language which has its roots dating back to the beginning of computing languages. Information about its history and development is readily available on the Internet. Common Lisp is similar to Fortran, C, and C++ in being specified by a formal international standard, rather than by a particular or "similar family" of implementations , as Perl, Python, and Ruby.

3.1 LispStat

While S was clearly the first platform for interactive statistical data analysis, it was closely followed by LispStat (Tierney, 1989) which provided a platform for both modeling and graphics. LispStat was built on top of a small portable Lisp system called XLisp. More information on this system can be found in the corresponding book as well as on the WWW.

Around this time, 1989-1990, Tierney also experimented with moving LispStat to the Common Lisp platform. The current work prospectively described in this paper starts with the preliminary LispStat implementation, with the goal of reusing some of the prior work implemented for XLispStat.

With apologies to John Chambers, Lisp is definitively the original **Programming with Data** language, though data is used here in the computer science rather than statistical sense. The distinction between operation and data is blurred to the point that one can be the other and the other can be one. This is no different than statistical practice, where some times the data varies over a particular statistical procedure, and other times the statistical procedures vary across a dataset.

3.2 CommonLisp Stat

CommonLisp Stat (CLS) is a statistical environment whose design is focused towards exploring interactive and high-throughput interfaces for statistically-oriented data analysis. It is being built to support the open source CLISP, SBCL, and CMUCL Common Lisp environments, striving for portability and clarity over efficiency and terseness.

As the system specification is currently in flux, we have tried to point out 2 pieces which are worth consideration. The first are the design goals, and the second is the approach of considering modules and patterns for prototyping a system to experiment with different approaches for specifying statistical expressions.

3.3 Design goals

Systems design of this environment is focused on expression of statistical concepts in a clean manner; we rely on the semantics of Lisp to ensure closure around the ideas conveyed in an computational expression. The use of the word closure can be considered as a play on words – both in the sense of knowledge completion as well as in the computer-science sense of the word closure, i.e. an environment with associated data and actions (Gentleman and Ihaka, 2000).

The list of primary design goals is:

- clear denotation of concepts: leverage Lisp syntax with parentheses to demarcate.
- flexibility and modularity of how data and statistical procedures interact.
- feature statistical concepts as first class stand-alone components: for example, a model for centrality should be separate from the inference engine (Frequentist / Bayesian / Decision-theoretic) that is employed, which should be separate from the optimization approach taken to make a particular philosophical inference on the centrality model.
- use of computer science concepts to cleanly describe the functionality and limitations of implementation approaches: the example of scoping by Gentleman and Ihaka (2000), as well as the use of such concepts as predicate-dispatch, can enhance the expressibility of the implemented algorithm.

Modularity of the system is a critical component for driving forward research in statistical computing. As described in the last goal, experimentation with how computer programming languages can work is a central component of statistical computing which can lead to comprehending what features of a language have better qualities for efficient execution, clean communication, and understanding of the implied limitations and required assumptions for both the expression and implementation within statistical research as well as in the applied practice of statistics.

3.4 Data-driven results

With respect to placing an emphasis on the translation between stages of a statistical activity, we need to let data structures which describe the statistical data, through relational schemas as found in SQL and similar database language, imply the possibility of statistical relatedness, as found in the concept of variance.n

4 Epilogue

We bring a sense of closure to this paper, but not to the goals behind this paper, claiming that knowledge, both contained in statistical research as well as due to the statistical analysis of substantive data, is generally poorly expressed in a computational setting. In fact, similar hopes are stated for technologies such as Web 2.0. While the selection of tools for this research program (Common Lisp as a platform, LispStat as a starting point, Statistical Design Patterns to quantify value) can be challenged, the importance of striving towards computational expressions which are clear and unambiguous should be obvious. Who hasn't had a problem understanding the conceptual and computational nuances of theory when it gets supported in a practical manner?

Fitted statistical models, both the best estimated versions along with their uncertainty, form the basis of knowledge both within the realm of statistics as well as by acting as a major contributor towards scientific inquiry. There is a dearth of research in this area, and new tools and ideas are critical to push the envelope towards better, self-expressed, statistical computing.

References

GENTLEMAN, R. and IHAKA, R. (2000) Lexical scope and statistical computing. *Journal of Computational and Graphical Statistics, 9:491-508.*

R DEVELOPMENT CORE TEAM (2008) R: A Language and Environment for Statistical Computing *R Foundation for Statistical Computing.* http://www.r-project.org/ ISBN 3-900051-07-0

ROSSINI, A.J. (2001): Literate Statistical Practice. In: Hornik, K. and Leisch, F. (eds): *Proceedings of the International Workshop on Distributed Statistical Computing (DSC), 2001.* URL http://www.ci.tuwien.ac.at/Conferences/DSC-2001. ISSN 1609-395X.

ROSSINI, A.J. and LEISCH, F. (2003) Literate Statistical Practice. *UW Biostatistics Working Paper Series, 194* http://www.bepress.com/uwbiostat/paper194

WICKHAM, H. and ROSSINI, A.J. (2008): Statistical Design Patterns. *in preparation.*

THEUS, M. (2002): Interactive Data Visualization using Mondrian. *Journal of Statistical Software, 7, 11:1–9.*

TIERNEY, L. (1991): *LispStat.* John Wiley & Sons, New York.

Implicit and Explicit Parallel Computing in R

Luke Tierney

Department of Statistics and Actuarial Science
University of Iowa, Iowa City, IA 52242, U.S.A., *luke@stat.uiowa.edu*

Abstract. This paper outlines two approaches to introducing parallel computing to the R statistical computing environment. The first approach is based on implicitly parallelizing basic R operations, such as vectorized arithmetic operations; this is suitable for taking advantage of multi-core processors with shared memory. The second approach is based on developing a small set of explicit parallel computation directives and is most useful in a distributed memory framework.

Keywords: shared memory parallel computing, distributed memory parallel computing, vectorized arithmetic

1 Introduction

With increasing availability of muiti-core processors as well as computational clusters it is useful to explore ways in which the R system can be modified or extended so statisticians can take advantage of these resources and speed up their computations or make computations feasible that would otherwise take too long to complete. While a major rewrite of R to take advantage of parallel resources would provide many interesting opportunities and challenges, the current objectives are more modest: to see what can be accomplished with reasonable levels of developer effort within the current R design. Several approaches are possible. One approach is to modify basic vectorized operations to automatically split computations among all available processors. This implicit approach tends to be most effective in a shared memory context. No user input is required, thus making this approach simple to use.

A second approach is to develop some simple explicit parallel computing constructs that can be used for expressing parallel computations explicitly. This approach requires some learning on the part of users but will allow coarser-grained parallelization and is also suitable for use in distributed memory settings that often allow much larger numbers of processors to be used.

This paper outlines efforts adding both forms of parallel computing to R. The next section discusses one approach to parallelizing some basic arithmetic operations. The third section outlines a simple explicit parallel computing framework. The final section provides some discussion and outlines future work.

2 Implicit parallel computing in R

The basic idea in implicit parallelization of a high level language like R is to identify operations that are computationally intensive and to arrange for the work of these operations to be divided up among several processors without requiring explicit intervention on the part of the user. This involves the use of multiple threads. In some cases libraries using this approach are readily available and can be used by linking them into R. For example, most linear algebra computations are based on the basic linear algepra subroutines library, or BLAS. R provides its own basic implementation derived from an open source reference implementation, but makes it easy to substitute an altrnate implementation, such as a harware vendor library or one from the ATLAS project (Whaley and Petitet (2005)). A number of BLAS implementations provide threaded versions that try to improve performance by using multiple threads. A major challenge is that there is overhead associated with synchronization of threads, among other things, that can result in threaded versions running slower than non-threaded ones. This has been observed in the use of threaded BLAS libraries.

Another candidate for implicit parallelization is R's vectorized arithmetic operations. The R math library includes many special functions, densities, cumulative distribution functions, and quantile functions. R level versions of these functions apply the functions to all elements of vector arguments. This is currently implemented by a simple loop. If the work of this loop is divided among several processors then the resulting computation may run faster. However, care is needed as there is synchronization overhead, and shared resources (memory, bus, etc.) impose bottlenecks. As a result, while parallelization of vectorized operations will be beneficial for large vectors, it can be harmful for short ones. Careful tuning is needed to ensure that parallelization is only used if it will be helpful.

Figure 1 shows performance measurements for a range of vector sizes and two functions on two 8-processor systems. Some simple empirical observations from these and similar plots for other functions: Times are roughly linear in vector length for each function/OS/thread number combination. The intercepts are roughly the same for all functions on a given platform. If the slope for P processors is s_P then, at least for $P = 2$ and $P = 4$, the approximation $s_p \approx s_1/P$ seems reasonable. Finally, the relative slopes for different functions seem roughly independent of OS/architecture.

These observations motivate a simple strategy: Relative slopes are computed using a number of runs on a range of platforms and recorded. The slope for the normal density function `dnorm` is used as a base line; thus timings are computed in units of single element `dnorm` calculations. Intercepts are estimated for each OS/architecture combination. The two-processor intercept for Linux/AMD/x86_64 is approximately 200 `dnorm` evaluations; for Mac OS X 10.4/Intel/i386 it is around 500. Using this information one can estimate for each function f the value $N_2(f)$ such that using $P = 2$ processors is faster

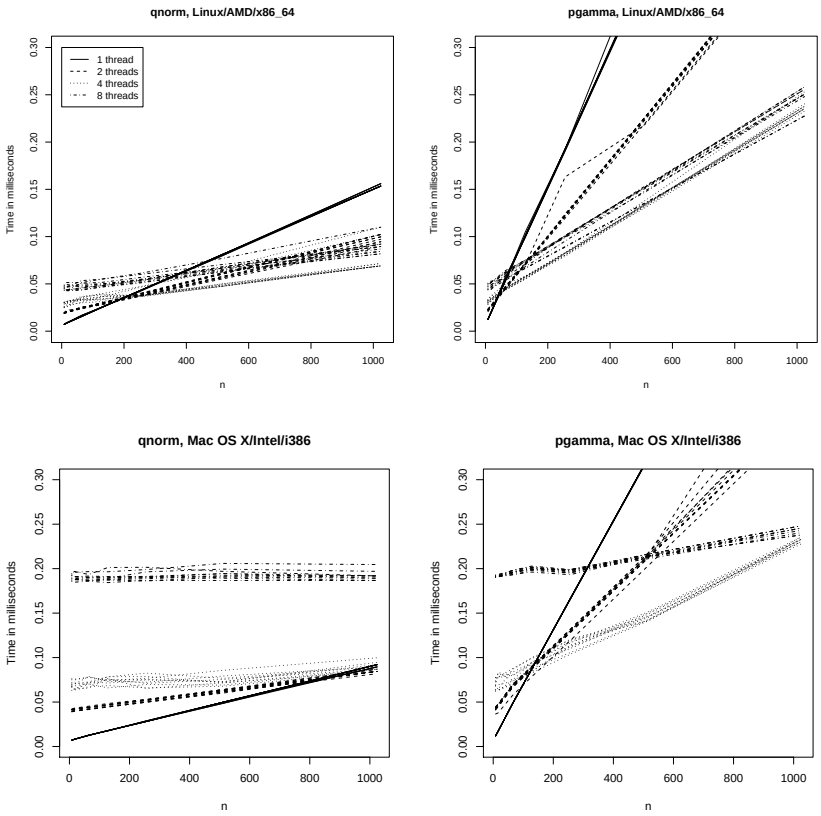

Fig. 1. Timings of vectorized function evaluations for `qnorm` and `pgamma` as a function of vector length for two 8-processor systems. Plots for 10 replications are shown.

than using a single processor for vectors of length $n > N_2(f)$. For $P = 4$ processors we use $N_4(f) = 2N_2(f)$ and for $P = 8$ we use $N_8(f) = 4N_2(f)$.

Figure 2 shows selected values of $N_2(f)$ for a Linux/AMD system. For simple functions like `sqrt` parallel evaluation does not pay for vectors with fewer that $n = 2000$ elements. For `dnorm` the cutoff is around 500. For some very computationally intensive functions, such as `qtukey`, parallel evaluation is useful for vectors of length $n = 2$.

Implementing the approach outlined above involves using threads to evaluate different parts of the basic vectorization loops. One possibility is to directly use a basic threading API, such as `pthreads`, but a better choice is to use Open MP (Chandra et al. 2000). Many commercial compilers as well as gcc 4.2 support Open MP; Redhat has back-ported the Open MP support

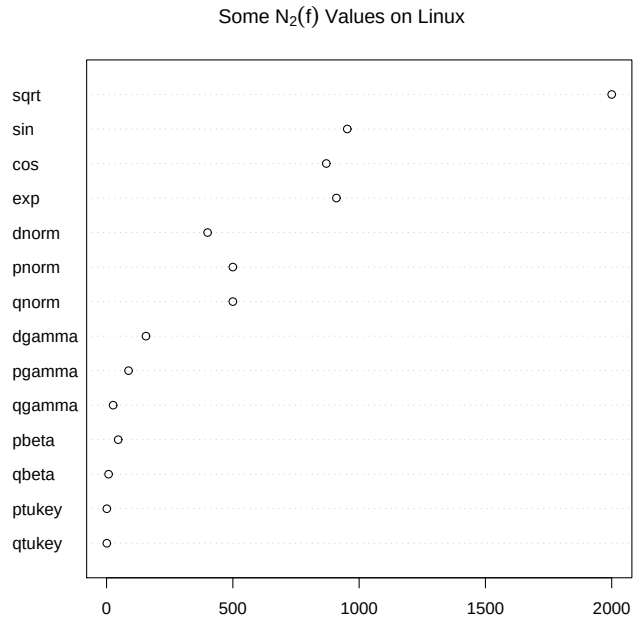

Fig. 2. Selected cutoff levels for switching to parallel evaluation of vectorized functions.

to gcc 4.1 in recent Fedora and Redhat Linux releases. The current MinGW Windows compiler suite also includes Open MP support.

Open MP uses compiler directives (`#pragma` statements in C; FORTRAN uses structured comments) to request parallel implementation of a loop. For example, Figure 3 shows the loop used for vectorizing a function of a single argument along with the Open MP parallelization directive. Functions of more

```
#pragma omp parallel for if (P > 0) num_threads(P) \
        default(shared) private(i) reduction(&&:naflag)
    for (i = 0; i < n; i++) {
        double ai = a[i];
        MATH1_LOOP_BODY(y[i], f(ai), ai, naflag);
    }
```

Fig. 3. Vectorization loop for function of one argument with Open MP parallelization directive.

than one argument are somewhat more complicated because of conventions

for recycling shorter arguments. A compiler that does not support Open MP will ignore the `omp` directive and compile this as a standard sequential loop. If the compiler supports Open MP and is asked to use it, then this will be compiled to use the number of threads specified by the variable P.

Use of Open MP eliminates the need to manually manage threads, but some effort is still needed. Only loops with simple control structure can be parallelized by Open MP, which requires rewriting some of the loops used in the standard R code. Also, it is essential that the functions being called are safe to call from multiple threads. For this to be true these functions cannot use read/write global variables, call R's memory manager, signal warnings or errors, or check for user interrupts. Even creating internationalized error messages can be problematic as the subroutines that do this are not guaranteed to be thread-safe. Almost all functions in the basic R math library are either thread-safe or easily modified to be thread-safe. Exceptions are the Bessel functions and the Wilcoxon and signed rank functions.

A preliminary implementation of the approach outlined here is available as a package `pnmath`. Loading this package replaces the standard vectorized functions in R by parallelized ones. For Linux and Mac OS X predetermined intercept calibrations are used; for other platforms a calibration test is run at package load time. The package requires a version of gcc that supports Open MP and allows `dlopen` to be used with the support library `libgomp`.

3 Explicit parallel computing in R

Several packages are available to support explicit parallel computing in R, including `Rmpi`, `rpvm`, `nws`, and `snow`. These packages are based on the idea of coordinating several separate R processes running either on a single multicore machine or on several machines connected by a network. The packages `Rmpi` and `rpvm` provide access from R to most features of the MPI (Pacheco, 1997) and PVM (Geist et al. 1994) message passing systems and as such are powerful frameworks for parallel computing. However they are not easy to use, and many parallel computations can be handled using a simpler framework. The goal of the `snow` package is to provide such a simple framework that is easy to use in an interactive context and is capable of expressing a number of interesting parallel computations. A particular goal is to provide a framework in which user code cannot create a deadlock situation, a common error in parallel code written with many very general parallel frameworks. A more extensive review of the `snow` framework is given in Rossini, Tierney, and Li (2007). The `snow` package is designed to operate on top of a more basic communication mechanism; currently supported mechanisms are sockets, PVM, and MPI.

Figure 4 shows a simple snow session. The call to the function `makeCluster` creates a cluster of two worker R processes. `clusterCall` calls a specified function with zero or more specified arguments in each worker process and

```
> cl <- makeCluster(2)
> clusterCall(cl, function() Sys,info()["nodename"])
[[1]]
[1] "node02"
[[2]]
[1] "node03"
> clusterApply(cl, 1:2, function(x) x + 1)
[[1]]
[1] 2
[[2]]
[1] 3
> stopCluster(cl)
```

Fig. 4. A minimal `snow` session.

returns a list of the results. `clusterApply` is a version of `lapply` that applies the specified function to each element of the list or vector argument, one element per worker process, and returns a list of the results. Finally, `stopCluster` shuts down the worker processes.

`clusterCall` and `clusterApply` are the two basic functions from which other functions are constructed. Higher level functions include `parLapply`, `parApply`, and `parMap`. These are parallel versions of the standard functions `lapply`, `apply`, and `Map`, respectively. A simple rule is used to partition input into roughly equal sized batches, with the number of batches equal to the number of worker processes. The process of converting a sequential R program to a parallel one using `snow` usually involves identifying a loop that can benefit from parallelization, rewriting the loop in terms of a function such as `lapply`, and, once the rewrite has been debugged, replacing the `lapply` call by a call to `parLapply`.

An important issue that needs to be addressed in parallel statistical computing is pseudo-random number generation. If one uses standard R generators then there is a very good chance, though no guarantee, that all R worker processes will see identical random number streams. If identical streams are desired, as they might be at times for blocking purposes, then this can be assured by setting a common seed. If, as is more commonly the case, one wants to treat the workers as producing independent streams then it is best to use R's ability to replace the basic random number generator along with one of several packages that are designed for parallel random number generation. `snow` supports using two of these packages, with the default being the `rlecuyer` interface to the random number streams library of L'Ecyuer et al. (2002). The function `clusterSetupRNG` is used to set up independent random number streams on all cluster processes.

Figure 5 shows a comparison of a sequential bootstrap calculation and a parallel one using a cluster of 10 worker processes. The elapsed time of the parallel version is approximately one tenth of the elapsed time for the

```
## sequential version:
> R <- 1000
> system.time(nuke.boot <-
+             boot(nuke.data, nuke.fun, R=R, m=1,
+                  fit.pred=new.fit, x.pred=new.data))
   user  system elapsed
 12.703   0.001  12.706
## Parallel version, using 10 processes:
> clusterEvalQ(cl,library(boot))
> clusterSetupRNG(cl)
> system.time(cl.nuke.boot <-
+             clusterCall(cl,boot,nuke.data, nuke.fun,
+                         R=R/length(cl), m=1,
+                         fit.pred=new.fit, x.pred=new.data))
   user  system elapsed
  0.009   0.004   1.246
```

Fig. 5. Bootstrap example from the `boot` help page.

sequential version. The function `clusterEvalQ` is a utility function used to evaluate an expression on all worker processes; in this case it is used to ensure that the `boot` package is loaded on all worker processes.

Linear performance speedup as seen in this bootstrap example is not always achievable. One issue is the cost of communication. Data is transferred to and from the workers. If the amount of computation on each worker is not large relative to the communication overhead then speedup will be less, and in extreme cases parallel versions can run slower than single process sequential versions. Another issue is that sometimes the time needed by each worker to perform its task may vary from worker to worker, either because of variations in tasks themselves or because of differing load conditions on the machines involved. This can be addressed by load balancing. Currently `snow` provides one function for doing this, `clusterApplyLB`, a load balancing version of `clusterApply`. For a cluster of P processes and a vector on $n > P$ elements this function assigns the first P jobs to the P processes, and then assigns job $P+1$ to the first process to complete its work, job $P+2$ to the second process to complete its work, and so on. As a result, the particular worker process that handles a given task is non-deterministic. This can create complications with simulations if random number streams are assigned to processes but can be very useful for non-stochastic applications.

4 Discussion and future work

Work on implicit parallelization within R is still in its early stages. The parallel vectorized math library package described in Section 2 above is a first step. Over the next few months this work will be folded into the base R distri-

bution, and we will explore other possibilities of using implicit parallelization implemented via Open MP. Some reduction operations, such as row or column sum calculations, may also be amenable to this approach. One aspect that will also be explored is whether the parallelization framework developed within the R internals, such as the loop shown in Figure 3, can be made available to package writers so package writers can easily define their own parallel vectorized functions without reimplementing what has already been done for the built-in functions.

Implicit parallelization is easiest from the user point of view as it requires no special user action. However in an interpreted framework such as R implicit parallelization is only easily applied at a very fine level of granularity of individual vector or linear algebra operations. This means that speed improvements are only achievable for large data sets. It is hoped that work currently underway on developing a byte code compiler for R may allow parallelization to be moved to a somewhat higher level of granularity by fusing together several vector operations. This should significantly reduce the synchronization overhead and allow parallel computation for much smaller data sets. Compilation may also help in automatically parallelizing certain simple uses of the `apply` family of functions.

Explicit parallel computing can more easily achieve substantial speedups both because it is possible to work at higher levels of granularity and because it is possible to bring to bear larger numbers of processors (though the number of processors available in multi-core computers is likely to increase in the next few years; see for example Asanovic et al. 2006). More work is needed on the interface provided by the `snow` framework. One area under consideration is a redesign of the approach to load balancing to make it possible for all parallel functions to optionally use load balancing. Another area is the development of tools for measuring and displaying the parallel computation, and the communication overhead in particular. Tools for doing this within PVM and certain MPI frameworks are available, but it should be possible to build on R's own facilities and develop some useful tools that work more easily and on all communications back ends.

The current `snow` framework is already quite effective for implementing a range of parallel algorithms. It can easily handle any computation expressible as a sequence of scatter-compute-gather steps. A useful addition would be to allow some intermediate results to remain on the worker processes between such scatter-compute-gather steps, but to be sure these results are cleaned up after a complete computation. Also useful would be the ability to request limited transfer of data between nodes. In Markov random field simulations for example, one might divide the field among workers and need to exchange boundary information in between iterations. Both of these ideas fit well within a formalism known as bulk synchronous parallel computing (BSP; Bisseling 2005). Like `snow`, the BSP model is designed so code using the model cannot create deadlock situations and is thus a good fit for generalizing the `snow`

model. Extensions to `snow` to support the BSP model are currently being explored.

More extensive rewriting of the R implementation might enable the integration of more advanced parallel libraries, such as ScaLAPACK (Blackford et al. (1997)), and more advanced parallel programming approaches. This is the subject of future research.

5 Acknowledgements

This work was supported in part by National Science Foundation grant DMS 06-04593. Some of the computations for this paper were performed on equipment funded by National Science Foundation grant DMS 06-18883.

References

ASANOVIC, K., BODIK, R., CATANZARO, B.C., GEBIS, J.J., HUSBANDS, P., KEUTZER, K., PATTERSON, D.A., PLISHKER, L.W., SHALF, J., WILLIAMS, S.W., YELICK, K. A. (2006): The landscape of parallel computing research: a view from Berkeley, EECS Department, University of California, Berkeley, Technical Report No. UCB/EECS-2006-183.

BISSELING, R.H. (2004): *Parallel Scientific Computation: A Structured Approach using BSP and MPI*, Oxford university Press, Oxford.

BLACKFORD, L.S., CHOI, J., CLEARY, A., D'AZEVEDO, E., DEMMEL, J., DHILLON, I., DONGARRA, J., HAMMARLING, S., HENRY, G., PETITET, A., STANLEY, K., WALKER, D., WHALEY, R.C. (1997): *ScaLAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia.

CHANDRA, R., MENON, R., DAGUM, L., KOHR, D. (2000): *Parallel Programming in OpenMP*. Morgan Kaufmann, San Francisco.

GEIST, A., BEGUELIN, A., DONGARRA, J., JIANG, W. (1994): *PVM: Parallel Virtual Machine*, MIT Press, Cambridge.

L'ECUYER, P., SIMARD, R., CHEN, E.J., KELTON, W.D. (2002): An objected-oriented random-number package with many long streams and substreams, *Operations Research, 50 (6), 1073–1075.*

PACHECO, P. (1997): *Parallel Programming with MPI*, Morgan Kaufmann, San Fransisco.

ROSSINI, A.J., TIERNEY, L., LI, N. (2007): Simple Parallel Statistical Computing in R, *Journal of Computational and Graphical Statistics, 16 (1), 399–420.*

WHALEY, R.C., PETITET, A. (2005): Minimizing development and maintenance costs in supporting persistently optimized BLAS, *Software: Practice and Experience, 35 (2), 101–121.*

Part III

Classification and Clustering of Complex Data

Probabilistic Modeling for Symbolic Data

Hans-Hermann Bock

Institute of Statistics, RWTH Aachen University, 52056 Aachen, Germany,
bock@stochastik.rwth-aachen.de

Abstract. Symbolic data refer to variables whose 'values' might be, e.g., intervals, sets of categories, or even frequency distributions. Symbolic data analysis provides exploratory methods for revealing the structure of such data and proceeds typically by heuristical, even if suggestive methods that generalize criteria and algorithms from classical multivariate statistics. In contrast, this paper proposes to base the analysis of symbolic data on probability models as well and to derive statistical tools by standard methods (such as maximum likelihood). This approach is exemplified for the case of multivariate interval data where we consider minimum volume hypercubes, average intervals, clustering and regression models, also with reference to previous work.

Keywords: symbolic data, interval data, probability models, minimum volume sets, average intervals, clustering, regression

1 Introduction

Starting with the seminal paper by Diday (1988), there is a large number of publications, reports and software tools dealing with the analysis of what is called 'symbolic data', i.e. collections of data vectors whose components are not (only) real numbers or labels (categories) as in classical statistics, but may be intervals, sets of categories, or (empirical, theoretical) frequency distributions (Symbolic Data Analysis, SDA). Most articles dealing with such data, e.g., with their descriptive characterization, the measurement of similarity and dissimilarity, clustering and discrimination methods, etc. (see, e.g., Bock and Diday 2000, Noirhomme and Diday 2007) proceed in an often surprisingly empirical (even if: suggestive) way without referring to any underlying formal structure or general principles as they are provided, e.g., by probabilistic models in classical multivariate statistics where interesting analytical results (e.g., on the performance of methods or the large sample behaviour of estimators) can be derived. Even the relationship to approaches such as 'statistics for set-valued data', 'fuzzy methods', 'imprecise data theory', etc. is often neglected.

In this paper we concentrate on the case of interval-type data and point to concepts and methods that rely on probabilistic models and may be useful or even better alternatives to corresponding 'symbolic' approaches. Throughout we consider a set $\mathcal{O} = \{1, ..., n\}$ of n objects (entities, items) whose

properties are described by p interval-type variables. We denote by $x_1, ..., x_n$ the resulting p-dimensional symbolic data vectors $x_k = (x_{k1}, ..., x_{kp})' = ([a_{k1}, b_{k1}], ..., [a_{kp}, b_{kp}])$ where the entries $x_{kj} = [a_{kj}, b_{kj}]$ are intervals from R^1 with upper/lower boundaries $a_{kj} \leq b_{kj}$ $(j = 1, ..., p,\ k = 1, ..., n)$. Then the k-th recorded object is represented by a hyper-rectangle (hypercube, rectangle, interval) $Q_k = [a_{k1}, b_{k1}] \times \cdots \times [a_{kp}, b_{kp}]$ in R^p[1].

2 Obtaining interval data from minimum volume hypercubes

Consideration of symbolic data is appropriate in situations where the objects under study are not n single individuals, but n specified *groups* (e.g., n cities) of *individuals* for which classical data vectors might be available, but cannot or should not be used for various reasons (e.g., for maintaining anonymity). Then the first step in SDA consists in summarizing, for each $k = 1, ..., n$, the properties of the k-th *group* G in the form of an interval-type vector x_k (a rectangle Q_k), on the basis of the $g = |G|$ single-valued classical data vectors $y_1, ..., y_g \in R^p$ that were recorded for the individuals in G. So we have to find the interval boundaries a_{kj}, b_{kj} within x_k from the individual vectors $y_s = (y_{s1}, ..., y_{sp})'$, $s = 1, ..., g$. For this task a range of empirical methods have been be proposed or implemented in SDA, e.g.:

(1) the **'min/max option'** with $a_{kj} = \min_{s \in G} y_{sj}$, $b_{kj} = \max_{s \in G} y_{sj}$ such that Q_k is the minimum hypercube in R^p that contains all individual vectors $y_1, ..., y_g$.

(2) the **'confidence interval option'**: Here the component-specific intervals are given by $x_{kj} = [m_j - u \cdot s_j, m_j + u \cdot s_j]$ where $m_j := (\sum_{s \in G} y_{sj})/g$ and $s_j^2 := (\sum_{s \in G} (y_{sj} - m_j)^2)/(g - 1)$ are the mean and empirical variance of the g data values of variable j and u an appropriate quantile of the t_{n-1} distribution.

(3) the **'quantile option'** where a_{kj}, b_{kj} are the lower and upper empirical β-quantiles of the g data values $y_{1j}, ..., y_{gj}$ (typically, the quartiles with $\beta = 1/4$); then Q_k contains at least $100 \cdot (1 - 2p\beta)$ % of the data points of G

(4) the **'$(1-\alpha)$-coverage option'** where Q_k is the smallest hypercube in R^p that contains $100 \cdot (1 - \alpha)$ % of the data points $y_1, ..., y_g$.

Whereas the min/max option is quite sensitive to outliers and will typically result in excessively large hypercubes for a large number g of individuals, the $(1 - \alpha)$-coverage option is robust in this respect and insofar much more adequate when defining 'symbolic data vectors' (but is computationally demanding). The following definition 1 shows that (4) is intimately related to various concepts of classical probability where the empirical distribution of

[1] For ease of notation, we will sometimes identify x_k and Q_k.

the values $y_1, ..., y_g$ is replaced by the distribution P of a p-dimensional random vector Y with density f in R^p (w.r.t. the Lebesgue measure λ_p). The subsequent definitions 2 and 3 modify these concepts with a view to SDA and and are illustrated for the ideal case of a two-dimensional standard normal distribution.

Def. 1: A *minimum volume set* (*modal set*) *of level* $\beta$for Y is a measurable subset $S \subset R^p$ that resolves the minimization problem:

$$\lambda_p(S) \rightarrow \min_{S \subset R^p} \quad \text{constrained by} \quad P(Y \in S) \geq \beta \tag{1}$$

where $0 < \beta < 1$ a given threshold (e.g., $\beta = 0.95$).

It was shown by Nuñez-Garcia et al. (2003) that all level sets $A_c := \{y \in \Re^p \mid f(y) \geq c\}$ are minimum volume sets (for the threshold $\alpha := P(Y \in A_c)$). These authors also determine conditions under which the inverse statement holds. Note that modal sets are related to possibility theory as well as to random set theory (see Nuñez-Garcia et al., 2003), and also to the classical 'high-density' or 'contour' clusters of Bock (1974, 1996a) and Hartigan (1975). Estimation methods are described in Scott and Nowak (2006).

In the context of SDA and interval data, the following modification might be considered:

Def. 2: A *minimum volume hypercube of level* β for Y is an interval $S = [a, b] \subset R^p$ that resolves the minimization problem:

$$\lambda_p(S) = \lambda_p([a, b]) \rightarrow \min_{a \leq b} \quad \text{constrained by} \quad P(a \leq Y \leq b) \geq \beta. \tag{2}$$

A third definition of an optimally representative ('prototype') hypercube for Y can be derived from work by Käärik and Pärna (2007). Starting from a distance measure $d(y, z)$ between points $y, z \in R^p$ (typically the Euclidean distance), they define the minimum distance between a point $y \in R^p$ and a set $Q \subset R^p$ by the nearest neighbour distance $D(y, Q) := \min_{z \in Q} d(y, z)$ (that is $= 0$ for all $y \in Q$) and look for a solution of the optimization problem

$$E[D(Y, Q)] = \int_{R^p} min_{z \in Q} \{d(y, z)\} \, dP(y) \quad \rightarrow \quad \min_{Q \in \mathcal{Q}} \tag{3}$$

under the constraint that the set Q belongs to a given (sufficiently large) family $\mathcal{Q}$ of subsets Q from R^p (typically: balls, hypercubes, unions of such sets,...) and has a given coverage $P(Y \in Q) = \beta$ (as an alternative: a given volume $\lambda_p(Q) = v$). In SDA the following special case might be appropriate:

Def. 3: A *prototype hypercube* Q *of level* β (of volume v) for Y is any hypercube $Q = [a, b] \subset R^p$ that resolves (3) with $\mathcal{Q}$ the set of all intervals in R^p, for a given coverage $0 < \beta < 1$ (a given volume $v > 0$).

At first sight, both definitions 2 and 3 might be equally useful for SDA. However, the following simple situation reveals that they can yield qualitatively

quite different solutions for Q (work together with K. Pärna): Let us assume that $Y = (Y_1, Y_2)$ has a two-dimensional standard normal distribution with independent components $Y_1, Y_2 \sim \mathcal{N}(0,1)$ with distribution function Φ and density ϕ. The following theorem shows that the optimum intervals have the form $Q = [-a, +a] \times [b, +b]$ in both cases, but that 'real' rectangles (i.e., with $a \neq b$) may result only for Def. 2, and only for small volumes v. Note that the criterion (3) has here the form:

$$E[D(Y,Q)] = 2(S(a) + S(b)) \tag{4}$$

with $S(a) := \{(a^2+1)(1-\Phi(a)) - a\phi(a)\}$.

Theorem:

(1) Under Def. 1 and for all parameter values $\beta \in (0,1)$ and $v > 0$, an optimum interval of level β (volume v) is a square $Q = [-a, +a]^2$ with equal side lengths $a = b = \Phi^{-1}((1+\sqrt{\beta})/2)$ (respectively $a = b = \sqrt{(v)}/2)$).

(2) Under Def. 2 a centered optimum interval of coverage β (volume v) has the form

$Q = ([-a, +a] \times [-b, +b])$.

(a) It is a square $Q = [-a, a]^2$ if $\beta \geq \beta^* = 0.352...$ ($v \geq v^* = 1.49819...$); then $a = b = \Phi^{-1}((1+\sqrt{\beta})/2)$ (resp. $a = b0\sqrt{(v)}/2)$.

(b) It is a 'typical' rectangle (i.e. with different side lengths) if $\beta < \beta^* = 0.352...$ (resp. $v < v^* = 1.49819...$). Then the side lengths are the solutions $a \neq b$ of the equation

$$a^2(1-\Phi(a)) - a\phi(a) = b^2(1-\Phi(b)) - b\phi(b).$$

The threshold v^* is the solution of $v\Phi(v/4) = 2\phi(v/4)$. The following tables illustrate the optimum configuration for some specifications of parameter values $\beta = P(Y \in Q)$ and $v = \lambda_p(Q)$. Note that for small β or v the solutions (a, b) are not very precise since small changes in a result in large changes of b, but lead only to very small changes in the coverage and probability values.

β	v	Sides $a = b$
0.146	1.00	0.50
0.467	4.00	1.00
0.950	20.07	2.24
0.990	31.58	2.81

Table 1: Optimum Q for Def. 1.

β	v	Side a	Side b
0.051	0.50	0.069	1.809
0.123	1.00	0.193	1.260
0.211	1.49819	0.612	0.612
0.466	4.00	1.000	1.000
0.500	4.46	1.052	1.052
0.950	20.07	2.236	2.236
0.990	31.58	2.806	2.806

Table 2: Optimum Q for Def. 2 and given v.

Remark: Various extensions exist for the case of non-spherical normal distributions.

3 Average intervals and class prototypes

This section deals with another major problem in SDA, i.e. the definition of an 'average' interval (and a 'variance') of n data intervals $Q_k = [a_k, b_k]$ in R^p that characterize a set $\mathcal{O} = \{1, ..., n\}$ of n objects. We recall the common approach in SDA and an approach from geometric probability, and then propose a parametric and probabilistic approach.

a) Centrocubes as optimum class representatives
A basic approach in SDA starts from a dissimilarity measure $D(Q, G)$ between two intervals Q, G in R^p and defines the '*average interval*' ('*class prototype*, '*centrocube*') as an interval $G = [u, v] \subset R^p$ with minimum average deviation in the sense

$$g(C, G) := \sum_{k \in \mathcal{O}} D(Q_k, G) \;\rightarrow\; \min_G \tag{5}$$

(or some modification therof). Typically there will be no explicit solution of this problem, but for some special choices of D (Hausdorff distance, Hausdorff-type L_1 or L_2 distance) an exact solution can be easily obtained (see Chavent & Lechevallier 2002, Chavent 2004, Bock 2005). The minimum value in (5) can be used as a variance measure.

b) Approaches starting from geometric probability
In the framework of geometric probability there exist various proposals to define the average of a random set, based on a measure-theoretical definition of a 'random (closed) set Q' in R^p and its distribution P (see, e.g., Mathéron 1975)[2]. The 'expectation' $E[Q]$ of Q is then defined in a way that retains some useful properties of the classical concept of the 'expectation of a random variable' or maintains their validity at least in an extended sense. A wealth of definitions and properties are presented and surveyed, e.g., in Molchanov (1997), Baddeley and Molchanov (1997, 1998), and Nordhoff (2003), e.g.:

Def. 4: The '*Aumann expectation*' of Q is defined as the set

$$E_{Au}[Q] := \{\; E[Y] \mid Y \text{ is a selection of } Q \;\text{ with } E[||Y||] < \infty] \;\} \;\subset\; R^p \tag{6}$$

where 'a selection of Q' is any random vector Y in R^p with $Y \in Q$ almost surely. – A corresponding variance definition is provided by Kruse (1987).

c) A parametric approach for defining an average interval
A hypercube $G = [a, b]$ is characterized either by its 'lower' and 'upper' vertices $a, b \in R^p$ or, equivalently, by its midpoint $m = (a+b)/2$ and the vector of (semi-)side lengths $\ell = (b-a)/2$. Similarly, a *random hypercube* Q is specified by its random midpoint $M = (\widetilde{M}_1, ..., \widetilde{M}_p) \in R^p$ and its random (semi-)side

[2] In the case of SDA, P will be the empirical probability measure on $\{Q_1, ..., Q_n\}$, assigning mass $1/n$ to each hypercube.

lengths vector $L = (\widetilde{L}_1, ..., \widetilde{L}_p) \in R_+^p$ with a joint distribution $P_\vartheta^{M,L}$ (eventually parametrized by a parameter ϑ). The expected midpoint and side length vectors $\mu := E[M] = (\widetilde{\mu}_1, ..., \widetilde{\mu}_p)'$ and $\lambda := E[L] = (\widetilde{\lambda}_1, ..., \widetilde{\lambda}_p)'$ are used in

Def. 5: The '*parametric average interval*' G of Q is given by the interval

$$PAI[Q] \;:=\; [\; E[M] - u \cdot E[L],\; E[M] + u \cdot E[L]\;] \;=\; [\; \mu - u \cdot \lambda, \mu + u \cdot \lambda\;] \quad (7)$$

where $u > 0$ is a specified constant (typically: $u = 1$).

Remark: For $u = 1$ the average interval (7) is optimum in the sense of minimizing the expected deviation $E[D(G,Q)]$ between $G = [a,b] \hat{=} (m, \ell)$ and the random interval $Q = [A,B] \hat{=} (M,L)$ in case of the dissimilarity measure $D(G,Q) := ||A-a||^2 + ||B-b||^2 = (1/2)\{||M-m||^2 + ||L-\ell||^2\}$ (vertex-type distance). – In the case of the distance $D(G,Q) := \sum_{j=1}^p \{|M_j - m_j| + |L_j - \ell_j|\}$ (Hausdorff-type L_1 distance) the optimum prototype interval is provided by $\widetilde{G} \hat{=} (med\{M\}, med\{L\})$ with coordinates given by the medians of $\widetilde{M}_j$ and $\widetilde{L}_j$ (similarly as in Chavent & Lechevallier 2002, Bock 2005).

Example: We illustrate Def. 5 by assuming that all $2p$ components of (M,L) are stochastically independent with $\widetilde{M}_j \sim \mathcal{N}(\widetilde{\mu}_j, \sigma^2)$ and $\widetilde{L}_j \sim \Gamma(a_j, b_j)$. Then (M,L) has a distribution $\mathcal{I}(\mu, \sigma^2; \alpha, \beta)$, say, with parameters μ, σ^2, $\alpha = (a_1, ..., a_p), \beta = (b_1, ..., b_p)$ and a product density of the form $f(m, \ell; \mu, \sigma^2, \alpha, \beta) = h_1(m; \mu, \sigma^2) \cdot h_2(\ell; \alpha, \beta)$. The expectations are given by $\widetilde{\mu}_j = E[\widetilde{M}_j]$ and $\widetilde{\lambda}_j := E[\widetilde{L}_j] = a_j / b_j$ for $j = 1, ..., p$. – In this case the parametric average interval of Q is given by

$$PAI[Q] := [\; \widetilde{\mu}_1 - u\frac{a_1}{b_1},\; \widetilde{\mu}_1 + u\frac{a_1}{b_1}\;] \times \cdots [\; \widetilde{\mu}_p - u\frac{a_p}{b_p},\; \widetilde{\mu}_p + u\frac{a_p}{b_p}\;]. \quad (8)$$

Parameter estimation: In practice, we have to estimate the unknown parameter vector ϑ in the distribution $P_\vartheta^{M,L}$ from n independent samples $Q_1, ..., Q_n$ of Q. In our example this amounts to estimating the parameters $\mu, \sigma^2, \alpha,$ and β If $m_1, ..., m_n \in R^p$ are the observed midpoints and $\ell_1, ..., \ell_n$ the observed side lengths vectors of $Q_1, ..., Q_n$ with $\ell_k = (\widetilde{\ell}_{k1}, ..., \widetilde{\ell}_{kp})'$, the m.l. estimates are given by

$$\hat{\mu} = \bar{m} := \frac{1}{n}\sum_{k=1}^n m_k, \quad \hat{\sigma}^2 = \frac{1}{np}\sum_{k=1}^n ||m_k - \bar{m}||^2, \quad \widehat{\widetilde{\lambda}}_j = \bar{\ell}_j := \frac{1}{n}\sum_{k=1}^n \widetilde{\ell}_{kj} \quad (9)$$

while the estimates $\hat{a}_j$ and $\hat{b}_j$ are the solutions of the m.l. equations

$$\ln \hat{a}_j - \psi(\hat{a}_j) = ln\left(\bar{\ell}_j / \ell_j^*\right) \qquad \hat{b}_j = \hat{a}_j / \bar{\ell}_j \quad (10)$$

where $\ell_j^* := (\Pi_k \widetilde{\ell}_{kj})^{1/n}$ is the geometric mean of $\widetilde{\ell}_{1j}, ..., \widetilde{\ell}_{nj}$ and $\psi(z) := \Gamma'(z)/\Gamma(z)$ the digamma function (for details see, e.g., Johnson et al. (1994), pp. 360, or Kotz et al. (2006), p. 2625).

The benefits from the proposed parametric approach resides in the fact that we can adapt the distribution type for, e.g., the side lenghts $\widetilde{L}_j$ to the procedure by which the boundaries a_{kj}, b_{kj} of the data intervals have been determined from the sampled individuals (see the options (1) to (4) in section 1). For example, the situation where the side lengths $\widetilde{L}_j$ have an exponential distribution $Exp(b_j)$ with expectation $\widetilde{\lambda}_j = E[\widetilde{L}_j] = 1/b_j$, corresponds to the case $a_j = 1$ in the model above. Then the m.l. estimates are given by (9) and $\hat{b}_j = 1/\overline{\ell}_j$. – If we assume that the side lengths $\widetilde{L}_j$ have a uniform distribution on an interval $[0, \Delta_j]$ from R^p_+, the m.l. estimates of the boundaries Δ_j are given by $\widehat{\Delta}_j := \max_k\{\widetilde{\ell}_{kj}\}$.

Another benefit from the probabilistic approach results insofar as we know or may derive the theoretical properties of the parameter estimates and insofar of the corresponding plug-in versions of the average interval (7) by the standard tools of mathematical statistics. This permits a statistical evaluation of the practically obtained results (which is not possible for the exploratory SDA approach).

4 Parametric probabilistic clustering models for interval data

The parametric approach is particularly useful when deriving clustering methods for symbolic data, i.e., here, for n observed hypercubes $Q_1, ..., Q_n$. In fact, by using parametric distributions for these intervals as in section 3.c, we can formulate probabilistic clustering models directly along the lines of classical cases with single-valued data vectors. Essentially three model options are available, i.e., a 'fixed-partition' model, a 'random partition' model, or a mixture model (see Bock 1996a, 1996b, 1996c).

As a representative example we will consider here the following *symbolic fixed-partition model* for n random intervals $Q_1, ..., Q_n$ characterizing the n objects in $\mathcal{O} := \{1, ..., n\}$:

(1) There exists a fixed unknown partition $\mathcal{C} = (C_1, ..., C_m)$ of $\mathcal{O}$ with a known number m of classes $C_i \subset \mathcal{O}$;

(2) For each class C_i there exist class-specific parameters μ_i, α_i, β_i with $\mu_i \in \Re^p$ and $\alpha_i = (a_{i1}, ..., a_{ip})', \beta_i = (b_{i1}, ..., b_{ip})'$ in $\Re^p_+$, and $\sigma^2 > 0$ such that

(3) all intervals Q_k from the same class C_i have the same distribution:

$$Q_k \sim \mathcal{I}(\mu_i, \sigma^2; \alpha_i, \beta_i) \qquad \text{for } k \in C_i, i = 1, ..., m. \tag{11}$$

Clustering is now conducted by maximizing the likelihood of the n observed data rectangles $Q_1, ..., Q_n$ with respect to the system $\vartheta = (\mu_1, ..., \mu_m, \sigma^2, \alpha_1, ..., \alpha_p, \beta_1, ..., \beta_m)$ of all parameters and to the unknown m-partition $\mathcal{C}$:

$$G(\mathcal{C}, \vartheta) := \Pi_{i=1}^m \; \Pi_{k \in C_i} \; h_1(m_k; \mu_i, \sigma^2) \cdot h_2(\ell_k; \alpha_i, \beta_i) \;\rightarrow\; max_{\mathcal{C}, \vartheta}$$

Taking logarithms leads to the *m.l. clustering criterion*:

$$g(\mathcal{C}, \vartheta) := -\sum_{i=1}^{m} \sum_{k \in C_i} \{\log h_1(m_k; \mu_i, \sigma^2) + \log h_2(\ell_k; \alpha_i, \beta_i)\} \ \rightarrow \ min_{\mathcal{C},\vartheta} \quad (12)$$

Since it is impossible to determine the exact optimum configuration for computational reasons, an optimum m-partition will be approximated by iterative methods, typically by the classical *k-means-type algorithm* (alternating optimization):

Starting from an initial partition $\mathcal{C}^{(0)} = \mathcal{C} = (C_1, ..., C_m)$
(a) we determine, in each class C_i, the m.l. estimates $\hat{\mu}_i$, $\hat{\alpha}_i$, $\hat{\beta}_i$, $\hat{\sigma}^2$ as in (9) and (10),
(b) then build m new classes $C_1^{(1)}, ..., C_m^{(1)}$ by assigning each object k (data interval Q_k) to the class with maximum likelihood such that for $i = 1, ..., m$:

$$C_i^{(1)} := \{\ k \in \mathcal{O} \mid f(m_k, \ell_k; \hat{\mu}_i, \hat{\sigma}^2, \hat{\alpha}_i, \hat{\beta}_i) = \max_{j=1,...,m} f(m_k, \ell_k; \hat{\mu}_j, \hat{\sigma}^2, \hat{\alpha}_j, \hat{\beta}_j)\ \}$$

(c) and iterate (a) and (b) until stationarity.

In contrast to clustering algorithms from common SDA which minimize, e.g., criteria of the type

$$g(\mathcal{C}, \mathcal{G}) := \sum_{i=1}^{m} \sum_{k \in C_i} D(Q_k, G_i) \ \rightarrow \ \min_{\mathcal{C},\mathcal{G}} \quad (13)$$

with respect to all systems $\mathcal{G} = (G_1, ..., G_m)$ of m interval-type class prototypes $G_1, ..., G_m$ and the m-partition $\mathcal{C}$, the clustering approach proposed here, exemplified by (12), avoids the definition of 'optimum class prototype intervals'. Instead it deals with optimum class-specific parameter constellations $(\widetilde{\mu}_i, \alpha_i, \beta_i)$ and may insofar attain a better adaption to the observed data configuration $Q_1, ..., Q_n$.

5 Probabilistic regression models for interval data

There were several attempts in SDA to extend classical regression methods to the symbolic interval data situation. Here we face the basic problem that it is not at all trivial to define a 'linear function of a hypercube', a 'linear structure' among n observed hypercubes $Q_1, ..., Q_n$, a 'linear dependence among two hypercubes' etc. Therefore some of the proposed 'symbolic' regression methods proceed mainly in a heuristic, empirical way without an underlying general principle (see also de Carvalho et al. 2004, Neto et al. 2005).

In contrast, Billard and Diday (2000, 2002) propose a mathematical model for the two-dimensional case that mixes probabilistic aspects with empirical considerations as follows: They consider, within each observed rectangle $Q_k \in$

R^2, a uniform distribution with density f_k, say, and introduce a (virtual) two-dimensional random vector $Z = (X, Y) \in R^2$ with the mixture distribution density $f(x, y) = (\sum_{i=1}^{n} \cdot f_k(x, y))/n$. Then the regression tools from classical statistics are applied to the linear prediction model '$Y = a + bX + e$' for the random variables X, Y (under f), and the resulting parameters (expectations, variances, regression coefficients, correlation,...) are interpreted with a view to $Q_1, ...Q_n$ (see also Billard 2004). - Obviously, this model does not really describe a linear dependence between intervals or rectangles.

In contrast, Gil et al. (2001, 2007), González-Rodríguez et al. (2007) have proposed a probabilistic regression model in the framework of random set theory, with reference to the system $\mathcal{K}$ of all random convex compact sets in R^p. They consider two alternative regression models for two random sets X, Y, both in R^p (typically based on the Minkowski addition *oplus* of sets):

Affine model 1:

There exists a fixed set $A \in \mathcal{K}$ and a scalar $b \in R^1$ (both unknown) such that

$$Y = A \oplus bX := \{\ y = a + bx \mid a \in A, x \in X\ \} \tag{14}$$

Regression model 2:

There exists a fixed set $A \in \mathcal{K}$ and a scalar $b \in R^1$ (both unknown) such that

$$Y|X = x\ \sim\ \epsilon_x \oplus bx\ =\ \{\ \eta + b\xi \mid \eta \in \epsilon_x, \xi \in x\ \} \tag{15}$$

where the random disturbance set ϵ_x belongs to $\mathcal{K}$ and has (for all values x of X) the set A as its *Aumann expectation*:

$$E_{Au}[\ \epsilon_x \mid X = x\ ]\ =\ A \qquad \text{for all } x \in \mathcal{K}. \tag{16}$$

For both models the statistical problem consists in determining a set $A \in \mathcal{K}$ and the scalar $b \in R^1$ such that the totality of predicted pairs $\{(x_k, A \oplus bx_k)\}_{k=1,...,n}$ is close to the totality of all observed pairs $\{(x_k, y_k)\}_{k=1,...,n}$ where for model 2 the data (hypercube) pairs (x_k, y_k) are supposed to fulfil $y_k = \epsilon_k \oplus bx_k$ with a convex set $\epsilon_k \in \mathcal{K}$. Gil et al. (2007) have used a least squares approach in the sense of minimizing the criterion

$$G(A, b)\ :=\ \sum_{k=1}^{n}\ D_W^2(y_k, A \oplus bx_k) \quad \to \min_{A,b} \tag{17}$$

where D_W is a distance between convex sets, and presented the explicit solution of this optimization problem in terms of (set) averages and covariance functions. – Note that in the SDA context, the convex set A will be a fixed hypercube and X a random one, and therefore the restricted hypercube model of Gonzalez-Rodriguez et al. (2006) and the related estimation method can be applied.

References

BADDELEY, A.J., MOLCHANOV, I.S. (1997): On the expected measure of a random set. In: D. Jeulin (Ed.): *Advances in theory and applications of random sets.* World Scientific, Singapore, 3-20.

BADDELEY, A.J., MOLCHANOV, I.S. (1998): Averaging of random sets based on their distance functions. *Journal of Mathematical Imaging and Vision* 8, 79-92.

BILLARD, L. (2004): Dependencies in bivariate interval-valued symbolic data. In: D. Banks, L. House, F.R. McMorris, Ph. Arabie, W. Gaul (Eds.): *Classification, clustering, and data mining applications.* Springer, Heidelberg, 319-324.

BILLARD, L., DIDAY, E. (2000): Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.-P., Groenen, P.J.F., Schader, M. (Eds.): *Data analysis, classification, and related methods.* Springer, Heidelberg, 2000, 369-374.

BILLARD, L., DIDAY, E. (2002): Symbolic regression analysis. In: K. Jajuga, A. Sokolowski, H.-H. Bock (Eds.): *Classification, clustering, and data analysis.* Springer, Heidelberg, 2002, 281-288.

BOCK, H.-H. (1974): *Automatische Klassifikation.* Vandenhoeck & Ruprecht, Göttingen.

BOCK, H.-H. (1996a): Probability models and hypotheses testing in partitioning cluster analysis. In: Ph. Arabie, L. Hubert, G. De Soete (Eds.): *Clustering and classification.* World Science, Singapore, 1996, 377-453.

BOCK, H.-H. (1996b): Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis* 23, 5-28.

BOCK, H.-H. (1996c): Probabilistic models in partitional cluster analysis. In: A. Ferligoj, A. Kramberger (Eds.): *Developments in data analysis.* FDV, Metodoloski zvezki, 12, Ljubljana, Slovenia, 1996, 3-25.

BOCK, H.-H. (2003): Clustering methods and Kohonen maps for symbolic data. *J. of the Japanese Society of Computational Statistics* 15 (2), 217-229.

BOCK, H.-H. (2005): Optimization in symbolic data analysis: dissimilarities, class centers, and clustering. In: D. Baier, R. Decker, L. Schmidt-Thieme (Eds.): *Data analysis and decision support.* Springer, Heidelberg, 3-10.

BOCK, H.-H (2007): Analyzing symbolic data: problems, methods, and perspectives. In: A. Okada, T. Imaizumi, W. Gaul, H.-H. Bock (Eds.): *Proc. of the German Japanese Workshops in Tokyo and Berlin 2005/2006.* Springer, Heidelberg, 2008 (to appear).

BOCK, H.-H., Diday, E. (Eds.) (2000): *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data.* Springer, Heidelberg.

CHAVENT, M. (2004): A Hausdorff distance between hyperrectangles for clustering interval data. In: D. Banks, L. House, F.R. McMorris, P. Arabie and W. Gaul (Eds.): *Classification, clustering, and data mining applications.* Springer, Heidelberg, 333-339.

CHAVENT, M., LECHEVALLIER, Y. (2002): Dynamical clustering of interval data: Optimization of an adequacy criterion based on Hausdorff distance. In: K. Jajuga, A. Sokołowski and H.-H. Bock (Eds.): *Classification, clustering, and data analysis.* Springer, Heidelberg, 53-60.

de CARVALHO, F., NETO, E., TENORIO, C.P. (2004): A new method to fit a linear regression model to interval data. In: S. Biundo, Frühwirth, T.W., Palm, G. (Eds.) *KI 2004: Advances in artificial intelligence.* Springer, Heidelberg, 295-306.

DIDAY, E. (1988): The symbolic approach in clustering and related methods of data analysis: the basic choices. In: H.-H. Bock (Ed.): *Classification and related methods of data analysis.* Proc. First Conference of the International Federation of Classification Societies (IFCS), RWTH Aachen University, Germany, June 29-July 1, 1987. Springer Verlag, Heidelberg, 673-684.

DIDAY, E., NOIRHOMME, M. (Eds.) (2008): *Symbolic data analysis and the SODAS software.* Wiley, New York.

GIL, M.A., LÓPEZ-GARCÍA, M.T., LUBIANO, M.A., MONTENEGRO, M. (2001): Regression and correlation analysis of a linear relation between random intervals. *Test* 10, 183-201.

GIL, M.A., GONZÁLEZ-RODRÍGUEZ, G., COLUBI, A., MONTENEGRO, M. (2007): Testing linear independence in linear models with interval-valued data. *Computational Statistics and Data Analysis* 51, 3002-3015.

GONZÁLEZ-RODRÍGUEZ, G., BLANCO, A., CORRAL, N., COLUBI, A. (2007): Least squares estimation of linear regression models for convex compact random sets. *Advances in Data Analysis and Classification* 1, 67-81.

HARTIGAN, J. (1975): *Clustering algorithms.* Wiley, New York.

JOHNSON, N.L., KOTZ, S., BALAKRISHNAN, N. (1994): *Continuous univariate distributions.* Vol. 1. Wiley, New York.

KÄÄRIK, M., PÄRNA, K. (2007): Approximating distributions by bounded sets. *Acta Applicandae Mathematica* 97, 15-23.

KOTZ, S., BALAKRISHNAN, N., READ, C.B., VIDAKOVIC, B. (2006): *Encyclopedia of statistical sciences.* Vol. 4. Wiley, New York.

KRUSE, R. (1987): On the variance of random sets. *J. of Mathematical Analysis and Applications* 122 (2), 469-473.

MATHÉRON, G. (1975): *Random sets and integral geometry.* Wiley, New York.

MOLCHANOV, I. (1997): Statistical problems for random sets. In: J. Goutsias (Ed.): *Random sets: theory and applications.* Springer, Berlin, 27-45.

NORDHOFF, O. (2003): *Expectation of random intervals.* Diploma thesis. Institute of Statistics, RWTH Aachen University.

NETO, E., de CARVALHO, F., FREIRE, E.S. (2005): Applying constrained linear regression models to predict interval-valued data. In: U. Furbach (Ed.): *KI 2005: Advances in Artificial Intelligence.* Springer, Berlin Heidelberg, 92-106

NUÑEZ-GARCIA, J., KUTALIK, Z., CHO, K.-H., WOLKENHAUER, O. (2003): Level sets and minimum volume sets of probability density functions. *Intern. J. of Approximate Reasoning* 34, 25-47.

SCOTT, C.D., NOWAK, R.D. (2006): Learning minimum volume sets. *J. of Machine Learning Research* 7, 665-704.

Monothetic Divisive Clustering with Geographical Constraints

Marie Chavent[1], Yves Lechevallier[2], Francoise Vernier[3], and Kevin Petit[3]

[1] Université Bordeaux 2, Institut de Mathématiques de Bordeaux, UMR 5251.
146, Rue Léo Saignat, 33076 Bordeaux cedex, France,
chavent@sm.u-bordeaux2.fr
[2] INRIA, Paris-Rocquencourt
78153 Le Chesnay cedex, France, *Yves.Lechevallier@inria.fr*
[3] CEMAGREF-Bordeaux, Unité de recherche ADER
50, Avenue de Verdun, 33612 Cestas, France,
francoise.vernier@bordeaux.cemagref.fr

Abstract. DIVCLUS-T is a descendant hierarchical clustering algorithm based on a monothetic bipartitional approach allowing the dendrogram of the hierarchy to be read as a decision tree. We propose in this paper a new version of this method called C-DIVCLUS-T which is able to take contiguity constraints into account. We apply C-DIVCLUS-T to hydrological areas described by agricultural and environmental variables, in order to take their geographical contiguity into account in the monothetic clustering process.

Keywords: divisive clustering, monothetic cluster, contiguity constraints

1 Introduction

DIVCLUS-T is a divisive and monothetic hierarchical clustering method which proceeds by optimization of a polythetic criterion (Chavent et al. (2007), Chavent (1998)). The bipartitional algorithm and the choice of the cluster to be split are based on the minimization of the within-cluster inertia. The complete enumeration of all possible bipartitions is avoided by using the same monothetic approach as Breiman et al. (1984) who proposed, and used, binary questions in a recursive partitional process, CART, in the context of discrimination and regression. In the context of clustering, there are no predictors and no response variable. Hence DIVCLUS-T is a DIVisive CLUStering method whose output is not a classification nor a regression tree, but a CLUStering-Tree. Because the dendrogram can be read as a decision tree, it simultaneously provides partitions into homogeneous clusters and a simple interpretation of those clusters.

This algorithm, design for classical data (either categorical or numerical), has also been proposed to deal with more complex data (see Chapter 11.2 of Bock and Diday (2000)). The modification concerns the within-cluster inertia criterion which is replaced by a distance-based criterion and the definitions of

binary questions. But with complex data, it is usually not possible to answer directly by *yes* or *no* to a binary question, and the solutions proposed are not always satisfactory.

In this paper we propose an extension of DIVCLUS-T, called C-DIVCLUS-T which is able to take contiguity constraints into account. Because the new criterion defined to include these constraints is a distance-based criterion, C-DIVCLUS-T will be able to deal with complex data. In order to avoid the problem pointed out below concerning the definition of binary questions for complex data, we impose to the variables used in the the binary questions, to be classical. The variables used in the calculation of the distance-based criterion can however have complex descriptions.

Several survey of constrained classification can be found in the literature (see for instance Murtagh (1985), Gordon (1996)). The method proposed here has the specificity to be monothetic and its main advantage is then the simple and natural interpretation of the dendrogram and the clusters of the hierarchy. Of course these monothetic descriptions are also constraints which may deteriorate the quality of the divisions. The price paid by construction in terms of inertia by DIVCLUS-T for this additional interpretation has been studied in Chavent et al. (2007) by applying DIVCLUS-T, Ward and the k-means on six databases from the UCI Machine Learning repository.

In this paper, we present an application of C-DIVCLUS-T to hydrological areas described by agricultural and environmental variables.

2 Definitions and notations

Let $\Omega = \{1, ..., i, ..., n\}$ be of n objects described by p variables $X^1, ..., X^j, ..., X^p$ in a matrix $\mathbf{X}$ of n rows and p columns:

$$\mathbf{X} = (x_i^j) = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} 1 \cdots j \cdots p \\ \left[\begin{array}{c} \cdot \\ \vdots \\ \cdots\ x_i^j\ \cdots \\ \vdots \\ \cdot \end{array}\right. \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \left.\begin{array}{c} \\ \\ \\ \\ \\ \end{array}\right] \end{array}.$$

For classical data, if X^j is numerical then $x_i^j \in \Re$ and if X^j is categorical then $x_i^j \in M^j$, with M^j the set of categories. For complex data, X^j can be described for instance by an interval $x_i^j = [a_i^j, b_i^j]$ or by a set of categories $x_i^j \subseteq M^j$.

A weight w_i is also associated to each object i. If the data result from random sampling with uniform probabilities, the weights are also uniform : $w_i = 1$ for all i. It can however be useful for certain applications, to work with non-uniform weights (reweighted sample, aggregate data).

Let V_1 be a subset of $\{X^1, ..., X^j, ..., X^p\}$ with either classical or complex descriptions. Let $\mathbf{D} = (d_{ii'})_{n\times n}$ be a distance matrix with $d_{ii'}$ a distance (or sometimes a dissimilarity) between two objects i and i'. This distances is calculated on the column of $\mathbf{X}$ corresponding to the subset V_1 of variables. In the rest of this paper, we assume that the matrix $\mathbf{D}$ is standardized ($\forall i, i' \in \Omega$, $d_{ii'} \leq 1$) in the following way: If δ is the distance used to compare i and i' on V_1 we have:

$$d_{ii'} = \frac{\delta(i, i')}{\delta_m}, \tag{1}$$

with $\delta_m = \max_{i,i' \in \Omega} \delta(i, i')$. The criterion W, used at each division to evaluate the homogeneity of the bi-partitions, will be defined from $\mathbf{D}$.

Let V_2 be an other subset of $\{X^1, ..., X^j, ..., X^p\}$. As V_1 is used to calculate the matrix distance $\mathbf{D}$ and then the criterion W, the variables in V_2 are used to define at each division, the set of binary questions inducing the finite number of admissible bi-partitions to evaluate. Thanks to the use of binary questions, the computational complexity of the algorithm is reduced and the best bi-partition, chosen according to the criterion W, is monothetic. We recommend to choose in V_2 variables with classical descriptions, such that the binary questions have clear definitions.

We can note that $V_1 \cap V_2$ is not necessarily empty: the same variable can be used to calculate W and the set of binary questions.

3 DIVCLUST-T algorithm

The goal of DIVCLUST-T algorithm is to split recursively a cluster into two sub-clusters, the algorithm starts from the set of objects Ω and the splitting process is stopped after a number of iterations which may be specified by the user. The output of this divisive clustering algorithm is an indexed hierarchy (dendrogram) which is also a decision tree. More precisely at each recursive step, the descendant hierarchical clustering algorithm DIVCLUS-T:

- splits a cluster C_ℓ into a bipartition $(A_\ell, \bar{A}_\ell)$ which minimizes the distance-based criterion W. In Edward and Cavalli-Sforza (1965) method one chooses the optimal bipartition $(A_\ell, \bar{A}_\ell)$ of C_ℓ among all 2^{n_i-1} possible partitions where n_i is the number of objects belonging C_ℓ. It is clear thant the amount of calculation needed when n_i is large will be prohibitive. DIVCLUST approach reduce the complexity by choosing the best bipartition among all the bipartitions induced by a set of all possible binary questions.
- chooses in the partition P_k the cluster C_ℓ to be split in such a way that the new partition P_{k+1} minimizes the distance-based criterion W.

In the complex data context the difficulty is to define (see Bock and Diday (2000))a distance on the set of complex variables included in the set V_1. In

the following chapter we propose a new distance-based criterion where the geographical contraints are added to the initial distances without changing their calculation.

The binary questions on a numerical or categorical variables of the set V_2 are easily defined (Chavent et al. (2007)). Some approaches, decribed in Chavent (1998) or in the chapter 11.2 of Bock and Diday (2000), give many strategies to construct a set of binary questions on the complex variables included in the set V_2.

4 A distance-based criterion

Let $P_K = \{C_1, \ldots, C_k, \ldots, C_K\}$ be a K-clusters partition of Ω and $\mathbf{D} = (d_{ii'})_{n\times n}$ the distance matrix. A distance-based homogeneity criterion can be defined as:

$$W(P_K) = \sum_{k=1}^{K} D(C_k),$$

with

$$D(C_k) = \sum_{i\in C_k} \sum_{i'\in C_k} \frac{w_i w_{i'}}{2\mu_k} d^2_{ii'}, \tag{2}$$

and $\mu_k = \sum_{i\in C_k} w_i$.

In case of numerical data with uniform weights compared with the Euclidean distance, $W(P_K)$ is the well-known within-clusters sum of squares criterion.

This distance-based criterion has the advantage to avoid centroids, often difficult to define explicitly in case of complex data. But because of the double sum in its definition, it has the drawback to increase the computational complexity.

Let now introduce geographical constraints in this criterion.

4.1 The geographical constraints

In the real application studied in this paper, the objects of Ω have geometrical constraints. Generally speaking, spatial constraints can be represented in a graph $G = (\Omega, E)$ where E is a set of edges (i, i') between two objects of Ω. There will be an edge between i and i' if i' is a neighbor of i.

Let $Q = (q_{ii'})_{n\times n}$ be the adjacency matrix of G where

$$\begin{aligned} &q_{ii'} = 1 \text{ if } (i, i') \in E \,(i' \text{ is a neighbor of } i') \\ &q_{ii'} = 0 \text{ otherwise.} \end{aligned} \tag{3}$$

4.2 The new distance-based criterion

The criterion $D(C_k)$ can be decomposed in the following way:

$$D(C_k) = \sum_{i \in C_k} \frac{w_i}{2\mu_k} D_i(C_k) \text{ where } D_i(C_k) = \sum_{i' \in C_k} w_{i'} d^2_{ii'} \tag{4}$$

The criterion $D_i(C_k)$ measures the proximity (dissimilarity) between the object i and the cluster C_k to which it belongs.

In order to take the geographical constraints into account, the criterion $D_i(C_k)$ is modified and re-written in the following way:

$$\tilde{D}_i(C_k) = \alpha a_i(C_k) + (1-\alpha) b_i(C_k) \tag{5}$$

with,

$$a_i(C_k) = \sum_{i' \in C_k} w_{i'}(1 - q_{ii'}) d^2_{ii'} \tag{6}$$

$$b_i(C_k) = \sum_{i' \notin C_k} w_{i'} q_{ii'} (1 - d^2_{ii'}), \tag{7}$$

and $\alpha \in [0, 1]$.

First we can notice that in the absence of constraints, the adjacency matrix Q is a $n \times n$ null matrix and that $\tilde{D}_i(C_k) = \alpha D_i(C_k)$. Otherwise $\tilde{D}_i(C_k)$ is decomposed into two parts. The first part $a_i(C_k)$ measures the coherence between i and its cluster C_k. It it small when i is similar to the objects in $C_k (d_{ii'} \approx 0)$ and when these objects are neighbor ($q_{ii'} = 0$) of i. The second part $b_i(C_k)$ measures the coherence between i and objects in other clusters than C_k. It is small when i is dissimilar from the objects not in C_k ($d_{ii'} \approx 1$) and when these objects are not neighbors of i ($q_{ii'} = 0$).

In other words, $a_i(C_k)$ measures of a dissimilarity between i and C_k by assigning the value 0 for the neighbors of i and the square of the distance for the other objects belonging to the clusters of i. The second part $b_i(C_k)$ represents a penalty for the neighbors of i which belongs to other clusters.

The new distance-based criterion taking the constraints into account in then:

$$\tilde{W}_\alpha(P_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \frac{w_i}{2\mu_k} (\alpha a_i(C_k) + (1-\alpha) b_i(C_k)). \tag{8}$$

4.3 Study of the parameter α

The parameter α can be chosen by the user (usually, $\alpha = 0.5$) or defined automatically. In this latter case, the idea is to chose α such that $\tilde{W}_\alpha(P_1) = \tilde{W}_\alpha(P_n)$. Indeed, if $\alpha = 1$,

$$\tilde{W}_1(P_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{i' \in C_k} \frac{w_i w_{i'}}{2\mu_k}(1 - q_{ii'})d^2_{ii'}, \tag{9}$$

and $\tilde{W}_1(P_n) = 0$. If $\alpha = 0$,

$$\tilde{W}_0(P_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{i' \notin C_k} \frac{w_i w_{i'}}{2\mu_k} q_{ii'}(1 - d^2_{ii'}), \tag{10}$$

and $\tilde{W}_0(P_1) = 0$.
A compromise is then to take α such that $\tilde{W}_\alpha(P_1) = \tilde{W}_\alpha(P_n)$ which gives:

$$\alpha = \frac{A}{A + B}, \tag{11}$$

and

$$\begin{aligned} A &= \sum_{i \in \Omega} \sum_{i' \in \Omega, i \neq i'} q_{ii'}(1 - d^2_{ii'}), \\ B &= \sum_{i \in \Omega} \sum_{i' \in \Omega} (1 - q_{ii'})d^2_{ii'}. \end{aligned} \tag{12}$$

5 Hydrological areas clustering

Agricultural policies have recently experienced major reformulations and became more and more spatialised. Defining policy priorities requires appropriate tools (indicators, models) with relevant results about ecological and social features of agricultural practices (CEC, 2001[1]). Agri-environmental indicators (AEIs) provide an essential tool for formalizing information from different sources and to address the impact of agricultural production on the environment. These indicators combine information about agricultural activity and environmental conditions (data on climate, soils, slopes, hydrology, etc.). In order to provide helpful results for decision makers, the statistical information on agricultural activity (mainly at the scale of administrative units) has to be transferred to environmentally relevant entities.

An important political issue is currently the implementation of WFD (Water Framework Directive) in European countries. It stresses that an assessment is required to implement efficient measurement programs to preserve or restore the good ecological status of water bodies. The spatial unity (hydrological unit) corresponds to the water body, which is the elementary partition of aquatic environments selected for the water status assessment.

[1] CEC, 2001. Statistical information needed for the indicators to monitor the integration of environmental concerns into the Common Agricultural Policy. Commission of the European Communities. Communication to the Council and the European Parliament, COM 2001, Brussels

A study is carrying out at Cemagref in the context of the SPICOSA[2] project and of the implementation of WFD: the purpose is to define the relevant spatial unit, helpful for the integrated management of the continuum "Pertuis Charentais Sea" and "Charente river basin". We have to define homogeneous areas within the Charente basin to calculate the spatialised AEIs and to implement an hydrological model (SWAT). The questions are: what type of spatial organization can be used to analyze the impact of agriculture on the freshwaters ? Are WFD existing ones (hydrographic units) relevant? Or should new spatial entities be created ?

In this first step, we decide to use the hydrological area (water bodies) as the relevant elementary spatial unit and to analyse all relevant variables at this scale. There are 140 hydrological units within the studied area. The goal is to partition the hydrological units and to obtain some clusters as homogeneous as possible in order to implement AEIs and the SWAT model. Two major types of variables are considered :

- Variables to characterize agricultural activities: because the territorial limits resulting from the environmental zonings established to support the implementation of the WFD are by construction independent of the French administrative geographical area (rgion, canton, commune), we used first the Ra-space method (Zahm and Vernier (2007)) to perform a spatial analysis of agricultural activities at the scale of the hydrological unit defined in the Water Framework Directive.
- Variables to characterize environmental conditions: some other variables are needed to assess the potential risk of agricultural pesticide or nutrients transfer towards surface waters. These data concern structural sensitivity (slope, soil, distance to river,..). We used GIS tools to intersect geographical layers and calculate the values for these variables at the hydrological unit scale.

The 140 hydrological units are then characterized by 14 types of soils (marshland soils, terraces, valleys, artificial area, lakes, different types of groies, clay soils, doucins, limestone soils, clay-limestone soils, and red lands), 17 types of soil occupation (forest, orchards, vineyard...) , 8 main crops, a mean slope and a drainage rate (sum of the length of rivers within the spatial unit/area of the spatial unit). The file provided by the GIS tools includes the calculation of the percentage of area for each variable (see Table 1 below). A second file, provided also by the GIS tools, includes for each hydological area the list of its neighbors.

The DIVCLUS-T method has been applied to the first data file, and C-DIVCLUS-T has been applied to the same data file taking into account the contiguity of the data given in the neighbors file. The five-clusters partition has been retained in both cases.

[2] SPICOSA project web site: www.spicosa.eu

Zhydro	Type of soil				Soil occupation				Crope				Mean slope	Drainage rate
	S_1	S_2	...	S_{14}	O_1	O_2	...	O_{17}	C_1	C_2	...	C_8		
R000	12	22	...	7.8	9.8	12.6	...	9.4	12	8.7	...	32.1	4.44	11.28
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1. The first rows of the data file.

Figures 1 and 2 give the map of the Charente basin and the clusters obtained with the two clustering methods for the 140 hydrological units . In order to illustrate the interest of using a monothetic approach for clustering, we have also reported on figure 2 the binary questions of the dendrogram obtained with C-DIVCLUS-T (see figure 3).

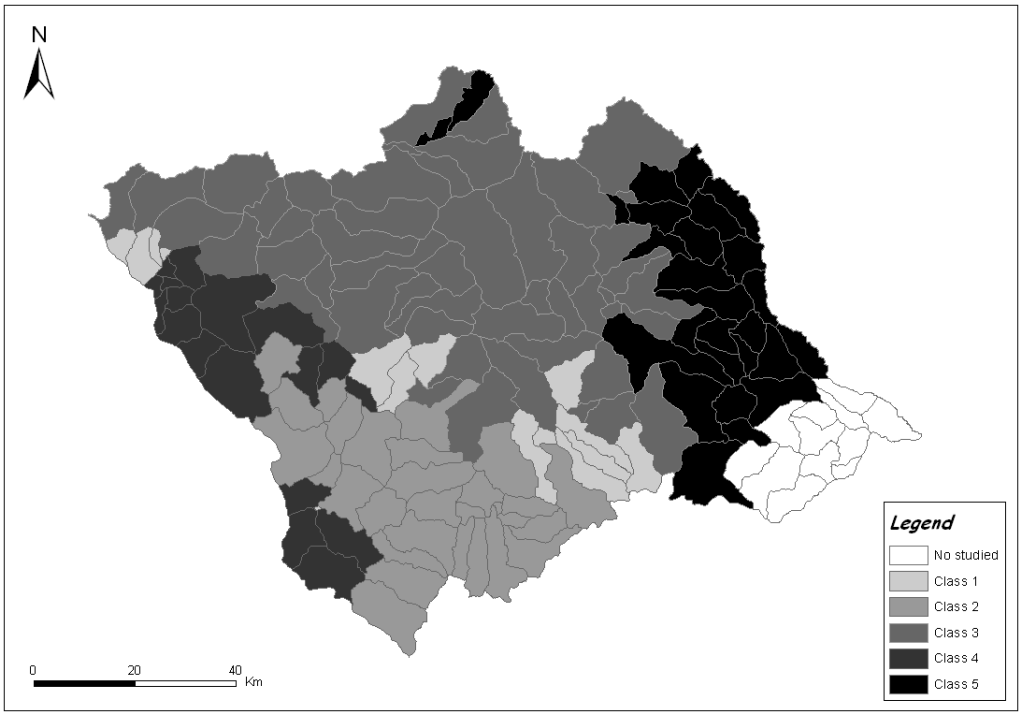

Fig. 1. The five-clusters partition obtained with DIVCLUS-T.

We can observe that the five clusters obtained with C-DIVCLUS-T are more interpretable than those obtained without spatial constraints. Indeed on the coastal zone three clusters are better delimited in figure 2 and a urban area (two hydrological units) is highlighted. Moreover, the hydrological unit of cluster 5 which was alone in the cluster 3 in figure 1, is merged to cluster 3 in figure 2.

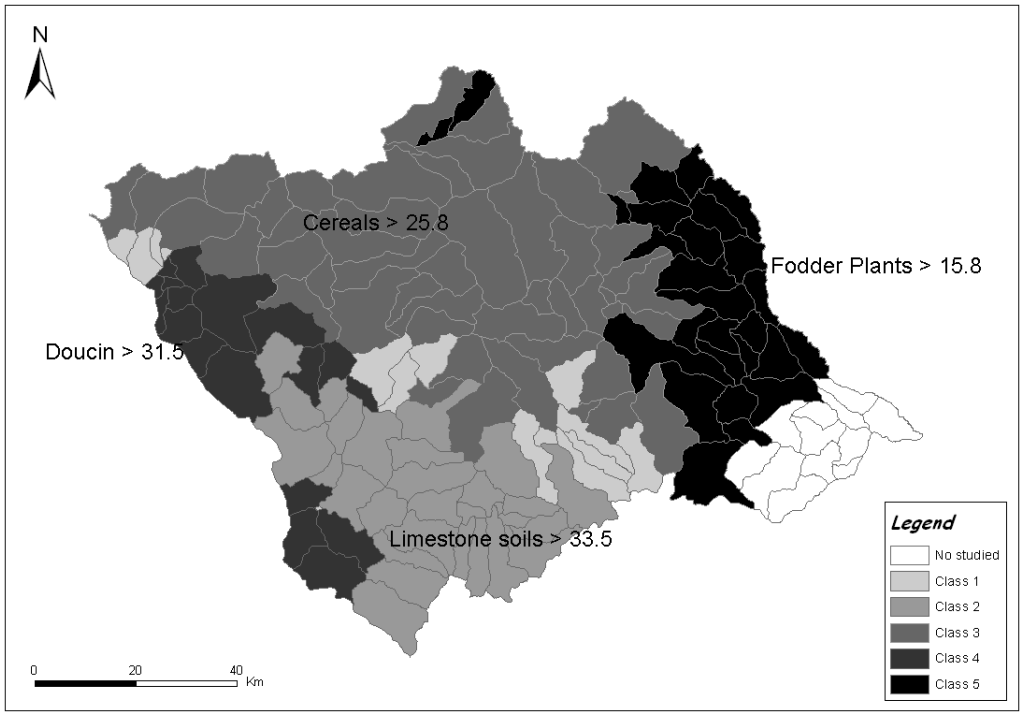

Fig. 2. The five-clusters partition obtained with C-DIVCLUS-T and $\alpha = 0.5$.

Figure 2 can then be read in the following way: a part of the coastal area can be linked to the presence of Doucins soils (moors). In the North of the river basin, an homogeneous area with cereal crops stands out and is not perturbed like in the previous classification. An other relevant area is delimited in the south of the basin with the variable "limestone soils" : we can find here vineyards and complex cultivation patterns. Finally, the cluster 1 can be linked to more artificialised areas.

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.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J. (1984): *Classification and regression Trees*. C.A:Wadsworth.

CHAVENT, M., BRIANT, O. and LECHEVALLIER, Y. (2007): DIVCLUS-T: a monothetic divisive hierarchical clustering method. *Computational Statistics and Data Analysis, 52 (2), 687-701.*

CHAVENT, M. (1998): A monothetic clustering method. *Pattern Recognition Letters, 19, 989-996.*

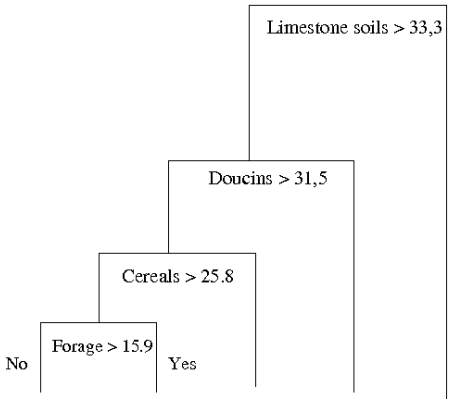

Fig. 3. Dengrogram obtained with C-DIVCLUS-T.

EDWARDS, A.W.F. and CAVALLI-SFORZA, L.L. (1965): A method for cluster analysis. *Biometrics, 21, 362-375.*

MURTAGH, F. (1985): A Survey of Algorithm for Contiguity-constrained clustering and Related Problems. *The computer journal, 28(1), 82-88.*

GORDON, A.D. (1996): A survey of constrained classification. *Computational statistics and data analysis, 21 (1), 17-29*

ZAHM, F. and VERNIER, F. (2007), *Contribution to the zoning of territorial agri-environmental measures within the context of the Rural Development Program for the 2007-2013 period: Application of the statistical model RA-SPACE to the river basin district of Adour-Garonne in order to implement a pesticide indicator.* Cemagref report to the French Ministry of Agriculture,122 p.

Comparing Histogram Data Using a Mahalanobis–Wasserstein Distance

Rosanna Verde and Antonio Irpino

Department of European and Mediterranean Studies
Second University of Naples
Via del Setificio 15, 81100 Caserta, Italy,
rosanna.verde@unina2.it , *antonio.irpino@unina2.it*

Abstract. In this paper, we present a new distance for comparing data described by histograms. The distance is a generalization of the classical Mahalanobis distance for data described by correlated variables. We define a way to extend the classical concept of inertia and codeviance from a set of points to a set of data described by histograms. The same results are also presented for data described by continuous density functions (empiric or estimated). An application to real data is performed to illustrate the effects of the new distance using dynamic clustering.

Keywords: histogram data, Wasserstein distance, Mahalanobis distance, inertia, dependence, dynamic clustering

1 Introduction

In many real experiences, data are collected and/or represented by frequency distributions. If $\mathbf{Y}$ is a numerical and continuous variable, many distinct values y_i can be observed. In these cases, the values are usually grouped in a smaller number H of consecutive and disjoint bins I_h (groups, classes, intervals, etc.). The frequency distribution of the variable $\mathbf{Y}$ is given considering the number of data values n_h falling in each I_h. A common way for representing this kind of data is the histogram representation of the variable $\mathbf{Y}$. The modeling of this kind of data has been proposed by Bock and Diday (2000) in the framework of "symbolic data analysis", where the concept of *histogram* variable is presented as a particular case of *modal* variable.
The interest in analyzing data expressed by empiric frequency distributions, as well as by histograms, is apparent in many fields of research. We refer to the treatment of experimental data that are collected in a range of values, whereas the measurement instrument gives only approximated (or rounded) values. An example, can be given by air pollution control sensors located in different zones of an urban area. The different distributions of measured data about the different levels of air pollutants in a day, allow us to compare and then to group into homogeneous clusters the different controlled zones.
In this paper, we propose to analyze data on the basis of the similarity of the frequency distributions.

After presenting histogram data and histogram variables in section 2, in section 3, we suggest using a distance based on the Wassertein metric (Gibbs and Su (2002)) for comparing two distributions, that is considered as an extension of the Euclidean distance between quantile functions. All the obtained results are generalizable to data described by density functions where the first two moments are finite.
Data can be described by several (histogram) variables. The first problem to solve in the analysis of multivariate data is the standardization of such data in order to balance their contribution to the results of the analysis. Other approaches dealing with the computation of the variability of a set of complex data can be found in Billard (2007), Bertrand and Goupil (2000) and Brito (2007). Billard (2007), Bertrand and Goupil (2000) apply the concept of variability to interval-valued data considering interval-valued datum realizations on $[a, b]$ that are uniformly distributed $U \sim (a, b)$. On this basis, Bertrand and Goupil (2000) developed some basic statistics to interval data and Billard (2007) extends them for the computation of dependence and interdependence measures for interval-valued data. In the case of histogram data, in section 4, we propose to extend the classical concept of inertia to a set of histogram data using the Wasserstein distance. The inertia computed for each variable will be used for the standardization of data. Further, in the analysis of a multivariate set of data it is important to take into consideration the interdependency structure of the variables. To this end, we propose a new distance for comparing multivariate histogram data extending the Wasserstein distance in order to obtain a generalization of the Mahalanobis distance (in section 6), based on (5) a generalization of the classical covariance measure for histogram variables.
In section 7, we present some results on a climatic dataset, using dynamic clustering. Section 8 ends the paper with some conclusions and perspectives.

2 Histogram data and histogram variables

Let $\mathbf{Y}$ be a continuous variable defined on a finite support $\mathbf{S} = [\underline{y}; \overline{y}]$, where $\underline{y}$ and $\overline{y}$ are the minimum and maximum values of the domain of $\mathbf{Y}$. The variable $\mathbf{Y}$ is partitioned into a set of contiguous intervals (bins) $\{I_1, \ldots, I_h, \ldots, I_H\}$, where $I_h = [\underline{y}_h; \overline{y}_h)$. Given N observations of the variable $\mathbf{Y}$, each semi-open interval, I_h is associated with a random variable equal to $\Psi(I_h) = \sum_{u=1}^{N} \Psi_{y_u}(I_h)$ where $\Psi_{y_u}(I_h) = 1$ if $y_u \in I_h$ and 0 otherwise. Thus, it is possible to associate with I_h an empirical distribution $\pi_h = \Psi(I_h)/N$.

A histogram of $\mathbf{Y}$ is the representation in which each pair (I_h, π_h) (for $h = 1, \ldots, H$) is represented by a vertical bar, with base interval I_h along the horizontal axis and the area proportional to π_h. Consider E as a set of n empirical distributions $\mathbf{Y(i)}$ $(i = 1, \ldots, n)$.

In the case of a histogram description it is possible to assume that $S(i) = [\underline{y}_i; \overline{y}_i]$, where $y_i \in \Re$. Considering a set of uniformly dense intervals $I_{hi} =$

$\left[\underline{y}_{hi}, \overline{y}_{hi}\right)$ such that:

$$i. \quad I_{li} \cap I_{mi} = \emptyset; \ l \neq m \ ; \qquad ii. \quad \bigcup_{s=1,\ldots,n_i} I_{si} = [\underline{y}_i; \overline{y}_i]$$

the support can also be written as $S(i) = \{I_{1i}, ..., I_{ui}, ..., I_{n_i i}\}$. In this paper, we denote with $\psi_i(y)$ the (empirical) density function associated with the description of i and with $\Psi_i(y)$ its distribution function. It is possible to define the description of $\mathbf{Y(i)}$ as:

$$Y(i) = \{(I_{ui}, \pi_{ui}) \mid \forall I_{ui} \in S(i); \ \pi_{ui} = \int_{I_{ui}} \psi_i(y)dy \geq 0\} \ \text{ where } \int_{S(i)} \psi_i(y)dy = 1.$$

According to Bock and Diday (2000), a histogram variable is a modal variable that associates a histogram with each observation.

3 Wasserstein distance between two histogram data

If F and G are the distribution functions of two random variables f and g respectively, with first moments μ_f and μ_g, and σ_f and σ_g their standard deviations, the Wasserstein L2 metric is defined as (Gibbs and Su, (2002))

$$d_M(F,G) := \left(\int_0^1 \left(F^{-1}(t) - G^{-1}(t)\right)^2 dt \right)^{1/2} \tag{1}$$

where F^{-1} and G^{-1} are the quantile functions of the two distributions. Irpino and Romano (2007) proved that the distance can be decomposed as:

$$d_W^2 = \underbrace{(\mu_f - \mu_g)^2}_{Location} + \underbrace{(\sigma_f - \sigma_g)^2}_{Size} + \underbrace{2\sigma_f\sigma_g(1 - \rho_{QQ}(F,G))}_{Shape} \tag{2}$$

where

$$\rho_{QQ}(F,G) = \frac{\int_0^1 \left(F^{-1}(t) - \mu_f\right)\left(G^{-1}(t) - \mu_g\right) dt}{\sigma_f\sigma_g} = \frac{\int_0^1 F^{-1}(t)G^{-1}(t)dt - \mu_f\mu_g}{\sigma_f\sigma_g} \tag{3}$$

is the correlation of the quantiles of the two distributions as represented in a classical QQ plot. It is worth noting that $0 < \rho_{QQ} \leq 1$ differently from the classical range of variation of the Bravais-Pearson's ρ correlation index. This decomposition allows us to take into consideration two aspects in the comparison of distribution functions. The first aspect is related to the location: two distributions can differ in position and this aspect is explained

by the distance between the mean values of the two distributions. The second aspect is related to the different variability of the compared distribution. This aspect is related to the different standard deviations of the distributions and to the different shapes of the density functions. While the former sub-aspect is taken into account by the distance between the standard deviations, the latter sub-aspect is taken into consideration by the value of ρ_{QQ}. Indeed, ρ_{QQ} is equal to one only if the two (standardized) distributions are of the same shape. We may consider a histogram as a particular case of a continuous density function built as a mixture of uniforms. Using this distance, we introduce an extended concept of inertia for a set of histogram data.

4 The inertia of a set data described by a histogram variable

A representative (prototype, barycenter) $\bar{f}_E$ associated with a set E of n described by a histogram variable X is an element of the space of description of E, i.e., it is a histogram. Extending the inertia concept of a set of points to a set of histograms, we may define such inertia as:

$$\begin{aligned} Inertia_E &= \sum_{i=1}^{n} d_W^2(y_i, \bar{f}_E) = \sum_{i=1}^{n} \int_0^1 \left(F_i^{-1}(t) - \bar{F}_E^{-1}(t)\right)^2 dt = \\ &= \sum_{i=1}^{n} \left[\left(\mu_{f_i} - \mu_{\bar{f}_E}\right)^2 + \left(\sigma_{f_i} - \sigma_{\bar{f}_E}\right)^2 + 2\sigma_{f_i}\sigma_{\bar{f}_E}(1 - \rho_{QQ}(F, \bar{F}_E))\right]. \end{aligned} \tag{4}$$

The $\bar{f}_E$ barycenter is obtained by minimizing the inertia criterion in (4), in the same way as the mean is the best least squares fit of a constant function to the given data points.

The $\bar{F}_E$ is a distribution where its t^{th} quantile is the mean of the t^{th} quantiles of the n distributions belonging to E. In this paper we introduce new measures of variability consistent with the classical concept of variability of a set of elements, without discarding any characteristics of the complex data (bounds, internal variability, shape, etc.).

It is interesting to note that the Wasserstein distance allows the Huygens theorem of decomposition of inertia for clustered data. Indeed, we showed (Irpino and Verde (2006) Irpino et al. (2006)) that it can be considered as an extension of the Euclidean distance between quantile functions. Reasoning by analogy, in Table 1 we introduce the sum of square deviation from the mean (DEV_F), and the extension of the variance (VAR_F) and of the standard deviation measure (STD_F) of a set of data described by a histogram variable.

4.1 The standardization of a histogram-valued dataset

A standardization of data is commonly applied when we need to analyze multivariate datasets where variables have different scales of measures or the

$$
\begin{array}{l|l}
Euclidean & Wasserstein \\
\hline
\sum_{i=1}^{n}\sum_{j=1}^{n} d_E^2(x_i,x_j) = & \sum_{i=1}^{n}\sum_{j=1}^{n} d_W^2\left(F_{1i}^{-1},F_{1j}^{-1}\right) = \\
= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2 = & = \sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^1\left(F_{1i}^{-1}(t)-F_{1j}^{-1}(t)\right)^2 dt = \\
= 2n\sum_{i=1}^{n}(x_i-\bar{x})^2 = & = 2n\sum_{i=1}^{n}\int_0^1\left(F_{1i}^{-1}(t)-\bar{F}_1^{-1}(t)\right)^2 dt = \\
= 2nDEV(X) & = 2nDEV_F(X_1) \\
DEV(X) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_E^2(x_i,x_j)}{2n} & DEV_F(X_1) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_W^2\left(F_{1i}^{-1},F_{1j}^{-1}\right)}{2n} \\
VAR(X) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_E^2(x_i,x_j)}{2n^2} & VAR_F(X_1) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_W^2\left(F_{1i}^{-1},F_{1j}^{-1}\right)}{2n^2} \\
STD(X) = \sqrt{VAR(X)} & STD_F(X_1) = \sqrt{VAR_F(X_1)}
\end{array}
\tag{5}
$$

Table 1. The inertia of a set of points and the inertia of a set of histograms using Wassertein distance.

same scale but different variability. Indeed, in several contexts of analysis it is important to homogenize data in order to treat them (clustering, classification, principal component analysis, etc.). Having computed the standard deviation, it is possible to introduce a method for the standardization of a set of histogram data. Let us consider the i^{th} deviation (D_i) from the prototype (mean, barycenter). It corresponds to a function of the quantile function of the i^{th} distribution minus the quantile distribution of the barycenter ($\bar{F}^{-1}$):

$$D_i(t) = F_i^{-1}(t) - \bar{F}^{-1}(t) \qquad 0 \le t \le 1. \tag{6}$$

It can be proven that $\sum_{i=1}^{n} D_i(t) = 0$ for $0 \le t \le 1$. By analogy, we introduce the standardized deviation–function as:

$$SD_i(t) = \frac{F_i^{-1}(t) - \bar{F}^{-1}(t)}{STD_F(X)} \qquad 0 \le t \le 1. \tag{7}$$

It can be proven that $n^{-1}\sum_{i=1}^{n}\int_0^1 (SD_i(t))^2\, dt = 1$ as in the classical case. Using standardization it is possible to extend some classic data analysis techniques to the data analysis of histogram data.

5 Interdependencies between histogram variables

In order to introduce a Mahalanobis version of the Wasserstein distance it is important to define an interdependence measure between two histogram variables. Given a set E described by p variables and the $\Sigma_{p\times p}$ covariance matrix, the Mahalanobis distance between two points described in $\Re^p$ is defined as:

$$d_m(\mathbf{x}_i,\mathbf{x}_{i'}) = \sqrt{(\mathbf{x}_i-\mathbf{x}_{i'})\Sigma^{-1}(\mathbf{x}_i-\mathbf{x}_{i'})^T} \tag{8}$$

where $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]$. In order to compare two histograms, we need to compute the covariance matrix of a histogram dataset. We propose to extend the covariance measure for histogram data as:

$$COVAR_F(X_j, X_{j'}) = \frac{CODEV_F(X_j, X_{j'})}{n} \tag{9}$$

where $CODEV_F$ is the codeviance of a set of data described by two histogram variables. In order to compute the $CODEV_F$, for each individual we know only the marginal distributions (the histograms) of the multivariate distribution that has generated it, and it is not possible to known the dependency structure between two histogram values observed for two variables. We assume that each individual is described by independent histogram variables. This is commonly used in the analysis of symbolic data (Billard (2007)). On this assumption, given two histogram variables X_j and $X_{j'}$, a set E of n histogram data with distribution F_{ij} and $F_{ij'}$ $(i = 1, \ldots, n)$, and considering the barycenter distributions $\bar{F}_j$ and $\bar{F}_{j'}$ of E for the two variables, we propose to extend the classical codeviance measure to histogram data as:

$$CODEV_F(X_j, X_{j'}) = \sum_{i=1}^{n} \int_0^1 \left(F_{ij}^{-1}(t) - \bar{F}_j^{-1}(t)\right)\left(F_{ij'}^{-1}(t) - \bar{F}_{j'}^{-1}(t)\right) dt. \tag{10}$$

Recalling equation (3), we may express it as:

$$\begin{aligned} &CODEV_F(X_j, X_{j'}) = \\ &= \sum_{i=1}^{n} \left[\alpha_i \cdot \sigma_{ij}\sigma_{ij'} - \beta_i \cdot \sigma_j\sigma_{ij'} - \gamma_i \cdot \sigma_{ij}\sigma_{j'}\right] + n \cdot \delta \cdot \sigma_j\sigma_{j'} + \\ &+ \left(\sum_{i=1}^{n} \mu_{ij}\mu_{ij'} - n\mu_j\mu_{j'}\right) \end{aligned} \tag{11}$$

where

- $\alpha_i = \rho_{QQ}\left(F_{ij}^{-1}, F_{ij'}^{-1}\right)$ is the QQ-correlation between the j^{th} histogram-value and the j'^{th} histogram-value of the i^{th} individual,
- $\beta_i = \rho_{QQ}\left(F_{ij'}^{-1}, \bar{F}_j^{-1}\right)$ is the QQ-correlation between the barycenter histogram of the j'^{th} variable and the j^{th} histogram of the i^{th} individual,
- $\gamma_i = \rho_{QQ}\left(F_{ij}^{-1}, \bar{F}_{j'}^{-1}\right)$ is the QQ-correlation between the j^{th} histogram of the i^{th} individual and the barycenter j'^{th} histogram,
- $\delta = \rho_{QQ}\left(\bar{F}_j^{-1}, \bar{F}_{j'}^{-1}\right)$ is the QQ-correlation between the barycenter j^{th} histogram and the barycenter j'^{th} histogram .

As a particular case, if all the distributions have the same shape (i.e., they are all normal or uniform distributed) then the ρ_{QQ}'s are equal to 1 and

$CODEV_F$ can be simplified as

$$CODEV_F\left(X_j, X_{j'}\right) = \left(\sum_i^n \sigma_{ij}\sigma_{ij'} - n\sigma_j\sigma_{j'}\right) + \left(\sum_i^n \mu_{ij}\mu_{ij'} - n\mu_j\mu_{j'}\right)$$

It is interesting to note that this approach is fully consistent with the classical decomposition of the codeviance. Indeed, we may consider a histogram as information related to a group of individuals. It can be proven that having a set of individuals grouped into k classes, the total codeviance can be decomposed in two additive components, the codeviance within and the codeviance between groups. With minimal algebra it is possible to prove that $|CODEV_F|$ cannot be greater than $\sqrt{DEV_F(X_1)DEV_F(X_2)}$. Then, we introduce the correlation measure for two histogram variables as:

$$CORR_F\left(X_j, X_{j'}\right) = \frac{CODEV_F\left(X_j, X_{j'}\right)}{\sqrt{DEV_F(X_1)DEV_F(X_2)}} \tag{12}$$

It is worth noting that if all histograms are identically distributed except for the first moments, $CORR_F$ depends only on the correlation of their first moments, and that if $CORR_F = 1$ (resp. -1) then all the histograms have their first moments aligned along a positive (resp. negative) sloped line and are identically distributed (except for the first moments).

6 A Mahalanobis–Wasserstein distance for histogram data

Given the vector $\mathbf{F_i} = [F_{i1}, \ldots, F_{ip}]$ and the inverse of $CODEV_F$ matrix $\Sigma_F^{-1} = [s_{hk}^{-1}]_{p\times p}$ we can introduce the Mahalanobis-Wasserstein distance as follows:

$$\begin{aligned} &d_{MW}(\mathbf{F_i}, \mathbf{F_{i'}}) = \\ &= \sqrt{\sum_{h=1}^{p}\sum_{k=1}^{p}\int_0^1 s_{hk}^{-1}\left(F_{ih}^{-1}(t) - F_{i'k}^{-1}(t)\right)\left(F_{ih}^{-1}(t) - F_{i'k}^{-1}(t)\right)dt}. \end{aligned} \tag{13}$$

The squared distance can be written:

$$\begin{aligned} &d_{MW}^2(\mathbf{F_i}, \mathbf{F_{i'}}) = \sum_{k=1}^{p} s_{kk}^{-1} d_W^2\left(F_{ik}, F_{i'k}\right) + \\ &+2\sum_{h=1}^{p-1}\sum_{k=h}^{p} s_{hk}^{-1}\left[\int_0^1 \left(F_{ih}^{-1}(t) - F_{i'h}^{-1}(t)\right)\left(F_{ik}^{-1}(t) - F_{i'k}^{-1}(t)\right)dt\right] = \\ &= \sum_{k=1}^{p} s_{kk}^{-1} d_W^2\left(F_{ik}, F_{i'k}\right) + \\ &+2\sum_{h=1}^{p-1}\sum_{k=h}^{p} s_{hk}^{-1}\left[\left(\alpha_{hk} - \beta_{hk} - \gamma_{hk} + \delta_{hk}\right) + \left(\mu_{ih} - \mu_{i'h}\right)\left(\mu_{ik} - \mu_{i'k}\right)\right] \end{aligned} \tag{14}$$

where:
$\alpha_{hk} = \rho_{QQ}\left(F_{ih}^{-1}, F_{ik}^{-1}\right) \cdot \sigma_{ih}\sigma_{ik}$; $\beta_{hk} = \rho_{QQ}\left(F_{ih}^{-1}, F_{i'k}^{-1}\right) \cdot \sigma_{ih}\sigma_{i'k}$;
$\gamma_{hk} = \rho_{QQ}\left(F_{ik}^{-1}, F_{i'h}^{-1}\right) \cdot \sigma_{ik}\sigma_{i'h}$; $\delta_{hk} = \rho_{QQ}\left(F_{i'h}^{-1}, F_{i'k}^{-1}\right) \cdot \sigma_{i'k}\sigma_{i'k}$.
If all distributions have the same shape (except for the first two moments) the distance can be simplified as:

$$\begin{aligned} d_{MW}^2(X_i, X_{i'}) &= \sum_{k=1}^{p} d_W^2\left(F_{ik}, F_{i'k}\right) \Sigma_{Fkk}^{-1} + \\ &+ 2\sum_{h=1}^{p-1} \left[\left(\sigma_{ih} - \sigma_{i'h}\right)\left(\sigma_{ik} - \sigma_{i'k}\right) + \left(\mu_{ih} - \mu_{i'h}\right)\left(\mu_{ik} - \mu_{i'k}\right)\right] \Sigma_{Fhk}^{-1}. \end{aligned} \tag{15}$$

7 An application

In this section, we show some results of clustering of data describing the mean monthly temperature, pressure, relative humidity, wind speed and total monthly precipitations of 60 meteorological stations of the People's Republic of China[1], recorded from 1840 to 1988. For the aims of this paper, we have considered the distributions of the variables for January (the coldest month) and July (the hottest month), so our initial data is a 60×10 matrix where the generic (i, j) cell contains the distribution of the values for the j^{th} variable of the i^{th} meteorological station. Figure 1 shows the geographic position of the 60 stations, while in Table 2 we have the basic statistics as proposed in section 4, and in Table 3 we show the interdependency measures as proposed in section 5. In particular, the upper triangle of the matrix contains the $COVAR_F$'s, while the bottom triangle contains the $CORR_F$'s for each pair of the histogram variables.

7.1 Dynamic clustering

The Dynamic Clustering Algorithm (DCA) (Diday (1971)) represents a general reference for partitioning algorithms. Let E be a set of n data described by p histogram variables X_j $(j = 1, \ldots, p)$. The general DCA looks for the partition $P \in P_k$ of E in k classes, among all the possible partitions P_k, and the vector $L \in L_k$ of k prototypes representing the classes in P, such that, the following Δ fitting criterion between L and P is minimized:

$$\Delta(P^*, L^*) = Min\{\Delta(P, L) \mid P \in P_k, L \in L_k\}. \tag{16}$$

Such a criterion is defined as the sum of dissimilarity or distance measures $\delta(x_i, G_h)$ of fitting between each object x_i belonging to a class $C_h \in P$ and the class representation $G_h \in L$:

$$\Delta(P, L) = \sum_{h=1}^{k} \sum_{x_i \in C_h} \delta(x_i, G_h). \tag{17}$$

[1] Dataset URL: `http://dss.ucar.edu/datasets/ds578.5/`

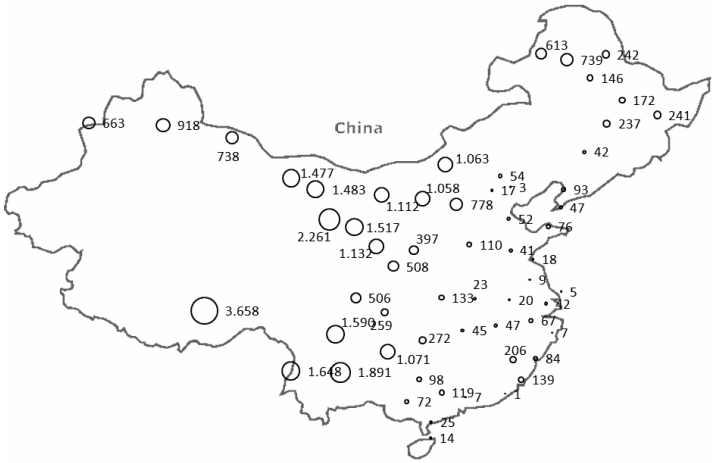

Fig. 1. The 60 meteorological stations of the China dataset; beside each point there is the elevation in meters.

♯ Variable	μ_j	σ_j	$VAR_F(X_j)$	$STD_F(X_j)$
X_1 Mean Relative Humidity (percent) Jan	67.9	7.0	127.9	11.3
X_2 Mean Relative Humidity (percent) July	73.9	4.5	114.2	10.7
X_3 Mean Station Pressure(mb) Jan	968.3	3.6	5864.7	76.5
X_4 Mean Station Pressure(mb) July	951.1	3.0	5084.4	71.3
X_5 Mean Temperature (Cel.) Jan	-1.2	1.7	114.8	10.7
X_6 Mean Temperature (Cel.) July	25.2	1.0	11.3	3.4
X_7 Mean Wind Speed (m/s) Jan	2.3	0.6	1.1	1.0
X_8 Mean Wind Speed (m/s) July	2.3	0.5	0.6	0.8
X_9 Total Precipitation (mm) Jan	18.2	14.3	519.6	22.7
X_{10} Total Precipitation (mm) July	144.6	80.8	499.9	70.7

Table 2. Basic statistics of the histogram variables: μ_j and σ_j are the mean and the standard deviation of the barycenter distribution of the j^{th} variable, while $VAR_F(X_j)$ and $STD_F(X_j)$ are the variability measures as presented in this paper.

Vars	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	128.0	49.1	510.2	486.1	34.0	20.0	0.7	1.6	109.9	97.9
X_2	**0.41**	114.2	392.6	376.4	53.5	11.2	4.2	1.2	72.3	475.6
X_3	**0.59**	**0.48**	5,864.7	5,455.2	162.9	198.3	32.0	24.3	672.4	1,570.2
X_4	**0.60**	**0.49**	**1.00**	5,084.4	158.9	183.1	29.9	22.5	634.8	1,504.6
X_5	**0.28**	**0.47**	**0.20**	**0.21**	114.8	22.5	0.0	-1.5	119.7	305.6
X_6	**0.52**	**0.31**	**0.77**	**0.76**	**0.62**	11.3	0.4	0.3	41.4	56.9
X_7	**0.06**	**0.38**	**0.40**	**0.40**	**0.00**	**0.13**	1.1	0.7	2.9	17.0
X_8	**0.17**	**0.14**	**0.39**	**0.39**	**-0.18**	**0.11**	**0.82**	0.6	1.6	-0.7
X_9	**0.43**	**0.30**	**0.39**	**0.39**	**0.49**	**0.54**	**0.12**	**0.09**	519.6	426.0
X_{10}	**0.12**	**0.63**	**0.29**	**0.30**	**0.40**	**0.24**	**0.23**	**-0.01**	**0.26**	4,999.3

Table 3. Covariances and correlations (in bold) of the ten histogram variables.

A prototype G_h associated with a class C_h is an element of the space of the description of E, and it can be represented as a vector of histograms. The algorithm is initialized by generating k random clusters or, alternatively, k random prototypes. We here present the results of two dynamic clustering using $k = 5$. The former considers δ as the squared Wasserstein distance among standardized data, while the latter uses the proposed squared Mahalanobis-Wasserstein distance. We have performed 100 initializations and we have considered the two partitions allowing the best quality index as defined in Chavent et al. (2003):

$$Q(P_k) = 1 - \frac{\sum_{h=1}^{k} \sum_{x_i \in C_h} \delta(x_i, G_h)}{\sum_{i \in E} \delta(x_i, G_E)}$$

where G_E is the prototype of the set E. The $Q(P_k)$ can be considered as the generalization of the ratio between the inter-cluster inertia and the total inertia of the dataset. Comparing the two clustering results, we may observe

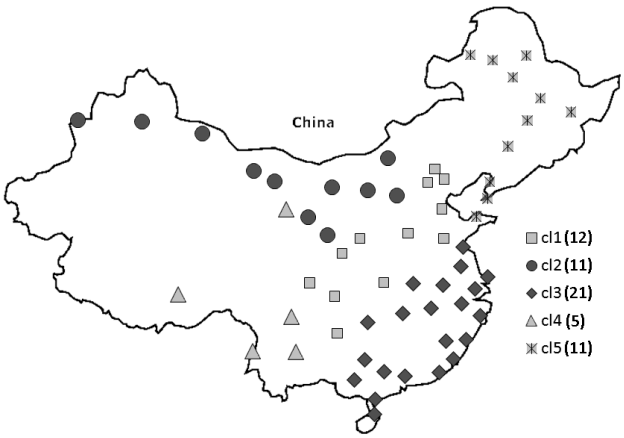

Fig. 2. Dynamic Clustering of the China dataset into 5 clusters (in brackets there is the cardinality of the cluster) using the Wasserstein distance on standardized data $Q(P_5) = 0.6253$.

that the two clusterings agree only on 65% of the observations (see Table 4): while the DCA using Wasserstein distance on standardized data allows a 61.53% of intra cluster inertia, the DCA using Mahalanobis-Wasserstein distance allows a 91.64%. Using covariance matrix, Mahalanobis distance removes redundancy between the variables. In this case, the DCA allows the definition of five clusters that collect stations at different elevations, respectively, stations between 0 and 140 meters (cluster 3), between 140 and 400 meters (cluster 5), between 500 and 900 meters (cluster 1), between 1000 and 1800 (cluster 2), and between 2,000 and 3,500 meters (cluster 4). Observing

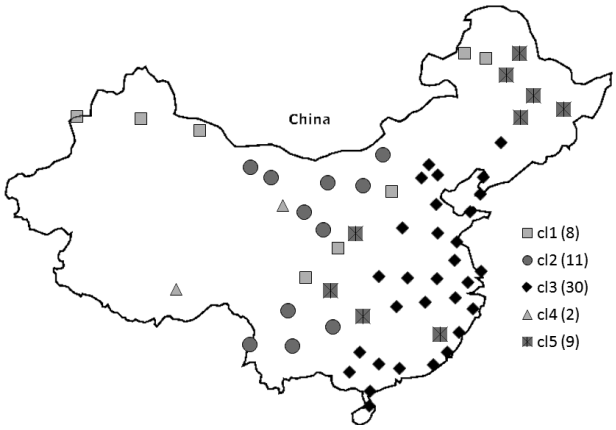

Fig. 3. Dynamic Clustering of the China dataset into 5 clusters (in brackets there is the cardinality of the cluster) using the Mahalanobis-Wasserstein distance, $Q(P_5) = 0.9164$.

		Clusters using Mahal.-Wass. distance					
		Cl 1	Cl 2	Cl 3	Cl 4	Cl 5	**Total**
	Cl 1	***2***	1	7		2	**12**
Clustering using	Cl 2	4	***7***				**11**
Wasserstein distance	Cl 3			***19***		2	**21**
between standardized	Cl 4		3		***2***		**5**
data	Cl 5	2		4		***5***	**11**
	Total	**8**	**11**	**30**	**2**	**9**	**60**

Table 4. Cross-classification table of the clusters obtained from the two dynamic clusterings.

a physical map of China, the obtained clusters seem more representative of the different typologies of meteorological stations considering their location and elevation. It is interesting to note that, also in this case, the use of a Mahalanobis metric for clustering data gives the same advantages of a clustering after a factor analysis (for example, a Principal Components Analysis), because it removes redundant information (in terms of linear relationships) among the descriptors.

8 Conclusions and future research

In this paper we have presented a new distance for comparing histogram data. The proposed method can be used in the interval data analysis whereas the intervals are considered as uniform densities according to Bertrand and Goupil (2000) and Billard (2007). Using the Wasserstein distance, we showed a way to standardize data, extending the classical concept of inertia for a set

of histogram data. The Mahalanobis–Wasserstein distance and the proposed interdependency measures between histogram variables can be considered as new useful tools for developing further analysis techniques for histogram data. The next step, considered very hard from a computational point of view (see Cuesta-Albertos and Matrán (1997)), is to find a way of considering the dependencies inside the histogram observations for multivariate histogram data in the computation of the Wasserstein distance.

References

BERTRAND, P. and GOUPIL, F. (2000): Descriptive statistics for symbolic data. In: H.-H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Berlin, 103–124.

BILLARD, L. (2007): Dependencies and Variation Components of Symbolic Interval–Valued Data. In: P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (Eds.): *Selected Contributions in Data Analysis and Classification*. Springer, Berlin, 3–12.

BOCK, H.-H. and DIDAY, E. (2000): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.

BRITO, P. (2007): On the Analysis of Symbolic Data. In: P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (Eds.): *Selected Contributions in Data Analysis and Classification*. Springer, Berlin, 13–22.

CHAVENT, M., DE CARVALHO, F.A.T., LECHEVALLIER, Y., and VERDE, R. (2003): Trois nouvelles méthodes de classification automatique des données symbolique de type intervalle. *Revue de Statistique Appliquée, LI, 4*, 5–29.

CUESTA-ALBERTOS, J.A., MATRÁN, C., TUERO-DIAZ, A. (1997): Optimal transportation plans and convergence in distribution. *Journ. of Multiv. An., 60*, 72–83.

DIDAY, E., and SIMON, J.C. (1976): Clustering analysis, In: K.S. Fu (Eds.),*Digital Pattern Recognition*, Springer Verlag, Heidelberg, 47–94.

DIDAY, E. (1971): Le méthode des nuées dynamiques, *Revue de Statistique Appliquée, 19, 2*, 19–34.

GIBBS, A.L. and SU, F.E. (2002): On choosing and bounding probability metrics. *Intl. Stat. Rev. 7 (3), 419–435*.

IRPINO, A., LECHEVALLIER, Y. and VERDE, R. (2006): Dynamic clustering of histograms using Wasserstein metric. In: A. Rizzi, M. Vichi, (Eds.) *COMPSTAT 2006*. Physica-Verlag, Berlin, 869–876.

IRPINO, A. and ROMANO, E. (2007): Optimal histogram representation of large data sets: Fisher vs piecewise linear approximations. *RNTI E-9, 99–110*.

IRPINO, A. and VERDE, R. (2006): A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. In: V. Batagelj, H.-H. Bock, A. Ferligoj, A. Ziberna (Eds.) *Data Science and Classification, IFCS 2006*. Springer, Berlin, 185–192.

VERDE, R. and IRPINO, A. (2007): Dynamic Clustering of Histogram Data: Using the Right Metric. In: P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (Eds.):

Selected Contributions in Data Analysis and Classification. Springer, Berlin, 123–134.

Part IV

Computation for Graphical Models and Bayes Nets

Iterative Conditional Fitting for Discrete Chain Graph Models

Mathias Drton

Department of Statistics, University of Chicago
5734 S. University Ave, Chicago, IL 60637, U.S.A.,
drton@uchicago.edu

Abstract. 'Iterative conditional fitting' is a recently proposed algorithm that can be used for maximization of the likelihood function in marginal independence models for categorical data. This paper describes a modification of this algorithm, which allows one to compute maximum likelihood estimates in a class of chain graph models for categorical data. The considered discrete chain graph models are defined using conditional independence relations arising in recursive multivariate regressions with correlated errors. This Markov interpretation of the chain graph is consistent with treating the graph as a path diagram and differs from other interpretations known as the LWF and AMP Markov properties.

Keywords: categorical data, chain graph, conditional independence, graphical model

1 Introduction

This paper considers models for categorical data that are analogous to Gaussian models induced by systems of linear regression equations with possibly correlated error terms. This analogy is of interest because systems of regression equations appear in many contexts, including structural equation modelling and graphical modelling, see e.g. Koster (1999), Wermuth and Cox (2004). For an example, consider the equations

$$X_1 = \beta_{10} + \varepsilon_1, \tag{1}$$
$$X_2 = \beta_{20} + \beta_{21}X_1 + \varepsilon_2, \tag{2}$$
$$X_3 = \beta_{30} + \varepsilon_3, \tag{3}$$
$$X_4 = \beta_{40} + \varepsilon_4, \tag{4}$$
$$X_5 = \beta_{50} + \beta_{51}X_1 + \varepsilon_5, \tag{5}$$

in which the coefficients β_{ij} may be arbitrary real numbers and the error terms ε_i have a centered joint multivariate normal distribution with positive definite covariance matrix $\Omega = (\omega_{ij})$. This matrix Ω is assumed to have a pattern of zero entries such that any pair of error terms is independent except for the pairs $(\varepsilon_i, \varepsilon_{i+1})$, $i = 2, 3, 4$, which may have possibly non-zero correlations. These assumptions lead to a particularly structured joint multivariate

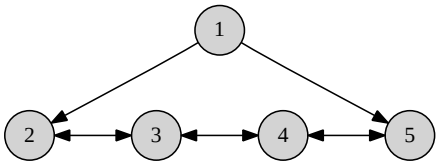

Fig. 1. Chain graph with directed and bi-directed edges. The use of bi-directed edges is in the tradition of path diagrams employed in structural equation modelling.

normal distribution $\mathcal{N}(\mu, \Sigma)$ for the random vector $X = (X_1, \ldots, X_5)$. While the mean vector μ is arbitrary, the covariance matrix is of the form

$$\Sigma = \begin{pmatrix} \omega_{11} & \beta_{21}\omega_{11} & 0 & 0 & \beta_{51}\omega_{11} \\ \beta_{21}\omega_{11} & \beta_{21}^2\omega_{11} + \omega_{22} & \omega_{23} & 0 & \beta_{21}\beta_{51}\omega_{11} \\ 0 & \omega_{23} & \omega_{33} & \omega_{34} & 0 \\ 0 & 0 & \omega_{34} & \omega_{44} & \omega_{45} \\ \beta_{51}\omega_{11} & \beta_{21}\beta_{51}\omega_{11} & 0 & \omega_{45} & \beta_{51}^2\omega_{11} + \omega_{55} \end{pmatrix}. \tag{6}$$

The normal model induced by (1)-(5) and the assumptions on the error terms ε_i comprises all distributions $\mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^5$ arbitrary and Σ of the form (6). This model can be represented using the graph shown in Figure 1. The directed edges in this graph represent covariate-response relations and, in the tradition of the path diagrams employed in the structural equation literature, bi-directed edges represent possible error correlations. The graph in Figure 1 is an instance of a chain graph. Chain graphs may have both oriented and unoriented edges, drawn here as directed and bi-directed edges, subject to acyclicity constraints. A brief introduction to chain graphs is given in §2. A more detailed treatment can be found, for example, in Andersson et al. (2001) and Lauritzen (1996). Note, however, that the so-called LWF and AMP chain graph models discussed in Andersson et al. (2001) and Lauritzen (1996) differ from the models considered in this paper.

It can be shown that the normal model associated with the graph in Figure 1 comprises all multivariate normal distributions in which

$$X_1 \perp\!\!\!\perp (X_3, X_4), \quad X_2 \perp\!\!\!\perp (X_4, X_5) \,|\, X_1 \quad \text{and} \quad (X_2, X_3) \perp\!\!\!\perp X_5 \,|\, X_1. \tag{7}$$

Here $\perp\!\!\!\perp$ denotes marginal or conditional independence depending on whether a conditioning set is specified. Having such a model characterization in terms of the non-parametric concept of conditional independence is useful because it is also meaningful outside the realm of normal distributions. Characterizations such as (7) are available more generally. In particular, if a system of linear regression equations with correlated errors corresponds to a chain graph, then the associated normal model can be characterized in terms of conditional independence. For the details of these results, we refer the reader to Koster (1999) and Richardson and Spirtes (2002).

This paper considers discrete chain graph models for categorical data obtained by imposing conditional independence relations such as (7). We review these models in §3, and in §4 we describe how the recently proposed 'iterative conditional fitting' algorithm can be modified for computation of maximum likelihood estimates. Different chain graphs can be Markov equivalent, i.e., lead to the same statistical model. In §5 we discuss how the choice of the graph can affect the computational efficiency of the associated fitting algorithm. Concluding remarks are given in §6.

2 Chain graphs

Let $G = (V, E)$ be a graph with finite vertex set V and edge set $E \subseteq (V \times V)$ such that there are no loops, i.e., $(v, v) \notin E$ for all $v \in V$. Two vertices v and w are *adjacent* if $(v, w) \in E$ or $(w, v) \in E$. If $(v, w) \in E$ and $(w, v) \in E$, then the edge $(v, w) \in E$ is without orientation and, reflective of pictures such as Figure 1, we refer to the edge as *bi-directed*. If $(v, w) \in E$ but $(w, v) \notin E$, then the edge (v, w) is *directed*. We will also write $v \to w$ and $v \leftrightarrow w$ to indicate directed and bi-directed edges, respectively. If $v \to w$ then v is a *parent* of w. The set of parents of v is denoted by $\mathrm{pa}(v)$, and for a set of vertices $\alpha \subseteq V$ we define the parents as

$$\mathrm{pa}(\alpha) = \{w \in V \mid \exists v \in \alpha : \ w \to v \text{ in } G\}\,.$$

A sequence of distinct vertices $\langle v_0, \ldots, v_k \rangle$ is a *path* if v_{i-1} and v_i are adjacent for all $1 \leq i \leq k$. A path $\langle v_0, \ldots, v_k \rangle$ is a *semi-directed cycle* if $(v_{i-1}, v_i) \in E$ for all $0 \leq i \leq k$ and at least one of the edges is directed as $v_{i-1} \to v_i$. Here, $v_{-1} \equiv v_k$. If the graph G has no semi-directed cycles, then G is a *chain graph*.

Define two vertices v_0 and v_k in a chain graph G to be equivalent if there exists a bi-directed path from v_0 to v_k, i.e., a path $\langle v_0, \ldots, v_k \rangle$ such that $v_i \leftrightarrow v_{i+1}$ for all $0 \leq i \leq k-1$. The equivalence classes under this equivalence relation are the *chain components* of G. For example, the chain graph in Figure 1 has the chain components $\{1\}$ and $\{2, 3, 4, 5\}$. The chain components $(\tau \,|\, \tau \in \mathcal{T})$ yield a partitioning of the vertex set

$$V = \bigcup_{\tau \in \mathcal{T}} \tau,$$

and the subgraph G_τ induced by each chain component τ is a connected graph with exclusively bi-directed edges. Moreover, the directed edges between two chain components τ_1 and τ_2 all have the same direction, i.e., if $(v_1, v_2) \in \tau_1 \times \tau_2$ and $(w_1, w_2) \in \tau_1 \times \tau_2$ are two pairs of adjacent vertices, then either $v_1 \to v_2$ and $w_1 \to w_2$, or $v_2 \to v_1$ and $w_2 \to w_1$.

3 Discrete chain graph models of multivariate regression type

Let $X = (X_v \,|\, v \in V)$ be a discrete random vector whose elements correspond to the vertices of a chain graph $G = (V, E)$. The graph G determines a list of conditional independence statements such as (7). The details of the process of determining the conditional independence statements are reviewed, for example, in Drton (2008). Let component X_v take values in $[d_v] = \{1, \ldots, d_v\}$, and define $\mathcal{I} = \times_{v \in V}[d_v]$. For $i = (i_v \,|\, v \in V) \in \mathcal{I}$, let

$$p(i) = P(X = i) = P(X_v = i_v \text{ for all } v \in V). \tag{8}$$

The joint distribution of X is determined by the array of probabilities $p = \big(p(i) \,|\, i \in \mathcal{I}\big)$, which is in the $(\prod_{v \in V} d_v) - 1$ dimensional probability simplex $\Delta \subset \mathbb{R}^{\mathcal{I}}$. Hence, the discrete chain graph model associated with G corresponds to a subset $\mathcal{P}(G) \subset \Delta$, which comprises exactly those arrays of probabilities p that lead to the desired conditional independence relations for the random vector X.

An array $p \in \mathcal{P}(G)$ obeys a factorization over the chain components of G. For a chain component $\tau \subseteq V$, let

$$p(i_\tau \,|\, i_{\pi(\tau)}) = P(X_\tau = i_\tau \,|\, X_{\pi(\tau)} = i_{\pi(\tau)}), \tag{9}$$

where $\pi(\tau)$ is the union of all chain components τ' in G that contain a vertex w that is a parent of a vertex in τ, i.e., all τ' such that $\tau' \cap \mathrm{pa}(\tau) \neq \emptyset$. If $\pi(\tau) = \emptyset$, then (9) refers to an unconditional probability. It then holds that each $p \in \mathcal{P}(G)$ factors as

$$p(i) = \prod_{\tau \in \mathcal{T}} p(i_\tau \,|\, i_{\pi(\tau)}), \qquad i \in \mathcal{I}. \tag{10}$$

The factorization in (10) is of the type usually encountered in directed graphical models (also known as Bayesian networks) but operates on the level of chain components rather than individual vertices. The factorization for the chain graph in Figure 1 is of the form

$$p(i) = p(i_1)p(i_2, i_3, i_4, i_5 \,|\, i_1), \qquad i \in \mathcal{I}. \tag{11}$$

The conditional independence relations that need to hold in an array of probabilities p in order for it to be in the model $\mathcal{P}(G)$ lead to constraints on the conditional probabilities $p(i_\tau \,|\, i_{\pi(\tau)})$. Drton (2008) describes a change of conditional probability coordinates that simplifies these constraints and yields in particular that the positive distributions in the model $\mathcal{P}(G)$ form a curved exponential family. This ensures regular large-sample asymptotics such as asymptotically normal maximum likelihood estimators.

Consider a chain component $\tau \in \mathcal{T}$ and let $i_{\pi(\tau)} \in \mathcal{I}_{\pi(\tau)}$. For a non-empty subset $\alpha \subseteq \tau$, define $\mathcal{J}_\alpha = \times_{v \in \alpha}[d_v - 1]$. The set $[d_v - 1] = \{1, \ldots, d_v - 1\}$

is the range for random variable X_v but with the highest-numbered element d_v removed. (Any other element could be chosen as baseline and be removed instead.) For each $j_\alpha \in \mathcal{J}_\alpha$, let

$$q(j_\alpha \,|\, i_{\pi(\tau)}) = P(X_\alpha = j_\alpha \,|\, X_{\pi(\tau)} = i_{\pi(\tau)}).$$

The probabilities $q(j_\alpha \,|\, i_{\pi(\tau)})$, $\emptyset \neq \alpha \subseteq \tau$, $j_\alpha \in \mathcal{J}_\alpha$, can be shown to be in one-to-one correspondence to the probabilities $p(i_\tau \,|\, i_{\pi(\tau)})$, $i_\tau \in \mathcal{I}_\tau$. This gives the above mentioned change of coordinates that simplifies the considered conditional independence relations to the form in Theorem 1.

A subset $\alpha \subseteq \tau$ is *disconnected* if there are two distinct vertices $v, w \in \alpha$ such that no path from v to w in G has all its vertices in α. Otherwise, α is a *connected* set. A disconnected set $\delta \subseteq \tau$ can be partitioned uniquely into inclusion-maximal disjoint connected sets $\gamma_1, \ldots, \gamma_r \subseteq \tau$,

$$\delta = \gamma_1 \cup \gamma_2 \cup \cdots \cup \gamma_r. \tag{12}$$

Theorem 1 (Drton, 2008). *Let G be a chain graph with chain components $(\tau \,|\, \tau \in \mathcal{T})$. An array p in the probability simplex Δ belongs to the discrete chain graph model $\mathcal{P}(G)$ if and only if the following three conditions hold:*

(i) The components of p factor as in (10).

(ii) For all $\tau \in \mathcal{T}$ and $i_{\pi(\tau)} \in \mathcal{I}_{\pi(\tau)}$, it holds that

$$q(j_\delta \,|\, i_{\pi(\tau)}) = q(j_{\gamma_1} \,|\, i_{\pi(\tau)}) q(j_{\gamma_2} \,|\, i_{\pi(\tau)}) \cdots q(j_{\gamma_r} \,|\, i_{\pi(\tau)})$$

for every disconnected set $\delta \subseteq \tau$ and $j_\delta \in \mathcal{J}_\delta$. Here $\gamma_1, \ldots, \gamma_r \subseteq \tau$ are the inclusion-maximal connected sets in (12).

(iii) For all $\tau \in \mathcal{T}$, connected subsets $\gamma \subseteq \tau$ and $j_\gamma \in \mathcal{J}_\gamma$, it holds that

$$q(j_\gamma \,|\, i_{\pi(\tau)}) = q(j_\gamma \,|\, k_{\pi(\tau)})$$

for every pair $i_{\pi(\tau)}, k_{\pi(\tau)} \in \mathcal{I}_{\pi(\tau)}$ such that $i_{\mathrm{pa}(\gamma)} = k_{\mathrm{pa}(\gamma)}$.

Example 1. For the graph from Figure 1, Theorem 1 only constrains the conditional probabilities $p(i_2, i_3, i_4, i_5 \,|\, i_1)$ for the second chain component $\{2, 3, 4, 5\}$. The constraints from condition (ii) are

$$q(j_2, j_4 \,|\, i_1) = q(j_2 \,|\, i_1) q(j_4 \,|\, i_1), \tag{13}$$
$$q(j_2, j_5 \,|\, i_1) = q(j_2 \,|\, i_1) q(j_5 \,|\, i_1), \tag{14}$$
$$q(j_3, j_5 \,|\, i_1) = q(j_3 \,|\, i_1) q(j_5 \,|\, i_1), \tag{15}$$
$$q(j_2, j_3, j_5 \,|\, i_1) = q(j_2, j_3 \,|\, i_1) q(j_5 \,|\, i_1), \quad \text{and} \tag{16}$$
$$q(j_2, j_4, j_5 \,|\, i_1) = q(j_2 \,|\, i_1) q(j_4, j_5 \,|\, i_1), \tag{17}$$

for all $i_1 \in [d_1]$ and $j_{2345} \in \mathcal{J}_{2345}$. Condition (iii) leads to the constraints

$$q(j_3 \,|\, i_1) = q(j_3 \,|\, k_1), \tag{18}$$
$$q(j_4 \,|\, i_1) = q(j_4 \,|\, k_1), \quad \text{and} \tag{19}$$
$$q(j_3, j_4 \,|\, i_1) = q(j_3, j_4 \,|\, k_1), \tag{20}$$

for all $i_1 < k_1 \in [d_1]$ and $j_{34} \in \mathcal{J}_{34}$.

4 Iterative conditional fitting

Suppose $X^{(1)}, \ldots, X^{(n)}$ are a sample of independent and identically distributed random vectors taking values in $\mathcal{I} = \times_{v \in V}[d_v]$. Suppose further that the probability array p for the joint distribution of the random vectors $X^{(k)}$ is in a chain graph model $\mathcal{P}(G)$. If we define the counts

$$n(i) = \sum_{k=1}^{n} 1_{\{X^{(k)}=i\}}, \quad i \in \mathcal{I},$$

then the *likelihood function* of $\mathcal{P}(G)$ is equal to

$$L(p) = \prod_{i \in \mathcal{I}} p(i)^{n(i)}.$$

For $\alpha \subseteq V$ and $i_\alpha \in \mathcal{I}_\alpha$, define the marginal counts

$$n(i_\alpha) = \sum_{j \in \mathcal{I}: j_\alpha = i_\alpha} n(j),$$

and the marginal probabilities

$$p(i_\alpha) = P(X_\alpha = i_\alpha) = \sum_{j \in \mathcal{I}: j_\alpha = i_\alpha} p(j).$$

Using the factorization in (10), the *log-likelihood function* $\ell(p) = \log L(p)$ can be written as the sum $\ell(p) = \sum_{\tau \in \mathcal{T}} \ell_\tau(p)$, where

$$\ell_\tau(p) = \sum_{i_{\pi(\tau)}} \sum_{i_\tau} n(i_\tau, i_{\pi(\tau)}) \log p(i_\tau \,|\, i_{\pi(\tau)}) \tag{21}$$

are the *component log-likelihood functions*. For different chain components, the conditional probabilities $p(i_\tau \,|\, i_{\pi(\tau)})$ are variation-independent. We can thus maximize ℓ over $\mathcal{P}(G)$ by separately maximizing the component log-likelihood functions over their respective domains.

Example 2. For the graph from Figure 1 we obtain that $\ell(p) = \ell_1(p) + \ell_{2345}(p)$ with component log-likelihood functions

$$\ell_1(p) = \sum_{i_1 \in [d_1]} n(i_1) \log p(i_1), \tag{22}$$

$$\ell_{2345}(p) = \sum_{i \in \mathcal{I}} n(i) \log p(i_2, i_3, i_4, i_5 \,|\, i_1). \tag{23}$$

The function ℓ_1 is maximized by $\hat{p}(i_1) = n(i_1)/n$ and only the maximization of ℓ_{2345} presents a challenge.

Since the component log-likelihood function ℓ_τ in (21) is a sum over $i_{\pi(\tau)}$, each term of this sum can be maximized separately if one has variation-independence of the probabilities appearing in the different terms.

Proposition 1. *Suppose τ is a chain component of the chain graph G such that $\mathrm{pa}(v) = \pi(\tau)$ for all $v \in \tau$. If $i_{\pi(\tau)}, j_{\pi(\tau)} \in \mathcal{I}_{\pi(\tau)}$ and $i_{\pi(\tau)} \neq j_{\pi(\tau)}$, then the two arrays of conditional probabilities $(p(i_\tau \,|\, i_{\pi(\tau)}) \,|\, i_\tau \in \mathcal{I}_\tau)$ and $(p(i_\tau \,|\, j_{\pi(\tau)}) \,|\, i_\tau \in \mathcal{I}_\tau)$ are variation-independent.*

Proof. Since $\mathrm{pa}(v) = \pi(\tau)$ for all $v \in \tau$, condition (iii) in Theorem 1 is void. Condition (ii) constrains each one of the arrays $(p(i_\tau \,|\, i_{\pi(\tau)}) \,|\, i_\tau \in \mathcal{I}_\tau)$ and $(p(i_\tau \,|\, j_{\pi(\tau)}) \,|\, i_\tau \in \mathcal{I}_\tau)$ separately. □

Clearly, Proposition 1 applies only in very special cases. It does not apply, for instance, to the chain component $\{2, 3, 4, 5\}$ of the graph from Figure 1 because $\mathrm{pa}(2) = \{1\}$ differs from $\mathrm{pa}(3) = \emptyset$.

Different approaches can be taken for maximization of a component log-likelihood function ℓ_τ in an arbitrarily structured chain graph. One approach is to express ℓ_τ in terms of parametrizations and then apply routines for unconstrained numerical optimization. Note, however, that care must be taken to avoid issues due to the fact that the involved parameters are generally variation-dependent. Here we will take an alternative approach by generalizing the 'iterative conditional fitting' (ICF) algorithm described in Drton and Richardson (2008a). This algorithm was proposed for binary marginal independence models and has a Gaussian version discussed in Chaudhuri et al. (2007). The marginal independence models studied in Drton and Richardson (2008a) are special cases of the chain graph models considered here. They are obtained from chain graphs with only one chain component, in which case conditions (i) and (iii) in Theorem 1 are void.

Generalized ICF for maximization of ℓ_τ from (21) starts with a choice of feasible estimates of the probabilities $p(i_\tau \,|\, i_{\pi(\tau)})$. These estimates are then improved iteratively. Each iteration cycles through all vertices in τ and when considering vertex $v \in \tau$ an update step with three parts is performed:

(a) Use the current feasible estimates to compute the conditional probabilities
$$p(i_{\tau\setminus\{v\}} \,|\, i_{\pi(\tau)}),$$
in which the variable X_v is marginalized out.

(b) Holding the probabilities computed in (a) fixed, solve a convex optimization problem to find updated estimates of the conditional probabilities
$$p(i_v \,|\, i_{(\tau\setminus\{v\})\cup\pi(\tau)}).$$

(c) Compute new feasible estimates according to
$$p(i_{\tau\setminus\{v\}} \,|\, i_{\pi(\tau)}) = p(i_v \,|\, i_{(\tau\setminus\{v\})\cup\pi(\tau)}) p(i_{\tau\setminus\{v\}} \,|\, i_{\pi(\tau)}).$$

The update step (a)-(c) mirrors the corresponding step in the original ICF algorithm, however, we now condition on the variables $X_{\pi(\tau)}$ throughout.

It remains to explain which convex optimization problem has to solved in part (b). The problem is to maximize ℓ_τ, treating the conditional probabilities from part (a) as fixed quantities and imposing the constraints from Theorem 1(ii) and (iii). To see the convexity of the problem, note that for fixed values of $p(i_{\tau\setminus\{v\}} \,|\, i_{\pi(\tau)})$ the function ℓ_τ is a concave function of the probabilities $p(i_v \,|\, i_{(\tau\setminus\{v\})\cup\pi(\tau)})$. Moreover, for fixed $p(i_{\tau\setminus\{v\}} \,|\, i_{\pi(\tau)})$, the constraints in Theorem 1(ii) and (iii) are linear in $p(i_v \,|\, i_{(\tau\setminus\{v\})\cup\pi(\tau)})$. Thus the feasible set for the problem is convex.

Example 3. We illustrate the outlined algorithm for the chain graph G in Figure 1 and the component log-likelihood function ℓ_{2345} from (23). For simplicity, we assume all five variables to be binary, i.e., $d_v = 2$ for $v = 1, \ldots, 5$. Up to symmetry, there are only two different update steps in ICF, namely, the one for $v = 2$ and the one for $v = 3$.

Update step for $v = 2$: In part (a) we compute the 16 conditional probabilities

$$p(i_3, i_4, i_5 \,|\, i_1), \quad i_1, i_3, i_4, i_5 = 1, 2.$$

These are then treated as fixed in part (b), in which we maximize

$$\sum_{i_1,\ldots,i_5=1}^{2} n(i_1, i_2, i_3, i_4, i_5) \log p(i_2 \,|\, i_1, i_3, i_4, i_5) \tag{24}$$

with respect to $p(i_2 \,|\, i_1, i_3, i_4, i_5)$. This maximization is done under constraints derived from (13), (14), (16) and (17). Since the probabilities $p(i_3, i_4, i_5 \,|\, i_1)$ are fixed, (15) is preserved automatically when updating the probabilities $p(i_2 \,|\, i_1, i_3, i_4, i_5)$. Moreover, since $\mathrm{pa}(2) = \pi(\tau)$ for $\tau = \{2, 3, 4, 5\}$, conditions (18)-(20) do not constrain the probabilities $p(i_2 \,|\, i_1, i_3, i_4, i_5)$. The important point is that (13), (14), (16) and (17) are linear constraints in $p(i_2 \,|\, i_1, i_3, i_4, i_5)$. The second factor on the right hand side of these equations is a function of the fixed quantities $p(i_3, i_4, i_5 \,|\, i_1)$ and thus also fixed. All other terms are linear combinations of $p(i_2 \,|\, i_1, i_3, i_4, i_5)$. For instance, (17) is linearized by writing

$$q(j_2, j_4, j_5 \,|\, i_1) = \sum_{i_3=1}^{2} p(j_2 \,|\, i_1, i_3, j_4, j_5) p(i_3, j_4, j_5 \,|\, i_1), \tag{25}$$

$$q(j_2 \,|\, i_1) = \sum_{i_3,i_4,i_5=1}^{2} p(j_2 \,|\, i_1, i_3, i_4, i_5) p(i_3, i_4, i_5 \,|\, i_1), \tag{26}$$

and noting that the probabilities $p(i_3, i_4, i_5 \,|\, i_1)$ are fixed quantities. In the resulting constrained maximization of (24) the two terms corresponding

to $i_1 = 1, 2$ can in fact be maximized separately because none of the constraints obtained from (13), (14), (16) and (17) involve two conditional probabilities $p(i_2 \mid i_1, i_3, i_4, i_5)$ and $p(k_2 \mid k_1, k_3, k_4, k_5)$ with $i_1 \neq k_1$.

Update step for $v = 3$: In part (a) we compute again 16 conditional probabilities, namely,

$$p(i_2, i_4, i_5 \mid i_1), \quad i_1, i_2, i_4, i_5 = 1, 2.$$

The objective function for part (b) is analogous to (24) but the probabilities $p(i_2 \mid i_1, i_3, i_4, i_5)$ are replaced by $p(i_3 \mid i_1, i_2, i_4, i_5)$. The relevant constraints are now (15), (16), (18) and (20). These can be linearized as in (25) and (26). The equations derived from (18) and (20) now involve $p(i_2 \mid i_1, i_3, i_4, i_5)$ and $p(k_2 \mid k_1, k_3, k_4, k_5)$ with $i_1 < k_1$. Hence, the optimization problem cannot be decomposed any further by splitting the strata $i_1 = 1$ and $i_1 = 2$.

In the update step for $v = 2$ in the above example, the terms in (24) indexed by different values of i_1 can be maximized separately. This observation is general in that analogous decompositions are possible if a vertex $v \in \tau$ satisfies $\mathrm{pa}(v) = \pi(\tau)$. This follows because $\mathrm{pa}(v) = \pi(\tau)$ implies that condition (iii) in Theorem 1 remains void; also recall Proposition 1.

5 Fitting marginal independence models via chain graphs

Graphical models based on graphs with directed edges generally lead to the problem of Markov equivalence, which arises if two different graphs induce the same statistical model. While such Markov equivalence poses challenges for the statistical interpretation of graphs, it can sometimes be exploited for more efficient computation of maximum likelihood estimates. Fitting algorithms, such as ICF, are often specified in terms of the underlying graph. The idea is then to find, among several Markov equivalent graphs, the one for which the associated fitting algorithm runs the fastest. Drton and Richardson (2008b) pursued this idea in the case of marginal independence models for the multivariate normal distribution and presented an algorithm that converts a graph with bi-directed edges into a Markov equivalent graph more suitable for optimization of the likelihood function. As we illustrate next, these constructions are also useful in the discrete case.

Table 2 in Drton and Richardson (2008a) presents data on seven binary variables from the US General Social Survey (sample size $n = 13\,486$). The data and software for their analysis are available for download at a supporting website for that article. A backward selection among bi-directed graphs for these data yields the graph G_a in Figure 2(a), which is identical to the one in Figure 4(a) in Drton and Richardson (2008a). Since all edges are bi-directed,

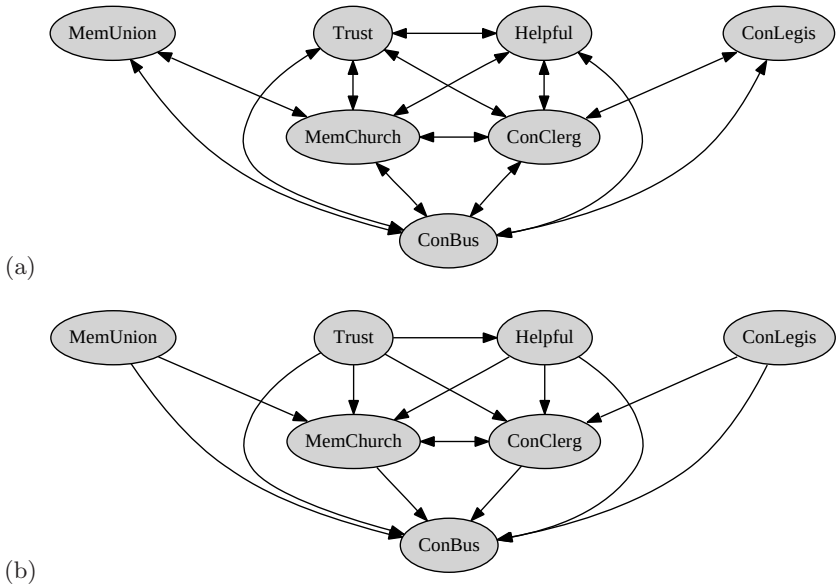

Fig. 2. (a) Bi-directed graph G_a for data from the US General Social Survey. (b) Chain graph G_b that is Markov equivalent to the bi-directed graph G_a.

the model $\mathcal{P}(G_a)$ can be characterized by marginal independences. The ICF algorithm for G_a iteratively estimates conditional probabilities of each of the seven variables given the remaining six variables. Running the algorithm for the given data, the deviance of $\mathcal{P}(G_a)$ was found to be 32.67 over 26 degrees of freedom, when compared with the saturated model of no independence. The asymptotic chi-square p-value is 0.172.

Using the results in Drton and Richardson (2008b), it is easily seen that the bi-directed graph G_a in Figure 2(a) is Markov equivalent to the chain graph G_b in Figure 2(b), i.e., $\mathcal{P}(G_a) = \mathcal{P}(G_b)$. When passing from G_a to G_b all but one of the bi-directed edges were substituted by directed edges. The remaining bi-directed edge in G_b between *MemChurch* and *ConClerg* cannot be replaced by a directed edge without destroying Markov equivalence to G_a. The graph G_b has six chain components, namely,

$$\begin{aligned} &\tau_1 = \{\mathit{Trust}\}, \quad &&\tau_2 = \{\mathit{Helpful}\}, &&\tau_3 = \{\mathit{MemUnion}\}, \\ &\tau_4 = \{\mathit{ConLegis}\}, \quad &&\tau_5 = \{\mathit{MemChurch}, \mathit{ConClerg}\}, &&\tau_6 = \{\mathit{ConBus}\}. \end{aligned}$$

The factorization (10) takes the form

$$\begin{aligned} p(i) = p(i_T)p(i_H \,|\, i_T)p(i_{MU})p(i_{CL})p(i_{MC}, i_{CC} \,|\, i_T, i_H, i_{MU}, i_{CL}) \times \\ p(i_{CB} \,|\, i_{MC}, i_{CC}, i_T, i_H, i_{MU}, i_{CL}), \end{aligned} \quad (27)$$

where the indices identify the variables via the capital letters in their names. The factorization in (27) reveals several closed-form maximum likelihood es-

timates, namely,

$$\hat{p}(i_T) = \frac{n(i_T)}{n},$$
$$\hat{p}(i_H \mid i_T) = \frac{n(i_T, i_H)}{n(i_T)},$$
$$\hat{p}(i_{MU}) = \frac{n(i_{MU})}{n},$$
$$\hat{p}(i_{CL}) = \frac{n(i_{CL})}{n},$$

and

$$\hat{p}(i_{CB} \mid i_{MC}, i_{CC}, i_T, i_H, i_{MU}, i_{CL}) = \frac{n(i_{CB}, i_{MC}, i_{CC}, i_T, i_H, i_{MU}, i_{CL})}{n(i_{MC}, i_{CC}, i_T, i_H, i_{MU}, i_{CL})};$$

see Drton (2008) where this observation is made in generality. The problem of computing maximum likelihood estimates in $\mathcal{P}(G_b)$ can thus be reduced to the simpler problem of maximizing the component log-likelihood function ℓ_{τ_5}. Moreover, since τ_5 is a complete set (there is an edge between all of its elements), ℓ_{τ_5} and thus also the likelihood function of $\mathcal{P}(G_b)$ have a unique local and global maximum if all counts are positive as is the case for the considered data; see again Drton (2008) for a general version of this result.

The maximization of ℓ_{τ_5} can be effected using the generalization of ICF developed in §4. It requires alternating between two update steps that estimate the conditional probabilities of *MemChurch* and of *ConClerg* given the respective five remaining variables in the components τ_m with $m \leq 5$. The variable *ConBus* in τ_6 is marginalized out in this computation. The constraints in part (b) of the update step for *MemChurch* arise from condition (iii) in Theorem 1. They take the form

$$\sum_{i_{CC}=1}^{2} p(j_{MC} \mid i_{CC}, i_T, i_H, i_{MU}, i_{CL}) p(i_{CC} \mid i_T, i_H, i_{MU}, i_{CL}) =$$
$$\sum_{i_{CC}=1}^{2} p(j_{MC} \mid i_{CC}, i_T, i_H, i_{MU}, k_{CL}) p(i_{CC} \mid i_T, i_H, i_{MU}, k_{CL}), \quad (28)$$

for $j_{MU} = 1, i_{CL} = 1, k_{CL} = 2$ and arbitrary combinations of i_T, i_H and i_{MU}. The variables being binary, (28) corresponds to a total of eight constraints. The symmetry present in G_b implies that the update step for *ConClerg* is analogous to that for *MemChurch*.

Running the ICF algorithm associated with G_b as opposed to G_a produced the same results with substantial reduction in computation time. When implementing the two algorithms in the statistical programming environment 'R' and using the same routine to solve the constrained optimization problem arising in the parts (b) of the respective update steps, we found that using G_b reduced the running time by a factor of about 70.

6 Conclusion

We have described a simple modification of the 'iterative conditional fitting' (ICF) algorithm proposed for marginal independence models in Drton and Richardson (2008a). This modification allows one to fit models associated with chain graphs. In future work it would be interesting to compare or even combine the ICF approach with other approaches to computation of maximum likelihood estimates such as that of Lupparelli et al. (2008).

As illustrated in §5, fitting algorithms for graphical models may take different forms for Markov equivalent graphs, and choosing the 'right' graph from a Markov equivalence class can be crucial for computationally efficient model fitting. Graphical constructions relevant for the models considered in this paper are given in Drton and Richardson (2008b) and Ali et al. (2005). However, these constructions generally do not return a chain graph but rather a graph from the larger class of ancestral graphs introduced in Richardson and Spirtes (2002). Generalizing ICF and other fitting algorithms to cover discrete ancestral graph models is an interesting topic for future research.

References

ALI, R.A., RICHARDSON, T.S., SPIRTES, P. and ZHANG, J. (2005): Towards characterizing Markov equivalence classes for directed acyclic graphs with latent variables. In: F. Bacchus and T. Jaakkola (Eds.): *Proc. 21st Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Corvallis, Oregon, 10-17.

ANDERSSON, S.A., MADIGAN, D. and PERLMAN, M.D. (2001): Alternative Markov properties for chain graphs. *Scand. J. Statist. 28, 33-85*.

CHAUDHURI, S., DRTON, M. and RICHARDSON, T.S. (2007): Estimation of a covariance matrix with zeros. *Biometrika 94, 199-216*.

DRTON, M. (2008): Discrete chain graph models. Manuscript.

DRTON, M. and RICHARDSON, T.S. (2008a): Binary models for marginal independence. *J. R. Stat. Soc. Ser. B. 70, 287-309*. (Software and data available at `http://www.blackwellpublishing.com/rss/Volumes/Bv70p2.htm`)

DRTON, M. and RICHARDSON, T.S. (2008b): Graphical methods for efficient likelihood inference in Gaussian covariance models. Preprint available at `arXiv:0708.1321`.

KOSTER, J.T.A. (1999): On the validity of the Markov interpretation of path diagrams of Gaussian structural equation systems with correlated errors. *Scand. J. Statist. 26, 413-431*.

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

LUPPARELLI, M., MARCHETTI, G.M. and BERGSMA, W.P. (2008): Parameterizations and fitting of bi-directed graph models to categorical data. Preprint available at `arXiv:0801.1440`.

RICHARDSON, T.S. and SPIRTES, P. (2002): Ancestral graph Markov models. *Ann. Statist. 30, 962-1030*.

WERMUTH, N. and COX, D.R. (2004): Joint response graphs and separation induced by triangular systems. *J. R. Stat. Soc. Ser. B 66, 687-717*.

Graphical Models for Sparse Data: Graphical Gaussian Models with Vertex and Edge Symmetries

Søren Højsgaard

Institute of Genetics and Biotechnology, Faculty of Agricultural Sciences, Aarhus University, Denmark, Research Center Foulum, DK-8830 Denmark,
sorenh@agrsci.dk

Abstract. The models we consider, generically denoted RCOX models, are a special class of graphical Gaussian models. In RCOX models specific elements of the concentration/partial correlation matrix can be restricted to being identical which reduces the number of parameters to be estimated. Thereby these models can be applied to problems where the number of variables is substantially larger than the number of samples. This paper outlines the fundamental concepts and ideas behind the models but focuses on model selection. Inference in RCOX models is facilitated by the `R` package `gRc`.

Keywords: concentration matrix, conditional independence, graphical model, graph, graph colouring, multivariate normal distribution, partial correlation

1 Introduction

Two types of graphical Gaussian models with edge and vertex symmetries are introduced by Højsgaard and Lauritzen (2007, 2008). The models generalize graphical Gaussian models, see for instance Whittaker (1990) and Lauritzen (1996). See Højsgaard and Lauritzen (2008) for additional references to related work.

In one type of models, denoted RCON models, selected elements of the concentration matrix (the inverse covariance matrix) are restricted to being identical. In the other class of models, denoted RCOR models, it is the partial *correlations* rather than the concentrations which are restricted to being equal. We use RCOX models as a generic term for both types. Tools for statistical inference a provided by the `gRc` package for `R`, R Development Core Team (2007).

A primary motivation for studying these models are potential applications in areas where high dimensional measurements are recorded on relatively few samples. A specific example is the analysis of gene expression data where the expression of a large number of genes (measured in thousands) is recorded on few samples (measured in tens or hundreds). For such data, parsimony in terms of the number of parameters can be essential and RCOX models provide a framework for obtaining this.

The focus of this paper is on outlining the ideas behind the models and discussing problems in relation to model selection for these models. For a comprehensive treatment of the models, their properties and estimation algorithms we refer to Højsgaard and Lauritzen (2007, 2008).

2 Preliminaries and notation

2.1 Graph colouring

Consider an undirected graph $\mathcal{G} = (V, E)$. Colouring the vertices of $\mathcal{G}$ with $R \leq |V|$ different colours induces a partitioning of V into disjoint sets $V_1, \ldots, V_R$ called *vertex colour classes* where all vertices in V_r have the same colour. Here $|V|$ denotes the number of elements in V. A similar colouring of the edges E with $S \leq |E|$ different colours yields a partitioning of E into disjoint sets $E_1, \ldots, E_S$ called *edge colour classes* where all edges in E_s have the same colour. We say that $\mathcal{V} = \{V_1, \ldots, V_R\}$ is a *vertex colouring* and $\mathcal{E} = \{E_1, \ldots, E_S\}$ is an *edge colouring*. A colour class with only one element is said to be *atomic*. A colour class which is not atomic is *composite*. A set $a \subset V$ is called *neutral* if its induced subgraph has only atomic colour classes.

When drawing vertices/edges we make the convention that black and white are used for atomic colour classes. Thus two edges displayed in black will be in different (atomic) colour classes.

Fig. 1 illustrates a graph colouring. To enable viewing in black and white we have also annotated vertices and edges with symbols to indicate the colourings. Vertices/edges with no annotation are in atomic colour classes. The edge between vertices `1` and `2` is written `1:2` etc. The coloured graph in (a) is given by $(\mathcal{V}, \mathcal{E})$ where

$$\mathcal{V} = [1, 4][2, 3], \quad \mathcal{E} = (1{:}2, 1{:}3)(2{:}4, 3{:}4)$$

whereas the graph in (b) is given by $\mathcal{V} = [1, 4][2][3]$ and $\mathcal{E} = (1{:}2, 1{:}3)(2{:}4)(3{:}4)$.

2.2 Graphical Gaussian models

Graphical Gaussian models are concerned with the distribution of a multivariate random vector $Y = (Y_\alpha)_{\alpha \in V}$ following a $N_d(\mu, \Sigma)$ distribution where $d = |V|$. For simplicity we assume throughout that $\mu = 0$. In the following we use Greek letters to refer to single variables and Latin letters to refer to sets of variables. We let $K = \Sigma^{-1}$ denote the inverse covariance, also known as the *concentration* with elements $(k_{\alpha\beta})_{\alpha,\beta \in V}$. The partial correlation between Y_α and Y_β given all other variables is then

$$\rho_{\alpha\beta|V\setminus\{\alpha,\beta\}} = -k_{\alpha\beta}/\sqrt{k_{\alpha\alpha}k_{\beta\beta}}. \tag{1}$$

Thus $k_{\alpha\beta} = 0$ if and only if Y_α and Y_β are conditionally independent given all other variables.

Fig. 1. Coloured graphs. (a): The edges `1:2` and `1:3` are in the same (light blue) edge colour class as also indicated by the "+"-sign. Likewise, `2:4` and `3:4` are in the same (green) edge colour class, also indicated by "++". The vertices `1` and `4` are in the red vertex colour class (also indicated by "*") while vertices `2` and `3` are in the blue vertex colour class (indicated by "**"). (b): Illustration of atomic colour classes. The vertices `2` and `3` are drawn in black and are atomic, so `2` and `3` are in different vertex colour classes. Likewise for edges `2:4` and `3:4`.

A *graphical Gaussian model* (hereafter abbreviated GGM) is represented by an undirected graph $\mathcal{G} = (V, E)$ where V is a set of vertices representing the variables and E is a set of undirected edges. The graph represents the model with K being a positive definite matrix having $k_{\alpha\beta} = 0$ whenever there is no edge between α and β in $\mathcal{G}$.

When the number n of samples is larger than the number of variables in the largest clique of $\mathcal{G}$, the maximum likelihood estimate exists and is unique. Thus fitting the saturated model requires $n > d$. For example in analysis of gene expression data, it is often the case that $n << d$.

Hence within GGMs attention must be restricted to graphs whose maximal clique size is small. In addition, parsimony can be obtained by considering RCON and RCOR models.

2.3 RCON models – restricted concentration models

An RCON model with vertex colour classes $\mathcal{V}$ and edge colour classes $\mathcal{E}$ is obtained by restricting the elements of $K = \Sigma^{-1}$ further as follows: 1) All partial variances (i.e. all diagonal elements of K) corresponding to vertices in the same vertex colour class are identical. 2) All off–diagonal entries of K corresponding to edges in the same edge colour class are identical. Thus, the diagonal of K can be specified by an R dimensional vector η while the off–diagonal elements are given by an S dimensional vector δ so we can write $K = K(\eta, \delta)$. Figure 1 (a) thereby represents the concentration matrix

$$K = \begin{bmatrix} \eta_1 & \delta_1 & \delta_1 & 0 \\ \delta_1 & \eta_2 & 0 & \delta_2 \\ \delta_1 & 0 & \eta_2 & \delta_2 \\ 0 & \delta_2 & \delta_2 & \eta_1 \end{bmatrix}.$$

To illustrate possible implications of such restrictions, let $a = \{1, 4\}$ and $b = \{2, 3\}$. The regression parameters when regressing b on a are given as

$-(K^{bb})^{-1}K^{ba}$. Thus the slope parameters for y_2 and y_3 become identical,

$$E(y_i|y_1, y_4) = a_i + (c_1/c_3)y_1 + (c_2/c_3)y_2 \text{ for } i = 2, 3,$$

where a_i is the intercept due to the mean. Thus the regression lines are parallel (and if the mean is zero then they coincide).

Another property of this model is that some partial correlations are restricted to being equal. For example it follows directly from (1) that

$$\rho_{12|34} = \rho_{13|24} = -\delta_1/\sqrt{\eta_1\eta_2} \text{ and } \rho_{32|14} = \rho_{42|13} = -\delta_2/\sqrt{\eta_1\eta_2}. \qquad (2)$$

2.4 RCOR models – restricted partial correlation models

An RCOR model with vertex classes $\mathcal{V}$ and edge classes $\mathcal{E}$ is obtained by restricting the elements of $K = \Sigma^{-1}$ as follows: 1) All partial variances corresponding to vertices in the same vertex colour class are identical. 2) All partial correlations corresponding to edges in the same edge colour class are identical.

As an RCOR model, Figure 1 (b) represents a concentration matrix K written as

$$K(\eta, \delta) = A(\eta)C(\delta)A(\eta),$$

where

$$A = \begin{bmatrix} \eta_1 & 0 & 0 & 0 \\ 0 & \eta_2 & 0 & 0 \\ 0 & 0 & \eta_3 & 0 \\ 0 & 0 & 0 & \eta_1 \end{bmatrix} \text{ and } C = \begin{bmatrix} 1 & \delta_1 & \delta_1 & 0 \\ \delta_1 & 1 & 0 & \delta_2 \\ \delta_1 & 0 & 1 & \delta_3 \\ 0 & \delta_2 & \delta_3 & 1 \end{bmatrix}.$$

Hence from (1), A contains the inverse partial standard errors on the diagonal while C contains minus the partial correlations on the off–diagonal. The vertex colour classes of an RCOR model is then restricting elements of A whereas the edge colour classes are restricting elements of C.

2.5 Example: Fret's heads

Mardia et al. (1979) report a study of heredity of head dimensions originating from Frets (1921). Graphical models for these were also considered, for instance by Whittaker (1990). Length and breadth of the heads of 25 pairs of first and second sons are measured. Previous analyses of these data support a model with conditional independence relations as in Figure 2, left. There is an obvious symmetry between the two sons so it makes sense to investigate a model where the joint distribution is unaltered if the two sons are interchanged. This model is obtained by adding restrictions as in Figure 2, right. In this model, the vertices `b1` and `b2` are in the same composite vertex colour class (and hence the corresponding diagonal elements of the concentration matrix are identical). The vertices `l1` and `l2` are also in the same composite

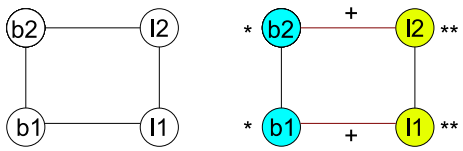

Fig. 2. Left: GGM describing the conditional independence structure for the Fret's heads data. Right: A corresponding RCON/RCOR model. Vertices/edges whose parameters are restricted to being identical are annotated with identical symbols.

vertex colour class. Likewise, the edges `b1:l1` and `b2:l2` are in the same composite edge colour class (and hence the corresponding off–diagonal elements of the concentration matrix are identical). The remaining edges, i.e. `b1:b2` and `l1:l2` are in two different atomic edge colour classes and are hence unconstrained.

It turns out that a model with this symmetry is both an RCON and RCOR model, see Højsgaard and Lauritzen (2008) for further details.

The saturated model gives $\log L = -252.7$ using 10 parameters. The model in Figure 2, left, gives $\log L = -254.2$ with 8 parameters. Finally, the model in Figure 2, right, gives $\log L = -255.2$ using 5 parameters, showing an excellent fit to data.

2.6 Maximum likelihood estimation

A detailed description of estimation algorithms are given by Højsgaard and Lauritzen (2007, 2008) and the algorithms are implemented in the `gRc` package. Hence we only outline methods here. There are two iterative methods in the package: 1) Newton scoring and 2) Iterative partial maximization.

When good starting values exist, Newton scoring generally requires fewest iterations, but with the drawback that Fisher information matrix needs to be calculated (and inverted) in each iteration. Each entry in the information matrix is based on calculation of traces of products of matrices, and this is somewhat expensive in terms of computing time.

Iterative partial maximization is based on iteratively maximizing the likelihood over one parameter at the time while keeping all other parameters fixed at their current values. This method generally requires more iterations than Newton scoring, but each iteration is cheaper in terms of computing time.

Empirical evidence suggests that it is advantageous (in terms of computing time) to combine the algorithms: First use iterative partial maximization for a small number of iterations to get near the maximum of the likelihood and and then use Newton scoring to obtain fast convergence.

3 Model selection issues

3.1 The need for selection strategies

The number of different models which can be formed by colouring edges and vertices in a graph is enormous. To illustrate the complexity, consider graphs with three vertices (for which there are 8 different graphs). A tedious enumeration shows that there are in total (over all 8 graphs) 15 possible edge colour classes. There are 5 possible vertex colour classes which gives $5 \times 15 = 75$ different models. This number grows very rapidly with the number of vertices.

In the following discussion we focus exclusively on edge colour classes. There are two ways of obtaining model reductions:

1. Join edge colour classes and/or
2. Delete edge colour classes.

Likewise, there are two ways of obtaining model expansions:

1. Partition edge colour classes and/or
2. Add new edge colour classes.

A brute–force approach for model reduction (facilitated by functions implemented in the `gRc` package) is as follows. Suppose the model under consideration has p edge colour classes.

With respect to joining edge colour classes, an approach is as follows: 1) Make all $p(p-1)/2$ possible pairwise comparisons of the edge colour classes, 2) Form a new model by joining those two edge colour classes whose parameters are not significantly different. If all colour classes are significantly different, then stop; else go to 1). In complexity, step 1) is quadratic in p and the number of times this step will be carried out is in the order of p. Hence this scheme becomes computationally too expensive unless p is small.

Moreover, such model reductions should be combined with a scheme for eliminating non–significant edge colour classes from the model, similiar to what is often done in a backward selection scheme for graphical models.

Such a scheme explodes in computational complexity for models with many variables. Therefore, alternative approaches to model selection become important. A contribution in this area is given below.

3.2 The FB^2 and FB^3 schemes

Fig. 3 illustrates an iterative forward model selection scheme consisting of an outer loop and an inner loop.

Let M^k denote the current model at the kth iteration of the outer loop. In the inner loop we make use of a sequence of temporary models $\mathrm{T1}, \mathrm{T2}, \ldots, \mathrm{T5}$. The inner loop takes the following form:

0: *Initialisation:* Set T1 $\leftarrow M^k$.
1: *Forward step:* Choose a set of candidate edges to add to T1. This forward step gives T2. If there are no candidate edges to choose from, then stop.
For example, in Fig. 3, (T2), the four dashed edges are added.
2: *Backward step 1:* From T2 remove those candidate edges which do not appear significant. This backward step gives T3 which is nested in T2. If all candidate edges are removed, then stop.
In Fig. 3, (T3), the edge `d:f` has been removed.
3: *Backward step 2:* Join the remaining candidate edges in T3 into composite colour classes whenever possible. This backward step gives T4 which is nested in T3.
Thus, in Fig. 3, (T4), edges `a:f` and `c:d` are joined into a composite edge colour class, while the edge `a:c` remains an atomic edge colour class.
4: Assign $M^{k+1} \leftarrow$ T4; go to 0.

This scheme, consisting of one forward step followed by two backward steps, will be denoted FB^2. The inner loop consists of a series of comparisons of nested models. Morover, the sequence of models created by the outer loop will also be nested; i.e. M^k is a submodel of M^{k+1}.

A modification of the scheme above is to replace step 4 by the following steps:

4a: *Backward step 3:* Join colour classes from T4 whenever possible. This backward step gives T5.
For example, in Fig. 3, (T5), the colour classes `(a:b,b:c)` and `(a:f,c:d)` are joined into one colour class.
4b: Assign $M^{k+1} \leftarrow$ T5; go to 0.

This scheme, consisting of one forward step followed by three backward steps will be denoted FB^3. As for FB^2, the inner loop of FB^3 also consists of a series of comparisons of nested models. However, the sequence of models created by the outer loop will in general not be nested.

In each step of the FB^2 and FB^3 schemes there are several specific choices to be made in each step. We will return to these in connection with an example on analyzing gene expression data.

3.3 Genes from breast cancer patients

Miller et al. (2005) investigated gene expression signatures for "p53" mutation status in 250 breast cancer samples. Of these, 58 samples have a mutation in the p53 sequence and data from these are considered in the following. The data have been standardized to have zero mean and unit variance. The dataset has expression values on 1000 genes.

Specific choices have to be made in relation to each step op the FB^3 scheme. Suggestions for these are described below. The starting model M^1 is taken to be the independence model. The selection criterion used is AIC.

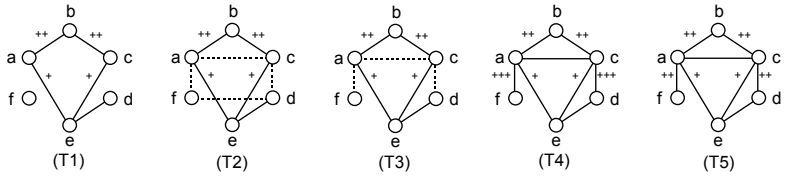

Fig. 3. The sequence of models T1-T4 constitutes the FB^2 scheme (forward + two times backward). The sequence T1-T5 is the FB^3 scheme (forward + three times backward).

Notice that an approximation to the difference in AIC between two nested models can be calculated by fitting the larger of the two models only: The Wald test statistics for testing the smaller model under the larger is asymptotically equivalent to the likelihood ratio statistic.

The `R` package `GeneNet` implements a method for estimating partial correlations in when there are more variables than samples, i.e. when $d > n$. The package also allows for testing which of these are significantly different for zero. Thereby we obtain a ranking of the possible edges in terms of the magnitude of p–values; i.e. a list of candidate edges E_{cand}. Below E_{cand} will be processed in chunks of N_{cand} edges starting with the most significant edges. For later use we let A denote the edges from E_{cand} which have already been considered for inclusion in the model.

1. Start by setting T1 $\leftarrow M^1$. Set $A = \emptyset$.
2. Let $C \subset E_{cand} \setminus A$ denote the N_{cand} most significant edges from $E_{cand} \setminus A$. Form T2 by adding the edges C to T1. Set $A = A \cup C$.
3. After fitting T2 we test whether each edge in C can be deleted from T2. The tests are based on Wald statistics, i.e. on the asymptotic covariance matrix under T2. It is here ignored that the test statistics in general are correlated. Insignificant edges are removed from T2 and this gives T3.
   (A computationally more expensive approach would be to apply a stepwise elimination scheme removing one candidate edge at the time.)
4. We fit T3 and make a hierarchical clustering of candidate edges based on their parameter estimates using `hclust()` of `R`. For a given number N_e the clustering induces an edge colouring with N_e edge colour classes. Because the clustering is hierarchical, increasing the values of N_e will then produce a sequence of nested models. In finding the opimal value of N_e we used Golden Section search. Based on this we take the model T4 to be the optimal model in terms of AIC.
   (A computationally more demanding approach would be to apply a stepwise scheme under which the two edge colour classes which are least different are joined to a new edge colour class. This process would stop when all edge colour classes are significantly different).

5. The model T4 consists of those n_1 edge colour classes inherited from T1 plus additionally n_2 new edge colour classes formed on the basis of candidate edges as described above. On these $n_1 + n_2$ colour classes we then apply the clustering scheme described above. This yields model T5. (An computationally more demanding alternative would be to compare each new edge colour class with those in T1 and join if possible. The complexity of this scheme is of the order $n_1 \times n_2$ so this can be feasible if these numbers are moderate in size).

It is beyond the scope of this paper to make a thorough investigation of whether models selected following the scheme above are biologically meaningful. Instead we shall focus on numerical comparisons.

3.4 Experiment 1

We partition the 58 cases into two datasets: A training dataset D_1 and a validation dataset D_2 each consisting of 29 cases. For simplicity we focus on 15 genes only, because this enables us to also make model search within GGMs and compare the results. Candidate edges are processed in chunks of $N_{cand} = 5$ edges. We focus on RCON models with all vertex colour classes being atomic. Hence the starting model M1 has 15 parameters.

On the basis of the dataset D_1 we consider the following models:

M^1: The independence model.
M^2: A GGM obtained by forward selection starting from the independence model.
M^3: An RCON model selected from the FB^3 scheme.
M^4: The GGM induced from the RCON model, that is a GGM with the same edges as in M^3 but where all edge colour classes are atomic.
M^5: The saturated model.

Table 1 shows the value of the log likelihood function under the respective models when evaluated on the training data D_1 and the validation data D_2. The interesting models to compare are M^2, M^3 and M^4. When evaluation the models on D_1, we see that the fit of the models are very similar. The same tendency can be seen on the validation data D_2. However, we note that M^3 (the RCON model) is the most parsimonious in terms of numbers of parameters.

The models M^2 and M^3 are shown in Fig. 4. The model M^2 has 17 edges and is close to being a tree while M^3 has 18 edges but these are described by only 10 parameters. The two graphs have 13 edges in common.

A tentative conclusion to be drawn from this is that when searching within RCON models we can obtain models which are widely different in structure from what we would find when searching among GGMs. The models fit about equally well to the data from which they are inferred. Moreover, the models

Table 1. Using 15 genes. Comparison of selected models based on calculating the log–likelihood on the training data (D_1) and on validation data (D_2).

Model	Dimension	$\log L$ (based on D_1)	$\log L$ (based on D_2)
M^1	15	−222.91	−247.01
M^2	32	−105.47	−206.54
M^3	25	−105.03	−210.29
M^4	33	−103.92	−214.54
M^5	120	−41.28	−460.28

also fit about equally well to an external dataset (as measured in terms of log likelihood). It remains a topic for further investigation whether models of the RCON/RCOR type are actually biologically meaningful.

3.5 Experiment 2

Next we consider a larger problem with 100 genes. Hence the number of genes is larger than the number of samples, so parsimoneous models are needed. We processed candidate edges in chunks of $N_{cand} = 8$ edges in each iteration. The initial model M^1 was the independence model, but with all vertices in the same vertex colour class. Hence the initial model contained one parameter only. The settings were otherwise as in Experiment 1. (Notice that with only one vertex colour class, all models are simultaneously both of the RCON and RCOR type.) Applying the FB^3 scheme for 100 iterations gave a resulting model M^3 with 558 edges but with only two edge colour classes (and hence three parameters). The average number of neighbours of each node is 11 with a standard deviation of 3, so the graph is far from having a tree structure. The graph has 123 cliques with 2 nodes, 214 cliques with 3 nodes and 22 cliques with 4 nodes. Hence the induced GGM M^4 (which has 658 parameters) can also be fitted to data.

Table 2 contains numerical summaries of the models. It follows that M^3 gave a very parsimonious description of data compared with the other models. When evaluating the models on the validation data we also see that M^3 appears to be the best of the models in terms of predictive ability.

Table 2. Using 100 genes. Comparison of selected models based on calculating the log–likelihood on the training data (D_1) and on validation data (D_2).

Model	Dimension	$\log L$ (based on D_1)	$\log L$ (based on D_2)
M^1	1	−1400.0	−1671.4
M^3	3	−307.6	−1328.7
M^4	658	−251.6	−2029.7

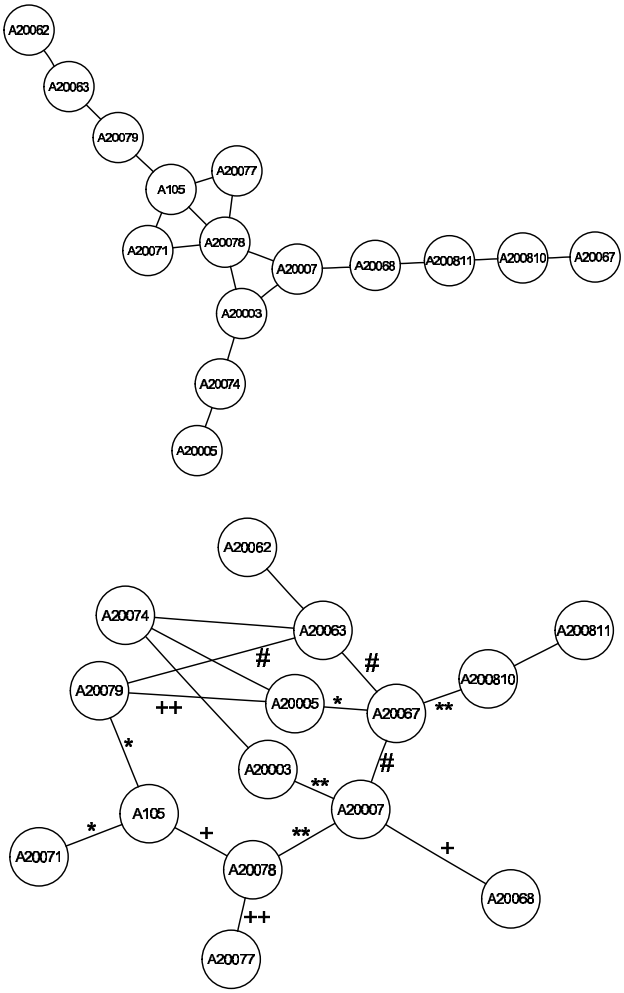

Fig. 4. Top: The GGM M^2 which has 17 edges and is close to being a tree. Bottom: The RCON model M^3 which has 18 edges but these are described by only 10 parameters. All vertices are in the same vertex colour class. Edges whose parameters are restricted to being identical are annotated with identical symbols.

4 Discussion

We have outlined the two new classes of graphical Gaussian models with edge and vertex symmetries as introduced by Højsgaard and Lauritzen (2007, 2008). These are denoted RCON and RCOR models and RCOX is used as a generic name for the two classes.

The main focus in this paper has been on model selection in these models. The space of RCOX models for a given set of variables is immensely

large. This makes model selection in RCOX models a complex issue. We have proposed a simple forward selection scheme which is based on first adding chunks of edges to a model (a large forward step) and then using clustering algorithms for collecting these into a smaller number of edge colour classes (several, typically smaller, backward steps). Empirical evidence suggests that this scheme leads to models with few edge colour classes and to graphs which are far from having a tree structure.

For RCOX models to be of practical use, further investigation of model selection strategies in high dimensional problems is needed.

References

FRETS, G.P. (1921): Heredity of head form in man. *Genetica 3, 193–400*

HØJSGAARD, S., LAURITZEN, S.L. (2007): Inference in graphical Gaussian models with edge and vertex symmetries with the gRc package for R *Journal of Statistical Software, 2007, 23 (6)*

HØJSGAARD, S., LAURITZEN, S.L. (2008): Graphical Gaussian models with edge and vertex symmetries. *To appear in Journal of the Royal Statistical Society, Series B*

MARDIA, K.V., KENT, J.T. and BIBBY, J.M. (1979): *Multivariate Analysis*, Academic Press.

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

MILLER, L.D., SMEDS, J., GEORGE, J., and VEGA, V.B., VERGARA, L., PAWITAN, Y., HALL, P., KLAAR, S., LIU, E.T., and BERGH, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences 102 (38), 13550-13555*

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/.

WHITTAKER, J. (1990): *Graphical Models in Applied Multivariate Statistics*, Wiley

Parameterization and Fitting of a Class of Discrete Graphical Models

Giovanni M. Marchetti[1] and Monia Lupparelli[2]

[1] Dipartimento di Statistica, Florence University
viale Morgagni, 59, 50134 Florence, Italy, *giovanni.marchetti@ds.unifi.it*
[2] Dipartimento di Economia Politica e Metodi Quantitativi, University of Pavia
via S. Felice, 7, 27100, Pavia, Italy, *mlupparelli@eco.unipv.it*

Abstract. Graphical Markov models are multivariate statistical models in which the joint distribution satisfies independence statements that are captured by a graph. We consider models for discrete variables, that are analogous to multivariate regressions in the linear case, when the variables can be arranged in sequences of joint response, intermediate and purely explanatory variables. In the case of one single group of variables, these models specify marginal independencies of pairs of variables. We show that the models admit a proper marginal log-linear parameterization that can accommodate all the marginal and conditional independence constraints involved, and can be fitted, using maximum likelihood under a multinomial assumption, by a general iterative gradient-based algorithm. We discuss a technique for determining fast approximate estimates, that can also be used for initializing the general algorithm and we present an illustration based on data from the U.S. General Social Survey.

Keywords: bi-directed graphs, covariance graphs, multivariate regression chain graphs, complete hierarchical parameterizations, reduced model estimates

1 Introduction

Joint-response chain graph models (Cox and Wermuth (1996), Wermuth and Cox (2004)) are a flexible family of models representing dependencies in systems in which the variables are grouped in blocks as responses, intermediate responses and purely explanatory factors. In this paper we discuss the properties of the subclass of multivariate regression chain graph models and their parameterization when all the variables are discrete. These models cannot be specified, in contrast with the classical chain graph models, within the standard log-linear parameterization. However, they can be expressed as marginal log-linear models (Bergsma and Rudas (2002)), and the paper discusses, mainly through examples, the definition of the parameters and of the constraints required. Maximum likelihood fitting of the models can be carried out with special iterative algorithms for constrained optimization, based on Aitchison and Silvey (1958). In the last section a method is proposed which approximates the maximum likelihood estimates, using a general method by

Cox and Wermuth (1990), and it is applied to the analysis of a data set from the U.S. General Social Survey. It is also shown that the approximate estimates, used as starting point, yield a faster convergence of the iterative algorithm.

2 Complete hierarchical marginal log-linear parameterizations

We review the definition of marginal log-linear models for discrete probability distributions, using complete hierarchical parameterizations; see Bergsma and Rudas (2002). Let $\boldsymbol{X} = (X_v)$, $v = 1, \ldots, d$ be a discrete random vector, with a strictly positive joint probability distribution $p(\boldsymbol{i}) > 0$, where $\boldsymbol{i}$ is a generic cell of the associated contingency table and let $p_M(\boldsymbol{i}_M)$ be any marginal probability distribution of a sub-vector $\boldsymbol{X}_M$, $M \subseteq V$. We call M a margin of the table. Each marginal probability distribution p_M has a log-linear expansion

$$\log p_M(\boldsymbol{i}_M) = \sum_{L \subseteq M} \lambda_L^M(\boldsymbol{i}_L)$$

where the parameters have the formal definition

$$\lambda_L^M(\boldsymbol{i}_L) = \sum_{A \subseteq L} (-1)^{|L \setminus A|} \log p_M(\boldsymbol{i}_A, \boldsymbol{i}^*_{M \setminus A})$$

with $\boldsymbol{i}^* = (1, \ldots, 1)$ denoting a baseline cell of the table; see Lauritzen (1996). By definition of the baseline, the function $\lambda_L^M(\boldsymbol{i}_L)$ is zero whenever at least one index in $\boldsymbol{i}_L$ is equal to 1. Also the parameter $\lambda_\phi^M = \log p(\boldsymbol{i}^*_M)$ is a function of the others, due to the sum to one constraint on the probabilities. The coding used here corresponds to the indicator coding, giving the parameters used for example by the R environment (R Development Core Team (2008)) with the function `contr.treatment`, but any other coding could be used as well. Let $\boldsymbol{\lambda}_L^M$ be the vector of the not structurally zero parameters, for each $M \subseteq V$ and for any non-empty subset $L \subseteq M$.

A *marginal log-linear parameterization* of the joint probability distribution p is obtained by combining log-linear parameters of several marginal distributions $p_{M_1}, \ldots, p_{M_s}$ of interest. The parameterization is characterized by a list of pairs

$$(M_1, \mathcal{L}_1), \ldots, (M_s, \mathcal{L}_s)$$

each composed of a margin M_j and a list $\mathcal{L}_j$ used to define the parameters. The essential requirements to obtain a proper parameterization are that:

(i) the sets $\mathcal{L}_j$ are a partition of *all possible non-empty subsets* L of the variables V;
(ii) if one log-linear parameter vector $\boldsymbol{\lambda}_L^M$ is defined within a margin M, then the log-linear parameters $\boldsymbol{\lambda}_{L'}^M$ for all supersets L' such that $L \subseteq L' \subseteq M$ must be defined in the same margin.

Table 1. A hierarchical partition of all the subsets of 1234.

Margin M_j	Parameters in $\mathcal{L}_j$	Minimal elements
14	$\{1, 4, 14\}$	$1, 4$
124	$\{2, 12, 24, 124\}$	2
134	$\{3, 13, 34, 134\}$	3
1234	$\{23, 123, 234, 1234\}$	23

If the parameterization satisfies the above requirements (i) and (ii), it is said *hierarchical and complete* and the *marginal log-linear parameters* $(\boldsymbol{\lambda}_L^{M_j}, L \subseteq \mathcal{L}_j), j = 1, \ldots, s$ define a *smooth* parameterization of the set of all strictly positive probability distributions p; see Bergsma and Rudas (2002). Notice that within each collection $\mathcal{L}_j$ we can find the minimal elements (with respect to inclusion), such that all the other elements are generated by taking all the supersets within $\mathcal{L}_j$.

Example 4. Table 1, defining the pairs $(M_j, \mathcal{L}_j)$ for $j = 1, \ldots, 4$, satisfies the requirements (i) and (ii) and thus defines a hierarchical and complete log-linear parameterization. For each set $\mathcal{L}_j$, the minimal elements are also reported. For instance, the minimal element of $\mathcal{L}_3$ is 3 because all the sets in $\mathcal{L}_3$ are the subsets of $M_3 = 134$ that include 3. The maximal element of each collection $\mathcal{L}_j$ is always the set M_j itself.

It can be shown that the sets $\mathcal{L}_j$ can be obtained from an ordered non-decreasing sequence of margins M_j, $j = 1, \ldots, s$, and with $M_s = V$, by defining

$$\mathcal{L}_1 = \mathcal{P}(M_1); \quad \mathcal{L}_2 = \mathcal{P}(M_2) \setminus \mathcal{L}_1; \quad \mathcal{L}_3 = \mathcal{P}(M_3) \setminus (\mathcal{L}_1 \cup \mathcal{L}_2), \ldots$$

and so on, where $\mathcal{P}(M)$ denotes the power set of M. Note that given two non-disjoint margins M_j and M_k, M_j precedes M_k if and only if the set $M_j \cap M_k$ is defined in $\mathcal{L}_j$.

Example 5. The *overall log-linear parameterization*, and the *multivariate logistic transformation* (Glonek and McCullagh (1995)) are two special cases of the marginal log-linear parameterization. The log-linear parameters are defined by $\boldsymbol{\lambda}_L^V$, for any $L \subseteq V$, and thus they are all considered with respect to the joint distribution. The multivariate logistic parameters, instead, are defined by $\boldsymbol{\lambda}_L^L$, for any subset $L \subseteq V$, i.e. they are the highest order log-linear parameters within each possible marginal distribution.

The model defined by the parameters $\boldsymbol{\lambda}_L^{M_j}$ for $L \in \mathcal{L}_j$, $j = 1, \ldots, s$ is a saturated marginal log-linear model. By a *marginal log-linear model* we mean a subset of the probability distributions defined by the constraints

$$\boldsymbol{\lambda}_L^{M_j} = \mathbf{0} \text{ for } L \in \mathcal{C}_j, j = 1, \ldots, s$$

where $\mathcal{C}_j$ is any collection of subsets of $\mathcal{L}_j$. Here we may require that the constraints are *hierarchical*, that is that if $L \in \mathcal{C}_j$, then $L' \in \mathcal{C}_j$, for any $L' \in \mathcal{L}_j$, such that $L \subseteq L'$. This class of log-linear marginal models is much wider and flexible than that of usual log-linear models. In the following we discuss its application to the fitting of discrete graphical models, with special emphasis to graphical models containing bi-directed edges.

3 Multivariate regression chain graph models

Discrete graphical Markov models are models for discrete distributions representable by graphs, associating nodes with the variables and using rules that translate properties of the graph into conditional independence statements between variables. There are several classes of graphical models, see Wermuth and Cox (2004) for a review. We consider here the two types called covariance graphs and multivariate regression chain graphs.

Covariance graphs represent marginal independencies and describe the associations between variables, all treated on the same footing, by a special type of edges: Cox and Wermuth (1996) use dashed edges, whereas Richardson (2003) uses bi-directed edges. Following the second convention we define a *covariance graph* as an undirected graph with node set V and edge set E, in which an edge uv is denoted by $u \longleftrightarrow v$. For this reason covariance graphs are also called *bi-directed graphs*. A discrete random vector $\boldsymbol{X} = (X_v)$ with $v \in V$ is associated with a covariance graph by defining a *Markov property*. The basic requirement for covariance graphs is that whenever two variables X_u and X_v associated with the nodes u and v are not adjacent, then they are marginally independent. We shall denote this statement by $u \perp\!\!\!\perp v$. The full set of requirements is summarized in the Appendix, equation (6), and if the probability distribution satisfies this set we say that the distribution is globally Markov with respect to the covariance graph.

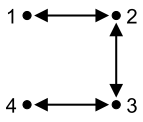

Fig. 1. A covariance graph.

Example 6. The graph in Figure 1 is a covariance graph with 4 nodes. A probability distribution is globally Markov with respect to this graph if

$$1 \perp\!\!\!\perp 4, \quad 2 \perp\!\!\!\perp 4|1, \quad 1 \perp\!\!\!\perp 3|4. \tag{1}$$

or, equivalently, if $12 \perp\!\!\!\perp 4$ and $1 \perp\!\!\!\perp 34$. This also implies that for any missing edge there is a marginal independence.

Multivariate regression chain graphs represent instead situations in which the variables can be arranged in an ordered series of groups and all the variables within a group are considered to be on an equal footing, while the relation between two variables in different groups is considered asymmetrically. The associated graph is a special type of chain graph in which the edges between components are arrows and the subgraphs within the chain components are covariance graphs, i.e. with bi-directed edges. More precisely, a *multivariate regression chain graph* is a graph $G = (V, E)$ with node set V and edge set E containing directed and/or bi-directed edges. Moreover the node set V can be partitioned into disjoint subsets τ called chain components, such that all edges in each subgraph G_τ are bi-directed and the edges between two different subsets $\tau_1 \neq \tau_2$ are directed, pointing in the same direction. The set of chain components is denoted by $\mathcal{T}$, with $V = \bigcup_{\tau \in \mathcal{T}} \tau$. Multivariate regression chain graphs, are also called *bi-directed chain graphs* by Drton et al. (2007).

In chain graphs, the chain components are ordered, and for any chain component τ, we can define the past of τ, composed of all the nodes contained in the previous chain components. The Markov property for a multivariate regression graph requires that any two variables X_u and X_v in a chain component τ will be considered conditionally on all the variables in the previous chain components. Thus, the meaning of a missing bi-directed edge between u and v is that

$$u \perp\!\!\!\perp v | \text{ the past of } \tau.$$

More generally, it is required that the conditional probability distribution of $\boldsymbol{X}_\tau$ conditional on all the variables in the past of τ is globally Markov with respect to the covariance subgraph G_τ. Instead, a missing arrow $u \leftarrow v$, with X_u and X_v belonging to two different chain components, is interpreted as the conditional independence

$$u \perp\!\!\!\perp v | \text{ the past of } \tau \setminus \{v\},$$

as is customary in multivariate regression models. These independencies are contained in the *block-recursive Markov property* for the multivariate regression chain graphs, which is detailed in the Appendix, properties (i) to (iii). Notice that a covariance graph is a trivial special case of a multivariate regression chain graph with only one chain component.

Example 7. The graph in Figure 2 is a multivariate regression chain graph with two chain components, $\tau_1 = \{1, 2, 3\}$ and $\tau_2 = \{4, 5\}$. The independencies implied by the block-recursive Markov property are

$$4 \perp\!\!\!\perp 5, \quad 1 \perp\!\!\!\perp 5|4, \quad 3 \perp\!\!\!\perp 4|5, \quad 2 \perp\!\!\!\perp 45, \quad 12 \perp\!\!\!\perp 5|4, \quad 23 \perp\!\!\!\perp 4|5, \quad 1 \perp\!\!\!\perp 3|45. \quad (2)$$

The model associated with this graph is a variant of the seemingly unrelated regression model with three responses and two explanatory variables.

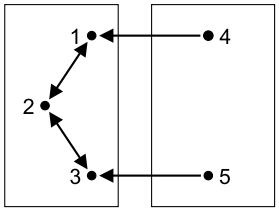

Fig. 2. A multivariate regression chain graph with two components.

4 Marginal log-linear parameterizations for multivariate regression chain graphs

Marginal log-linear models can be used to fit several classes of discrete graphical models. See Rudas et al. (2006), for directed acyclic graph models and chain graphs, Bartolucci et al. (2007) for block recursive models with ordinal variables and Lupparelli, et al. (2008), for bi-directed graph models. In this paper we discuss the parameterization of multivariate regression chain graph models, in the marginal log-linear approach, mainly through simple examples.

In marginal log-linear models, the independence constraints are obtained by zero restrictions on the parameters $\boldsymbol{\lambda}_L^M$, for a specific collection of subsets $L \subseteq M$. Given two subsets A, B of a set M define the collection $\mathcal{Q}_M(A, B)$ of subsets of M containing at least one element of A and at least one element of B. This set can be defined formally by

$$\mathcal{Q}_M(A, B) = \mathcal{P}(M) \setminus [\mathcal{P}(A \cup C) \cup \mathcal{P}(B \cup C)], \text{ with } C = M \setminus (A \cup B),$$

where, as before, $\mathcal{P}(M)$ denotes the power set of M. For example,

$$\mathcal{Q}_{1234}(1, 23) = 12, 13, 123, 124, 134, 1234, \text{ and } \mathcal{Q}_{123}(1, 23) = 12, 13, 123.$$

If $\boldsymbol{\lambda}_L^M$, $L \in \mathcal{L}$ are the marginal log-linear parameters calculated within a margin M, and if A, B are two disjoint subsets of M, then it can be shown that

$$A \perp\!\!\!\perp B \mid M \setminus (A \cup B) \iff \boldsymbol{\lambda}_L^M = \mathbf{0} \text{ for } L \in \mathcal{Q}_M(A, B). \tag{3}$$

This means that, for example, a conditional independence between two variables u and v given $M \setminus \{u, v\}$ is equivalent to the vanishing of the log-linear parameters indexed by all subsets of M containing u and v. Another useful result is that, when $M = A \cup B$,

$$A \perp\!\!\!\perp B \iff \boldsymbol{\lambda}_L^L = \mathbf{0} \text{ for } L \in \mathcal{Q}_M(A, B) \tag{4}$$

showing that marginal independencies are equivalent to zero restrictions on selected multivariate logistic parameters; see Kauermann (1997) and Lupparelli, et al. (2008). Using relations (3) and (4), we can define a conditional

Table 2. Parameterization of the discrete bi-directed graph model of Figure 1.

Margin M_j	Terms $\mathcal{L}_j$	Independence Constraints $\mathcal{C}_j$	
14	1, 4, 14	$1 \perp\!\!\!\perp 4$	$\mathcal{Q}_{14}(1,4) = 14$
124	2, 12, 24, 124	$2 \perp\!\!\!\perp 4 \mid 1$	$\mathcal{Q}_{124}(2,4) = 24, 124$
134	3, 13, 34, 134	$1 \perp\!\!\!\perp 3 \mid 4$	$\mathcal{Q}_{134}(1,3) = 13, 134$
1234	23, 123, 234, 1234		

independence $A \perp\!\!\!\perp B|C$ within a margin $M = A \cup B \cup C$, using marginal log-linear parameters, provided that the set $\mathcal{L}$ contains $\mathcal{Q}_M(A, B)$. Notice that the set of independence constraints is hierarchical in the sense explained in Section 2.

Example 8. Consider the covariance graph models described by the graph of Figure 1 and by the independencies (1). This model can be parameterized by the marginal log-linear model defined by Table 1 and with the constraints of Table 2. The constraints $\boldsymbol{\lambda}_L^{M_j} = \mathbf{0}$ for $L \in \mathcal{C}_j$, $j = 1, \ldots, 3$ are exactly equivalent to the three independencies stated in equation (1). For other marginal log-linear parameterizations see Lupparelli et al. (2008) and for a different approach to covariance graph models, see Drton and Richardson (2008).

Sometimes a conditional independence is defined by specifying constraints on parameters of more than one margin, by combining independencies obtained in previous margins with zero constraints on higher-order log-linear parameters in the current margin, as shown in the following Example 9.

Example 9. The multivariate regression chain graph model shown in Figure 2, specifying the conditional independencies of equation (2), can be parameterized as detailed by Table 3. The first four independence statements can be

Table 3. Parameterization of the discrete multivariate regression chain graph model of Figure 2.

Margin M_j	Terms $\mathcal{L}_j$	Independence Constraints	$\mathcal{C}_j$
45	4, 5, 45	$4 \perp\!\!\!\perp 5$	$\mathcal{Q}_{45}(4,5) =$ 45.
145	1, 14, 15, 145	$1 \perp\!\!\!\perp 5 \mid 4$	$\mathcal{Q}_{145}(1,5) =$ 15, 145.
345	3, 34, 35, 345	$3 \perp\!\!\!\perp 4 \mid 5$	$\mathcal{Q}_{345}(3,4) =$ 34, 345.
245	2, 24, 25, 245	$2 \perp\!\!\!\perp 45$	$\mathcal{Q}_{245}(2,45) =$ 24, 25, 245.
1245	12, 124, 125, 1245	$12 \perp\!\!\!\perp 5 \mid 4$	125, 1245.
2345	23, 234, 235, 2345	$23 \perp\!\!\!\perp 4 \mid 5$	234, 2345.
1345	13, 134, 135, 1345	$1 \perp\!\!\!\perp 3 \mid 45$	$\mathcal{Q}_{1345}(1,3) =$ 13, 134, 135, 1345.
12345	all others		

defined by separate constraints using the operator $\mathcal{Q}$. The conditional independence $12 \perp\!\!\!\perp 5|4$ cannot be defined by the constraints $\mathcal{Q}_{1245}(12, 5)$ that are

not feasible with the available parameters. Instead, we use the equivalence

$$1 \perp\!\!\!\perp 5|4, \quad 2 \perp\!\!\!\perp 5|4 \text{ and } \boldsymbol{\lambda}^{1245}_{125} = \mathbf{0}, \quad \boldsymbol{\lambda}^{1245}_{1245} = \mathbf{0} \iff 12 \perp\!\!\!\perp 5|4.$$

Therefore, as the first two conditional independencies are already defined by the constraints in the margins 145 and 245, we need just the remaining constraints on the three-factor and four-factor log-linear parameters in margin 1245. The same argument is used to define the independence $23 \perp\!\!\!\perp 4|5$.

The parameterization uses a first margin 45 containing the explanatory variables, and then margins defined by the union of 45 and all possible subsets of the responses 123. The meaning of the free parameters is in general complicated by the presence of higher-level log-linear parameters. Thus, in this case, it is tempting to say that the parameters $\boldsymbol{\lambda}^{145}_{14}$ and $\boldsymbol{\lambda}^{345}_{35}$ are conditional association measures between a response and its direct explanatory variable, i.e. between X_1 and X_4 given X_5, and between X_3 and X_5 given X_4, respectively. However this interpretation is obscured by the presence of the non-zero log-linear parameters defined by the supersets of 12 or 23, i.e. indexed by 12, 23 and all higher level terms including them.

5 Maximum likelihood fitting and applications

Assuming a multinomial sampling scheme with sample size N, the vector of cell frequencies $n(\boldsymbol{i})$ for $\boldsymbol{i} \in \mathcal{I}$, has a multinomial distribution with parameters N and $p(\boldsymbol{i})$. If $\boldsymbol{\omega}$ denotes the vector of $\omega(\boldsymbol{i}) = \log E\{n(\boldsymbol{i})\}$, the full vector $\boldsymbol{\lambda}$ of all marginal log-linear parameters in the saturated model may be written in the form

$$\boldsymbol{\lambda} = \boldsymbol{C} \log \boldsymbol{T} \exp(\boldsymbol{\omega}),$$

where $\boldsymbol{T}$ and $\boldsymbol{C}$ are suitable matrices defining the marginal expected counts and the log-linear parameters, respectively. For details on the computation of such matrices see Bartolucci et al. (2007). Given any marginal log-linear model defined by a sequence of margins M_j and a sequence of constraints $\mathcal{C}_j$, we can always split $\boldsymbol{\lambda}$ in two parts $(\boldsymbol{\lambda}_u, \boldsymbol{\lambda}_c)$, such that the reduced model is obtained by

$$\boldsymbol{\lambda}_c = \boldsymbol{h}(\boldsymbol{\omega}) = \mathbf{0},$$

for a suitable function $\boldsymbol{h}$. Thus, the inference problem is based on a log-likelihood function

$$l(\boldsymbol{\omega}) = \sum_{\boldsymbol{i} \in \mathcal{I}} n(\boldsymbol{i}) \omega(\boldsymbol{i}) - \sum_{\boldsymbol{i} \in \mathcal{I}} \exp(\omega(\boldsymbol{i})),$$

to be maximized under the constraints

$$\boldsymbol{h}(\boldsymbol{\omega}) = \mathbf{0}, \quad \sum_{\boldsymbol{i} \in \mathcal{I}} \exp(\omega(\boldsymbol{i})) = N.$$

This constrained estimation problem can be solved by the method of Lagrange multipliers, studied by Aitchison and Silvey (1958), and specifically developed for marginal log-linear models by Bergsma (1997) among others. Accounts of this approach can be found in Rudas et al. (2007) and Lupparelli et al. (2008). These authors discuss an iterative gradient-based procedure, with step size adjustment, which usually converges if the model and the chosen starting point are well specified. Here we discuss the related issue of determining a good starting point of the algorithm, based on a technique, developed by Cox and Wermuth (1990), for generating close approximations to maximum likelihood estimates, assuming that the independence constraints of the model are well compatible with the observations. This theory concerns general situations in which a curved exponential family model is expanded to a saturated form, i.e. to the full family; see Cox (2006). Denoting by $\widehat{\boldsymbol{\lambda}}^{sat}$ the maximum likelihood estimates under the saturated (i.e. covering) model and by $\widehat{\boldsymbol{\lambda}}$, the maximum likelihood estimates under the constrained (i.e. originating) model, the latter can be approximated by the following explicit reduced model estimates $\widetilde{\boldsymbol{\lambda}}$, with

$$\widetilde{\boldsymbol{\lambda}}_u = \widehat{\boldsymbol{\lambda}}_u^{sat} - \boldsymbol{\Sigma}_{uc}\boldsymbol{\Sigma}_{cc}^{-1}\widehat{\boldsymbol{\lambda}}_c^{sat}, \quad \widetilde{\boldsymbol{\lambda}}_c = \mathbf{0}, \tag{5}$$

where $\boldsymbol{\Sigma}_{uc}$ and $\boldsymbol{\Sigma}_{cc}$ are sub-matrices of the asymptotic covariance matrix $\mathrm{cov}(\widehat{\boldsymbol{\lambda}})$, under the saturated model. This representation has the attractive feature of expressing the constrained estimate of the parameters as a modification of the corresponding unconstrained estimate. Though the explicit reduced model estimates are asymptotically efficient, they are recommended only for observations which strongly support the reduced model. Wermuth et al. (2006), use the approximate estimates (5) in case of Gaussian covariance graph models and Roddam (2004) shows an application to multivariate discrete data regression models. In our framework of chain graph models for discrete data, denoting by $\boldsymbol{p}$ the vector of the cell probabilities, the asymptotic covariance matrix of the observed proportions $\widehat{\boldsymbol{p}}$ is

$$\mathrm{cov}(\widehat{\boldsymbol{p}}) = N^{-1}\{\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T\} = \boldsymbol{\Omega}(\boldsymbol{p}),$$

where $\mathrm{diag}(\boldsymbol{p})$ is a diagonal matrix operator. Therefore, the explicit formula can be used with

$$\boldsymbol{\Sigma} = \boldsymbol{J}\boldsymbol{\Omega}(\widehat{\boldsymbol{p}})\boldsymbol{J}^T, \text{ where } \boldsymbol{J} = \boldsymbol{C}\mathrm{diag}(\boldsymbol{T}\widehat{\boldsymbol{p}})^{-1}\boldsymbol{T}.$$

The reduced model estimates $\widetilde{\boldsymbol{\lambda}}$ are relatively easy to obtain, provided that the computation of the matrix $\boldsymbol{J}$ is carried out efficiently, because the matrices involved are usually large and sparse. To use the approximation, we back transform the estimates to the fitted counts, and this cannot be done explicitly in general, but by some iterative scheme. A Newton-Raphson algorithm is usually fast (cf. Glonek and McCullagh (1995)). The approximate estimates

$\widetilde{\boldsymbol{n}}$ of the cell counts can be used as starting points $\widetilde{\boldsymbol{\omega}} = \log \widetilde{\boldsymbol{n}}$ for the general constrained optimization algorithm discussed for example by Lupparelli et al. (2008). Usually, the number of further iterations to convergence is around 20.

Example 10. We show an application to data collected on 13798 individuals concerning 5 variables obtained from the U.S. General Social Survey (Davis et al. (2007)), for years 1972-2006. The variables are reported below with the original name from the GSS Codebook.

- A ABRAPE: do you think it should be possible for a pregnant woman to obtain legal abortion if she became pregnant as a result of rape? (1= yes, 2 = no).
- J SATJOB: how satisfied are you with the work you do? (1 = very satisfied, 2= moderately satisfied, 3 = a little dissatisfied, 4= very dissatisfied). Categories 3 and 4 were merged together.
- F CONFINAN: confidence in banks and financial institutions (1= a great deal, 2= only some, 3= hardly any).
- G GUNLAW: would you favor or oppose a law which would require a person to obtain a police permit before he or she could buy a gun? (1=favor, 2=oppose).
- S SEX: Gender.

The data concern only individuals with complete observations, and we did not attempt to correct for the high number of missing values. Therefore, the following exploratory analysis is intended only as an illustration. We arranged the variables in two components putting gender S in a block of explanatory variables and treating the other variables as joint responses. A preliminary analysis of data suggested the independencies $J \perp\!\!\!\perp S, \quad G \perp\!\!\!\perp JA|S, \quad A \perp\!\!\!\perp FG|S$. The first resulted from a Pearson's chi-squared test, whereas the others were obtained after fitting two separate logit models for the responses G and A. The multivariate regression chain graph model shown in Figure 3, represents the stated independencies under the block-recursive Markov property, as de-

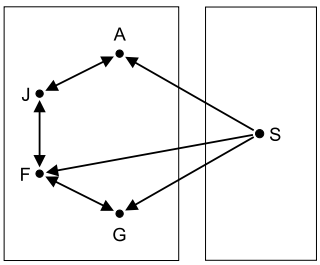

Fig. 3. A multivariate regression chain graph for the data from the U.S. General Social Survey.

tailed in the Appendix. These are equivalent to

$$J \perp\!\!\!\perp S, \quad A \perp\!\!\!\perp G|S, \quad A \perp\!\!\!\perp F|GS, \quad G \perp\!\!\!\perp J|AS,$$

and thus an appropriate marginal log-linear parameterization is generated by the sequence of margins $(AS, JS, FS, GS, AGS, AFGS, AGJS, AFGJS)$, and the independencies can be specified by the constraints

$$\mathcal{C}_1 = \mathcal{Q}_{JS}(J,S),\ \mathcal{C}_2 = \mathcal{Q}_{AGS}(A,G),\ \mathcal{C}_3 = \mathcal{Q}_{AFGS}(A,F),\ \mathcal{C}_4 = \mathcal{Q}_{AGJS}(G,J).$$

The model was fitted, by maximum likelihood, using the Aitchison and Silvey approach, obtaining an adequate fit with a deviance of 19.14, on 20 degrees of freedom. Starting from a default initial point, the algorithm takes 564 iterations to converge, but only 11 from the approximate reduced model estimates (5). For comparison, the deviance of the reduced model fit is 19.16.

6 Appendix A : Separation and Markov properties

In a covariance graph given three disjoint subsets A, B and C of the node set V, A and B are said to be *separated by* C if for any node u in A and any v in B all paths from u to v have at least one inner node in C. A joint probability distribution p of the discrete random vector $\boldsymbol{X} = (X_v)$, for $v \in V$, is said to be *globally Markov* with respect to a covariance graph G if, for any triple of disjoint sets A, B and C,

$$A \perp\!\!\!\perp B \mid V \setminus (A \cup B \cup C) \text{ whenever } A \text{ is separated from } B \text{ by } C \text{ in } G, \quad (6)$$

(see Kauermann (1996)). Associated with any chain graph and thus also for multivariate regression chain graphs is a directed acyclic graph $\mathcal{D}$ over the chain components with nodes $\mathcal{T}$ and having an edge $\tau \leftarrow \tau'$ if there exist an arrow $u \leftarrow v$ in G with $u \in \tau$ and $v \in \tau'$. We can thus define the two concepts of parent of a node v in the chain graph G, $\mathrm{pa}(v)$ and of parent of a chain component τ, $\mathrm{pa}_{\mathcal{D}}(\tau)$. Similarly we define the concept of non-descendants of a component $\mathrm{nd}_{\mathcal{D}}(\tau)$ as the set of all components τ' such that there is no direction-preserving path from τ to τ' in $\mathcal{D}$.

Then we say that a probability distribution p of the discrete random vector $\boldsymbol{X} = (X_v)$ satisfies the *block recursive Markov property* with respect to the multivariate regression chain graph G if for any chain component τ,

(i) $\tau \perp\!\!\!\perp (\mathrm{nd}_{\mathcal{D}}(\tau) \setminus \mathrm{pa}_{\mathcal{D}}(\tau)) \mid \mathrm{pa}_{\mathcal{D}}(\tau)$
(ii) the conditional distribution $\tau \mid \mathrm{pa}_{\mathcal{D}}(\tau)$ is globally Markov with respect to the covariance subgraph G_τ;
(iii) for any subset A of τ: $A \perp\!\!\!\perp (\mathrm{pa}_{\mathcal{D}}(\tau) \setminus \mathrm{pa}(A)) \mid \mathrm{pa}(A))$.

References

AITCHISON, J. and SILVEY, S.D. (1958): Maximum likelihood estimation of parameter subject to restraints. *Annals of Mathematical Statistics 29*, 813–828.

BARTOLUCCI, F., COLOMBI, R. and FORCINA, A. (2007): An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica 17, 691– 711.*

BERGSMA, W.P. (1997): *Marginal models for categorical data*, Ph.D Thesis, Tilburg.

BERGSMA, W.P. and RUDAS, T. (2002): Marginal log-linear models for categorical data. *Annals of Statistics 30, 140 – 159.*

COX, D.R. (2006): *Principles of Statistical Inference.* Cambridge University Press, Cambridge.

COX, D.R. and WERMUTH, N. (1990): An approximation to maximum likelihood estimates in reduced models. *Biometrika 77*, 747–761.

COX, D.R. and WERMUTH, N. (1996): *Multivariate dependencies. Models, analysis and interpretation.* Chapman and Hall, London.

DAVIS, J.A., SMITH, T.W. and MARDSEN, J.A. (2007): *General Social Surveys Cumulative Codebook: 1972-2006.* NORC, Chicago. URL http://sda.berkeley.edu/GSS/.

DRTON, M., EICHLER, M. and RICHARDSON, T.S. (2007): Identification and Likelihood Inference for Recursive Linear Models with Correlated Errors. *arXiv:0601631.*

DRTON, M. and RICHARDSON, T.S. (2008): Binary models for marginal independence, *J. Royal Statist. Soc., Ser. B 70*, 287–309.

GLONEK, G.F.V. and McCULLAGH, P. (1995): Multivariate logistic models. *J. Royal Statist. Soc., Ser. B 57*, 533–546.

KAUERMANN, G. (1996): On a dualization of graphical Gaussian models. *Scand. J. of Statistics 23*, 105–116.

KAUERMANN, G. (1997): A note on multivariate logistic models for contingency tables. *Australian Journal of Statistics 39(3)*, 261–276.

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

LUPPARELLI, M., MARCHETTI, G.M. and BERGSMA, W.P. (2008): Parameterizations and fitting of bi-directed graph models to categorical data. *arXiv:0801.1440.*

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.

RICHARDSON, T.S. (2003): Markov property for acyclic directed mixed graphs, *Scand. J. of Statistics 30*, 145–157.

RODDAM, A.W. (2004): An approximate maximum likelihood procedure for parameter estimation in multivariate discrete data regression models. *J. of Applied Statistics 28*, 273–279.

RUDAS, T., BERGSMA, W.P. and NÉMETH, R. (2006): Parameterization and estimation of path models for categorical data. In: A. Rizzi & M. Vichi, (Eds.): *Compstat 2006 Proceedings.* Physica-Verlag, Heidelberg, 383–394.

WERMUTH, N. and COX, D.R. (2004): Joint response graphs and separation induced by triangular systems. *J. Royal Statist. Soc. B 66*, 687-717.

WERMUTH, N., COX, D.R. and MARCHETTI, G.M. (2006): Covariance chains. *Bernoulli, 12*, 841–862.

Part V

Computational Econometrics

Exploring the Bootstrap Discrepancy

Russell Davidson[1,2]

[1] Department of Economics,
McGill University,
855 Sherbrooke Street West,
Montreal, Quebec, Canada H3A 2T7, *russell.davidson@mcgill.ca*
[2] GREQAM,
Centre de la Vieille Charité,
13236 Marseille cedex 02, France

Abstract. Many simulation experiments have shown that, in a variety of circumstances, bootstrap tests perform better than current asymptotic theory predicts. Specifically, the discrepancy between the actual rejection probability of a bootstrap test under the null and the nominal level of the test appears to be smaller than suggested by theory, which in any case often yields only a rate of convergence of this discrepancy to zero. Here it is argued that the Edgeworth expansions on which much theory is based provide a quite inaccurate account of the finite-sample distributions of even quite basic statistics. Other methods are investigated in the hope that they may give better agreement with simulation evidence. They also suggest ways in which bootstrap procedures can be improved so as to yield more accurate inference.

Keywords: bootstrap discrepancy, bootstrap test, Edgeworth expansion

1 Introduction

Since the bootstrap was introduced by Efron (1979), its use by statisticians and econometricians has grown enormously; see for instance Horowitz (2001) for a useful survey. Asymptotic theory for the bootstrap has not been in short supply; after Bickel and Freedman (1981), landmark contributions have been Beran (1987) and (1988), and especially Hall (1992), in which a profound connection is established between bootstrap inference and Edgeworth expansions.

Although current asymptotic theory for the bootstrap accounts for many of the properties of bootstrap inference as discovered by simulation experiments, a recurrent phenomenon is that the bootstrap performs better than theory indicates. In this paper, I argue that the approximations provided by Edgeworth expansions are quite inadequate to describe the behaviour of bootstrap tests, and look at other methods which, while still inadequate, give quite different results. My hope is that the approach outlined here can give better insights into the properties of the bootstrap. One suggestion developed in this paper leads to the possibility of designing improved bootstrap schemes.

2 The Bootstrap discrepancy

Suppose that a test statistic τ is designed to test a particular null hypothesis. The set of all DGPs that satisfy that hypothesis is denoted as $\mathbb{M}_0$; this set constitutes what we may call the null model. A bootstrap test based on the statistic τ approximates the distribution of τ under a DGP $\mu \in \mathbb{M}_0$ by its distribution under a bootstrap DGP that also belongs to $\mathbb{M}_0$ and can be thought of as an estimate of the true DGP μ.

We define the *bootstrap discrepancy* as the difference, as a function of the true DGP and the nominal level, between the actual rejection probability and the nominal level. In order to study it, we suppose, without loss of generality, that the test statistic is already in approximate P value form, so that the rejection region is to the left of a critical value.

The *rejection probability function*, or RPF, depends both on the nominal level α and the DGP μ. It is defined as

$$R(\alpha, \mu) \equiv \Pr_\mu(\tau < \alpha). \tag{1}$$

We assume that, for all $\mu \in \mathbb{M}$, the distribution of τ has support $[0, 1]$ and is absolutely continuous with respect to the uniform distribution on that interval. For given μ, $R(\alpha, \mu)$ is just the CDF of τ evaluated at α. The inverse of the RPF is the *critical value function*, or CVF, which is defined implicitly by the equation

$$\Pr_\mu\big(\tau < Q(\alpha, \mu)\big) = \alpha. \tag{2}$$

It is clear from (2) that $Q(\alpha, \mu)$ is the α-quantile of the distribution of τ under μ. In addition, the definitions (1) and (2) imply that

$$R\big(Q(\alpha, \mu), \mu\big) = Q\big(R(\alpha, \mu), \mu\big) = \alpha \tag{3}$$

for all α and μ.

In what follows, we assume that the distribution of τ under the bootstrap DGP, which we denote by μ^*, is known exactly. The bootstrap critical value for τ at level α is then $Q(\alpha, \mu^*)$. If τ is approximately (for example, asymptotically) pivotal relative to the model $\mathbb{M}_0$, realisations of $Q(\alpha, \mu^*)$ under DGPs in $\mathbb{M}_0$ should be close to α. This is true whether or not the true DGP belongs to the null model, since the bootstrap DGP μ^* does so. The bootstrap discrepancy arises from the fact that, in a finite sample, $Q(\alpha, \mu^*) \neq Q(\alpha, \mu)$.

Rejection by the bootstrap test is the event $\tau < Q(\alpha, \mu^*)$. Applying the increasing transformation $R(\cdot, \mu^*)$ to both sides and using (3), we see that the bootstrap test rejects whenever

$$R(\tau, \mu^*) < R\big(Q(\alpha, \mu^*), \mu^*\big) = \alpha. \tag{4}$$

Thus the bootstrap P value is just $R(\tau, \mu^*)$, which can therefore be interpreted as a bootstrap test statistic.

We define two random variables that are deterministic functions of the two random elements, τ and μ^*, needed for computing the bootstrap P value $R(\tau, \mu^*)$. The first of these random variables is distributed as U(0, 1) under μ; it is

$$p \equiv R(\tau, \mu). \tag{5}$$

The uniform distribution of p follows from the fact that $R(\cdot, \mu)$ is the CDF of τ under μ and the assumption that the distribution of τ is absolutely continuous on the unit interval for all $\mu \in \mathbb{M}$. The second random variable is

$$q \equiv R(Q(\alpha, \mu^*), \mu) - \alpha = R(Q(\alpha, \mu^*), \mu) - R(Q(\alpha, \mu), \mu). \tag{6}$$

Let the CDF of q under μ conditional on the random variable p be denoted as $F(q \mid p)$. Then it is shown in Davidson and MacKinnon (2006) that the bootstrap discrepancy can be expressed as

$$\int_{-\alpha}^{1-\alpha} x \, \mathrm{d}F(x \mid \alpha + x). \tag{7}$$

that is, the expectation of q conditional on p being at the margin of rejection at level α.

The random variable $q + \alpha$ is the probability that a statistic generated by the DGP μ is less than the α-quantile of the bootstrap distribution, conditional on that distribution. The expectation of q can thus be interpreted as the bias in rejection probability when the latter is estimated by the bootstrap. The actual bootstrap discrepancy, which is a nonrandom quantity, is the expectation of q conditional on being at the margin of rejection.

2.1 An asymptotically normal statistic

In some approaches to approximating the bootstrap discrepancy, it is assumed that the statistic is in asymptotically N(0,1) rather than approximately U(0,1) form. This is the case for the Edgeworth expansion approach considered in the next section. It is useful to define the random variables p and q in terms of new functions R_N and Q_N that respectively express the CDF and quantile function of the approximately normal statistic. Thus $R_N(x, \mu)$ is the CDF of the statistic under DGP μ, while $Q_N(\alpha, \mu)$ is the α-quantile. It is easy to see that $R_N(x, \mu) = R(\Phi(x), \mu)$ and $Q_N(\alpha, \mu) = \Phi^{-1}(Q(\alpha, \mu))$, where Φ is the CDF of the N(0,1) distribution. If now we denote the approximately normal statistic by τ_N, we see that $q = R_N(Q_N(\alpha, \mu^*), \mu) - \alpha$ and $p = R_N(\tau_N, \mu)$; compare (6) and (5). Here we assume that the rejection region is to the left, as it would be for a statistic in P value form. Straightforward modifications can handle two-tailed tests or tests that reject to the right.

3 Approximations to the Bootstrap discrepancy

3.1 Edgeworth expansion

Suppose that the statistic τ_N is computed using data generated by a DGP μ. Under the null hypothesis that τ_N is designed to test, we suppose that its distribution admits a valid Edgeworth expansion; see Hall (1992) for a complete treatment of Edgeworth expansions in connection with the bootstrap. The expansion takes the form

$$R_N(x,\mu) = \Phi(x) - n^{-1/2}\phi(x)\sum_{i=1}^{\infty} e_i(\mu)\mathrm{He}_{i-1}(x). \tag{8}$$

Here ϕ is the density of the N(0,1) distribution, $\mathrm{He}_i(\cdot)$ is the Hermite polynomial of degree i (see for instance Abramowitz and Stegun (1965), Chapter 22, for details of these polynomials), and the $e_i(\mu)$ are coefficients that are at most of order 1 as the sample size n tends to infinity. The Edgeworth expansion up to order n^{-1} then truncates everything in (8) of order lower than n^{-1}.

The $e_i(\mu)$ can be related to the moments or cumulants of the statistic τ_N as generated by μ by means of the equation

$$n^{-1/2}e_i(\mu) = \frac{1}{i!}\mathrm{E}_\mu\big(\mathrm{He}_i(\tau_N)\big). \tag{9}$$

The bootstrap DGP, μ^*, is realised jointly with τ_N, as a function of the same data. We suppose that this CDF can be expanded as in (8), with the $e_i(\mu)$ replaced by $e_i(\mu^*)$, and so the CDF of the bootstrap statistics is $R_N(x,\mu^*)$. We consider a one-tailed test based on τ_N that rejects to the left. Then, from (8), the random variable $p = R_N(\tau_N,\mu)$ is approximated by the expression

$$\Phi(\tau_N) - n^{-1/2}\phi(\tau_N)\sum_{i=1}^{\infty} e_i(\mu)\mathrm{He}_{i-1}(\tau_N) \tag{10}$$

truncated so as to remove all terms of order lower than n^{-1}. Similarly, the variable q of (6) is approximated by $R'_N(Q_N(\alpha,\mu),\mu)\big(Q_N(\alpha,\mu^*)-Q_N(\alpha,\mu)\big)$, using a Taylor expansion where R'_N is the derivative of R_N with respect to its first argument.

It is convenient to replace μ and μ^* as arguments of R_N and Q_N by the sequences $\boldsymbol{e}$ and $\boldsymbol{e}^*$ of which the elements are the $e_i(\mu)$ and $e_i(\mu^*)$ respectively. Denote by $\mathbf{D}_e R_N(x,\boldsymbol{e})$ the sequence of partial derivatives of R_N with respect to the components of $\boldsymbol{e}$, and similarly for $\mathbf{D}_e Q_N(\alpha,\boldsymbol{e})$. Then, on differentiating the identity $R_N(Q_N(\alpha,\boldsymbol{e}),\boldsymbol{e}) = \alpha$, we find that

$$R'_N(Q_N(\alpha,\boldsymbol{e}),\boldsymbol{e})\mathbf{D}_e Q_N(\alpha,\boldsymbol{e}) = -\mathbf{D}_e R_N(Q_N(\alpha,\boldsymbol{e}),\boldsymbol{e}). \tag{11}$$

To leading order, $Q_N(\alpha, \boldsymbol{e}^*) - Q_N(\alpha, \boldsymbol{e})$ is $\mathbf{D}_e Q_N(\alpha, \boldsymbol{e})(\boldsymbol{e}^* - \boldsymbol{e})$, where the notation implies a sum over the components of the sequences. Thus the variable q can be approximated by

$$-\mathbf{D}_e R_N(Q_N(\alpha, \boldsymbol{e}), \boldsymbol{e})(\boldsymbol{e}^* - \boldsymbol{e}). \tag{12}$$

The Taylor expansion above is limited to first order, because, in the cases we study here, $Q_N(\alpha, \mu^*) - Q_N(\alpha, \mu)$ is of order n^{-1}. This is true if, as we expect, the $e_i(\mu^*)$ are root-n consistent estimators of the $e_i(\mu)$. From (8) we see that component i of $\mathbf{D}_e R(x, \boldsymbol{e})$ is $-n^{-1/2}\phi(x)\mathrm{He}_{i-1}(x)$. To leading order, $Q_N(\alpha, \boldsymbol{e})$ is just z_α, the α-quantile of the N(0,1) distribution. Let $l_i = n^{1/2}(e_i(\mu^*) - e_i(\mu))$. In regular cases, the l_i are or order 1 and are asymptotically normal. Further, let $\gamma_i(\alpha) = \mathrm{E}(l_i \,|\, p = \alpha)$. Then the bootstrap discrepancy (7) at level α is a truncation of

$$n^{-1}\phi(z_\alpha) \sum_{i=1}^{\infty} \mathrm{He}_{i-1}(z_\alpha)\gamma_i(\alpha). \tag{13}$$

3.2 Approximation based on asymptotic normality

If the distribution of a statistic τ_N has an Edgeworth expansion like (8), then it is often the case that τ_N itself can be expressed as a deterministic function of a set of asymptotically jointly normal variables of expectation 0; the special case of the next section provides an explicit example. If so, then the distribution of τ_N can be approximated by that of the same function of variables that are truly, and not just asymptotically, normal. This distribution depends only on the covariance matrix of these variables, and so can be studied at moderate cost by simulation.

In order to study the bootstrap discrepancy, one looks at the covariance matrix under the bootstrap DGP. This is normally an estimate of the true covariance matrix, and can often be expressed as a function of asymptotically normal variables, including those of which τ_N is a function. The joint distribution of the approximate p and q can then be used to approximate the bootstrap discrepancy, in what is of course a very computationally intensive procedure.

3.3 Approximation by matching moments

The Edgeworth expansion (8) is determined by the coefficients $e_i(\mu)$. These coefficients are enough to determine the first four moments of a statistic τ_N up to the order of some specified negative power of n. Various families of distributions exist for which at least the first four moments can be specified arbitrarily subject to the condition that there exists a distribution with those moments. An example is the Pearson family of distributions, of which more

later. A distribution which matches the moments given by the $e_i(\mu)$, truncated at some chosen order, can then be used to approximate the function $R_N(\tau_N, \mu)$ for both the DGP μ and its bootstrap counterpart μ^*. An approximation to the bootstrap discrepancy can then formed in the same way as (13), with a different expression for $\mathbf{D}_e R_N(z_\alpha, \boldsymbol{e})$.

4 A special case: I. The distribution

One of the simplest tests imaginable is a test that the expectation of a distribution is zero, based on an IID sample of n drawings, u_t, $t = 1, \dots, n$, from that distribution. We suppose that the expectation is indeed zero, and that the variance exists. The sample mean is $\hat{a} = n^{-1} \sum_t u_t$, and the sample variance, under the null that the expectation is zero, is $\hat{\sigma}^2 = n^{-1} \sum_t u_t^2$. A statistic that is asymptotically standard normal under the null is then $n^{1/2}\hat{a}/\hat{\sigma}$. Since this is homogeneous of degree 0 in the u_t, we may without loss of generality suppose that their true variance is 1. If we define the asymptotically normal variables $w_i = n^{-1/2} \sum_{t=1}^n \big(\mathrm{He}_i(u_t) - \mathrm{E}(\mathrm{He}_i(u_t))\big)$, $i = 1, 2, \dots$, then the statistic can be written as

$$w_1/(1 + n^{-1/2} w_2)^{1/2}. \tag{14}$$

On expanding the denominator by use of the binomial theorem, and truncating everything of order lower than n^{-1}, we can study the approximate test statistic

$$\tau_N = w_1 - \tfrac{1}{2} n^{-1/2} w_1 w_2 + \tfrac{3}{8} n^{-1} w_1 w_2^2. \tag{15}$$

4.1 The Edgeworth expansion

In order to apply the methodologies of Section 3.1 or Section 3.3, we have first to compute the expectations of the Hermite polynomials evaluated at τ_N. The quantities $e_i(\mu)$ can then be computed using (9) – here μ is the DGP that generates samples of n IID drawings from the given distribution. Working always only to order n^{-1} means that we need the $e_i(\mu)$ only to order $n^{-1/2}$. We see that

$$\begin{aligned} e_1(\mu) &= -\tfrac{1}{2}\kappa_3, & e_2(\mu) &= n^{-1/2}\kappa_3^2, \\ e_3(\mu) &= -\tfrac{1}{3}\kappa_3, & e_4(\mu) &= \tfrac{1}{12} n^{-1/2}(8\kappa_3^2 - 3 - \kappa_4) \\ e_5(\mu) &= 0, & e_6(\mu) &= \tfrac{1}{144} n^{-1/2}(9 + 8\kappa_3^2 - 3\kappa_4), \end{aligned} \tag{16}$$

where κ_3 and κ_4 are the third and fourth cumulants respectively of the distribution from which the u_t are drawn. All $e_i(\mu)$ for $i > 6$ are zero to order $n^{-1/2}$. The Edgeworth expansion of the distribution of τ_N is then, from (8),

$$\begin{aligned} R_N(x, \mu) = \Phi(x) + \phi(x) \Big(& \tfrac{1}{6} n^{-1/2} \kappa_3 (1 + 2x^2) + n^{-1} \big(\tfrac{1}{48} x (8\kappa_3^2 + 3\kappa_4 - 81) \\ & + \tfrac{1}{72} x^3 (63 - 8\kappa_3^2 - 9\kappa_4) - \tfrac{1}{144} x^5 (9 + 8\kappa_3^2 - 3\kappa_4) \big) \Big) \end{aligned} \tag{17}$$

For many numerical illustrations, we will use the Pearson family of distributions. By adjusting four parameters, the first and second moments can be set to 0 and 1 respectively, and the third and fourth cumulants can be chosen from a wide range. In Table 1 are shown the maximum differences between the true CDF of a statistic of the form (14), as estimated using 100,000 simulations, and the asymptotically normal approximation (d_0), the approximation given by (17) through the $n^{-1/2}$ term (d_1), and through the n^{-1} term (d_2), for a range of sample sizes, and values of κ_3 and κ_4. It can be seen that for large values of κ_3 and κ_4, the Edgeworth approximations are not close to the true distribution until the standard normal approximation is also fairly close. What the table does not show is that the Edgeworth approximations are not necessarily bounded between 0 and 1, and are not necessarily increasing.

4.2 The asymptotic normality approximation

The distributions of the statistics (14) and (15), both functions of the asymptotically normal w_1 and w_2, can be approximated by those of the same functions of two genuinely normal variables z_1 and z_2, with the same first and second moments as those of w_1 and w_2. We have $\mathrm{var}(w_1) = 1$, $\mathrm{var}(w_2) = 2 + \kappa_4$, and $\mathrm{cov}(w_1, w_2) = \kappa_3$. Measures of the maximum differences between the true CDF and the approximation based on (14) are shown as d_3 in Table 1. The d_3 are smaller than the d_1, especially for large values of κ_3 and κ_4, and are of similar magnitude to the d_2. Of course, the approximations are themselves true distributions, unlike the Edgeworth expansions.

4.3 Matching moments

The first four moments of the statistic (14) are implicitly given to order n^{-1} by (16). They are as follows:

$$\begin{aligned} \mathrm{E}(\tau) &= -\tfrac{1}{2} n^{-1/2} \kappa_3 & \mathrm{E}(\tau^2) &= 1 + 2n^{-1}\kappa_3^2 \\ \mathrm{E}(\tau^3) &= -\tfrac{7}{2} n^{-1/2} \kappa_3 & \mathrm{E}(\tau^4) &= 3 + 2n^{-1}(14\kappa_3^2 - \kappa_4 - 3). \end{aligned} \tag{18}$$

The distribution of (14) can be approximated by a Pearson distribution with those cumulants. Again, this is a true distribution. The maximum differences between the true CDF and this approximation are given as d_4 in Table 1.

5 A special case: II. The Bootstrap discrepancy

5.1 The Edgeworth approximation

In order to compute the approximate bootstrap discrepancy (13), we make use of the differences $e_i(\mu^*) - e_i(\mu)$ between the coefficients of the Edgeworth expansion of the bootstrap statistic and that of the statistic itself. The

$e_i(\mu^*)$ are given by the expressions in (16) with κ_3 and κ_4 replaced by the estimates $\hat{\kappa}_3$ and $\hat{\kappa}_4$ used, explicitly or implicitly, in the bootstrap DGP. The most obvious bootstrap DGP is a resampling DGP in which the elements of a bootstrap sample are drawn at random, with replacement, from the u_t after centring. Since the statistic is scale invariant, the distribution of the bootstrap statistics would be the same if we resampled the $(u_t - \hat{a})/\hat{\sigma}$. The third cumulant of the bootstrap distribution is then the third moment of the rescaled quantities, and the fourth cumulant is their fourth moment minus 3. Some algebra then shows that

$$n^{1/2}(\hat{\kappa}_3 - \kappa_3) = w_3 - \tfrac{3}{2}\kappa_3 w_2 - \tfrac{3}{8}n^{-1/2}\big(8w_1w_2 + 4w_2w_3 - 5\kappa_3 w_2^2\big) \tag{19}$$

$$\begin{aligned} n^{1/2}(\hat{\kappa}_4 - \kappa_4) = {} & w_4 - 4\kappa_3 w_1 - 2\kappa_4 w_2 \\ & - n^{-1/2}\big(6w_1^2 - 8\kappa_3 w_1 w_2 + 4w_1 w_3 + 3(1-\kappa_4)w_2^2 + 2w_2 w_4\big) \end{aligned} \tag{20}$$

Thus, from the formulas (16), we can see that

$$l_1 = n^{1/2}\big(e_1(\mu^*) - e_1(\mu)\big) = -\tfrac{1}{4}(2w_3 - 3\kappa_3 w_2) + O_p(n^{-1/2}); \quad l_3 = \tfrac{2}{3}l_1$$

while all l_i for $i \neq 1, 3$ are of order lower than unity. By definition, the w_i are (jointly) asymptotically normal. The variance of w_1 is 1, and so $\mathrm{E}(w_i \mid w_1) = w_1 \mathrm{E}(w_1 w_i)$. Now

$$\mathrm{E}(w_1 w_2) = n^{-1}\sum_{t=1}^{n} \mathrm{E}\big(\mathrm{He}_1(u_t)\mathrm{He}_2(u_t)\big) = \mathrm{E}(u_t^3 - u_t) = \kappa_3.$$

Similarly, $\mathrm{E}(w_1 w_3) = \kappa_4$. The $\gamma_i(\alpha)$ used in the approximate expression (13) for the bootstrap discrepancy are the expectations of the l_i conditional on the event $p = \alpha$. By (10), the variable p is approximated to leading order by $\Phi(\tau_N)$, and, from (15), this is $\Phi(w_1)$, again to leading order. Thus the conditioning event can be written as $w_1 = z_\alpha$. It follows that

$$\gamma_1(\alpha) = -\tfrac{1}{4}z_\alpha(2\kappa_4 - 3\kappa_3^2) \quad \text{and} \quad \gamma_3(\alpha) = \tfrac{2}{3}\gamma_1(\alpha)$$

with error of order lower than 1, all other $\gamma_i(\alpha)$ being of lower order. For our special case, therefore, the bootstrap discrepancy at level α, as approximated by (13), is

$$\tfrac{1}{12}n^{-1}\phi(z_\alpha)(3\kappa_3^2 - 2\kappa_4)z_\alpha(1 + 2z_\alpha^2) \tag{21}$$

We see that this expression vanishes if $3\kappa_3^2 - 2\kappa_4 = 0$. This is true for the normal distribution of course, for which all cumulants of order greater than 2 vanish. But it is true as well for many other commonly encountered distributions. Among these, we find the central chi-squared, exponential, Pearson Type III, and Gamma distributions.

Table 2 gives the maximum differences (d_1) between the actual discrepancy, as evaluated using a simulation with 100,000 replications with 399 bootstrap repetitions, and the approximate discrepancy (21), again for various

sample sizes and various cumulant values. When both κ_3 and κ_4 are zero, the data are standard normal. Some other distributions for which $3\kappa_3^2 - 2\kappa_4 = 0$ are also given.

The approximate discrepancy (21) is an odd function of z_α. The discrepancies approximated by simulation often do not seem to share this property even roughly.

5.2 Matching moments

In this approach, the function $R_N(x, \mu)$ is approximated by the CDF of a Pearson distribution, characterised by the four moments (18). Denote this approximation by $R_N(x, \kappa_3, \kappa_4)$. An approximation to the bootstrap discrepancy can be found exactly as in the preceding subsection. Analogously to (12), we approximate q by

$$-\sum_{i=3}^{4} \frac{\partial R_N}{\partial \kappa_i}(Q_N(\alpha, \kappa_3, \kappa_4), \kappa_3, \kappa_4)(\kappa_i^* - \kappa_i). \tag{22}$$

But, of the four moments (18), only the fourth depends on κ_4, with κ_4 multiplied by n^{-1}. From (20), $\hat{\kappa}_4 - \kappa_4 = O_p(n^{-1/2})$, and so only the term with $i = 3$ contributes to (22) to order n^{-1}. To leading order, $Q_N(\alpha, \kappa_3, \kappa_4) = z_\alpha$, and so the approximate bootstrap discrepancy is

$$\tfrac{1}{2} n^{-1/2} \frac{\partial R_N}{\partial \kappa_3}(z_\alpha, \kappa_3, \kappa_4)(3\kappa_3^2 - 2\kappa_4) z_\alpha, \tag{23}$$

since, from (19), $\mathrm{E}(\hat{\kappa}_3 - \kappa_3 \,|\, w_1 = z_\alpha) = (2\kappa_4 - 3\kappa_3^2) z_\alpha / 2$. Column d_2 in Table 2 gives the maximum differences between the actual discrepancy and (23). Of course, it coincides with column d_1 for all cases with $3\kappa_3^2 - 2\kappa_4 = 0$.

6 Designing a better Bootstrap DGP

6.1 Theoretical considerations

It can be seen both from (7) and the discussion in the previous section of the Edgeworth approximation of the bootstrap discrepancy that its rate of convergence to 0 as $n \to \infty$ is faster if the bootstrapped statistic is uncorrelated with the determinants of the bootstrap DGP. This is often easy to realise with a parametric bootstrap, since a statistic that tests a given null hypothesis is often asymptotically independent of parameters estimated under that null; see Davidson and MacKinnon (1999). But with a nonparametric bootstrap like the resampling bootstrap studied in section 4, it is not obvious how to achieve approximate independence of the statistic and the bootstrap DGP, as shown by the fact the cumulant estimates given in (19) and (20) are correlated with the statistic (15), with the result that the discrepancy (21) is of order n^{-1}.

However, the fact that the Edgeworth approximation (17) depends on just two parameters of the DGP, κ_3 and κ_4, suggests that it might be possible to construct a parametric bootstrap using just these parameters. For instance, the elements of a bootstrap sample could be drawn from the Pearson distribution with expectation 0, variance 1, and third and fourth cumulants given by those estimated using the u_t. The Edgeworth approximation of the bootstrap discrepancy (21) would be unchanged, although the actual bootstrap discrepancy could be smaller or greater than that of the ordinary resampling bootstrap. Another possibility, that would involve no bootstrap simulations at all, would be to use for the bootstrap distribution the Pearson distribution with the moments (18) with the estimated cumulants. We will shortly explore these possibilities by simulation.

We now turn to the questions of why (19) and (20) are correlated with the statistic (15), and whether it is possible to find other cumulant estimates that are approximately uncorrelated with it. First, we look at estimation of the second cumulant, that is, the variance. The sample variance, $\hat{\sigma}^2$, always assuming that the true variance is 1, can be see to be equal to $1 + n^{-1/2}w_2$, and, since $\mathrm{E}(w_1 w_2) = \kappa_3$, it too is correlated with τ_N unless $\kappa_3 = 0$. In fact, $\hat{\sigma}^2$, as a variance estimator, is *inefficient*, since it does not take account of the fact that, under the null, the expectation is 0.

An efficient estimator can be found by various means. Let m_k denote the uncentred moment of order k of the u_t. It can be shown that m_2, m_3, and m_4 can be estimated efficiently by $\tilde{m}_k \equiv \hat{m}_k - (\hat{m}_1 \hat{m}_{k+1})/\hat{m}_2$, $k = 2, 3, 4$. Some algebra then shows that, to leading order,

$$n^{1/2}(\tilde{\kappa}_3 - \kappa_3) = w_3 - \kappa_4 w_2 - \tfrac{3}{2}\kappa_3(w_2 - \kappa_3 w_1) \tag{24}$$

$$n^{1/2}(\tilde{\kappa}_4 - \kappa_4) = w_4 - \kappa_5 w_1 - 4\kappa_3 w_1 - 2\kappa_4(w_2 - \kappa_3 w_1). \tag{25}$$

Here κ_5 is the fifth cumulant. It can be shown that $\mathrm{E}(w_1 w_4) = \kappa_5 + 4\kappa_3$, and that, consequently, (24) and (25) are uncorrelated with w_1. Since $\tilde{\sigma}^2$ is more efficient that $\hat{\sigma}^2$, it makes sense to bootstrap the statistic $n^{1/2}\hat{a}/\tilde{\sigma}$ rather than $n^{1/2}\hat{a}/\hat{\sigma}$. To leading order, this statistic is also equal to w_1, and is thus uncorrelated with $\tilde{\kappa}_3$ and $\tilde{\kappa}_4$.

A bootstrap DGP that uses $\tilde{m}_3$ and $\tilde{m}_4$ can be constructed by using a Pearson distribution parametrised with first and second moments 0 and 1 respectively, and these estimators as third and fourth moments.

6.2 Simulation evidence

The first set of experiments concerns the bootstrap without simulation, in which the moments (18) are used to set up a Pearson distribution, which is then used to obtain a bootstrap P value. There is little point in reporting the simulation results, which, while they confirm that the procedure is possible, show that the distortions are greater than for any other bootstrap procedure considered here.

n	κ_3	κ_4	d_0	d_1	d_2	d_3	d_4
50	2	3	0.042	0.011	0.005	0.017	0.012
100	2	3	0.027	0.005	0.004	0.009	0.007
50	6	40	0.180	0.127	0.095	0.059	0.096
100	6	40	0.106	0.060	0.040	0.038	0.052
50	13	175	0.617	0.599	0.585	0.326	0.604
100	13	175	0.437	0.415	0.400	0.210	0.395
1000	13	175	0.064	0.027	0.016	0.029	0.024

Table 1: Maximum differences between true distribution and various approximations: d_0 for N(0,1), d_1 for $n^{-1/2}$ Edgeworth approximation, d_2 for n^{-1} approximation, d_3 for asymptotic normality approximation, d_4 for matching moments.

n	κ_3	κ_4	$3\kappa_3^2 - 2\kappa_4$	distribution	d_1	d_2
20	0	0	0	N(0,1)	0.003	0.003
50	0	0	0	N(0,1)	0.003	0.003
20	2.828	12	0	χ_1^2	0.022	0.022
50	2.828	12	0	χ_1^2	0.010	0.010
20	2	6	0	exponential	0.017	0.017
50	2	6	0	exponential	0.009	0.009
20	1	1.5	0	Gamma(4)	0.009	0.009
50	1	1.5	0	Gamma(4)	0.004	0.004
20	0	-1.2	-2.4	uniform	0.003	0.005
50	0	-1.2	-2.4	uniform	0.004	0.004
50	2	3	6	Pearson I	0.008	0.008
50	3	11	5	Pearson I	0.013	0.016
50	6	40	28	Pearson I	0.070	0.133
50	9	83	77	Pearson I	0.199	0.434
50	12	175	82	Pearson I	0.228	0.467
100	12	175	82	Pearson I	0.120	0.233
500	12	175	82	Pearson I	0.017	0.025

Table 2: Maximum differences between bootstrap discrepancy and Edgeworth approximation (d_1) and moment-matching approximation (d_2).

n	distribution	d_1	d_2	d_3
20	χ_1^2	0.029	0.033	0.052
50	χ_1^2	0.010	0.011	0.018
20	exponential	0.006	0.024	0.035
50	exponential	0.006	0.011	0.014
20	uniform	0.012	0.006	0.007
50	uniform	0.007	0.003	0.006

Table 3: Maximum P value discrepancies: resampling (d_1), Pearson with inefficient (d_2) and efficient (d_3) moment estimates.

The next set of experiments again uses a Pearson distribution, but this time for the bootstrap disturbances. The moments of the distribution of the residuals determine a Pearson distribution, and the bootstrap disturbances are drawings from this. In a further set of experiments, the moments were estimated with the zero expectation imposed, as discussed in the previous subsection.

7 Conclusions

We have investigated various types of approximations to the bootstrap discrepancy, including the traditional Edgeworth expansion approximations, but not restricted to them. We find that all approaches that are implicitly or explicitly based on estimates of the moments of the disturbances are quantitatively not at all accurate, although their inaccuracies take on very different forms.

We consider bootstrap DGPs based on both unrestricted and restricted estimates of the first few moments of the disturbances, and find that these essentially parametric bootstraps compete well with the conventional resampling bootstrap. It appears that much remains to be learned about the determinants of the bootstrap discrepancy for any given procedure, as well as about different procedures.

References

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

BERAN, R. (1987): Prepivoting to Reduce Level Error of Confidence Sets. *Biometrika* 74, 457-468.

BERAN, R. (1988): Prepivoting Test Statistics: a Bootstrap View of Asymptotic Refinements. *Journal of the American Statistical Association* 83, 687–697.

BICKEL, P.J. and FREEDMAN, D.A. (1981): Some Asymptotic Theory for the Bootstrap. *Annals of Statistics* 9, 1196-1217.

DAVIDSON, R. and MACKINNON, J.G. (1999): The Size Distortion of Bootstrap Tests. *Econometric Theory* 15, 361-376.

DAVIDSON, R. and MACKINNON, J.G. (2004): *Econometric Theory and Methods*. Oxford, New York.

DAVIDSON, R. and MACKINNON, J.G. (2006): The Power of Asymptotic and Bootstrap Tests. *Journal of Econometrics* 133, 421-441.

EFRON, B. (1979): Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics* 7, 1–26.

HALL, P. (1992): *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

HOROWITZ, J.L. (2001): The Bootstrap. In J.J. Heckman and E.E. Leamer (Eds.): *Handbook of Econometrics* Vol. 5, Amsterdam: North-Holland, 3159–3228.

On Diagnostic Checking Time Series Models with Portmanteau Test Statistics Based on Generalized Inverses and {2}-Inverses

Pierre Duchesne[1] and Christian Francq[2]

[1] Département de mathématiques et statistique, Université de Montréal C.P. 6128 Succursale Centre-Ville, Montréal, Québec, H3C 3J7, Canada, *duchesne@dms.umontreal.ca*

[2] Université Lille 3, EQUIPPE-GREMARS, BP 60149, 59653, Villeneuve d'Ascq, cedex, France, *christian.francq@univ-lille3.fr*

Abstract. A class of univariate time series models is considered, which allows general specifications for the conditional mean and conditional variance functions. After deriving the asymptotic distributions of the residual autocorrelations based on the standardized residuals, portmanteau test statistics are studied. If the asymptotic covariance of a vector of fixed length of residual autocorrelations is non singular, portmanteau test statistics could be defined, following the approach advocated by Li (1992). However, assuming the invertibility of the asymptotic covariance of the residual autocorrelations may be restrictive, and, alternatively, the popular Box-Pierce-Ljung test statistic may be recommended. In our framework, that test statistic converges in distribution to a weighted sum of chi-square variables, and the critical values can be found using Imhof's (1961) algorithm. However, Imhof's algorithm may be time consuming. In view of this, we investigate in this article the use of generalized inverses and {2}-inverses, in order to propose new test statistics with asymptotic chi-square distributions, avoiding the need to implement Imhof's algorithm. In a small simulation study, the following test statistics are compared: Box-Pierce-Ljung test statistic, the test statistic based on the proposal of Li (1992), and the new test statistics relying on generalized inverses and {2}-inverses.

Keywords: conditional heteroscedasticity, diagnostic checking, generalized inverses, portmanteau test statistics, residual autocorrelations

1 Introduction

Let $\{Y_t\}$ be a stationary stochastic process. We consider the following univariate time series model:

$$Y_t = m_{\boldsymbol{\theta}_0}(Y_{t-1}, Y_{t-2}, \dots) + \sigma_{\boldsymbol{\theta}_0}(Y_{t-1}, Y_{t-2}, \dots)\eta_t, \tag{1}$$

where $\boldsymbol{\theta}_0$ denotes a s dimensional vector of unknown parameters belonging to a subset Θ, where $\Theta \subset \mathbb{R}^s$. The error process $\{\eta_t\}$ is an independent and identically distributed (iid) sequence of random variables with mean zero and unit variance. It is assumed that the random variable η_t is independent of

$\{Y_{t-i}, i > 0\}$. The nonlinear model (1) represents a very general class of time series models with a general specification for the error term. It includes the classical autoregressive moving-average (ARMA) time series model, with possible [general] conditional heteroskedasticity ([G]ARCH) in the error process, and also nonlinear models, such as threshold autoregressive models (TAR), self-exciting TAR models (SETAR), and smooth versions of TAR models. Tong (1990) and Granger and Teräsvirta (1993) provide surveys of univariate nonlinear models.

Let Y_t, $t = 1, \ldots, n$ be a finite realization of the stochastic process $\{Y_t\}$. An important practical aspect is to validate an adjusted model such as (1), using estimation procedures such as quasi-maximum likelihood (QML) and nonlinear least squares (NLS) methods (the latter being obtained by assuming a constant conditional variance). Klimko and Nelson (1978) investigated general properties of conditional least squares estimators in univariate nonlinear time series. See also Potscher and Prucha (1997) and Taniguchi and Kakizawa (2000), amongst others.

Residual autocorrelations have been found useful for checking model adequacy of many time series models (see, e.g., Li (2004)). In view of this fact, we first derive the asymptotic distributions of the residual autocorrelations based on the standardized residuals. As an application of that result, portmanteau test statistics are studied. If the asymptotic covariance of a vector of fixed length of residual autocorrelations is non singular, portmanteau test statistics could be defined, following the approach advocated by Li (1992). However, assuming the invertibility of the asymptotic covariance of the residual autocorrelations may be somewhat restrictive. For example, in validating an ARMA model with an iid error term, it is well-known that the asymptotic covariance matrix of a vector of fixed length of residual autocorrelations is approximatively idempotent, with rank $n - p - q$, where p and q correspond to the autoregressive and moving average orders, respectively. On the other hand, if model (1) represents a nonlinear time series model, such as the TAR model considered in Li (1992), then, under some conditions, the asymptotic covariance matrix is expected to be non-singular. See also Li (2004, pp. 79-80). For a given model, the precise conditions which guarantee the invertibility of the asymptotic covariance matrix may be hard to obtain. Alternatively, the popular Box-Pierce-Ljung test statistic may be recommended (see Li (2004), amongst others). In our framework, this test statistic converges in distribution to a weighted sum of chi-square variables, where, in practice, the weights are determined with the data (see Francq, Roy and Zakoïan (2005) for general results in the context of ARMA models with weak errors). Interestingly, the range of applicability of Box-Pierce-Ljung test statistic appears to be more general, in the sense that if the asymptotic covariance matrix is non-singular, then all weights are strickly positive. However, contrary to the test procedure of Li (1992, 2004), the Box-Pierce-Ljung test statistic is still appropriate in linear time series models: for an AR(1) time series model, say, one weight

in the weighted sum of chi-square variables is identically equal to zero, and the others weights are strictly positives. In practice, the critical values of the Box-Pierce-Ljung test statistic can be found using Imhof's algorithm. Even today, it may be still time consuming to implement this algorithm, since, to the best of our knowledge, Imhof's algorithm is not actually available in popular softwares such as S-PLUS or R.

Since the asymptotic covariance matrix of a vector of fixed length of residual autocorrelations may be essentially singular in linear time series models, and, under certain assumptions, invertible in non-linear time series models, we investigate here the use of generalized inverses, such as the Moore-Penrose inverse, and also of $\{2\}$-inverses of that covariance matrix. This leads us to propose new portmanteau test statistics with asymptotic chi-square distributions. These new test statistics avoid the need to implement Imhof's algorithm. In a small simulation study, the following test statistics are compared with respect to level and power: Box-Pierce-Ljung test statistic, the test statistic based on the proposal of Li (1992), a new test statistic relying on the Moore-Penrose inverse, and several new proposals relying on $\{2\}$-inverses. The rest of the paper is organized as follows. In Section 2, we derive the asymptotic distribution of the residual autocorrelations. Classical portmanteau test statistics are discussed in Section 3. In Section 4, modified test statistics are presented. A small simulation study is conducted in Section 5.

2 Asymptotic distribution of the residual autocorrelations

Consider model (1). The first and second conditional moments are given by:

$$m_t(\boldsymbol{\theta}_0) := m_{\boldsymbol{\theta}_0}(Y_{t-1}, Y_{t-2}, \dots) = E(Y_t \mid Y_{t-1}, Y_{t-2}, \dots),$$
$$\sigma_t^2(\boldsymbol{\theta}_0) := \sigma_{\boldsymbol{\theta}_0}^2(Y_{t-1}, Y_{t-2}, \dots) = \mathrm{Var}(Y_t \mid Y_{t-1}, Y_{t-2}, \dots),$$

respectively. Given the time series data $Y_1, \dots, Y_n$, and the initial values $Y_0 = y_0, Y_{-1} = y_{-1}, \dots$, at any $\boldsymbol{\theta} \in \Theta$ the conditional moments $m_t(\boldsymbol{\theta})$ and $\sigma_t^2(\boldsymbol{\theta})$ can be approximated by the measurable functions defined by $\tilde{m}_t(\boldsymbol{\theta}) = m_{\boldsymbol{\theta}}(Y_{t-1}, \dots, Y_1, y_0, \dots)$ and $\tilde{\sigma}_t^2(\boldsymbol{\theta}) = \sigma_{\boldsymbol{\theta}}^2(Y_{t-1}, \dots, Y_1, y_0, \dots)$, respectively. A natural choice for the initial values is to specify $Y_i = 0$ for all $i \leq 0$. A QML estimator of $\boldsymbol{\theta}_0$ is defined as any measurable solution $\hat{\boldsymbol{\theta}}_n$ of

$$\hat{\boldsymbol{\theta}}_n = \arg \inf_{\boldsymbol{\theta} \in \Theta} \tilde{Q}_n(\boldsymbol{\theta}),$$

where $\tilde{Q}_n(\boldsymbol{\theta}) = n^{-1} \sum_{t=1}^n \tilde{\ell}_t$ and $\tilde{\ell}_t = \tilde{\ell}_t(\boldsymbol{\theta}) = (Y_t - \tilde{m}_t)^2 / \tilde{\sigma}_t^2 + \log \tilde{\sigma}_t^2$. It can be shown that the QML estimator is consistent and asymptotically normal under Assumption A.

Assumption A: *(i) Θ represents a compact set and the functions $\boldsymbol{\theta} \to \tilde{m}_t(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \to \tilde{\sigma}_t^2(\boldsymbol{\theta}) > 0$ are continuous; (ii) $\{Y_t\}$ corresponds to a non anticipative strictly stationary and ergodic solution of (1); (iii) $E \log^- \sigma_t^2(\boldsymbol{\theta}) <$*

∞ for all $\boldsymbol{\theta} \in \Theta$, and $E \log^+ \sigma_t^2(\boldsymbol{\theta}_0) < \infty$; (iv) $\sup_{\boldsymbol{\theta}\in\Theta} \left|\ell_t - \tilde{\ell}_t\right| \to 0$ a.s. as $t \to \infty$, where $\ell_t(\boldsymbol{\theta}) = (Y_t - m_t)^2/\sigma_t^2 + \log \sigma_t^2$; (v) if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ then $m_t(\boldsymbol{\theta}) \neq m_t(\boldsymbol{\theta}_0)$ or $\sigma_t^2(\boldsymbol{\theta}) \neq \sigma_t^2(\boldsymbol{\theta}_0)$ with non zero probability; (vi) $\boldsymbol{\theta}_0$ belongs to the interior $\overset{\circ}{\Theta}$ of Θ; (vii) $\boldsymbol{\theta} \to m_t(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \to \sigma_t(\boldsymbol{\theta})$ admit continuous third order derivatives, and

$$E \sup_{\boldsymbol{\theta}\in\Theta} \left| \frac{\partial^3 \ell_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_k} \right| < \infty \quad \forall i, j, k.$$

(viii) The moments $\mu_i = E\eta_t^i$, $i \leq 4$, and the information matrices $\boldsymbol{I} = E[\{\partial \ell_t(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}\}\{\partial \ell_t(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}'\}]$ and $\boldsymbol{J} = E\{\partial^2 \ell_t(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'\}$ exist. Furthermore, $\boldsymbol{I}$ and $\boldsymbol{J}$ are supposed to be non singular.

Write $a \overset{c}{=} b$ when $a = b + c$. Under Assumption A:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \overset{o_P(1)}{=} -\boldsymbol{J}^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{Z}_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n}\right) \tag{2}$$

as $n \to \infty$, where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n} := \boldsymbol{J}^{-1}\boldsymbol{I}\boldsymbol{J}^{-1}$ and

$$\boldsymbol{Z}_t = -2\frac{\eta_t}{\sigma_t}\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \left\{1 - \eta_t^2\right\} \frac{1}{\sigma_t^2} \frac{\partial \sigma_t^2(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}.$$

Following the current practice, the information matrices $\boldsymbol{I}$ and $\boldsymbol{J}$ are consistently estimated by their empirical counterparts, that is by the formula $\hat{\boldsymbol{I}} = n^{-1}\sum_{t=1}^{n}\{\partial \tilde{\ell}_t(\hat{\boldsymbol{\theta}}_n)/\partial \boldsymbol{\theta}\}\{\partial \tilde{\ell}_t(\hat{\boldsymbol{\theta}}_n)/\partial \boldsymbol{\theta}'\}$ and $\hat{\boldsymbol{J}} = n^{-1}\sum_{t=1}^{n} \partial^2 \tilde{\ell}_t(\hat{\boldsymbol{\theta}}_n)/\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'$, respectively.

Define the following standardized residuals:

$$\hat{\eta}_t = \frac{Y_t - \tilde{m}_t(\hat{\boldsymbol{\theta}}_n)}{\tilde{\sigma}_t(\hat{\boldsymbol{\theta}}_n)}, \quad t = 1, \ldots, n.$$

Portmanteau test statistics based on the autocorrelations of the residuals are routinely performed for model adequacy checking. In order to derive the asymptotic distribution of the residual autocorrelations, some additional notations are needed. Let

$$\eta_t(\boldsymbol{\theta}) = \frac{Y_t - m_t(\boldsymbol{\theta})}{\sigma_t(\boldsymbol{\theta})}, \quad \tilde{\eta}_t(\boldsymbol{\theta}) = \frac{Y_t - \tilde{m}_t(\boldsymbol{\theta})}{\tilde{\sigma}_t(\boldsymbol{\theta})},$$

so that $\eta_t = \eta_t(\boldsymbol{\theta}_0)$ and $\hat{\eta}_t = \tilde{\eta}_t(\hat{\boldsymbol{\theta}}_n)$. For any fixed integer $m \geq 1$, let

$$\boldsymbol{\gamma}_m = (\gamma(1), \ldots, \gamma(m))', \quad \boldsymbol{\rho}_m = (\rho(1), \ldots, \rho(m))',$$

where, for $\ell \geq 0$,

$$\gamma(\ell) = \frac{1}{n}\sum_{t=1}^{n-\ell} \eta_t \eta_{t+\ell} \quad \text{and} \quad \rho(\ell) = \frac{\gamma(\ell)}{\gamma(0)}.$$

Note that $\boldsymbol{\gamma}_m \overset{o_P(1)}{=} n^{-1}\sum_{t=1}^{n}\boldsymbol{\Upsilon}_t\boldsymbol{\Upsilon}_t'$ where $\boldsymbol{\Upsilon}_t = \eta_t\boldsymbol{\eta}_{t-1:t-m}$ and $\boldsymbol{\eta}_{t-1:t-m} = (\eta_{t-1},\ldots,\eta_{t-m})'$. In view of (2), the central limit theorem applied to the martingale difference $\left\{(\boldsymbol{Z}_t',\boldsymbol{\Upsilon}_t')';\sigma(\eta_u,\, u\le t)\right\}$ implies the following asymptotic distribution:

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\\ \boldsymbol{\gamma}_m\end{pmatrix}\xrightarrow{\mathcal{L}}\mathcal{N}\left\{\mathbf{0},\begin{pmatrix}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n} & \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n\boldsymbol{\Upsilon}_m}\\ \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n\boldsymbol{\Upsilon}_m}' & I_m\end{pmatrix}\right\}, \tag{3}$$

where I_m denotes the identity matrix of order m and:

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n\boldsymbol{\Upsilon}_m} &= -\boldsymbol{J}^{-1}E\boldsymbol{Z}_t\boldsymbol{\Upsilon}_t',\\ &= 2\boldsymbol{J}^{-1}E\frac{1}{\sigma_t}\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}_{t-1:t-m}' + \boldsymbol{J}^{-1}\mu_3E\frac{1}{\sigma_t^2}\frac{\partial\sigma_t^2(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\boldsymbol{\eta}_{t-1:t-m}'.\end{aligned}$$

We now turn to the residual autocorrelation function $\hat{\rho}(\cdot)$ obtained by replacing η_t by $\hat{\eta}_t$ in $\rho(\cdot)$. Similarly, define the function $\hat{\gamma}(\cdot)$ and the vectors $\hat{\boldsymbol{\gamma}}_m$ and $\hat{\boldsymbol{\rho}}_m$. A Taylor expansion of the function $\boldsymbol{\theta}\mapsto n^{-1}\sum_{t=1}^{n-\ell}\eta_t(\boldsymbol{\theta})\eta_{t+\ell}(\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$ gives $\hat{\gamma}(\ell)\overset{o_P(1)}{=}\gamma(\ell)+\boldsymbol{c}_\ell'(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$, where

$$\begin{aligned}\boldsymbol{c}_\ell &= E\eta_{t-\ell}\frac{\partial\eta_t}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0) = -E\frac{\eta_{t-\ell}}{\sigma_t(\boldsymbol{\theta}_0)}\left\{\frac{\partial m_t}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0)+\frac{\eta_t}{2\sigma_t(\boldsymbol{\theta}_0)}\frac{\partial\sigma_t^2}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right\}\\ &= -E\frac{\eta_{t-\ell}}{\sigma_t(\boldsymbol{\theta}_0)}\frac{\partial m_t}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0).\end{aligned}$$

Using (3) and the notation $\boldsymbol{C}_m = (\boldsymbol{c}_1\ \boldsymbol{c}_2\ \cdots\ \boldsymbol{c}_m)$, it follows that:

$$\sqrt{n}\hat{\boldsymbol{\gamma}}_m\xrightarrow{\mathcal{L}}\mathcal{N}\left\{\mathbf{0},\boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}_m}\right\}$$

as $n\to\infty$, where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}_m} = I_m + \boldsymbol{C}_m'\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n}\boldsymbol{C}_m + \boldsymbol{C}_m'\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n\boldsymbol{\Upsilon}_m} + \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n\boldsymbol{\Upsilon}_m}'\boldsymbol{C}_m$. Some simplifications are possible. First, we note that $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n\boldsymbol{\Upsilon}_m} = -2\boldsymbol{J}^{-1}\boldsymbol{C}_m + \mu_3\boldsymbol{J}^{-1}\boldsymbol{D}_m$, where $\boldsymbol{D}_m = (\boldsymbol{d}_1\ \boldsymbol{d}_2\ \cdots\ \boldsymbol{d}_m)$ and $\boldsymbol{d}_\ell = E\eta_{t-\ell}\sigma_t^{-2}\partial\sigma_t^2(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$. Thus, the asymptotic covariance matrix can be written as:

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\gamma}_m} = I_m &+ \boldsymbol{C}_m'\boldsymbol{J}^{-1}\boldsymbol{I}\boldsymbol{J}^{-1}\boldsymbol{C}_m - 4\boldsymbol{C}_m'\boldsymbol{J}^{-1}\boldsymbol{C}_m\\ &+\mu_3\left(\boldsymbol{C}_m'\boldsymbol{J}^{-1}\boldsymbol{D}_m + \boldsymbol{D}_m'\boldsymbol{J}^{-1}\boldsymbol{C}_m\right).\end{aligned} \tag{4}$$

Since $\hat{\gamma}(0) = 1 + o_P(1)$, the asymptotic distribution of the residual autocorrelations follows easily:

$$\sqrt{n}\hat{\boldsymbol{\rho}}_m\xrightarrow{\mathcal{L}}\mathcal{N}\left\{\mathbf{0},\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}\right\}, \tag{5}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}_m}$. One can define empirical estimates $\hat{\boldsymbol{C}}_m$ and $\hat{\boldsymbol{D}}_m$ by replacing $\boldsymbol{c}_\ell$ and $\boldsymbol{d}_\ell$ in $\boldsymbol{C}_m$ and $\boldsymbol{D}_m$ by

$$\hat{\boldsymbol{c}}_\ell = -\frac{1}{n}\sum_{t=\ell+1}^{n}\frac{\hat{\eta}_{t-\ell}}{\sigma_t(\hat{\boldsymbol{\theta}}_n)}\frac{\partial m_t}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n)\quad\text{and}\quad\hat{\boldsymbol{d}}_\ell = \frac{1}{n}\sum_{t=\ell+1}^{n}\frac{\hat{\eta}_{t-\ell}}{\sigma_t^2(\hat{\boldsymbol{\theta}}_n)}\frac{\partial\sigma_t^2}{\partial\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n).$$

We then obtain an estimator $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}$ of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ by replacing μ_3 and the matrices $\boldsymbol{I}$, $\boldsymbol{J}$, $\boldsymbol{C}_m$ and $\boldsymbol{D}_m$ by their empirical counterparts in (4).

3 Classical portmanteau tests

3.1 Box-Pierce-Ljung test statistics

For checking the adequacy of an ARMA(p,q) model, it is customary to employ the so-called portmanteau tests, such as the Box-Pierce-Ljung test statistic $Q_m^{BPL} = n(n+2)\sum_{i=1}^{m} \hat{\rho}^2(i)/(n-i)$. When diagnosing ARMA models, the null hypothesis of an ARMA(p,q) model is rejected at the nominal level α when $Q_m^{BPL} > \chi^2_{m-(p+q)}(1-\alpha)$, where $m > p+q$ and $\chi^2_\ell(1-\alpha)$ denotes the $(1-\alpha)$-quantile of a χ^2 distribution with ℓ degrees of freedom.

More generally, when the conditional mean $m_{\boldsymbol{\theta}_0}(\cdot)$ and the conditional variance $\sigma_{\boldsymbol{\theta}_0}(\cdot)$ are well specified in (1), the residual autocorrelations $\hat{\rho}(h)$ are expected to be close to zero for all $h \neq 0$. Therefore, it is natural to reject the null hypothesis H_0 that the data generating process (DGP) is the model (1) when $\|\sqrt{n}\hat{\boldsymbol{\rho}}_m\|^2$ is larger than a certain critical value. More precisely, (5) shows that, under the null hypothesis H_0 of model adequacy:

$$Q_m^{BPL} \xrightarrow{\mathcal{L}} \sum_{i=1}^{m} \lambda_i \mathrm{Z}_i^2 \quad \text{as } n \to \infty, \tag{6}$$

where $\mathrm{Z}_1, \ldots, \mathrm{Z}_m$ correspond to independent $\mathcal{N}(0,1)$ random variables and $\lambda_1, \ldots, \lambda_m$ represent the eigenvalues of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$. For an ARMA$(p,q)$ model with iid errors, it is shown in McLeod (1978) that the $p+q$ smallest eigenvalues λ_i are close to zero and that the other eigenvalues are equal to one. Thus, we obtain a χ^2_{m-s} approximation, where $s = p+q$ is the number of estimated parameters, for the asymptotic distribution of Q_m^{BPL} when the DGP is an ARMA (p,q) model with iid errors.

When the errors are uncorrelated but not independent, and when the ARMA coefficients are estimated by least squares, it is shown in Francq, Roy and Zakoïan (2005) that the asymptotic distribution of Q_m^{BPL} is poorly approximated by the chi-square distribution χ^2_{m-s}.

In this paper, the framework is different from the one considered in Francq, Roy and Zakoïan (2005): here, a more general model is permitted than the classical ARMA model. However, in the present set-up, the error process $\{\eta_t\}$ is assumed to be iid and the error term in (1) represents a martingale difference sequence.

It is clear that all the eigenvalues of the matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ are positive and that, when $\mu_3 = 0$ and $m > s$, at least $m-s$ of its eigenvalues are equal to one. When $\mu_3 = 0$ and $m > s$, we then have $\lim_{n\to\infty} P(Q_m^{BPL} > x) \geq P(\chi^2_{m-s} > x)$. Consequently, the test statistic defined by the critical region $\{Q_m^{BPL} > \chi^2_{m-s}(1-\alpha)\}$ is expected to be liberal at the nominal level α. In the sequel, this test statistic will be referred to as the χ^2_{m-s}-based ($\mathrm{BPL}_{\chi^2_{m-s}}$) Box-Pierce-Ljung portmanteau test statistic.

It is possible to evaluate the distribution of the Gaussian quadratic form in (6) by means of Imhof's algorithm. Following Francq, Roy and Zakoïan

(2005), one can thus propose a modified portmanteau test statistic based on the following steps: 1) compute the eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$ of a consistent estimator $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}$ of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$, 2) evaluate the $(1-\alpha)$-quantile $c_\alpha(\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$ of $\sum_{i=1}^m \hat{\lambda}_i \mathrm{Z}_i^2$ using Imhof's algorithm, 3) reject the null that the DGP is (1) when $n\hat{\boldsymbol{\rho}}_m' \hat{\boldsymbol{\rho}}_m \geq c_\alpha(\hat{\lambda}_1, \ldots, \hat{\lambda}_m)$. For further reference, this test statistic will be referred to as the Imhof-based (BPL_{Imhof}) Box-Pierce-Ljung portmanteau test statistic. Compared to the $\mathrm{BPL}_{\chi^2_{m-s}}$ method, the BPL_{Imhof} version is asymptotically more accurate (since the χ^2_{m-s} distribution is only a crude approximation of the true asymptotic distribution), but the BPL_{Imhof} test statistic is relatively more involved to implement since an estimator of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ is required, and Imhof's algorithm must be implemented, which is relatively complicated and may be time consuming. Section 4 below proposes alternatives to $\mathrm{BPL}_{\chi^2_{m-s}}$ and BPL_{Imhof} portmanteau test statistics.

3.2 An example

The autoregressive (AR) and the autoregressive conditional heteroscedastic (ARCH) models are among the most widely used models for the conditional mean and conditional variance. We combine the simplest versions of these two models to obtain the AR(1)-ARCH(1) model:

$$\begin{cases} Y_t = a_0 Y_{t-1} + \epsilon_t, \\ \epsilon_t = \sigma_t \eta_t, \quad \sigma_t^2 = \omega_0 + \alpha_0 \epsilon_{t-1}^2. \end{cases} \tag{7}$$

Under very general assumptions, Assumption **A** holds true (see Francq and Zakoïan (2004), who discuss Assumption **A** in the framework of ARMA-GARCH models). The unknown parameter is $\boldsymbol{\theta}_0 = (a_0, \omega_0, \alpha_0)$. In order to be able to compute explicitly the information matrices $\boldsymbol{I}$ and $\boldsymbol{J}$, we assume $\alpha_0 = 0$. By (4.12), (4.13) and (4.40)–(4.42) in Francq and Zakoïan (2004):

$$\frac{\partial \ell_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = -2\frac{\eta_t}{\sqrt{\omega_0}} \begin{pmatrix} Y_{t-1} \\ 0 \\ 0 \end{pmatrix} + (1-\eta_t^2)\frac{1}{\omega_0} \begin{pmatrix} 0 \\ 1 \\ \epsilon_{t-1}^2 \end{pmatrix}$$

and

$$\frac{\partial^2 \ell_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{pmatrix} \frac{2Y_{t-1}^2}{\omega_0} & \frac{2\eta_t Y_{t-1}}{\omega_0^{3/2}} & \frac{2\eta_t Y_{t-1}\epsilon_{t-1}^2}{\omega_0^{3/2}} + \frac{2(\eta_t^2-1)\epsilon_{t-1}Y_{t-2}}{\omega_0} \\ . & \frac{2\eta_t^2-1}{\omega_0^2} & \frac{(2\eta_t^2-1)\epsilon_{t-1}^2}{\omega_0^2} \\ . & . & \frac{(2\eta_t^2-1)\epsilon_{t-1}^4}{\omega_0^2} \end{pmatrix}.$$

Thus we have:

$$\boldsymbol{I} = \mathrm{Var}\,\frac{\partial \ell_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{4}{1-a_0^2} & 0 & 2\mu_3^2 \\ 0 & \frac{\mu_4-1}{\omega_0^2} & \frac{\mu_4-1}{\omega_0} \\ 2\mu_3^2 & \frac{\mu_4-1}{\omega_0} & \mu_4(\mu_4-1) \end{pmatrix},$$

$$\boldsymbol{J} = E\frac{\partial^2 \ell_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \begin{pmatrix} \frac{2}{1-a_0^2} & 0 & 0 \\ 0 & \frac{1}{\omega_0^2} & \frac{1}{\omega_0} \\ 0 & \frac{1}{\omega_0} & \mu_4 \end{pmatrix}$$

and

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n} = \begin{pmatrix} 1-a_0^2 & -\frac{\omega_0\mu_3^2(1-a_0^2)}{\mu_4-1} & \frac{\mu_3^2(1-a_0^2)}{\mu_4-1} \\ -\frac{\omega_0\mu_3^2(1-a_0^2)}{\mu_4-1} & \omega_0^2\mu_4 & -\omega_0 \\ \frac{\mu_3^2(1-a_0^2)}{\mu_4-1} & -\omega_0 & 1 \end{pmatrix}.$$

Note that, when $\eta_t \sim \mathcal{N}(0,1)$, we have $\boldsymbol{I} = 2\boldsymbol{J}$ and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_n} = 2\boldsymbol{J}^{-1}$. We also have

$$\boldsymbol{C}_m = -\begin{pmatrix} 1 & a_0 & \dots & a_0^{m-1} \\ & & 0_{2\times m} & \end{pmatrix}, \quad \boldsymbol{D}_m = \begin{pmatrix} & 0_{2\times m} & \\ \mu_3 & & 0'_{m-1} \end{pmatrix}.$$

Note that $\boldsymbol{D}'_m\boldsymbol{J}^{-1}\boldsymbol{C}_m = \boldsymbol{0}$, $\boldsymbol{J}^{-1}\boldsymbol{C}_m = \boldsymbol{C}_m(1-a_0^2)/2$ and $\boldsymbol{I}\,\boldsymbol{C}_m = \boldsymbol{C}_m 4/(1-a_0^2) + 2\mu_3^3\boldsymbol{C}^*_m$, where $\boldsymbol{C}^*_m$ is obtained by permuting the rows 1 and 3 of $\boldsymbol{C}_m$, so that $\boldsymbol{C}'_m\boldsymbol{C}^*_m = 0$. It follows that

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}_m} = I_m - (1-a_0^2)\boldsymbol{C}'_m\boldsymbol{C}_m.$$

When m is large or a_0 is close to 0, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}_m} \simeq I_m - \boldsymbol{C}'_m\left(\boldsymbol{C}_m\boldsymbol{C}'_m\right)^{-1}\boldsymbol{C}_m$ is close to a projection matrix with $m-1$ eigenvalues equal to 1, and one eigenvalue equals to 0. Therefore, in this particular situation where $\alpha_0 = 0$, the asymptotic distribution of the Box-Pierce-Ljung test statistics can be approximated by a χ^2_{m-1} distribution. Note that this is not the approximation usually employed in the ARMA case, namely the χ^2_{m-s} where s is the number of estimated parameters.

3.3 Test statistic based on a proposal of Li (1992)

Assume $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ to be non-singular. A natural approach considered by Li (1992, 2004) in non-linear time series with independent errors consists to define the following test statistic:

$$Q_m^{INV} = n\hat{\boldsymbol{\rho}}'_m\hat{\boldsymbol{\Sigma}}^{-1}_{\hat{\boldsymbol{\rho}}_m}\hat{\boldsymbol{\rho}}_m, \tag{8}$$

which follows asymptotically a χ^2_m distribution under the null hypothesis H_0 that the DGP satisfies (1), provided $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}$ corresponds to a consistent estimator of the nonsingular matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$.

However, as suggested by the example of the preceding section, the matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ is not invertible in the ARMA case and conditions which guaranty the invertibility of that asymptotic matrix seem difficult to find. If $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ is singular, this invalidates the asymptotic χ^2_m distribution. In practice, numerical instability is expected in the computation of Q_m^{INV} when $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ is singular.

In the next section, we investigate the use of several generalized inverses of the matrix $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}$. Basic results on generalized inverses are reviewed in the next section.

4 Modified portmanteau tests using generalized inverses and {2}-inverses

4.1 Generalized inverses and {2}-inverses

A generalized inverse (g-inverse) of a matrix $\boldsymbol{\Sigma}$ is a matrix $\tilde{\boldsymbol{\Sigma}}$ satisfying $\boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$. Usually this condition is the first of the four conditions defining the (unique) Moore-Penrose inverse of $\boldsymbol{\Sigma}$, and $\tilde{\boldsymbol{\Sigma}}$ is called a {1}-inverse (Getson and Hsuan (1988)). On the other hand, a {2}-inverse of $\boldsymbol{\Sigma}$ is any matrix $\boldsymbol{\Sigma}^*$ satisfying the second relation defining the Moore-Penrose inverse of $\boldsymbol{\Sigma}$, that is $\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}^*$. When both requirements are satisfied, the resulting matrix is sometimes called a reflexive g-inverse or a $\{1,2\}$-inverse (Rao (1973, p. 25)). Let $\boldsymbol{\Sigma} \neq \mathbf{0}$ be a positive semidefinite symmetric matrix of order m, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. The spectral decomposition of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}' = \sum_{i=1}^{m} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i'$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ and the columns $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ of the matrix $\boldsymbol{P}$ constitute an orthonormal basis of $\mathbb{R}^m$. If $\lambda_{m-s} > 0$ and $\lambda_{m-s+1} = \cdots = \lambda_m = 0$, then the matrix $\boldsymbol{\Sigma}^- = \boldsymbol{P}\boldsymbol{\Lambda}^-\boldsymbol{P}'$ where $\boldsymbol{\Lambda}^- = \mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_{m-s}^{-1}, \mathbf{0}_s')$ is the Moore-Penrose inverse (or pseudo-inverse) of $\boldsymbol{\Sigma}$. For $k = 1, \ldots, m-s$, let the matrix $\boldsymbol{\Sigma}^{-k} = \boldsymbol{P}\boldsymbol{\Lambda}^{-k}\boldsymbol{P}'$ where $\boldsymbol{\Lambda}^{-k} = \mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_k^{-1}, \mathbf{0}_{m-k}')$. The matrix $\boldsymbol{\Lambda}^{-k}$ is always a {2}-inverse, but this is not a g-inverse of $\boldsymbol{\Sigma}$ when $k < m - s$. Now suppose that $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}_m, \boldsymbol{\Sigma})$. Then, using natural notations, we have $\boldsymbol{\Lambda}^{-k\,1/2}\boldsymbol{P}'\boldsymbol{Z} \sim \mathcal{N}\left\{\mathbf{0}_m, \mathrm{diag}(\mathbf{1}_k', \mathbf{0}_{m-k}')\right\}$ and thus $\boldsymbol{Z}'\boldsymbol{\Sigma}^{-k}\boldsymbol{Z} = \|\boldsymbol{\Lambda}^{-k\,1/2}\boldsymbol{P}'\boldsymbol{Z}\|^2 \sim \chi_k^2$. Now suppose that $\boldsymbol{Z}_n \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}_m, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_n \to \boldsymbol{\Sigma}$ almost surely, as $n \to \infty$. For $k = 1, \ldots, \mathrm{rank}(\boldsymbol{\Sigma})$, the matrix $\boldsymbol{\Sigma}^{-k}$ exists and can be approximated by $\boldsymbol{\Sigma}_n^{-k}$, for all large enough n (note, however, that the matrices $\boldsymbol{\Sigma}^{-k}$ and $\boldsymbol{\Sigma}_n^{-k}$ are not unique, because they depend on the particular choice of the orthonormal basis in the decompositions $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$ and $\boldsymbol{\Sigma}_n = \boldsymbol{P}_n\boldsymbol{\Lambda}_n\boldsymbol{P}_n'$). Using the continuity property of the eigenvalues and eigenprojections (see Tyler (1981)), it can be shown under mild regularity assumptions that:

$$\boldsymbol{Z}_n'\boldsymbol{\Sigma}_n^{-k}\boldsymbol{Z}_n \xrightarrow{\mathcal{L}} \chi_k^2, \qquad \forall k \leq \mathrm{rank}(\boldsymbol{\Sigma}). \tag{9}$$

The condition $k \leq \mathrm{rank}(\boldsymbol{\Sigma})$ appears to be essential. For example, consider: $\boldsymbol{\Sigma}_n = \boldsymbol{P}_n\boldsymbol{\Lambda}_n\boldsymbol{P}_n'$, $\boldsymbol{\Lambda}_n = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{n} \end{pmatrix}$, $\boldsymbol{P}_n = \begin{pmatrix} \sqrt{\frac{1}{n}} & -\sqrt{\frac{n-1}{n}} \\ \sqrt{\frac{n-1}{n}} & \sqrt{\frac{1}{n}} \end{pmatrix}$, and the bivariate normal distribution $\boldsymbol{Z}_n \sim \mathcal{N}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{c}{n} & 0 \\ 0 & 1 \end{pmatrix}\right\}$, $c \geq 0$. Then $\boldsymbol{Z}_n \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}_2, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_n \to \boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma} = \mathrm{diag}(0, 1)$, but $\boldsymbol{Z}_n'\boldsymbol{\Sigma}_n^{-2}\boldsymbol{Z}_n \sim (c + \frac{c}{n^2} - \frac{c}{n})Z_1^2 + (2 - \frac{1}{n})Z_2^2 \not\xrightarrow{\mathcal{L}} \chi_2^2$. In the next subsection, test statistics will be constructed, relying on an appropriate estimator of the Moore-Penrose inverse, and on estimators of the {2}-inverses considered in this section.

4.2 Generalized portmanteau test statistics

Consider a consistent estimator $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}$ of the matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$. Since $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ in (5) may be singular, one can propose:

$$Q_m^{MP} = n\hat{\boldsymbol{\rho}}_m' \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}^{-} \hat{\boldsymbol{\rho}}_m, \tag{10}$$

where $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}^{-}$ is an estimator of the Moore-Penrose inverse of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$. At the nominal level α, the null hypothesis H_0 that the DGP follows a nonlinear model of the form (1) is rejected when $Q_m^{MP} > \chi^2_{k_n}(1-\alpha)$, with k_n the number of eigenvalues of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}$ larger that a certain tolerance ϵ (e.g, $\epsilon =$ `sqrt(.Machine$double.eps)` $= 1.49 \times 10^{-8}$ with the `R` software). In view of (5) and (9), test statistics relying on estimators $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}^{-k}$ of the $\{2\}$-inverses $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}^{-k}$ introduced in Section 4.1 can be proposed. For $k \in \{1, \ldots, m\}$ fixed, they are defined by:

$$Q_m^{-k} = n\hat{\boldsymbol{\rho}}_m' \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\rho}}_m}^{-k} \hat{\boldsymbol{\rho}}_m. \tag{11}$$

At the nominal level α, the null hypothesis H_0 is rejected when $Q_m^{-k} > \chi^2_k(1-\alpha)$. The test statistics (10) and (11) constitute interesting alternatives to BPL_{Imhof}; they do not require the use of Imhof's algorithm. However, the range of application of the test statistics relying on $\{2\}$-inverses is more limited because the test statistic Q_m^{-k} presumes the assumption that $\text{rank}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}) \geq k$. The range of application of the Q_m^{-k}-test decreases as k increases from one to m. The test obtained with $k = m$, which is actually that based on $Q_m^{-m} = Q_m^{INV}$, that is the test statistic proposed by Li (1992), is the most restrictive one, in the sense that the invertibility of $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\rho}}_m}$ is required. On the other hand, the set of the alternatives for which the Q_m^{-k}-test is consistent should increase with k: under appropriate regularity conditions $n^{-1}Q_m^{-1} \to \lambda_1^{-1} \left(\boldsymbol{\rho}_m' \boldsymbol{v}_1\right)^2$ with probability one as $n \to \infty$. Thus, the Q_m^{-1}-test should not have much power against alternatives such that $\boldsymbol{\rho}_m' \boldsymbol{v}_1 = 0$. The next section provides an empirical comparison of the different test statistics.

5 Numerical illustrations

Here, we compare empirically the following portmanteau tests: BPL_{Imhof} and the liberal test statistic $\text{BPL}_{\chi^2_{m-s}}$ described in Section 3.1, $\text{BPL}_{\chi^2_{m-1}}$ advocated in Section 3.2, and the test statistics Q_m^{MP} (with $\epsilon = 1.49 \times 10^{-8}$) and Q_m^{-k} introduced in Section 4.2; $Q_m^{INV} = Q_m^{-m}$ of Section 3.3 is included in our experiments. We concentrate on the case $m = 4$, which leads to comparing eight tests. In a first set of Monte Carlo experiments, $N = 1000$ independent trajectories of the AR(1)-ARCH(1) model (7) are simulated. The lengths of the trajectories are $n = 200, 2000$. The code is written in `R` and `FORTRAN`.

Table 1 displays the empirical sizes. For the nominal level $\alpha = 5\%$, the empirical size over the $N = 1000$ independent replications should belong to the interval [3.6%, 6.4%] with probability 95%. When the relative rejection

Table 1. Empirical size of portmanteau tests: relative frequencies (in %) of rejection of the AR(1)-ARCH(1) model (7), when the DGP follows the same model. The number of replications is $N = 1000$.

Model	n	α	$\text{BPL}_{\chi^2_{m-1}}$	$\text{BPL}_{\chi^2_{m-s}}$	Q_4^{-1}	Q_4^{-2}	Q_4^{-3}	Q_4^{-4}	Q_4^{MP}	BPL_{Imhof}
		1%	1.0	**7.3**	0.5	0.6	1.0	1.0	1.2	1.0
I	200	5%	4.2	**24.9**	4.1	**2.7**	4.0	**2.9**	3.7	3.7
		10%	8.3	**40.1**	8.3	8.2	**8.0**	**4.8**	**6.8**	**7.9**
		1%	1.0	**9.1**	1.2	1.0	1.0	1.3	1.2	1.0
I	2000	5%	5.9	**26.9**	5.3	4.8	5.9	3.6	6.3	5.9
		10%	10.9	**44.1**	9.1	9.6	10.9	**7.2**	9.9	10.7
		1%	1.2	**9.6**	0.6	0.6	0.5	0.9	0.9	0.8
II	200	5%	5.1	**31.3**	4.3	**3.4**	**3.1**	**3.3**	**3.3**	**3.3**
		10%	11.7	**48.7**	8.7	**7.8**	**6.9**	**6.8**	**6.8**	**6.2**
		1%	1.6	**16.0**	0.8	1.2	0.9	0.8	0.8	0.8
II	2000	5%	**9.8**	**38.9**	5.2	5.0	5.1	5.0	5.0	4.5
		10%	**18.0**	**58.9**	10.7	10.5	10.4	**12.1**	**12.1**	**12.0**

I: $a_0 = 0$, $\omega_0 = 1$ and $\alpha_0 = 0$ II: $a_0 = 0.95$, $\omega_0 = 1$ and $\alpha_0 = 0.55$

Table 2. Empirical power of portmanteau tests: relative frequencies (in %) of rejection of the AR(1)-ARCH(1) model (7), when the DGP follows an AR(3)-ARCH(1) model (model III) or an AR(1)-ARCH(3) model (model IV).

Model	n	α	Q_4^{-1}	Q_4^{-2}	Q_4^{-3}	Q_4^{-4}	Q_4^{MP}	BPL_{Imhof}
		1%	20.0	31.2	**38.4**	29.2	**32.2**	**38.3**
III	200	5%	36.1	53.2	**65.3**	52.6	**57.0**	**65.8**
		10%	46.3	64.2	**75.4**	63.4	**67.7**	**75.3**
		1%	35.6	61.0	**80.8**	71.4	**73.9**	**81.2**
III	400	5%	49.5	73.4	**92.0**	84.6	**86.0**	**91.9**
		10%	58.7	79.3	**95.4**	88.7	**89.2**	**95.2**
		1%	1.5	3.5	**4.7**	3.4	**4.3**	**4.4**
IV	200	5%	6.1	8.8	**10.0**	8.4	**9.5**	**8.8**
		10%	11.1	14.1	**16.5**	11.3	**14.9**	**14.9**
		1%	2.7	5.8	**7.6**	6.4	**7.7**	**7.4**
IV	400	5%	8.3	13.5	**15.3**	11.3	**13.8**	**14.5**
		10%	13.9	**20.9**	**22.7**	16.2	19.9	**21.4**

III: $Y_t = 0.2Y_{t-3} + \epsilon_t$ where $\epsilon_t^2 = \sqrt{1 + 0.2\epsilon_{t-1}^2}\eta_t$

IV: $Y_t = 0.2Y_{t-1} + \epsilon_t$ where $\epsilon_t^2 = \sqrt{1 + 0.5\epsilon_{t-3}^2}\eta_t$

frequencies are outside the 95% significance limits, they are displayed in bold in Table 1. When the relative rejection frequencies are outside the 99% significance limits [3.2%, 6.9%], they are underlined. It can be seen that: 1) the rejection frequency of $\text{BPL}_{\chi^2_{m-s}}$ is definitely too high, 2) as expected, $\text{BPL}_{\chi^2_{m-1}}$

works well when $a_0 = 0$ and $\alpha_0 = 0$, but not when $a_0 \neq 0$ or $\alpha_0 \neq 0$, 3) the empirical levels of $Q_4^{-4} = Q_4^{INV}$ are far from the nominal levels for Model I, which is explained by the singularity of $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m}$, 4) the errors of the first kind of the test statistics Q_4^{MP}, Q_4^{-k}, $k < 4$, and BPL_{Imhof} are well controlled when n is large. Table 2 compares the empirical powers, excluding $\text{BPL}_{\chi^2_{m-s}}$ and $\text{BPL}_{\chi^2_{m-1}}$, which display unsatisfactory empirical levels. The three highest powers are displayed in bold, and the highest one is underlined. Note that misspecification of the conditional mean (model III) seems easier to detect than misspecification of the conditional variance (model IV). As expected, the power of Q_m^{-k} is function of k. From Table 2, Q_4^{-3} and BPL_{Imhof} are the most powerful portmanteau test statistics, at least in our experiments. Interestingly, Q_4^{MP} offers an empirical power very close to the one of BPL_{Imhof}, and slightly better than the one of Q_4^{-4}. In general, BPL_{Imhof} seems to be advisable in view of its good theoretical and finite sample performance, but given its computational simplicity, Q_4^{MP} appears to be a close competitor.

References

FRANCQ, C., ROY, R. and ZAKOÏAN, J.-M. (2005): Diagnostic checking in ARMA models with uncorrelated errors. *Journal of the American Statistical Association 100, 532-544.*

FRANCQ, C. and ZAKOÏAN, J.-M. (2004): Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli 10, 605-637.*

GETSON, A.J. and HSUAN, F.C. (1988): *{ 2 }-Inverses and Their Statistical Application.* Lecture Notes in Statistics 47, Springer-Verlag, New York.

GRANGER, C.W.J. and TERÄSVIRTA, T. (1993): *Modelling Nonlinear Economic Relationships.* Oxford University Press, Oxford.

IMHOF, J.P. (1961): Computing the distribution of quadratic forms in normal variables. *Biometrika 48, 419-426.*

KLIMKO, L.A. and NELSON, P.I. (1978): On conditional least squares estimation for stochastic processes. *The Annals of Statistics 6, 629-642.*

LI, W.K. (2004): *Diagnostic Checks in Time Series.* Chapman & Hall/CRC, New York.

LI, W.K. (1992): On the asymptotic standard errors of residual autocorrelations in nonlinear time series modelling. *Biometrika 79, 435-437.*

MCLEOD, A.I. (1978): On the distribution of residual autocorrelations in Box-Jenkins method. *Journal of the Royal Statistical Society B 40, 296-302.*

POTSCHER, B.M. and PRUCHA, I.R. (1997): *Dynamic Nonlinear Econometric Models.* Springer, Berlin.

RAO, C.R. (1973): *Linear Statistical Inference and Its Applications.* Wiley, New York.

TANIGUCHI, M. and KAKIZAWA, Y. (2000): *Asymptotic Theory of Statistical Inference for Time Series.* Springer, New York.

TONG, H. (1990): *Non-linear Time Series: A Dynamical System Approach.* Oxford University Press, Oxford.

TYLER, D.E. (1981): Asymptotic inference for eigenvectors. *The Annals of Statistics 9, 725-736.*

New Developments in Latent Variable Models: Non-Linear and Dynamic Models

Irini Moustaki[1,2]

[1] London School of Economics and Political Science
Department of Statistics, Houghton Street, London WC2A 2AE, U.K.
[2] Athens University of Economics and Business
Department of Statistics, 76 Patission street, Athens 104 34, Greece
i.moustaki@lse.ac.uk

Abstract. The paper reviews recent work on latent variable models for ordinal longitudinal variables and factor models with non-linear terms. The model for longitudinal data has been recently proposed by Cagnone, Moustaki and Vasdekis (2008). The model allows for time-dependent latent variables to explain the associations among ordinal variables within time where the associations among the same items across time are modelled with item-specific random effects. Rizopoulos and Moustaki (2007) extended the generalized latent variable model framework to allow for non-linear terms (interactions and higher order terms). Both models are estimated with full information maximum likelihood. Computational aspects, goodness-of-fit statistics and an application are presented.

Keywords: latent variable models, ordinal data, longitudinal data, non-linear terms.

1 Introduction

Latent variable models aim to explain the interrelationships (covariance structure) among a set of observed variables using a small number of latent variables. They are widely used in social sciences for measuring unobserved constructs such as ability, wealth, conservatism, political efficacy etc. Models and methods have been developed to account for needs in Psychology (Psychometrics), Sociology, Biostatistics, Educational Testing, Economics etc. Latent variable models have different objectives depending on the area of application. Some of the objectives can be summarized as follows:

- Reduce the original dimensionality of the data.
- Score subjects on the identified latent dimensions.
- Construction of scales.
- Test of a specific hypothesis (confirmatory factor analysis).

Latent variables can be either continuous, discrete or mixed leading to a factor analysis type model, a latent class model or a hybrid model respectively. Similarly manifest variables can be either continuous, categorical (nominal/ ordinal) or mixed (see Bartholomew and Knott, 1999).

For the treatment of categorical responses, different estimation methods have been investigated in the literature. We mainly divide them into full information maximum likelihood (FIML) and limited information methods (LI). Classical as well Bayesian estimation methods have been used for FIML. FIML use all the information in the data where LI methods use the information from the lower order margins (univariate and bivariate margins). FIML methods are known to be more computationally intensive but generally more efficient than LI methods.

In this paper, we will review factor models for ordinal longitudinal data as they have been developed in Cagnone et. al. (2008) and factor models that allow for non-linear terms (Rizopoulos and Moustaki, 2007) using FIML.

Longitudinal data are collected for mainly studying changes across time. Factor models for longitudinal data are known as growth curve models (see e.g. Bollen and Curran, 2006), transition models as well as dynamic factor analysis models. Here we present a latent variable model for longitudinal ordinal items with non-linear terms such as interaction terms between the latent variables as well as higher order terms. In multivariate longitudinal data, the model needs to account for correlations among items within time as well as correlations among the same items between time. If one considers a continuous observed variable then the total variance can be decomposed into common variance (variance explained by the latent variables common to all items at a specific time point), the item-specific variance and the error variance (see Raffalovich, and Bohrnstedt, 1987, Marsh and Grayson, 1994 and Eid, 1996). This decomposition of total variance can only be achieved when longitudinal data are analyzed. In structural equation modelling (SEM) literature (see e.g. Jöreskog and Sörbom, 1999) item-specific variance is taken into account by considering correlated measurement errors for the same items across time or item-specific factors (see e.g. multitrait-multimethod models). Jöreskog (2002) discusses a SEM for longitudinal ordinal data where differences in means and covariances of latent variables over time are measured by assuming measurement invariance of thresholds and factor loadings and by correlating the measurement errors of same items over time.

Usually the effect of latent variables or covariates is additive. However, the inclusion of interaction or quadratic terms might be necessary. Applications of latent variable models with nonlinear terms in different disciplines such as marketing, social psychology, political theory, etc. can be found in Schumacker and Marcoulides (1998) and references therein. The idea of nonlinear factor analysis goes back to Gibson (1960) and Mcdonald (1962). It has been pointed out by Bartlett (1953) that the inclusion of the interaction between two latent variables in the linear factor analysis model will produce the same correlation properties of the manifest variables as if a third genuine factor had been included. That opens some discussion on the identifiability of latent variable models with non-linear terms. More references on the technical developments of non-linear factor analysis type models can be found in

Rizopoulos and Moustaki (2007). In SEM, Kenny and Judd (1984) proposed a methodology for handling non-linear terms. Their paper led to a series of papers by other researchers discussing the limitation of their model and ways of improving it. In the SEM approach, products of observed variables are used as indicators for the interaction terms between the latent variables. Recently, a series of papers by Lee and Zhu (2002), Lee and Song (2004) and Song and Lee (2004) discuss Bayesian and ML estimation methods for nonlinear factor models with mixed data that allow for missing outcomes and hierarchical structures. In all those papers dichotomous and ordinal variables are assumed to be manifestations of underlying normally distributed continuous variables.

Rizopoulos and Moustaki (2007) incorporated nonlinear terms in the factor analysis model without making complex distributional assumptions for the joint distribution of the manifest variables. Their method is a full maximum likelihood estimation and therefore the computational burden is quite significant.

The paper is organized as follows: section 2 reviews the model for longitudinal data that has been developed in the paper by Cagnone et. al. (2008), section 3 reviews the model with non-linear terns that has been developed in Rizopoulos and Moustaki (2007), section 4 outlines the estimation method, section 5 discusses goodness-of-fit issues and section 6 presents an application.

Notation

Let $y_{1t}, y_{2t}, \ldots, y_{pt}$ be the ordinal observed variables measured at time t, $(t = 1, \ldots, T)$. The whole response vector for an individual m is denoted by $\mathbf{y}_m = (\mathbf{y}'_{1m}, \ldots, \mathbf{y}'_{tm}, \ldots, \mathbf{y}'_{Tm})'$. In the case of cross-sectional data $T = 1$. Small letters are used to denote both the variables and the values that these variables take. Let c_i denote the number of categories for the ith variable, $(i = 1, \ldots, p)$.

The c_i ordered categories have probabilities $\pi_{i1(t)}(\mathbf{z}), \pi_{i2(t)}(\mathbf{z}), \cdots, \pi_{ic_i(t)}(\mathbf{z})$, which are functions of the vector of latent variables $\mathbf{z}$. As we will see in the following section, the vector of latent variables in the longitudinal case consists of two different types of latent variables. The time-dependent latent variables ξ_t $(t = 1, \ldots, T)$ and the random effects u_i, $(i = 1, \ldots, p)$. The time-dependent latent variables account for association among items within time where the random effect terms account for association between the same items across time. For the latent variable model with non-linear terms the vector $\mathbf{z}$ consists of the main effects of the latent variables $\xi_j, (j = 1, \ldots, q)$ and of higher order terms of those latent variables (quadratic or interaction terms).

The latent variables $\mathbf{z}$ affect directly the manifest ordinal variables or to be more precise the probability of responding into a specific category.

2 Models for multivariate longitudinal ordinal responses

We present here the model framework developed in Cagnone et. al. (2008) for analyzing ordinal longitudinal manifest variables with latent variables. At each time point, the association among the ordinal observed variables $\mathbf{y}_t$ is explained by a set of latent variables $\boldsymbol{\xi}_t$. Those latent variables are time dependent. In addition to the vector of latent variables an item-specific random effect u_i is included to capture the dependencies between same items measured across time.

The linear predictor is written as:

$$\eta_{it(s)} = \tau_{i,s,t} - \boldsymbol{\alpha}'_{it}\boldsymbol{\xi}_t + u_i \ , \quad i = 1, \ldots, p; \ s = 1, \ldots, c_i; \ t = 1, \ldots, T. \quad (1)$$

where $\tau_{i,s,t}$ are item, time, category specific fixed intercepts and $\boldsymbol{\alpha_{it}}$ are factor loadings measuring the effects of $\boldsymbol{\xi_t}$ on some function of the response probability. Finally, the link between the systematic component and the conditional means of the random component distributions is $\eta_{it(s)} = v_{it(s)}(\gamma_{it(s)})$ where $\gamma_{it(s)} = P(y_{it} \leq s \mid \boldsymbol{\xi}_t, u_i)$ and $v_{it(s)}(.)$ is the link function which can be any monotonic differentiable function. Defining the linear predictor in this way, the associations among the items measured at time t are explained by the vector of latent variables $\boldsymbol{\xi}_t$. The associations among the same item measured across time $(y_{i1}, \ldots, y_{iT})$ are explained by the item-specific random effect u_i. We also assume that the time dependent latent variables are linked with a first order autoregressive structure

$$\boldsymbol{\xi}_t = \boldsymbol{\phi}'\boldsymbol{\xi}_{t-1} + \boldsymbol{\delta}_t \quad (2)$$

where for identification purposes $\boldsymbol{\delta}_t \sim N(\mathbf{0}, I)$. The assumption that $u_i \sim N(0, \sigma_u^2)$, completes the model definition. Equation (2) accounts for serial correlation in the latent variables and expresses the dynamic nature of latent variables. The random effects are assumed to have variances equal to $\sigma^2_{u_i}$ and covariances equal to zero, $i = 1, \ldots, p$.

3 Factor models for ordinal responses with non-linear terms

The model presented here is discussed in detail in Rizopoulos and Moustaki (2007). Let us consider the case of cross-sectional survey data. When non-linear terms are added into the model the linear predictor is written as:

$$\eta_{i(s)} = \tau_{i,s} - \boldsymbol{\alpha}'_i\boldsymbol{\xi} \ , \quad i = 1, \ldots, p; \ s = 1, \ldots, c_i. \quad (3)$$

where now the vector of latent variables contains both main effects as well as quadratic and interaction terms between the latent variables. In the case of

two latent variables with an interaction term, the vector of latent variables becomes $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_1 \times \xi_2)'$. The parameters $\boldsymbol{\alpha}_i$ remain the factor loadings. The latent variables $\boldsymbol{\xi}= (\xi_1, \xi_2)$ are assumed to follow a bivariate standard normal distribution with zero correlation. The assumption of zero correlation can be also relaxed if necessary. The inclusion of higher-order terms allows the analysis of complex real situations. The model has been estimated with ML using a hybrid integration-maximization algorithm and standard errors are corrected for model mis-specification via the sandwich estimator.

4 Estimation

We denote with $\mathbf{y}_m$ the response vector for the mth individual. The vector $\mathbf{z}_m$ will contain both the latent variables $\boldsymbol{\xi}_m$ (those will be time-dependent in the model for longitudinal responses) and the random effect component term $\mathbf{u}_m$ required again for the longitudinal case.

All out inference is based on the joint density function of the data written as:

$$f(\mathbf{y}_m) = \int \cdots \int h(\mathbf{z}_m)\; g(\mathbf{y}_m \mid \mathbf{z}_m)\; d\mathbf{z}_m, \tag{4}$$

The model parameters are the factor loadings and the variances of the random effects. In the vector $\mathbf{z} = (\boldsymbol{\xi}, \mathbf{u})$, the $\boldsymbol{\xi}$ latent variables are taken to be independent from the random effects ($\mathbf{u}$).

If the EM algorithm is going to be used for estimating the model parameters one needs to write down the complete data log-likelihood. Therefore, for a random sample of size n the complete log-likelihood is written as:

$$\begin{aligned} L &= \sum_{m=1}^{n} \log f(\mathbf{y}_m, \mathbf{z}_m) \\ &= \log \sum_{m=1}^{n} \left[\log g(\mathbf{y}_m \mid \mathbf{z}_m) + \log h(\mathbf{z}_m)\right] \end{aligned} \tag{5}$$

where g is the likelihood of the data conditional on the latent variables and the random effects and h is the common distribution function of the latent variables and the random effects. Note that factor loadings will be estimated from the first component of the log-likelihood in (5) where the variances of the random effects and the autoregressive parameter for the longitudinal model will be estimated from the second component.

Responses across items and time are independent conditional on the latent variables ($\boldsymbol{\xi}_m$) or some function of those latent variables and the random effects u_{im} (conditional independence):

$$g(\mathbf{y}_m \mid \mathbf{z}_m) = \prod_{i=1}^{pT} g(y_{mi} \mid \mathbf{z}_m), \tag{6}$$

where $T = 1$ in cross-sectional studies and the conditional distribution for each item is taken to be the multinomial. In the longitudinal case, the latent variables $\boldsymbol{\xi}_m$ are dependent through the autoregressive structure but they are independent from the random effects u_{im} according to the model specification given in (1).

Note that in the case of the model with non-linear terms, there are some computational advantages. First, their approach is based on conditional distributions of observed variables given the latent variables and that keep their form within the known distributions of the exponential family. Second, for the two-factor model with nonlinear terms the marginal distribution of the data $f(\mathbf{y})$ is computed using a double integral instead of a higher order integration, while allowing for more complex latent structures compared with a two-factor model.

The E-M algorithm requires the following steps:

E-M algorithm

step1 Choose initial estimates for all model parameters.
step2 Compute the expected score function (E-step).
step3 Obtain improved estimates for the parameters by solving the maximum likelihood equations for the parameters. (M-step)
step4 Return to step 2 and continue until convergence is attained.

The integration involved in the E-step of the algorithm can be achieved either by using adaptive Gauss-Hermite quadrature or Monte Carlo integration. Details of the estimation of the models are given in Cagnone et. al. (2008) and Rizopoulos and Moustaki (2007). Accelerated tools such as the PX-EM and hybrid estimation algorithms have been proposed for speeding up the estimation. The convergence of the E-M algorithm is monitored through the change in the values of the log-likelihood. Initial estimates are chosen arbitrarily.

Model used for ordinal data

The linear predictor given in equations (1) and (3) is linked with the cumulative probability of responding below a category s through a link function. Possible links function are the logit, the complementary log-log function, the inverse normal function, the inverse Cauchy, or the log-log function. The logit and the inverse normal function also known as probit are the link functions most often used in practice. The probit and the logit link function have very similar shapes and therefore give similar results.

When the logit link is used the model is known as the proportional odds model written as:

$$\log\left[\frac{\gamma_{i,s,t}(\mathbf{z})}{1-\gamma_{i,s,t}(\mathbf{z})}\right] = \tau_{i,s,t} - \boldsymbol{\alpha}_i'\mathbf{z} \tag{7}$$

where $s = 1, \ldots, c_i - 1; i = 1, \ldots, p; t = 1, \ldots, T$. From (7) we get that:

$$\gamma_{i,s,t} = P(y_{i,t} \leq s \mid \mathbf{z}) = \frac{\exp(\tau_{i,s,t} - \boldsymbol{\alpha}_i' \mathbf{z})}{1 + \exp(\tau_{i,s,t} - \boldsymbol{\alpha}_i' \mathbf{z})}, \tag{8}$$

where $s = 1, 2, \ldots, c_i - 1$ and $\gamma_{i,m_i,t} = 1$. In cross-sectional studies, $T = 1$.

5 Goodness-of-fit tests and measures

Goodness-of-fit of the models can be checked in three different ways. One way is to consider overall goodness-of-fit tests such as the Pearson X^2 and the likelihood ratio test. However, those statistics can rarely be used in practice due to the fact that are both greatly distorted in the presence of sparseness in multi-way contingency tables.

Alternatively, one can use goodness-of-fit measures such as those suggested by Jöreskog and Moustaki (2001). They have proposed an alternative procedure that does not provide a test statistic but rather focuses on measurement of fit. They investigate how well the model fits the univariate and bivariate marginal distributions. A Likelihood ratio (LR-fit) and a Pearson goodness-of-fit statistic (GF-fit) are computed from the univariate and bivariate contingency tables. Those statistics are not χ^2 distributed but they can be used as measures of fit indicating items and pair of items where the fit is not good.

Finally, one can use differences in the deviance as well as model selection criteria such as the AIC and BIC information criteria for choosing among alternative models.

6 Applications: Efficacy data

The data consists of the USA sample of the political action survey[1] (Barnes and Kaase, 1979). Initially, six variables were analyzed as being indicators of *political efficacy*. Political efficacy has been defined as "*the feeling that individual political action does have, or can have, an impact upon the political process... The feeling that political and social change is possible, and that the individual citizen can play a part in bringing about this change*" (Campbell et al., 1954, p.187). The data set has been extensively analyzed in the literature. Jöreskog and Moustaki (2001) has shown using different approaches for analyzing ordinal data that the six items are not unidimensional. Therefore, for the purpose of our analysis we chose just three items that form a unidimensional scale. Those items are given below:

[1] The data was made available by the Zentralarchiv für Empirische Sozialforschung, University of Cologne. The data was originally collected by independent institutions in different countries.

NOSAY People like me have no say in what the government does
COMPLEX Sometimes politics and government seem so complicated that a person like me cannot really understand what is going on
NOCARE I don't think that public officials care much about what people like me think

Permitted responses to these questions were *agree strongly (***AS***), agree (***A***), disagree (***D***), disagree strongly (***DS***), don't know*, and *no answer*. Missing values have been eliminated from the analysis.

We fitted the non-stationary model with the item-specific random effects. The results are given in Table 1. The estimated loadings are close to the solution obtained from LISREL when correlated errors are estimated between the same items across the two time points. The variances of the random effects are found significant indicating that there is strong correlation between the same items across time.

Table 1. Parameter estimates and standard errors in brackets, Efficacy data.

Items	$\hat{\alpha}_{i1}$ (time 1)	$\hat{\alpha}_{i2}$ (time 2)	$\hat{\sigma}_{ui}$
NOSAY	1.00	1.00	0.29 (0.07)
COMPLEX	0.73 (0.16)	0.73 (0.16)	1.70 (0.44)
NOCARE	1.33 (0.17)	1.33 (0.17)	0.19 (0.03)

The estimated unrestricted covariance matrix of the time-dependent attitudinal latent variables is:

$$\begin{bmatrix} 3.92 & 2.55 \\ 2.55 & 2.66 \end{bmatrix}$$

The estimated ϕ parameter ($\hat{\phi} = 0.65$, with an estimated standard error equal to 0.08) shows a strong and significant correlation between the latent variable across the two time points.

The overall fit measure given in Table 2 indicate a very satisfactory fit. The big chi-square value between item NOSAY and NOCARE at time 1 is due to just one large discrepancy between category 1 for item NOSAY and category 4 for item NOCARE where the observed frequency is only 3.

7 Conclusion

The paper reviews latent variable models for longitudinal ordinal data and latent variable models with non-linear terms. Both models try to explain complex data structures. The model for longitudinal data needs to use a large number of factors including the random effects to be able to explain

Table 2. Univariate and Bivariate GF-Fits, Efficacy data.

NOSAY(t = 1)	–					
COMPLEX(t = 1)	15.13	–				
NOCARE(t = 1)	37.55	18.36	–			
NOSAY(t = 2)	16.08	7.59	7.41	–		
COMPLEX(t = 2)	12.91	7.91	6.47	25.93		
NOCARE(t = 2)	6.19	7.76	6.49	19.87	23.70	–
OVERALL GF-FIT=**14.62**						

the associations among the ordinal responses. Linking the time-dependent latent variables with an autoregressive model complicates even more the model assumptions and structure. Similarly complicated is the model with the non-linear terms that tries to incorporate higher-order terms of the latent variables in the predictor. The proposed estimation method for both models is FIML. FIML requires heavy integrations that limits the number of items and factors that can be analyzed. The use of adaptive quadrature routines have allowed one to use a very small number of points for approximating the integrals. We hope that with the further advances of computer power those computational limitations will be eliminated and the use of FIML will become a common practice among researchers for estimating complex models.

Issues of goodness-of-fit and model identification remain still open for investigation.

References

BARNES, S. and KAASE, M.E. (1979): *Political action: Mass participation in five western democracies*. Beverley Hills and London: Sage Publications.

BARTHOLOMEW, D.J. and KNOTT, M. (1999): *Latent Variable Models and Factor Analysis* (2nd ed.). London: Arnold.

BARTLETT, M.S. (1953): Factor analysis in psychology as a statistician sees it. In *Uppsala Symposium of Psychology and Factor Analysis*, pp. 23–34. Uppsala: Almqvist & Wicksell.

BOLLEN, K.A. and CURRAN, P.J. (2006): *Latent Curve Models: a Structural Equation Perspective*. New York: John Wiley.

CAGNONE, S., MOUSTAKI, I. and VASDEKIS, V.: Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology*. (To appear).

EID, M. (1996): Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychoilogical Research 1*(4), 65–85.

GIBSON, W.A. (1960): Nonlinear factors in two dimensions. *Psychometrika 25*, 381–392.

JÖRESKOG, K.G. (2002): Structural equation modeling with ordinal variables using LISREL. Available at http://www.ssicentral.com/lisrel/ordinal.htm.

JÖRESKOG, K.G. and MOUSTAKI, I. (2001): Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research 36*, 347–387.

JÖRESKOG, K.G. and SÖRBOM, D. (1999): *LISREL 8 User's Reference Guide*. Chicago: Scientific Software International.

KENNY, D.A. and JUDD, C.M. (1984): Estimating the non-linear and interactive effects of latent variables. *Psychological Bulletin 96*, 201–210.

LEE, S.Y. and SONG, X.Y. (2004): Maximum likelihood analysis of a generalized latent variable model with hierarchically mixed data. *Biometrics 60*, 624–636.

LEE, S.Y. and ZHU, H. (2002): Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika 67*, 189–210.

MARSH, H. and GRAYSON, D. (1994): Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling 2*, 116–145.

McDONALD, R.P. (1962): A note on the derivation of the general latent class model. *Psychometrika 27*(2), 203–206.

RAFFALOVICH, L. and BOHRNSTEDT, G. (1987): Common, specific, and error variance components of factor models. *Sociological Methods and Research 15(4)*, 385–405.

RIZOPOULOS, D. and MOUSTAKI, I. (2007): Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*. in press.

SCHUMACKER, R. and MARCOULIDES, G.E. (1998): *Interaction and Nonlinear Effects in Structural Equation Models*. New Jersey: Lawrence Erlbaum Associates.

SONG, X.Y. and LEE, S.Y. (2004): Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology 57*, 29–52.

Part VI

Computational Statistics and Data Mining Methods for Alcohol Studies

Estimating Spatiotemporal Effects for Ecological Alcohol Systems

Yasmin H. Said[1,2]

[1] Isaac Newton Institute for Mathematical Sciences
University of Cambridge, Cambridge CB3 0EH UK
[2] Department of Computational and Data Sciences
George Mason University, Fairfax, VA 22030 USA
ysaid99@hotmail.com

Abstract. In this paper, I consider data on fatal automobile crashes, DWI arrests, and alcohol addiction admissions in Virginia, USA and use these as a basis for estimating the hourly, weekly, monthly, and annual cycles associated with alcohol consumption. In addition, I use surveys carried out by the Department of Alcoholic Beverage Control in Virginia to establish geospatial patterns of purchases of distilled spirits. This data analysis allows me to conjecture spatiotemporal patterns that can be incorporated into calibration of a more complex ecological alcohol systems model.

Keywords: social networks, geospatial and temporal effects, adjacency matrix, one-mode networks, two-mode networks, bipartite graphs, allegiance

1 Introduction

Alcohol use and abuse contributes to both acute and chronic negative health outcomes and represents a major source of mortality and morbidity in the world as a whole (Ezzati et al., 2002) and in the developed world, such as the United States, where alcohol consumption is one of the primary risk factors for the burden of disease. It ranks as a leading risk factor after tobacco for all disease and premature mortality in the United States (Rehm et al., 2003; Said and Wegman, 2007). Said and Wegman (2006) outlined a graph-theoretic agent-based simulation tool and applied this tool to examine alcohol-related violence in Northern Virginia. That work did not incorporate spatial or temporal factors except with respect to residence locations of the agents. In a companion piece to this paper, Wegman and Said (2008) expand the model to accommodate the temporal and geospatial dimensions of acute outcomes. In contrast with the modeling paper, Wegman and Said (2008), this paper focuses on methods to exploit temporal and spatial data with the idea of calibrating the model to include these effects. The model proposed in Wegman and Said (2008) incorporates a social network component that was not included in Said and Wegman (2006). The overall goal of the models is not just to simulate known data, but to provide a policy tool that would allow decision makers to examine the feasibility of alcohol-related interventions.

Such interventions are designed to reduce one or more acute outcomes such as assault, murder, suicide, sexual assault, domestic violence, child abuse and DWI-related injuries and deaths. Unfortunately, interventions can sometimes lead to unintended consequences, while suppressing one acute outcome, other acute outcomes may be enhanced. By adjusting conditional probabilities, the effect of interventions can be explored without actually introducing major societal policy actions.

2 Estimating temporal and cyclic effects

It is well known that there are substantial seasonal and temporal effects associated with alcohol use and its acute outcomes (Fitzgerald and Mulford 1984; Cho et al. 2001; and Carpenter 2003). For purposes of analysis of interventions, it is desirable to understand when and where interventions may be most effective. Alcohol use shows patterns on multiple scales including time-of-day, day-of-week, and week-of-year as well as month-of-year. The detailed construction of probabilities conditioned on job class, ethnicity, gender, socioeconomic status, residential location, age, and a host of other demographic factors results in many thousands of conditional probabilities to be estimated. Adding temporal effects specific to each of the many thousands of combinations of relevant variables and acute outcomes is unrealistic with existing data sets. In order to incorporate temporal effects in the model outlined in Wegman and Said (2008), I need to develop proxies for the temporal effects. In this paper, I consider data from the Virginia Department of Motor Vehicles concerning alcohol-related fatal crashes. In addition, I consider DWI arrest data from Fairfax County, Virginia and, finally, alcohol treatment admission data from Fairfax County, Virginia. I investigate the cyclic effects over the period 2000-2005.

The main data resource consists of state-wide data from Virginia Department of Motor Vehicles on all alcohol-related crashes from 2000 to 2005. These data were made available with time-of-day as well as date information. Based on this, I am able to assess day-of-week and month-of-year information. These data are aggregated for each year and plotted in Figure 1 by time of day. There is remarkable consistency in these plots suggesting that there is comparatively little yearly effect. As one might expect, drinking behaviors increase during late afternoon and early evening and this is reflected in the increases in alcohol-related fatal crashes. The very early morning events (1 to 3 am) after closing hours of most alcohol outlets are associated with attempts to drive after an extended evening of drinking.

Figure 2 plots the crashes by day-of-week. Again the data are aggregated across the years by day of the week. There is once again remarkable consistency over the annual period. As one might reasonably expect drinking behaviors increase with the onset of the weekend. Typically, drinking behav-

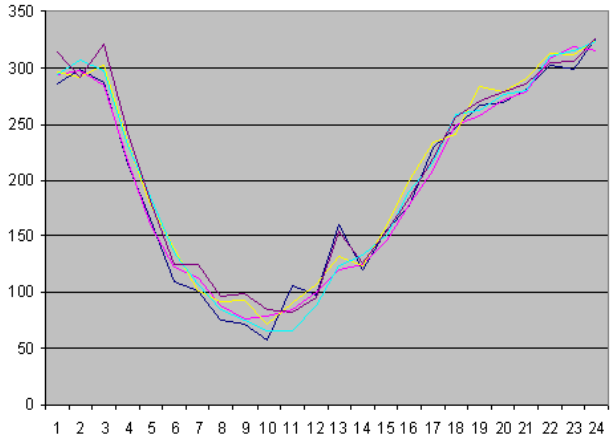

Fig. 1. Alcohol Related Crashes by Time-of-Day.

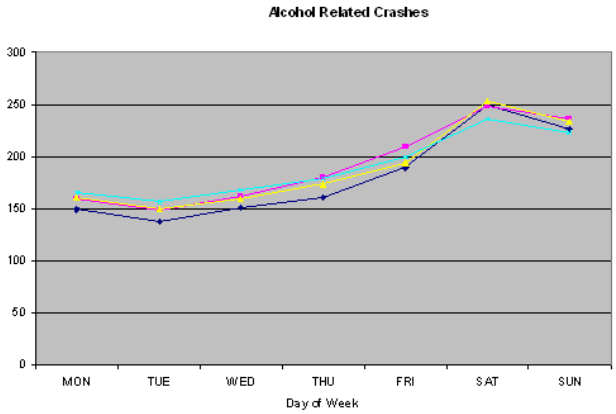

Fig. 2. Alcohol Related Crashes by Day-of-Week.

iors begin on Friday are most intense on Saturday and begin to subside on Sunday.

Examining the month-of-year effect, I find that the lowest crash rates are in January and February. The traditional holiday period, October, November, and December incur the greatest number of fatal crash events. Surprisingly, the peak of the summer driving/summer holiday season is comparatively low. Even more surprising is the relative peak in May. I conjecture that this is associated with the end of the semester for high schools and colleges with proms, graduations, and other social events for high schools and collegiate celebrations at the end of final exams and commencements.

Figures 1-3 suggest that there are very substantial time-of-day effects, substantial day-of-week effects, a somewhat less pronounced monthly effects,

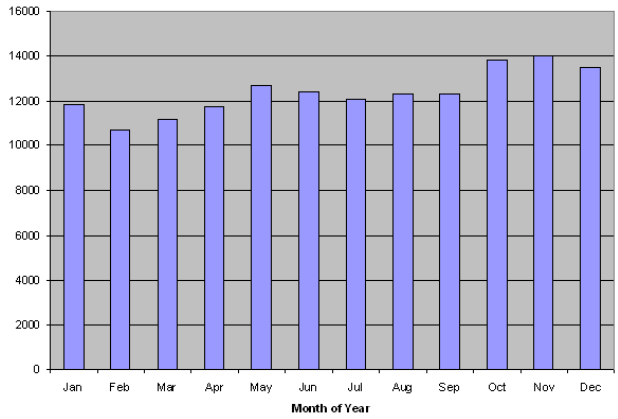

Fig. 3. Alcohol Related Crashes by Month of Year.

and virtually no variation over the years. In order to investigate this more analytically, I consider mixed effects linear models. The first part of this analysis examines data from the Virginia Department of Motor Vehicles alcohol-related fatal crashes for the period 2000-2005 (896,574 instances). After summarization, the data set has 2,192 instances (365×6). The date of the crash is used to extract the year, month, week, and day of the week. Analysis is done on the number of crashes on a particular day of the week. The second part of my analysis focuses specifically on Fairfax County.

The response variable, alcohol-related fatal crashes, is skewed to the left. This indicates non-normality. The square root transformation is used for this data set and the resulting plots are shown in Figure 4.

I consider the mixed effects model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$$

where y_{ijk} is the observation of the ith day of the jth week of the kth year. In this case, $i = 1, \ldots, 7$, $j = 1, \ldots, 52$, and $k = 1, \ldots, 6$ where α_i is the day-of-week effect, β_j is the week-of-year effect, γ_k is year, μ is the fixed intercept, and ϵ_{ijk} is the noise. Day-of-week variations are highly significant, week-of-year variations marginally significant, and the year effect is not significant.

While the alcohol-related crash data are state-wide data, they do not capture temporal characteristics of DWI arrests and alcohol-related hospital admissions. In Figure 5, we have the boxplot of DWI arrests by day-of-week. This strongly resembles the day-of-week plot for alcohol-related fatal crashes state-wide. Of course, there are many more DWI arrests than there are alcohol-related fatal crashes. In these figures, day 1 is Monday and, as before, both fatal crashes and DWI arrests are lowest on Tuesdays. Figure 6 is the boxplot of alcohol-related treatment admissions. This essentially is

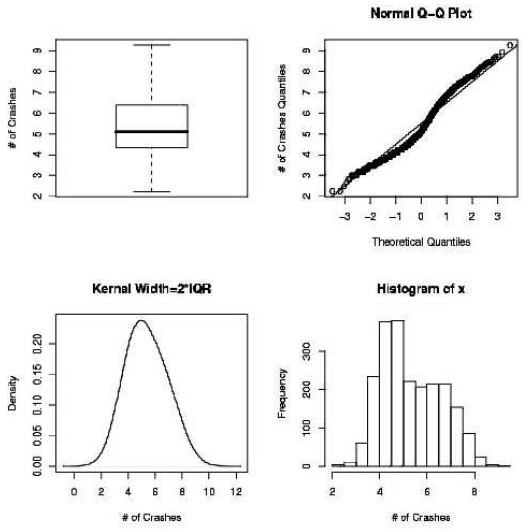

Fig. 4. Number of Crashes Plots Transformed by Square Root.

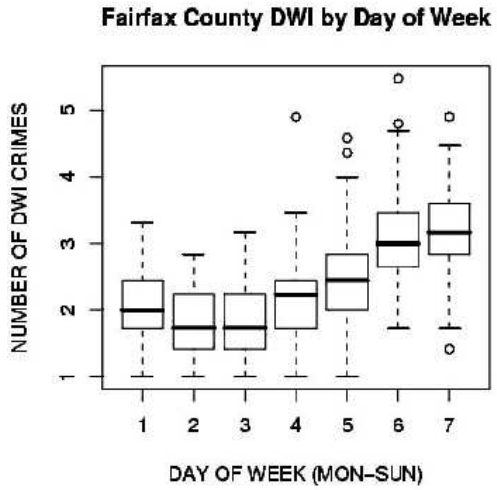

Fig. 5. Fairfax County DWI Arrests by Day of Week.

inversely correlated with the DWI and crash data. Here, Tuesday is the peak day for admissions to treatment.

In summary, the time-of-day and day-of-week alcohol-related crashes provide an excellent proxy for the critical temporal effects needed in my model and allow us to incorporate these temporal effects.

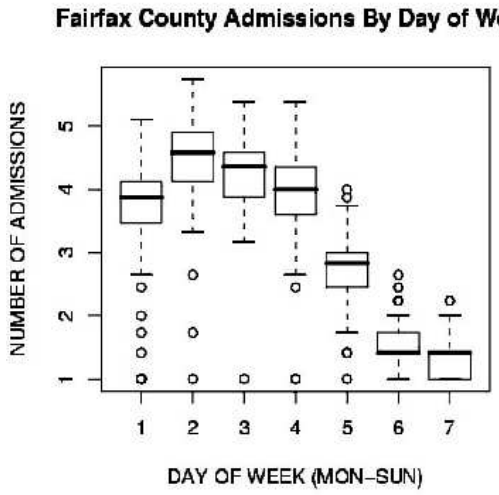

Fig. 6. Fairfax County Hospital Admissions for Alcohol Treatment.

3 Estimating geospatial effects

Fundamental to understanding the alcohol ecological system is an understanding of the transportation system, of the geospatial distribution of alcohol availability, and of scenarios by which alcohol users acquire and use their alcohol. This obviously is a data-intensive exercise and can become exceptionally complex. Ideally, simulating the entire transportation system and its implications for the alcohol system would be the most desirable course of action. Such transportation models are used to generate activities as part of more comprehensive transportation simulation. A prime example is the TRANSIMS system (Smith et al., 1995), which was a comprehensive model developed under federal support at Los Alamos National Laboratory and which has been commercialized by IBM. The TRANSIMS activity generator constructs activities, their locations, and method of travel between locations, for each member of every simulated household in a city. The program was used in a demonstration project in Portland, Oregon, a city whose greater metropolitan area was approximately 1.6 million residents and is currently approximately 2.5 million residents.

Transportation planners regularly conduct extensive travel surveys. An example was the Portland Travel/Activity Survey of 1994-1995 (Cambridge Systematics, 1996; Shiftan and Suhrbier, 2002; Buliung and Kanaroglou, 2004). Participants in the survey kept diaries and recorded every activity with duration and location for a two-day period for each member of the survey households. In addition, extensive demographic information was recorded. There were 4,451 households and 10,048 individuals in the survey. 129,188

activities were recorded. It should be noted that drinking behavior was not recorded in this survey. A number of activities, presumably including drinking, were aggregated in a leisure activity category. However, the survey did provide valuable data on daily patterns of activity, periods of work, meals, shopping, and leisure activities.

Because a large percentage of alcohol-related incidents take place between alcohol outlets and the alcohol user's residence, either at the outlet itself where drinking behaviors take place, at the residence, the other principal site for alcohol use, or in between for such incidents as DWI and alcohol-related crashes, I take advantage of data collected by the Department of Alcoholic Beverage Control in Virginia. The ABC survey is conducted periodically at ABC Stores. Customers are asked their residential postal code and this information is recorded for planning purposes in order to determine the best locations for new alcohol outlets. Presumably, best is judged relative to enhanced sales, not in terms of optimal societal or public health benefits. While not as elaborate as something like the TRANSIMS simulation effort or the Portland Travel/Activity Survey, it is more germane to geospatial aspects of alcohol-related behaviors.

The ABC surveys essentially provide a two-mode social network and arranged in the form of an adjacency matrix, I can form associations between alcohol outlets and postal codes. A gray-scale rendition of the ABC by postal code adjacency matrix is given in Figure 7.

There is a fundamental duality between the graph representation of a social network and the adjacency matrix. Because of this, the two-mode network can easily be transformed into two one-mode networks. If $\boldsymbol{X}_{\boldsymbol{m\times n}}$ is the m by n adjacency matrix of the two-mode network, then $\boldsymbol{A}_{\boldsymbol{m\times m}} = \boldsymbol{X}_{\boldsymbol{m\times n}}\ \boldsymbol{X}^{\boldsymbol{T}}_{\boldsymbol{n\times m}}$, the matrix product of the $\boldsymbol{X}$ matrix with its transpose is the one-mode adjacency matrix for the alcohol outlets relative to the postal codes of purchasers. Similarly, $\boldsymbol{Z}_{\boldsymbol{n\times n}} = \boldsymbol{X}^{\boldsymbol{T}}_{\boldsymbol{n\times m}}\ \boldsymbol{X}_{\boldsymbol{m\times n}}$ is the one-mode adjacency matrix for the postal codes relative to the alcohol outlets. It is now desirable to cluster these networks in order to determine which postal codes and which ABC stores behave in a similar way.

Clustering in social networks begins with dividing the set of actors into discrete, non-overlapping subsets called partitions. The partition determines the block model (Wasserman and Faust, 1994). The block model is the device that clusters the network data. The block, B_{ij}, is formed from the ties of actors in partition i, to the actors in partition j. If $i \neq j$, the block B_{ij} represents ties between partitions i and j. If $i = j$, B_{ij} represents internal ties of actors within the block. These latter diagonal blocks represent a clustering of the actors. Generally speaking, we like to see blocks that are cliques or nearly cliques in the usual graph-theoretic sense of a clique.

Rigsby (2005) developed the concept of allegiance in order to have a systematic way to form the partition and the blocks. As with the usual clustering methods, the appropriate number of clusters is not usually known. A quanti-

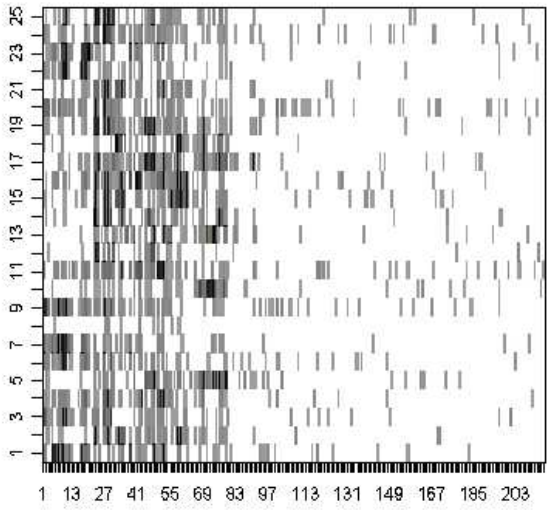

Fig. 7. ABC outlets versus postal codes. There are 25 ABC outlets in Fairfax County, Virginia and more than 230 residential postal codes representing the homes of people who purchase distilled spirits in Fairfax County. Fairfax County contains 48 postal codes, so a considerable amount of alcohol is purchased in Fairfax County by non-county residents. In this Figure, the rows represent the alcohol outlets and the columns represent the residential postal codes.

tative measure of block model strength allows us to estimate the true number of partitions. Allegiance measures the support that an actor provides for the structure of his block. An actor supports his block by having internal block edges. A measure of this is the total number of edges that an actor has internal to his block. An actor supports his block by not having external edges from the block to other actors or blocks. A measure of this is the total number of possible external edges minus the total number of existing external edges. The allegiance for a block is a weighted sum of the measure of internal allegiance and the measure of external allegiance. The overall allegiance for a social network is the sum of the allegiances for the individual blocks. If the overall allegiance is positive then a good partition was made. The partitioning continues recursively until a new partition no longer contributes to a positive allegiance. A more technical discussion can be found in Said et al. (2008).

Using the concept of allegiance, we can create block models for the $\boldsymbol{A}$ and the $\boldsymbol{Z}$ matrices. These are represented respectively in Figures 8 and 9.

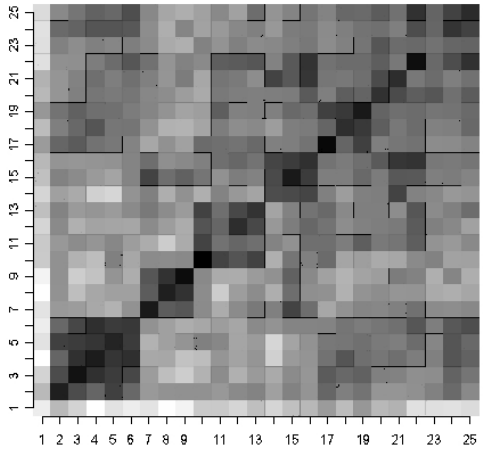

Fig. 8. ABC Stores Block-Model Matrix Rendered in Gray Scale.

The blocks along the diagonal in Figure 8 represent ABC stores that behave similarly with respect to their customers. That is to say, these stores tend to serve the same customer postal code base. Similarly the block along the diagonals of Figure 9 represent postal codes that behave the same way with respect to alcohol outlets, that is, the residents tend to purchase alcohol from the same outlets. For many purposes, the block concept is a way of aggregating actors in a network that behave the same way, in effect, reducing the number of actors.

An interesting statistic associated with Figure 9 is that purchasers of alcohol in Fairfax County, Virginia have their residence as far away as 340 miles. There are a large number of purchasers of alcohol in Fairfax County that live approximately 100 miles away from Fairfax County. I conjecture that these are commuters whose local ABC stores would be closed by the time they arrive at their residences. Fairfax County is West of Washington, DC and would be on a commuter route for those living in the more Western parts of Virginia and West Virginia.

Still, if I focus on the 48 postal codes in Fairfax County and the 25 ABC outlets in Fairfax County, I can establish approximate commuter routes for purchasers of alcohol and thus determine what may well be local hotspots for alcohol-related incidents, by virtue of the amount of alcohol purchased as well as the likely commuter routes. Of course, ABC stores are off-license establishments so that restaurants and bars that are on-license establishments are not accounted for in this exercise. Still, the ABC stores are a reasonable

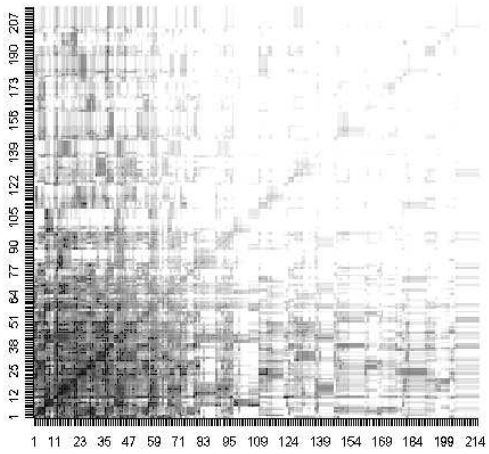

Fig. 9. Postal Codes Block-Model Matrix Rendered in Gray Scale.

proxy for geospatial effects. Figure 10 represents the map of Fairfax County with postal codes.

Based on the adjacency matrix of the two mode network weighted by the amount of alcohol purchased (in terms of dollar sales) and the centroids of the postal codes, I can estimate the distribution of distances traveled between outlet and residence and to at least a crude approximation, estimate the routes alcohol users take between outlet and residence. These are relatively crude proxies, but reasonable first approximations for purposes of calibrating my model.

4 Conclusions

In this paper, I develop an estimation procedure for a comprehensive modeling framework based on agent-based simulation, social networks, and directed graphs that captures some of the complexity of the alcohol ecological system. I provide a framework for temporal and geospatial effects and discuss proxies for approximating these effects. To completely specify the framework outlined in Wegman and Said (2008), considerable additional data collection efforts must be undertaken. Nonetheless, with opportunistically available data sources, I am able to reproduce with reasonable accuracy the actual experiences for a diversity of alcohol-related acute outcomes in Fairfax County, Virginia. The novelty of the framework I have proposed lies in the ability to modify the conditional probabilities on the various paths through the directed graph and thus experimentally determine what social interventions are most

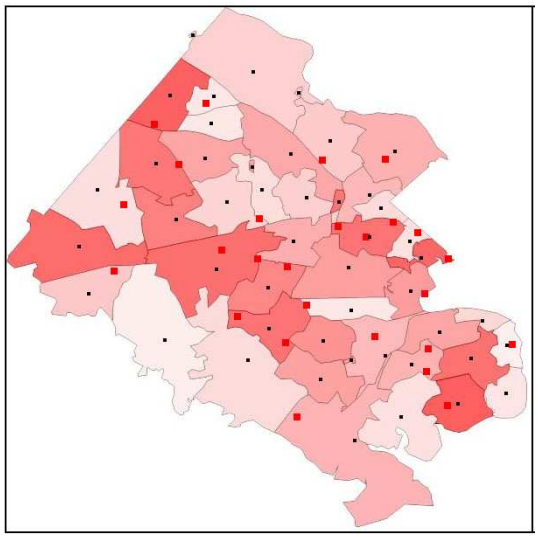

Fig. 10. Fairfax County, Virginia with Postal Codes Outlined. The Color Shading is Proportional to Actual Number of Acute Events in Each Postal Code. The Red Markers are the Locations of ABC stores. The Black Markers are the Centroids of Each Postal Code. From the Two-Mode Adjacency Matrix Restricted to Fairfax County we can Determine Approximate Distances and Routes.

likely to have a beneficial effect for reducing acute outcomes. Although this model is cast in terms of acute alcohol-related outcomes, there is an obvious applicability to expanding this framework to other settings such as homeland security and criminal activities.

Acknowledgements

First, I would like to acknowledge Dr. Edward Wegman, my frequent coauthor and mentor, for his continuing support and stimulating discussions. I would like to acknowledge the contributions of Dr. Rida Moustafa and my students, Peter Mburu and Walid Sharabati, who assisted me with the development of various figures and computations in this paper. I also acknowledge the contributions of John Rigsby whose ideas on allegiance in social networks provided me with insight in the analysis of the geospatial component of my discussion. My work is supported in part by Grant Number F32AA015876 from the National Institute on Alcohol Abuse and Alcoholism. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health. I am a Visiting Fellow at the Isaac Newton Institute for Mathematical Sciences at University of Cambridge in

Cambridge, England. I am in debt for the support provided by the Newton Institute, which has contributed to the successful completion of this work.

References

BULIUNG, R.N. and KANAROGLOU, P.S. (2004): On design and implementation of an object-relational spatial database for activity/travel behaviour research. *Journal of Geographical Systems 6, 237-262.*

CAMBRIDGE SYSTEMATICS (1996): *Data Collection in the Portland, Oregon Metropolitan Area: Case Study.* U.S. Department of Transportation: DOT-T-97-09.

CARPENTER, C. (2003): Seasonal variation in self-reports of recent alcohol consumption: racial and ethnic differences. *J Stud Alcohol 64(3), 415-8.*

CHO, Y.I., JOHNSON, T.P., FENDRICH, M. (2001): Monthly variations in self-reports of alcohol consumption. *J Stud Alcohol 62(2), 268-272.*

EZATTI, M., LOPEZ, A., RODGERS, A., VANDER HOORN, S., and MURRAY, C. (2002): Comparative risk assessment collaborating group. Selected major risk factors and global and regional burden of disease. *Lancet 360, 1347-1360.*

FITZGERALD, J.L. and MULFORD, H.A. (1984): Seasonal changes in alcohol consumption and related problems in Iowa, 1979-1980. *J Stud Alcohol 45(4), 363-368.*

REHM, J., ROOM, R., MONTEIRO, M., GMEL, G., REHN, N., SEMPOS, C.T., FRICK, U., and JERNIGAN, D. (2004): Alcohol. In EZATTI, M., LOPEZ, A.D., RODGERS, A., and MURRAY, C.J.L. (Eds.) *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Due to Selected Major Risk Factors*, WHO, Geneva.

RIGSBY, J.T. (2005): *Block Models and Allegiance.* A thesis submitted to George Mason University in partial fulfillment of the M.S. in Statistical Science, Fairfax, Virginia.

SAID, Y.H. and WEGMAN, E.J. (2006): Geospatial distribution of alcohol-related violence in Northern Virginia. *COMPSTAT2006 197-208.*

SAID, Y.H. and WEGMAN, E.J. (2007): Quantitative assessments of alcohol-related outcomes. *Chance 20(3), 17-25.*

SAID, Y.H., WEGMAN, E.J., SHARABATI, W.K., and RIGSBY, J.T. (2008): Social networks of author-coauthor relationships. *Computational Statistics and Data Analysis 52(4), 2177-2184.*

SHIFTAN, Y. and SUHRBIER, J. (2002): The analysis of travel and emission impacts of travel demand management strategies using activity-based models. *Transportation 29(2), 145-168.*

SMITH, L., BECKMAN, R., ANSON, D., NAGEL, K., and WILLIAMS, M. (1995): TRANSIMS: Transportation analysis and simulation system. *Proceedings of the 5th National Transportation Planning Methods Applications Conference*, Seattle.

WASSERMAN, S. and FAUST, K. (1994): *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge, UK.

WEGMAN, E.J. and SAID, Y.H. (2008): A directed graph model of ecological alcohol systems incorporating spatiotemporal effects. This Volume.

A Directed Graph Model of Ecological Alcohol Systems Incorporating Spatiotemporal Effects

Edward J. Wegman[1,2] and Yasmin H. Said[1,2]

[1] Isaac Newton Institute for Mathematical Sciences
University of Cambridge, Cambridge CB3 0EH, UK
[2] Department of Computational and Data Sciences
George Mason University, Fairfax, VA 22030, USA
ewegman@gmail.com, ysaid99@hotmail.com

Abstract. Users of alcohol are incorporated into a societal system, which for many purposes resembles an ecological system. We have previously modeled such systems using a directed graph with acute outcomes reflecting undesirable individual and societal outcomes. In this paper, we expand the model to a hybrid social network directed graph model for modeling the acute outcomes associated with alcohol use and abuse. We describe the approximate estimates of conditional probabilities based on available data. In the present model, we also approximate geospatial effects related to transportation as well as temporal effects. Acute outcomes include assault, murder, suicide, sexual assault, infection with STDs or HIV, domestic violence, child abuse, and DWI and alcohol related fatal crashes. The model is calibrated using demographic, crime, alcohol-related, and alcohol outlet data from Virginia and Fairfax County in Virginia. We propose proxy data and methods for capturing temporal and geospatial effects. The goal is to investigate methods for simultaneous suppression of multiple negative public health consequences. The model may be used as a public policy decision tool by adjusting conditional probabilities in order to investigate the effect of interventions.

Keywords: directed graph, social networks, acute consequences, public health, interventions, geospatial effects, temporal effects

1 Introduction

Alcohol use is a major cause of morbidity and mortality in all areas of the world (Ezzati et al., 2002). In the developed world alcohol consumption is one of the primary risk factors for the burden of disease; it ranks as the second leading risk factor (after tobacco) for all disease and premature mortality in the United States (Said and Wegman, 2007). Alcohol is such a potent risk factor because it is widely used, inexpensive, and causes both acute and chronic consequences. A major difficulty in developing and assessing alcohol-related public health interventions is that drinking behaviors and their consequences form, in effect, a complex ecological system of individual behaviors embedded in socio-cultural settings and interactions that are distributed over space and

time. Interventions focused on any specific aspect of this system (e.g., drinking and driving) may result in unintended consequences (e.g., a potential increase in domestic violence as the result of high volume drinkers spending more time at home due to the anti-drinking-driving programs). Another example of unintended consequences is alcohol-related promiscuous behavior resulting in infections with sexually transmitted diseases (STD) and Human Immunodeficiency Virus (HIV).

The research that we report here is designed to model simultaneously acute alcohol-related outcomes among all individuals by specific geographic areas. Thus, the potential impact of interventions that change any aspect of the entire alcohol ecological system can be examined across population subgroups and locations. This information can be used to guide the development of interventions that are most likely to have the impact intended, and to assist in the selection of interventions that are most worthy of deserving the large amount of resources necessary for conducting a public health field trial.

We develop a model of the alcohol ecological system based on population subgroups, age, gender, ethnicity, socioeconomic status, composed of individuals (i.e., the agents in the model), embedded in space (geographic location identified by postal code areas), time (i.e., time-of-day, day-of-week), and socio-cultural context (e.g., alcohol availability, neighborhood). The model focuses on acute alcohol-related outcomes (e.g., domestic violence, homicide, crashes, STD, HIV, etc.). The model is agent-based, meaning that individual outcomes are simulated; this software design allows us to incorporate specific agent interactions. This approach focuses on ecological systems, i.e., behaviors and interactions of individuals and their setting, not on aggregate groups as the unit of analysis. The model is stochastic in that every agent in a specific demographic subgroup will have an individual outcome, rather than all subgroups members having the same outcome.

The specific intent for this research is: 1) To create a spatiotemporal social network stochastic directed-graph model of a population accounting for agents' alcohol use and the associated acute outcomes. We initially model an "average" day (including activities by time-of-day), and then move to models of weekend days and other days of the week. The model runs for the entire population for a full year (365 days) to examine the distribution of acute outcomes; 2) To collect and identify the data that are specifically required to calibrate, validate, and verify the model. This includes assessing a large number and variety of data sets describing alcohol-related behaviors. In situations where there was no data, we consulted with experts in the field so we could develop expert opinions, which, in turn, have guided us to estimate the conditional marginal probabilities; and 3) To incorporate a policy tool in the model in order to examine intervention strategies to understand where policies may be most effective in suppressing simultaneously all of the acute outcomes. This initial policy tool allows for examinations of the impact of

changes in alcohol availability (by outlet type and by geographic location) and population demographics on the number, type, and location of acute alcohol-related outcomes.

2 Building the model

Graph theory allows for representation of many organizational and relational structures (Bondy and Murty, 1976; Diestel, 2005). Special forms of graph theory are the directed graph and stochastic or random graphs. (Chartrand and Lesniak, 1986; Bang-Jensen and Gutin, 2000; Bollobos, 2001; Marchette, 2004). A directed graph (often called a digraph) G is a pair (V, E) where V is a set of elements called vertices or nodes and E is a subset of the set of all ordered pairs (a, b), where a and b are vertices. An element of E is called a directed edge, an arc, or an arrow of G. An edge $e = (a, b)$ is considered to be directed from a to b; b is called the head and a is called the tail of the edge. The pair (a, b) is not the same as the pair (b, a). Typically, we regard the direction of the edge (a, b) as flowing from a to b. Conventionally, an edge of the digraph is represented as two points representing a and b with an arrow whose tail is at a and whose head is at b. A directed graph allows both (a, b) and (b, a). Sometimes E is allowed to be multisets, so that there can be more than one edge between the same two vertices. This is useful distinction when the relationship between a pair of nodes is asymmetric.

Our model of the alcohol system is an agent-dependent, time-dependent stochastic digraph. The concept is that the vertices of the digraph will represent the state of the agent (e.g. including such factors as physical location, present activity, and level of Blood Alcohol Content (BAC)) and the edges represent a decision/action that takes the agent into a new state. The agent represents any individual in the population including the alcohol users as well as the non-users. The edge going from one state to another will have a conditional probability attached to it, hence, the notion of a stochastic digraph. The conditional probability attached to a given edge will depend on the specific sub-population from which the agent is drawn, hence is agent-dependent, and the conditional probability will, in principle, also depend on the time-of-day, hence it is time-dependent.

Implicit in this description is that individuals represented in the network relate to each other and to social institutions. Thus, overlaying the directed graph model is a social network. Figure 1 is a representation of the social network for the alcohol user. We develop an agent-based social network digraph model of the alcohol system. Implicit in the social network is that each node (each agent) carries with him or her covariates that characterize such factors as residence, ethnicity, job class, gender, age, socioeconomic status, disposition to alcohol, and the like. While social networks are often represented as graphs, they can also be characterized by adjacency matrices. Figure 2 represents the adjacency matrix characterizing Figure 1.

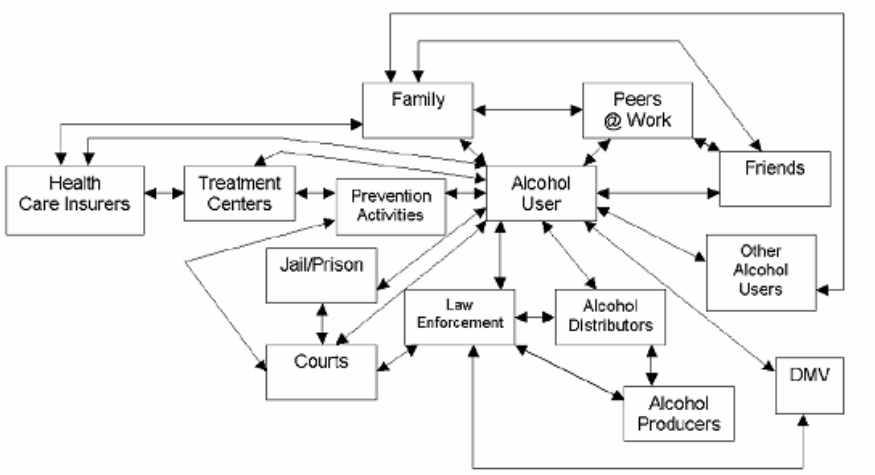

Fig. 1. Social network for the alcohol user. This same network may apply to nonusers with some edges removed.

The concept is that relatively homogeneous clusters of people (agents) will be identified along with their daily activities. Agents can be alcohol users, nonusers, and the whole suite of participants in the social network mentioned above. Agents can interact with each other in asymmetric ways. Their activities will be characterized by different states in the directed graph, and decision resulting in actions by an agent will move the agent from state to state in the directed graph. The leaf nodes in the graph will represent a variety of outcomes, most of which are benign, but a number of which will be acute alcohol-related outcomes. Specifically, we study simultaneously the following acute outcomes: assault, suicide, domestic violence, child abuse, sexual assault, STD and HIV, murder, DWI (drinking and driving episodes, including fatalities).

Figure 3 represents the essentially complete model. In each of the "time planes" we have the social network along with the attendant covariates. The conditional probabilities are, of course, time and agent dependent. For example, a white collar employee who has a normal day job and is not alcohol dependent is unlikely to have an episode of domestic violence in the 6 to 8 am time sector, but could have a higher likelihood of DWI after a three martini lunch. An unemployed worker, however, is more likely to get involved in an assault situation, especially in the late afternoon or early evening out of frustration associated with low income. We represent time as being quantized in Figure 3 although it need not essentially be so. Indeed, while gender, ethnicity, job class are inherently categorical, such covariates as age, socioeconomic status as represented by income level and other covariates may be

	Alcohol User	Family	Peers @ Work	Friends	Other Users	Alcohol Distributors	Alcohol Producer	Law Enforcement	Courts	Prison	Rehab	Treatment	Health Insurance	DMV
Alcohol User		X	X	X	X	X		X	X	X	X	X	X	X
Family	X		X	X	X								X	
Peers @ Work	X	X		X										
Friends	X	X	X											
Other Users	X	X												
Alcohol Distribution	X						X	X						
Alcohol Producers						X		X						
Law Enforcement	X					X	X		X				X	
Courts	X							X		X	X			
Prison	X								X					
Rehab	X								X			X		
Treatment	X										X		X	
Health Insurance	X	X										X		
DMV	X							X						

Fig. 2. Adjacency matrix summarizing strengths of connectivity in the alcohol users directed graph. The x's are place holders for numerical probability values. There is a general symmetry, but the bidirectional probabilities between any two nodes are likely to be unequal.

continuously variable. In the next section we will address compromises made related to data available to calibrate the data.

Now, as agents are introduced into the directed graph model, their ultimate outcomes whether benign or acute will be accumulated so that a (multinomial) probability distribution over all outcomes can be estimated. The ultimate goal is to introduce interventions that will alter the structure of the directed graph or its associated probabilities and assess how those interventions affect the probability distribution over the outcomes. It is possible that an intervention may reduce the incidence of one acute outcome, but increase the prevalence of other acute outcomes. The goal of our research is to study the alcohol system as a whole in order to evaluate best interventions for reducing the overall incidence of acute outcomes.

3 Choosing the agents

Calibrating the model described in Section 3 would ideally involve calculating the conditional probability of each acute event conditioned on residence location, ethnicity, job class, gender, age, economic status, time-of-day, alcohol availability, whether the individual had an alcohol abuse problem or not, and conceivably many more factors that would affect the likelihood of that acute

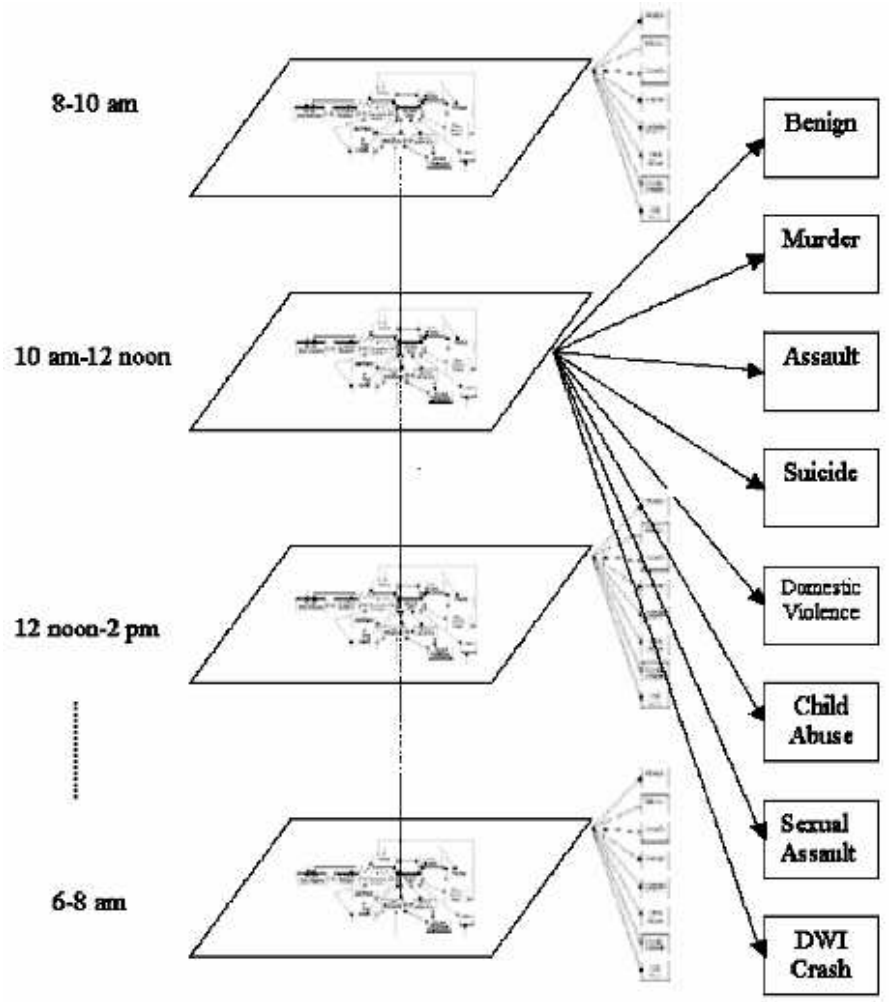

Fig. 3. Time-of-day directed graph with acute outcomes. For illustrative purposes, we have blocked the day into twelve two-hour blocks. Note that any time of day, all outcomes are possible, but because drinking patterns are time-dependent, the likelihood changes throughout the day.

outcome. Unfortunately, there is no direct data to support estimation of the conditional probabilities. Part of the reason, therefore, for turning to a directed graph model is to decompose the desired probabilities into components that can be justified by available data.

The first goal is to devise a procedure that chooses agents in a manner that is representative of and consistent with the population of the geospatial area of interest. Note that if we considered only the factors mentioned above

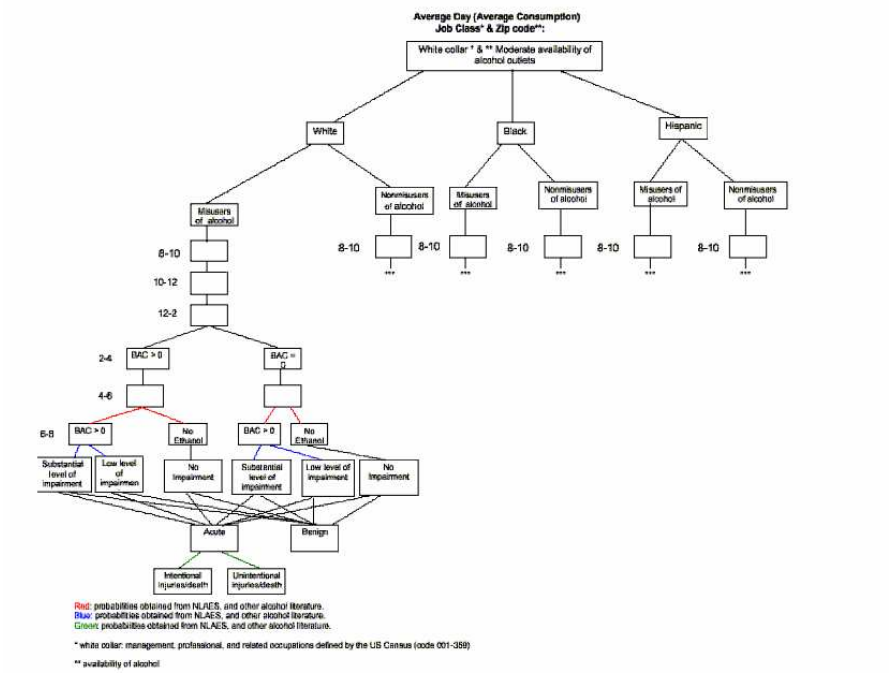

Fig. 4. Directed graph model for a single population class. Note that only one tree is shown; trees exist for each population class.

and quantized them at reasonable levels, say residence location: 50 postal codes in an area, ethnicity: 5 ethnic classes, job class: 3 broad job classes, gender: 2, age: 10 age blocks, socioeconomic status: 10, time-of-day: 12 two-hour blocks, alcohol availability: perhaps 3, and alcohol abuse problem or not: 2, we could be seeking to draw agents from 360,000 categories. And, of course, for each of these 360,000 categories, we would have to know the conditional probability of each of the eight acute events we are attempting to model. If we knew all of these probabilities, modeling could be captured by a relatively simple Bayesian network. Of course, we don't have all of these probabilities, so we turn to the directed graph model. What we are interested in is estimating the multinomial distribution of acute outcomes together with the benign outcome. Ultimately, we are interested in assessing interventions, which will modify the probabilities we use in the directed graph.

For illustrative purposes and admittedly because of data availability, we have chosen to use data from Fairfax County, Virginia, USA. Fairfax County has a population of approximately 1,000,000 individuals living in the 48 postal code areas in the county. A specimen directed graph model is presented in Figure 4. As an approximation, we use a manageable number of population subgroups (classes) that are based on ethnicity (White, African American,

and Hispanic) and job type (based on Census job codes, i.e. blue collar, white collar, and unemployed). For these classes, the probability of being an alcohol misuser (alcohol abuse and alcohol dependence) or nonmisuser (abstainer and drinkers who do not have a diagnosis) are estimated based on data from population surveys for each class (e.g., National Longitudinal Alcohol Epidemiological Survey, NLAES). Note that although the model is based on classes, the stochastic approach simulates outcomes for each individual, meaning that persons in the same class do not necessarily have the same outcome.

We know the population distribution in Fairfax County based on Census data. The simulation begins by choosing an agent at random from one of the 48 postal codes. Because we know the distribution of job class and ethnicity within each postal code also from U.S. Census data and data from the U.S. Bureau of Labor Statistics, we can attach to the selected agent information about the agent's ethnicity and job class. In principle, we could also attach information about age, gender, and socioeconomic status based on Census data, but we will not do so in this illustration. While the NLAES data does not provide information on whether someone is an alcohol misuser or not conditioned on postal code, it does provide this information conditioned on ethnicity and job class. This we can estimate the probability that our agent will be a misuser of alcohol. We also have information from the Virginia Department of Alcoholic Beverage Control (ABC) on the availability of alcohol principally in terms of sales volumes of distilled spirits for each ABC store within each postal code, but also more generically sales tax figures for outlets selling beer and wine. Thus, we can estimate the alcohol availability by postal code. It is known from studies such as Gorman et al. (2001) and Gruenewald et al. (2002) that alcohol availability directly influences drinking behaviors enhancing the probabilities of acute outcomes. The prevalence of acute outcomes is dependent on the nature of the acute outcome, murder and suicide are fairly rare, DWI and assaults are fairly common. However, we have obtained annual aggregate figures in each category for Fairfax County for a period from 1999-2005. Finally, being an alcohol misuser can enhance the probability of engaging in violence by a factor of five or more (Collins and Schlenger, 1988 and Leonard et al., 2003). Thus, except for the time-of-day component, we can estimate the probability that our agent will engage in some acute outcome, although we have no information on which specific acute outcome. However, because we do know the distribution of acute outcomes annually, we can estimate by proportion the probability that our agent for each acute or benign outcome. This process will associate with each agent a likelihood of specific type of acute outcome and simultaneously give us aggregated information for each postal code, job class, ethnic type, and so on. Thus, for purposes of our approximation, we compute the following probability:

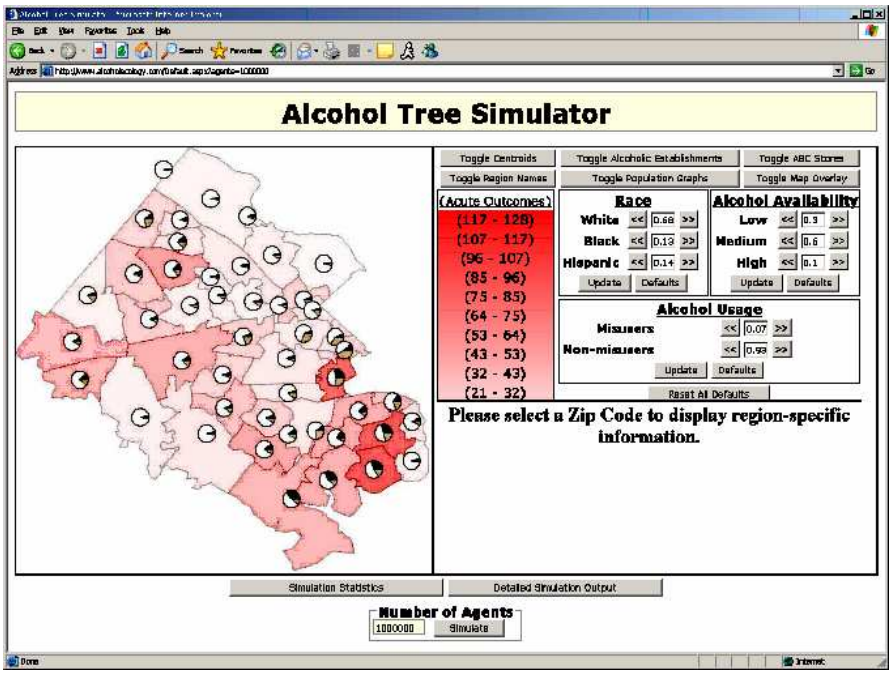

Fig. 5. Example of Software Interface for the Initial Model.

$$P(Acute\ Event_i|Agent_j) =$$

$$P(Z)P(E,J|Z)P(M|E,J)P(A|Z)P(Some\ Acute\ Event|M)P(Acute\ Event_i)+$$

$$P(Z)P(E,J|Z)P(M^c|E,J)P(A|Z)P(Some\ Acute\ Event|M^c)P(Acute\ Event_i),$$

where $P(Z)$ is the probability of choosing agent j from postal code Z, $P(E,J|Z)$ is the conditional probability that the agent is of ethnicity E and job class J given postal code Z, $P(M|E,J)$ is the conditional probability of the agent being a misuser of alcohol given he is of ethnicity E and job class J, $P(A|Z)$ is the probability of alcohol availability A in postal code Z, $P(Some\ Acute\ Event|M)$ is the probability that the agent will create some acute event given that he is a misuser of alcohol, and, finally, $P(Acute\ Event_i)$ is the unconditional probability that Acute Event i occurs (in a given day). Of course, M^c indicates the agent is not a misuser of alcohol.

The model simulates an average day for each postal code and for each ethnicity and job class. Figure 4 shows the tree diagram for only a single postal code; the entire digraph model is composed of similar decision trees for each of the 48 postal code areas separately for low, medium, and high levels of alcohol availability. In the general model, the complexity has increased substantially because population classes will be further split by gender and age

categories. Also, instead of an "average" day, we estimate weekend and non-weekend days separately. Note that Figure 4 also shows the paths through the model (which are stochastic conditional probabilities) leading to drinking (or not drinking) and having a benign outcome or a pathological/negative outcome. The pathways will allow us to also include time-of-day factors where the boxes are blank in the figure although we do not do so in this first approximation. More on time-of-day in Section 5. The model was implemented as a web application as shown in Figure 5. This simulation is available at http://alcoholecology.com. This website presents a JAVA-based implementation and allows the user to explore adjustments in alcohol availability and ethnic balance for the whole county and within individual postal codes.

Outcome Type	Actual/Year	Simulated/Year	Mean	MSE
DWI	722	658	708	19.6
Assault	133	107	132	2.0
Murder	6	4	7	1.0
Sexual Assault	32	38	34	4.1
Domestic Violence	161	168	168	16.1
Child Abuse	221	213	216	17.4
Suicide	97	84	100	2.6
Benign	998993	998728	998635	888.4

Table 1: Actual and Simulated Results for the Model

As can be seen from Table 1, even with fairly simple assumptions, the directed graph model reasonably approximates the actual data. The column labeled "Mean" is calculated by running the simulation 100 times and transforming this estimate to a yearly average value. The "MSE" column represents the error associated with the simulation results when compared with the actual values using the probabilities for the year of the actual values.

4 Conclusions

Alcohol studies have traditionally focused on specific acute outcomes and intervention strategies for mitigating those particular acute effects. Such approaches often have unintended consequences so that intervening to reduce one acute effect may increase other undesirable outcomes. Geospatial and temporal effects have been considered separately, but are not traditionally integrated with intervention models. Statistical analysis in alcohol studies is done often at a relatively elementary level applying relatively elementary hypothesis testing procedures. Indeed, it is difficult to find outlets for modeling innovations within the traditional alcohol literature. In the present paper, we outline a comprehensive modeling framework based on agent-based simulation, social networks, and directed graphs that captures some of the

complexity of the alcohol ecological system. We aggregate agents into relatively homogeneous clusters, but provide diversity by stochastic modeling of outcomes. We provide a framework for temporal and geospatial effects and discuss proxies for approximating these effects. We calibrate the model, again using approximations based on actual data.

Acknowledgements

The work of Dr. Wegman is supported in part by the U.S. Army Research Office under contract W911NF-04-1-0447. The work of Dr. Said is supported in part by Grant Number F32AA015876 from the National Institute On Alcohol Abuse And Alcoholism. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health. Drs. Wegman and Said are Visiting Fellows at the Isaac Newton Institute for Mathematical Sciences at University of Cambridge in Cambridge, United Kingdom. The authors are in debt for the support provided by the Newton Institute, which has made the successful completion of this work possible.

References

BANG-JENSEN, J. and GUTIN, G. (2000): *Digraphs: Theory, Algorithms and Applications*. London, Springer-Verlag .

BOLLOBAS, B. (2001): *Random Graphs*. Cambridge, UK, Cambridge University Press.

BONDY, J.A. and MURTY, U.S.R. (1976): *Graph Theory with Applications*. North Holland, Amsterdam.

CHARTRAND, G. and LESNIAK, L. (1986): *Graphs and Digraphs*. Wadsworth Publishing Co., Belmont, California.

COLLINS, J.J. and SCHLENGER, W.E. (1988): Acute and chronic effects of alcohol use on violence. *Journal of Studies on Alcohol 49(6), 516-521*.

DIESTEL, R. (2005): *Graph Theory: Third Edition*. Springer Verlag, Berlin.

EZATTI, M., LOPEZ, A., RODGERS, A., VANDER HOORN, S., and MURRAY, C. (2002): Comparative risk assessment collaborating group. Selected major risk factors and global and regional burden of disease. *Lancet 360, 1347-1360*.

GORMAN, D.M., SPEER, P.W., GRUENEWALD, P.G., and LABOUVIE, E.W. (2001): Spatial dynamics of alcohol availability, neighborhood structure and violent crime. *Journal of Studies on Alcohol 62, 628-636*.

GRUENEWALD, P.J., REMER, L., and LIPTON, R. (2002): Evaluating the alcohol environment: Community geography and alcohol problems. *Alcohol Research and Health 26*.

LEONARD, K.E., COLLINS, R.L., and QUIGLEY, B.M. (2003). Alcohol consumption and the occurrence and severity of aggression: An event-based analysis of male-to-male barroom violence. *Aggressive Behavior 29, 346-365*.

MARCHETTE, D.J. (2004): *Random Graphs for Statistical Pattern Recognition*. John Wiley and Sons, Hoboken.

NATIONAL INSTITUTE ON ALCOHOL ABUSE AND ALCOHOLISM (1998): *Drinking in the United States: Main findings from the 1992 national longitudinal alcohol epidemiologic survey (NLAES)*. NIH Publication No. 99-3519. U.S. Alcohol Epidemiologic Data Reference Manual. 6.

SAID, Y.H. and WEGMAN, E.J. (2007): Quantitative assessments of alcohol-related outcomes. *Chance 20(3), 17-25.*

Spatial and Computational Models of Alcohol Use and Problems

William F. Wieczorek[1], Yasmin H. Said[2,3], and Edward J. Wegman[2,3]

[1] Center for Health and Social Research
Buffalo State College
State University of New York
Buffalo, NY 14222 USA,
wieczowf@buffalostate.edu
[2] Isaac Newton Institute for Mathematical Sciences
Cambridge University
Cambridge, CB3 0EH UK
[3] Department of Computational and Data Sciences
George Mason University MS 6A2
Fairfax, VA 22030 USA
ysaid99@hotmail.com, ewegman@gmail.com

Abstract. This paper focuses on multivariate and computational approaches that are being developed in the alcohol field. There is substantial monetary support for conducting alcohol research. Alcohol use and problems are complex behaviors by individuals, across their life spans, while embedded in a number of social and economic networks. This complexity, coupled with the research support primarily from the National Institutes of Health (NIH), has led to numerous data collection and research projects, many of which require sophisticated multivariate and spatial statistical approaches. Some of the methods used to model alcohol use and problems are latent growth curves, multilevel models, and latent class analysis. These techniques allow for the examination and modeling of both individual and group level factors. However, these types of models are not suitable for mining large data sets. In this paper, we exploit regional data in Erie County, NY to illustrate the use of multivariate and spatial analysis tools in alcohol studies.

Keywords: GIS, social indicators, public health, interventions, CrystalVision, CCmaps

1 Introduction

Statistical methods are often used in health studies including alcohol studies in order to test hypotheses about health risks. However, these relatively elementary techniques do not exploit the newer methods of multivariate data visualization and spatial statistics. The ability to manipulate multivariate spatial data offers the possibility of extracting additional meaning and suggests, in addition to the confirmatory role for statistical methods, also an

exploratory role. Statistical spatial analysis often begins with spatial analysis using a geographic information systems (GIS). Such systems allow the analysis of distance and connectivity including the measures of distances between points and between points and centroids, analysis of adjacency, analysis of networks including roads and other transportation systems, and analysis of buffer areas between otherwise adjacent areas. Spatial analysis of this sort can give insight into effective distances which may be substantially different from apparent Euclidean distances.

Key components of geographic information systems in spatial analysis are the ability to access and enhance spatial data and the ability to present these relationships in terms of maps, graphics, and data files. Spatial dependencies define the relationships among spatially diverse entities, including non-random patterns in geographic space, clusters, dispersion, and spatial autocorrelation. Spatial factors are integral to the development of alcohol simulation models such as those presented in Wegman and Said (2008) and Said (2008). Spatial analysis contributes to hypothesis generation, spatial epidemiology, multi-level/multi-resolution modeling, spatial interaction and travel models, and understanding spatial processes in small areas. The latter capability allows the development and testing of psychosocial models, especially with respect to spatial interactions among alcohol and drug users.

2 Risk factors and social indicators

Traditionally, health studies including alcohol studies collect data by surveys, which provide data at the individual level. However, it is not always possible to collect data at the individual level because of cost, privacy, or lack of resources. In many situations it is impractical or impossible to measure a specific outcome such as early drinking, adolescent drug use, or alcohol dependence. In contrast, information may be easily available on factors associated with these phenomena such as poverty, immigration status, language facility, and alcohol availability. These are examples of social indicators. Social indicators are numerical data, usually archival in nature, that measure the well-being of a population. There are frequently issues of data quality including reliability and validity. Is the indicator a stable measure? Is the indicator actually related to the phenomenon of interest? The advantage of using social indicator data include the use of substantial amounts of administratively available data, the ability to make data-driven decisions on topics that are impractical to measure directly, and the fact that specific indicators have conceptual and evidential relationships to difficult-to-measure outcomes. Disadvantages of using social indicator data include the fact that data are collected for purposes other than their use as indicators, hence, may not have statistical validity, that there are few direct indicators (relationships of indicator to outcome are indirect), that there are usually few indicators at local geographic level (postal code or census tract, most are at county,

state, or national levels), and that there are a huge number of indicators from which to select many of which may be overlapping and collinear. Social indicators provide an indirect method of needs assessment for public health services. They show relative need for services and may be used to estimate actual need for services in some situations. In addition, coupled with demographic information, social indicator analysis allows for tailoring services to population characteristics.

Indicators can fall into a number of categories including neighborhood indicators, family indicators, and individual indicators. Neighborhood indicators would include availability of drugs and firearms, community attitudes toward laws and social norms, attitudes favorable to drug use, firearms and crime, state of transition and mobility within the neighborhood, levels of neighborhood attachment, and community disorganization. Family-level indicators include extreme economic privation, family history of problem behaviors, family management problems, family conflict, and lack of commitment to schools. At the level of individual indicators, indicators include alienation and rebeliousness, early academic failure, substance abuse, delinquency, lack of parental involvement in problem behaviors, and teen pregnancy. This categorization of indicators is aligned with the well-known risk and protective factor model presented by Hawkins and colleagues (1992).

3 Erie County risk indicators

Wieczorek and Delmerico (2005) assembled a database of risk indicators for Erie County, NY using several sources. Erie County includes the city of Buffalo, New York. This database provides a data-rich snapshot of a relatively small county-level geographic area. The sources include the U.S. Census 2000, the New York State Education Department, and the New York State Department of Criminal Justice Services. At the local level, sources include the Center for Health and Social Research, the City of Buffalo Police Department, the Erie County Board of Elections, the Erie County Department of Health, the Erie County Department of Mental Health, and the Roswell Park Cancer Institute. Indicators are calculated at two spatial levels: census tract and five-digit zip code area. An indicator may be available at both census tract and zip code levels or just one level as dictated by the availability of data. Most indicators are available at the zip code level, although for a few, there are just no data available for areas smaller than the entire county. Even when an indicator is available, not every tract or zip code record will have an associated value. For some, the value will be missing. Common reasons for missing data are data availability and small populations.

Because all indicators are essentially ratios of the form: cases/population (expressed as percent or per 10,000), it is important to avoid unreliable indicator values due to small populations. For this reason, an arbitrary threshold of population greater than 100 was set. Records for zip codes and tracts with

populations below 100 have been removed from the database. Sometimes the source data for calculation of the indicators were available at a spatial level other than census tract or zip code area. In these cases, risk indicators were first calculated at the available level, and then imputed to the zip level. The imputation was performed using a population-based weighting method. Three imputation schemes were utilized in calculating the risk indicators:

(a) From school districts to zip code areas. This scheme was used to transfer data collected for school districts, e.g., performance on mathematics tests, dropouts, to zip code areas and to calculate corresponding risk indicators.

(b) From police departments' areas of responsibility to zip code areas. Crime statistics obtained from New York State Department of Criminal Justice Services (DCJS) are tabulated by law enforcement agencies in Erie County. Areas served by each law enforcement agency (usually a town or an incorporated place) were delineated and data were interpolated to zip code areas for ease of use and for compatibility with crime data from Buffalo Police Department.

(c) From 1994 census tracts to zip code areas (current, i.e. from census 2000). This scheme was needed to transfer crime data from Buffalo Police Department to zip code areas. After the interpolation to zip code areas, Buffalo Police Department data were integrated into DCJS crime data to provide better spatial detail of crime within Buffalo.

4 Multivariate analysis

Alcohol use and abuse can be thought of in terms of both a cause and an effect. Alcohol use and abuse is a cause insofar as it leads to acute outcomes such as DWI/DUI, DWI with fatal crashes, assault, domestic violence, child abuse, sexual assault, murder, suicide as well as chronic outcomes such as cirrhosis of the liver and other alcohol-induced diseases. See Ezzati et al. (2002), Room et al. (2005), and Said and Wegman (2007). Some social indicators for these outcomes from the Erie County Risk Indicators Database include *crm.dwi* (DWI crime), *de.traffic* (fatal crash deaths), *crm.viol* (violent crime), *de.trauma* (trauma deaths), *jar.viol* (juvenile crime), *crm.drug* (drug-related crimes), *de.suicide* (suicide deaths), and *de.cirrohsis* (cirrhosis deaths). Conversely, alcohol use and abuse can be thought of as being caused by poverty, marital unhappiness, poor education, drug and alcohol availability, neighborhood factors, parental alcoholism, and other demographic factors. Some social indicators in the Erie County Risk Indicators Database include *fam.pov* (family poverty), *med.income* (median income), *unem* (unemployment), *divorce* (divorce rates), *nv.married* (never married), *edu.g8* (education below 8th grade level), *edu.col.d* (educated beyond college), *dropout* (dropout rates), *alc.all* (all alcohol outlets), *alc.off* (off license outlets, i.e., stores that sell alcohol to be consumed elsewhere), *tobacco* (tobacco outlets), *vacant* (neighborhood vacancies), *vote.gen* (general voting registrations) and *poor.eng* (poor house-

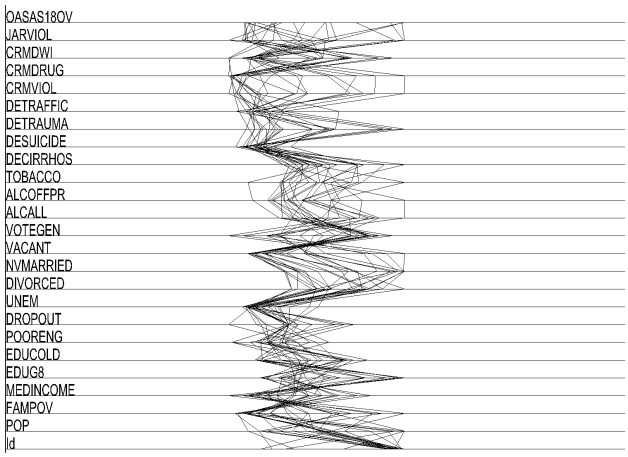

Fig. 1. Parallel coordinate display of 25 variables associated with alcohol use in Erie County, NY. This figure shows postal codes with high levels of admissions to treatment for alcohol and drugs for individuals over 18 years old. See the text for interpretation.

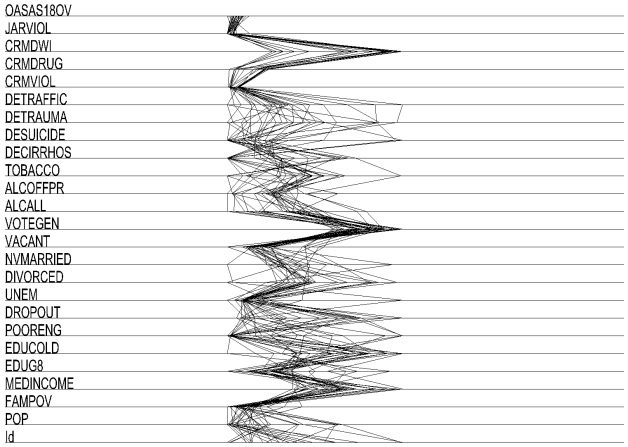

Fig. 2. Parallel coordinate display of 25 variables associated with alcohol use in Erie County, NY. This figure shows postal codes with low levels of admissions to treatment for alcohol and drugs for individuals over 18 years old. See the text for interpretation.

hold English usage rates). An indicator of overall alcohol problems for Erie County is the rate of admissions to treatment for alcoholism and substance abuse. The appropriate indicator is *oasas.18ov*, which is the rate per 10,000 by zip code for individuals over the age of 18. We use this as the primary indicator of severe alcohol use problems.

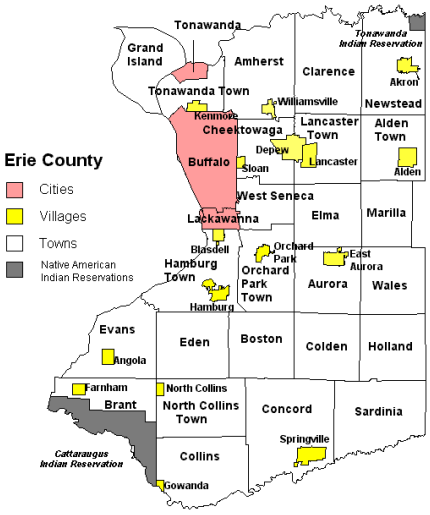

Fig. 3. Map of Erie County, NY showing major subdivisions.

Figures 1 and 2 are parallel coordinate displays of the 25-dimensional multivariate indicator data. Parallel coordinate multivariate representations have become a standard way of visualizing multivariate data. See Inselberg (1985), Wegman (1990, 1998, 2003), Wegman et al. (1993), and Wegman and Solka (2002). See especially Wegman (1990, 2003) for interpretation of these displays as a exploratory data analysis tool. In Figures 1 and 2, we have provided a parallel coordinate display keyed on the *oasas.18ov* indicator, the rate of admissions by zip code of individuals over 18. For zip codes with high rates of alcohol and drug related admissions are shown in Figure 1, while those with lower rates of admission are shown in Figure 2. This gives a very clear picture of associations. High admission rates are associated with high levels of drug-related crimes, violent crime, trauma deaths, cirrhosis deaths, high numbers of tobacco outlets, high number of alcohol off-license outlets, high number of all alcohol outlets, high fraction of vacant houses, high percentage of never married, high percentage of divorces, high levels of unemployment, high percentage of population educated to the 8th grade or less, and high levels of family poverty. Conversely, high admission rates are associated with low levels of DWI arrests, low levels of voter registrations, low percentages of college educated population, and low levels of median income. None of these is surprising except possible the low levels associated with DWI arrests. In view of the association with poverty in Erie County, one suspects that the low correlation between alcohol and drug admissions and DWI arrests is because poor people are **less likely to drive frequently or to have automobiles**. Spatial analysis will give additional insight into this situation.

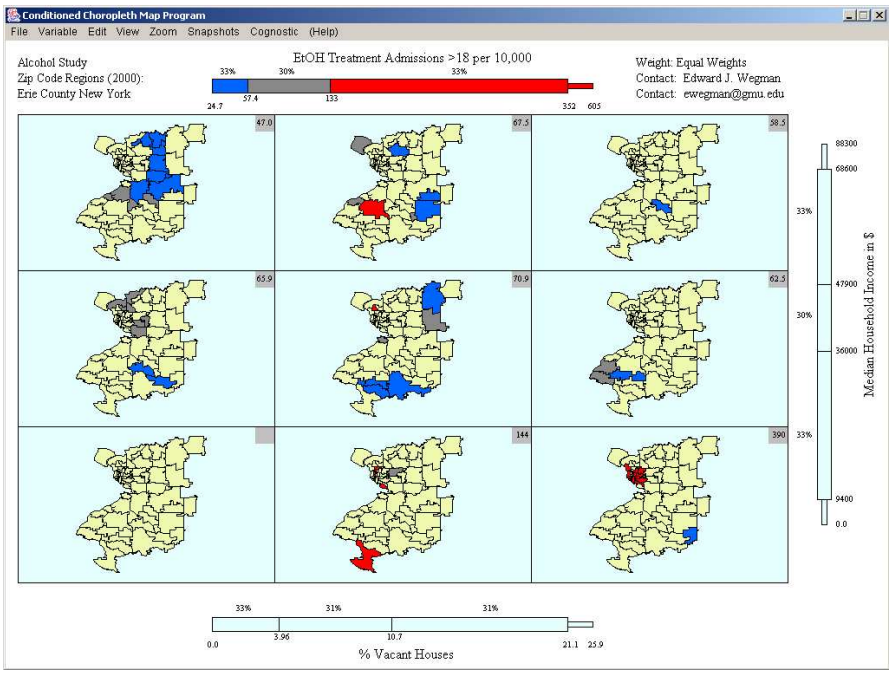

Fig. 4. CCmap display with *oasas.18ov* as the response variable and *vacant* and *med.income* as the independent predictor variables. For CCmaps, it is conventional to start with the sliders adjusted so that about 1/3 is in the low range, 1/3 in the middle range, and 1/3 in the high range. Further adjustments can isolate extremes.

5 Spatial analysis using CCmaps

Exploratory multivariate data analysis is suggestive of some further spatial exploration. We have hypothesized that alcohol use and abuse has its roots in poverty and related social indicators. Conversely, alcohol causes other adverse societal effects. Indeed, causality is uncertain in many situations. Do neighborhoods degenerate because of alcohol and drug availability or do degenerating neighborhoods provide an easy environment for drugs and alcohol to flourish? Do people commit crimes because of alcohol and drugs or does a criminal life-style encourage use of alcohol and drugs? In any case, the multivariate visualization in Figures 1 and 2 strongly suggests there is an association. We include Figure 3, which is a map representation of Erie County for the spatial orientation of the reader.

A spatial analysis tool that allows us to dynamically explore the relationships among these indicator variables is the conditioned choropleth maps (CCmaps). See Carr et al. (2000, 2002, 2004, 2005) and Carr (2005). The CCmaps software allows dynamic exploration of spatially-oriented data. The software essentially supports a five-dimensional view of the data with two of

the dimensions being geospatial and the other three being in our case indicator variables. The indicator variables are essentially controlled by sliders so that proportions may be dynamically adjusted. There are sliders on top, to the right, and at the bottom of the display. The sliders are divided into three regions, perhaps best thought of as high, medium, and low. The bottom slider divides the display into three vertical regions. The right slider divides the display into three horizontal regions. Thus in the CCmaps display, there are nine panels. See Figure 4-6 for examples. The top slider is color-coded into blue, gray, and red and these color codes show up in the map displays. Conceptually, the top slider controls a response variable while the right and bottom sliders control independent predictor variables. Due to the reproduction of the paper in gray scale, these color nuances are not available in the printed version.

In Figure 4, we use Vacant Houses and Median Household Income as the independent variables and the EtOH Treatment Admissions > 18 per 10,000 as the response variable. EtOH is short-hand for ethanol or ethyl alcohol. This plot shows alcohol problems as a function of neighborhood characteristics. The bottom right-hand panel shows a hot spot of alcohol problems in Buffalo corresponding to neighborhoods that have a high percentage of vacant houses and low median household income. Conversely, the upper left-hand panel shows a more suburban region with low levels of alcohol problems corresponding to low levels of vacant houses and relatively high levels of income. By examining all nine panels, one can isolate spatial regions with high and with low levels of alcohol problems.

In Figure 5, we replace the Median Household Income variable with an alcohol availability variable. Again in the upper right panel, we see that there is a hotspot in Buffalo for alcohol admissions associated with high levels of housing vacancy and high levels of alcohol availability. As in Figure 4, the lower left panel shows a spatial cluster of suburban locations with low levels of alcohol problems associated with low levels of vacant houses and low levels of alcohol availability. Figures 4 and 5 are very suggestive that poverty, poor neighborhood infrastructure and high alcohol availability are associated with high levels of alcohol problems as reflected by alcohol and drug treatment admissions. The spatial analysis allows us to pinpoint problem areas where interventions may make sense, as well as identify areas where interventions may be less necessary.

In Figure 6, we choose alcohol and drug treatment over 18 and alcohol off license availability as independent predictor variables and examine the violent crimes variable as the response variable. Figures 4 and 5 focused on environmental variables as predictors of alcohol problems. In Figure 6, we change the perspective somewhat, and examine crime as a response to alcohol problems. As before, we see a high crime area in the upper right panel associated with high levels of alcohol availability and high levels of individuals over 18 in treatment for alcohol and drug problems. Of course,

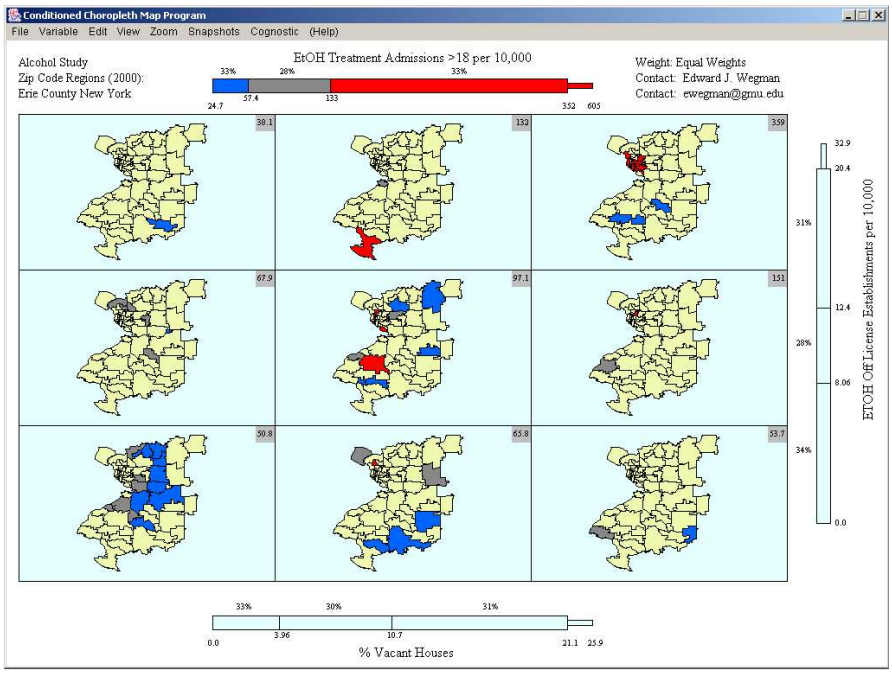

Fig. 5. CCmap display with *oasas.18ov* as the response variable and *vacant* and *alc.off* as the independent predictor variables. In both Figures 4 and 5, we examine alcohol problems as a function of neighborhood quality.

the usual statistical caveat of association does not imply causation, but such plots are suggestive. Please note that the problem areas highlighted in Figure 4, 5, and 6 are not identical.

Finally in Figure 7, we consider the same independent predictor variables as in Figure 6, but consider DWI arrests as the response variable. In this case, we have a startling difference. In what was a problem area before in the upper right panel, we see that Buffalo is seemingly on the low end of DWI arrests, while the previously non-problem areas in the more rural areas of Erie County are now the apparent location of problem areas. It is clear that the type of interventions required for distinct alcohol-related problems are highly dependent on the spatial environments where the problems take place and that the geospatial analysis presented here reinforce this point.

6 Conclusions

Wegman and Said (2008) and Said (2008) make the point that proposed interventions for alcohol-related problems are highly dependent on multiple factors including spatial and temporal analysis and that interventions not carefully thought out may have unintended consequences. The analysis in this

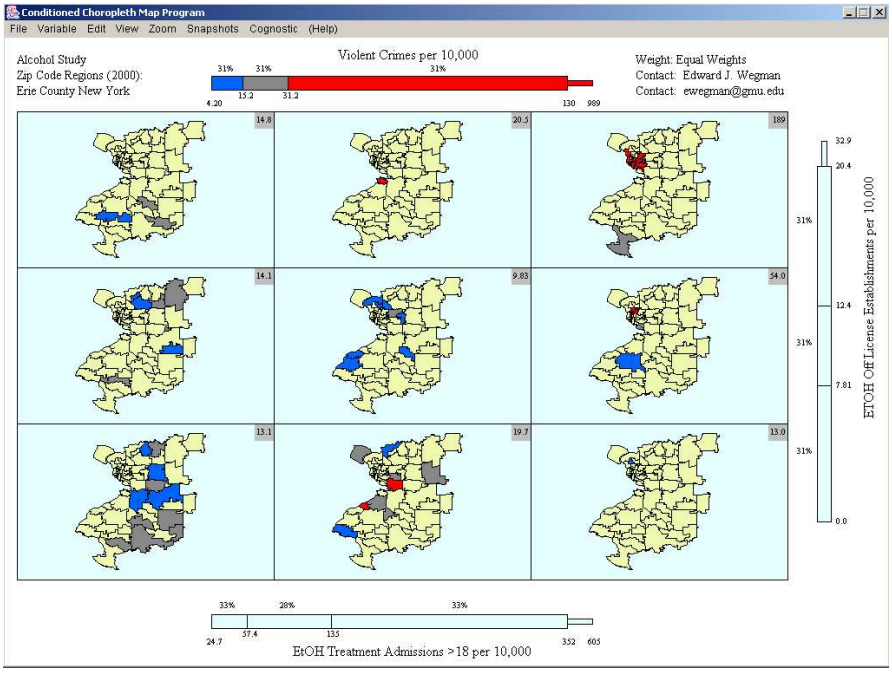

Fig. 6. In this figure we examine violent crime as a function of alcohol problems. Here *crm.viol* is the response variable and *oasas.18ov* and *alc.off* are the independent predictor variables.

paper focuses strongly on the multivariate spatial visualization and, by doing so, strongly reinforces the points made by Wegman and Said. Although they consider Fairfax County, VA and, in this paper, we consider Erie County, NY the concepts presented are general in nature. Indeed, their geospatial analyses also show relative hotspots as well. The data available for Erie County, NY have a much larger collection of variables, hence, a more subtle analysis can be made. Interestingly enough, there are common elements as well as differences. Both locations have large university communities. Fairfax County is a largely suburban community while Erie County contains a major city, with suburbs. Fairfax County has a large military presence, which is not the case with Erie County. The differences only highlight the commonality of the approaches.

The results of the analyses presented in this paper also have substantive implications for more typical statistical models in alcohol research that use multiple levels of data (e.g., data from individuals and indicator data at a group level). The results of the CCmaps analyses suggest that it may be possible to use weighted combinations of group level variables to explore interactions effects at the group level. This could prove to be more enlightening in multi-level models than using only simple multiplicative interactions and should be explored in future research.

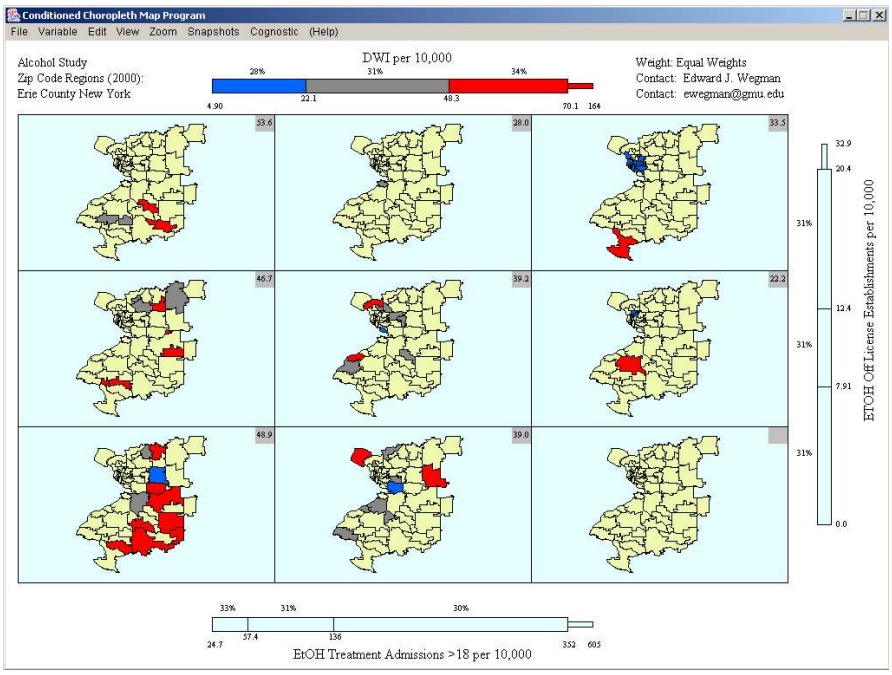

Fig. 7. Here we examine DWI arrests as a function of alcohol problems. In this figure, *crm.dwi* is the response variable and *oasas.18ov* and *alc.off* are the independent predictor variables.

Acknowledgements

The work of Dr. Wieczorek is supported in part by grant number R01AA016161 from the National Institute on Alcohol Abuse and Alcoholism and by a contract from Western New York United Against Alcohol and Drug Abuse/Erie County Department of Mental Health. The work of Dr. Said is supported in part by grant number F32AA015876 from the National Institute on Alcohol Abuse and Alcoholism. The work of Dr. Wegman is supported in part by the U.S. Army Research Office under contract W911NF-04-1-0447. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Alcohol Abuse and Alcoholism or the National Institutes of Health. Drs. Wegman and Said are Visiting Fellows at the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge in Cambridge, England. They are indebted for the support provided by the Newton Institute, which has made the successful completion of this work possible.

References

CARR, D.B. (2005): Some recent graphics templates and software for showing statistical summaries. In: C.R. Rao, E.J. Wegman and J.L. SOLKA (Eds.): *Handbook of Statistics 24: Data Mining and Data Visualization*, 415-436.

CARR, D.B., CHEN, J., BELL, B.S., PICKLE, L., and ZHANG, Y. (2002): Interactive linked micromap plots and dynamically conditioned choropleth maps. *ACM Proceedings of the 2002 Annual National Conference on Digital Government Research 1-7.*

CARR, D.B., WALLIN, J.F., and CARR, D.A. (2000): Two new templates for epidemiology applications: linked micromap plots and conditioned choropleth maps. *Statistics in Medicine 19(17-18), 2521-2538.*

CARR, D.B., WHITE, D., and MACEACHREN, A.M. (2005): Conditioned choropleth maps and hypothesis generation. *Annals of the Association of American Geographers 95(1), 32-53.* doi:10.1111/j.1467-8306.2005.00449.

CARR, D.B., ZHANG, Y., LI, Y., and ZHANG, C. (2004): Dynamic conditioned choropleth maps: Examples and innovations. *ACM Proceedings of the 2004 Annual National Conference on Digital Government Research 1-2.*

EZATTI, M., LOPEZ, A., RODGERS, A., VANDER HOORN, S., and MURRAY, C. (2002): Comparative risk assessment collaborating group. Selected major risk factors and global and regional burden of disease. *Lancet 360, 1347-1360.*

HAWKINS, J.D., CATALANO, R. F., and MILLER, J. Y. (1992): Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. *Psychological Bulletin 112(1), 64-105.*

INSELBERG, A. (1985): The plane with parallel coordinates. *The Visual Computer 1(4), 69-91.* DOI 10.1007/BF01898350.

ROOM, R., BABOR, T., and REHM, J. (2005): Alcohol and public health. *The Lancet 365, 519-530.*

SAID, Y.H. (2008): Estimating spatiotemporal effects for ecological alcohol systems. This Volume.

SAID, Y.H. and WEGMAN, E.J. (2007): Quantitative assessments of alcohol-related outcomes. *Chance 20(3), 17-25.*

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

WEGMAN, E.J. (1998): Parallel coordinate and parallel coordinate density plots. *Encyclopedia of Statistical Sciences, Update Volume 2.* S. KOTZ, C.B., READ and D.L. BANKS(Eds.), 518-525 + color plates.

WEGMAN, E.J. (2003): Visual data mining. *Statistics in Medicine 22, 1383-1397 + 10 color plates.*

WEGMAN, E.J., CARR, D.B., and LUO, Q. (1993): Visualizing multivariate data. In: C.R. Rao (Ed.): *Multivariate Analysis: Future Directions.* North Holland, Amsterdam 423-466.

WEGMAN, E.J. and SAID, Y.H. (2008): A directed graph model of ecological alcohol systems incorporating spatiotemporal effects. This Volume.

WEGMAN, E.J. and SOLKA, J.L. (2002): On some mathematics for visualizing high dimensional data. *Sanhkya (A) 64(2), 429-452.*

WIECZOREK, W.F. and DELMERICO, A. (2005): *Planning for Prevention: The Erie County Risk Indicator Database.* CHSR, Buffalo, New York.

Part VII

Finance and Insurance

Optimal Investment for an Insurer with Multiple Risky Assets Under Mean-Variance Criterion

Junna Bi[1] and Junyi Guo[2]

[1] School of Mathematical Sciences, Nankai University
Tianjin 300071, China, *bijunna@mail.nankai.edu.cn*
[2] School of Mathematical Sciences, Nankai University
Tianjin 300071, China, *jyguo@nankai.edu.cn*

Abstract. This paper considers the optimal investment strategy for an insurer under the criterion of mean-variance. The risk process is a compound Poisson process and the insurer can invest in a risk-free asset and multiple risky assets. We obtain the optimal investment policy using the stochastic liner-quadrant (LQ) control theory. Then the efficient strategy (optimal investment strategy) and efficient frontier are derived explicitly by a verification theorem with the classical solution of Hamilton-Jacobi-Bellman (HJB) equation.

Keywords: M-V portfolio selection, optimal investment, efficient frontier

1 Introduction

The problem of optimal investment for an insurer has attracted more and more attention since the work of Browne (1995), where the risk process was approximated by a Brownian motion with drift and the stock price process was modeled by a geometric Brownian motion. The expected constant absolute risk aversion (CARA) utility from the terminal wealth was maximized. Browne (1995) also showed that the target of minimizing the ruin probability and the target of maximizing the exponential utility of the reserve produce the same type of strategy when the interest rate of the risk-free asset is zero. In Hipp and Plum (2000), the risk process was described by the classical Cramer-Lundberg model and the insurer can invest in a risky asset to minimize the ruin probability. This set-up was adopted by most of the works on this subject since 2000. Later, there were some papers considering the same optimal investment problem, such as Gaier et al. (2003), Hipp and Schmidli (2004), Yang and Zhang (2005), Wang (2007), and Wang, Xia and Zhang (2007). In Wang (2007), the claim process was supposed to be a pure jump process and the insurer had the option of investing in multiple risky assets. In the sense of maximizing the exponential utility of the terminal wealth, the optimal strategy was to put a fixed amount of money in each risky asset if

there was no risk-free asset.

The mean-variance portfolio selection problem was firstly proposed by Markowitz (1952). From then on, it became a rather popular criterion to measure the risk in finance theory; see Merton (1972), Zhou and Li (2000), and Li, Zhou and Lim (2002) etc. Recently, Wang, Xia and Zhang (2007) pointed out that the mean-variance problem was also of interest in insurance applications. They considered the optimal investment problem under the criterion of mean-variance using a martingale approach and the insurer could invest in a risk-free asset and a risky asset.

In this paper, we use stochastic LQ control as the framework for studying the mean-variance portfolio selection problem. Compared with Wang (2007), the criterion is mean-variance. Wang, Xia and Zhang (2007) considered one risky asset and they used martingale approach, here we consider multiple risky assets and use the stochastic LQ control theory. We give a new verification theorem of our jump-diffusion model, and then the optimal value function and optimal strategy are derived explicitly.

2 Problem formulation

Let $(\Omega, \mathcal{F}, P)$ be a probability space with filtration $\{\mathcal{F}_t\}$ containing all objects defined in the following.

The risk process $(R(t))$ of the insurer is modeled by

$$dR(t) = cdt - d\sum_{i=1}^{N(t)} Y_i, \quad R(0) = u, \tag{1}$$

where constant c is the premium rate. $\{N(t)\}$ is a Poisson process with intensity $\lambda > 0$ and $N(t)$ represents the number of claims occurring in time interval $[0, t]$. Y_i is the size of the ith claim and Y_i is independent of $\{N(t)\}$. Thus the compound Poisson process $\sum_{i=1}^{N(t)} Y_i$ represents the cumulative amount of claims in time interval $[0, t]$. The claims' sizes $Y = \{Y_i, i \geq 1\}$ are assumed to be an i.i.d sequence with a common distribution function F. The ith claim occurs at time T_i. The expectation of Y is $EY = \mu_1 > 0$ and the second moment of Y is $E(Y^2) = \mu_2 > 0$. The risk process defined in equation (1), from the perspective of the insurer is really a pay-off process associated with the (insurance) contracts he (or she) has entered.

Suppose that the insurer is allowed to invest all of his (or her) wealth in a financial market consisting of one risk-free asset (bond) and m risky assets (stocks). We consider the financial market where $m + 1$ assets are traded continuously on a finite time horizon $[0, T]$. The price of the bond is given by

$$\begin{cases} dP_0(t) = r(t)P_0(t)dt, & t \in [0, T], \\ P_0(0) = p_0, \end{cases} \tag{2}$$

where $r(t)(>0)$ is the interest rate of the bond.

The prices of the stocks are modeled by the following stochastic differential equations

$$\begin{cases} dP_i(t) = P_i(t)[b_i(t)dt + \sum_{j=1}^{m} \sigma_{ij}(t)dW^j(t)], & t \in [0,T], \\ P_i(0) = p_i, & i = 1,2,\cdots,m, \end{cases} \tag{3}$$

where $b_i(t)(> r(t))$ is the appreciation rate and $\sigma_{ij}(t)$ is the volatility coefficient. We denote $\sigma(t) := (\sigma_{ij}(t))$. $W(t) := (W^1(t), W^2(t), \cdots, W^m(t))'$ is a standard $\{\mathcal{F}_t\}_{t\geq 0}-$adapted $m-$dimensional Brownian motion. The sign $'$ here means the transposition. We assume that $r(t)$, $b(t)$ and $\sigma(t)$ are deterministic, Borel-measurable and bounded on $[0,T]$. In addition, we assume that the non-degeneracy condition

$$\sigma(t)\sigma(t)' \geq \delta I, \quad \forall t \in [0,T], \tag{4}$$

where $\delta > 0$ is a given constant, is satisfied.

Let $u_i(t), i = 0,1,\cdots,m$ denote the total market value of the agent's wealth in the ith bond/stock, and $u_0(t) + u_1(t) + \cdots + u_m(t) = X(t)$. We call $u(t) := (u_1(t), u_2(t), \cdots, u_m(t))'$ a strategy. Thus, the resulting surplus process $X(t)$ is given by

$$\begin{cases} dX(t) = [r(t)X(t) + B(t)'u(t) + c]dt + u(t)'\sigma(t)dW(t) - d\sum_{i=1}^{N(t)} Y_i \\ X(0) = X_0, \end{cases} \tag{5}$$

where $B(t) := (b_1(t) - r(t), \cdots, b_m(t) - r(t))' \in \mathbb{R}_+^m$.

A strategy $u(t)$ is said to be admissible if $u(t)$ is $\mathcal{F}_t$-progressively measurable, and satisfies $E\int_0^T (u_1^2(t) + u_2^2(t) + \cdots + u_m^2(t))dt < +\infty$. We denote the set of all admissible strategies by Π.

Let $X^u(T)$ denote the terminal wealth when the strategy $u(\cdot)$ is applied. Then the problem of mean-variance portfolio choice is to maximize the expected terminal wealth $E[X^u(T)]$ and, in the meantime, to minimize the variance of the terminal wealth $\mathrm{Var[X^u(T)]}$ over $u(\cdot) \in \Pi$. This is a multi-objective optimization problem with two conflicting criteria. The trading strategy $u^* \in \Pi$ is said to be mean-variance efficient if there does not exist a strategy $u \in \Pi$ such that

$$E[X^u(T)] \geq E[X^{u^*}(T)] \quad \text{and} \quad \mathrm{Var[X^u(T)]} \leq \mathrm{Var[X^{u^*}(T)]}$$

with at least one inequality holding strictly. In this case, we call $(\mathrm{Var[X^{u^*}(T)]}, \mathrm{E[X^{u^*}(T)]}) \in \mathbb{R}^2$ an efficient point. The set of all efficient points is called the efficient frontier.

We firstly consider the problem of finding an admissible investment policy such that the expected terminal wealth satisfies $EX^u(T) = d$, where d is a constant, while the risk measured by the variance of the terminal wealth

$$\mathrm{Var[X^u(T)]} = \mathrm{E\{X^u(T) - E[X^u(T)]\}^2} = \mathrm{E\{[X^u(T) - d]^2\}}$$

is minimized. We impose throughout this paper the following assumption

Assumption 1 *The value of the expected terminal wealth d satisfies*

$$d \geq X_0 e^{\int_0^T r(s)ds} + (c - \lambda\mu_1) \int_0^T e^{\int_v^T r(s)ds} dv.$$

Assumption 1 states that the investor's expected terminal wealth d can not be less than $X_0 e^{\int_0^T r(s)ds} + (c - \lambda\mu_1) \int_0^T e^{\int_v^T r(s)ds} dv$, which coincides with the expected amount that he/she would earn if all of the initial wealth is invested in the bond for the entire investment period. Clearly, this is a reasonable assumption, for the solution of the problem under $d < X_0 e^{\int_0^T r(s)ds} + (c - \lambda\mu_1) \int_0^T e^{\int_v^T r(s)ds} dv$ is foolish for rational investors. We will discuss this assumption additionally later in this paper.

Definition 1. The above variance minimizing problem can be formulated as the following optimization problem:

$$\begin{aligned} &\min \quad \mathrm{Var[X^u(T)]} = \mathrm{E[X^u(T) - d]^2} \\ &\text{subject} \quad \text{to} \quad \begin{cases} EX^u(T) = d \\ u \in \Pi \\ (X(\cdot), u(\cdot)) \quad \text{satisfy} \quad (5). \end{cases} \end{aligned} \tag{6}$$

The optimal portfolio for this problem (corresponding to a fixed d) is called a variance minimizing portfolio, and the set of all points $(\mathrm{VarX^*(T)}, \mathrm{d})$, where $\mathrm{VarX^*(T)}$ denotes the optimal value of (6) corresponding to d and d runs over $[X_0 e^{\int_0^T r(s)ds} + (c - \lambda\mu_1) \int_0^T e^{\int_v^T r(s)ds} dv, +\infty)$, is called the variance minimizing frontier.

Since (6) is a convex optimization problem, the equality constraint $EX^u(T) = d$ can be dealt with by introducing a Lagrange multiplier $\beta \in \mathbb{R}$. In this way, problem (6) can be solved via the following optimal stochastic control problem(for every fixed β)

$$\begin{aligned} &\min \quad \mathrm{E\{[X^u(T) - d]^2 + 2\beta[EX^u(T) - d]\}}, \\ &\text{subject} \quad \text{to} \quad \begin{cases} u \in \Pi \\ (X(\cdot), u(\cdot)) \quad \text{satisfy} \quad (5), \end{cases} \end{aligned} \tag{7}$$

where the factor 2 in the front of β is introduced in the objective function just for convenience. Clearly, this problem is equivalent to (letting $b = d - \beta$)

$$\begin{aligned} &\min \quad \mathrm{E\{[X^u(T) - b]^2\}}, \\ &\text{subject} \quad \text{to} \quad \begin{cases} u \in \Pi \\ (X(\cdot), u(\cdot)) \quad \text{satisfy} \quad (5), \end{cases} \end{aligned} \tag{8}$$

in the sense that the two problems have exactly the same optimal control.

3 Solution to an auxiliary stochastic LQ-problem

Problem (8) formulated in the previous section is a stochastic optimal LQ control problem. In the following, we will show how to solve this problem with the help of the HJB equation.

Firstly, we need to solve a auxiliary problem. We consider the following controlled linear stochastic differential equation:

$$\begin{cases} dx(t) = [r(t)x(t) + B(t)'u(t) + c(t)]dt + u(t)'\sigma(t)dW(t) - d\sum_{i=1}^{N(t)} Y_i \\ x(0) = x_0, \end{cases} \tag{9}$$

and the problem:

$$\begin{aligned} &\min \quad \mathrm{E}\{\frac{1}{2}[\mathrm{x(T)}]^2\}, \\ &\text{subject} \quad \text{to} \quad \begin{cases} u \in \mathit{\Pi} \\ (x(\cdot), u(\cdot)) \quad \text{satisfy} \quad (9). \end{cases} \end{aligned} \tag{10}$$

Note that if we set $x(t) = X(t) - (d - \beta)$, $c(t) = c + (d - \beta)r(t)$ and $X(0) = x(0) + (d - \beta)$, (9) is equivalent to (5).

We define the optimal value function by

$$J(t, x) = \inf_{u \in \mathit{\Pi}} E\{\frac{1}{2}[x(T)]^2 | x(t) = x\}. \tag{11}$$

Before starting, we recall the following lemma.

Lemma 1. *Let h be a continuous, strictly convex quadratic functions, $h(u) := \frac{1}{2}u'DD'u + \alpha B'u$ over $u \in \mathbb{R}$, where $B \in \mathbb{R}^m$, and $D \in \mathbb{R}^{m\times m}$. Then h has a minimizer $u^* = -\alpha(DD')^{-1}B$ and*

$$h(u^*) = h(-\alpha(DD')^{-1}B) = -\frac{1}{2}\alpha^2 B'(DD')^{-1}B.$$

Now we study the corresponding HJB equation of problem (9)-(10), which is the following partial differential equation:

$$\begin{cases} \inf_u \{v_x(t,x)[r(t)x + B(t)'u + c(t)] + \frac{1}{2}v_{xx}(t,x)u'\sigma(t)\sigma(t)'u\} \\ \qquad + v_t(t,x) + \lambda E[v(t, x - Y) - v(t,x)] = 0 \\ v(T,x) = \frac{1}{2}x^2. \end{cases} \tag{12}$$

Here $\sigma(t)'$ means the transposition of $\sigma(t)$ and $v_t(t,x)$ means the partial derivative of $v(t,x)$.

Suppose that (12) has a solution which has the following form

$$V(t,x) = \frac{1}{2}P(t)x^2 + Q(t)x + R(t). \tag{13}$$

The boundary condition in (12) implies that $P(T) = 1$, $Q(T) = 0$, and $R(T) = 0$. Inserting this trivial solution into (12) and rearranging, we have

$$\begin{aligned}&\inf_u\{\frac{1}{2}u'\sigma(t)\sigma(t)'u + B(t)'u[x + \frac{Q(t)}{P(t)}]\}P(t) + [\frac{1}{2}P_t(t) + P(t)r(t)]x^2\\ &+ [Q_t(t) + Q(t)r(t) + P(t)c(t) - \lambda\mu_1 P(t)]x + R_t(t) + Q(t)c(t)\\ &- \lambda\mu_1 Q(t) + \frac{1}{2}\lambda\mu_2 P(t) = 0.\end{aligned} \tag{14}$$

Using Lemma 1 by letting $\alpha = x + \frac{Q(t)}{P(t)}$, $u = u$, $D = \sigma(t)$, and $B = B(t)$, then $u^*(t) = -[x + \frac{Q(t)}{P(t)}]\Sigma(t)$, where $\Sigma(t) = (\sigma(t)\sigma(t)')^{-1}B(t)$ and (14) becomes

$$\begin{aligned}&[\tfrac{1}{2}P_t(t) + P(t)r(t)]x^2 + [Q_t(t) + Q(t)r(t) + P(t)c(t) - \lambda\mu_1 P(t)]x + R_t(t)\\ &+Q(t)c(t) - \lambda\mu_1 Q(t) + \tfrac{1}{2}\lambda\mu_2 P(t) - \tfrac{1}{2}[x + \tfrac{Q(t)}{P(t)}]^2 B(t)'\Sigma(t)P(t) = 0.\end{aligned}$$

Comparing the coefficients of x^2, x, and the constants respectively, and adding to the boundary conditions, we have the following differential equations

$$\begin{cases} P_t(t) = [-2r(t) + B(t)'\Sigma(t)]P(t) \\ P(T) = 1 \end{cases} \tag{15}$$

$$\begin{cases} Q_t(t) = [-r(t) + B(t)'\Sigma(t)]Q(t) + [\lambda\mu_1 - c(t)]P(t) \\ Q(T) = 0 \end{cases} \tag{16}$$

$$\begin{cases} R_t(t) = [\lambda\mu_1 - c(t)]Q(t) - \dfrac{1}{2}\lambda\mu_2 P(t) + \dfrac{1}{2}B(t)'\Sigma(t)\dfrac{Q(t)^2}{P(t)} \\ R(T) = 0. \end{cases} \tag{17}$$

Solving equations (15), (16), and (17), substituting the solutions into (13), and rearranging, we obtain the following theorem

Theorem 1. *A classical solution of the HJB equation (12) is*

$$\begin{aligned}V(t,x) &= \frac{1}{2}\lambda\mu_2\int_t^T e^{\int_v^T [2r(s) - B(s)'\Sigma(s)]ds}dv\\ &+ \frac{1}{2}e^{-\int_t^T B(s)'\Sigma(s)ds}\{xe^{\int_t^T r(s)ds} + \int_t^T [c(s) - \lambda\mu_1]e^{\int_s^T r(z)dz}ds\}^2\end{aligned} \tag{18}$$

and the value $u^*(t,x)$ that minimize the left side of the first equation in (12) is

$$u^*(t,x) = -\Sigma(t)\{x + e^{-\int_t^T r(s)ds}\int_t^T [c(s) - \lambda\mu_1]e^{\int_s^T r(z)dz}ds\}.$$

4 Verification theorem

The classical verification theorem described by Fleming and Soner (1993) for diffusion model can not be applied to our jump-diffusion model. In the following, a new verification theorem is given for our model.

Theorem 2. *Let $V(t,x)$ be defined by Theorem 1. Then $V(t,x) = J(t,x)$. Furthermore, the optimal strategy $u^*(v,x(v))$ is equal to*

$$-(\sigma(v)\sigma(v)')^{-1}B(v)\{x(v) + e^{-\int_v^T r(s)ds}\int_v^T [c(s) - \lambda\mu_1]e^{\int_s^T r(z)dz}ds\}. \quad (19)$$

Proof: For any admissible strategy u, applying Ito's formula for jump-diffusion process (9), we have

$$\begin{aligned}
V(T,x^u(T)) &= V(t,x) + \sum_{i=N(t)+1}^{N(T)} [V(T_i,x^u(T_i)) - V(T_i-,x^u(T_i-))] \\
&+ \int_t^T \{V_s(s,x^u(s)) + [r(s)x^u(s) + B(s)'u(s) + c(s)]V_x(s,x^u(s))\}ds \\
&+ \int_t^T u(s)'\sigma(s)V_x(s,x^u(s))dW(s) + \frac{1}{2}\int_t^T u(s)'\sigma(s)\sigma(s)'u(s)V_{xx}(s,x^u(s))ds.
\end{aligned}$$

Since $V(t,x)$ fulfills HJB equation (12), we have

$$\begin{aligned}
\frac{1}{2}[x^u(T)]^2 \geq \lambda\int_t^T\int_0^\infty [V(s,x^u(s)) - V(s,x^u(s)-y)]dF(y)ds \\
+ \sum_{i=N(t)+1}^{N(T)} [V(T_i,x^u(T_i)) - V(T_i-,x^u(T_i-))] \quad (20) \\
+ \int_t^T u(s)'\sigma(s)V_x(s,x^u(s))dW(s) + V(t,x)
\end{aligned}$$

We know that

$$\begin{aligned}
&\sum_{i=0}^{N(t)} [V(T_i,x^u(T_i)) - V(T_i-,x^u(T_i-))] + \int_0^t u(s)'\sigma(s)V_x(s,x^u(s))dW(s) \\
&+ \lambda\int_0^t\int_0^\infty [V(s,x^u(s)) - V(s,x^u(s)-y)]dF(y)ds
\end{aligned}$$

is a martingale. Taking expectations on both sides of inequality (20), it follows that

$$\frac{1}{2}E\{[x^u(T)]\}^2 \geq V(t,x),$$

which implies $J(t,x) \geq V(t,x)$. For the optimal strategy u^*, the inequality becomes equality, that is $\frac{1}{2}E\{[x^{u^*}(T)]^2\} = V(t,x)$. Thus $J(t,x) \leq V(t,x)$, then $J(t,x) = V(t,x)$, which completes the proof.

5 Efficient strategy and efficient frontier

In this section, we apply the results established in the previous section to the mean-variance problem. First of all, we give the following definition

Definition 2. The mean-variance portfolio selection problem is formulated as the following multi-objective optimization problem

$$\begin{aligned} &\min \quad (\mathrm{J_1(u(\cdot)), J_2(u(\cdot))} := (\mathrm{Var[X^u(T)], -E[X^u(T)]}) \\ &\text{subject to} \quad \begin{cases} u \in \Pi \\ (X(\cdot), u(\cdot)) \quad \text{satisfy} \quad (5). \end{cases} \end{aligned} \tag{21}$$

An admissible portfolio $u^*(\cdot)$ is called an efficient portfolio if there exists no admissible portfolio $u(\cdot)$ such that

$$J_1(u(\cdot)) \leq J_1(u^*(\cdot)), \quad J_2(u(\cdot)) \leq J_2(u^*(\cdot)) \tag{22}$$

with at least one of the inequalities holding strictly. In this case, we call $(J_1(u^*(\cdot)), -J_2(u^*(\cdot))) \in \mathbb{R}^2$ an efficient point. The set of all efficient points is called the efficient frontier.

In words, an efficient portfolio is one for which there does not exist another portfolio that has higher mean and no higher variance, and/or has less variance and no less mean at the terminal time T. In other words, an efficient portfolio is one that is Pareto optimal. It is well known that the variance minimizing portfolio is really the efficient portfolio (see Bielecki, Jin et al (2005)). So problem (6) is equivalent to problem (21).

Now, we present the optimal value of problem (7). For convenience, we omit the superscript u of $X^u(t)$. Set $x(t) = X(t) - (d - \beta)$ (then $X(t) = x(t) + (d - \beta)$, $X(0) = x(0) + (d - \beta)$) and $c(t) = c + (d - \beta)r(t)$ in (9). Then (9) is equivalent to (5). We have

$$\begin{aligned} E\{\frac{1}{2}[x(T)]^2\} &= E\{\frac{1}{2}[X(T) - (d - \beta)]^2\} \\ = E\{\frac{1}{2}[X(T) - d]^2\} &+ \beta[EX(T) - d] + \frac{1}{2}\beta^2. \end{aligned}$$

Hence, for every fixed β, we have

$$\begin{aligned}&\min_{u\in\Pi} E\{\frac{1}{2}[X(T)-d]^2\}+\beta[EX(T)-d]\\ =&\min_{u\in\Pi} E[\frac{1}{2}(x(T))^2]-\frac{1}{2}\beta^2=V(0,x(0))-\frac{1}{2}\beta^2.\end{aligned}$$

Because $c(s)=c+(d-\beta)r(s)$ in (19), the optimal investment strategy of problem (7) is

$$\begin{aligned}&u^*(t,X(t))=(u_1^*(t,X(t)),u_2^*(t,X(t)),\cdots,u_m^*(t,X(t)))\\ &=-\Sigma(t)[x+(d-\beta)(1-e^{-\int_t^T r(s)ds})+(c-\lambda\mu_1)\int_t^T e^{-\int_t^s r(z)dz}ds]\\ &=-\Sigma(t)[X(t)-(d-\beta)e^{-\int_t^T r(s)ds}+(c-\lambda\mu_1)\int_t^T e^{-\int_t^s r(z)dz}ds].\end{aligned}\tag{23}$$

Therefore, we conclude that the optimal value of problem (7) is

$$\begin{aligned}&\min_{u\in\Pi} E[X(T)-d]^2+2\beta[EX(T)-d]\\ =&P(0)[X_0-(d-\beta)]^2+2Q(0)[X_0-(d-\beta)]+2R(0)-\beta^2\\ =&e^{-\int_0^T B(s)'\Sigma(s)ds}[X_0e^{\int_0^T r(s)ds}+(c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv-d]^2\\ &+2e^{-\int_0^T B(s)'\Sigma(s)ds}[X_0e^{\int_0^T r(s)ds}+(c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv-d]\beta\\ &+[e^{-\int_0^T B(s)'\Sigma(s)ds}-1]\beta^2+\lambda\mu_2\int_0^T e^{\int_v^T[2r(s)-B(s)'\Sigma(s)]ds}dv.\end{aligned}\tag{24}$$

Note that the above value still depends on the Lagrange multiplier β, we denote it as $W(\beta)$. To obtain the optimal value (i.e., the minimum variance Var $X(T)$) and optimal strategy for the original portfolio selection problem (6), we need to maximize the value in (24) over $\beta\in\mathbb{R}$ according to the Lagrange duality theorem (see Luenberger (1968)).

From (24) we can see that, $W(\beta)$ is a concave function, so $W(\beta)$ attains its maximum

$$\begin{aligned}W(\beta^*)=&\frac{[X_0e^{\int_0^T r(s)ds}+(c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv-d]^2}{e^{\int_0^T B(s)'\Sigma(s)ds}-1}\\ &+\lambda\mu_2\int_0^T e^{\int_v^T[2r(s)-B(s)'\Sigma(s)]ds}dv\end{aligned}$$

at $\beta^*=\frac{X_0e^{\int_0^T r(s)ds}+(c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv-d}{e^{\int_0^T B(s)'\Sigma(s)ds}-1}$.

The above discussion leads to the following theorem

Theorem 3. *The efficient strategy of portfolio selection problem (6) (or (21)) corresponding to the expected terminal wealth $EX(T)=d$, as a function*

of time t and wealth $X(t)$, is

$$\begin{aligned} u^*(t,X(t)) &= (u_1^*(t,X(t)), u_2^*(t,X(t)), \cdots, u_m^*(t,X(t))) \\ &= -\Sigma(t)[X(t) - (d-\beta^*)e^{-\int_t^T r(s)ds} + (c-\lambda\mu_1)\int_t^T e^{-\int_t^s r(z)dz}ds], \end{aligned} \tag{25}$$

where

$$\beta^* = \frac{X_0 e^{\int_0^T r(s)ds} + (c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv - d}{e^{\int_0^T B(s)'(\sigma(s)\sigma(s)')^{-1}B(s)ds} - 1}.$$

Moreover, the efficient frontier is

$$\begin{aligned} \mathrm{Var[X(T)]} =& \frac{[X_0 e^{\int_0^T r(s)ds} + (c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv - EX(T)]^2}{e^{\int_0^T B(s)'(\sigma(s)\sigma(s)')^{-1}B(s)ds} - 1} \\ &+ \lambda\mu_2 \int_0^T e^{\int_v^T [2r(s) - B(s)'(\sigma(s)\sigma(s)')^{-1}B(s)]ds}dv. \end{aligned} \tag{26}$$

The expected terminal wealth $EX(T)$ satisfies

$$EX(T) \geq X_0 e^{\int_0^T r(s)ds} + (c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv.$$

Remark 1. From (25) and (26), we can see that if we do not consider the premium income and the claim payout, that is $c = 0$ and $\lambda = 0$, the efficient strategy and efficient frontier in this paper are the same as those in Zhou and Li (2000).

Remark 2. If there is a strategy u_1 leading to $EX^{u_1}(T) < X_0 e^{\int_0^T r(s)ds} + (c - \lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv$, we can get another strategy u_2 satisfying $Var[X^{u_1}(T)] = Var[X^{u_2}(T)]$ and $EX^{u_2}(T) = 2[X_0 e^{\int_0^T r(s)ds} + (c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv] - EX^{u_1}(T) > X_0 e^{\int_0^T r(s)ds} + (c-\lambda\mu_1)\int_0^T e^{\int_v^T r(s)ds}dv > EX^{u_1}(T)$ from (26). u_2 is mean-variance efficient. So Assumption 1 is reasonable.

Example 11. Let $m = 3$, $T = 1$, $\mu_1(= EY) = 0.1$, $r(t) \equiv r$, $B(t) \equiv B = \begin{pmatrix} 0.02 \\ 0.03 \\ 0.04 \end{pmatrix}$, and $\sigma(t) \equiv \sigma = \begin{pmatrix} 1 & 0 & \frac{2}{3} \\ 0 & 1 & 0 \\ 0 & 0 & \frac{2}{3} \end{pmatrix}$.

(1) Let $\mu_2 = 0.02$ (that is, $Var(Y) = \mu_2 - \mu_1^2 = 0.01$), $\lambda = 3$, $r = 0.04$ and c takes values 0.32, 0.30, and 0.28, then the efficient frontier is given by Figure 1 (the upside of the parabola). From Figure 1 we conclude that c does not effect the shape of the efficient frontier and the higher c, the higher $EX(T)$ with the same Var(X(T)).

(2) Let $c = 0.32$, $\lambda = 3$, $r = 0.04$ and μ_2 takes values 0.025, 0.020, and 0.015 (that is $Var(Y)$ take values 0.015, 0.010, and 0.005), then the efficient frontier is given by Figure 2 (the upside of the parabola). From Figure 2 we

can see that the higher μ_2, the lower $EX(T)$ with the same Var(X(T)) and this phenomenon is not obvious when Var(X(T)) is big enough.

(3) Let $\mu_2 = 0.02$ (that is, $Var(Y) = 0.01$), $c = 0.32$, $r = 0.04$ and λ takes values 3.5, 3, and 2.5, then the efficient frontier is given by Figure 3 (the upside of the parabola). From Figure 3 we can see that the higher λ, the lower $EX(T)$ with the same Var(X(T)).

(4) Let $\mu_2 = 0.02$ (that is, $Var(Y) = 0.01$), $c = 0.32$, $\lambda = 3$, and r takes values 0.05, 0.04, and 0.03, then the efficient frontier is given by Figure 4 (the upside of the parabola).

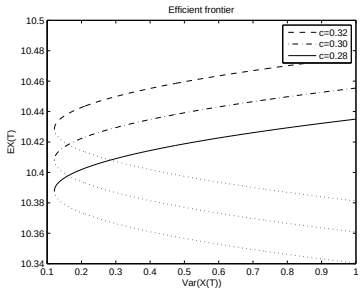

Fig. 1. Efficient frontier of different c.

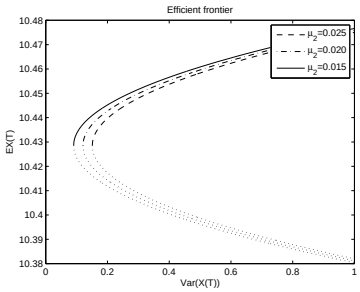

Fig. 2. Efficient frontier of different μ_2.

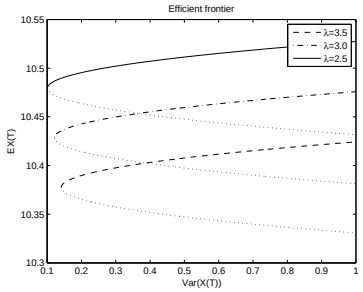

Fig. 3. Efficient frontier of different λ.

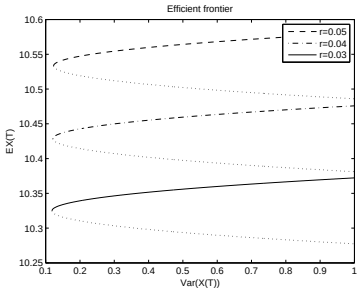

Fig. 4. Efficient frontier of different r.

6 Some other criteria

In Wang (2007), their objective is maximizing the expected exponential utility of the terminal wealth, that is $\max_{u\in\Pi} E[1 - e^{-\eta X(T)}]$. We can consider

it as maximizing the benefit. They do not consider the risk. Then the optimal strategy is $u(t) = \eta^{-1}e^{rt}((\sigma(t)\sigma(t)')^{-1})B(t)$, which is independent of the current wealth $X(t)$.

In Yang and Zhang (2005), they minimize the ruin probability which is a measure of the company's risk and they do not consider the benefit of the company. Their optimal investment strategy is also not related to the current wealth.

In this paper, we consider both the benefit (mean) and the risk (variance) under the mean-variance criterion. Our optimal investment strategy is related to the current wealth.

Acknowledgements. This research was supported by NNSF of China (Grant No. 10571092) and National Basic Research Program of China (973 program) 2007CB814905.

References

BIELECKI, T.R., JIN, H., PLISKA, S.R., ZHOU, X.Y. (2005): Continuous-time mean-variance portfolio selection with bankrutcy probihition. *Mathematical Finance 15, 213-244.*

BROWNE, S. (1995): Optimal investment policies for a firm with a random risk process: Exponential utility and minimizing the probability of ruin. *Mathematics of Operations Research 20, 937-957.*

FLEMING, W.H., SONER, H.M. (1993): *Controlled Markov processes and viscosity solutions.* Springer-Verlag: Berlin, New York.

GAIER, J., GRANDITS, P. and SCHACHERMAYER, W.(2003): Asymptotic ruin probabilities and optimal investment. *Annals of Applied Probability 13, 1054-1076.*

HIPP, C., PLUM, M. (2000): Optimal investment for insurers. *Insurance: Mathematics and Economics 27, 215-228.*

HIPP, C., SCHMIDLI, H. (2004): Asymptotics of ruin probabilities for controlled risk processes in the small claims case. *Scandinavian Actuarial Journal 5, 321–335.*

LI, X., ZHOU, X.Y., LIM, A.E.B. (2002): Dynamic mean-variance portfolio selection with no-shorting constraints. *SIAM J. Contr. Optim. 40, 1540-1555.*

LUENBERGER, D.G. (1968): *Optimization by vector space methods.* Wiley, New York.

MARKOWITZ, H. (1952): Portfolio selection. *Journal of Finance 7, 77-91.*

MERTON, R.C. (1972): An analytical derivation of the efficient portfolio frontier. *J. Financial and Economics Anal. 7, 1851-1872.*

WANG, N. (2007): Optimal investment for an insurer with exponential utility preference. *Insurance: Mathematics and Economics 40, 77-84.*

WANG, Z., XIA, J., ZHANG, L. (2007): Optimal investment for insurer with jump-diffusion risk process. *Insurance: Mathematics and Economics 40, 322-334.*

YANG, H.L., ZHANG, L.H. (2005): Optimal investment for insurer with jump-diffusion risk process. *Insurance: Mathematics and Economics 37, 615-634.*

ZHOU, X.Y., LI, D. (2000): Continuous-time mean-variance portfolio selection: a stochastic LQ framework. *Appl. Math. Optim 42, 19-33.*

Inhomogeneous Jump-GARCH Models with Applications in Financial Time Series Analysis

Chunhang Chen[1] and Seisho Sato[2]

[1] Department of Mathematical Sciences, University of the Ryukyus
Nishihara-cho, Okinawa 903-0213, Japan, *chench@sci.u-ryukyu.ac.jp*
[2] The Institute of Statistical Mathematics
Minami-Azabu 4-6-7, Minato-ku, Tokyo 106-8569, Japan, *sato@ism.ac.jp*

Abstract. This paper discusses the statistical behaviors and applicability of the jump-GARCH model proposed by Chen and Sato (2007), in which jump arrivals are time inhomogeneous and also state dependent. We discuss maximum likelihood estimation and likelihood ratio tests for the model. We investigate the statistical behaviors of the jump-GARCH model through financial time series analysis and showing comparisons of this model with GARCH and traditional jump models. Our results indicate that this model can reveal many important characteristics related with jump dynamics and volatility structures in asset prices.

Keywords: jump dynamics, jump-GARCH model, volatility

1 Introduction

Nowadays it has become a stylized fact that many financial asset prices possess jumps and the importance of taking considerations of jumps into the models has been widely recognized in finance literature. How do jumps occur? Empirical evidence indicates that jump arrivals tend to cluster, among others. As such, jump arrivals do not follow a traditional homogeneous Poisson process, which is a common assumption in many models used so far. Financial modeling with jump models and processes is thoroughly discussed by a recent work of Cont and Tankov (2004), in which Lévy processes with applications in modeling and option pricing of asset prices are fully treated. These authors also show some drawbacks of Lévy processes and indicate the importance of considering more general processes such as time inhomogeneous jump processes and stochastic volatility models with jumps. Recent researches have begun to develop jump models to cope with time inhomogeneous effects in jump arrivals by introducing a time-varying jump intensity into the model. Bates (2000), Andersen et al. (2002) and Pan (2002) considered continuous time jump-SV processes with a time-varying jump intensity depending on volatility. Maheu and Mccudy (2004) proposed a discrete time jump-GARCH model in which the conditional jump intensity is autoregressive.

Another important characteristic of jump dynamics revealed by empirical study of stock prices is that jump arrivals tend to be state-dependent, or asymmetric, which means that jump arrivals in a downside period may possess different characteristics with those in an upside period. It seems that the state-dependent structure of jump dynamics has not been taken into consideration in most of the models proposed so far. To incorporate state dependent jump dynamics, Chen and Sato (2007) proposed a nonlinear jump-GARCH model in which jump arrivals are allowed to be time-varying and also state-dependent. In this research we investigate in some details the statistical behaviors of the jump-GARCH model. We discuss maximum likelihood estimation and likelihood-ratio tests for this model. The model is used to analyze the jump dynamics of some financial asset prices. We will report the results related with the characteristics of jump dynamics and structures of volatilities of financial asset prices, which include some individual stocks, TOPIX, Yen-Dollar exchange rate and yields of bonds. Finally, we indicate possibilities for extending the jump-GARCH model to achieve more generality, flexibility and applicability.

2 Jump-GARCH models

Let S_t be the price of some financial asset at time t and $r_t = \log S_t - \log S_{t-1}$ its return. In practice the return series is often fitted by some simple models such as an AR(1) model:

$$r_t = \mu + \phi_1 r_{t-1} + \varepsilon_t,$$

where ε_t is the return shock, representing the one-step forecast error of the return. In many cases ε_t exhibits the behaviors of heteroscedastic weak white noise with heavy-tailed distribution, and the most widely used models for the return shock are ARCH and the GARCH-zoo models. However, GARCH models do not presume the existence of jumps and are not suitable for financial asset prices that actually have jumps. To remedy the problem of GAHCH models and also drawbacks of time homogeneous jump models, Maheu and Mccudy (2004) proposed a jump-GARCH model by adding an inhomogeneous jump component into the GARCH models. Although the model of Maheu and Mccudy (2004) shows nice behaviors for modeling inhomogeneous jump dynamics and volatility structures, it does not specify explicitly how jump dynamics depend on the state of asset prices. To incorporate both time inhomogeneous and state dependent structures of jump dynamics, Chen and Sato (2007) proposed another jump-GARCH model, which is introduced in this section.

Taking jumps of financial prices into consideration, now the return shock is expressed as a sum of two conditionally independent components:

$$\varepsilon_t = \varepsilon_{t,1} + \varepsilon_{t,2}, \qquad \{\varepsilon_{t,1}|\mathcal{F}_{t-1}\} \perp \{\varepsilon_{t,2}|\mathcal{F}_{t-1}\},$$

where $\varepsilon_{t,1}$ represents the one-step forecast error of the part of the return that changes smoothly, while $\varepsilon_{t,2}$ is the one-step forecast error of the part of the return that is due to a jump. Here $\mathcal{F}_{t-1}$ is the information of the prices up to time $t-1$. We model the component $\varepsilon_{t,1}$ in the following way:

$$\varepsilon_{t,1} = \sigma_{t,1} z_t, \qquad \{z_t\} \sim \mathrm{NID}(0,1),$$

$$\sigma_{t,1}^2 := \mathrm{Var}(\varepsilon_{t,1}|\mathcal{F}_{t-1}) = \omega + \beta\sigma_{t-1,1}^2 + \alpha\varepsilon_{t-1}^2,$$

where $\omega,\ \beta,\ \alpha > 0$ and $\beta + \alpha < 1$. We call $\varepsilon_{t,1}$ the GARCH-shock and $\sigma_{t,1}^2$ the GARCH-volatility. This model is close to a GARCH(1,1) model. It would become a genuine GARCH(1,1) model if the shock ε_{t-1} was substituted with the GARCH-shock $\varepsilon_{t-1,1}$. Here the consideration behind this is that the GARCH-shock can not be observed, and that jumps occurred previously cause impacts on GARCH-volatility.

In order to model the jump component $\varepsilon_{t,2}$, we need to make some assumptions about the jump dynamics. Suppose in the t-th day there will be N_t jumps to occur, among which the jump size of the k-th jump is $Y_{t,k}$. Let J_t denote the jump amplitude in the t-th day. We make the following major assumptions:

$$N_t|\mathcal{F}_{t-1} \sim \mathrm{Poisson}(\lambda_t), \qquad \{Y_{t,k}\}_{k=1}^{\infty} \sim \mathrm{NID}(\theta, \delta^2).$$

We further assume that N_t and $\{Y_{t,k}\}$ are independent. Here λ_t is the conditional jump intensity which controls the jump dynamics. We propose to model the conditional intensity in the following way:

$$\lambda_t = \nu_0 + \nu_1\lambda_{t-1} + \tau\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 I_{(\varepsilon_{t-1}<0)},$$

where $\nu_0,\ \tau,\ \gamma \geq 0,\ 0 \leq \nu_1 < 1$, and I_A is the indicator function of the event A. The main roles of the parameters are as follows: ν_1 controls the persistence of jump-clustering effect, τ and γ control the time-varying and state-dependent (asymmetric) effects in jump dynamics. By definition, the jump component $\varepsilon_{t,2}$ represents the one-step forecast error of the jump amplitude J_t. Under the above assumptions we easily understand that $\varepsilon_{t,2}$ is expressed as

$$\varepsilon_{t,2} := J_t - E(J_t|\mathcal{F}_{t-1}) = \sum_{k=1}^{N_t} Y_{t,k} - \theta\lambda_t.$$

We call $\varepsilon_{t,2}$ the jump-shock. It induces another source of risk (volatility), which is given by

$$\sigma_{t,2}^2 := \mathrm{Var}(\varepsilon_{t,2}|\mathcal{F}_{t-1}) = (\theta^2 + \delta^2)\lambda_t.$$

We call $\sigma_{t,2}^2$ the jump-volatility. The total volatility is the sum of the GARCH-vol. and the jump-vol.:

$$\sigma_t^2 := \mathrm{Var}(\varepsilon_t|\mathcal{F}_{t-1}) = \sigma_{t,1}^2 + \sigma_{t,2}^2.$$

In the jump-GARCH model, if we set $\nu_0 = \nu_1 = \tau = \gamma = 0$, the model turns to be a GARCH(1,1) model. Furthermore, if we let $\nu_1 = \tau = \gamma = 0$, the model becomes a jump-SV model with time homogeneous jump arrivals. Thus GARCH and essential jump-SV models are embedded into the jump-GARCH model, which enables us to perform likelihood ratio tests for GARCH or time homogeneous jump-SV models vs. jump-GARCH model.

The major contribution of this research is in adding the time-varying and state-dependent term $\tau\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 I_{(\varepsilon_{t-1}<0)}$ into the conditional jump intensity, which allows us to cope with jump dynamics more flexibly. Our empirical study to be given in the next section will show that the proposed jump-GARCH model can reveal much more important and interesting characteristics of jump dynamics and structures of volatility for various financial assets than GARCH and traditional jump models.

□ **Maximum likelihood estimation of the model**

Suppose an AR(1) model is fitted to the return series and the return shock is fitted by the jump-GARCH model. Then the model includes the following unknown parameters:

$\boldsymbol{\theta} = (\mu, \phi_1, \omega, \beta, \alpha, \theta, \delta^2,\ \nu_0, \nu_1, \tau, \gamma)' \in \mathbb{R}^{11}$.

To estimate these unknown parameters, we consider the maximum likelihood estimation. Through some conditioning arguments, the likelihood function is evaluated as follows:

$\varepsilon_t = r_t - \phi_1 r_{t-1} - \mu,\ t = 1, \ldots, T$;

$\varepsilon_t|(N_t = j,\ \mathcal{F}_{t-1}) \sim \mathrm{N}(j\theta - \theta\lambda_t,\ \sigma_{t,1}^2 + j\delta^2)$;

$N_t|\mathcal{F}_{t-1} \sim \mathrm{Poisson}(\lambda_t)$;

$$
\begin{aligned}
f(\varepsilon_t|\mathcal{F}_{t-1}) &= \sum_{j=0}^{\infty} f(\varepsilon_t|N_t = j,\ \mathcal{F}_{t-1}) P(N_t|\mathcal{F}_{t-1}) \\
&= \sum_{j=0}^{\infty} \frac{1}{\sqrt{2\pi(\sigma_{t,1}^2 + j\delta^2)}} \exp\left\{ -\frac{(\varepsilon_t - j\theta + \theta\lambda_t)^2}{2(\sigma_{t,1}^2 + j\delta^2)} \right\} \cdot \frac{\lambda_t^j}{j!} e^{-\lambda_t} \\
&=: \sum_{j=0}^{\infty} g(t, j) =: L_t(\boldsymbol{\theta});
\end{aligned}
$$

Likelihood function: $L(\boldsymbol{\theta}) = \prod_1^T L_t(\boldsymbol{\theta})$;

Log-likelihood function: $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_1^T \log L_t(\boldsymbol{\theta})$.

The score of the model is given in the Appendix. Based on the score, numerical calculation of the MLE $\hat{\boldsymbol{\theta}}_T$ can be undertaken by using Newton-Raphson method:

$$
\hat{\boldsymbol{\theta}}_T^{(k)} = \hat{\boldsymbol{\theta}}_T^{(k-1)} + \left[\sum_{t=1}^{T} S_t\Big(\hat{\boldsymbol{\theta}}_T^{(k-1)}\Big) \Big(S_t(\hat{\boldsymbol{\theta}}_T^{(k-1)})\Big)' \right]^{-1} \sum_{t=1}^{T} S_t\Big(\hat{\boldsymbol{\theta}}_T^{(k-1)}\Big),
$$

where $S(\boldsymbol{\theta}) = \frac{1}{T}\sum_1^T S_t(\boldsymbol{\theta})$ is the score. Under suitable assumptions, it can be shown that the MLE $\hat{\boldsymbol{\theta}}_T$ is asymptotically normally distributed:

$$\hat{\boldsymbol{\theta}}_T \sim \mathrm{AN}\left(\boldsymbol{\theta},\ \left[\sum_{t=1}^{T} S_t\Big(\hat{\boldsymbol{\theta}}_T\Big)\Big(S_t(\hat{\boldsymbol{\theta}}_T)\Big)'\right]^{-1}\right).$$

Using these results, we can perform likelihood-ratio tests and t-tests for the parameters and calculate the standard errors of the MLE.

3 Empirical study

We investigate statistical behaviors of the jump-GARCH model through an empirical study. Then we use this model to reveal characteristics of jump dynamics and volatility structures for various financial assets.

□ **Data**
Data sets include the following daily time series:

(1) TOPIX (From Jan. 4, 1985 to Jun. 30, 2005; $T = 5177$);
(2) TEPCO —Tokyo Electricity Production Co., traded in the 1st division of TSE (From Jan. 4, 1990 to Dec. 16, 2005; $T = 3932$);
(3) WOWOW, an equity traded in the newly raised Mothers-Market of TSE (From Apr. 20, 2001 to Aug. 1, 2005; $T = 1039$);
(4) Yields of the Japanese 10 years bond (From Mar. 1, 1989 to Sep. 30, 2004; $T = 3844$);
(5) Japanese Yen-US Dollar exchange rate (From Jan. 1, 1990 to Dec. 22, 2005; $T = 4169$).

□ **Estimation and likelihood-ratio tests**
For these series, we fit an AR(1) model to the return series and a jump-GARCH model to the return shock. All of the parameters are estimated via maximum likelihood estimation. Table 1 shows the results, where $(\cdot)$ is the corresponding standard error. The four parameters θ, ν_1, τ and γ are important in characterizing the different structures of the jump dynamics among the various financial assets. Before discussing this, we undertake some likelihood-ratio tests for the parameters in the jump part. Here we focus on the following likelihood-ratio tests:

- Test I: $\nu_0 = \nu_1 = \tau = \gamma = 0$ — A test of GARCH vs. jump-GARCH, or existence of a jump;
- Test II: $\nu_1 = \tau = \gamma = 0$ — A test of inhomogeneity of jump arrivals;
- Test III: $\nu_1 = 0$ — A test of jump-cluster effect;
- Test IV: $\tau = 0$ — A test of state-dependent effect in jump dynamics.
- Test V: $\gamma = 0$ — A test of asymmetric state-dependent effect in jump dynamics.

Table 1. Maximum likelihood estimates of the jump-GARCH model.

	TOPIX	TEPCO	WOWOW	Bond	Yen-Dollar
μ	0.000146	−0.00015	0.000238	−0.00183	−0.00004
	(0.000125)	(0.000179)	(0.001043)	(0.000536)	$(9.82E-05)$
ϕ_1	0.139983	−0.03947	0.002016	0.004708	−0.01915
	(0.01492)	(0.017589)	(0.033106)	(0.017995)	(0.016531)
ω	$1.21E-06$	$7.66E-06$	$1.32E-05$	$1.55E-05$	$3.59E-07$
	$(2.12E-07)$	$(1.49E-06)$	$(8.15E-06)$	$(3.66E-06)$	$(1.00E-07)$
β	0.911955	0.708497	0.925189	0.901333	0.963008
	(0.008758)	(0.0338)	(0.038663)	(0.011993)	(0.006586)
α	0.052896	0.128189	0.011300	0.060886	0.017217
	(0.007555)	(0.01504)	(0.009652)	(0.009833)	(0.003567)
θ	−0.00235	0.001590	0.013822	0.012108	−0.00315
	(0.001012)	(0.000682)	(0.00316)	(0.005372)	(0.000885)
δ^2	0.000266	0.000219	0.001274	0.003445	0.000120
	$(3.86E-05)$	$(2.65E-05)$	(0.000233)	(0.000523)	$(1.97E-05)$
ν_0	0.005842	$9.99E-09$	0.100161	0.031170	0.019717
	(0.004837)	(0.000926)	(0.049801)	(0.016251)	(0.009185)
ν_1	0.680868	0.973631	0.635636	0.443532	0.621479
	(0.061651)	(0.005459)	(0.105102)	(0.180614)	(0.100486)
τ	$1.81E-05$	0.000010	100.9098	25.92714	96.12646
	(51.97723)	(8.290643)	(38.74718)	(9.853342)	(153.1217)
γ	707.8508	112.7074	6.970889	$3.50E-12$	790.3226
	(170.2787)	(31.40688)	(41.03209)	(13.46923)	(282.9626)

Table 2. The p-values of the likelihood-ratio tests.

	TOPIX	TEPCO	WOWOW	Bond	Yen-Dollar
Test I	0.0000	0.0000	0.0000	0.0000	0.0000
Test II	0.0000	0.0000	0.0015	0.0007	0.0000
Test III	0.0000	0.0000	0.0015	0.2920	0.0000
Test IV	1.0000	0.9989	0.0008	0.0000	0.5074
Test V	0.0000	0.0000	0.8608	1.0000	0.0008

The p-values of these likelihood-ratio tests are given in Table 2. The results for Test I show that, in all the cases the null hypothesis of a GARCH model is rejected even at a significance level 1%, and thus provide strong evidence for existence of a jump. Furthermore, the results for Test II show that, jump arrivals in the asset prices are not time homogeneous. Tests III, IV and V reveal some clear differences in structures of the time inhomogeneous and state dependent jump dynamics of the various financial assets, which we summarize as follows:

- Among the various assets, jump arrivals in yields of the Japanese 10-year bond appear to have the simplest structure: they are weakly state dependent, but not asymmetric and not persistent;
- Jump arrivals in WOWOW are state dependent and persistent, but not asymmetric.
- Jump arrivals in TOPIX, TEPCO and Yen-Dollar exchange rate are persistent and only asymmetrically state dependent, which mean that only a large fall in prices lead to a jump cluster. However, we observe that the jump-size parameter $\theta < 0$ in TOPIX and Yen-Dollar, while $\theta > 0$ in TEPCO. These indicate that, large falls of TOPIX and Yen-Dollar exchange rate tend to cause a cluster of downward jumps, while a large fall of TEPCO tends to lead to a cluster of upward jumps. The diametrical way in jump dynamics of TEPCO against TOPIX and Yen-Dollar exchange rate reflects the fact that this equity has been traded as a defensive asset in TSE. On the other hand, the jump dynamics of TOPIX and Yen-Dollar exchange rate seem to real the mechanism of the well-known leverage effect: If jump arrivals are persistent and only asymmetrically state-dependent, then a large fall in price tends to lead to a downward jump cluster, which in turn increases the jump-vol. and thus the total volatility. However, as in the cases of TEPCO, WOWOW and Bond, leverage effects may not be a common and significant effect for various assets, and our empirical study shows that the proposed jump-GARCH model is much more flexible for treating more general effects than the leverage effect.

□ **Goodness-of-fit**
It is not easy to perform a test of goodness-of-fit for the jump-GARCH model, since the distributions of the residuals are much more complicated. Here we show initially goodness-of-fit by comparing the marginal distribution of the model with the empirical distribution and the marginal distribution of a GARCH(1,1) model. Fig.1 shows comparisons among the distributions. We find that the jump-GARCH model fits remarkably well and much more better than a GARCH(1,1) model.

□ **Jump dynamics**
Our empirical study based on the jump-GARCH model shows that jumps do occur in asset prices and that in general jump arrivals are time inhomogeneous and state dependent, and may be also asymmetric. To deepen understanding for jump dynamics, we show time series plots of conditional jump intensities for the various assets in Fig.2. We find that jumps seem to occur more often in the individual equities, namely TEPCO and WOWOW, than in TOPIX, Yen-Dollar exchange rate and the Bond. We can also observe that jump arrivals in TEPCO are strongly persistent.

The jump-GARCH model provides an efficient way to obtain an ex post estimate of the probability of the event that jumps occurred in a certain day,

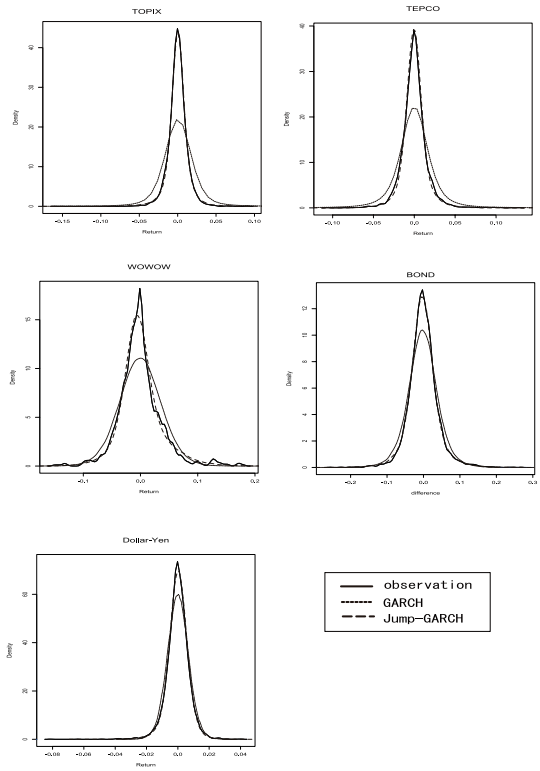

Fig. 1. Density.

which is important in risk management. The ex post estimate of the jump probability in the t-th day is given as follows:

$$\begin{aligned} P_t &:= P\{N_t \geq 1|\mathcal{F}_t\} = 1 - P\{N_t = 0|\mathcal{F}_t\} \\ &= 1 - \frac{f(\varepsilon_t|N_t = 0, \mathcal{F}_{t-1})P\{N_t = 0|\mathcal{F}_{t-1}\}}{f(\varepsilon_t|\mathcal{F}_{t-1})} \\ &= 1 - \frac{\dfrac{1}{\sqrt{2\pi\sigma_{t,1}^2}} \exp\left\{-\dfrac{(\varepsilon_t + \theta\lambda_t)^2}{2\sigma_{t,1}^2} - \lambda_t\right\}}{\displaystyle\sum_{j=0}^{\infty} g(t,j)} \end{aligned}$$

Ex post estimates of jump probabilities for the various assets are given in Fig.3. Paying attentions to the days whose ex post estimates of jump probabilities are close to 1, we find that jumps occurred more often in the individual asset than the other assets. Again, we find that jump arrivals in TEPCO are strongly persistent.

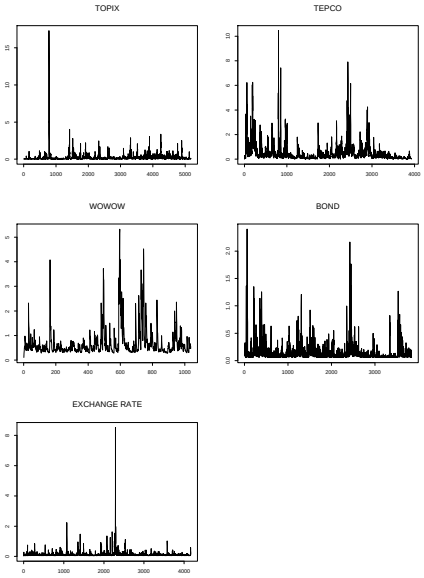

Fig. 2. Intensity.

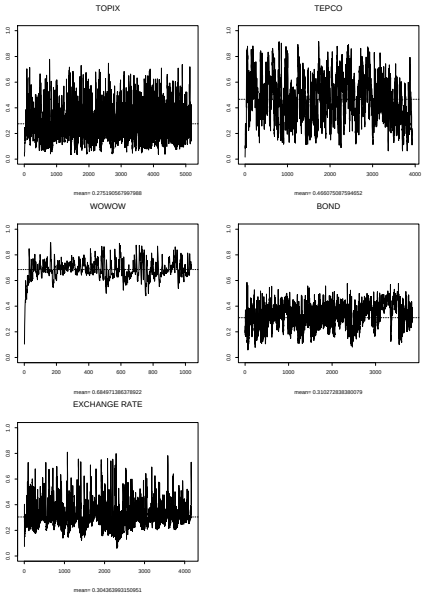

Fig. 4. Ratio of jump volatility.

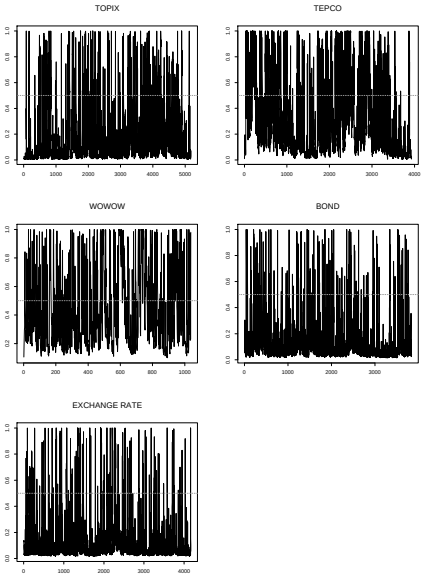

Fig. 3. Jump probability.

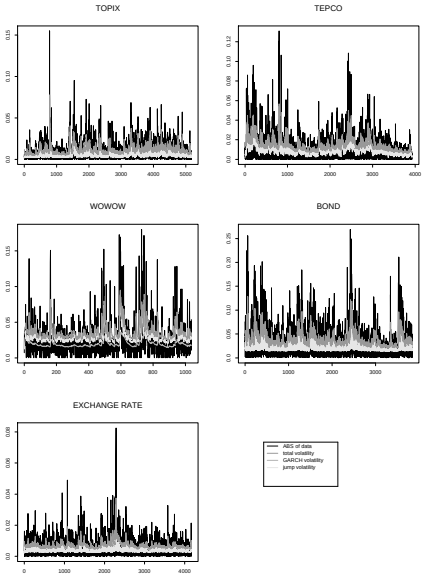

Fig. 5. Volatility.

□ **Structures of volatility**

The presence of jumps causes another source of volatility — the jump-volatility. Since jump arrivals are due to the impacts of news arrivals, the jump-vol. can be explained as an external or nonsystematic risk in some sense, while the GARCH-vol. an essential or systematic risk. From a point of view of risk management, it is important to measure the contributions of jump-vol. to total-vol. in an asset. Fig.4 shows ratios of jump-vol. to total-vol. for the various assets, where the average ratios are also given. In the cases of TOPIX, Bond and Yen-Dollar exchange rate, the jump-vol. contributes in average about 30% to the total volatility. However, in these assets we have found that when a jump occurred, the contributions of jump-vol. may increase to 60% or 70%. On the other hand, contributions of jump-vol. in individual equities are much higher than those in TOPIX, bond and Yen-Dollar exchange rate. In particular, jump-vol. in WOWOW turns to be the dominating part of volatility in most of the trading days. We think that volatileness of a financial asset is best characterized by the contribution of the jump-vol. to the total-vol., which can be evaluated by using the jump-GARCH model.

Since GARCH models have widely used in financial engineering, it is useful to compare the volatilities measured by GARCH and jump-GARCH models. Fig.5 shows such comparisons, where absolute returns and the jump volatilities are also plotted. We have found that there are considerably differences in the volatilities measured by the two models. Since our empirical study strongly support the jump-GARCH model, we think that the jump-GARCH model is more reasonable than a GARCH model for measuring risks.

□ **Further extensions**

The jump-GARCH model can be further extended along some directions to gain its flexibility and applicability. We provide some possibilities here. Firstly, we can input some explanatory variables related with macro economic market and worldwide political situations into the conditional jump intensity. If we have additional information such as opening prices, minimum/maximum prices or traded volumes, we can also input these information into the jump intensity, which may improve behaviors of the jump-GARCH model.

Other possibility to gain applicability of the model is to combine the use of the jump-GARCH model with some nonlinear mean time series models, such as threshold AR models, Markov switching models and simultaneous switching AR models (Kunitomo and Sato (1996)).

4 Summary

In this research we have investigated statistical behaviors of the jump-GARCH model proposed by Chen and Sato (2007). This model is efficient in revealing structures of jump dynamics in asset prices. Based on this model, we

have found that jump arrivals in asset prices are time inhomogeneous and state dependent, and may be persistent and asymmetrically state dependent, which together characterize the nature of jump dynamics of an asset. As a result, this model can detect the fine structure of volatility dynamics, which is important for financial risk measurement and management. .

◯ Appendix: The score of the jump-GARCH model

$$S(\boldsymbol{\theta}) := \frac{1}{T}\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{T}\sum_{1}^{T}\frac{1}{L_t(\boldsymbol{\theta})}\frac{\partial L_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} =: \frac{1}{T}\sum_{1}^{T} S_t(\boldsymbol{\theta})$$

$$S_t(\boldsymbol{\theta}) = \frac{1}{L_t(\boldsymbol{\theta})}\cdot\frac{\partial L_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$\frac{\partial L_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial L_t}{\partial \mu}, \frac{\partial L_t}{\partial \phi_1}, \frac{\partial L_t}{\partial \omega}, \frac{\partial L_t}{\partial \beta}, \frac{\partial L_t}{\partial \alpha}, \frac{\partial L_t}{\partial \theta}, \frac{\partial L_t}{\partial (\delta^2)}, \frac{\partial L_t}{\partial \nu_0}, \frac{\partial L_t}{\partial \nu_1}, \frac{\partial L_t}{\partial \tau}, \frac{\partial L_t}{\partial \gamma}\right)'$$

- $$\frac{\partial L_t}{\partial \mu} = -\frac{1}{2}\cdot\frac{\partial(\sigma_{t,1}^2)}{\partial \mu}\sum_{j=0}^{\infty}\frac{1}{\sigma_{t,1}^2 + j\delta^2}\cdot g(t,j) + \frac{\partial \lambda_t}{\partial \mu}\sum_{j=0}^{\infty}\left(\frac{j}{\lambda_t} - 1\right) g(t,j)$$
$$+\frac{1}{2}\sum_{j=0}^{\infty}\frac{\varepsilon_t - j\theta + \theta\lambda_t}{(\sigma_{t,1}^2 + j\delta^2)^2}\left((\varepsilon_t - j\theta + \theta\lambda_t)\frac{\partial(\sigma_{t,1}^2)}{\partial \mu} + 2(\sigma_{t,1}^2 + j\delta^2)\Big(1 - \theta\frac{\partial \lambda_t}{\partial \mu}\Big)\right) g(t,j)$$

Here,

$$\frac{\partial(\sigma_{1,1}^2)}{\partial \mu} = 0; \quad \frac{\partial(\sigma_{t,1}^2)}{\partial \mu} = \beta\frac{\partial(\sigma_{t-1,1}^2)}{\partial \mu} - 2\alpha\varepsilon_{t-1},\ t = 2,\ldots,T$$

$$\frac{\partial \lambda_1}{\partial \mu} = 0; \quad \frac{\partial \lambda_t}{\partial \mu} = \nu_1\frac{\partial \lambda_{t-1}}{\partial \mu} - 2\tau\varepsilon_{t-1} - 2\gamma\varepsilon_{t-1}I_{(\varepsilon_{t-1}<0)},\ t = 2,\ldots,T$$

- $$\frac{\partial L_t}{\partial \phi_1} = -\frac{1}{2}\cdot\frac{\partial(\sigma_{t,1}^2)}{\partial \phi_1}\sum_{j=0}^{\infty}\frac{1}{\sigma_{t,1}^2 + j\delta^2}\cdot g(t,j) + \frac{\partial \lambda_t}{\partial \phi_1}\sum_{j=0}^{\infty}\left(\frac{j}{\lambda_t} - 1\right) g(t,j)$$
$$+\frac{1}{2}\sum_{j=0}^{\infty}\frac{\varepsilon_t - j\theta + \theta\lambda_t}{(\sigma_{t,1}^2 + j\delta^2)^2}\left((\varepsilon_t - j\theta + \theta\lambda_t)\frac{\partial(\sigma_{t,1}^2)}{\partial \phi_1} + 2(\sigma_{t,1}^2 + j\delta^2)\Big(r_{t-1} - \theta\frac{\partial \lambda_t}{\partial \phi_1}\Big)\right) g(t,j)$$

Here,

$$\frac{\partial(\sigma_{1,1}^2)}{\partial \phi_1} = 0; \quad \frac{\partial(\sigma_{t,1}^2)}{\partial \phi_1} = \beta\frac{\partial(\sigma_{t-1,1}^2)}{\partial \phi_1} - 2\alpha\varepsilon_{t-1}r_{t-1},\ t = 2,\ldots,T$$

$$\frac{\partial \lambda_1}{\partial \phi_1} = 0; \quad \frac{\partial \lambda_t}{\partial \phi_1} = \nu_1\frac{\partial \lambda_{t-1}}{\partial \phi_1} - 2\tau\varepsilon_{t-1}r_{t-1} - 2\gamma\varepsilon_{t-1}r_{t-1}I_{(\varepsilon_{t-1}<0)}$$

- $$\frac{\partial L_t}{\partial \omega} = \frac{\partial(\sigma_{t,1}^2)}{\partial \omega}\sum_{j=0}^{\infty}\frac{1}{2}\left[\left(\frac{\varepsilon_t - j\theta + \theta\lambda_t}{\sigma_{t,1}^2 + j\delta^2}\right)^2 - \frac{1}{\sigma_{t,1}^2 + j\delta^2}\right] g(t,j)$$
$$=: \frac{\partial(\sigma_{t,1}^2)}{\partial \omega}\sum_{j=0}^{\infty} h(t,j)$$

Here,

$$\frac{\partial(\sigma_{1,1}^2)}{\partial \omega} = 1; \quad \frac{\partial(\sigma_{t,1}^2)}{\partial \omega} = 1 + \beta\frac{\partial(\sigma_{t-1,1}^2)}{\partial \omega},\ t = 2,\ldots,T$$

- $$\frac{\partial L_t}{\partial \beta} = \frac{\partial(\sigma_{t,1}^2)}{\partial \beta}\sum_{j=0}^{\infty} h(t,j)$$

Here,

$$\frac{\partial(\sigma_{1,1}^2)}{\partial \beta} = \sigma_{0,1}^2; \quad \frac{\partial(\sigma_{t,1}^2)}{\partial \beta} = \sigma_{t-1,1}^2 + \beta\frac{\partial(\sigma_{t-1,1}^2)}{\partial \beta},\ t = 2,\ldots,T$$

- $$\frac{\partial L_t}{\partial \alpha} = \frac{\partial(\sigma_{t,1}^2)}{\partial \alpha}\sum_{j=0}^{\infty} h(t,j)$$

Here,

$$\frac{\partial(\sigma_{1,1}^2)}{\partial \alpha} = \varepsilon_0^2; \quad \frac{\partial(\sigma_{t,1}^2)}{\partial \alpha} = \varepsilon_{t-1}^2 + \beta\frac{\partial(\sigma_{t-1,1}^2)}{\partial \alpha},\ t = 2,\ldots,T$$

- $$\frac{\partial L_t}{\partial \theta} = \sum_{j=0}^{\infty}\frac{\varepsilon_t - j\theta + \theta\lambda_t}{\sigma_{t,1}^2 + j\delta^2}\cdot(j - \lambda_t)\cdot g(t,j)$$

- $\dfrac{\partial L_t}{\partial(\delta^2)} = \sum_{j=0}^{\infty} j \cdot h(t,j)$
- $\dfrac{\partial L_t}{\partial \nu_0} = \dfrac{\partial \lambda_t}{\partial \nu_0} \sum_{j=0}^{\infty} \left[g_1(t,j) - g(t,j) \left(\theta \cdot \dfrac{\varepsilon_t - j\theta + \theta\lambda_t}{\sigma_{t,1}^2 + j\delta^2} + 1 \right) \right]$
  $=: \dfrac{\partial \lambda_t}{\partial \nu_0} \sum_{j=0}^{\infty} h_1(t,j)$

Here

$$g_1(t,j) := \frac{1}{\sqrt{2\pi(\sigma_{t,1}^2 + (j+1)\delta^2)}} \exp\left\{ -\frac{(\varepsilon_t - (j+1)\theta + \theta\lambda_t)^2}{2(\sigma_{t,1}^2 + (j+1)\delta^2)} \right\} \cdot \frac{\lambda_t^j}{j!} e^{-\lambda_t}$$

$$\frac{\partial \lambda_1}{\partial \nu_0} = 1; \quad \frac{\partial \lambda_t}{\partial \nu_0} = 1 + \nu_1 \frac{\partial \lambda_{t-1}}{\partial \nu_0}, \ t = 2, \ldots, T$$

- $\dfrac{\partial L_t}{\partial \nu_1} = \dfrac{\partial \lambda_t}{\partial \nu_1} \sum_{j=0}^{\infty} h_1(t,j)$

Here

$$\frac{\partial \lambda_1}{\partial \nu_1} = \lambda_0; \quad \frac{\partial \lambda_t}{\partial \nu_1} = \lambda_{t-1} + \nu_1 \frac{\partial \lambda_{t-1}}{\partial \nu_1}, \ t = 2, \ldots, T$$

- $\dfrac{\partial L_t}{\partial \tau} = \dfrac{\partial \lambda_t}{\partial \tau} \sum_{j=0}^{\infty} h_1(t,j)$

Here

$$\frac{\partial \lambda_1}{\partial \tau} = \varepsilon_0^2; \quad \frac{\partial \lambda_t}{\partial \tau} = \varepsilon_{t-1}^2 + \nu_1 \frac{\partial \lambda_{t-1}}{\partial \tau}, \ t = 2, \ldots, T$$

- $\dfrac{\partial L_t}{\partial \gamma} = \dfrac{\partial \lambda_t}{\partial \gamma} \sum_{j=0}^{\infty} h_1(t,j)$

Here

$$\frac{\partial \lambda_1}{\partial \gamma} = \varepsilon_0^2 I_{(\varepsilon_0 < 0)}; \quad \frac{\partial \lambda_t}{\partial \gamma} = \varepsilon_{t-1}^2 I_{(\varepsilon_{t-1} < 0)} + \nu_1 \frac{\partial \lambda_{t-1}}{\partial \gamma}, \ t = 2, \ldots, T$$

Acknowledgment

Helpful comments by the anonymous referee and the editor are gratefully acknowledged. This research was partially supported by the Grant-in-Aid for Scientific Research by the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

ANDERSEN, T.G., BENZONI, L. and LUND, J. (2002): An empirical investigation of continuous-time equity return models. *Journal of Finance 62*, 1239-1284.

BATES, D.S. (2000): Post-'87 crash fears in the S&P 500 futures option market. *Journal of Econometrics 94*, 181-238.

CHEN, C. and SATO, S. (2007): Jump-GARCH models and jump dynamics in financial asset prices. *Bulletin of the International Statistical Institute*.

CONT, R. and TANKOV, P. (2004): *Financial Modelling With Jump Processes*. Chapman & Hall/CRC.

KUNITOMO, N. and SATO, S. (1996) : Asymmetry in economic time series and simultaneous switching autoregressive model, *Structural Change and Economic Dynamics 7*, 1-34

MAHEU, J.M. and MCCURDY, T.H. (2004): News arrival, jump dynamics, and volatility components for individual stock returns. *Journal of Finance 64*, 755-793.

PAN, J. (2002): The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of Financial Economics 63*, 3-50.

The Classical Risk Model with Constant Interest and Threshold Strategy

Yinghui Dong[1] and Kam C. Yuen[2]

[1] Department of Mathematics, Suzhou Technology University
Suzhou 215006, P.R. China, *dongyinghui1030@163.com*
[2] Department of Statistics and Actuarial Science, The University of Hong Kong
Pokfulam Road, Hong Kong, *kcyuen@hku.hk*

Abstract. In recent years, insurance risk models with dividend payments have been studied extensively. The threshold dividend strategy assumes that dividends are paid out at the maximal admissible rate whenever the surplus exceeds a certain threshold. In this paper, we consider the classical risk model with constant interest under the threshold strategy. We derive integro-differential equations for the expected discounted penalty function. In some special cases with exponential claims, we are able to obtain closed-form expressions for the expected discounted penalty function.

Keywords: classical risk model, dividend payments, threshold strategy

1 Introduction

Suppose that the surplus process of an insurer follows the classical risk model given by

$$U(t) = u + ct - \sum_{k=1}^{N(t)} Z_k = u + ct - S(t), \quad t \geq 0, \tag{1}$$

where $u \geq 0$ is the initial surplus, $c > 0$ is the rate of premium, $N(t)$ is a Poisson process with intensity $\lambda > 0$, and $\{Z_k, k = 1, 2, \cdots\}$ is a sequence of independent and identically distributed non-negative random variables with common distribution F. It is assumed that $F(0) = 0$ and that $N(t)$ and Z_k's are independent. Since $N(t)$ indicates the number of claims up to time t and Z_k's represent the claim amounts, the compound Poisson process $S(t)$ is usually called the aggregate claims process. Note that the surplus process (1) is also known as the compound Poisson risk model.

In the classical risk model (1), the assumption that the claim amounts Z_k's are independent is often not met in many non-life insurance problems. In view of this, the study of risk models with various dependence relations among claim amounts as well as different classes of insurance business has become one of the popular actuarial topics in the past decade. Hence, one may try to extend the main results presented in this paper to a risk model

with both correlated claims and dividend payments. Undoubtedly, such an extension would be a challenging one.

Assume that the insurer pays out certain amount of his surplus as dividends to the policyholders according to some dividend strategy. Let $D(t)$ be the total dividends paid up to time t, and $X(t)$ be the resulting surplus of the insurer at time t. Thus,

$$X(t) = U(t) - D(t), \quad t \geq 0. \tag{2}$$

Here, we also assume that the insurer receives interest from his surplus at a constant rate $\delta > 0$. Then, the surplus process (2) becomes

$$\begin{aligned} Y(t) &= e^{\delta t}\left(u + \int_0^t e^{-\delta s} dX(s)\right) \\ &= e^{\delta t}\left(u + \int_0^t e^{-\delta s} d\left(U(s) - D(s)\right)\right), \quad t \geq 0. \end{aligned} \tag{3}$$

In the actuarial literature, the issue of dividend strategies has received remarkable attention recently. The study of the optimal dividend problem goes back to De Finetti (1957). Due to its practical importance, much research on dividend-payment problems has been carried out for various surplus processes since then. For example, see Gerber and Shiu (2006), Lin and Pavlova (2006), Yuen et al. (2007, 2008a, 2008b), and references therein.

Under the threshold dividend strategy, dividends are paid at the maximal admissible rate $\alpha < c$ whenever the surplus is above the threshold level b, and that no dividends are paid whenever the surplus is below b. For the surplus process (2) under the threshold strategy, Gerber and Shiu (2006) examined the optimal dividend problems and derived a rule for deciding between plowback and dividend payout while Lin and Pavlova (2006) derived and solved two integro-differential equations for the expected discounted penalty function. For the surplus process (3), Fang and Wu (2007) studied the optimal dividend problems under the threshold strategy.

In this paper, we extend the work of Fang and Wu (2007) to investigating the expected discounted penalty function which embraces many important actuarial functions. In Section 2, we derive integro-differential equations for the expected discounted penalty function. In Section 3, a few examples with closed-form expressions for the expected discounted penalty function are presented.

2 Expected discounted penalty function under the threshold strategy

Under the threshold strategy, the surplus process (3) can be rewritten as

$$dY(t) = \begin{cases} cdt - dS(t) + \delta Y(t)dt, & \text{if} \quad Y(t) < b, \\ (c-\alpha)dt - dS(t) + \delta Y(t)dt, & \text{if} \quad Y(t) > b, \end{cases} \tag{4}$$

for $t \geq 0$, where $\alpha < c$ is the dividend rate, and $b > 0$ is the threshold level. Let $T = \inf\{t : Y(t) < 0\}$ be the time of ruin. Then, for the surplus process (4), the expected discounted penalty function introduced by Gerber-Shiu (1998) has the form

$$L_b(u) = E^u \left[e^{-\gamma T} w(Y(T-), |Y(T)|) I(T < \infty) \right], \tag{5}$$

where the penalty function w is a nonnegative measurable function on $[0, \infty) \times [0, \infty)$, $Y(T-)$ is the surplus just prior to ruin, $|Y(T)|$ is the deficit at ruin, $I(B)$ is the indicator function of event B, and the parameter $\gamma \geq 0$ can be interpreted as a force of interest. It is obvious that (5) is the expectation of the discounted value of the the penalty function depending on the surplus just prior to ruin and the deficit at ruin.

The expected discounted penalty function is a very useful technical tool for studying various ruin problems in modern risk theory. It includes many important actuarial functions, for example, the probability of ruin ($\gamma = 0$ and $w(Y(T-), |Y(T)|) \equiv 1$), the Laplace transform of the time of ruin ($w(Y(T-), |Y(T)|) \equiv 1$), the distribution of the surplus just prior to ruin ($\gamma = 0$ and $w(Y(T-), |Y(T)|) = I(Y(T-) \leq x)$, the distribution of the deficit at ruin ($\gamma = 0$ and $w(Y(T-), |Y(T)|) = I(|Y(T)| \leq y)$), and the joint distribution of the surplus just prior to ruin and the deficit at ruin ($\gamma = 0$ and $w(Y(T-), |Y(T)|) = I(Y(T-) \leq x, |Y(T)| \leq y)$).

To show the continuity of $L_b(u)$ in u, especially, the continuity at $u = b$, we first derive some integral equations for (5). Define

$$\phi_1(u, t) = e^{\delta t} \left(u + c \int_0^t e^{-\delta s} ds \right),$$
$$\phi_2(u, t) = e^{\delta t} \left(u + (c - \alpha) \int_0^t e^{-\delta s} ds \right).$$

Let

$$t_b = \frac{1}{\delta} \ln \left(\frac{\delta b + c}{\delta u + c} \right),$$

such that $\phi_1(u, t_b) = b$. By conditioning on the time and the amount of the first claim, we obtain the following integral equations for $L_b(u)$

$$L_b(u) = \int_0^{t_b} \lambda e^{-(\lambda+\gamma)t} \xi(\phi_1(u, t), b) dt + \int_{t_b}^{\infty} \lambda e^{-(\lambda+\gamma)t} \xi(\phi_2(b, t - t_b), b) dt, \quad 0 \leq u < b, \tag{6}$$

and

$$L_b(u) = \int_0^{\infty} \lambda e^{-(\lambda+\gamma)t} \xi(\phi_2(u, t), b) dt, \quad u \geq b, \tag{7}$$

where

$$\xi(u,b) = \int_0^u L_b(u-z)dF(z) + \zeta(u),$$

and

$$\zeta(u) = \int_u^\infty w(u, z-u)dF(z). \tag{8}$$

By changing variables, we derive from (6) and (7) that

$$\begin{aligned} L_b(u) = &\int_u^b \lambda \left(\frac{\delta u + c}{\delta y + c}\right)^{\frac{\lambda+\gamma}{\delta}} \frac{\xi(y,b)}{\delta y + c} dy \\ &+ \int_b^\infty \lambda \left(\frac{(\delta u + c)(\delta b + c - \alpha)}{(\delta b + c)(\delta y + c - \alpha)}\right)^{\frac{\lambda+\gamma}{\delta}} \frac{\xi(y,b)}{\delta y + c - \alpha} dy, \end{aligned} \tag{9}$$

for $0 \le u < b$, and

$$L_b(u) = \int_u^\infty \lambda \left(\frac{\delta u + c - \alpha}{\delta y + c - \alpha}\right)^{\frac{\lambda+\gamma}{\delta}} \frac{\xi(y,b)}{\delta y + c - \alpha} dy, \tag{10}$$

for $u \ge b$. Then, it follows from (9) and (10) that $L_b(u)$ is continuous on $[0, \infty)$. In particular, we have

$$L_b(b^-) = L_b(b) = L_b(b^+), \quad b > 0. \tag{11}$$

An integro-differential equation is an equation that involves both integrals and derivatives of an unknown function. The theory of integro-differential equations is close in spirit to the classical ordinary differential equations. In the actuarial literature, many ruin problems were often studied by means of integro-differential equations. For instance, see Lin and Pavlova (2006) and Yuen et al. (2007, 2008a, 2008b). Here, we also employ this method for studying the expected discounted penalty function $L_b(u)$.

Differentiating both sides of (9) and (10) with respect to u shows that $L_b(u)$ satisfies the integro-differential equations

$$(c + \delta u)L_b'(u) - (\lambda + \gamma)L_b(u) + \lambda \int_0^u L_b(u-z)dF(z) + \lambda\zeta(u) = 0, \tag{12}$$

for $0 \le u < b$, and

$$(c - \alpha + \delta u)L_b'(u) - (\lambda + \gamma)L_b(u) + \lambda \int_0^u L_b(u-z)dF(z) + \lambda\zeta(u) = 0, \tag{13}$$

for $u \ge b$. It is obvious that

$$\lim_{u\to\infty} L_b(u) = 0. \tag{14}$$

Hence, the expected discounted penalty function $L_b(u)$ can be determined using (12) and (13) with the boundary conditions (11) and (14). Moreover, (12) and (13) imply that

$$(c+\delta b)L_b'(b^-) = (c-\alpha+\delta b)L_b'(b^+).$$

Thus, $L_b'(u)$ may not be continuous at $u = b$.

3 Examples

In this section, we present a few special cases of the expected discounted penalty function (5) in which closed-form solutions to (12) and (13) can be derived.

Denote the probability of ruin for the surplus process (4) by $\Psi(u) = P(T < \infty)$. It is obvious that if $\gamma = 0$ and $w \equiv 1$, $L_b(u) = E[I(T < \infty)] = \Psi(u)$. Hence, it follows from (12) and (13) that $\Psi(u)$ satisfies the integro-differential equations

$$(c+\delta u)\Psi'(u) - \lambda\Psi(u) + \lambda\int_0^u \Psi(u-z)dF(z) + \lambda\overline{F}(u) = 0,\ 0 \le u < b, \quad (15)$$

and

$$(c-\alpha+\delta u)\Psi'(u) - \lambda\Psi(u) + \lambda\int_0^u \Psi(u-z)dF(z) + \lambda\overline{F}(u) = 0,\ u \ge b, \quad (16)$$

where $\overline{F}(u) = 1 - F(u)$. The boundary conditions for determining $\Psi(u)$ are

$$\lim_{u\to\infty}\Psi(u) = 0 \quad \text{and} \quad \Psi(b^-) = \Psi(b^+). \quad (17)$$

Example 3.1. Let $f(z)$ be exponential with mean β^{-1}. It follows from (15) and (16) that $\Psi(u)$ satisfies the differential equations

$$(c+\delta u)\Psi''(u) + (\beta(c+\delta u) + \delta - \lambda)\Psi'(u) = 0, \quad 0 \le u < b, \quad (18)$$

and

$$(c-\alpha+\delta u)\Psi''(u) + (\beta(c-\alpha+\delta u) + \delta - \lambda)\Psi'(u) = 0, \quad u \ge b. \quad (19)$$

Define

$$K(u,c) = \int_0^u e^{-\int_0^t \left(\beta + \frac{\delta-\lambda}{c+\delta s}\right)ds}dt, \quad u \ge 0. \quad (20)$$

Let K' be the derivative of K with respect to u. So,

$$K'(b,c) = (c+\delta b)^{\frac{\lambda}{\delta}-1}c^{1-\frac{\lambda}{\delta}}e^{-\beta b}.$$

Put

$$M = -(c-\alpha+\delta b)(c+\lambda K(b,c))K'(b,c-\alpha) \\ +\lambda(c+\delta b)K'(b,c)\left(K(b,c-\alpha)-K(\infty,c-\alpha)\right). \quad (21)$$

Solving the differential equations (18) and (19) with the boundary conditions in (17) as well as

$$(c+\delta b)\Psi'(b^-) = (c-\alpha+\delta b)\Psi'(b^+), \quad \text{and} \quad c\Psi'(0^+)-\lambda\Psi(0) = -\lambda,$$

we have

$$\Psi(u) = \begin{cases} a_1K(u,c)+a_2, & 0 \le u < b, \\ a_3K(u,c-\alpha)+a_4, & u \ge b, \end{cases} \quad (22)$$

where

$$a_1 = M^{-1}\lambda(c-\alpha+\delta b)K'(b,c-\alpha), \quad (23)$$

$$a_2 = M^{-1}\big(-\lambda(c-\alpha+\delta b)K(b,c)K'(b,c-\alpha) \\ +\lambda(c+\delta b)K'(b,c)(K(b,c-\alpha)-K(\infty,c-\alpha))\big), \quad (24)$$

$$a_3 = M^{-1}\lambda(c+\delta b)K'(b,c), \quad (25)$$

$$a_4 = -M^{-1}\lambda(c+\delta b)K'(b,c)K(\infty,c-\alpha). \quad (26)$$

From (21)-(26), closed-form expressions for the probability of ruin $\Psi(u)$ on $[0,\infty)$ can be obtained.

Let $c = 1.6, \alpha = 0.4, \lambda = \beta = 1, \delta = 0.02$, and $b = 8$. Then,

$$\Psi(u) = \begin{cases} -0.2414K(u,1.6)+0.6137, & 0 \le u < b, \\ -0.0724K(u,1.2)+0.3149. & u \ge b. \end{cases}$$

From Table 1 ,we see that the ruin probability $\Psi(u)$ is a decreasing function of u (for each b) and also a decreasing function of b (for each u).

$u\backslash b$	1	3	5	8	10	15	25	40
0	0.7404	0.6725	0.6349	0.6137	0.6095	0.6071	0.6069	0.6069
1	0.6061	0.5031	0.4460	0.4138	0.4073	0.4038	0.4035	0.4035
2	0.4741	0.3889	0.3187	0.2791	0.2711	0.2668	0.2664	0.2664
5	0.2136	0.1818	0.1397	0.0897	0.0797	0.0742	0.0737	0.0737
8	0.0882	0.0751	0.0577	0.0362	0.0256	0.0198	0.0193	0.0193
12	0.0239	0.0203	0.0156	0.0098	0.0072	0.0035	0.0030	0.0029
20	0.0012	0.0010	0.0008	0.0005	0.0004	0.0002	0.0001	0.0001

Table 1. Ruin probabilities with $c = 1.6, \alpha = 0.4, \lambda = \beta = 1, \delta = 0.02$, and $b = 8$.

□

Example 3.2. Let $f(x) = A_1\mu e^{-\mu x} + A_2\nu e^{-\nu x}, x > 0$, where $\mu, \nu, A_1, A_2 > 0$ are constants and $A_1 + A_2 = 1$. Without loss of generality, let $\theta = \mu - \nu > 0$.

Because

$$\left(\frac{d}{du}+\nu\right)\left(\frac{d}{du}+\mu\right)\int_0^u \Psi(u-z)f(z)dz = \mu\nu\Psi(u),$$

applying the operator $(d/du+\nu)(d/du+\mu)$ to (15) and (16) yields

$$\begin{aligned}&(c+\delta u)\Psi'''(u)+(2\delta-\lambda+(\mu+\nu)(c+\delta u))\,\Psi''(u)\\&\quad+((c+\delta u)\mu\nu-\lambda(A_2\mu+A_1\nu)+(\mu+\nu)\delta)\,\Psi'(u)=0,\quad 0\le u<b,\end{aligned}\tag{27}$$

and

$$\begin{aligned}&(c-\alpha+\delta u)\Psi'''(u)+(2\delta-\lambda+(\mu+\nu)(c-\alpha+\delta u))\,\Psi''(u)\\&\quad+\ ((c-\delta+\delta u)\mu\nu-\lambda(A_2\mu+A_1\nu)+(\mu+\nu)\delta)\,\Psi'(u)=0,\quad u\ge b.\end{aligned}\tag{28}$$

Substituting $c+\delta u=-\delta z/\theta$ and $\Psi'(u)=e^{-\mu z/\theta}y(z)$ into (27), we get the confluent hypergeometric form

$$zy''(z)+\left(2-z-\frac{\lambda}{\delta}\right)y'(z)+\left(1-\frac{A_1\lambda}{\delta}\right)y(z)=0.$$

Hence, from Slater (1960) (also see Example 4.2 of Yuen and Wang (2005) and Example 4.2 of Yuen et al. (2007)), we get two independent solutions

$$\begin{aligned}\Psi_1'(u,c)&=e^{-\mu u}\left(\frac{c}{\delta}+u\right)^{\frac{\lambda}{\delta}-1}G_1\left(\frac{A_2\lambda}{\delta},\frac{\lambda}{\delta},\theta\left(\frac{c}{\delta}+u\right)\right),\\ \Psi_2'(u,c)&=e^{-\mu u}\left(\frac{c}{\delta}+u\right)^{\frac{\lambda}{\delta}-1}G_2\left(\frac{A_2\lambda}{\delta},\frac{\lambda}{\delta},\theta\left(\frac{c}{\delta}+u\right)\right),\end{aligned}$$

where

$$G_1(a,d,u)=\frac{\Gamma(d)}{\Gamma(d-a)\Gamma(a)}\int_0^1 e^{ut}t^{a-1}(1-t)^{d-a-1}dt,\ d>a>0,\ u\ge 0,$$

is the standard confluent hypergeometric function, and its second form is

$$G_2(a,d,u)=\frac{1}{\Gamma(a)}\int_0^\infty e^{-ut}t^{a-1}(1+t)^{d-a-1}dt,\quad a>0,u\ge 0.$$

Since

$$\begin{aligned}G_1'(a,d,u)&=\frac{a}{d}G_1(a+1,d+1,u),\\ G_2'(a,d,u)&=-aG_2(a+1,d+1,u),\end{aligned}$$

for $u > 0$, we have

$$\Psi_1''(u,c) = \left(\frac{\lambda-\delta}{c+\delta u} - \mu\right)\Psi_1'(u,c) + A_2 e^{-\mu u}\left(\frac{c}{\delta}+u\right)^{\frac{\lambda}{\delta}-1} G_1\left(1+\frac{A_2\lambda}{\delta}, 1+\frac{\lambda}{\delta}, \theta\left(\frac{c}{\delta}+u\right)\right),$$

and

$$\Psi_2''(u,c) = \left(\frac{\lambda-\delta}{c+\delta u} - \mu\right)\Psi_2'(u,c) - \frac{A_2\lambda}{\delta} e^{-\mu u}\left(\frac{c}{\delta}+u\right)^{\frac{\lambda}{\delta}-1} G_2\left(1+\frac{A_2\lambda}{\delta}, 1+\frac{\lambda}{\delta}, \theta\left(\frac{c}{\delta}+u\right)\right).$$

Set

$$\Psi_i(u) = \int_0^u \Psi_i'(x,c)dx, \quad i = 1,2, \quad 0 \le u < b.$$

Then, the general solution to (27) has the form

$$\Psi(u) = c_0 + c_1\Psi_1(u) + c_2\Psi_2(u), \quad 0 \le u < b, \tag{29}$$

where c_0, c_1, and c_2 are arbitrary constants. Along the same line, a similar expression for (28) can be derived. Following the steps in the derivation of (29), we obtain the following general solution to (28)

$$\Psi(u) = c_3 + c_4\overline{\Psi}_1(u) + c_5\overline{\Psi}_2(u), \quad u \ge b,$$

where

$$\overline{\Psi}_i(u) = \int_u^\infty \Psi_i'(x, c-\alpha)dx, \quad i = 1,2, \quad u \ge b,$$

and c_3, c_4, and c_5 are arbitrary constants.

It follows from (15)-(17) that

$$\lim_{u\to\infty} \Psi(u) = 0, \tag{30}$$

$$\Psi(b^-) = \Psi(b^+), \tag{31}$$

$$(c+\delta b)\Psi'(b^-) = (c-\alpha+\delta b)\Psi'(b^+), \tag{32}$$

$$c\Psi'(0^+) - \lambda\Psi(0) = -\lambda. \tag{33}$$

It is easy to see that (30) implies $c_3 = 0$.

Differentiating (27) and (28) with respect to u yields

$$\begin{aligned}&(c+\delta u)\Psi''(u) + (\delta-\lambda)\Psi'(u) + \lambda\Psi(u)(A_1\mu + A_2\nu)\\ &\quad - \lambda(A_1\mu e^{-\mu u} + A_2\nu e^{-\nu u})\\ &\quad - \lambda\int_0^u \Psi(z)\left(A_1\mu^2 e^{-\mu(u-z)} + A_2\nu^2 e^{-\nu(u-z)}\right)dz = 0, \quad 0 \le u < b,\end{aligned} \tag{34}$$

and

$$\begin{aligned}&(c-\alpha+\delta u)\Psi''(u)+(\delta-\lambda)\Psi'(u)+\lambda\Psi(u)(A_1\mu+A_2\nu)\\&\quad-\lambda(A_1\mu e^{-\mu u}+A_2\nu e^{-\nu u})\\&\quad-\lambda\int_0^u\Psi(z)\left(A_1\mu^2e^{-\mu(u-z)}+A_2\nu^2e^{-\nu(u-z)}\right)dz=0,\quad u\geq b.\end{aligned}\tag{35}$$

Because of the continuity of $\Psi(u)$ at b, (34) and (35) lead to the boundary condition

$$(c+\delta b)\Psi''(b^-)+(\delta-\lambda)\Psi'(b^-)=(c-\alpha+\delta b)\Psi''(b^+)+(\delta-\lambda)\Psi'(b^+).\tag{36}$$

Furthermore, (33) and (35) with $u=0$ give the boundary condition

$$c\Psi''(0^+)+(\delta-\lambda+c(A_1\mu+A_2\nu))\Psi'(0^+)=0.\tag{37}$$

Let

$$\begin{aligned}
d &= \delta-\lambda+c(A_1\mu+A_2\nu),\\
d_{12} &= \Psi_1(b^-), & d_{13} &= \Psi_2(b^-),\\
d_{14} &= -\overline{\Psi}_1(b^+), & d_{15} &= -\overline{\Psi}_2(b^+),\\
d_{22} &= -c\Psi_1'(0^+), & d_{23} &= -c\Psi_2'(0^+),\\
d_{32} &= (c+\delta b)\Psi_1'(b^-), & d_{33} &= (c+\delta b)\Psi_2'(b^-),\\
d_{34} &= -(c-\alpha+\delta b)\overline{\Psi}_1'(b^+), & d_{35} &= -(c-\alpha+\delta b)\overline{\Psi}_2'(b^+),\\
d_{42} &= c\Psi_1''(0^+)+d\Psi_1'(0^+), & d_{43} &= c\Psi_2''(0^+)+d\Psi_2'(0^+),\\
d_{52} &= (c+\delta b)\Psi_1''(b^-)+(\delta-\lambda)\Psi_1'(b^-),\\
d_{53} &= (c+\delta b)\Psi_2''(b^-)+(\delta-\lambda)\Psi_2'(b^-),\\
d_{54} &= -(c-\alpha+\delta b)\overline{\Psi}_1''(b^+)-(\delta-\lambda)\overline{\Psi}_1'(b^+),\\
d_{55} &= -(c-\alpha+\delta b)\overline{\Psi}_2''(b^+)-(\delta-\lambda)\overline{\Psi}_2'(b^+).
\end{aligned}$$

To determine $c_i,\ i=0,1,2,4,5$, we define the matrix $\mathbf{M}$ and the column vector $\mathbf{B}$ as

$$\begin{pmatrix}1 & d_{12} & d_{13} & d_{14} & d_{15}\\ \lambda & d_{22} & d_{23} & 0 & 0\\ 0 & d_{32} & d_{33} & d_{34} & d_{35}\\ 0 & d_{42} & d_{43} & 0 & 0\\ 0 & d_{52} & d_{53} & d_{54} & d_{55}\end{pmatrix},$$

and

$$\mathbf{B}=(0,-\lambda,0,0,0)^T.$$

Let $\mathbf{M}_i$ denote the matrix with the form of $\mathbf{M}$ except that the i-th column of $\mathbf{M}$ is replaced by $\mathbf{B}$. Denote the determinant of a matrix by $\det(\cdot)$. Using boundary conditions (31)-(33), (36) and (37), we get

$$c_i=\frac{\det(\mathbf{M}_i)}{\det(\mathbf{M})},\quad i=0,1,2,4,5.$$

□

Let $H_b(u) = E^u[h(|Y(T)|)I(T < \infty)]$ denotes the expectation of the penalty of the deficit at ruin for the surplus process (4). It can be obtained by letting $\gamma = 0$ and $w(x_1, x_2) = h(x_2)$ in (5) of $L_b(u)$. In this case, $\zeta(u)$ of (8) becomes

$$\zeta(u) = \int_0^\infty h(z)f(u+z)dz.$$

From (12) and (13), we see that $H_b(u)$ satisfies the integro-differential equations

$$(c+\delta u)H_b'(u) - \lambda H_b(u) + \lambda \int_0^u H_b(u-z)f(z)dz + \lambda \int_0^\infty h(z)f(u+z)dz = 0, \tag{38}$$

for $0 \le u < b$, and

$$(c-\alpha+\delta u)H_b'(u) - \lambda H_b(u) + \lambda \int_0^u H_b(u-z)f(z)dz + \lambda \int_0^\infty h(z)f(u+z)dz = 0, \tag{39}$$

for $u \ge b$, with the boundary conditions

$$\lim_{u\to\infty} H_b(u) = 0 \quad \text{and} \quad H_b(b^-) = H(b^+). \tag{40}$$

Therefore, one can use (38)-(40) to find closed-form expressions for $H_b(u)$.

□

Example 3.3. Let $f(z)$ be exponential with mean β^{-1}. Hence, $\zeta(u) = \overline{H}e^{-\beta u}$ with $\overline{H} = \int_0^\infty h(z)\beta e^{-\beta z}dz$. Then, it follows from (38) and (39) that $H_b(u)$ satisfies the differential equations

$$(c+\delta u)H_b''(u) + (\beta(c+\delta u) + \delta - \lambda)H_b'(u) = 0, \quad 0 \le u < b \tag{41}$$

and

$$(c-\alpha+\delta u)H_b''(u) + (\beta(c-\alpha+\delta u) + \delta - \lambda)H_b'(u) = 0, \quad u \ge b, \tag{42}$$

with the boundary conditions in (40) as well as

$$(c+\delta b)H_b'(b^-) = (c-\alpha+\delta b)H_b'(b^+), \quad \text{and} \quad cH_b'(0^+) - \lambda H_b(0^+) = -\lambda\overline{H}. \tag{43}$$

Similar to (22), the general solutions to (41) and (42) are given by

$$H_b(u) = \begin{cases} e_1K(u,c) + e_2, & 0 \le u < b \\ e_3K(u, c-\alpha) + e_4, & u \ge b, \end{cases}$$

where K is defined in (20). Using the boundary conditions in (40) and (43), we get

$$e_i = \overline{H} a_i, \quad i = 1, 2, 3, 4,$$

where a_i's are given in (23)-(26). Hence,

$$H_b(u) = \overline{H}\Psi(u), \quad u \geq 0. \tag{44}$$

From (44) and the result in Example 3.1, one can obtain closed-form expressions for $H_b(u)$.

From (44), we see that the expected penalty of the deficit at ruin is proportional to the probability of ruin if the claims are exponentially distributed. Applying (44) yields

$$H_b(u) = \begin{cases} n!\beta^{-n}\Psi(u), \text{ if } & h(x) = x^n, \\ \frac{\beta}{\eta+\beta}\Psi(u), \quad \text{if } & h(x) = e^{-\eta x}, \text{ and } \eta > 0, \end{cases} \tag{45}$$

for $u \geq 0$. That is, (45) gives the nth moment and the Laplace transform of the deficit at ruin when the claims are exponentially distributed. □

Acknowledgements

The research of Yinghui Dong is supported by the National Natural Science Foundation of China (Grant No. 10571132). The research of Kam C. Yuen was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 7475/05H).

References

DE FINETTI, B. (1957): Su un' impostazione alternativa dell teoria collettiva del rischio. *Transactions of the XVth International Congress of Actuaries 2, 433-443.*

FANG, Y. and WU, R. (2007): Optimal Dividend strategy in the Compound Poisson Model with constant interest. *Stochastic Model 23, 149-166.*

GERBER, H.U. and SHIU, E.S.W. (1998): On the time value of ruin. *North American Actuarial Journal 2 (1), 48-78.*

GERBER, H.U. and SHIU, E.S.W. (2006): On optimal dividend strategies in the compound Poisson model. *North American Actuarial Journal 10 (2), 76-93.*

LIN X.S. and PAVLOVA, K.P. (2006): The compound Poisson risk model with a threshold dividend strategy. *Insurance: Mathematics and Economics 38, 57-80.*

SLATER, L.J. (1960): *Confluent Hypergeometric Functions.* Cambridge University Press, London.

YUEN, K.C. and WANG, G. (2005). Some ruin problems for a risk process with stochastic interest. *North American Actuarial Journal 9 (3), 129-142.*

YUEN, K.C., LU, Y. and WU, R. (2008a): The compound Poisson process perturbed by diffusion with a threshold dividend strategy. *Applied Stochastic Models in Business and Industry*, to appear.

YUEN, K.C., WANG, G. and LI, W.K. (2007): The Gerber-Shiu expected discounted penalty function for risk processes with interest and a constant dividend barrier. *Insurance: Mathematics and Economics 40, 107-112*.

YUEN, K.C., ZHOU, M. and GUO, J. (2008b): On a risk model with debit interest and dividend payments. *Statistics and Probability Letters*, to appear.

Estimation of Structural Parameters in Crossed Classification Credibility Model Using Linear Mixed Models

Wing K. Fung and Xiaochen Xu

Department of Statistics and Actuarial Science, The University of Hong Kong
Pokfulam Road, Hong Kong, *wingfung@hku.hk*

Abstract. In this paper, the linear mixed model is used under the Dannenburg's two-way crossed classification model. Maximum likelihood (ML) and restricted maximum likelihood (REML) methods are employed to estimate the structural parameters with both independent and exchangeable error structures. Evidenced by results of simulation studies, the proposed linear mixed effects estimators appear to outperform those given by Dannenburg with both independent and exchangeable error structures.

Keywords: linear mixed model, crossed classification credibility model, maximum likelihood estimator, restricted maximum likelihood estimator, SAS

1 Introduction

In credibility context, the credibility data can be treated as longitudinal data. Therefore the longitudinal data interpretation suggests additional techniques that actuaries can use in credibility ratemaking. The theoretical development of credibility theory has been linked with random effects models. Linear mixed models can be used to capture the data structure of repeated measurements and longitudinal data. Hence the implementation of linear mixed models can help us to capture the within panel correlation structure under credibility context.

Frees et al. (1999) and Antonio and Beirlant (2006) have demonstrated the implementation of the linear mixed model and generalized linear mixed model in the classical credibility models with the assumption that the consecutive error terms with regard to the same risk entity are independent. Later developments given by Cossette and Luong (2003), Lo, Fung and Zhu (2006) and Lo, Fung and Zhu (2007) have employed regression models in credibility context, and proposed the weighted least squares method and the generalized estimating equations (GEE) respectively in order to estimate the stuctural parameters with the presence of an assumed correlation structure for error terms. In both literatures, their proposed methods have been examined by simulation studies.

Since the Dannenburg's two-way crossed classification model is of the form of a linear mixed model, we simply apply the maximum likelihood (ML) and restricted maximum likelihood (REML) estimation methods to estimate the structural parameters. It can be shown that the linear mixed effects estimation method can outperform the Dannenburg's method in general with both independent and exchangeable error structures.

The structure of this paper is organized as follows. In section 2, linear mixed models are introduced, and the definition of the exchangeable correlation is provided. Section 3 gives a brief introduction on ML and REML methods, and their applications to the linear mixed model. Section 4 discusses the Dannenburg's crossed classification model. Several simulation study results are analyzed in section 5.

2 Model specification

2.1 Linear mixed model

In this paper, we employ the linear mixed model which extends the classical linear model by incorporating the random effects. Linear mixed models have the following form:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, 2, \ldots, n. \tag{1}$$

Each element y_{ij} in the $n_i \times 1$ vector $\mathbf{y}_i$ corresponds to the observed value or the realized value of some measurable characteristic as regards observation j of the i^{th} sector or group. $\mathbf{X}_i$, of dimensions $n_i \times m$, enters the model as a known constant matrix. $\mathbf{Z}_i$, of dimensions $n_i \times q$, is another explanatory matrix. The dimension of the vector of regression coefficients $\boldsymbol{\alpha}_i$, labeled q, is essentially the size of the covariance matrix present in our model. $\boldsymbol{\alpha}_i$'s are assumed to be independent and normally distributed, with common mean $\mathbf{0}$ and covariance matrix $\mathbf{F}$ for all i. And the vector $\boldsymbol{\beta}$ captures the mean of the random variable $\mathbf{y}_i$. The error vectors $\boldsymbol{\varepsilon}_i$'s are assumed to be independently distributed from normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{V}_i = \sigma^2\mathbf{W}_i^{-1/2}\boldsymbol{\Gamma}_i\mathbf{W}_i^{-1/2}$, where $\mathbf{W}_i$ is a weight matrix and $\boldsymbol{\Gamma}_i$ is a correlation matrix. $\mathbf{W}_i^{-1/2}$ is a square matrix with known positive constants along the principal diagonal and zero elements elsewhere. We assume $\boldsymbol{\Gamma}_i$, which describes the correlation between the error terms ε_{ij}'s for entity i, to be positive definite and depend on some fixed unknown parameters which are to be estimated. Aided by the specifications stated above, readers may then easily derive the following about $\mathbf{y}_i$:

a) $\mathbf{y}_i$ and $\mathbf{y}_j$ are statistically independent for $i \neq j$;

b) $E(\mathbf{y}_i|\boldsymbol{\alpha}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\alpha}_i$ and $\boldsymbol{\mu}_i = E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$;

c) $V(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{F}\mathbf{Z}_i^{'} + \sigma^2\mathbf{W}_i^{-1/2}\boldsymbol{\Gamma}_i\mathbf{W}_i^{-1/2}$.

2.2 Exchangeable correlation matrix

The exchangeable type of correlation is commonly used to model the error structure. It is also known as the uniform correlation. By using relatively few unknown parameters, it can help us to capture the correlation structure well. Under the Dannenburg's model, the observations are classified into different sectors. Therefore the observations in the same sector have certain similarities. If we assume the same correlation among observations in the same sector, it is reasonable to consider exchangeable correlation structure. There are also other correlation structures that we can consider, namely the moving average and the autoregressive types of correlation. In our empirical studies we have only incorporated the exchangeable error correlation structure into the Dannenburg's model for brevity.

The correlation matrix $(\Gamma_{ij})_{n\times n}$ of the exchangeable type of error can be written as:

$$\Gamma_{ij} = \begin{cases} 1, & \text{for } i = j, \\ \rho, & \text{otherwise.} \end{cases}$$

So the exchangeable correlation matrix takes the explicit form of

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \ddots & \\ \rho & \rho & 1 & \ddots & \\ & \ddots & \ddots & \ddots & \\ \rho & \cdots & \rho & \rho & 1 \end{bmatrix}.$$

3 The MLE and REML methods

In the regression credibility model, the variance and covariance parameters can be estimated using the well-known maximum likelihood (ML) and the restricted maximum likelihood (REML) estimation methods. Under normality the likelihood is portioned into two parts, one of which is free of fixed effects. REML estimators are obtained by maximizing the part that is free from fixed effects.

From our assumption, the error vectors, $\boldsymbol{\varepsilon}_i$, and regression coefficient vectors, $\boldsymbol{\alpha}_i$, are normally distributed, this implies $\mathbf{y}_i$ follows a multivariate normal distribution with derivable mean and variance-covariance matrix

$$\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta},\ \mathbf{X}_i\mathbf{F}\mathbf{X}_i' + \sigma^2\mathbf{W}_i^{-1/2}\boldsymbol{\Gamma}_i\mathbf{W}_i^{-1/2}).$$

Hence we can derive the log likelihood and the restricted log likelihood function of $\mathbf{y}_i$. They have been shown as

$$L_{ML} = c_1 - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{V}(\mathbf{y}_i)| - \frac{1}{2}\sum_{i=1}^{n}\mathbf{r}_i'\mathbf{V}(\mathbf{y}_i)\mathbf{r}_i, \tag{2}$$

$$L_{REML} = c_2 - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{V}(\mathbf{y}_i)| - \frac{1}{2}\log(\sum_{i=1}^{n}|\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i|) - \frac{1}{2}\sum_{i=1}^{n}\mathbf{r}_i'\mathbf{V}(\mathbf{y}_i)\mathbf{r}_i, \tag{3}$$

where

$$\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\left(\sum_{i=1}^{n}\mathbf{X}_i'\cdot\mathbf{V}^{-1}(\mathbf{y}_i)\cdot\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i'\cdot\mathbf{V}^{-1}(\mathbf{y}_i)\cdot\mathbf{y}_i\right),$$

and c_1, c_2 are appropriate constants.

To estimate the parameters that we are interested in, we define the vector $\boldsymbol{\gamma}$ to contain all of them. For example $\boldsymbol{\gamma} = (b, \sigma^2, \rho)'$, where b indicates the element of the covariance matrix $\mathbf{F}$ of $\boldsymbol{\alpha}_i$. We could solve $\boldsymbol{\gamma}$ by maximizing the log likelihood function with respect to $\boldsymbol{\gamma}$ or by solving the score function

$$\left.\frac{\partial L_{ML}}{\partial\boldsymbol{\gamma}}\right|_{\gamma=\hat{\gamma}} = 0$$

for the ML approach, and

$$\left.\frac{\partial L_{REML}}{\partial\boldsymbol{\gamma}}\right|_{\gamma=\hat{\gamma}} = 0$$

for the REML approach. More details about the derivation of the likelihood and restricted likelihood functions, fixed and random effects, estimates of the variance and covariance components can be found in Laird and Ware (1982), McCulloch (1997) and Verbeke and Molenberghs (2000).

Computationally there are various ways to obtain the ML and the REML estimators, for example, the Newton-Raphson method and the simplex algorithm. Details of those methods can be found in Lindstrom and Bates (1988) and Nelder and Mead (1965) respectively. There are also many statistical packages available that can be used to perform such estimation, for example, *SAS*, *Matlab*, *R* and *S+*.

4 Dannenburg's model and method

Dannenburg et al. (1996) proposed the two-way crossed classification model. In Dannenburg's model, the risk factors are treated in a symmetrical way, instead of fully nested. The model is a two-way analysis model with interaction

terms and having random effects. It takes the following form:

$$y_{ijt} = \beta + \alpha_i^{(1)} + \alpha_j^{(2)} + \alpha_{ij}^{(12)} + \epsilon_{ijt}, \quad t = 1, \cdots, T_{ij}. \tag{4}$$

In this model, there are two risk factors, 1 and 2. The number of categories of the first factor is I and of the second risk factor is J. An insurance portfolio which is subdivided by these two risk factors can be viewed as a two-way table. Suppose I is 2, J is 3. We have

$$\begin{array}{rc} \alpha_1^{(1)} + \rightarrow & \begin{array}{|c|c|c|} \hline +\alpha_{11}^{(12)} & +\alpha_{12}^{(12)} & +\alpha_{13}^{(12)} \\ \hline +\alpha_{21}^{(12)} & +\alpha_{22}^{(12)} & +\alpha_{23}^{(12)} \\ \hline \end{array} \\ \alpha_2^{(1)} + \rightarrow & \end{array}$$

$$\begin{array}{ccc} \uparrow & \uparrow & \uparrow \\ +\alpha_1^{(2)} & +\alpha_2^{(2)} & +\alpha_3^{(2)} \end{array}$$

The first risk factor $\alpha_i^{(1)}$ can be called the row factor. The second risk factor $\alpha_j^{(2)}$ can be called the column factor. The structural parameters are defined as follows:

$$Var(\alpha_i^{(1)}) = b^{(1)}, \quad Var(\alpha_j^{(2)}) = b^{(2)},$$

$$Var(\alpha_{ij}^{(12)}) = a, \quad Var(\epsilon_{ijt}) = s^2/w_{ijt}.$$

The credibility estimator of $y_{ij,T_{ij}+1}$ is equal to (Dannenburg et al., 1996):

$$\begin{aligned} y_{ij,T_{ij}+1} = \beta + z_{ij}(y_{ijw} - \beta) + (1 - z_{ij}) z_i^{(1)}(x_{izw} - \beta) \\ + (1 - z_{ij}) z_j^{(2)}(x_{zjw} - \beta), \end{aligned} \tag{5}$$

where the credibility factors are

$$z_{ij} = \frac{a}{a + \sigma^2/w_{ij\Sigma}}, \quad \text{with} \quad w_{ij\Sigma} = \sum_t w_{ijt}, \tag{6}$$

$$z_i^{(1)} = \frac{b^{(1)}}{b^{(1)} + a/z_{i\Sigma}}, \quad \text{with} \quad z_{i\Sigma} = \sum_j z_{ij}, \tag{7}$$

$$z_j^{(2)} = \frac{b^{(2)}}{b^{(2)} + a/z_{\Sigma j}}, \quad \text{with} \quad z_{\Sigma j} = \sum_i z_{ij}. \tag{8}$$

x_{izw}, x_{zjw} are the adjusted weighted average, which can give us a much clearer view on the risk experience with regard to different risk factors,

$$x_{izw} = \sum_j \frac{z_{ij}}{z_{i\Sigma}} (y_{ijw} - \Xi_j^{(2)*}), \tag{9}$$

$$x_{zjw} = \sum_i \frac{z_{ij}}{z_{\Sigma j}} (y_{ijw} - \Xi_i^{(1)*}), \tag{10}$$

where

$$y_{ijw} = \sum_t \frac{w_{ijt}}{w_{ij\Sigma}} y_{ijt}.$$

And $\Xi_i^{(1)*}$, $\Xi_j^{(2)*}$ are the row effect and the column effect respectively. They can be found as the solution of the following $I + J$ linear equations using iterative approach.

$$\Xi_i^{(1)*} = z_i^{(1)} \Big[\sum_j \frac{z_{ij}}{z_{i\Sigma}} (y_{ijw} - \Xi_j^{(2)*}) - \beta \Big], \tag{11}$$

$$\Xi_j^{(2)*} = z_j^{(2)} \Big[\sum_i \frac{z_{ij}}{z_{\Sigma j}} (y_{ijw} - \Xi_i^{(1)*}) - \beta \Big]. \tag{12}$$

In Dannenburg's approach, the structural parameters β and s^2 can be found by the following equations (Dannenburg et al., 1996):

$$\beta = x_{www} = \sum_i \sum_j \frac{w_{ij\Sigma}}{w_{\Sigma\Sigma\Sigma}} y_{ijw}, \tag{13}$$

$$s^{2\bullet} = \frac{\sum_i \sum_j \sum_t w_{ijt} (y_{ijt} - y_{ijw})^2}{\sum_i \sum_j (T_{ij} - 1)_+}. \tag{14}$$

To obtain the estimators a, $b^{(1)}$ and $b^{(2)}$, we can solve the following linear equations (Dannenburg et al., 1996):

$$\begin{aligned} E\left[\tfrac{1}{I} \textstyle\sum_i \left(\sum_j \frac{w_{ij\Sigma}}{w_{i\Sigma\Sigma}} (y_{ijw} - y_{iww})^2 - s^{2\bullet}(J-1)/w_{i\Sigma\Sigma} \right)\right] \\ = (b^{(2)} + a) \left(1 - \tfrac{1}{I} \textstyle\sum_i \sum_j (\frac{w_{ij\Sigma}}{w_{i\Sigma\Sigma}})^2 \right), \end{aligned} \tag{15}$$

$$\begin{aligned} E\left[\tfrac{1}{J} \textstyle\sum_j \left(\sum_i \frac{w_{ij\Sigma}}{w_{\Sigma j\Sigma}} (y_{ijw} - y_{wjw})^2 - s^{2\bullet}(I-1)/w_{\Sigma j\Sigma} \right)\right] \\ = (b^{(1)} + a) \left(1 - \tfrac{1}{J} \textstyle\sum_j \sum_i (\frac{w_{ij\Sigma}}{w_{\Sigma j\Sigma}})^2 \right), \end{aligned} \tag{16}$$

$$\begin{aligned} E\Big[\textstyle\sum_i \sum_j \frac{w_{ij\Sigma}}{w_{\Sigma\Sigma\Sigma}} (y_{ijw} - y_{www})^2 - s^{2\bullet}(IJ-1)/w_{\Sigma\Sigma\Sigma} \Big] \\ = b^{(1)} \left(1 - \textstyle\sum_i (\frac{w_{i\Sigma\Sigma}}{w_{\Sigma\Sigma\Sigma}})^2 \right) + b^{(2)} \left(1 - \textstyle\sum_j (\frac{w_{\Sigma j\Sigma}}{w_{\Sigma\Sigma\Sigma}})^2 \right) \\ + a \left(1 - \textstyle\sum_i \sum_j (\frac{w_{ij\Sigma}}{w_{\Sigma\Sigma\Sigma}}) \right), \end{aligned} \tag{17}$$

where $y_{iww} = \sum_j \frac{w_{ij\Sigma}}{w_{i\Sigma\Sigma}} y_{ijw}$ and $y_{wjw} = \sum_i \frac{w_{ij\Sigma}}{w_{\Sigma j\Sigma}} y_{ijw}$. To find the "unbiased estimator" of a, $b^{(1)}$ and $b^{(2)}$, we can drop the expectation operation of the above linear equations.

Table 1
Estimation results for Study 1 associated with independent error structure.

Parameter		Method REML-I	REML-EX	Dannenburg
β	Bias	0.047	0.047	0.047
	MSE	2.051(1.00[a])	2.051 (1.00)	2.051
y^b	Bias	0.149	0.149	0.149
	MSE	21.050 (1.00)	21.050 (1.00)	21.050
z_{ij}	Bias	−0.049	−0.049	−0.049
	MSE	0.021 (1.00)	0.021 (1.00)	0.021
$z_i^{(1)}$	Bias	−0.018	−0.018	−0.018
	MSE	0.003 (1.00)	0.003 (1.00)	0.003
$z_j^{(2)}$	Bias	0.055	−0.055	−0.066
	MSE	0.029 (2.14)	0.029 (2.14)	0.062
$b^{(1)}$	Bias	0.031	0.031	0.031
	MSE	29.929 (1.00)	29.928 (1.00)	29.930
$b^{(2)}$	Bias	0.135	0.135	0.133
	MSE	9.174 (1.00)	9.174 (1.00)	9.186
a	Bias	−0.057	−0.057	−0.056
	MSE	0.207 (1.00)	0.208 (1.00)	0.208
s^2	Bias	0.124	0.124	0.124
	MSE	0.015 (1.00)	0.015 (1.00)	0.015

[a] Relative efficiency of the estimator. Dannenburg estimator serves as the baseline.
[b] Credibility premium, which can be written as $y_{ij,T_{ij}+1}$

5 Empirical studies

Since the Dannenburg's crossed classification model is of the form of linear mixed models, we could make use of the statistical packages that are designed especially for the parameter estimation in linear mixed models. One possible software is SAS. In our simulation studies, the results are obtained from the SAS procedure PROC MIXED. Since the simulation results for the ML and REML estimators are very similar, we only present the results for REML in this paper.

5.1 Study 1

In this study, we assign equal weight to each observation. The simulation study is based on the following choice of parameters:

$$I = 7, \quad J = 5, \quad T_{ij} = n = 20,$$

$$b^{(1)} = 9, \quad b^{(2)} = 4, \quad a = 1, \quad s^2 = 16.$$

Table 2
Estimation results for Study 1 associated with exchangeable error structure.

Parameter		Method REML-I	REML-EX	Dannenburg
β	Bias	−0.041	−0.041	−0.041
	MSE	2.182 (1.00)	2.182 (1.00)	2.182
y	Bias	0.150	0.150	0.150
	MSE	12.365 (1.00)	12.368 (1.00)	12.368
z_{ij}	Bias	0.357	0.357	0.357
	MSE	0.128 (1.00)	0.127 (1.00)	0.127
$z_i^{(1)}$	Bias	−0.112	−0.098	−0.098
	MSE	0.028 (0.86)	0.023 (1.04)	0.024
$z_j^{(2)}$	Bias	−0.252	−0.232	−0.275
	MSE	0.121 (2.83)	0.110 (3.09)	0.345
$b^{(1)}$	Bias	0.741	2.104	2.102
	MSE	37.061 (1.42)	52.619 (1.00)	52.626
$b^{(2)}$	Bias	−0.689	−0.205	−0.218
	MSE	8.004 (1.29)	10.169 (1.01)	10.278
a	Bias	4.222	4.201	4.214
	MSE	18.661 (1.00)	18.475 (1.01)	18.591
s^2	Bias	−6.248	−6.248	−6.248
	MSE	39.035 (1.00)	39.036 (1.00)	39.036

The error structure is independent for Table 1, and exchangeable with $\rho = 0.4$ for Table 2. The above scenario is simulated 500 times. We use two approaches to estimate the structural parameters.

1. Dannenburg: The unbiased estimators for s^2, a, $b^{(1)}$ and $b^{(2)}$ are computed using Equations (14), (15), (16) and (17).
2. REML: The restricted maximum likelihood estimation is used to compute the structural parameters. Two REML estimators are used in this paper. They are linked with the independent and exchangeable error structures and are denoted by REML-I and REML-EX respectively.

After we obtain the estimators of a, $b^{(1)}$ and $b^{(2)}$, we can use Equations (6), (7) and (8) to find the credibility factors z_{ij}, $z_i^{(1)}$ and $z_j^{(2)}$. The next step is to find the credibility estimator $y_{ij,T_{ij}+1}$ using Equation (5), while the parameters x_{izw}, x_{zjw}, $\Xi_i^{(1)*}$ and $\Xi_j^{(2)*}$ that are used in Equation (5) can be estimated using Equations (9), (10), (11) and (12).

As for the simulation results, we show the bias and mean square error (MSE) of the Dannenburg's and REML estimators for β, $y_{ij,T_{ij}+1}$, z_{ij}, $z_i^{(1)}$, $z_j^{(2)}$, $b^{(1)}$, $b^{(2)}$, a and s^2.

Table 3
Estimation results for Study 2 associated with independent error structure.

Parameter		Method REML-I	REML-EX	Dannenburg
β	Bias	8.35×10^{-1}	8.45×10^{-1}	-3.11×10^{-2}
	MSE	8.04×10^{1}	8.04×10^{1}	9.04×10^{1}
		(1.12)	(1.12)	
y	Bias	-8.19×10^{-1}	-8.20×10^{-1}	7.14×10^{-1}
	MSE	1.46×10^{3}	1.46×10^{3}	2.65×10^{4}
		(18.15)	(18.15)	
z_{ij}	Bias	-1.09×10^{-2}	-5.44×10^{-3}	-4.43×10^{-1}
	MSE	4.03×10^{-3}	5.25×10^{-3}	6.50×10^{2}
		$(> 10^{5})$	$(> 10^{5})$	
$z_i^{(1)}$	Bias	-3.68×10^{-3}	-3.05×10^{-3}	3.89×10^{-3}
	MSE	9.60×10^{-5}	9.27×10^{-5}	2.36×10^{-3}
		(24.58)	(25.46)	
$z_j^{(2)}$	Bias	-3.71×10^{-2}	-3.51×10^{-2}	5.16×10^{-2}
	MSE	1.29×10^{-2}	1.27×10^{-2}	1.47×10^{2}
		$(> 10^{4})$	$(> 10^{4})$	
$b^{(1)}$	Bias	-6.10×10^{1}	4.49	5.11
	MSE	1.28×10^{5}	1.46×10^{5}	1.50×10^{5}
		(1.17)	(1.03)	
$b^{(2)}$	Bias	-1.39×10^{-1}	3.76	3.21×10^{-1}
	MSE	3.35×10^{3}	3.73×10^{3}	1.78×10^{4}
		(5.31)	(4.77)	
a	Bias	-1.22×10^{-1}	1.85	-2.01
	MSE	1.31×10^{2}	1.95×10^{2}	1.18×10^{4}
		(90.08)	(60.51)	
s^2	Bias	2.18×10^{2}	2.61×10^{2}	2.60×10^{2}
	MSE	6.06×10^{4}	6.80×10^{4}	6.79×10^{4}
		(1.12)	(1.00)	

Table 1 summarizes the results when the data are independent, i.e. $\rho = 0$ when generating data. In term of MSE, the three estimation methods are quite close to each other, except that both REML-I and REML-EX methods show obvious advantage in estimating $z_j^{(2)}$. The results show that the "unbiased" estimator proposed by Dannenburg may not be the best, even when its independent error structure assumption coincides with the actual error structure.

When the assumption of independent error structure is violated, with reference to Table 2, we can see that REML-I still performs well in comparing with the Dannenburg's method, especially in reducing the bias and the mean squared error for $b^{(1)}$ and $b^{(2)}$. With the correct assumption about the error

Table4
Estimation results for Study 4 associated with exchangeable error structure.

Parameter		Method REML-I	REML-EX	Dannenburg
β	Bias	1.17	8.07×10^{-1}	-3.80×10^{-1}
	MSE	8.65×10^{1}	8.71×10^{1}	9.58×10^{1}
		(1.11)	(1.10)	
y	Bias	-1.35	-1.17	-1.49
	MSE	9.99×10^{2}	1.13×10^{3}	9.70×10^{2}
		(0.98)	(0.86)	
z_{ij}	Bias	3.39×10^{-1}	8.65×10^{-2}	4.17×10^{-1}
	MSE	1.21×10^{-1}	1.73×10^{-2}	2.17×10^{-1}
		(1.79)	(12.54)	
$z_i^{(1)}$	Bias	-3.28×10^{-2}	-2.20×10^{-3}	-6.92×10^{-2}
	MSE	1.60×10^{-3}	7.91×10^{-5}	6.87×10^{-3}
		(4.29)	(86.85)	
$z_j^{(2)}$	Bias	-1.75×10^{-1}	-8.41×10^{-2}	2.32×10^{-1}
	MSE	8.06×10^{-2}	6.06×10^{-2}	7.66×10^{-1}
		(9.50)	(12.64)	
$b^{(1)}$	Bias	-8.21×10^{1}	-3.17×10^{1}	-1.04×10^{2}
	MSE	1.28×10^{5}	1.36×10^{5}	1.45×10^{5}
		(1.13)	(1.07)	
$b^{(2)}$	Bias	-5.78	-7.26	-2.78×10^{2}
	MSE	5.13×10^{3}	5.62×10^{3}	1.01×10^{5}
		(19.69)	(17.97)	
a	Bias	1.52×10^{2}	6.95	3.51×10^{2}
	MSE	2.40×10^{4}	3.72×10^{2}	1.42×10^{5}
		(5.92)	(381.72)	
s^2	Bias	-9.41×10^{3}	-9.67×10^{3}	-9.67×10^{3}
	MSE	8.85×10^{7}	9.36×10^{7}	9.36×10^{7}
		(1.06)	(1.00)	

structure, REML-EX outperforms the Dannenburg's estimator in the estimation of all the parameters that we are interested in.

5.2 Study 2

The setting of Study 2 is similar to Study 1; the values of the parameters are chosen as follows:

$$I = 12, \quad J = 8, \quad T_{ij} = n = 10$$

$$b^{(1)} = 900, \quad b^{(2)} = 100, \quad a = 36, \quad s^2 = 25600.$$

In contrast to Study 1, the weight of the observations are not always the same. In this study, the observations are divided into $I \times J$ cells (96 cells).

We randomly select 48 cells, and this 48 cells have weight $w_{ijt} = 150$; the other cells have weight $w_{ijt} = 10$. The above scenario is simulated 500 times. And the sectors retain its weight which has been assigned during the first replicate. The error structure is independent for Table 3, and exchangeable with $\rho = 0.4$ for Table 4. The methods we considered in this study are the same as in Study 1.

With reference to Table 3, we can see that with independent error structure, significant advantage has been recorded for both the REML-I and REML-EX methods in estimating the structural parameters, especially in estimating $b^{(2)}$ and a. Hence the accuracy of the estimation for the credibility estimators $z_i^{(1)}$, $z_j^{(2)}$ and z_{ij} has been largely improved. With regard to Table 3, enormous discrepancies between the REML approach and the Dannenburg's approach have been recorded as regards to the estimation of the credibility estimators. Therefore remarkable improvement on the estimation of the credibility premium occurs.

For exchangeable error structure with $\rho = 0.4$, we can see from Table 4 that the REML approach can largely improve the estimation efficiency. With the correct assumption on the error structure, REML-EX shows apparent advantage in estimating the structural parameters, especially in estimating a (relative efficiency beyond 380). As a result, better estimations of the structural parameters improve the accuracy of estimations of the credibility factors $z_i^{(1)}$, $z_j^{(2)}$ and z_{ij}. While REML-I has also largely improved the estimation efficiencies comparing to the Dannenburg's approach.

From Tables 3 and 4, we can see that REML approach can outperform the Dannenburg's approach in general even when its assumption of the error structure does not coincide with the actual error structure.

References

ANTONIO, K., and BEIRLANT, J. (2006): Actuarial Statistics with Generalized Linear Mixed Models. *Insurance: Mathematics and Economics 40, 58-76.*

COSSETTE, H., and LUONG, A. (2003): Generalized Least Squares Estimators for Covariance Parameters for Credibility Regression Models with Moving Average Errors. *Insurance: Mathematics and Economics 32, 281-93.*

DANNENBURG, D.R., KAAS, R. and GOOVAERTS, M.J. (1996): *Practical Actuarial Credibility Model.* Leuven: Institution of Actuarial Science and Econometrics.

FREES, E.W., YOUNG, V.R. and LOU, Y. (1999): A Longitudinal Data Analysis Interpretation of Credibility Models. *Insurance: Mathematics and Economics 24, 229-47.*

LAIRD, N.M. and WARE, J.H. (1982): Random-Effect Models for Longitudinal Data. *Biometrics 38, 963-74.*

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 83, 1014-22.*

LO, C.H., FUNG, W.K. and ZHU, Z.Y. (2006): Generalized Estimating Equations for Variance and Covariance Parameters in Credibility Models. *Insurance: Mathematics and Economics 39, 99-113*.

LO, C.H., FUNG, W.K. and ZHU, Z.Y. (2007): Structural Parameter Estimation Using Generalized Estimating Equations for Regression Credibility Models. *ASTIN Bulletin 37, 323-43*.

MCCULLOCH, C.E. (1997): Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association 92, 162-170*.

NELDER, J.A., and MEAD, R. (1965): A Simplex Algorithm for Function Minimization. *Computer Journal 7, 308-313*.

VERBEKE, G., and MOLENBERGHS, G. (2000): *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Part VIII

Information Retrieval for Text and Images

A Hybrid Approach for Taxonomy Learning from Text

Ahmad El Sayed[1] and Hakim Hacid[2]

[1] University of Lyon 2 - ERIC Laboratory
5, avenue Pierre Mendès-France - 69676 Bron cedex - France,
asayed@eric.univ-lyon2.fr
[2] University of New South Wales
Sydney NSW 2052, Australia,
*hakimh@cse.unsw.edu.au****

Abstract. Ontology learning from text is considered as an appealing and challeging alternative to address the shortcomings of the hand-crafted ontologies. In this paper, we present OLea, a new framework for ontology learning from text. The proposal is a hybrid approach combining the pattern-based and the distributionnal approaches. It addresses key issues in the area of ontology learning: context-dependency, low recall of the pattern-based approach, low precision of the distributionnal approach, and finally ontology evolution. Experiments performed at each stage of the learning process show the advantages and drawbacks of the proposal.

Keywords: taxonomy learning, knowledge acquisition, relevance feedback

1 Introduction

In spite of the great efforts to elaborate tools and normalized methodologies for building ontologies with help of engineers and domain experts, the task still requires an incredible amount of human labour for the intellectual encoding of "semantics". Further, hand-crafted ontologies will always suffer from a poor coverage comparing to the enormous amount of information available today in real-world repositories, like the Web.

An appealing and challenging approach is thus to build such ontologies automatically from wealthy resources like texts. This led to the emergence of the field of *ontology learning* as an important sub-field of ontology engineering Maedche and Staab (2001). In litterature, approaches for ontology learning can be generally classified as linguistic (pattern-based) or stastical approaches (distributionnal).

In this paper, we present OLEA (Ontology LEArning), a new framework for ontology learning from text[1]. The proposal is a hybrid approach that aims to deal with key issues in the area of ontology learning:

*** This work has been done when the author was a PhD student at the University of Lyon 2.

[1] The accomplished work concerns though the concepts learning and the concepts hierarchies learning, so we will rather refer to the task as *taxonomy learning*

On Low Recall of the Pattern-Based Approach. The pattern-based approach Hearst (1992), though yealding "'acceptable" precision, suffers from very low recall since detecting relations depends on the rare appearance of a set of *rigid* lexico-syntactic patterns (e.g., *NP such as* {*NP,NP..*}). Some authors face this difficulty by matching patterns on larger text resources like the Web Cimiano et al. (2005), Etzioni et al. (2004) or Wikipedia Maria Ruiz-Casado and Castellas (2005). We argue that this can resolve only partially the problem, and a more flexible method taking into account the complex and sparse nature of text is still needed. Our framework deals with this drawback, and proposes a technique able to capture and match more "flexibily" patterns in text.

On Low Precision of the Distributionnal Approach. The distributionnal approach consisting mainly of clustering terms basing on their similarities, lacks generally from low precision. This is due to two main reasons. (1) The commonly used hierarchical methods are not quiet adaptive to build taxonomies Cimiano et al. (2005), Grefenstette (1994), Caraballo (1999) since they provide binary trees of crisp clusters. (2) Methods lack of reliability since they rely, in most cases, on a single semantic relation (e.g., synonymy). That is, we present a learning procedure involving more semantic relations, and thus supplying us with more reliable decisions while building hierarchies.

On Ontology Evolution. It is known that an ontology should be subject of continuous refinements in order to adapt it to new users' and applications' requirements. However, existing approaches either ignore the evolution issue, or require regular human interventions, which is a tedious task. That is, we propose a preminalary approach that places the learned taxonomy at the core of a search engine, in order to adapt the taxonomy to users vision over text, without any manual effort.

The remainder of this paper is organized as follows: we start by presenting the general architecture of the proposed framework in Section 2. Next, we show how we build the taxonomy in Sections 3, 4, and then how we evolve the learned taxonomy in Section 5. Evaluations of the different proposed techniques are outlined at the end of each section. Finally, we conclude and draw some future works in Section 6.

2 OLea: general architecture

The general architecture of OLea, illustrated in Figure 1, is composed of three principal stages:

1. we estimate confidence rates for a set of semantic relations between cooccuring terms in a corpus. To achieve this, semantic relations are first estimated between terms that could be matched in WordNet (i.e., concepts). This will build the learning base, that will serve for mining the

relations' confidence rates between terms uncovered by WordNet (Section 3);
2. semantic relations are used as input for a concept learning algorithm that will output terms clustered into senses that will be regarded as concepts. Then, discovered 'concepts' along with the semantic relations between them will be the input of a concepts hiearchy learning algorithm (Section 4).
3. the learned taxonomy is involved in an IR environment, where users interactions with the search engine are taken into account in order to launch a relevance feedback mechanism able to adapt the taxonomy to the user vision over text (Section 5).

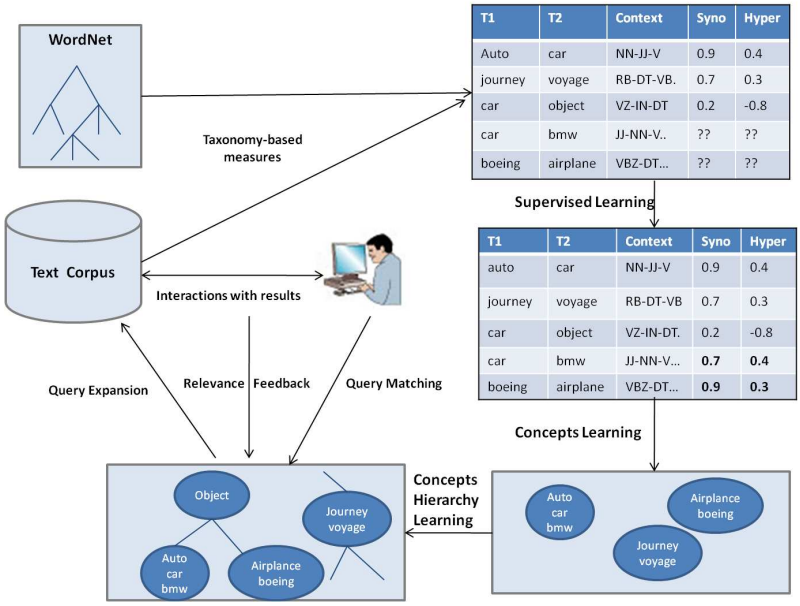

Fig. 1. OLea: General Architecture.

3 Estimating semantic relations

First, we present a technique able to to capture and match more "flexibily" semantic relations in text. The purpose is to identify larger number of semantic relations in text, thus resulting in a greater recall. The overall technique is described as follows. Each pair of terms occuring in a corpus is represented by a set of lexico-syntactic features. Pairs that could be matched in WordNet will be augmented by confidence rates for each of their semantic relation (e.g,

synonymy, hypernymy). This will construct the learning base that will serve to predict the semantic relation rates between pairs uncovered by WordNet.

3.1 Calculating relations between concepts

For pairs of concepts that could be matched in WordNet, we calculate a confidence rate for each of their semantic relations basing on the semantic structure of the taxonomy. What we are seeking at the end, is statements assessing, for instance, that *"object"* and *"car"* are *0.1-synonyms, 0.8-hypernyms, and 0-meronyms*. The calculation of such rates depends on the target relation. While hypernymy confidence relies on the edges count along the shortest path separating two concepts, confidences for antonymy and meronymy are boolean, depending simply on the presence/absence of such relations. All values will be then normalized between [-1, 1] for the asymmetric relations (e.g., hypernymy) and [0,1] for the symetric ones (e.g., antonymy, meronymy)[2].

Fortunately, synonymy relations can be expressed more significantly using semantic similarity measures. Note that the goal of ontology learning from text is not to construct an output structure that will be "an extended miror" of an existing structure like WordNet, but must reflect the context of the target corpus/domain. That is, an important point is that our semantic distance (synonymy) defined between terms must be *context-dependent*, which is not the case with the existing measures in the literature. This pushed us to develop a context-dependent semantic distance measure between concepts El Sayed et al. (2007). The proposed measure led us to a very promising correlation rate of 0.876 with human rankings on a set of words pairs.

3.2 Mining relations between terms

Semantic relations calculated so far, although 'enough' accurate, are limited to the terms covered by WordNet. An option is indeed to use text as a computing resource for terms uncovered by WordNet. However, it is known that such approach lacks reliability, since it relies on distributional frequency of words in text. The problem would get even more complicated when it comes to compute other relations, such as hypernymy or part-meronymy.

Instead of relying solely on unstructured text, we will make use of the obtained rates from the previous step (relying on a taxonomy) as a "reference" for predicting semantic relations between the uncovered pairs in WordNet. The assumption is that term pairs appearing in similar contexts tend to have similar semantic relations. Our motivation for using taxonomy-based rates as a "reference" cames essentially from the high correlation ratio obtained in our approach which we believe can hardly be exceeded by a "pure" corpus-based measure.

[2] In this paper, we assume that these relations are symmetric which is true in most cases since we are dealing with disambiguated terms , i.e., concepts

Consider an uncovered pair P represented by a set of lexico-syntactic features F (context), and associated to a set of semantic relations R, whose confidence rates are *null*. We wish to predict the values of R. To do this, we retrieve for P its K Nearest Neighbors (K-NN) from the set of covered pairs, thus where confidence rates are known. Subsequently, each semantic relation in R is predicted by means of a weighted average of the relations among the K closest pairs[3].

Similarity between two pairs, P_1 and P_2, is computed on the basis of their respective lexico-syntactic features, F_1 and F_2. Lexico-syntactic features F that characterize a pair P are issued from another area of research (i.e., Semantic Role Labeling Gildea and Jurafsky (2002), Pradhan et al. (2004)) and are defined as follows: *Terms POS, Head Word lemma and POS, partial path, chunk path, lexical path, path length.*

In order to compare two contexts, a distance is calculated as a combination of many sub-distances between each of these features. Sub-distances all range in the [0,1] interval. Some sub-distances are based on a simple integer or a string comparison. However, since paths are highly sparse (e.g., 68.4% of distinct partial paths in our experiments), we turned to the Waterman alignment algorithm Smith and Waterman (1981), basically created for comparing sequences of proteins, DNA, RNA in bioinformatics. Consequently, our final distance between two contexts is defined as the combination of sub-distances between the different features:

$$Dist(F_1, F_2) = \beta_1.POSDist + \beta_2.HWDist + \beta_3.HWPOSDist + \quad (1)$$

$$\beta_4.PartPathDist + \beta_5.FullDist + \quad (2)$$

$$\beta_6.LexiPathDist + \beta_7.PathLenDist + \epsilon \quad (3)$$

where $\beta_1 ... \beta_7$ denotes the coefficients assigned for each sub-distance during the distance calculation, and ϵ denotes the intercept or the model error.

Consider a relation r for a pair P. Finding its K best confidence rates depends on how much we can "optimize" the distances between pairs of features. These distances can be optimized when reaching a maximal correlation with distances between pairs of semantic relations (response variables). That is, we applied a multiple linear regression model to find the parameters that optimize the correlation between $Dist(r_1, r_2)$ and $Dist(F_1, F_2)$. Then, we apply the discovered optimal parameters on the previous equation in order to find more accurately the K-NN.

Following that, for each P uncovered by WordNet, a set of K relation confidences along with their respective distances with P are retrieved from the learning base. The final relation confidence is then calculated by means of the weighted average of the K-nearest relation confidences:

[3] Each semantic relation is treated separately.

$$r(P_j) = \frac{1}{K} \sum_{i=1..K} (\frac{1}{dist(F_j, F_i)} * r(P_i)) \quad (4)$$

where $r(P_j)$ denotes the relation confidence of relation r in pair P_j; $dist(F_j, F_i)$ denotes the distance between the features sets F_j and F_i of the two pairs P_j and P_i respectively.

Finally, same pairs appearing in several contexts will have their semantic relations averaged. Using the proposed hybrid method, semantic relations can be detected between any couple of co occurring terms in text, and do not depend on matching *exactly* a set of predefined patterns.

3.3 Evaluation and results

Our experiments have been carried out on a benchmark composed of 1000 documents picked from the Reuters corpus[4] along with the WordNet taxonomy. The key question of this evaluation is to figure out to which extent the predicted relation confidences (by means of supervised learning) can approach the taxonomy-based relation confidences (by means of WordNet). For this, we divided the set of concept pairs into 80% for the training set, and 20% for the test set.

From Figure 2, we can notice that K has no significant effect on the results, and that models performance depends more on the obtained Pearson's coeficient of the regression model R^2. Without using this model, thus by setting an equal coefficient of 1 to all variables, we obtained a best correlation ratio of 0.32 for synonymy. However, when incorporating the regression model with KNN, we could dramatically increase correlation, attaining a rate of 0.82 for synonymy. Figure 3 shows clearly that the final performance depends on how successful the regression model is.

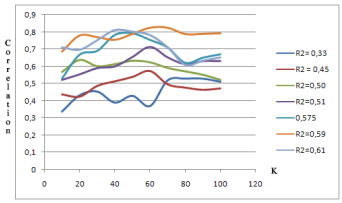

Fig. 2. Effect of K variation on the final correlation rate for each regression model.

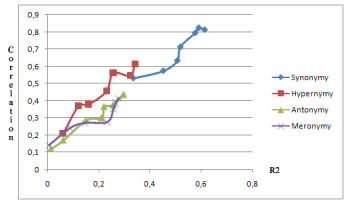

Fig. 3. Effect of R^2 on the final correlation rate for each semantic relation.

[4] Reuters corpus, volume 1, english language, release date: 2000-11-03

4 Taxonomy learning

In this section, we present a two-phases procedure that takes as an input the semantic relations rates, and provides as an output a hierarchy of concepts. It includes concepts learning, and concepts hiearchy learning.

4.1 Concepts learning

The goal here is to group terms into a set of sense-bearing clusters, where each cluster will be perceived as a new concept. Hence, we define a soft hierarchical-based clustering algorithm able to deal with polysemous words.

Algorithm 1 Concept Learning Process.

input: Initialize each term t_i in the set T as a cluster c_i in the set C
repeat
  Identify the closest pair P of clusters c_p and c_k in C having synonymy exceeding a threshold θ_1 and having no other relation exceeding a threshold θ_2
  Create a new cluster c_n containing the instances of c_p and c_k
  if c_p is not basically a term **then**
    Remove c_p from C
  end if
  if c_k is not basically a term **then**
    Remove c_k from C
  end if
  for all other clusters oc in C **do**
    if $oc_i \subset c_n$ OR $c_n \subset oc_i$ **then**
      continue
    end if
    Compute relations R_i between c_n and oc_i
    if $R_i(synonymy)$ is above θ_1 and all other R_i are below θ_2 **then**
      Merge oc_i instances in c_n
      if oc_i is not basically a term **then**
        Remove oc_i from C
      end if
    end if
  end for
  Mark c_p and c_k as a considered pair (to not be considered again)
until P is empty
return the set of created clusters along with the set of terms that were not added to any cluster

Rather than clustering terms by relying solely on semantic similarity which is error-prone, our algorithm offers more reliable decisions by taking into account a larger set of relations. The point is that two related clusters will be merged only if they are found "purely" synonyms, therefore do not have any other relation with a confidence rate greater than a specified threshold.

4.2 Concepts hierarchy learning

Following concepts learning, the goal is to learn taxonomic *is-a* relations in order to build a hierarchy of concepts. The Algorithm 2 used for this purpose is somehow similar to the previous one in the sense that hypernyms relations

are created recursively by considering at each iteration the "best" concepts pair with respect to predefined thresholds.

Algorithm 2 Taxonomic-Relations Learning Process.

input: Let P be the set of concepts pairs with their relations confidence obtained from the previous phase
input: Define direct-hypernymy confidence $dirhyp(cp_i)$ for a concept pair cp_i in P as synonymy(cp_i)*hypernymy(cp_i)
1: **repeat**
2: Identify the concept pair cp_k with the highest $absolute(dirhyp(cp_k))$ that must be above a threshold α_1 and having the other relations confidence (except synonymy and hypernymy) below a threshold α_2
3: Create a hypernymy link for cp_k
4: **until** cp_k is empty
5: **for** each remaining concept c_r sharing no link with any other concept **do**
6: find the K closest concepts for c_r by means of synonymy
7: **for** each close concept c_i **do**
8: calculate a score s_i as a function of $synonymy(c_r, c_i)$ and $synonymy(c_r, hyponyms(c_i))$
9: **end for**
10: create a hypernym link between c_r and the c_i with the highest score $MAX(c_i)$
11: **end for**

The formula $dirhyp(c_i, c_j) = synonymy(c_i, c_j).hypernymy(c_i, c_j)$ is designed under the simple assumption that a concept c_i is judged as a "good" direct hypernym for another concept c_j if c_i and c_j share a high hypernymy and a high synonymy confidence. In addition to that, as shown in the algorithm, we ensure that *'is-a'* links are created between concept pairs that are "purely" hypernyms, thus sharing any other relation (except synonymy) with a confidence rate greater than θ_2.

As hypernyms occur rarely between pairs of terms, lot of concepts will remain unlinked. To overcome this shortcoming, we start by identifying for each unliked concept c_r its K closest concepts C_k by means of synonymy. Next, in order to identify which of the C_k is the most appropriate hypernym for c_r, a direct-hypernymy score is calculated for each c_i in C_k basing on the synonymy relations that c_r shares with the children of c_i :

$$dirhyp(c_i, c_r) = \frac{AvgSyno(c_r, hypo(c_i))}{(1 + VarSyno(c_r, hypo(c_i)))} \tag{5}$$

$AvgSyno(c_r, hypo(c_i))$ denotes the average of synonyms confidences that c_r has with the children of c_i, while $VarSyno(c_r, hypo(c_i))$ denotes the variance of synonyms confidences that c_r has with the children of c_i.

The underlying assumption of this measure is that the ideal hypernym c_i for a concept c_r is found when: 1) c_i has the highest semantic similarity with c_r, 2) no children of c_i has a highest semantic similarity with c_r than with c_i itself, 3) c_r has exactly the same semantic similarity (or a minimum variance) with all the children of c_i. These assumptions are the result of observation of the optimal behavior for a semantic similarity measure in a taxonomy.

After learning taxonomic relations between concepts, multiple disparate hierarchies are created that will be merged by means of an expert[5]. At the end of this phase, we obtain a fuzzy taxonomy in the sense that related terms within a concept are assigned a synonymy confidence between each others, and that concepts are related to each others by an *'is-a'* relation being assigned a hypernymy confidence as well.

4.3 Evaluation and results

In literature, one of the approaches for evaluating a learned ontology is to compare it with another reference ontology. Actually, ontology learning community lacks of common evaluation frameworks which leads to a lack of comparative results showing the effectiveness and efficiency for each technique. Concerning our work, we performed a preminelary evaluation against human judgments. Typically, after specifying the actual context of *newspapers*, we asked a human subject to group and organize in one or many trees a set of 50 terms. Finally, we compare the human-made tree with our learned tree in terms of precision of recall by means of the number of correct *vs* incorrect learned relations. Furthermore, all instances in a concept are considered inter-related and will have their relations evaluated as well.

Concepts Learning Evaluation. Since precision and recall depends on the predefined thresholds θ_1 and θ_2 (Algorithm 1), we altered θ_1 in the range of [0.88, 0.97], while fixing $\theta_2$to 0.05. As we can see in Figure 4, while precision tends to drop dramatically when reducing θ_1, recall tends to be somehow stable along the different parameter values.

Taxonomic-Relations Learning Evaluation. We consider here the number of correct *vs* incorrect links between validated concepts by the user, which discarded 6 out of 37 learned concepts. We fixed the parameter θ_1 at 0,95, since it gave the optimal trade-off between precision and recall. Then, we applied Algorithm 2 by alternating α_1 in the interval [0.1, 0.2] and fixing α_2 at 0.05. Results illustrated in Figure 5 show interesting performance, especially from a recall point of view.

Let's quote here that a comparison with other methods is still needed to assess the added-value of the proposed method. This is not an easy task due to the lack of common frameworks for evaluation. We argue that an application-oriented evaluation is the most meanigfull way to compare different methods.

5 Improving taxonomy with relevance feedback

Involving human subjects in the learning process, although extremly benefic, can be a very tedious and time-consuming task Faure and Poibeau (2000),

[5] This a very complex task to perform automatically. This issue will be addressed in future works.

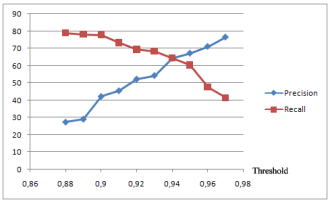

Fig. 4. Concept learning performance in terms of precision and recall with different parameter values.

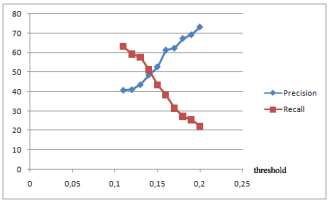

Fig. 5. Taxonomic-relations learning performance in terms of precision and recall with different parameter values.

Moldovan and Girju (2001). What we propose here is to add supervision to the learning process without any manual effort: Since our taxonomy seeks essentially to integrate a IR environment, we involve human visions implicitly in the learned taxonomy by considering their interactions with a search engine using the relevance feedback mechanism Ruthven and Lalmas (2003).

Therefore, we placed our learned taxonomy at the core of our IR system El Sayed et al. (2007b). Keywords queries, after being mapped in the taxonomy, will be expanded to other related terms by means of the synonymy and hypernymy relations. At the end, relevant documents will be returned to the user by their titles and two-lines outlines. Four possible situations are present here:

- A document viewed by the user and containing an expansion term $\Rightarrow$ confidence $\nearrow$
- A document viewed by the user and not containing an expansion term $\Rightarrow$ confidence$\searrow$
- A document not viewed by the user and containing an expansion term $\Rightarrow$ confidence $\searrow$
- A document not viewed by the user and not containing an expansion term $\Rightarrow$ confidence$\nearrow$

Subsequently, an implicit relevance feedback mechanism Ruthven and Lalmas (2003) is involved in order to strengthen/weaken the relations confidences between the query terms and the expansion terms. Since our query expansion depends on the taxonomy's fuzzy relations, feedbacks will enable the system to take more subjective decisions about accepting or rejecting a specific expansion term in future queries. Note that only synonymy and hypernymy relations are covered by the mechanism, since other relations do not contribute to any query expansion, playing the simple role of "prevention" during taxonomy learning. Moreover, the mechanism will not cover concepts detected in Wordnet because we consider that they yield the optimal accuracies that one could reach, and that updates would tend to deteriorate them.

5.1 Evaluation and results

To evaluate the effect of relevance feedback on taxonomy learning, we observe the concepts and relations learning accuracy that is likely to increase as new queries are sent to the IR system. Thus, we took as a starting point the results given by the learned taxonomy obtained using the optimal parameters (Section 4.3). Next, 100 keywords queries (related to the selected hierarchy for evaluation) are sent consecutively to the system. At the end of session of each query, viewed and unviewed documents by the user are considered for the feedback. Taxonomy is updated at the end of each 20 queries in order to be reevaluated against the hand-built taxonomy of 50 terms (Section 4.3). Figure 6 shows the precision and recall values for both concepts and relations learning along the 100 sent queries. We can notice the slight but sure improvement in the final results (especially in precision). Yet, we argue that the improvement can be seen more clearly with larger set of queries.

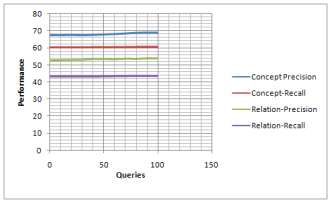

Fig. 6. Performance evolution along queries using Relevance Feedback.

6 Conclusion and future works

To wrap up, we presented in this paper OLea, a framework for learning ontology from a text corpus. It has the advantage to be able to deal with the sparse nature of text, offering more flexible recognition for semantic relations between terms. In addition to that, we presented two algorithms that make use of the detected relations in order to output a taxonomy while dealing with polysemic terms. At the end, we showed how results can be improved by placing our learned taxonomy in the core of an Information Retrieval environment. As for future works, we are seeking possible solutions for merging different disparate hierarchies into one final hierarchy. Then, we are planning to extend the current approach in order to learn non-taxonomic semantic relations from text. Since no "correct" ontology exists for any domain, we argue that a learned ontology is better evaluated by assessing its positive/negative contribution for the environment and the task(s) that it was intended for (e.g. Agirre et al. (2000)) for Word Sense Disambiguation.). Thus, we are intending to perform a task-oriented evaluation in environments like Information Retrieval and Text Classification.

References

AGIRRE, E., ANSA, O., HOVY, E., and MARTINEZ, D. (2000): Enriching very large ontologies using the www. In: *Proceedings. of the Ontology Learning Workshop, ECAI,* Berlin, Germany.

CARABALLO, S.A. (1999), Automatic construction of a hypernym-labeled noun hierarchy from text, In: *Proc. of the Conference of the Association for Computational Linguistics*, 120–126.

CIMIANO, P., HOTHO, A., and STAAB, S. (2005): Learning concept hierarchies from text corpora using formal concept analysis, In: *J. Artif. Intell. Res. (JAIR)*, 24, 305–339.

EL SAYED, A., HACID, H., and ZIGHED, D. (2007): A multisource context-dependent approach for semantic distance between concepts. In: Roland Wagner and Norman Revell and Günther Pernul (Eds.): *Database and Expert Systems Applications, 18th International Conference, DEXA 2007*. Springer, Berlin, 54–63.

EL SAYED, A., HACID, H., and ZIGHED, D. (2007b): Combining text and image for content-based information retrieval. In: *Proceedings of the International Conference on Information and Knowledge Engineering*, 47–53.

ETZIONI, O., CAFARELLA, M., DOWNEY, D., KOK, S., POPESCU, A.-M., SHAKED, T., SODERLAND, S., WELD, D.S., and YATES, A. (2004): Web-scale information extraction in knowitall: preliminary results. In: *WWW '04: Proceedings of the 13th international conference on World Wide Web*. ACM Press, 100–110.

FAURE, D. and POIBEAU, T. (2000): First experiments of using semantic knowledge learned by asium for information extraction task using intex. In: Werner Horn (Ed.): *ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence*, Berlin, Germany, August 20-25, 2000.

GILDEA, D. and JURAFSKY, D. (2002): Automatic labeling of semantic roles. In: *Computational Linguistic*, 28(3), 245–288.

GREFENSTETTE, G. (1994): *Explorations in automatic thesaurus construction*, Kluwer.

HEARST, M.A. (1992): *Automatic acquisition of hyponyms from large text corpora*, no. S2K-92-09, 8.

MAEDCHE, A. and STAAB, S. (2001): Ontology learning for the semantic web, *IEEE Intelligent Systems*, 16(2), 72–79.

MARIA RUIZ-CASADO, E.A. and CASTELLS, P. (2005): Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia, In: *Natural Language Processing and Information Systems*. Springer Berlin, 67–79.

MOLDOVAN, D.I. and GIRJU, R. (2001): An interactive tool for the rapid development of knowledge bases. In: *International Journal on Artificial Intelligence Tools*, 10(1-2), 65–86.

PRADHAN, S., HACIOGLU, K., KRUGLER, V., WARD, W., MARTIN, J. H., and JURAFSKY, D. (2005): Support vector learning for semantic argument classication. In: *Machine Learning*, 60(1-3), 11–39.

RUTHVEN, I. and LALMAS, M. (2003): A survey on the use of relevance feedback for information access systems. In: *Knowl. Eng. Rev.*, 18(2), 95–145.

SMITH, T. and WATERMAN, M. (1981): Identication of common molecular subsequences. In: *Journal of Molecular Biology*. 147, 195–197.

Image and Image-Set Modeling Using a Mixture Model

Charbel Julien[1] and Lorenza Saitta[2]

[1] Laboratoire ERIC, Université Lumière Lyon2
5 avenue Pierre Mendès-France, 69676 Bron, France, *charbeljulien@hotmail.com*
[2] Dipartimento di Informatica, Università del Piemonte Orientale
Vie Bellini 25/G, 15100 Alessandria, Italy, *saitta@mfn.unipmn.it*

Abstract. Modeling an image or an image-set, which share similar visual contents, by means of a discrete distribution (such as a signature) or by means of a mixture model (such as a Gaussian mixture-model) has a major utility, and may serve as a basis for Content Based Image Retrieval and other related areas. Mixture model can encode information about color, texture, and spatial relationships between colored/textured regions. Image modeling is used in several tasks, such as Image retrieval, Automatic annotation, Unsupervised or Semi-supervised Clustering. Linear optimization techniques offer a reliable and efficient way to compute distance, in both cases, discrete distributions and mixture models. Linear optimization can be also used for modeling image-sets, by computing a mixture model that minimizes distances.

Keywords: image modeling, image-set modeling, discrete distribution, Gaussian mixture model, linear optimization

1 Introduction

Early approaches in Content Based Image Retrieval (CBIR) were based on extracting low-level visual content from images by using histograms of color, texture, and other low-level features. Another common approach in the literature is based on fixed-size feature vector. A feature vector codes information about color (for instance, color moments and histogram of color), texture and shape. Evaluating the similarity between feature vectors is worked out by using a weighted linear combination of subsets of features. The distance between subsets of feature is computed by using a simple distance such as the Euclidean distance.
More recent techniques use discrete distributions like *signatures*, and mixture model like *Gaussian Mixture Models* (GMMs) to model the low-level visual content. Mixture model are well suited to abstract the content of images, and they can be adapted to the image content complexity. Simple images have a few mixture components while complex images have many components. Mixture distributions can serve also for modeling image-sets in a compact form for supervised, unsupervised, and semi-supervised learning.

Models are computed directly from the images by unsupervised learning, using, for example, the k-Means algorithm in the case of discrete distributions, or the Expectation Maximization (EM) algorithm with GMM in the case of mixture model. Unlike fixed-size feature vector, where the centroid that minimizes the distance to a set of vectors can be computed by averaging the values in the feature vectors, mixture model's centroid needs a more complex technique to be computed.

2 Related work

Rubner et al. (2000) use signatures as discrete distributions to model image content, instead of histograms, leading to better results in CBIR. They use a k-d tree to perform clustering using color content to extract the signature of an image. The Earth Mover's Distance (EMD), a transportation algorithm, has been proposed to calculate distance between signatures.
Li and Wang (2006), seeking automatic annotation through supervised learning, use signatures of color and texture to model the image contents. *Mallows distance* was used as a metric to evaluate distance between signatures. A linear optimization algorithm, called *D2-clustering*, was proposed to compute a set of centroids for every image-set category. Every categorical class is modeled by a set of signatures for every visual content (color/texture) using *D2-clustering*, an algorithm in the same spirit as the K-Means algorithm.
Datta et al. (2006) build two models to capture different visual aspects. A structure-composition model, which uses Beta distributions to capture color interactions, and a Gaussian mixture model in the joint color-texture feature space. These two models are used in a supervised learning in order to categorize unseen images. The image is examined from two separate viewpoints in order to place it in a category.
Goldberger et al. (2006) propose an information theoretic framework for unsupervised image-set clustering. They model the color content of images by a mixture of Gaussian distributions by applying the EM algorithm. Monte-Carlo simulation is used to approximate the KL-divergence distance between GMM distributions. An information framework called *Information Bottleneck* (IB) was used afterwards for agglomerative unsupervised clustering of the images. The image-set is modeled by averaging all the models within the set, so that all the information present in the image-set is conserved; the main disadvantage is the growing if the complexity with the growing of the number of images in the set.
Image-set modeling requires compressing mixture models of single images into a compact form. Zhang and Kwok (2006) reduce the model size by first grouping the components into clusters, and then perform local function approximation that minimizes an upper bound of the approximation error. The L_2 norm is used as the error criterion. Goldberger at al. (2007) propose to reduce a large Gaussian mixture to smaller one by minimizing a KL-based

distance between the two mixtures through a clustering based on unscented transform.

3 Image modeling

In this work we use the color of an image as low-level content. Color has been proved to be a more powerful feature, in the case of natural color images, than texture or shape features. Obviously, using discrete distribution or Gaussian mixture model we can model other low-level image content like texture and edges. Discrete distribution and Gaussian mixture model are considered as alternative rather then the simple statistical representation by histograms.
A pre-processing of the image is need to extract color content. We first smooth each band of the image's RGB representation using a 2D-Gaussian filter, with the aim of reducing the possible color quantization and dithering artifacts. We then transform the image representation into the LUV perceptual color space. Images are partitioned into blocks of 4 x 4 pixels. The block size is chosen as a compromise between the color details and the computation time, owing to the large number of images to be processed in an image database. Three features are extracted from each block, concerning the color, namely the average value of the *luminance* L, and of the *chrominance*, captured by the parameters U and V, which encode color information. It has been proved that the LUV color space has a good correlation with human perception. Each image is represented by a set of feature vectors extracted from each block. These feature vectors are then used to model the content of images by discrete distributions/GMMs. The feature vector can include either three features per block (L,U and V) or five features (x, y, L, U and V) where x and y are the coordinate of block in the image.

3.1 Discrete distribution

Modeling images by a discrete distribution is realized by computing its signature. In this work we use the color content to compute a signature. We use a method similar to the one used by Li and Wang (2006). The K-Means algorithm is used to cluster the feature vectors of color into several classes. The number of clusters in the algorithm is determined dynamically by thresholding the average within-cluster variation.

Let us suppose that the observations (color vectors) are $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$. The role of the K-Means algorithm is to partition the observations into k groups with means $\hat{x}_1, \hat{x}_2, ..., \hat{x}_k$ such that the following objective function $D(k)$ (the distortion) is minimized:

$$D(k) = \sum_{i=1}^{M} min_{1 \leq j \leq K} (x_i - \hat{x}_j)^2$$

K-Means does not specify how many clusters to choose. We adaptively choose the number of clusters by gradually increasing k and stopping when some halt condition is met. We start with $k = 2$ and stop increasing k when one of the following conditions is satisfied:

1. The distortion $D(k)$ is below a given threshold. A low $D(k)$ indicates high purity in the clusters.
2. The discrete approximation of the first derivative of the distortion with respect to k, *i.e.*, $[D(k) - D(k-1)]$, is below a threshold. A low $[D(k) - D(k-1)]$ indicates convergence in the clustering process.
3. The number k exceeds an upper bound. We force an image signature to only consists of no more than 8 elements. Usually, the segmentation process generates a much lower number of classes.

Once the clusters of color have been found, the color signature $\mathcal{S}_l = \{(S_i^l, w_i^l) |\ i = 1, ..., k_l\}$ of image $\mathcal{I}_l$ are computed; this amounts to calculate the fraction of pixels that belong to each cluster. This fraction is computed by finding, for each pixel in the image, its nearest-neighbor set.
The EMD distance is used to compute the distance between signatures. Euclidean distance is used internally by the EMD distance. Mallows distance (Levina and Bickel (2001)) can be used instead of EMD distance, and this leads to the same results, because we normalized the signature weights in such a way that have $\sum_{i=1}^{k_l} w_i^l = 1$.

3.2 Gaussian mixture model

In order to include spatial information, the (x, y) position is appended to the color feature vectors extracted from blocks. The image is represented by a collection of feature vectors with five dimensions. Afterward, images $\mathcal{I}_l$ are modeled using a mixture of Gaussian. The expectation-maximization (EM) algorithm is used to determine the maximum likelihood parameters of the model. The minimum description length (MDL) principle is useful to select among values of k. k is the number of Gaussian components in the mixture model that ranges, in this work, from 2 to 8.
Blocks are grouped into homogeneous regions which are represented by a Gaussian mixture model. The distribution of a 5-dimensional random variable y, representing a feature vector, is a mixture of k Gaussians with density function:

$$f_l(y) = \sum_{i=1}^{k_l} \alpha_i \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} exp\{-\frac{1}{2}(y - \mu_i)^T \Sigma_i^{-1} (y - \mu_i)\}$$

The EMD is the minimum cost of changing one mixture into another one when the cost is that of moving a probability mass from one component in the first mixture to another component in the second mixture. A common choice of the cost is the symmetrized KL-distance (Goldberger et al. (2006));

with this cost the EMD does not obey the triangle inequality. Instead we use normalized L_2 distance. Let $f_l^{'}(y) = f_l/\sqrt{\int f_l(y)^2 dy}$. The normalized L_2 distance $d_{nL_2}(\mathcal{I}_a, \mathcal{I}_b) = \int_{R^d} (f_a^{'}(y) - f_b^{'}(y))^2 dy$, and the normalized L_2 distance is a continuous version of cosine distance (Jensen et al. (2007)):

$$d_{nL_2}(\mathcal{I}_a, \mathcal{I}_b) = 2(1 - \int_{R^d} f_a^{'}(y) f_b^{'}(y) dy)$$

For Gaussian Model, closed form expression for the normalized L_2 distance can be computed (Ahrendt (2005)):

$$d_{nL_2}(\mathcal{I}_a, \mathcal{I}_b) = 2(1 - \int_{R^d} f_a^{'}(y) f_b^{'}(y) dy)$$

$$= 2(1 - \int_{R^d} \frac{N_y(\mu_a, \Sigma_a)}{\sqrt{\int_{R^d} N_y^2(\mu_a, \Sigma_a) dy}} * \frac{N_y(\mu_b, \Sigma_b)}{\sqrt{\int_{R^d} N_y^2(\mu_b, \Sigma_b) dy}} dy)$$

$$= 2(1 - \frac{\int_{R^d} N_y(\mu_a, \Sigma_a) * N_y(\mu_b, \Sigma_b) dy}{\sqrt{\int_{R^d} N_y^2(\mu_a, \Sigma_a) dy} * \sqrt{\int_{R^d} N_y^2(\mu_b, \Sigma_b) dy}})$$

$$= 2(1 - \frac{|2\pi(\Sigma_a + \Sigma_b)|^{-\frac{1}{2}} exp(-\frac{1}{2}(\mu_a - \mu_b)^T(\Sigma_a + \Sigma_b)^{-1}(\mu_a - \mu_b))}{(\sqrt{\int_{R^d} N_y^2(\mu_a, \Sigma_a) dy} * \sqrt{\int_{R^d} N_y^2(\mu_b, \Sigma_b) dy})})$$

4 Image-set modeling

Image modeling can be extended to image-set modeling using mixture models. By image-set, we mean a collection of images that exhibit visual similarity in color content and/or in spatial relationships between colored regions. Image-sets are generated either by supervised categorization, unsupervised or by semi-supervised clustering of image collection into groups. Let $\mathcal{I}_l$ ($l \in \mathcal{C}_n = \{1, ..., N\}$) denote the images within the image-set. Modeling an image-set can be done by computing a mixture model that minimizes the distance to all mixture models of images within the image-set, as can be modeled by a mixture of mixture models; in this case the image-set is partitioned into homogenuous subsets, and for every subset a prototype is computed.

The prototype is a mixture model, as well as the image-set and every component of the image-set distribution. In this paper we are concerned with the case of computing a mixture model as centroid of other mixture models. We first choose the mixture model that minimizes the distance to all other models. Afterwards, we use linear optimization algorithm, similar to *D2-clustering*, to compute a mixture model that minimizes the distance to all models in the image-set.

Let $\alpha = \{(z_1, q_1), (z_2, q_2), ..., (z_m, q_m)\}$ be the mixture model to be computed as a centroid. Let $\{(v_1^l, p_1^l), (v_2^l, p_2^l), ..., (v_{m_l}^l, p_{m_l}^l)\}$ be the mixture model of

$\mathcal{I}_l$, where z_k $(k = 1, ..., m)$ and v_j^l $(j = 1, ..., m_l)$ are the parameters of the mixture model, namely a vector of features, in the case of discrete distributions (signatures), and the mean vector μ and covariance matrix Σ of a Gaussian distribution in the case of GMMs. Let q_k and p_j^l be the probabilities associated to each component with:

$$\sum_{k=1}^{m} q_k = \sum_{j=1}^{m_l} p_j^l = 1$$

In the case of discrete distribution, the algorithm proceeds as follows:

1. Fix z_k, $k = 1, .., m$. Update q_k and $w_{k,j}^l$ by using linear optimization technique:

   Minimize: $\sum_{l \in C_n} \sum_{k=1}^{m} \sum_{j=1}^{m_l} w_{k,j}^l d(z_k, v_j^l)$

   subject to:

   $\sum_{k=1}^{m} q_k = 1$; $q_k \geq 0$, $k = 1, ..., m$
   $\sum_{j=1}^{m_l} w_{k,j}^l = q_k$; $k = 1, ..., m$, for every $l \in C_n$
   $\sum_{k=1}^{m} w_{k,j}^l = p_j^l$; $j = 1, ..., m_l$, for every $l \in C_n$
   $w_{k,j}^l \geq 0$, $l \in C_n$ $j = 1, ..., m_l$ and $k = 1, ..., m$
2. Fix q_k, $w_{k,j}^l$ update z_k, $\forall$ $k \in \{1, ..., m\}$

   $z_k = \frac{\sum_{l \in C_n} \sum_{j=1}^{m_l} w_{k,j}^l * v_j^l}{\sum_{l \in C_n} \sum_{j=1}^{m_l} w_{k,j}^l}$
3. Compute $\sum_{k=1}^{m} \sum_{j=1}^{m_l} w_{k,j}^l d(z_k, v_j^l)$

   if the rate of decrease from the previous iteration is below a threshold, return (optimization ended); otherwise go to Step 1.

($d(z_k, v_j)$ is the distance between components of discrete distributions, i.e. distance between two vectors)

In the case of GMMs the algorithm operates as follows:

1. Fix z_k, $k = 1, .., m$. Update q_k and $w_{k,j}$ by using linear optimization technique:

   Minimize: $\sum_{l \in C_n} \sum_{k=1}^{m} \sum_{j=1}^{m_l} w_{k,j} d(z_k, v_j^l)$

   subject to:

   $\sum_{k=1}^{m} q_k = 1$; $q_k \geq 0$, $k = 1, ..., m$
   $\sum_{j=1}^{m_l} w_{k,j} = q_k$; $k = 1, ..., m$, for every $l \in \mathcal{C}_n$

$\sum_{k=1}^{m} w_{k,j} = p_j^l$; $l \in \mathcal{C}_n$, $j = 1, ..., m_l$
$w_{k,j} \geq 0$, $l \in \mathcal{C}_n$ $j = 1, ..., m_l$ and $k = 1, ..., m$

2. Fix q_k, $w_{k,j}$ update z_k, $\forall\ k \in \{1, ..., m\}$

$$\mu_k = \frac{\sum_{l \in \mathcal{C}_n} \sum_{j=1}^{m_l} w_{k,j} * \mu_j^l}{\sum_{l \in C_n} \sum_{j=1}^{m_l} w_{k,j}}$$

$$\Sigma_k = \frac{\sum_{l \in \mathcal{C}_n} \sum_{j=1}^{m_l} w_{k,j} * (\Sigma_j^l + (\mu_j^l - \mu_k)(\mu_j^l - \mu_k)^T)}{\sum_{l \in \mathcal{C}_n} \sum_{j=1}^{m_l} w_{k,j}}$$

3. Compute $\sum_{k=1}^{m} \sum_{j=1}^{m_l} w_{k,j} d(z_k, v_j^l)$
if the rate of decrease from the previous iteration is below a threshold, return (optimization ended); otherwise go to step 1.

5 Experiments

In order to test image modeling, we used Wang's dataset, which contains one thousand general purpose images, manually selected from the Corel database. The dataset contain 10 classes of 100 images each. Images are 384×256 or 256×384 pixels, compressed in JPEG format. Afterwards, these mixture distributions are used separately in unsupervised clustering. First, a matrix of distances between images is computed. We use simple k-Means algorithm for unsupervised clustering, using the distance matrix as input. Multi-dimensional scaling is used afterward within clusters, in order to reflect the similarity between images. Images that have smaller distance among themselves are displayed near each other.

5.1 Quality evaluation

The capacity of modeling can be evaluated by the quality of clustering. The quality of clustering is dependent in part on the quality of the model representing items and on the distance measure. Items are known to the clustering algorithm only by their distances. The quality of clustering results is difficult to measure. In particular, one needs to find a quality measure that is not dependent on the technique used in the cluster generation process, on the representation scheme, and on the distance measure. Using a ground truth clustering database gives an independent evaluation of clustering quality. We use *Normalized Mutual Information* (NMI) (Strehl at el. (2000)) between true and predicted labels to measure the quality of clustering. The NMI measures the amount of information that the knowledge of one variable value provides about another one. The NMI ranges between 0 and 1:

$$NMI = 2\ \frac{H(L) - H(L|\hat{L})}{H(L) + H(\hat{L})} \tag{1}$$

A high value of NMI indicates a strong content resemblance inside clusters. In (1) L and $\hat{L}$ are random variables corresponding to the ground-truth labels

and to the labels assigned by the clustering algorithm, respectively. $H(L)$ and $H(\hat{L})$ are the marginal entropies of L and $\hat{L}$, whereas $H(L|\hat{L})$ is the conditional entropy. Using the NMI measure, we get 0.46 as result of clustering using color signatur, and 0.41 using the GMM of color plus coordinates (x,y). The results of quality clustering of discrete and continuous distributions can also be measured using the "F-measure":

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

The F-measures concerns every class (Africa, Beach, Bus, etc.) in contrast to NMI, which provides an overall evaluation. Then, for every cluster, we compute the "F-measure" (Yang and Liu (1999)) regarding a class (Africa, Beach, Bus, etc.), and afterward we take the maximum value found. This value is considered as a quality clustering regarding a specific categorical class. The results of the experiments are reported in Table 1.

class	**Discrete Distribution**	**Color+XY GMM**
Africa	0.37	0.57
Beach	0.32	0.35
Historical Building	0.32	0.29
Bus	0.36	0.48
Dinosaurs	0.84	0.73
Elephants	0.54	0.35
Flowers	0.55	0.78
Horses	0.77	0.72
Mountains	0.49	0.33
Foods	0.46	0.66

Table 1. Clustering evaluation for discrete and GMMs.

6 Conclusion

In this paper we presented image and image-set modeling by mixture of distributions. Mixture models serve as an efficient way to summarize visual contents of images. We proposed linear optimization algorithms as a reliable way to calculate distance between mixture distributions, as well as for computing a centroid of mixture models sets in both cases discrete distributions and GMMs.
Clustering enables us to evaluate the modeling capacity. Using a fully labeled, ground truth image database we have evaluated the clustering quality.
The results show that color discrete distributions give the best overall clustering results in term of *NMI*. Gaussian mixture models of color plus coordinates

give good results in modeling class categories as *flowers* and *foods* that exhibit colored regions correlated to spatial locations. In spite of the large number of works in image representation, we are still far from handling high semantic level. Image features still are closely related to low-level such as color and texture.

References

AHRENDT, P. (2005): *The Multivariate Gaussian Probability Distribution*. IMM, Technical University of Denmark.

DATTA, R., Ge, W., LI, J. and WANG, J. (2006): Toward Bridging the Annotation-Retrieval Gap in Image Search. In: *MM' 06*. ACM, Sanata Barbara.

GOLDBERGER, J., GORDON, S. and GREENSPAN, H. (2006): Unsupervised Image-Set Clustering Using an Information Theoretic Framework. In: *Transactions on Image Processing*. IEEE.

GOLDBERGER, J., GREENSPAN, H. and DREYFUSS J. (2007): An Optimal Reduced Representation of MoG With Applications to Medical Image Database Classification. In: *Computer Vision and Pattern Recognition*. CVPR.

JENSEN, J., ELLIS, D., CHRISTENSEN, M., and JENSEN, S. (2007): Evaluation of Distance Measures between Gaussian Mixture Models of MFCCS. In: *Austrian Computer Society*. OCG.

LI, J. and WANG, J.Z. (2006): Real-Time Computerized Annotation of Pictures. In: *MM' 06*. ACM, Sanata Barbara.

LEVINA, E. and BICKEL, P. (2001): The earth mover's distance is the Mallows distance: Some insights from statistics. In: Proceedings of *Int. Conf. on Computer Vision,* Vancouver, Canada, 251-256.

RUBNER, Y., TOMASI, C. and GUIBAS, L. (2000): The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision 40 (2), 99-121.*

STREHL, A., GHOSH, J. and MOONEY R.J. (2000): Impact of similarity measure on web-page Clustering. In: AAAI.

YANG, Y. and LIU, X. (1999): A re-examination of text categorization methods. In: Proceedings of *the ACM SIGIR Conference on Research and Development in Information Retrieval.*

ZHANG, K. and KWOK, J.T. (2006): Simplifying mixture models through function approximation. In: *Neural Information Processing Systems*. NIPS.

Strategies in Identifying Issues Addressed in Legal Reports

Gilbert Ritschard[1], Matthias Studer[1], and Vincent Pisetta[2]

[1] Department of Econometrics, University of Geneva, Switzerland,
{*gilbert.ritschard,matthias.studer*}*@metri.unige.ch*
[2] ERIC Laboratory, University of Lyon 2, France, *vincent.pisetta@univ-lyon2.fr*

Abstract. This paper deals with the automatic retrieval of issues reported in legal texts and presents an experience with expert's reports on the application of ILO Conventions. The aim is to provide the end user, i.e. the legal expert, with a set of rules that permits her/him to find among a predefined list of issues those addressed by any new text. Since the end user is not supposed to be able to pre-process the text, we need rules that can be directly applied on raw texts. We present the strategy followed for generating the rules in this ILO legal setting and single out a few possible improvements that should significantly improve the performance of the retrieval process. Our approach consists in characterizing in a first stage a list of descriptor concepts, which are then used to get a quantitative representation of the texts. In the learning phase, using a sample of texts labeled by legal experts with the issues they actually address, we build the rules by means of induced decision trees.

Keywords: information retrieval, content prediction, quantitative text representation, legal texts

1 Introduction

The concern of the paper is the automatic identification of the type of issues reported by given legal texts, for example which violations are pointed out in experts' comments on the application of ILO (International Labor Office) Conventions. Such an automatic text mining process becomes necessary when we face a large number of texts for either 1) pointing out the most relevant texts when one wants to investigate a given issue, or 2) drawing synthetic analyses of the relationships between issues as well as with other factors. The objective is then essentially to provide the end user, i.e. the legal expert, with prediction rules of the issues addressed by each text. We consider the case where the issues of interest have been previously specified. We assume thus that we have a closed list of issues.

The paper describes the process followed for building such rules within a joint research project between the ILO, the University of Geneva and the University of Lyon 2 (Ritschard et al., 2007) on the Social dialogue regimes prevailing in democratic countries. We also single out the main weaknesses of the approach and propose a series of strategies for improving the process.

The approach followed consists in characterizing in a first stage a list of descriptor concepts from which we derive then a quantitative representation of the texts. In the learning phase, using a sample of texts labeled with the addressed issues by legal experts we build the rules by means of induced decision trees. A separate tree is grown for each of the issue. Each time a binary variable indicating whether the issue is present or not is used as target variable and the descriptor concepts serve as predictive attributes. The characterization of the descriptor concepts and the quantification of their importance within each text is obviously a crucial stage in our process. Once the rules were obtained, we had to provide the end user (legal expert) a simple piece of software that 1) builds in an automatic way the quantitative representation of any new text, i.e. evaluates the importance of each descriptor concept in the text, and 2) determines from that representation what the probability is that the text addresses each issue of interest.

To make our presentation less abstract, a few words are worth on the application context for which the described text mining strategy was developed. The aim of text mining was to help us identify the nature of issues raised by a Committee of experts (CEACR) regarding the application of ILO Conventions. Due to space constraints we consider here only Convention 87 on Freedom of association and protection of right to organize. What we want to know is what types of violations of this Convention does the Committee identify in its reports. Using a priori knowledge, we categorized the possible violations in the form of a list of 9 key concepts — types of violations — (Table 1) themselves derived from a more detailed list of 27 key concepts listed in Ritschard et al. (2007).

v_1	Right to life and physical integrity (not observed)
v_2	Right to liberty and security of person / Right to a fair trial (not observed)
v_3	Right to establish and join workers' organizations
v_4	Trade union pluralism
v_5	Dissolution or suspension of workers' organizations (not observed)
v_6	Election of representatives / Eligibility criteria
v_7	Organization of activities / Protection of property / Financial independence
v_8	Approval and registration of workers' organizations
v_9	Restrictions on the right to industrial action

Table 1. Retained key concepts, i.e. types of violation.

The paper is organized as follows. In Section 2 we discuss the usefulness and inconvenience of text pre-processing and explain in Section 3 the semantic preserving text representation that was retained. The learning process itself is described in Section 4, where we give also some experimentation results illustrating the efficiency of the process. Concluding remarks are given in Section 5.

2 Text pre-processing

Text mining (Feldman and Dagan, 1995; Fan et al., 2006) refers to the process of analysing text to extract information that is useful for particular purposes (Witten and Frank, 2005, pp 351-356). It is supposed to be more than just finding documents or pages containing a given keyword — which is what simple indexing or search engines do well. For instance, if we are looking for texts commenting on violations of the freedom to organize the election of trade union representatives, we will not be satisfied with just texts containing the keyword "election", but we may want to consider also all terms or expressions more or less related to this notion such as for example "elected workers' representative" or "union leader".

As opposed to numerical data, text data are essentially unstructured. Synonymy (different expressions with same meaning) and polysemy (different meanings for a same expression), among others, make them hard to analyse in an automatic way and necessitate heavy pre-processing. The aim of the pre-processing is to transform the essentially unstructured text data into a suitable structured representation for further automatic processing. By structured representation we mean a representation where each useful notion is uniquely and unambiguously defined so that we can surely rely on the counts of its occurrences.

There are basically two main ways of representing a text: through n-grams and as a bag of words. The former ignores the meaning of the words and considers each subsequence of say 3 letters — 3-gram — that can be found in the words as a countable characteristic (Damashek, 1995; Mayfield and McNamee, 1998). The second (Salton et al., 1992, 1996) retains each different observed word as a characteristic and focuses essentially on its frequency in the text and among the texts. The latter approach is best suited for our supervised classification purpose where the semantic content of the text is of primary importance.

Now, texts contain a huge number of different words. Some of them may have a same or similar meaning (synonyms), may have a context dependent meaning (polysemy), or, as in the case of function or stop words (the, to, from, or, and, ...), will clearly be useless for discrimination purposes. The general practice is then to reduce the number of descriptors by dropping useless stop words and by merging synonyms into equivalence classes.

A first step for solving ambiguities is tagging words grammatically, which can be done automatically using for instance freely available tools such as Brill (Brill, 1995) or TreeTagger (Schmid, 1994). The grammatical tag permits indeed to distinguish for example between the noun, verb or adjective usage of the word "trade", or the conjunction, verb or adjective usage of the word "like". This grammatical tagging will also pinpoint stop words that could be dropped from the list of descriptors.

To avoid bothering with the various inflected forms of nouns, verbs and adjectives, other often applied pre-processing operations are lemmatization

and stemming (Plisson et al., 2004). The former consists in retaining just the base form — e.g. the infinite of a conjugated verb — of each encountered word, and the latter in extracting the lemma — the root — of each word. This can again be done almost automatically with freely available tools such as TreeTagger (Schmid, 1994).

In our case, since the goal is to facilitate the processing of new additional texts by legal experts with no experience in these pre-processing steps, we opted for an approach that avoids in its application phase any pre-processing operation that could not be fully automatized. Therefore, we chose to not lemmatize the texts, and resorted to grammatical tagging only during the learning phase in order to facilitate the extraction of the useful terminology.

3 The chosen text representation

For the purpose of our analysis, we decided to represent the CEACR comments by means of a limited set of descriptor concepts. These concepts were defined in a partially automated process consisting in first extracting the useful terminology, then grouping the terms into concepts and eventually refining the description of the concepts. We begin by commenting the terminology extraction process.

3.1 Extracting the useful terminology

The terminology that could be used for predicting violations reported in the Committees's observations includes not only single words, but also composite expressions such as "trade union" or "right to organize". It is then essential to find and list the terms useful for the analysis.

Several tools can be used for this. Some of them, such as Xtract (Smadja, 1993), ATR (Frantzi et al., 2000), Lexter (Bourigault and Jacquemin, 1999) proceed automatically either by comparison with a pre-specified lexicon or by seeking frequent sub-sequences of words. Others, such as Exit (Heitz et al., 2005), are semi-automatic and require a domain expert to guide the process. The latter are best suited when, as in our case, we do not have access to a lexicon of the considered specialized language. Since we had the possibility to interact with legal experts, we chose to extract the useful terminology with the aid of the Exit software.

The input data provided to Exit is the grammatically tagged text (the set of all comments merged into a single file). We then select the useful terms in an iterative way. First, we chose successively among single words or pairs of a given type — noun-noun, noun-adjective, adjective-noun, verb-noun, noun-verb, etc. — that satisfy a minimal frequency criterion those that the expert considers relevant for the analysis. For example, "worker organization" and "national security" are two retained pairs, the former being of the noun-noun type and the latter of the "adjective-noun" type. A grammatical tag is

assigned to each new retained term according to rules that could be changed by the user. For instance, adjective-noun terms such as "national security" are automatically tagged as noun. Then by iterating the process we single out terms that include themselves previously defined terms. We get thus terms composed of more than two words such as "minimum level of service".

3.2 Descriptor concepts

There is a huge number of different terms — words and composite expressions — used in the CEACR comments and it is not convenient to use all of them as text descriptors. We therefore, decided to represent texts through a small number of descriptor concepts that: (i) Characterize the conceptual content of the text; (ii) Are useful for predicting the issues — violation or key concepts — reported in the observations.

A first entirely statistical possibility of characterizing descriptor concepts (Kumps et al., 2004) would be to seek the words that best discriminate the key concepts we want to predict, and then to group them according to their co-occurrences. Lemmatization would be necessary in that case.

However, since we had the possibility to interact with legal experts, we preferred to rely on a linguistic approach. Such an approach where terms — words and expressions — are grouped according to both their statistical characteristics and the similarity of their meaning, provide concepts that are semantically better founded.

Thus, the approach followed consists in three steps carried out on the overall corpus: i) a preliminary set of concepts is built during the terminology extraction with Exit; ii) this preliminary set and the concept definitions are refined through an extensional induction process (Kodratoff, 2004) with the legal experts; and iii) the experts' amended list is once again compared with the text content for a final coherence check.

The preliminary concept set is obtained in a semi-automatic way by starting the term extraction process with a high threshold, which provides a relatively short list of terms. Those terms may be considered as initial representatives of the main conceptual axes that can be found inside the texts. We obtain a starting set of concepts after possibly grouping terms with similar semantic meaning. Then, we repeat the process by lowering successively the minimal frequency threshold. At every iteration, we get additional terms and then assign each one of them to the most appropriate preexisting concept. In case there is no reasonable preexisting concept with which the new term could be associated, a new concept is created. At the end of the terminology extraction we get our preliminary list of concepts, where each concept is characterized by its list of associated terms.

This preliminary list of descriptor concepts serves then as a starting list for the experts who may either confirm the relevance of the concepts or change them to fit their overall knowledge of the domain. The preliminary list is thus transformed into an expert's amended list of concepts.

In order to increase even further the coherence of the amended descriptor concepts, we carried out some additional checking. Indeed, we observed that the overall corpus of CEACR comments contains some infrequent terms that clearly belong to one of the retained descriptor concepts. Ignoring them would undoubtedly be a source of errors. The goal of the additional checking is to browse the corpus for such relevant but infrequent terms. More specifically, for each term already associated to a concept, we look for the presence in the corpus of synonyms and alternative inflection forms as well as for the presence of extended terms obtained by inserting one or more words in the term. For example, the term "call a strike" is frequent in the corpus and was detected as representing the strike action descriptor concept. Less frequent expressions such as "calling a strike" or "calling of a strike", were not detected however. The search of such alternative forms is easily done by browsing the terms found with regular expressions. For example, using the two strong words "call" and "strike", all three aforementioned terms were found with the PCRE regular search expression:[1]

```
"/[^;\.]*call[^;\.\,]{0,45}strike[^;\.]*/i"
```

.

As for synonyms, a lexicon such as the online WordNet may be useful for usual terms. For a specialized corpus such as the one formed by our legal texts, it is more helpful to ask experts in the domain. This is what was done in our analysis. Good sense may also prove useful. For example, we noticed in the reports that experts used independently and equivalently the terms "trade union" and "workers organization". Hence, each time a concept definition list included a term such as "registration of a trade union", we augmented, when it made sense, the list with "registration of a workers organization", even when this new expression was infrequent in the corpus.

The final list of descriptor concepts is given in Table 2 and examples of their list of associated terms can be found in Ritschard et al. (2007).

The designing of the descriptor concepts is clearly a crucial stage of our text mining process. It is also time-consuming and requires clever tuning through individual interventions from both the domain experts and the text mining experts. Furthermore, because of these multiple personal interventions, the resulting descriptor concepts remain somewhat subjective. Improvement and systematization of the process is possible and would here be necessary. It requires, however, an access to a detailed ontology of the concerned legal domain which does not yet exist. The designing of such an ontology that puts together the characteristic terminology of the domain, organizes it in terms of concepts and sub-concepts, and also describes the interrelation between concepts would then be our next development priority.

[1] The regular expression searches the text for expressions in which the word "call" is preceded by any sequence of characters other than a semi-column or a dot, the word "strike" is followed by any sequence of characters other than a semi-column or a dot, and the two words are separated by any sequence of at most 45 characters other than a semi-column, a dot or a comma.

c_1	Life and physical integrity	c_{10}	Industrial action
c_2	Liberty and security of persons	c_{11}	Essential service
c_3	Property and financial independence	c_{12}	Arbitration
c_4	Service	c_{13}	Strike action
c_5	Pluralism	c_{14}	Union establishment limitations
c_6	Election	c_{15}	Specific workers
c_7	Opinion and expression freedom	c_{16}	Number of workers
c_8	Restrictions on trade union activities	c_{17}	Supervision
c_9	Trade union approval		

Table 2. Retained descriptor concepts.

3.3 The quantitative text representation

Having now defined our descriptor concepts, we get a quantitative representation of the texts by assigning for each document (comment) a load on each concept. A classical way is to use the $tf \times idf$, which is the term frequency (tf) — indeed the term count — in the document weighted by the inverse of the document frequency (idf), the document frequency being the number of documents in which the concept has been observed (Salton and Buckley, 1988). The general idea of this $tf \times idf$ is that a term — a concept in our case — is characteristic of a text when it is frequently mentioned in it (high tf) and only few other documents mention it (high idf). Let tf_{ij} be the term frequency of concept j in document i, and idf_j be the inverse term frequency of concept j. Formally, the inverse document frequency is defined as $\log(d/d_j)$, where d is the total number of documents and d_j the number of documents mentioning concept j. The $tf \times idf$ weight of concept j in a document i, is then

$$w_{ij} = tf_{ij}\, idf_j = tf_{ij} \log\left(\frac{d}{d_j}\right) .$$

With this formulation, the lengthier a document i the greater chance it has to have large tf_{ij}'s and hence important weights. To avoid this size effect, Salton et al. (1992) propose the length normalized form $\tilde{w}_{ij} = w_{ij}/||\boldsymbol{w}_i||$, with $\boldsymbol{w}_i$ the vector of the $tf_{ij} \times idf_j$'s of the document i.

For our objectives, what matters is the absolute place devoted to a given concept in a comment whatever other issues the comment addresses. In that sense, the normalized $tf \times idf$ is not useful in our setting. In other words, we consider that the importance of a concept in a text is reflected by its number of occurrences independently of the document's length.

Using the $tf \times idf$'s of the retained descriptor concepts, our text data set can be put in the form of a classical quantitative data table as illustrated in Table 3, which exhibits an extract of the data for comments on the application of Convention 87.

CEACR Comment	Descriptor Concepts							
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	$\cdots$
Algeria 1991	0	0	0	0	2.75	0	0.8	$\cdots$
Argentina 1991	0	0	0	0	20.59	2.39	0.8	$\cdots$
Bangladesh 1991	1.0	0.77	2.35	1.24	0	1.59	5.59	$\cdots$
$\cdots$								

Table 3. Extract of data representing comments in terms of descriptor concepts.

4 Learning process

Through the previous steps, i.e. extracting useful terms, organizing them into a limited number of relevant descriptor concepts and finally measuring the importance devoted to each descriptor concept by each CEACR comment with the $tf \times idf$ weight, we were able to code the comments numerically. What remains now is to learn the prediction rules.

This learning phase requires a learning sample of texts — comments — previously labeled in accordance with the type of violation they report. The labeling was done by a legal expert for 78 out of 671 CEACR texts concerning Convention 87. The labels are represented by a set of ℓ 0-1 indicator variables v_k, $k = 1, \ldots, \ell$ that take value 1 when the text mentions violation k, and zero otherwise. Remember that the violations we are interested in correspond to the key concepts listed in Table 1.

Using this learning sample the aim is to find rules for predicting each key concept (violation) from the quantified descriptor concepts. We then consider successively each key concept in turn, and build the prediction rule for it. Letting c_j denote the $tf \times idf$ of the jth descriptor concept, we look for each k for a prediction rule $\hat{v}_k = f_k(c_1, \ldots, c_c)$.

Since our texts are numerically coded, classical supervised statistical or machine learning techniques may be considered. We used induced classification trees, which produce usually good classification results and have the advantage of being easily applicable, of detecting automatically interaction effects of the predictors and of providing easily interpretable rules.

Classification trees are grown by seeking, through recursive splits of the learning data set, some optimal partition of the predictor space for predicting the outcome class, i.e. whether the comment does or does not report a violation of type k. Each split is done according to the values of one predictor — descriptor concept —. The process is greedy. At first step, it tries all predictors to find the "best" split using, for quantitative predictors as those we face here (the concept $tf \times idf$'s), an automatic local optimal discretisation. Then, the process is repeated at each new node until some stopping rule is reached. This requires a local criterion to determine the "best" split at each node. The choice of the criterion is the main difference between the various tree growing methods that have been proposed in the literature.

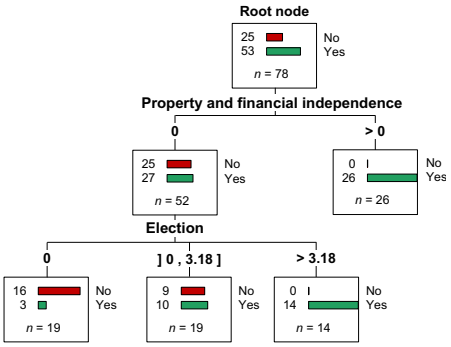

Fig. 1. Induced tree for v_7, Restrictions on organization of trade union activities.

Figure 1 shows the tree grown for violation 7 — restrictions on the organization of trade union activities — using Exhaustive CHAID (the improved CHAID method by Biggs et al., 1991) with a significance threshold of 5%, the Bonferroni correction, a minimal leaf size of 10 and a minimal parent node size of 30. The descriptors retained are whether the comment explicitly refers to property and financial independence and whether it talks about election.

The tree has 4 terminal nodes, which are called *leaves*. We associate to each of them a rule taking the form *condition* $\Rightarrow$ *conclusion*. The condition is defined by the path from the root node to the leaf, and the conclusion is, for a classification tree, usually the most frequent class in the leaf.

A similar tree is grown for each type of violation, which results in 6 sets of rules. Some violations (v_1, v_2 and v_5 for instance), are not covered by any comment in the learning sample, and no tree is grown for them. In two cases, we did not rely on the mere statistical criterion and forced the algorithm to split at the first step using the second best variable that seemed theoretically better sounded from our knowledge base.

The classification performance of each tree may be evaluated by means of its classification error, i.e. the percentage of cases which are misclassified

Key concept (violation)	Learning error rate	Cross-validation error rate	Cross-validation std err	Test sample (size 21) number of errors
v_3	14.10%	n.a.*	n.a.*	3
v_4	5.13%	5.13%	2.50%	0
v_6	12.82%	14.1%	3.94%	4
v_7	15.38%	n.a.*	n.a.*	7
v_8	7.69%	7.69%	3.01%	4
v_9	2.56%	2.56%	1.79%	2

*Cross-validation is not available for v_3 and v_7, because first split is enforced.

Table 4. Error rates, Convention 87.

Key	Positives		Negatives		% with key concept			
Concept	true	predicted	true	predicted	reported	predicted	Recall	Precision
v_3	30	32	37	46	50.0%	41.0%	76.9%	93.8%
v_4	29	31	45	47	39.7%	39.7%	93.5%	93.5%
v_6	35	38	33	40	53.8%	48.7%	83.3%	92.1%
v_7	50	59	16	19	67.9%	75.6%	94.3%	84.7%
v_8	29	30	43	48	43.6%	38.5%	85.3%	96.7%
v_9	57	59	19	19	73.1%	75.6%	100.0%	96.6%

Table 5. False Positives, False Negatives, Recall and Precision, Convention 87.

by the derived classification rules. Table 4 shows learning error rates (i.e. rates computed on the learning sample) and 10-fold cross-validation error rates with their standard error. It gives in addition the number of errors on a small test sample of 21 comments about the application of Convention 87.

Table 5 exhibits some additional useful indicators. Column 'True positives' gives the number of comments classified as reporting a violation of type k that effectively report it, and column 'Predicted positives' the total number of comments classified as reporting the violation. For key concept v_7, for example, 50 out of 57 comments classified as reporting the violation actually report it. The number of true and predicted negatives is also shown. Table 5 gives the percentage of the 78 comments that report on the relevant key concept and the percentage of comments that are classified as reporting the key concept. For v_7 again, we may check that $59 = 75.6\% \times 78$, are classified as reporting the violation, while there is actually a total $53 = 67.9\% \times 78$ reporting v_7. The 'Recall' is the percentage of this total that is classified as reporting the violation — true positives —, e.g. $94.7\% = 50/53$ for v_7. The 'Precision' is the ratio of the number of true positives on the number of predicted positives, e.g. $84.7\% = 50/59$ for v_7.

These results are quite good when compared with those obtained with other classifiers. For instance, we experimented with support vector machine (SVM) as well as with neighboring graphs. These methods did not produce significantly better results, while producing much less explicit rules. Nevertheless, error rates above 10% as well as recall and precision percentages below 90% may look unsatisfactory. Remember, however, that the learning was done with a sample of only 78 texts. It is also worth mentioning that errors may be more or less important depending on the research objectives. In our case, as stated in the introduction, the text mining has two main purposes: To help the legal expert interested in a given issue in identifying texts reporting this issue (it is not supposed to replace the expert in this task), and to provide material for analysing synthetically the relationship between issues, i.e. types of violations. With such objectives, it is not dramatic to make false predictions for a small number of texts. If the end user wants to find all texts dealing with an issue of interest, false positive cases will generally

be easier to identify than false negative ones. Hence we should in that case favor a strategy that limits false negatives even if it is at the cost of more false positives. This can easily be done by lowering for instance the probability threshold used for assigning the outcome class to the rules. For synthetic analyses, on the other hand, we may prefer to retain only the most reliable predictions. We would then primarily limit the number of false positives.

5 Conclusion

We described in this paper an ad hoc text mining process for identifying issues reported in legal texts. The process described is semi-automatic. The building of the prediction rules relies on an interaction with the domain expert at several points and especially for defining relevant descriptor concepts. This stage of the process could, however, be improved on at least two sides. First, the interest of the descriptor concepts for the targets (each associated to one of the considered violations) is based solely on the opinion of the domain expert. By specifying a global criterion taking simultaneously into account all considered targets, it should be possible to measure the global discriminating power of terms and hence select objectively the most discriminating ones. Likewise, we should be able to measure the similarity in the discriminating capacity of the terms and use these similarities as a guide for grouping them into descriptor concepts. Second, organizing the descriptor concepts into hierarchical ontology would allow for some freedom for choosing between concepts and sub-concepts. It would also produce reusable knowledge material for other applications in similar domains. Beside the systematization of the descriptor concept definition stage, significant improvement may also be expected at the learning level. For instance, taking account of the preference for limiting false positives rather than false negatives (or conversely) during learning and not only during class assignment should most probably generate better suited rules.

References

BIGGS, D., DE VILLE, B. and SUEN, E. (1991): A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics 18*(1), 49–62.

BOURIGAULT, D. and JACQUEMIN, C. (1999): Term extraction + term clustering: An integrated platform for computer-aided terminology. In *Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pp. 15–22.

BRILL, E. (1995): Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics 21*(4), 543–565.

DAMASHEK, M. (1995): Gauging similarity with ngrams: Language-independent categorization of text. *Science 267*, 843–848.

FAN, W., WALLACE, L., RICH, S. and ZHANG, Z. (2006): Tapping the power of text mining. *Communications of the ACM 49*(9), 76–82.

FELDMAN, R. and DAGAN, I. (1995): Knowledge discovery in textual databases (KDT). In *KDD '95*, pp. 112–117.

FRANTZI, K.T., ANANIADOU, S. and MIMA, H. (2000): Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries 3*(2), 115–130.

HEITZ, T., ROCHE, M. and KODRATOFF, Y. (2005): Extraction de termes centrée autour de l'expert. *Revue des nouvelles technologies de l'information RNTI E-5*, 685–690.

KODRATOFF, Y. (2004): Induction extensionnelle: définition et application l'acquisition de concepts à partir de textes. *Revue des nouvelles technologies de l'information RNTI E-2*, 247–252.

KUMPS, N., FRANCQ, P. and DELCHAMBRE, A. (2004): Création d'un espace conceptuel par analyse de donnés contextuelles. In G. Purnelle, C. Fairon, and A. Dister (Eds.), *Le Poids des Mots (JADT 2004)*, Volume 2, pp. 683–691. Presse Universitaire de Louvain.

MAYFIELD, J. and MCNAMEE, P. (1998): Indexing using both n-grams and words. In *TREC*, pp. 361–365.

PLISSON, J., LAVRAČ, N. and MLADENIĆ, D. (2004): A rule based approach to word lemmatization. In *Proceedings of IS04*.

RITSCHARD, G., ZIGHED, D.A., BACCARO, L., GEORGIOU, I., PISETTA, V. and STUDER, M. (2007): Mining expert comments on the application of ILO Conventions on freedom of association and collective bargaining. Working Papers 2007.02, Department of Econometrics of the University of Geneva.

SALTON, G., ALLAN, J. and SINGHAL, A. (1996): Automatic text decomposition and structuring. *Information Processing and Management 32*(2), 127–138.

SALTON, G. and BUCKLEY, C. (1988): Term-weighting approaches in automatic text retrieval. *Information Processing and Management 24*(5), 513–523.

SALTON, G., BUCKLEY, C., and ALLAN, J. (1992): Automatic structuring of text files. *Electronic Publishing—Origination, Dissemination, and Design 5*(1), 1–17.

SCHMID, H. (1994): Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing, Manchester*.

SMADJA, F.A. (1993): Retrieving collocations from text: Xtract. *Computational Linguistics 19*(1), 143–177.

WITTEN, I.H. and FRANK, E. (2005): *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Amsterdam: Morgan Kaufman (Elsevier).

Part IX

Knowledge Extraction by Models

Sequential Automatic Search of a Subset of Classifiers in Multiclass Learning

Francesco Mola and Claudio Conversano

Department of Economics, University of Cagliari,
Viale Frá Ignazio 17, I-09123, Cagliari, Italy, *mola@unica.it, conversa@unica.it*

Abstract. A method called *Sequential Automatic Search of a Subset of Classifiers* is hereby introduced to deal with classification problems requiring decisions among a wide set of competing classes. It utilizes classifiers in a sequential way by restricting the number of competing classes while maintaining the presence of the true (class) outcome in the candidate set of classes. Some features of the method are discussed, namely: a cross-validation-based criteria to select the best classifier in each iteration of the algorithm, the resulting classification model and the possibility of choosing between an heuristic or probabilistic criteria to predict test set observations. Furthermore, the possibility to cast the whole method in the framework of unsupervised learning is also investigated. Advantages of the method are illustrated analyzing data from a letter recognition experiment.

Keywords: classification, subset selection, decision tree, cross-validation, Bayes rule

1 Introduction

Multiclass classification is a difficult task in statistical learning. It requires a classifier to discriminate instances (objects) among several classes of an outcome (response) variable. Typical examples are handwritten character recognition (Lee and Seung, 1997), POS-tagging (Even-Zohar and Roth, 2001) and Customer Relationship Management (Prinzie and Van den Poel, 2005).
To get *K-class classifiers*, the most common approach is to construct a set of binary classifiers each trained to separate one class from the rest (the so-called *One Versus the Rest*), and to combine them according to a model averaging criteria as, for example, by voting (Cutzu, 2003). Main shortcomings of this *winner-takes-all* strategy is that the resulting binary classifiers are obtained from different asymmetric binary classification problems. Thus, it is unclear whether they are comparable or not. Difficulty in class-assignment may arise when each classifier assigns an instance to its reference-"one"-class leading to a situation in which no final class can be chosen. An alternative (voting-based) approach is *Pairwise Classification* (Hastie and Tibshirani, 1998). It uses an *all-versus-all* strategy by training a classifier for each possible pairs of response classes. Although larger training time is required compared to the previous case, the individual classification problems are significantly smaller

because the training sets are smaller and the learning goals are easier since the classes are less overlapped. A somewhat similar approach is *Error-Correcting Output Coding* (Dietterich and Bakiri, 1995). Here, a large number of binary classification problems is generated by splitting the original set of clsses into two subsets and by training a binary classifier for each possible dichotomization. The final classification derives from a synthesis of the results obtained from each binary classification example which are stored in a decoding matrix composed of $\{\pm 1\}$. Finally, other approaches directly cast the multiclass classification problem into an objective function that simultaneously allows the computation of a multiclass classifier as, for instance, in the Weston and Herbrich's (2000) approach based on Support Vector Machines (SVM). Despite of its elegant formulation and high accuracy, this approach lacks of feasibility in same situations because it has to deal simultaneously with many SVs. All the aforementioned approaches are quite effective, but it is fair to say that there is probably no multiclass approach that generally outperforms the others (Scholkopf and Smola, 2001: p. 214).

In the following, we introduce a multiclass classification algorithm called *Sequential Automatic Search of a Subset of Classifiers* (SASSC) that build on some of the advantages of the main approaches proposed in the literature. It adaptively and sequentially aggregates subset of instances related to a proper aggregation of a subset of the response classes, that is, to a *superclass*. In each iteration, aggregations are based on the search of the subset of instances whose response classes generate a classifier presenting the lowest generalization error compared to other alternative aggregations. Crossvalidation is used to estimate such generalization errors. User can choose a final number of subsets of the response classes (*superclasses*) obtaining a final classifier-based model for multiclass classification presenting an high level of accuracy without neglecting parsimony. In this respect, this approach is inspired by the *model-based knowledge discovery* paradigm since the number of classifiers included in the final model is relatively small so that interpretability of results is strictly preserved.

The motivation underlying the formalization of the SASSC algorithm derives from the following intuition: basically, since standard classifiers unavoidably lead to prediction inaccuracy in the presence of multiclass response, it would be favourable to look for a relatively reduced number of classifiers each one relating to a subset of classes of the response variable (superclasses). Reducing the number of response classes for each of those classifiers naturally leads to improve the overall prediction accuracy. To further enforce this guess, an appropriate criterion to derive the correct number of superclasses and the most parsimonious classifier for each of them has to be found. To this purpose, a sequential approach that automatically proceeds through subsequent aggregations of the response classes might be a natural starting point.

The remainder of the paper is as follows. Section 2 describes the SASSC algorithm and introduces the related classification model. The results from

both the application of such a model on the Letter Recognition dataset and the comparison of the performance of SASSC with respect to alternative approaches are summarized in section 3. Section 5 briefly introduces a dendrogram-based visualization of the aggregations produced by the algorithm and cast the method in the framework of unsupervised learning. Section 5 discusses the possible advantages connected to the use of the proposed method.

2 Sequential aggregation of a subset of classifiers

2.1 Algorithm

SASSC produces a partition of the set of the response classes into a reduced number of superclasses. It is applicable to a dataset $\mathbf{D}$ composed of N instances characterized by a set of J (numerical or categorical) inputs X_j $(j = 1, \ldots, J)$ and an outcome Y presenting K classes. Such response classes identify the initial set of classes $C^{(0)} = (c_1, c_2, \ldots, c_K)$. Partitioning $\mathbf{D}$ with respect to $C^{(0)}$ allows to identify K disjoint subsets $\mathbf{D}_k^{(0)}$, such that: $\mathbf{D}_k^{(0)} = \{\mathbf{x}_s : \mathbf{y}_s \in c_k\}$, with $s = 1, \ldots, N_k$. In practice, $\mathbf{D}_k^{(0)}$ is the set of instances presenting the k-th class of Y. The algorithm works by aggregating the K classes in pairs and by learning a classifier for each subset of corresponding instances. The "best" aggregation (*superclass*) is chosen as the one minimizing the generalization error estimated using V-fold cross-validation. Suppose that, in the ℓ-th iteration, such a best aggregation is found for the pair of classes c_{i^*} and c_{j^*} (with $i^* \neq j^*$ and $i^*, j^* \in (1, \ldots, K)$) that allow to aggregate subsets $\mathbf{D}_{i^*}$ and $\mathbf{D}_{j^*}$. Denoting with $T_{(i^*,j^*)}$ the classifier minimizing the cross-validated generalization error $\theta_{cv}^{(\ell)}$, the criteria for selecting the best classifier can be formalized as follows:

$$(i^*, j^*) = \arg\min_{(i,j)} \left\{ \theta_{cv}^{(\ell)} \left(T_{(i,j)} | \mathbf{D}_i \cap \mathbf{D}_j \right) \right\} \tag{1}$$

The SASSC algorithm is analytically described in Table 1. It proceeds by learning all the possible classifiers obtainable by joining in pairs the K subgroups, retaining the one satisfying the selection criteria introduced in (1). After the ℓ-th aggregation, the number of subgroups is reduced to $K^{(\ell)} = K^{(\ell-1)} - 1$, since the subgroups of instances presenting the response classes c_{i^*} and c_{j^*} are discarded from the original partition and replaced by the subset $\mathbf{D}_{(i^*,j^*)}^{(\ell)} = (\mathbf{D}_{i^*} \cap \mathbf{D}_{j^*})$ identified by the super-class $c^{(\ell)} = (c_{i^*} \cap c_{j^*})$. The initial set of classes C is replaced by $C^{(\ell)}$, the latter being composed of a reduced number of classes since some of the original classes form the superclasses coming out from the ℓ aggregations. Likewise, also $\mathbf{D}_k^{(\ell)}$ is formed by a lower number of subsets as a consequence of the ℓ aggregations. The algorithm proceeds sequentially in the iteration $\ell + 1$ by searching for the

Table 1. The SASSC algorithm.

Input:	$C = \{c_1, \ldots, c_K\}_{c_i \cap c_j = \emptyset; i \neq j; i,j \in (1,\ldots,K)}$
Set:	$C^{(0)} = C; \quad K^{(0)} = K; \quad \theta_{cv}^{(0)} = 0;$ $\mathbf{D}_k^{(0)} = \{\mathbf{x}_s : y_s \in c_k\}_{s=1,\ldots,N_k; k=1,\ldots,K}$
For: ℓ in 1 to K	$c^{(\ell)} = \{c_{i^*} \cap c_{j^*}\} : \theta_{cv}^{(\ell)}(T_{(i^*,j^*)} \vert \mathbf{D}_{i^*} \cap \mathbf{D}_{j^*}) = \min$ $K^{(\ell)} = K^{(\ell-1)} - 1$ $C^{(\ell)} = \{c_1, \ldots, c_{K^{(\ell)}-2+1} = c^{(\ell)}\}$ $\mathbf{D}_k^{(\ell)} = \{\mathbf{x}_s : y_s \in c_k\}_{k=1,\ldots,K^{(\ell)}-1}$
end For	
Output:	$C^{(1)}, \ldots, C^{(K-1)}; \quad T_{(1)}, \ldots, T_{(K-1)}; \quad \Theta_{cv}^{(1)}, \ldots, \Theta_{cv}^{(K-1)}$

most accurate classifier over all the possible ones obtainable by joining in pairs the $K^{(\ell)}$ subgroups. The sequential search is repeated until the number of subgroups reduces to one in the K-th iteration. The classifier learned on the last subgroup corresponds to the one obtainable by learning a classifier on the original dataset.

The output of the procedure is a sequence of sets $C^{(1)}, \ldots, C^{(K-1)}$ of response classes with the associated sets of classifiers $T_{(1)}, \ldots, T_{(K-1)}$. The latter are derived by learning $K - k$ classifiers ($k = 1, \ldots, K-1$) on disjoint subgroups of instances whose response classes complete the initial set of classes $C^{(0)}$: these response classes identify the superclasses relating to the sets of classifiers $T_{(k)}$. An overall generalization error is associated to each $T_{(k)}$: such an error is also based on V-fold cross-validation and it is computed as a weighted average of the generalization errors obtained from each of the $K - k$ classifiers composing a set $C^{(k)}(k = 1, \ldots, K-1)$. In accordance to the previously introduced notation, the overall generalization errors can be denoted as $\Theta_{cv}^{(1)}, \ldots, \Theta_{cv}^{(k)}, \ldots, \Theta_{cv}^{(K-1)}$. Of course, by decreasing the number of classifiers composing a sequence $T_{(k)}$ (that is, when moving k from 1 to $K-1$) the corresponding $\Theta_{cv}^{(k)}$ increases, since the number of superclasses associated to $T_{(k)}$ is also decreasing. This means that a lower number of classifiers are learned on more heterogeneous subsets of instances, because each of those subsets pertain to a relatively large number of response classes. Taking this inverse relationship into account, the analyst can be aware of the overall prediction accuracy of the final model on the basis of the relative increase in $\Theta_{cv}^{(k)}$ when moving from 1 to $K-1$. In this respect, he can select accordingly a suitable number of classifiers to be included in the final classification model.

2.2 Tree-based classification model

The SASSC algorithm can be applied by using as classifier one of the prediction model typically used in the statistical learning framework such as, among others: discriminant analysis, logistic regression, kernel methods or SVM. The current implementation uses CART-like decision trees (Breiman et al., 1984) as classifier, because of their nonparametric and distribution-free characteristics as well as of their effectiveness in dealing with nonlinearly associated inputs. Besides their functional flexibility, classification trees are also useful in dealing with heterogeneity, as separate models can be automatically fit to previously identified subsets of data. Last but not least, they are very efficient at selecting from large number of inputs, which is a typical prerequisite in data mining applications. Despite these advantages, decision trees suffer from instability: we account for such a drawback using cross-validation.

Supposing a final subset $C^{(*)}$ of g classifiers has been selected ($g << K-1$), the estimated classification model can be represented as:

$$\hat{f}(X) = \sum_{i=1}^{g-1} \sum_{m_i=1}^{M_i} \hat{\tau}_i \hat{c}_{k,i} I((X_1, \ldots, X_J) \in R_{m_i}) \tag{2}$$

This notation is consistent with that used in Hastie et al. (2001). The parameter τ is called "vehicle parameter". It allows to assign a new instance to the most suitable classifier in the subset $C^{(*)}$. It is defined by a set of $g-1$ dummy variables. Each of them equals 1 if the object belongs to the i-th classifier ($i = 1, \ldots, g-1$) and zero otherwise. The M_i regions, corresponding to the number of terminal nodes of the classifier i, are created by splits on inputs $(X_1, \ldots, X_J)$. The classification tree i assigns a new object to the class $\hat{c}_{k,i}$ of Y according to the region R_{m_i}. $I(\cdot)$ is an indicator function with value 1 if an instance belongs to R_m and value 0 if not. R_{m_i} is identified by the inputs used in the splits leading to that terminal node. The modal class of the instances in a region R_{m_i} (also called the m^{th} terminal node of the i-th classifier) is usually taken as an estimate for $\hat{c}_{k,i}$ in the training step of the algorithm.

Using decision trees as classifiers, another option of the algorithm is the possibility to learn classifiers and to select the suitable pair of response classes satisfying (1) by using alternative splitting criteria. As for CART, either the Gini index or Twoing can be used as alternative splitting rules. It is known that, unlike Gini rule, Twoing searches for two classes that make up together more than 50% of the data and allows us to build more balanced trees even if the resulting recursive partitioning algorithm works slower. As an example, if the total number of classes is equal to K, Twoing uses 2^{K-1} possible splits. Since it has been proved (Breiman et al., pag.95) that decision tree accuracy is insensitive to the choice of the splitting rule, it can be interesting to investigate how different splitting criteria works in the SASSC's framework characterized by the search of the most accurate classifiers.

2.3 Decision rule

To classify test-set instances it is useful to consider the prediction accuracy of each classifier in the final subset $C^{(*)}$: each decision tree allows to estimate the predicted class and to derive the test conditions used for such a prediction. At the same time, the assigned class $\hat{c}_{k,i}$ lead to the estimation of τ_i for each test set instance. Once that a new instance is slipped into each of the g classifiers of $C^{(*)}$, two alternative criteria can be applied:

a) a *heuristic estimation* of $c_{k,i}$: the assigned class $\hat{c}_{k,i}$ is found with respect to the tree whose terminal node better classifies the new instance. In other words, a new instance is assigned to the purest terminal node among all the g classifiers and the estimated class is the modal class of Y for the training instances falling in that node.
b) a *probabilistic estimation* of $c_{k,i}$: Supposing, in the simplest case, that $C^{(*)}$ is composed of two subsets identified by the corresponding decision tree, a new object can fall into one terminal node for each tree. These two terminal nodes assign new objects to two disjoint subsets of the response classes. Remembering K is the original number of response classes, we can define $P(c_k)$ as the prior probability for class k $(k = 1, \ldots, K)$ and $P(\tau_i = 1|c_k)$ as the likelihood of k given the i-th classifier, corresponding to the proportion of training instances that are correctly classified in the assigned terminal node of the i-th classifier. To decide which is the assigned class for the new object a measure of the (posterior) probability that the assigned class for the terminal node of the i-th decision tree is k can be obtained using the Bayes rule as follows:

$$P(c_k|\tau_i = 1) = \frac{P(c_k)P(\tau_i = 1|c_k)}{\sum_{k=1}^{K} P(c_k)P(\tau_i = 1|c_k)} \tag{3}$$

$P(c_k|\tau_i = 1)$ refers to the response classes related to the assigned terminal node of the i-th classifier. By computing such a probability for each assigned terminal node of all the g classifiers of $C^{(*)}$, the final estimated class is obtained as the maximum value of the posterior probability measured with respect to k and i, namely:

$$\hat{c}_{k,i} = \arg\min_{(k,i)} P(c_k|\tau_i = 1) \qquad (k = 1, \ldots, K);\ (i = 1, \ldots, g) \tag{4}$$

In a nutshell, this probabilistic criteria derives from averaging the proportion of correct-classified training instances of each terminal node with respect to the prior probabilities.

3 Analyzing the *Letter Recognition* dataset

In the following, SASSC is applied on the "Letter Recognition" dataset from the UCI Machine Learning Repository (Asuncion and Newman, 2008). This

dataset is originally analyzed in Frey and Slate (1991), who not achieve a good performance in terms of prediction accuracy. Later on, the same dataset is analyzed in Fogarty (1992) using nearest neighbours classification. Obtained results give over 95.4% accuracy compared to the best result of 82.7% reached in Frey and Slate (1991). Nevertheless, no information about the interpretability of the nearest neighbour classification model is provided and the computational inefficiency of such a procedure is deliberately admitted by the authors.

In the Letter Recognition analysis, the task is to classify 20,000 black-and-white rectangular pixel displays into one of the 26 letters in the English alphabet on the basis of 16 numerical attributes. Dealing with $K = 26$ response classes, SASSC provides 25 sequential aggregations. Classification trees aggregated at each single step were chosen according to 10-fold cross validation. A tree was aggregated to the sequence if it provided the lowest cross validated generalization error with respect to the other trees obtainable from different aggregations of (subgroups of) response classes. The results of the SASSC

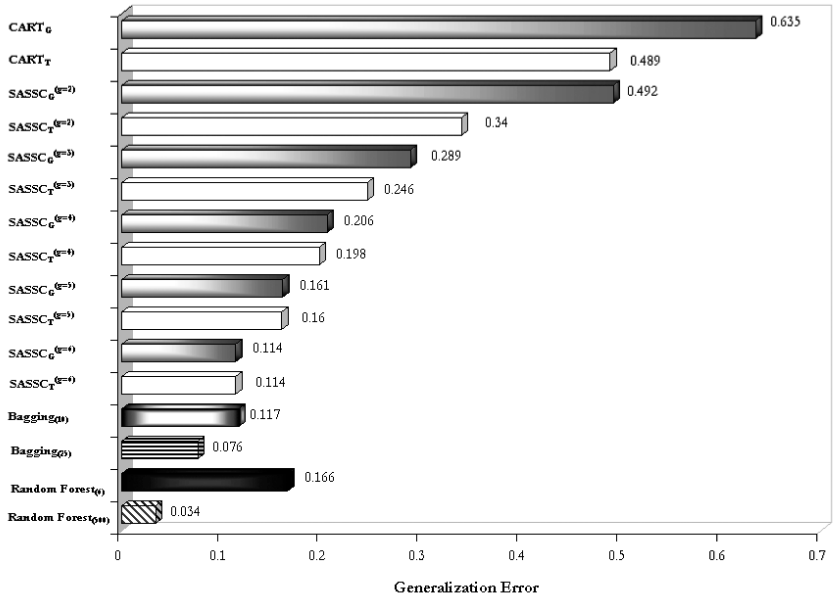

Fig. 1. The generalization errors for the Letter Recognition dataset provided by alternative approaches: as for SASSC, subscript G (T) indicates the Gini (Twoing) splitting rule, whereas apex g indicates the number of superclasses (i.e., classifiers) identified by the algorithm. The subscript for Bagging and Random Forest indicates the number of classifiers used to obtain the classification by majority voting.

algorithm are summarized in Figure 1. It compares the performance of the

SASSC model formed by $g = 2$ up to $g = 6$ subsets of the response classes with that of the CART algorithm using in both cases either Gini or Twoing as splitting rules. Bagging (Breiman, 1996) and Random Forest (Breiman, 2001) are used as benchmarking methods. Computations have been carried out using the R software for statistical computing (R Development Core Team, 2008).
The SASSC model using 2 superclasses consistently improves the results of CART using the Gini (Twoing) splitting rule since the generalization error reduces to 0.49 (0.34) from 0.52 (0.49). As expected, the choice of the splitting rule (Gini or Twoing) is relevant when the number of superclasses g is relatively small ($2 \leq g \leq 4$), whereas it becomes negligible for higher values of g (reuslts for $g \geq 5$ are almost identical). Focusing on Gini splitting criterion, the SASSC's generalization error further reduces to 0.11 when the number of superclasses increases to 6. For comparative purposes, Bagging and Random Forest have been trained using 6 and 10 classifiers respectively and, in these cases, obtained generalization errors are worse than those deriving from SASSC with $g = 6$. As for Bagging and Random Forest, as far as the number of trees used to classify each subset of randomly drawn objects increases, the performance of these two methods in terms of prediction accuracy improves. The reason is that their predictions derive form ("in-sample") independent bootstrap replications. Instead, cross-validation predictions in SASSC derives from aggregations of classifications made on "out-of-sample" observations that are excluded from the tree growing procedure. Thus, it is natural to expect that cross-validation predictions are more inaccurate than bagged ones. Of course, as far as the number of subsets of the response classes in SASSC increases, the cross-validated generalization error reduces but, at the same time, the complexity of the final classification model also increases. In spite of a relatively lower accuracy, interpretability of the results in SASSC with $g = 6$ is strictly preserved. Figure 2 shows the classifiers obtained for $g = 6$. As in standard classification trees, the user can easily understand the more influential inputs and their relative split points for each classifier. Whereas, the same kind of interpretation is not easily achievable in the case of Bagging (Random Forest) with 25 (500) bootstrap replications.

4 A partially unsupervised learning perspective

Under a different perspective SASSC consists in building a taxonomy of classes in an ascendant manner: this is done by the solution of a multiclass problem obtained by decomposing it to several r-nary problems ($r \geq 2$) in an agglomerative way. The algorithm begins with every class representing a singleton object. At each of the $K - 1$ steps the closest two (less dissimilar) classes (or subsets of classes) are merged into a single superclass, producing one less class at the next higher level. To define a measure of dissimilarity,

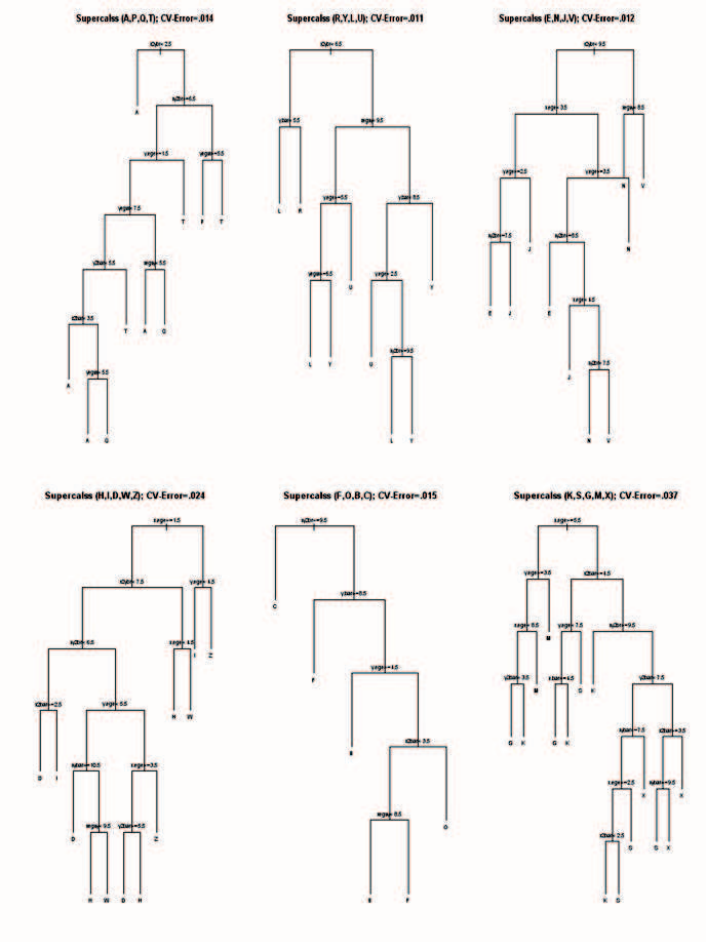

Fig. 2. The six classifiers obtained from the SASSC algorithm with $g = 6$ superclasses using the Gini splitting rule and 10-fold cross validation.

we refer to the generalization error obtained with cross-validation once that two classes (or superclasses) are aggregated. The best aggregation is the one producing the lowest generalization error. Consequently, an overall measure of quality for the superclasses obtained in each iteration is provided by the weighted average of the generalization errors obtained from the classifiers trained on each superclass: such a measure has been denoted with $\Theta_{cv}^{(k)}$ in section 2.1.

Figure 3 provides a graphical insight into a dendrogram-based representation of the SASSC's results. A somewhat similar representation for multiclass classification is in Benabdeslem and Bennani (2006). Dendrogram provides information about the overall aggregation process by illustrating the classes

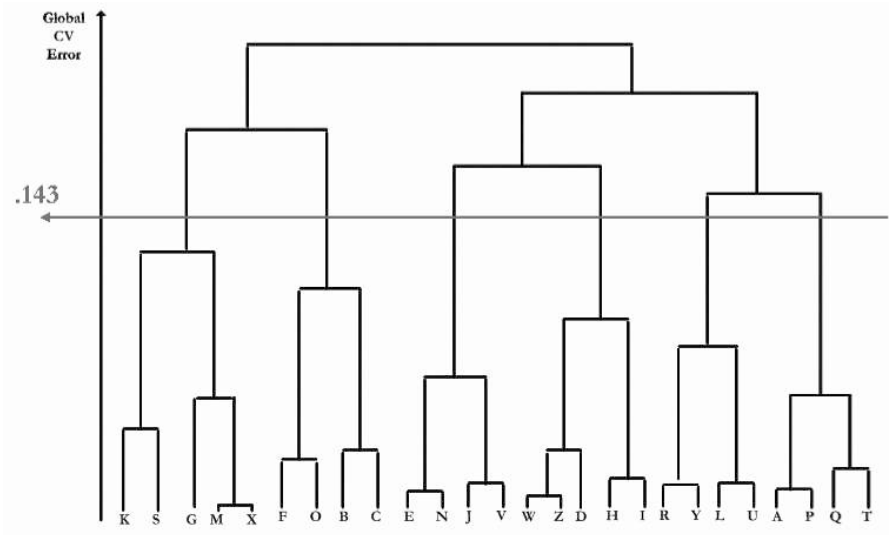

Fig. 3. A dendrogram-based representation of the SASSC's results: The horizontal line cuts the dendrogram and provides information about the global CV error provided by the six classifiers obtained by choosing $g = 6$ superclasses and by using the Gini splitting rule and 10-fold cross validation.

(English alphabet letters) aggregated at each step as well as the overall measure of impuity (global CV error) associated to each possible set of superclasses. As in standard hierarchical classification, the visual inspection of such a dendrogram can influence the user in deciding which is the final partition to take as a result of the method and to cut the dendrogram accordingly.
In this respect, SASSC's approach partially resembles what happens in standard unsupervised hierarchical clustering (see, among others, Duda et al., 2001: p. 550) with a main distinguishing issue: aggregations in SASSC involve groups of observations identified by a unique response class. As a consequence, the SASSC's classification model can not be considered as a completely unsupervised learning approach, as it uses an apriori classification of instances induced by the response classes. Nevertheless, instances are aggregated with respect to an overall goodness of fit measure (the global CV error) that partially depends on measurements of the outcome variable. As a result, SASSC can be defined as a *partially unsupervised learning* method.

5 Discussion

A large number of applications in statistical learning can be viewed as problems of resolving ambiguity based on the properties of the surrounding context. These, in turn, can all be viewed as classification problems in which the goal is to select a class label from a large collection of classes. In many

of these classification problems a significant source of difficulty is the fact that the number of candidate classes is very large. Since general purpose learning algorithms do not handle this multiclass classification problems in a proper manner, most of the studies do not address the whole problem; rather, a small set of classifiers is trained to choose among these. While this approach is important in that it allows the research community to develop better learning methods and evaluate them in a range of applications, it is worth realizing that an important stage is missing: in some situation, like in the Letter Recognition example, the small set of candidates is not fixed and it could be difficult to determine it.

Our SASSC method can be meant as a general *data driven* approach to the study of multiclass classifiers. It results in a sequential learning model that utilizes general purpose classifiers to sequentially restrict the number of competing candidate classes while maintaining the presence of the true class in the candidate set. In fact, in each iteration of the algorithm the sought-after classifier has to choose a single response class (or a small set of them) from among a large set of response classes. The method works by sequentially applying simple classifiers. In fact, the classifiers in the sequence are selected to be simple in the sense that they typically work only on subgroup of observations, where the detection of such subgroups is done according to an overall goodness of fit criteria. Simple classifiers are chosen so that the presence of the true class in the candidate set of classes is maintained: one of the option of the method allows for a Bayesian-like probabilistic assessment of such a presence. The order of the sequence is determined so as to maximize the rate of decreasing accuracy of the overall model, estimated through cross-validation.

Finally, the analysis of the Letter Recognition dataset has demonstrated that the SASSC algorithm can be applied by pursuing two complementary goals: 1) a *content-related* goal, resulting in the specification of a classification model that provides a good interpretation of the results without disregarding accuracy; 2) a *performance-related* goal, dealing with the development of a model which is effective in terms of predictive accuracy without neglecting interpretability.

Taking all of the above mentioned considerations into account, SASSC appears as a valuable alternative to evaluate whether a restricted number of independent classifiers improves the generalization error of a classification model.

References

ASUNCION, A., NEWMAN, D.J. (2008): *UCI Machine Learning Repository.* University of California, Irvine, School of Information and Computer Sciences. Available via FTP: http://mlearn.ics.uci.edu/MLRepository.html.

BENABDESLEM, K., BENNANI, Y. (2006): Dendogram-based SVM for Multi-Class Classification. *Journal of Computing and Information Technology - CIT 14, 283-289.*

BREIMAN, L.(2001): Random Forests. *Machine Learning 45, 5-32.*

BREIMAN, L. (1996): Bagging Predictors. *Machine Learning 24, 123-140.*

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J. (1984): *Classification and regression trees.* Wadsworth, Belmont (CA).

CUTZU, F. (2003): Polychotomous Classification with Pairwise Classifiers: A New Voting Principle. In: Windeatt, T., Roli, F. (Eds.): *Multiple Classifier System, Proceedings of the Fourth International Workshop MCS 2003.* Springer-Verlag, New York, 115–124.

DIETTERICH, T.G., BAKIRI, G. (1995): Solving multi-class learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research 2, 263-286.*

DUDA, R.O., HART, P.E., STORK, D.G. (2001): *Pattern classification.* John Wiley & Sons, New York.

EVEN-ZOHAR, Y., ROTH, D. (2001): A Sequential Model for Multi-Class Classification. In: Lee, L., Harman, D. (Eds.): *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing.* Available via FTP: www.cs.cornell.edu/home/llee/emnlp.html, 10–19.

FOGARTY, T. (1992): First Nearest Neighbor Classification on Frey and Slate's Letter Recognition Problem (Technical Note). *Machine Learning 9, 387-388.*

FREY, P.W., SLATE, D.J. (1991): Letter Recognition Using Holland-style Adaptive Classifiers. *Machine Learning 6, 161-182.*

HASTIE, T.J., FRIEDMAN, J., TIBSHIRANI, R.J. (2001): *The Elements of Statistical Learning.* Springer, New York.

HASTIE, T.J., TIBSHIRANI, R.J. (1998): Classification by pairwise coupling. *The Annals of Statistics 26(1), 451-478.*

LEE, D., SEUNG, H. (1997): Unsupervised learning by convex and conic coding. In: Mozer, M.C., Jordan, M.I., Petsche, T. (Eds.): *Advances in Neural Information Processing Systems.* MIT press, Cambridge (MA), 9, 515–521.

PRINZIE, A., VAN DEN POEL, D. (2005): Constrained optimization of datamining problems to improve model performance: a direct-marketing application. *Expert Systems with Applications 29(3), 630-640.*

R Development Core Team (2008): *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. Available via FTP: www.R-project.org.

SCHOLKOPF, B., SMOLA, A.H. (2001): *Learning with Kernels.* MIT press, Cambridge (MA).

WESTON, J., HERBICH, R. (2000): Adaptive margin support vector machines. In: Smola, A.J., Bartlett, P.L., Scholkopf, B., Schuurmans, D. (Eds.): *Advances in Large Margin Classifiers.* MIT press, Cambridge (MA), 281–295.

Possibilistic PLS Path Modeling: A New Approach to the Multigroup Comparison

Francesco Palumbo[1] and Rosaria Romano[2]

[1] Dipartimento di Istituzioni Economiche e Finanziarie, Universitá di Macerata
Via Crescimbeni 20, 62100 Macerata, Italy, *palumbo@unimc.it*
[2] Department of Food Science, University of Copenhagen
Rolighedsvej 30, Frederiksberg, Denmark, *rro@life.ku.dk*

Abstract. Structural equation models are reference techniques for measuring cause-effect relationships in complex systems. In many real cases observations are *a priori* grouped into homogeneous segments according to a specific characteristic, so that different models can be assessed for each segment. The present paper proposes to adopt an Euclidean metric based on the model parameters: the aim is to determine differences among models. However, estimated models assess the relation structures in different proportions, i.e. the residual component can vary with respect to the different models. In order to overcome this shortcoming, the present work proposes alternative models with fuzzy parameters.

Keywords: imprecise data, fuzzy regression, PLS-path modeling

1 Introduction

The study of a modern socio-economic system, and in wide sense the study of complex systems, requires to measure several structured relationships: evaluate relationships among and within two or more sets of variables. Structural Equation Models (SEM) (Bollen, (1989)) cope with the complex relations statistical analysis inside a complex system. SEM basic principles consist in assuming that there are some not directly measurable variables (*latent variables*) measured by means of a number of observable indicators (*manifest variables*).

In the complex multivariate relationships modelling framework, Partial Least Squares Path Modeling (PLS-PM) (Tenenhaus et al., (2005)) represents a statistical approach to SEM, with an increasing popularity in several areas. PLS-PM formulates the causality dependencies between latent variables in terms of linear conditional expectations. This approach privileges a prediction-oriented discovery process to the statistical testing of causal hypotheses.

In many real situations, e.g. customer satisfaction analysis and sensory data analysis, statistical units may belong to different *a priori* defined homogeneous groups, which are distinct according to one specific characteristic. Assuming that groups have the same structural connections, the same model

can be replicated and estimated, for each group. Evaluating the differences in the model parameters leads to understand behavioral differences in the groups: multigroup comparison.

The present paper presents a novel approach to multigroup comparison in the multivariate case: the technique is based on the comparison of (local) fuzzy parameterized models. Fuzzy PLS-PM (FPLS-PM) (Romano; 2007) combines *Fuzzy Possibilistic Regression* (FPR) (Tanaka et al.; 1982) and PLS-PM. Differently from crisp parameters-based models, structural and residual information are gathered together in fuzzy models. Exploiting this special property, the paper proposes to compare the local FPLS-PM. A suitable distance measure for interval data is introduced to compare models. Moreover, to get easier interpretation and more intuitive results, distances among the local models are graphically displayed by means of hierarchical trees.

The procedure is based on the following assumptions: 1) variables belong to disjoint sets and each set of variables is well represented by a corresponding latent variable; 2) the resulting latent structure is identical in all groups of units; 3) the structural relationships among the latent variables are represented by appropriate multiple regression models linking the adjacent latent variables.

The paper consists of the following sections: this introduction; some basic notations and definitions are provided in section (2); section (3) comprises two sub-sections that shortly introduce the Partial Least Squares Path-Modeling and the Fuzzy Possibilistic Regression Model (FPRM); section (4) shows the combining of Fuzzy approach and PLS-PM, which is used in section (5.1) for comparing different PLS-PM models. Last section summarizes most significant results coming from an application on a real dataset.

2 Notation and basic definitions

Let Y and $\{X\} \equiv \{X_1, X_2, \ldots, X_P\}$ be a quantitative dependent variable and a set of independent variables, respectively. Regression analysis studies the statistical dependence of Y with respect to $\{X\}$.
Given the generic model:

$$Y = f(X, \beta) + \varepsilon$$

the aim is to find the set of unknown parameters β so that $\hat{Y} = f(X, \hat{\beta})$ is a good prediction of Y. *(Multiple) Linear Regression Model* (MLRM) assumes that the dependent variable Y would be expressed as the weighted sum of the independent variables $\{X_1, X_2, \ldots, X_P\}$. Least Squares allow us to identify the unknown weights $\{\beta_1, \beta_2, \ldots, \beta_P\}$ according to the minimization of the sum of squared deviations: $\{\sum (Y - f(X, \hat{\beta}))^2\}$. The term ε indicates the deviation of Y from the model.
As the system complexity increases the MLRM may become ineffective because more dependencies must be taken into account to explain the involved

relationships. SEM and PLS-PM offer satisfactory solutions to take into account higher complexity degrees.
Formally, the variables $\{X\} = \{X_1, X_2, \ldots, X_P\}$ are divided into J disjoint sets (blocks) of manifest variables $\mathbf{X_j}$ $(j = 1, \ldots, J)$; each block related to a latent variable ξ_j. Variable blocks $\{X_1, X_2, \ldots, X_P\}$ are observed on N statistical units $(n = 1, \ldots, N)$ belonging to G *a priori* defined disjoint groups whose sizes are indicated by N_g, with $g = 1, \ldots, G$ and not necessarily assumed to be equal.
The resulting data structure is a multiple table where the general term is indicated by x^g_{np}, $g \in [1, \ldots, G]$ refers to the group, $p = 1, \ldots, P$ refers to the variable and $n = 1, \ldots, N$ refers to the statistical unit. Notice that $\sum_g N_g = N$.

3 Methodological framework

3.1 Partial Least Squares Path Modeling (PLS-PM)

PLS-PM approach aims to study the relationships among two or more sets of variables: each set is summarized by a latent variable; relationships among latent variables are defined through a system of interdependent equations based on simple and multiple regressions. Causal relationships among the latent variables define the *structural model*, while relations between latent variables and related manifest variables define the *measurement model*.
The *structural model* is formalized as follows:

$$\xi_j = \beta_{j0} + \sum_{\substack{j'=1 \\ j \neq j'}}^{J} \beta_{jj'} \xi_{j'} + \psi_j, \tag{1}$$

where ξ_j and $\xi_{j'}$ are adjacent latent variables, j assumes value in $[1, \ldots, J]$ and ψ_j represent the error term. An iterative procedure permits to estimate the outer weights (ω) for constructing the latent variable scores (ξ). The path coefficients (β) then come from a regular regression between the estimated latent variables. Differently from SEM, the estimation PM procedure solves blocks iteratively and separately. It is named *partial* because of that.

PLS-PM is a *soft* modeling approach to be preferred when the traditional assumptions related to the distributions, the measurement scale and the sample size are not satisfied.

3.2 Fuzzy possibilistic regression models

The basic concept of the fuzzy theory is the *fuzzy set* (Zadeh, (1965)). Formally, given the *set of objects* Ω, ω_1 as the generic element, a *fuzzy set* $\tilde{A}$ in Ω is defined as a set of ordered pairs:

$$\tilde{A} = \{(\omega_i, \mu_{\tilde{A}}(\omega i)) | \omega_i \in \Omega\} \tag{2}$$

where the value $\mu_{\tilde{A}}(\omega_i) = h_i$ expresses the *membership degree* for a generic element $\omega_i \in \Omega$. The larger the value of $\mu_{\tilde{A}}(\omega_i)$, the higher the grade of membership of ω_i in $\tilde{A}$. If the *membership function* is permitted to have only the values 0 and 1 then the *fuzzy* set is reduced to a classical *crisp* set. If a fuzzy set $A \subset \mathbb{R}$ and satisfies the following properties then it is a *fuzzy number*:

i) $\sup(\mu_A(\omega_i)) = 1 \longrightarrow$ the fuzzy set is normal;
ii) $\mu_A[\lambda\omega_i + (1-\lambda)\omega_{i'}] \geq \mu A(\omega_i) \wedge \mu_A(\omega_{i'}) \longrightarrow$ all $h-$cuts are convex and bounded.

Where $\{\omega_i, \omega_{i'}\} \in \mathbb{R}$, i and i' vary in $[1, \ldots, I]$ and $\lambda \in [0, 1]$.

The *triangular* fuzzy number is the most popular fuzzy number for its easy codification. An *interval value* is a special case of fuzzy number where $\mu_A(\omega_i) = 1$, $\forall i \in [1, \ldots, I]$ (Tanaka and Guo (1999)).

According to the above definitions, symmetrical triangular fuzzy numbers and interval values can be numerically represented by two values: *midpoint* (or *center*) and *spread* (or *radius*), alternatively $[\min, \max]$.

When a phenomenon under consideration has not stochastic variability but is uncertain in the fuzzy way, it is more natural to seek a fuzzy model for the given data.

In this work, attention is given to the Tanaka's possibilistic fuzzy regression (FPR). Specifically, to the FPR for crisp input and fuzzy parameters which is based on the following fuzzy linear model:

$$Y = \widetilde{\beta}_0 + \widetilde{\beta}_1 X_1 + \ldots + \widetilde{\beta}_p X_p, + \ldots + \widetilde{\beta}_P X_P, \tag{3}$$

where the coefficients are symmetric triangular fuzzy numbers denoted by $\widetilde{\beta}_p$. In terms of midpoint and spread $\widetilde{\beta}_p = (c_p; a_p)$.
Differently from statistical regression, the deviations between data and linear models are assumed to depend on the vagueness of the parameters and not on measurement errors. The basic idea of Tanaka's approach was to minimize the uncertainty of the estimates by minimizing the total spread of the fuzzy coefficients. Spread minimization must be pursued under the constraint of the inclusion of the whole given data set, which satisfies a degree of belief h $(0 < h < 1)$ defined by the decision maker. The above analysis leads to the following linear programming problem:

$$\underset{a_p}{\text{minimize}} \sum_{p=0}^{P} \left(\sum_{n=1}^{N} a_p |x_{np}| \right) \tag{4}$$

subject to the following constraints:

$$\sum_{p=0}^{P} c_p x_{np} + (1-h) \sum_{p=0}^{P} a_p |x_{np}| \geq y_n, \forall n = 1, \ldots, N;$$

$$\sum_{p=0}^{P} c_p x_{np} - (1-h)\sum_{p=0}^{P} a_p |x_{np}| \leq y_n, \forall n = 1, \ldots, N;$$

satisfying the following conditions: *i*) $a_p \geq 0$, *ii*) $c_p \in R$, *iii*) $x_{n0} = 1$.

There are no restrictive assumptions on the model. Increasing the h–coefficient expands the fuzzy intervals as well as increasing the confidence level in statistical regression expands the confidence interval width.

Kim et al. (1996) have carried out that fuzzy linear regression is more useful when the data set is too small to support statistical regression analysis and/or the aptness of the model is poor due to vague relationships among variables or a poor model specification.

Romano and Palumbo (2006), comparing the classical statistical linear regression model with the FPR model, have pointed out that fuzzy estimators are unbiased. Moreover, they have shown that FPR estimators are not affected by *quasi* multi-collinearity.

4 Fuzzy PLS path modeling

PLS-PM and FPR present many similar characteristics so that a combination of these two methodologies seems to be very appropriate. They are well suited methodologies for analyzing *phenomena* where human judgment is influential. For instance in *consumer analysis*, where consumers give their opinions on a certain number of products and/or services. In this framework such as in many other decision processes the major source of uncertainty is fuzziness rather than randomness. In addition, both PLS-PM and FPR are *soft modeling* approaches, i.e. sample size does not influence the quality of the estimators and there are no constraints on distributions and measurement scale. This connection implies the following two stage estimation procedure, where FPR joins PLS-PM in its final step allowing for a *fuzzy structural model* but a still crisp measurement model:

***stage* 1** : latent variable scores are estimated according to the PLS-PM estimation procedure;

***stage* 2** : FPR on the estimated latent variable scores is performed so that the following *fuzzy structural model* is obtained:

$$\xi_j = \widetilde{\beta}_{j0} + \sum_{j'} \widetilde{\beta}_{jj'} \xi_{j'} \tag{5}$$

where $\widetilde{\beta}_{jj'}$ refers to the generic fuzzy path coefficient and j and j' vary as described in section 3.1.

It is worth noticing that the *structural model* from this procedure is different with respect to the traditional *structural model* in section 3.1. Here path coefficients are fuzzy numbers and there is no error term, as a natural consequence of FPR: the error term is reflected in the model via fuzzy parameters.

5 Comparing models

The analysis of complex systems, characterized by particularly heterogeneous statistical populations, leads to split the whole population into more homogeneous groups or segments. Like in classical inferential problems, sampling from heterogeneous populations, the stratified sampling is preferred to the random sampling. The segments are based on some predetermined criteria such as geographic location, size or any demographic characteristic. It is important the segments are as heterogeneous as possible according to the predetermined criterion. Population segmentation leads to estimate the same model as many times as the segments identified in the target population. Several approaches have been proposed to compare the sub-populations. One of the main approaches consists in comparing the estimated parameters (Clogg et al.; 1995). However, estimated models assess the relation structures in different proportions; in fact the residual component can vary with respect to the different models. It is important to stress that comparing models in such a way could lead to biased results. Let us consider the simple linear regression analysis. Specifically, let us consider two models with equal parameters (slope and location). Such models should be considered statistically equivalent, according to the approach based on the parameters comparison. However, the models could have a different fit.
In the analysis of a statistical model one should always, in one way or another, take into account the goodness of fit, above all in comparing different models. The estimation of fuzzy parameters, instead of single-valued (crisp) parameters, permits us to gather both the structural and the residual information. In fact, FPR embeds the residual in the model via fuzzy parameters allowing a full comparison among the models.

5.1 Comparing PLS-PM models

In the specific framework of SEM, the model comparison problem is considered as a special case of *moderating effects*. Moderating effects (also called interaction effects) arise when some variables influence a direct effect between the latent variables inside the model. In particular if the *moderator variable* is categorical, it becomes a grouping variable involving group comparisons, i.e. comparisons of model estimates for different groups of observations. Once the observations are grouped according to the moderator variable, the strategy is then to estimate local models with direct effects for each group and look for differences in path coefficients across groups. To this aim, non-parametric

approaches may be used to test for different path coefficients among groups (Chin and Dibbern; 2007).
If there is no difference between the parameters from the different models then there is no reason for considering local models. In other words, a global model is effective for the whole population. If there exist differences between parameters, these are evaluated as differences between local models.
It should be noticed how the moderating effects only concern the *structural models*. That is equivalent to assuming that the differences among *measurement models* are not significant.
Under this hypothesis, in this work, model differences are gathered comparing the *fuzzy structural parameters* in terms of distances. The strategy consists of three basic steps:

a) estimate local fuzzy structural models for each group

$$\xi_j^g = \tilde{\beta}_{j0}^g + \sum_{j'} \tilde{\beta}_{jj'}^g \xi_{j'}^g$$

b) gather model differences comparing the related fuzzy path coefficients, describing the local models

	$\tilde{\beta}_{j,1}$	...	$\tilde{\beta}_{J,1}$	...	$\tilde{\beta}_{J,j}$
model[1]	$\tilde{b}_{j,1}$	...	$\tilde{b}_{J,1}$	...	$\tilde{b}_{J,j}$
...	...		...		...
model[g]	$\tilde{b}_{j,1}$	...	$\tilde{b}_{J,1}$	...	$\tilde{b}_{J,j}$
...	...		...		...
model[G]	$\tilde{b}_{j,1}$	...	$\tilde{b}_{J,1}$	...	$\tilde{b}_{J,j}$

Table 1. Data matrix.

c) displaying distances for fuzzy/interval data by hierarchical trees or pyramids.

The G estimated fuzzy structural models are characterized by fuzzy path coefficients. That means there are no residual terms, because in the fuzzy model the error terms are embedded in the parameters themselves. This peculiarity confers to the G fuzzy structural models the same explicative power, making the comparison, that is based on the estimated fuzzy path coefficients, meaningful. The hierarchical classification (step c) is used to easily visualize similarities/differences between models. Alternative methods like Multidimensional Scaling can be used for visualizing results.

5.2 PLS-PM distances visualization by hierarchical pyramids

This proposal exploits a fuzzy classification algorithm for interval data to compare the groups with respect to the identified PLS-PM fuzzy structural

parameters. In particular the paper refers to the HIPYR hierarchical classification algorithm (Brito (2000)). The procedure is implemented in the SODAS© software (Symbolic Official Data Analysis System, rel. 2.5) and is designed to cluster symbolic datasets (Bock and Diday (2000)). It is worth noticing that fuzzy data and interval data can also be defined as special cases of symbolic data, only characterized by quantitative continuous variables.
HIPYR displays hierarchical tress or pyramids starting from a distance matrix computed on an interval data matrix. HIPYR clustering procedure determines the $G(G-1)/2$ distances between models by the Hausdorff metric in $\mathbb{R}^p$. This allows us to appreciate the differences in the $\mathbb{R}^p$ parameters space.

6 Example

A typical application for PLS-PM is the estimation of *customer satisfaction*. Within this framework, a widely adopted model is the one specified for the European Customer Satisfaction Index (ECSI) (Tenenhaus et al. (2005)).
ECSI model allows us to estimate latent variable scores from their respective manifest variables and to build individual indexes of satisfaction.
The global model contains the following latent variables: *perceived quality*, *expectation*, *perceived value*, *satisfaction index*, *image*, *loyalty* and *complaints*. In particular, *customer satisfaction* is explained by the drivers *perceived quality*, *expectation*, *perceived value* and *image*.
The present application has been performed on a data set used to estimate the customer satisfaction of a service industry. Data contains 23 variables observed on 366 units. Variables are grouped in 6 blocks: *perceived quality* (7 manifest variables), *expectations* (4 manifest variables), *perceived value* (3 manifest variables), *satisfaction index* (3 manifest variables), *image* (3 manifest variables) and *loyalty* (3 manifest variables). According to the moderator variable *sector of activity*, statistical units are groped into 8 classes (labels are indicated in round brackets): hospital (Hosp), health authority (HeAu), school (Scho), university (Univ), local administration (LoAd), Red Cross (RCro), social security (SoSe), public administration (PaAd). Variable names cannot be revealed because of confidential constraints on the data.
First, results from the global Fuzzy PLS-PM on the whole dataset are presented. It is important to remember how in such a procedure the FPR is introduced just in the last step of the algorithm. In other words, as this approach provides a *fuzzy structural model* but a still *crisp measurement model*, all the results of the algorithm are the same as in the traditional PLS-PM except those related to the *inner model*. The five fuzzy structural equations may be written as follows:

$$\text{expectation} = \tilde{\beta}_{21}\text{image}$$

$$\begin{aligned}
\text{perceived quality} &= \tilde{\beta}_{32}\text{customer expectation} \\
\text{perceived value} &= \tilde{\beta}_{42}\text{customer expectation} + \tilde{\beta}_{43}\text{perceived quality} \\
\text{loyalty} &= \tilde{\beta}_{61}\text{image} + \tilde{\beta}_{65}\text{CSI} \\
\text{satisfaction} &= \tilde{\beta}_{51}\text{customer expectation} + \tilde{\beta}_{52}\text{perceived quality} \\
&\quad + \tilde{\beta}_{53}\text{perceived value} + \tilde{\beta}_{54}\text{image}
\end{aligned}$$

Table (2) shows the path coefficients of the classical PLS-PM and the fuzzy path coefficients of the fuzzy PLS-PM, with the *possibility level* $\alpha = 0.5$. In order to compare results from classical PLS-PM with results from the

Latent Variable	*path coeff.*	*fuzzy path coeff.* center, spread	*fuzzy path coeff.* min, max
expectation	0.5077	[3.8923, 3.6378]	[0.2544, 7.5301]
perceived quality	0.5269	[2.7954, 2.4719]	[0.3235, 5.2673]
perceived value	0.2987	[0.1833, 0.1350]	[0.0483, 0.3183]
	0.6067	[0.6604, 0.4797]	[0.1807, 1.1401]
loyalty	0.3570	[0.7374, 0.7191]	[0.0183, 1.4565]
	0.5248	[0.3872, 0.1400]	[0.2472, 0.5273]
satisfaction	0.0829	[0.0500, 0.1772]	[−0.127, 0.227]
	0.1940	[0.2582, 0.0758]	[0.1824, 0.3339]
	0.3687	[0.0203, 0.0071]	[0.0132, 0.0273]
	0.3273	[0.3895, 0.2220]	[0.1675, 0.6115]

Table 2. Inner model weights.

fuzzy PLS-PM, some outputs of the classical approach are shown in table (3): the R^2 of each structural equation, the bootstrap confidence intervals and the value of test-statistic t for each path coefficient. Comparing results in table (2) and (3) it can be seen that fuzzy path coefficients with wider spread

Latent Variable	R^2	*confidence interval*	*T-Statistic*
expectation	0.2578	[0.4227, 0.5827]	12.0471
perceived quality	0.2777	[0.4607, 0.6084]	13.5374
perceived value	0.6484	[0.2309, 0.3768]	7.9238
		[0.5371, 0.6594]	20.9333
loyalty	0.6951	[0.2620, 0.4590]	8.4573
		[0.4176, 0.6159]	11.5011
satisfaction	0.7479	[0.0118, 0.1594]	2.27795
		[0.1161, 0.2579]	5.25735
		[0.2405, 0.4874]	5.80815
		[0.2319, 0.4318]	6.91954

Table 3. PLS-PM output.

correspond to structural equations with lower R^2. This is a natural consequence of the FPR, which is extremely sensitive to outliers. Furthermore, the relation between the *expectation* and *satisfaction* seems to be not significant in both approaches. In fact, in the fuzzy approach the corresponding fuzzy interval embed the value 0, whereas in the classical approach the t-value is close to 2. The t-values are computed on the basis of the percentiles of the bootstrap distribution: the relation between two latent variables is considered statistically significant if the t-value of the corresponding path coefficient is more or less higher than 2.

The multi-group structure data suggests significantly different models: 8 local models have been estimated according to the procedure in section (4).

The eight estimated models are compared on the basis of their fuzzy path coefficients. In other words, the distances among the different local models are considered for a pyramidal classification procedure.

Results show two fuzzy groups:

Cluster 1 [Scho, Hosp, SoSe, RCro, HeAu]
Cluster 2 [Hosp, SoSe, RCro, HeAu, Univ, LoAd, PaAd]

The overlapping Hosp, SoSe, RCro, HeAu appears in both clusters.
It is quite evident that Cluster 1 is mainly characterized by public bodies having autonomy of expenditure. On the contrary Cluster 2 is mainly characterized by Local and Central administrations. Notice that these two later groups only appear in the Cluster 2. In order to compare results with classical multi-group analysis in PLS-PM, the multi-group t test (for any number of groups) implemented in the XL-STAT-PLSPM software has been applied. This test uses the estimates obtained from the bootstrap sampling in a parametric sense via t-tests for the difference in path coefficients between groups.

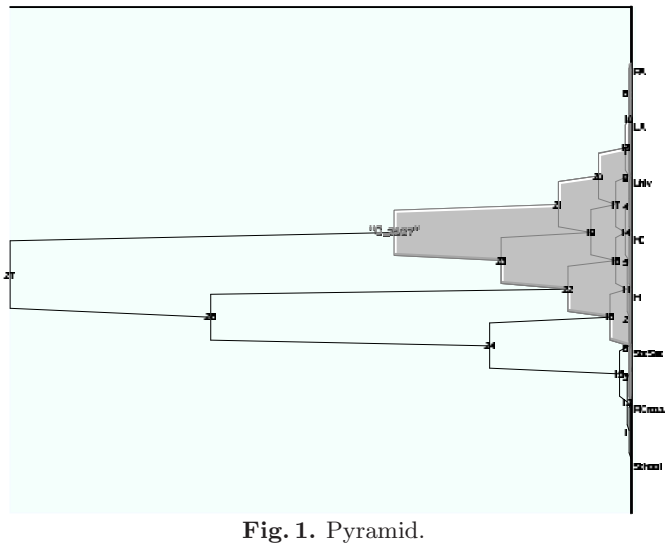

Fig. 1. Pyramid.

Results from the multi-group t-test 6 are coherent with the proposed analytical approach. In fact, SoSe shows a significant different model from the group {PuAd, LoAd, Univ, HeAu, Hosp}. In addition, SoSe present a similar model to Scho, which is different from Rcro. In other words it is confirmed the presence of two groups with similar model.

Groups	**Difference**	**p-value**	**alpha**	**Significant**
Path coefficient (E − > PV)				
HeAu vs PuAd	0,278	0,033	0,050	Yes
Path coefficient (PQ − > PV)				
Hosp vs SoSe	0,364	0,044	0,050	Yes
Path coefficient (I − > S)				
LoAd vs SoSe	0,495	0,041	0,050	Yes
Scho vs LoAd	0,360	0,017	0,050	Yes
Scho vs Univ	0,356	0,043	0,050	Yes
HeAu vs SoSe	0,568	0,029	0,050	Yes
HeAu vs Scho	0,433	0,008	0,050	Yes
Hosp vs HeAu	0,381	0,045	0,050	Yes
Path coefficient (E − > S)				
Rcro vs PuAd	0,576	0,002	0,050	Yes
Rcro vs SoSe	0,496	0,050	0,050	Yes
LoAd vs Rcro	0,451	0,011	0,050	Yes
Univ vs Rcro	0,515	0,027	0,050	Yes
Scho vs Rcro	0,401	0,021	0,050	Yes
HeAu vs Rcro	0,470	0,037	0,050	Yes
Hosp vs PuAd	0,272	0,043	0,050	Yes
Path coefficient (PQ − > S)				
HeAu vs LoAd	0,323	0,040	0,050	Yes
Path coefficient (PV − > S)				
SoSe vs PuAd	0,627	0,012	0,050	Yes
LoAd vs SoSe	0,658	0,008	0,050	Yes
Univ vs SoSe	0,834	0,009	0,050	Yes
Scho vs PuAd	0,271	0,040	0,050	Yes
Scho vs LoAd	0,302	0,024	0,050	Yes
Scho vs Univ	0,478	0,007	0,050	Yes
HeAu vs PuAd	0,434	0,023	0,050	Yes
HeAu vs SoSe	1,061	0,005	0,050	Yes
HeAu vs Rcro	0,771	0,026	0,050	Yes
HeAu vs LoAd	0,402	0,036	0,050	Yes
HeAu vs Scho	0,705	0,001	0,050	Yes
Hosp vs HeAu	0,653	0,026	0,050	Yes
Path coefficient (I − > L)				
SoSe vs PuAd	0,554	0,013	0,050	Yes
LoAd vs SoSe	0,688	0,008	0,050	Yes
HeAu vs SoSe	0,571	0,040	0,050	Yes
Hosp vs SoSe	0,764	0,023	0,050	Yes
Path coefficient (S − > l)				
SoSe vs PuAd	0,523	0,023	0,050	Yes
LoAd vs SoSe	0,695	0,006	0,050	Yes
Univ vs SoSe	0,638	0,041	0,050	Yes
HeAu vs SoSe	0,827	0,019	0,050	Yes
Hosp vs SoSe	0,639	0,035	0,050	Yes

Table 4. Multi-group t-test results.

References

BOCK, H. H. and DIDAY, E. (2000): *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg.

BOLLEN, K. A. (1989): *Structural equations with latent variables*. Wiley, New York.

BRITO, P. (2000): Hierarchical and Pyramidal Clustering with Complete Symbolic Objects. In: *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer, Heidelberg, 312–341.

CHIN, W. W. and DIBBERN, J. (2007): A permutation based procedure for multi-group PLS analysis: Results of tests of differences on simulated data and a cross cultural analysis of the sourcing of information system services between Germany and the USA. In: V. Esposito Vinzi, W. W. Chin, J. Henseler, and H. Wang (Eds.): *Handbook of Partial Least Squares - concepts, methods and applications*. Springer.

CLOGG, C. C., PETKOVA, E. and HARITOU, A. (1995): Statistical methods for comparing regression coefficients between models. *The American Journal of Sociology 100 (5), 1261–1293*.

KIM, K. J., MOSKOWITZ, H. and KOKSALAN, D. (1996): Fuzzy versus statistical linear regression. *European J. Oper. Res. 92, 417–434*.

ROMANO, R. (2007): *Fuzzy regression and PLS path modeling: a combined two stage approach for multi-block analysis*. Phd thesis in Statistics, Dep. of Mathematics and Statistics, University Federico II, Naples.

ROMANO, R. and PALUMBO, F. (2006): Fuzzy regression and least squares regression: the relationship between two different fitting criteria. In: *Abstracts of the SIS2006 Conference*. Vol. 2. CLUEP, Torino, 693–696.

TANAKA, H. and GUO, P. (1999): *Possibilistic Data Analysis for Operations Research*. Physica-Verlag, Wurzburg.

TANAKA, H., UEJIMA, S. and ASAI, K. (1982): Linear regression analysis with fuzzy model. *IEEE Transactions Systems Man Cybernet 12, 903–907*.

TENENHAUS, M., ESPOSITO VINZI, V., CHATELIN, Y-M. and LAURO. C. (2005): PLS path modeling. *Comp. Stat. and Data Analysis 48, 159-205*.

ZADEH, L. (1965): Fuzzy sets. *Information and Control 8, 338–353*.

Models for Understanding Versus Models for Prediction

Gilbert Saporta

Chaire de statistique appliquée & CEDRIC, CNAM
292 rue Saint Martin, Paris, France, *gilbert.saporta@cnam.fr*

Abstract. According to a standard point of view, statistical modelling consists in establishing a parsimonious representation of a random phenomenon, generally based upon the knowledge of an expert of the application field: the aim of a model is to provide a better understanding of data and of the underlying mechanism which have produced it. On the other hand, Data Mining and KDD deal with predictive modelling: models are merely algorithms and the quality of a model is assessed by its performance for predicting new observations. In this communication, we develop some general considerations about both aspects of modelling.

Keywords: model choice, data mining, complexity, predictive modelling

1 Models for understanding

A statistical model consists usually in the formulation of a parametric formula for the distribution of a (multidimensional) random variable. When the interest lies in a particular response, the usual form is $y = f(x;\theta) + \varepsilon$. When the model is completely specified by an expert (economist, biologist, etc.) the statistical work consists in estimating the unknown parameters, and (or) refute the model according to a goodness of fit test. If the model is rejected, the expert should think of an other one.
Generally a model should be simple, and parameters should be interpretable in terms of the application field : elasticity, odds-ratio, etc. The need for interpretation explains why for instance logistic regression is preferred to discriminant analysis by biostatisticians and econometricians, the coefficients being uniquely estimated and having the meaning of log-odds.
The purpose of a model is to give insights in the nature of a stochastic phenomenon, not necessarily to give accurate predictions. This may be viewed as a paradox, since *eg* in natural sciences, a good model must give good predictions, otherwise the model is replaced by an other one. It is due to the importance of the random term ε . In epidemiology for instance, it is more important to find risk indicators than having an accurate individual prediction of getting some disease.

1.1 Model estimation

Maximum likelihood estimation is by far the standard technique : the likelihood principle which comes back to R.A.Fisher says that among several values of a parameter θ, one must choose the one which maximizes the probability density function which is equal for iid observations to

$$L\left(x_1,..,x_n;\theta\right)=\prod_{i=1}^{n}f\left(x_i;\theta\right)$$

considered as a function of θ.
Advantages of ML estimation are in availability of asymptotic standard errors, as well as tests. The extensive use of ML estimation is recent: the first technique for estimating the logistic regression model proposed by Berkson in the 40's was the minimum chi-squared, see Berkson (1980).
Least squares estimation is often more robust and need less assumption (both are related) and is of common use, especially in exploratory analysis, including PLS structural equation modelling.

1.2 Model choice

Choosing between several models occurrs when the "expert" hesitates between several formulations. Statistics may help to choose among several models using some parsimony principle. This is conform to Occam's razor, which is often considered as a scientific principle against unnecessary hypotheses . The major use of model selection is for variable selection including interaction selection.
A considerable amount of literature has been devoted to model selection by minimizing penalized likelihood criteria like AIC, BIC, see Burnham and Anderson (2000).

$$AIC=-2\ln\left(L(\hat{\theta})\right)+2k\quad BIC=-2\ln\left(L(\hat{\theta})\right)+\ln(n)k \tag{1}$$

BIC favourizes more parsimonious models than AIC due to its penalization. AIC, but not BIC, is biassed in the following sense: if the true model belongs to the family M_i, the probability that AIC chooses the true model does not tend to one when the number of observations goes to infinity.
AIC and BIC have similar formulas but originates from different theories and there is no rationale to use simultaneously AIC and BIC: AIC is an approximation of the Kullback-Leibler divergence between the true model and the estimated one, while BIC comes from a bayesian choice based on the maximisation of the posterior probability of the model, given the data.

1.3 Some limitations

Even if we knew the *"true"* model, parameter estimation could be a difficult task when the number of cases is low. For example in a multiple regression model, this can lead to severe multicollinearity. If we want to estimate all parameters, without discarding variables (and we should not discard variables if we believe in our model), it is necessary to put some constraints or in other words to do some regularization. Ridge regression which is a direct application of Tikhonov regularisation is a well known remedy to multicollinearity. Projecting onto a lower dimension space is another kind of regularization and includes principal components regression as well as PLS regression.
Bayesian statistics provides an elegant solution: it balances the lack of observations by using prior information. For the normal regression model with normal priors on the parameters, bayesian estimation comes down to ridge regression and provides an enlightening interpretation of this technique.

Model choice by penalized likelihood suffers from practical limitations: penalized likelihood cannot be applied to ridge or PLS regression, since there may be no simple likelihood nor a simple number of parameters: what is eg the right number of parameters for a ridge regression? There are still p parameters but since they are constrained by a condition like $\|\beta\|^2 < k$, we need an equivalent number of parameters less than p depending on k but the exact formula is unknown. No need to say that penalized likelihood is hard to apply to choose a decision tree, or between a decision tree and a logistic regression. Let us also remark that the underlying hypothesis for BIC of having a uniform prior on models is not very realistic.

The ambition of finding the *"true"* model belonging to the family of distributions is questionable and we must remind of the famous dictum from George Box (Box and Draper,1987, p.424): *"Essentially, all models are wrong, but some are useful."* This is especially true for very large data sets where no simple parsimonious model can fit to the data: it is well known that significance or goodness of fit tests always reject any precise null hypothesis when one has millions of observations: a correlation of 0.01 will be considered as significatively different from zero, but the point of interest is in the strength of the relationship, not in its existence (assuming that a correlation different from zero is a proof of existence...)

1.4 Models for exploration

Most of what have been said before is about *regression* in the broad sense of relating a response to some inputs. There is a slightly different use of models, applied to some kind of exploratory problems.
Let us suppose that we analyze a sample drawn *iid* from a population defined by some model, then it is possible to do some inference for the outputs of

the analysis: confidence intervals for the eigenvalues of a PCA, confidence regions for the positions of points in multidimensional scaling, PCA, or Correspondence Analysis etc. The model is here an help to interpret and provide additional knowledge to what can be seen from graphical displays of exploratory analysis.
However the validity of the model is is often dubbious for numerical variables. The standard model for inference in PCA is the multivariate normal, as one can find in any textbook, and for most data sets this model is wrong : there is here a contradiction in using both an exploratory technique, which has for goal to reveal the underlying structure of frequently heterogenous data and a single simple model. The situation is more comfortable for categorical data where a multinomial scheme is realistic. For instance in contingency tables analysis it is true, provided that the sampling scheme is simply random, that the joint distribution of the frequencies n_{ij}, is the multinomial $M(n; p_{ij})$ where the p_{ij} are unknown parameters. This lead to exact results (however difficult) for the distribution of eigenvalues in correspondence analysis. This may be extended to the case where the margins of one variable are fixed (stratified sampling).
Many experiences have proved that resampling, eg bootstrap, provides more reliable results than using unrealistic models, see Hatabian and Saporta (1986) or Lebart (2006). Let us however mention the neglected fixed-effect model for PCA (Besse et al. , 1988) which is less demanding in its hypotheses than the multivariate normal and lead to inferences on dimensionality and on the displays.

2 Models for prediction

In data mining applications, a model is merely an algorithm, coming more often from the data than from a theory. The focus here is not on an accurate estimation of the parameters, or on the adequacy of a model on past observations but on the predictive ability, ie the capacity of making good predictions for new observations : forecasting differs from fitting.
The "black-box model" (Vapnik, 2006, p.416) illustrates the differences with the previous conception of a model, while keeping the same mathematical formulation $y = f(x; \theta) + \varepsilon$. Statistical modelling (understanding data) look for a parsimonious function $f(x; \theta)$ belonging to a prespecified set. On the other hand, in predictive modelling, the aim is not to approximate the true function but to get a function which gives as accurate predictions as the model, since it is a stochastic one. The question is not to discover the hidden mechanism but to perform as well.
In many operational applications, like in Customer Relationship Management or pattern recognition, understanding the phenomenon would be a too complex and vain task: a banker does not need a theory for predicting if a loan will at risk or not, but only a good score function. In predictive inference,

models could be very complex like multilayer perceptrons or non-linear SVM, and we have the paradox that a good prediction does not need a deep understanding of what is observed. This may not be confused with the readability of a model: a decision tree is very simple to use for the end-user but is not a model of the hidden mechanism producing the data.

2.1 Risk minimization

Let L be a loss fonction and R=$E(L)$ its expectation, called here "risk".

$$R = E(L) = \int L(z, \theta) dP(z) \tag{2}$$

The risk is the average loss for new observations. The ideal would be to choose the model among some family of models in order to minimize R but it is an impossible task, since we do not know the true distribution P. Choosing the model which minimizes the empirical risk (ie the risk on observed data, or learning set)

$$R_{emp} = \frac{1}{n} \sum_{i=1}^{n} L(y_i; f(x_i; \theta)) \tag{3}$$

usually leads to overfitting.
For binary classification where one chooses as loss function the misclassification rate, Vapnik's inequality gives an upper bound relying on the VC-dimension h :

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(ln\left(2n/h\right)+1\right) - \ln\left(\alpha/4\right)}{n}} \tag{4}$$

Based on the upper bound of R, the structured risk minization principle or SRM provides a model choice methodology different from penalized likelihood, since no distributional assumptions are necessary. Given a family of models, the principle is (for fixed n) to choose the model which minimizes the upper bound : this realizes a trade-off between the fit and the generalization capacity.
This inequality proves that (provided h is finite) one may increase the complexity of a family of models (eg in a simple case increase the degree of polynomials) when the number of learning cases increases, since it is the ratio h/n which is of interest. This shows a strong difference between SRM and model choice based on BIC, since the penalization in BIC increases with n and tends to choose simpler models for large n. Devroye et al. (1996) proved that SRM is strongly universally consistent.

2.2 AUC-like measures of efficiency

Error rate estimation corresponds to the case where one applies a strict decision rule and depend strongly on prior probabilities and on group frequencies.

But in many applications one just needs a score S which is a rating of the risk to be a member of one group, and any monotonic increasing transformation of S is also a score like

$$S' = \frac{\exp(S)}{1+\exp(S)}$$

Usual scores are obtained with linear classifiers (Fisher's discriminant analysis, logistic regression) but since a probability is also a score (ranging from 0 to 1), almost any technique, even decision trees gives a score. SVM classifier also provide scores.

The ROC curve is a popular measure of efficiency which synthesizes the performance of a score for any threshold s such that if $S(x) > s$ then x is classified in group 1. Using s as a parameter, the ROC curve links the true positive rate to the false positive rate. The true positive rate (or specificity) is the probability of being classified in G_1 for a member of G_1 :$P(S > s|G_1)$. The false positive rate (or 1- sensitivity) is the probability of being wrongly classified to $G_1 : P(S > s|G_2)$. In other words, the ROC curve links the power of the procedure $1-\beta$ to α the probability of error of first kind. ROC curve is invariant with respect to increasing transformations of S.

Since the ideal curve is the one which sticks to the edges of the unit square, the most popular measure is given by the area under the ROC curve (AUC); another measure is the so-called Gini index which is equal to twice the area between the ROC curve and the diagonal: $Gini = 2AUC-1$. Since theoretical AUC is equal to the probability of concordance : $AUC = P(X_1 > X_2)$ when one draws at random two observations independently from both groups, AUC reduces to an old measure of nonparametric comparison: Mann-Whitney's U statistic.

ROC curve and AUC are extensively used in the banking industry to assess the quality of the credit risk rating system and are recommanded by the Basel Committee on Banking Supervision (see BCBS 2005).

Model choice using AUC should of course not be based on the learning sample. Inequalities similar to (4) , may be obtained for AUC but are not very useful in practice. Moreover ROC curve and AUC do not take into account some elements of interest in business applications like the error costs and the fact that very often the two subpopulations are not balanced at all.

2.3 Empirical model choice

Even if Vapnik's inequality is not directly applicable, for it is often difficult to evalute the VC dimension, SRM theory gives a way to handle methods where penalized likelihood is not applicable. One important idea is that one has to realize a trade-off between the fit and the robustness of a model.

An empirical way of choosing a model in the spirit of Statistical Learning Theory is the following (Hastie et al. 2001): Split the available data into 3 parts: the first set (training) is used to fit the various models of a family, the

second set (validation set) is used to estimate the prediction error of each previously estimated model and choose the best one, the last set (test set) is reserved to assess the generalization error rate of the best model. This last set is necessary, because using repeatedly the validation step is itself a learning step.
However split only once the data set into 3 parts is not enough, and may lead to unexpected sampling variations, see Niang and Saporta (2007). In order to avoid too specific patterns, all this process should be repeated a number of times to get mean values and standard errors. For measuring the prediction error in regression, Borra and Di Ciaccio (2007) compared several resampling technique including bootstrap and .632 bootstrap; they showed by simulation that a resampled 10-fold cross-validation technique outperformed other estimators. Since Fisher's supervised classification in 2 classes is a special case of the linear model, the latter results may be also valid for discrimination.

3 Conclusions

Two very different conceptions correspond to the same name of "model", and this may be a cause of misunderstanding. As Cherkassky and Mulier (1998) wrote: *"Too many people use different terminology to solve the same problem; even more people use the same terminology to address completely different issues"*. Models for understanding data correspond to the part of statistics considered as an auxiliary of science. Models for prediction belong to the other face of statistics as an help for decision. There are more job opportunities for graduate students in predictive modelling but also more competitors coming from other disciplines.
But one may question this opposition between science and action : when a technique gives really good predictions, it is also an improvement of the knowledge we have on data. Predictive modelling belongs to empiricism which is itself a theory of knowledge.

References

BCBS (2005): Studies on the Validation of Internal Rating Systems, *Basel Committee on Banking Supervision, Bank of International Settlements*, *http* : *//www.bis.org/publ/bcbs_wp*14.*htm*

BERKSON, J. (1980): Minimum chi-square, not maximum likelihood! *Annals of Mathematical Statistics* 8, 457-487.

BESSE, P., CAUSSINUS, H., FERRÉ, L. and FINE, J. (1988): Principal Components Analysis and Optimization of Graphical Displays, *Statistics*, 19, 301-312.

BORRA, S. and Di CIACCIO, A.(2007): Measuring the prediction error. A comparison of cross-validation, bootstrap and hold-out methods, in Ferreira, C., Lauro, C., Saporta, G. and Souto de Miranda, M. (eds), *Proceedings IASC 07, Aveiro, Portugal*

BOX, G.E.P. and DRAPER, N.R. (1987): *Empirical Model-Building and Response Surfaces*, Wiley

BURNHAM, K.P. and ANDERSON, D.R. (2000): *Model Selection and Inference*, Springer

CHERKASSKY, V. and MULIER, F. (1998): *Learning from data* , Wiley

DEVROYE, L., GYÖRFI L. and LUGOSI, G. (1996): *A Probabilistic Theory of Pattern Recognition*, Springer

HAND, D.J. (2000): Methodological issues in data mining, in J.G.Bethlehem and P.G.M. van der Heijden (eds), *Compstat 2000 : Proceedings in Computational Statistics*, Physica-Verlag, 77-85

HASTIE, T., TIBSHIRANI, F. and FRIEDMAN, J. (2001): *Elements of Statistical Learning*, Springer

HATABIAN, G. and SAPORTA, G. (1986): Régions de confiance en analyse factorielle, in Diday E. (ed) *Data Analysis and Informatics IV*, North-Holland, 499-508

LEBART, L. (2006): Validation Techniques in Multiple Correspondence Analysis, in Greenacre M. and Blasius J. (eds) *Multiple Correspondence Analysis and related techniques*, Chapman and Hall/CRC, 179-196

NIANG, N. and SAPORTA, G. (2007): Resampling ROC curves, in Ferreira, C., Lauro, C., Saporta, G. and Souto de Miranda, M. (eds), *Proceedings IASC 07, Aveiro, Portugal*

VAPNIK, V. (2006): *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer

Posterior Prediction Modelling of Optimal Trees

Roberta Siciliano, Massimo Aria, Antonio D'Ambrosio

Department of Mathematics and Statistics, University of Naples Federico II
Monte Sant'Angelo, Via Cinthia, I-80126 Naples, Italy,
{*roberta, aria, antdambr*}*@unina.it*

Abstract. The framework of this paper is classification and regression trees, also known as tree-based methods, binary segmentation, tree partitioning, decision trees. Trees can be fruitfully used either to explore and understand the dependence relationship between the response variable and a set of predictors or to assign the response class or value for new objects on which only the measurements of predictors are known. Since the introduction of two-stage splitting procedure in 1992, the research unit in Naples has been introducing several contributions in this field, one of the main issues is combining tree partitioning with statistical models. This paper will provide a new idea of knowledge extraction using trees and models. It will deal with the trade off between the interpretability of the tree structure (i.e., exploratory trees) and the accuracy of the decision tree model (i.e., decision tree-based rules). Prospective and retrospective view of using models and trees will be discussed. In particular, we will introduce a tree-based methodology that grows an optimal tree structure with the posterior prediction modelling to be used as decision rule for new objects. The general methodology will be presented and a special case will be described in details. An application on a real world data set will be finally shown.

Keywords: tree-based model, model prediction, optimal partitioning

1 Introduction

This paper is about tree-based methods. Trees result from a supervised learning approach where a response or output variable is predicted by more predictors or inputs. Two main directions of research can be distinguished according to the aim of the analysis, namely exploratory and confirmatory (Siciliano, 1998).

1.1 Exploratory trees

In the exploratory context, binary segmentation can be understood as a recursive partitioning of objects due to some splitting variables derived from available predictors such to obtain internally homogeneous and externally heterogeneous subgroups with respect to a target or response variable. The final result is an exploratory tree, either classification or regression tree, that

can be analysed to understand the dependence relationship between the response variable (of categorical or numerical type) and the predictors (of any type). At any internal node of the tree, a predictor generates the splitting variable (i.e., dummy variable) to discriminate the objects falling into the left subnode from those falling into the right subnode. Usually, the best splitting variable is chosen among all possible splits generated by the predictors such to maximize the decrease of impurity of the response variable within the two subnodes (Breiman et al. 1984), where the impurity is a measure of deviance or variation for numerical responses (i.e., regression tree) and a measure of hetereogeneity or entropy for categorical responses (i.e., classification tree). Two-stage splitting criterion finds first the best predictor(s) and then the best split of the best predictor(s) on the basis of predictability measures using in cross-classifications (Mola and Siciliano, 1992, 1996). Fast algorithm allows to find the best split without trying out all possible splits, thus using a subset of best predictors (Mola and Siciliano, 1997; Siciliano et al. 1998). Terminal nodes include disjoint and homogeneous subgroups of objects, defining a partition of the starting group of objects with respect to the response variable.

1.2 Decision tree-based rules

In the confirmatory context, the aim is to predict a response class/value of new objects in which only the measurements of predictors are known. One approach is to define a set of nested pruned subtrees by removing at turn the most unreliable branches and then select the most accurate subtree for new cases (Breiman et al., 1984; Mola and Siciliano, 1994). As measure of accuracy the error rate for classification problem and the mean squared error for regression problem are evaluated on independent test sample as well as according to a cross-validated sampling procedure. Model selection procedures can be also considered to select the best among alternative decision trees. Another approach consists in ensemble methods, such as boosting and bagging algorithms. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a vote of their predictions. The most popular ensemble methods work by manipulating the training examples through re-sampling methods. Bagging (Breiman, 1996) uses V-fold bootstrap replication, whereas Boosting (Freund and Schapire, 1997) uses V-fold weighted-bootstrap replications in the sense that the probability of each object to be included in the subsequent sample increases if the same object is misclassified by the learning algorithm. Both algorithms aggregate the object decisions by voting. This approach improves accuracy of decision tree-based rules but it does not yield to a prediction tree structure.

1.3 This paper

This paper provides a one-step methodology for both exploratory and confirmatory analysis. Main idea is to obtain a prediction tree structure based on the optimal partitioning of known objects and a posterior prediction model for unknown objects. The proposed approach can be understood as a retrospective view of using trees and models, that is an alternative to the prospective view where models are considered either in the splitting criterion definition (Siciliano and Mola, 1994) or in the terminal node description (Conversano et al. 2001). The concepts of optimal partitioning and posterior prediction modeling will be defined. In the following, after recalling some previous and related work, we will describe the general methodology and a special case when using the logit-linear modelling. An application on a dataset available in Machine Learning Repository will be shown for completeness.

2 Key issues of Tree Harvest

Tree Harvest is the name of our specialized software for tree-based methods proposed by our research group in Naples during the last sixteen years. Pioneer method has been the two-stage splitting procedure (Mola and Siciliano, 1992). Benchmarking is the well known CART methodology and related approaches (Breiman et al. 1984; Breiman, 1996; Hastie et al. 2002; Berthold and Hand, 2003). In the following, we outline the key issues characterizing any tree-based method: the task of the analysis, as we have above discussed, either exploratory or confirmatory; the type of target variable, namely a dummy response variable, a multi-class response, an ordinal response, a numerical response, more recently we are dealing with preference rankings as well as the number of responses, i.e., univariate and multivariate trees (Siciliano and Mola, 2000); the type of predictors (i.e., all categorical, all numerical, mixed); the type of splitting variable, i.e., either simple or multiple split (Siciliano and Mola, 2002); the prior information, so that the predictors can be either active or illustrative if they generate splitting variables or not (Siciliano and Mola, 2004); the data structure, such as data can be classified either in two ways, i.e., measurements of variables on a sample of objects, or in three ways, i.e., variables, objects and occasions, where the latter can be understood as an instrumental variable to stratify either groups of predictors or groups of objects (Tutore et al., 2008); the optimality criteria, specifying the splitting criterion, the model selection, etc.; the context application, as for instance incremental tree-based methods have been proposed in Data Editing, such as missing data imputation (Conversano et al., 2004; D'Ambrosio et al. 2007a), data fusion (D'Ambrosio et al. 2007b), data validation (Petrakos et al., 2004).

3 Prospective view of trees and models

3.1 Prospective splitting by model fit

A greedy searching procedure is usually applied *to look forward for* the best partition of objects into two groups trying out all possible (or a suitable subset of) splits, that can be deduced from the predictor space. This results in the best prospective split at a given node. The optimality criterion usually takes into account the maximization of the decrease of impurity (or variation) of the response variable. As measure of impurity, the residual sum of squares obtained from the model fitting to the parent node as well as to the two sub-nodes can be considered. In this sense, some approaches using generalized linear regression models (Ciampi 1991) and Cox models (Ciampi 1995) as partitioning criteria have been proposed. The goodness of fit of a given model in the two subnodes allows to select the best prospective split of the predictor space. Main idea is that sub-groups at the left and the right nodes are internally homogeneous and externally heterogeneous with respect to the selected submodels. A hierarchy of models is associated to the nodes of the binary tree according to the following rules: (a) if a model at the left sub-node is accepted, then the model at the parent node has been accepted, and (b) if a model at the parent node is rejected, then the models at the left and the right sub-nodes are also rejected. While rule (a) ensures that the hierarchy is respected, rule (b) ensures to stop the procedure when a model is rejected. This methodology is known as model selection tree (Siciliano and Mola, 2004). Two special cases have proved to given interesting results: the use of logistic regression for classification trees (Siciliano et al. 1996) and the use of test for total heterogeneity in regression for regression trees (Siciliano and Mola, 1996).

3.2 Prospective splitting by model parameters

An alternative approach is to choose the best prospective split using model parameters in place of the model fit. Basically, a two-stage criterion can be considered: first, the best predictor is selected to fit the dependence relation with a given model and, then, the best split is chosen on the basis of the fitted model parameters. As an example, the latent budget model has been fruitfully considered to improve classification when the response variable is multi-class (Siciliano, 1999). In particular, Multi-Class Latent Budget Trees (Aria, 2005) selects the best predictor maximizing the goodness of fit of the latent budget model and then the best split on the basis the estimated mixing parameters. These describe the conditional probability of an object to fall into either the left subnode or the right subnode given the predictor feature, thus tho object falls into the subnode where the estimated mixing parameter is the highest. Another example is Ternary Trees by NonSymmetric Correspondence Analysis (Siciliano and Mola, 1998) where the model parameters

are the factorial scores of the objects on the first axis. The idea is to partition the objects into three groups: two of them include those objects with a positive score and those with a negative score respectively, but both induced by predictor features resulting with the highest predictability power, the remaining group includes objects with scores nearly to zero. This procedure allows to identify outliers as well as well predicted tree paths. Other factorial approaches to classification trees are based on discriminant linear functions (Cappelli and Conversano, 2002) and nonlinear canonical correlation analysis such as optimal scaling trees (Tutore et al., 2007).

4 Retrospective view of trees and models

4.1 Retrospective split

The concept of retrospective splits has been introduced in Two-Stage Discriminant Trees for regression trees (Mola and Siciliano, 2002) and also considered in Clockwork Trees through Visual Multivariate Splitting (Conversano et al. 2003). The split induced by splitting the objects on the basis of the response variable, discarding the predictors, is called retrospective. For instance, the measurement values of a numerical response variable can be classified into two groups on the basis of the highest between deviation. The result is the optimal (theoretical) retrospective split of the objects at a given node. Two properties can be mentioned: 1) the set of prospective splits is not necessarily coincident with the set of retrospective splits, being the former included into the latter; 2) a split of objects based just on the response (i.e., a retrospective split) could not present any split of the predictor features (i.e., a prospective split) yielding to the same partition. As a result, at any node, for any optimal partition of the objects due to the best retrospective split it might not exist an observed partition of the predictor features provided by any observed prospective split. Thus, for interpretative purposes, it is necessary *to look backward to* find the most coherent observed split of the predictor features to be associated to such an optimal split of the objects. A tree growing procedure based on the best retrospective splits yields to the optimal splitting tree retrospectively, where the optimality is referred to the optimal partition of objects obtained at each node of the tree.

4.2 Optimal tree growing algorithm

A tree is optimal if it is recursively partitioned on the basis of the optimal (theoretical) retrospective split at each node. Thus, partitioning is only governed by the response variable, in this respect ensuring at each node the most internally homogeneous subgroups as well as the most heterogeneous subgroups. Table 1 shows the pseudo-code of the optimal tree growing algorithm. Once in the preliminary step the response variable is sorted, all its

$n-1$ possible binary partition are considered. The split is then generated by maximizing the between sum of squares. Stopping rules to tree growing are node size and tree depth.

Let $\mathbf{X}$ be a $n \times p$ matrix of numerical and categorical predictors. Let Y be a $n \times 1$ vector of numerical responses. Sort the data matrices X and Y according to the response variable distribution. Let t the number of a tree-node; Let $s \in S$ be the generic binary split of Y. Let Db_s the Y between sum of squares measure calculated between the two sub-distributions obtained by the s split. Start from root node. • while a tree growing stopping rule does not occur, repeat for each child node – find the optimal binary split S^* of Y maximizing Db_s respect to $s \in S$ – generate the two children nodes according to the split s^* • Output: Optimal tree according to the Y distribution.

Table 1. Optimal tree growing algorithm.

4.3 Posterior model prediction criterion

The most coherent prediction model with respect to the optimal retrospective split needs to be found. In particular, a class of models has to be used to generate the prediction rules. Table 2 shows the pseudo-code of the prediction modelling algorithm. For each node, a suitable model generates the rules to explain the partitioned response variable. This process continues until a pruning rule does not occur. The pruning rule at a node t is defined by the significance of the selected model. In other words, if the model is not significant the branch is pruned and the node t becomes terminal. As a result, the prediction modelling is assigned a-posteriori the optimal partition.

5 An application using the logistic regression

Concrete Compressive Strength data set (available on UCI machine learning repository) consists on nine numerical variables and 1030 instances. The goal is to predict the concrete compressive strength given a set of predictors. The complete set of variables is the following:

- X_1: cement, kg in a m^3 mixture;
- X_2: blast furnace slag, kg in a m^3 mixture;

Let $\mathcal{M}$ a class of models according to the splitting criterion of the optimal tree;
Let $m \in \mathcal{M}$ a generic model function;
Let $m_t \in \mathcal{M}$ a generic model function selected at the node t ;

Start from root node.

- while a tree pruning rule does not occur, repeat for each child node
  - estimate m_s, the best model $m \in \mathcal{M}$ at the father node t.
  - generate by the model m_s the prediction rule H_t for new observations.
- Output: Prediction rules set by the class of models $\mathcal{M}$.

Table 2. Prediction modelling algorithm.

- X_3: fly ash, kg in a m^3 mixture;
- X_4: water, kg in a m^3 mixture;
- X_5: superplasticizer, kg in a m^3 mixture;
- X_6: coarse aggregate, kg in a m^3 mixture;
- X_7: fine aggregate, kg in a m^3 mixture;
- X_8: age, day (1 - 365);
- Response variable: concrete compressive strength, MPa.

Once the optimal tree was built, response variable was iteratively re-coded in a binary one with respect to the node in which its values fallen down. Starting form root node, values of the response variable were re-coded as 0 if they fallen down child node 2 (the left one) and as 1 otherwise (the right one); considering node 2, response variable values were re-coded as 0 if they fallen down child node 4 (its left child node) and as 1 otherwise; and so on. Since in building the optimal tree the response variable needs to be sorted, for each split its lower values are re-coded as 0 as well as its higher values are re-coded as 1.
The posterior prediction model chosen was the logistic regression model. In each node the logistic regression was fitted using a backward procedure to select the predictors to be included in the model. Goodness of fit measure of the models was the McCullagh and Nelder deviance (McCullagh and Nelder, 1990), which compares the actual model with the saturated one: the higher is the p-value associated to this statistics, the better is the model because the null hypothesis concerns the difference between the saturated and the actual model. Figure 1 shows the resulting tree, as well as table 3 shows the detailed output. The numbers above nodes indicate node number, whereas under the nodes the p-value of the model deviance at node is reported.

Table 3 consists of several columns which describe the characteristics of each node. Columns *model coefficients at node* show the coefficients of the logistic regression. As an example, by considering the root node the first logistic model is the following:

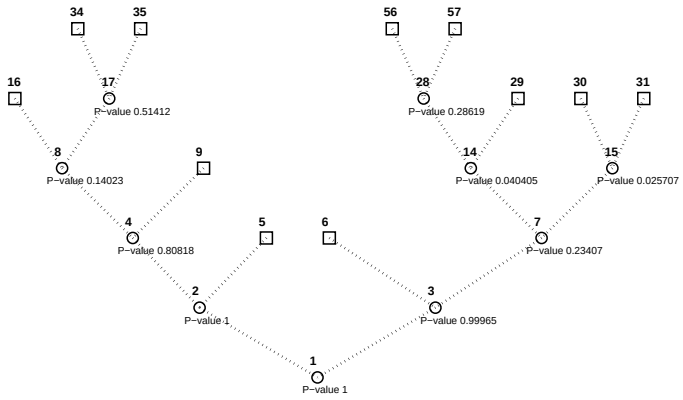

Fig. 1. Posterior prediction modelling tree.

$$\pi_Y = (1 + exp - (0.0171X_1 + 0.0132X_2 + 0.0129X_3 - 0.0391X_4 \\ -0.0015X_7 + 0.0341X_8))^{-1}$$

This means that the probability of each observation to belong to class 1 increases when values of X_1, X_2, X_3, X_8 variables (Cement, Blast Furnace Slag, Fly Ash, Age) increases, as well as probability to belong to the same class decreases when values of both X_4 (Water) and X_7 (Fine Aggregate) variables increase. In several non-terminal nodes, the model suggests an interaction among more than one variable which explains the best partition determined by the tree-growing phase (in fact in all internal nodes except for the nodes number 15 and 28).
The column named *cutting point* shows the assignment rule of the logistic model in terms of probability to belong to either class 0 or class 1. This assignment rule works in this way:
Start from the minimum probability value assigned by the model at each individual at the generic t node.

- assign response class according to the given probability representing the actual cutting point and compute the relative misclassification ratio;
- increase the cutting point by 0.01, assign response class and compute the relative misclassification ratio;
- repeat the last step until the cutting point reaches the minimum misclassification ratio.

The column called *model error rate* shows the misclassification ratio of the logistic model for each node in predicting the re-coded binary response variable. As an example, the previously described logistic model in the root node makes an error of 15.73% in classifying the re-coded response variable. With few exceptions, the model error rates tend to increase when node size decreases. An important remark is that if there is a bad performance of a model in terms of misclassification ratio in a given node, there is not propagation of this error in the subsequent nodes because the initial tree partition is optimal in the sense that tree-structure has been built through an optimal retrospective split.

6 Conclusions

The optimal partition of objects obtained at each node of the tree is achieved through the retrospective splitting procedure. The model prediction criterion introduced in this paper tries to confirm this optimal partition with the use of predictors. Once a class of suitable models is chosen, the risk of an error propagation through the tree structure is minimum because the models are independent between them.
Indeed in a CART-like process if the splitting rules do not sufficiently discriminate the partition in terms of the response variable, this lack of discrimination can be updated in the overall tree-based structure. An advantage of this procedure is that if within a given node a model does not work in the best way, in any case the subsequent partition is the optimal one because it is not governed by predictors.
In addiction, the posterior prediction modelling criterion considers interactions among more variables in a given node whereas CART-like procedure considers only one variable at turn in determining the partition.

Node Number	Model coefficients at node			Deviance p-value	Cutting point	Parent node	Model error rate	Node size
	intercept	positive coefficient	negative coefficient					
1	—	$\beta_1 = 0.0171$ $\beta_2 = 0.0132$ $\beta_3 = 0.0129$ $\beta_8 = 0.0341$	$\beta_4 = -0.0391$ $\beta_7 = -0.0015$	1.0000	0.3059	*root node*	0.1573	1030
2	$\beta_0 = -43.0641$	$\beta_1 = 0.0483$ $\beta_2 = 0.0385$ $\beta_3 = 0.0392$ $\beta_6 = 0.0093$ $\beta_7 = 0.00184$ $\beta_8 = 0.1839$	—	1.0000	0.4513	1	0.1252	583
4	$\beta_0 = -65.5352$	$\beta_1 = 0.0588$ $\beta_2 = 0.0471$ $\beta_3 = 0.0428$ $\beta_6 = 0.0209$ $\beta_7 = 0.0304$ $\beta_8 = 0.2235$	—	0.8082	0.4542	2	0.1941	237
8	—	$\beta_1 = 0.0165$ $\beta_8 = 0.0951$	$\beta_4 = -0.0195$	0.1402	0.4293	4	0.2523	107
17	$\beta_0 = 89.1956$	$\beta_1 = 0.0171$	$\beta_4 = -0.2168$ $\beta_6 = -0.0245$ $\beta_7 = -0.0340$	0.5141	0.4798	8	0.1846	65
3	$\beta_0 = -23.7535$	$\beta_1 = 0.0276$ $\beta_2 = 0.0281$ $\beta_3 = 0.0282$ $\beta_6 = 0.0089$ $\beta_7 = 0.0090$ $\beta_8 = 0.0146$	$\beta_4 = -0.0379$	0.9996	0.6508	1	0.1655	447
7	—	$\beta_1 = 0.0219$ $\beta_2 = 0.0315$ $\beta_3 = 0.0223$ $\beta_8 = 0.0144$	$\beta_4 = -0.0739$ $\beta_5 = -0.1501$	0.2341	0.5620	3	0.2252	151
14	—	$\beta_1 = 0.0061$ $\beta_2 = 0.0072$ $\beta_7 = 0.0080$	$\beta_4 = -0.446$ $\beta_5 = -0.1790$	0.0404	0.4279	7	0.3000	90
28	—	—	$\beta_4 = -0.0071$	0.2862	0.2665	14	0.2549	51
15	—	—	$\beta_4 = -0.0019$	0.0257	0.4107	7	0.5574	61
5	Tn, $\widehat{y} = 30.2373$	—	—	—	—	2	—	346
9	Tn, $\widehat{y} = 18.4120$	—	—	—	—	4	—	130
34	Tn, $\widehat{y} = 11.2834$	—	—	—	—	17	—	27
35	Tn, $\widehat{y} = 13.3061$	—	—	—	—	17	—	38
16	Tn, $\widehat{y} = 7.7974$	—	—	—	—	8	—	42
6	Tn, $\widehat{y} = 44.5802$	—	—	—	—	3	—	296
29	Tn, $\widehat{y} = 62.0715$	—	—	—	—	14	—	39
30	Tn, $\widehat{y} = 68.5765$	—	—	—	—	15	—	35
31	Tn, $\widehat{y} = 77.0979$	—	—	—	—	15	—	26
56	Tn, $\widehat{y} = 55.8595$	—	—	—	—	28	—	39
57	Tn, $\widehat{y} = 58.2646$	—	—	—	—	28	—	12
Tn = Terminal node								

Table 3. Posterior prediction modelling tree: main results.

Acknowledgements: Financial support from MIUR of Italy and from European FP6 Project iWebCare IST-4-02-8055 (Scientific Responsible: Prof. Roberta Siciliano).

References

ARIA, M., (2005): Multi-Class Budget Exploratory Trees. In *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag, 3-8.

BERTHOLD, M., and HAND, D.J. (Eds.) (2003). *Intelligent Data Analysis*. Second edition, Springer, Berlin.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE C.J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont CA.

BREIMAN, L. (1996). Bagging Predictors, *Machine Learning*, 24, 123-140.

CAPPELLI, C., and CONVERSANO, C. (2002). Canonical Discriminant Function for Recursive Partitioning in Data Mining, in Härdle W. and Rönz B (eds), *COMPSTAT 2002 Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, pag. 213-218.

CAPPELLI, C., MOLA, F., and SICILIANO, R. (2002). A Statistical Approach to Growing a Reliable Honest Tree, *Computational Statistics and Data Analysis*, 38, 285-299.

CIAMPI, A. (1991). Generalized Regression Trees. *Computational Statistics and Data Analysis* 12, 57-78.

CIAMPI, A., NEGASSA, A., LOU, Z. (1995) Tree-structured prediction for censored survival data and the Cox model. *Journal of Clinical Epidemiology* 48(5), 675-689.

CONVERSANO C., and CAPPELLI C., (2002). Missing Data Incremental Imputation through Tree Based Methods, in Härdle W. et al. (eds.), Proceedings in Computational Statistics COMPSTAT 2002-, Physica-Verlag, pag. 455-460.

CONVERSANO, C., DI BENEDETTO, D., and SICILIANO, R. (2003). The Clockwork Trees through Visual Multivariate Splitting, in *Proceedings of CLADAG 2003*, Physica Verlag, 113-116.

CONVERSANO, C., MOLA, F., and SICILIANO, R. (2001). Partitioning and Combined Model Integration for Data Mining. *Journal of Computational Statistics*, 16, 323-339, Physica Verlag, Heidelberg (D).

CONVERSANO C., and SICILIANO R., (2004). Incremental Tree-Based Imputation with lexicographic ordering, Interface 2003 Proceedings, Minotte M., Swzychak A. (eds.), Interface Foundation of North America, Washington, CD-ROM.

D'AMBROSIO, A., ARIA, M., and SICILIANO, R. (2007). Robust Tree-based Incremental Imputation Method for Data Fusion, in *Proceedings of the 7th IDA2007 Conference (Ljubljana, 6-8 September, 2007), Lecture Notes in Computer Science Series of Springer*, 174-183.

FREUND Y., and SCHAPIRE R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1) 119-139.

HASTIE, T., FRIEDMAN, J. H., and TIBSHIRANI, R. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.

MC CULLAGH, P., and NELDER, J. A. (1990). *Generalized Linear Models*. Second edition, Chapman and Hall.

MOLA, F., and SICILIANO, R. (1992). A two-stage predictive splitting algorithm in binary segmentation, in Y. Dodge, J. Whittaker. (Eds.): *Computational Statistics: COMPSTAT 92*, 1, Physica Verlag, Heidelberg (D), 179-184.

MOLA, F., and SICILIANO, R. (1994). Alternative strategies and CATANOVA testing in two-stage binary segmentation, in E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy (Eds.): *New Approaches in Classification and Data Analysis: Proceedings of IFCS '93*, Springer Verlag, Heidelberg (D), 316-323.

MOLA, F., and SICILIANO, R. (1997). A Fast Splitting Procedure for Classification and Regression Trees, *Statistics and Computing*, 7, Chapman Hall, 208-216.

MOLA, F., and SICILIANO, R. (2002). Discriminant Analysis and Factorial Multiple Splits in Recursive Partitioning for Data Mining, in Roli, F., Kittler, J. (eds.): *Proceedings of International Conference on Multiple Classifier Systems* (Chia, June 24-26, 2002), 118-126, Lecture Notes in Computer Science, Springer, Heidelberg.

PETRAKOS, G., CONVERSANO, C., FARMAKIS, G., MOLA, F., SICILIANO, R., and STAVROPULOS, P. (2004): New ways to specify data edits. *Journal of the Royal Statistical Society*, Series A Statistics in Society, Ser. A, 167, Part 2, 249-274.

SICILIANO, R. (1998). Exploratory versus decision trees. In: Payne, R., Green, P. (Eds.), *Proceedings in Computational Statistics*. Physica-Verlag, 113-124.

SICILIANO, R. (1999). Latent budget trees for multiple classification, in M. Vichi, P. Optitz (Eds.): *Classification and Data Analysis: Theory and Application*, Springer Verlag, Heidelberg (D), 121-130.

SICILIANO, R., and MOLA, F. (1994). Modelling for Recursive Partitioning and Variable Selection, in R. Dutter, W. Grossmann (Eds.): *Proceedings in Computational Statistics: COMPSTAT '94* (Vienna, August 24-28, 1994), Physica Verlag, Heidelberg (D), 172-177.

SICILIANO, R., MOLA, F., and KLASCHKA, J. (1996). Logistic Classification Trees, in A. Prat (Ed.): *Proceedings in Computational Statistics: COMPSTAT '96* (Barcellona, August 24-28, 1996), , Physica-Verlag, Heidelberg (D), 373-378.

SICILIANO, R., and MOLA, F. (1996). A Fast Regression Tree Procedure, in Forcina, A., Marchetti, G.M., Hatzinger, R., Galmacci, G. (Ed.): *Statistical Modelling, Proceedings of the 11th International Workshop on Statistical Modeling* (Orvieto, 15-19 luglio), 332-340, Graphos, Citt di Castello.

SICILIANO, R., and MOLA, F. (1998). A general splitting criterion for classification trees, *Metron*, 56, 3-4.

SICILIANO, R., and MOLA, F. (1998). Ternary Classification Trees: a Factorial Approach, in M. Greenacre, J. Blasius (Eds.): *Visualization of Categorical Data*, cap. 22, , Academic Press, San Diego (CA), 311-323.

SICILIANO, R. and MOLA, F (2000): Multivariate Data Analysis through Classification and Regression Thees, *Computational Statistics and Data Analysis*, 32, Elsevier Science, 285-301.

SICILIANO, R., ARIA, M., and CONVERSANO, C. (2004). Tree Harvest: Methods, Software and Applications, in Antoch J. (ed.): *COMPSTAT 2004 Proceedings*, Springer, 1807-1814.

TUTORE, V.A., SICILIANO, R. and ARIA, M. (2007), Conditional Classification Trees using Instrumental Variables. *Advances in Intelligent Data Analysis*, Springer-Verlag, pp 163-173.

Part X

Model Selection Algorithms

Selecting Models Focusing on the Modeller's Purpose

Jean-Patrick Baudry[1], Gilles Celeux[2], and Jean-Michel Marin[2]

[1] Université Paris-Sud
91450 Orsay France, *Jean-Patrick.Baudry@math.u-psud.fr*
[2] Inria Saclay
91450 Orsay, France, *Gilles.Celeux@inria.fr*, *Jean-Michel.Marin@inria.fr*

Abstract. Model selection is a difficult task for which it is often profitable to take into account the modeller point of view. Hidden structure models are a good example for which this point of view can be dealt with in a simple way. In the model-based clustering context, we present model selection criteria focussing on the clustering purpose. Their rationale and theoretical features are given and their practical behavior in comparison with classical penalized likelihood criteria is discussed from numerical experiments.

Keywords: mixture model, BIC, entropy, slope heuristic

1 Selection of hidden structure models

Standard model selection criteria as AIC or BIC implicitly assume that the sampling distribution is belonging to, at least, one of the models in competition (see for instance Burnham and Anderson 1998). This assumption is most often unrealistic and can lead to underpenalize complex models. Taking into account the modelling purpose can counter efficiently this tendency. This point of view is much sensible for hidden structure models. In this setting, discovering the hidden structure is often of primary interest for the user to derive a reliable clustering of his data set. In this article, we present model selection criteria aiming at favoring models for which the ratio "observed information/complete information" is small.

In hidden structure models, complete data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ take the form $\mathbf{x}_i = (\mathbf{y}_i, \mathbf{z}_i)$, $\mathbf{y}_i$ being the observed data for unit i and $\mathbf{z}_i$ being a hidden state in $\{1, \ldots, K\}$. Denoting respectively $\mathbf{p}(\mathbf{x}|\theta)$ and $\mathbf{p}(\mathbf{y}|\theta)$ the parameterized density of complete data and observed data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ and $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ the vector of missing labels, we get

$$\mathbf{p}(\mathbf{x}|\theta) = \mathbf{p}(\mathbf{y}|\theta)\mathbf{p}(\mathbf{z}|\mathbf{y}, \theta) \tag{1}$$

and

$$\mathbf{p}(\mathbf{y}|\theta) = \sum_{\mathbf{z}} \mathbf{p}(\mathbf{y}|\mathbf{z}, \theta)\mathbf{p}(\mathbf{z}|\theta). \tag{2}$$

Finite mixture models, hidden Markov chains models, competing risk models are examples of hidden structure models (cf. McLachlan and Peel, 2000, Cappé, Moulines and Ryden, 2005, and Crowder, 2001). An important issue when using such models is to choose an appropriate number K of labels. In that purpose, when focussing on a good estimation of the density of the observed data, standard penalized likelihood criteria such as BIC can be recommended (Schwarz 1978). But, when the purpose is to get the "best" clustering of the observed data, it is highly desirable to take this focus into account to design a relevant model selection criterion.

The aim of this paper is to present such criteria and discuss their interest. To be more specific, we concentrate the presentation on finite mixture models used in the model-based clustering context.

Model-based clustering (MBC) consists of assuming that the observed data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ in $\mathbf{R}^{nd}$ arise from a mixture

$$p(\mathbf{y}_i \mid K, \theta_K) = \sum_{k=1}^{K} p_k \phi(\mathbf{y}_i \mid \mathbf{a}_k)$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k p_k = 1$) $\phi(. \mid \mathbf{a}_k)$ denotes a parameterized density (often the d-dimensional Gaussian density) with parameter $\mathbf{a}_k$, and the vector parameter to be estimated is $\theta_K = (p_1, \ldots, p_K, \mathbf{a}_1, \ldots, \mathbf{a}_K)$. A mixture model involves missing label data $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ where the $\mathbf{z}_i's$ are binary vectors with $z_{ik} = 1$ if and only if $\mathbf{y}_i$ arises from component k. Those indicator vectors define a partition $P = (P_1, \ldots, P_K)$ of data $\mathbf{y}$ with $P_k = \{\mathbf{y}_i \mid z_{ik} = 1\}$.

2 The integrated complete likelihood criterion

Two likelihoods can be defined with finite mixture models. The observed loglikelihood of $\theta = (p_1, \ldots, p_K, \mathbf{a}_1, \ldots, \mathbf{a}_K)$ is

$$L(K) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} p_k \phi(\mathbf{y}_i \mid \mathbf{a}_k) \right].$$

The complete loglikelihood of θ for the complete sample $\mathbf{x}$ is

$$CL(K) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(p_k \phi(\mathbf{y}_i \mid \mathbf{a}_k)).$$

Those loglikelihoods are linked by the following relation (Hathaway 1986)

$$CL(K) = L(K) - E(K) \tag{3}$$

$$\text{where } E(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \log t_{ik} \geq 0,$$

$$t_{ik} = \frac{p_k \phi(\mathbf{y}_i \mid \mathbf{a}_k)}{\sum_{j=1}^{K} p_j \phi(\mathbf{y}_i \mid \mathbf{a}_j)}$$

being the conditional probability that $\mathbf{y}_i$ arises from component k.

In a Bayesian perspective, a classical way for choosing a model is to select the model maximizing the integrated likelihood,

$$\mathbf{f}(\mathbf{y} \mid K) = \int \mathbf{f}(\mathbf{y} \mid K, \theta)\pi(\theta \mid K)d\theta, \tag{4}$$

$$\mathbf{f}(\mathbf{y} \mid K, \theta) = \prod_{i=1}^{n} f(\mathbf{y}_i \mid K, \theta),$$

$\pi(\theta \mid K)$ being a non or weakly informative prior distribution on θ. For n large enough, formula (4) can be approximated with the BIC criterion (see for instance Raftery, 1995)

$$\log \mathbf{f}(\mathbf{y} \mid K) \approx \log \mathbf{f}(\mathbf{y} \mid K, \hat{\theta}) - \frac{\nu_K}{2} \log n,$$

where $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ derived from $\mathbf{y}$ and ν_K is the number of free parameters in the model. Simulation experiments (see Roeder and Wasserman 1997) show that BIC works well when K is to be chosen in a density estimation purpose.

But when the aim is to choose K to get the mixture giving rise to partitioning data with the greatest evidence, it makes sense to take into account the conditional distribution of the missing labels which define a fuzzy classification matrix: $\mathbf{t} = (\{t_{ik}\}, i = 1, \ldots, n; k = 1, \ldots, K)$. As shown in Biernacki *et al.* (2000), the mixture entropy

$$\mathrm{ENT}(K) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \log t_{ik} \geq 0, \tag{5}$$

is a measure of the ability of the K-component mixture to provide a relevant "hard" partition of the data. If the mixture components are well-separated, the classification matrix $\mathbf{t}$ tends to define a partition of $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ and $\mathrm{ENT}(K) \approx 0$. But if the mixture components are poorly separated, $\mathrm{ENT}(K)$ has a large value.

The integrated likelihood of the complete data $(\mathbf{y}, \mathbf{z})$ is

$$\mathbf{f}(\mathbf{y}, \mathbf{z} \mid K) = \int \mathbf{f}(\mathbf{y}, \mathbf{z} \mid K, \theta)\pi(\theta \mid K)d\theta.$$

It can be approximated using a BIC-like approximation (Biernacki *et al.*, 2000):

$$\log \mathbf{f}(\mathbf{y}, \mathbf{z} \mid K) \approx \log \mathbf{f}(\mathbf{y}, \mathbf{z} \mid K, \hat{\theta}^*) - \frac{\nu_K}{2} \log n$$

with

$$\hat{\theta}^* = \arg\max_{\theta} \mathbf{f}(\mathbf{y}, \mathbf{z} \mid K, \theta).$$

But $\mathbf{z}$ and $\hat{\theta}^*$ are unknown. However, assuming that, for n large enough,$\hat{\theta} \approx \hat{\theta}^*$, Biernacki *et al.* (2000) propose, in an empirical way, to replace the missing data $\mathbf{z}$ with $\hat{\mathbf{z}} = \text{MAP}(\hat{\theta})$ defined by

$$\hat{z}_{ik} = \begin{cases} 1 \text{ if } \arg\max_{\ell} t_{i\ell}(\hat{\theta}) = k \\ 0 \text{ otherwise.} \end{cases}$$

This leads to the criterion

$$\text{ICL}(K) = \log \mathbf{f}(\mathbf{y}, \hat{\mathbf{z}} \mid K, \hat{\theta}) - \frac{\nu_K}{2} \log n. \tag{6}$$

¿From formulae (6) and (3), criterion ICL can be interpreted as BIC penalized by the estimated entropy $\text{ENT}(K)$ defined in (5). Because of this additional entropy term, ICL can be helpful to identify a number of clusters which need not be the same as the number of mixture components which provides a good fit of the data. Thus, ICL is expected to provide a stable and reliable estimate of K for real data sets and also for simulated data sets from mixtures when the components are not too much overlapping. But ICL, which is not aiming at discovering the true number of mixture components, can underestimate the number of components for simulated data arising from mixture with poorly separated components. In the model-based clustering context, ICL can be expected to be mode robust than BIC : as a matter of fact, the additional entropic term $\text{ENT}(K)$ in ICL can be regarded as a term counterbalancing the model mispecification.

3 Minimum contrast estimation and slope heuristic

In the following, $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ denotes the random variables, and $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ a realization.

3.1 Minimum contrast estimation

Let s^* be a target distribution belonging to a universe $\mathcal{U}$, related to the sampling distribution f of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Minimum constrast estimation consists of estimating s^* in a model $S_m \subset \mathcal{U}$, S_m belonging to a family of models $\{S_m\}_{m \in \mathcal{M}}$, by considering some empirical contrast function γ_n, depending on $(\mathbf{Y_1}, \ldots, \mathbf{Y_n})$, such that $s \in \mathcal{U} \longmapsto \mathbb{E}[\gamma_n(s)]$ attains its minimum at s^*. Examples of standard empirical contrasts are mean squared error and $-$loglikelihood of a model. The target in model S_m is $s_m = \arg\min_{s \in S_m} \mathbb{E}[\gamma_n(s)]$. It is estimated by

$$\hat{s}_m = \arg\min_{s \in S_m} \gamma_n(s).$$

In this framework, the oracle model is

$$\hat{m}(s^*) = \arg\min_{m\in\mathcal{M}} \mathbb{E}_{\mathbf{Y}'}\left\{\mathbb{E}_{\mathbf{Y}}\left[\gamma_n^{\mathbf{Y}}(\hat{s}_m^{\mathbf{Y}'})\right] - \mathbb{E}_{\mathbf{Y}}\left[\gamma_n^{\mathbf{Y}}(s^*)\right]\right\}.$$

Remark 3. In this formula, $\gamma_n^{\mathbf{Y}}$ denotes the empirical contrast associated to $\mathbf{Y}$, and the expectation with respect to $\mathbf{Y}'$, which is supposed to be independent from $\mathbf{Y}$, is related to $\hat{s}_m^{\mathbf{Y}'}$. Here, $\hat{s}_m^{\mathbf{Y}'} = \arg\min_{s\in S_m} \gamma_n^{\mathbf{Y}'}(s)$.

The penalized minimum contrast estimation procedure consists of considering some proper penalty function pen : $\mathcal{M} \longmapsto \mathbf{R}^+$ and choosing $\hat{m}$ minimizing $\gamma_n(\hat{s}_m) + \text{pen}(m)$ over $\mathcal{M}$. It can be shown that, under regularity conditions, the criterion to be minimized takes the form

$$\gamma_n(\hat{s}_m) + 2V_m,$$

where V_m is a variance term of the form $V_m \approx \beta D_m$, D_m being the number of free parameters in the model S_m (see Massart, 2007, Section 8.5). With this approach, it remains to choose the constant β from an heuristic point of view...

In the model-based cluster analysis context, we consider CCL, the conditional expectation of the complete likelihood contrast, rather than the likelihood contrast. As a matter of fact, the approximation made in Biernacki *et al.* (2000), $\hat{\theta} \approx \hat{\theta}^*$ (see page 340), can be misleading. The setting is as follows:

- $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \in \mathbb{R}^d$ are iid random variables with density $\mathbf{p}$,
- $\mathcal{U}$ is the set of finite Gaussian mixtures on $\mathbb{R}^d$, and s is denoting the density of a finite Gaussian mixture,
- The family of models is $\{S_m\}_{m\in\mathcal{M}}$, with $S_m = \{\mathbf{p}(\cdot\,;\theta_m)|\theta_m \in \Theta_m\} \subset \mathcal{U}$, Θ_m being the set of parameters for model S_m,
- $\forall s \in \mathcal{U}$, the empirical contrast is the opposite of the conditional expectation of the complete loglikelihood

$$\gamma_n(s) = -\frac{1}{n}\sum_{i=1}^{n} \log s(\mathbf{y}_i) + \text{ENT}(\mathbf{y}_1, \ldots, \mathbf{y}_n; s),$$

$$\text{where} \quad \text{ENT}(\mathbf{y}_1, \ldots, \mathbf{y}_n; s) = -\sum_{i=1}^{n}\sum_{k\in s} t_i^k(s) \log t_i^k(s)$$

$$\text{with} \quad t_i^k(s) = \mathbb{P}_s[\mathbf{Y}_i \text{ arose from component } k \text{ of mixture } s \,|\, \mathbf{y}_i]$$

Thus, in model S_m, $\hat{s}_m = \mathbf{p}(\cdot\,;\hat{\theta}_m^{MccLE})$, $\hat{\theta}_m^{MccLE}$ minimizing in S_m:

$$-CCL(\theta_m) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbf{p}(\mathbf{y}_i\,;\theta_m) + \text{ENT}(\mathbf{y}_1, \ldots, \mathbf{y}_n; \theta_m).$$

The model selection criterion to be minimized is then

$$-\frac{1}{n}\sum_{i=1}^{n} \log \mathbf{p}(\mathbf{y}_i\,;\hat{\theta}_m^{MccLE}) + \text{ENT}(\mathbf{y}_1, \ldots, \mathbf{y}_n; \hat{\theta}_m^{MccLE}) + 2\beta D_m,$$

β being a constant to be chosen.

3.2 Slope heuristic

The constant β has to be chosen in a heuristic way. The slope heuristic is directly deduced from the form of the model selection criterion to be designed. In good settings, the bias

$$\mathbb{E}[\gamma_n(s_m)] - \mathbb{E}[\gamma_n(s^\star)]$$

is expected to be constant –if not null– for models with high complexities. Therefore the empirical constrast is expected to behave linearly with respect to the number of parameters for those models. Then, the heuristic consists of choosing β to be the slope of the linear part of the empirical contrast as illustrated in Figure 1.

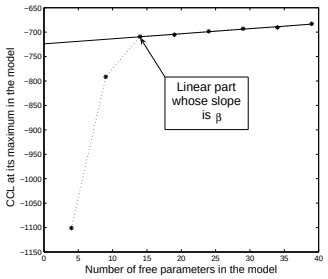

Fig. 1. Calibration of the slope.

In practice, this slope is computed in the following way: The maximum values of CCL for $K = 1, \ldots, K_{\max}$ are computed. Then, for $i = 1, \ldots, K_{\max} - 1$, a slope β_i is derived from a linear regression on the values $\max \mathrm{CCL}$ for $\{i, \ldots, K_{\max}\}$. The idea is to choose $\beta = \beta_{i_0}$, where i_0 is the index from which the sequence (β_i) becomes stable. In the following applications, the choice of i_0 such that $|\beta_{i_0+1} - \beta_{i_0}| < 2 \min_i |\beta_{i+1} - \beta_i|$ works well.

Remark 4. On the graph of the complete loglikelihood at its maxima, the emergence of a linear part when the numbers of components increases confirms that the heuristic can be applied (Massart 2007). Nevertheless, because of possible numerical difficulties (spurious or unsensible maxima), some points out of this linear part could appear. In such a case, using a robust linear regression method (see Rousseuw and Leroy 2003) instead of standard linear regression is recommended.

3.3 Computing MccLE

Computing the MLE for a Gaussian mixture is usually done with the EM algorithm. Computing the parameter value maximizing CCL is more difficult.

However, it is possible to adapt the EM algorithm along the lines described in Lange (1999, chapter 12) to obtain a Bayesian version of EM to compute the maximum CCL estimator (MccLE) for a Gaussian mixture model S_m. The iteration ℓ of the adapted EM algorithm is as follows (the m index is omitted):

- *E step:* Compute $\mathbb{E}_{\theta^\ell}[\mathrm{CCL}(\theta)|\mathbf{y}] = L(\theta) + \sum_{i=1}^n \sum_{k=1}^K t_i^k(\theta^\ell) \log t_i^k(\theta)$.
- *M step:* Maximize $\mathbb{E}_{\theta^\ell}[\mathrm{CCL}(\theta)|\mathbf{y}] - \mathrm{ENT}(\mathbf{y}; s(\cdot;\theta))$

$$\begin{aligned}\theta^{\ell+1} &= \arg\max_{\theta\in S_m} \left\{ L(\theta) + \sum_{i=1}^n \sum_{k=1}^K t_i^k(\theta^\ell) \log t_i^k(\theta) - \mathrm{ENT}(\mathbf{y}; s(\cdot;\theta)) \right\} \\ &= \arg\max_{\theta\in S_m} \underbrace{\left\{ L(\theta) + \sum_{i=1}^n \sum_{k=1}^K (t_i^k(\theta^\ell) + t_i^k(\theta)) \log t_i^k(\theta) \right\}}_{A(\theta)}.\end{aligned}$$

The difference with EM lies in the M step: the M step for mixtures is an easy step with EM. But maximizing the conditional expectation of the complete loglikelihod, the M step needs to call a gradient-maximization program. Gradient optimisation could be applied directly on the conditional expectation of the complete loglikelihood, but the adapted EM algorithm improves the performances. Moreover, it can be expected to do the job because of the following result:

Proposition 2 (Adapted EM algorithm).
If

$$A(\theta^{\ell+1}) > A(\theta^\ell) \quad \textit{(see M step)},$$

Then

$$CCL(\theta^{\ell+1}) < CCL(\theta^\ell).$$

Proof. ¿From the M step

$$\begin{aligned}L(\theta^{\ell+1}) + \textstyle\sum_{i=1}^n \sum_{k=1}^K (t_i^k(\theta^\ell) + t_i^k(\theta^{\ell+1})) \log t_i^k(\theta^{\ell+1}) > \\ L(\theta^\ell) + \sum_{i=1}^n \sum_{k=1}^K (t_i^k(\theta^\ell) + t_i^k(\theta^\ell)) \log t_i^k(\theta^\ell).\end{aligned}$$

Then, denoting

$$\mathrm{ENT}(\mathbf{y}_1, \ldots, \mathbf{y}_n; \theta) = -\sum_{i=1}^n \sum_{k=1}^K t_i^k(\theta) \log t_i^k(\theta),$$

we have

$$L(\theta^{\ell+1}) - \mathrm{ENT}(\mathbf{y}; \theta^{\ell+1}) > L(\theta^\ell) - \mathrm{ENT}(\mathbf{y}; \theta^\ell) + \sum_{i=1}^n \sum_{k=1}^K t_i^k(\theta^\ell) \log \frac{t_i^k(\theta^\ell)}{t_i^k(\theta^{\ell+1})}.$$

But $\sum_{i=1}^{n}\sum_{k=1}^{K} t_i^k(\theta^\ell)\log\frac{t_i^k(\theta^\ell)}{t_i^k(\theta^{\ell+1})} = \sum_{i=1}^{n} d_{KL}(t_i(\theta^\ell), t_i(\theta^{\ell+1}))$. Kullback-Leibler distances between probability distributions being non-negative quantities, the proposition is proved. ◇

Another difficulty to be discussed is the initialization step: we suggest the following strategy, which works well in our experiments:

- Run the adapted EM algorithm a lot of times with a few iterations from random initial positions.
- Choose the solution for which the complete loglikelihood is the largest, to initialize the adapted EM algorithm with a large number of iterations.

4 Numerical comparisons

The two model selection criteria presented in Section 2 and 3 are compared with BIC on Monte Carlo numerical experiments in an illustrative purpose. Two different simulated mixture models are considered.

4.1 Experiments with overlapping clusters

A four-component Gaussian mixture model depicted in Figures 2 and 3 has been simulated with a sample size $n = 200$. The experiments have been repeated 100 times. A diagonal Gaussian mixture model has been fitted with the EM algorithm, the Table 1 provides the frequencies of the resulting number of components K with criteria BIC, ICL and MccLE (slope heuristic) among 100 experiences.

K	2	3	4	5	6	7	8
BIC	0	11	87	2	0	0	0
ICL	0	98	2	0	0	0	0
Slope heuristic (MccLE)	2	79	8	8	3	0	0

Table 1. Frequencies of resulting number of components for each criterion.

The true distribution of the data belongs to the model with four components. BIC does its job: it mostly selects the true number of components, and achieves good performances in a density estimation purpose. At the opposite, ICL mostly selects three components. It corresponds to the three clusters appearing in this data set: ICL behaves as expected. The slope heuristic for MccLE behaves analogously, but is not as good as ICL. It might be the consequence of optimisation difficulties.

Typical solutions on an example of 200 simulated observations (overlapping case):

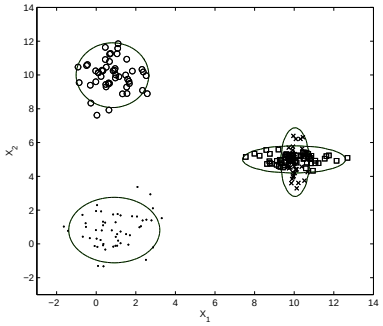

Fig. 2. BIC.

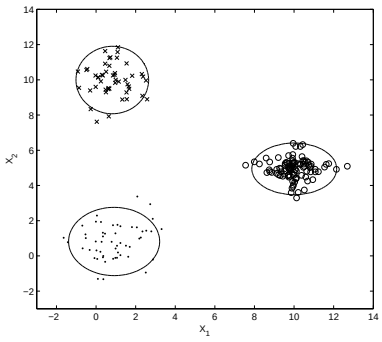

Fig. 3. ICL and slope heuristic (MccLE).

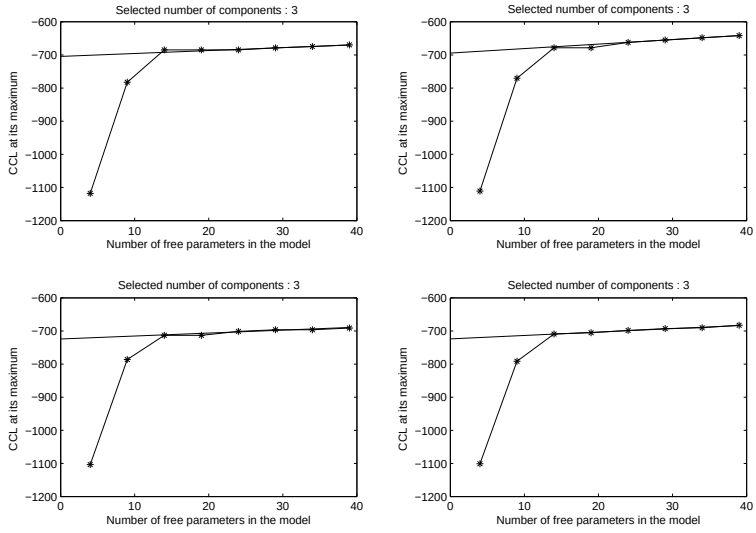

Fig. 4. A few examples of linear parts of the empirical contrast.

4.2 Experiments with biased models

Data arise from a four component Gaussian mixture whose covariance matrices are not diagonal: see Figures 5 or 6. But the fitted models are diagonal Gaussian mixture models.

Since the sampling distribution does not belong to considered models, BIC has a strong tendency to overestimate the complexity of the model because it is aiming at discovering the true distribution. On the contrary, ICL, taking into account the entropy, tends to select the right number of clusters. Here,

K	3	4	5	6	7	8
BIC	0	16	37	28	14	5
ICL	0	49	41	5	3	2
Slope heuristic (MccLE)	0	81	17	2	0	0

Table 2. Frequencies of resulting number of components for each criterion.

'Typical' solutions on an example of 200 simulated observations (biased case):

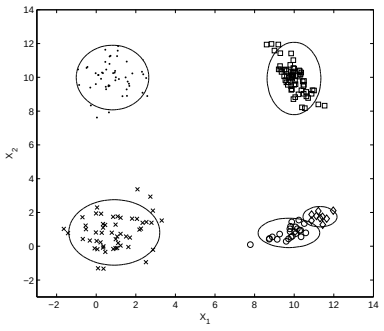

Fig. 5. BIC.

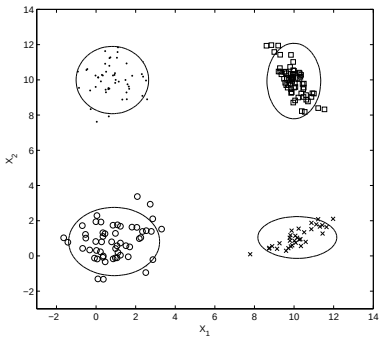

Fig. 6. ICL and slope heuristic (MccLE).

the slope heuristic has the best behavior in a clustering purpose, as it mostly selects the four clusters solution. See Table 2.

4.3 Experiments with a disturbing component

Data arise from a Gaussian mixture with three big components and a smaller disturbing component with a small proportion and attached to one of them. See Figure 7. We fitted diagonal mixture models to this data set.

K	3	4	5	6	7	8
BIC	42	57	0	0	0	1
ICL	93	7	0	0	0	0
Slope heuristic (MccLE)	78	17	4	1	0	0

Table 3. Frequencies of resulting number of components for each criterion.

All criteria have a satisfactory behavior: BIC hesitates between three and four components, ICL clearly points out three clusters and SH(MccLE) has an intermediate position between BIC and ICL. It expresses well that the oracle achieves close values for three and four components (results not reported here).

'Typical' solutions on an example of 200 simulated observations (biased case):

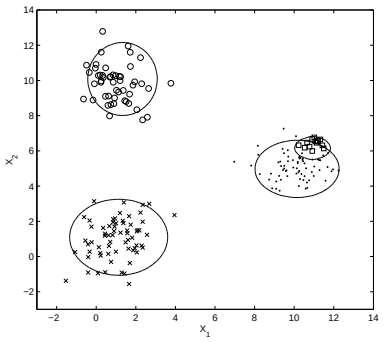

Fig. 7. BIC.

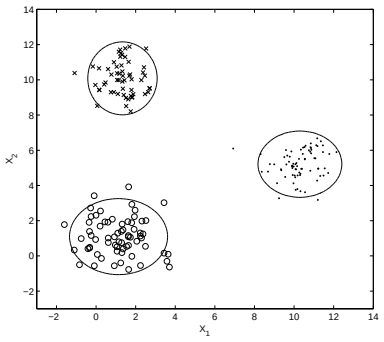

Fig. 8. ICL and slope heuristic (MccLE).

5 Discussion

We have exhibited two promising criteria, ICL and SH(MccLE) (Slope Heuristic on the conditional expectation of the complete likelihood) to assess the number of clusters in a model-based clustering framework. ICL is a BIC-like criterion, designed on a heuristic ground. The framework in which SH(MccLE) has been designed, namely penalized model selection for contrast minimization, seems to be more convenient in the purpose of a theoretical study. From a practical point of view, both criteria seem to behave analogously. Nevertheless, ICL is much easier to compute and, up to now, computing SH(MccLE) can involve numerical difficulties. On the basis of our experience, we claim that when the deviation of the model family from the true distribution is negligible, ICL should be preferred to SH(MccLE). But when the bias is negligible for no model, SH(MccLE) should be preferred to ICL. And in the real world, all models are wrong, therefore SH(MccLE) could be expected to be useful...

References

BIERNACKI C., CELEUX, G. and GOVAERT, G. (2000): Assessing a mixture model for clustering with the integrated complete likelihood. *IEEE Trans. on PAMI 22, 719-725.*

BURNHAM, K.P. and ANDERSON, D.R. (1998): *Model selection and inference.* Springer-Verlag, New York.

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005): *Inference in hidden Markov models.* Springer-Verlag, New York.

CROWDER, M. (2001): *Classical competing risks.* Chapman & Hall/CRC Press, London.

HATHAWAY, R. J. (1986): Another interpretation of the EM algorithm for mixture distributions. *Statistics and & Probability Letters 4, 53-56.*

LANGE, K. (1999): *Numerical analysis for statisticians* Springer-Verlag, New York.
MASSART, P. (2007): *Concentration inequalities and model selection*. Lecture Notes in Mathematics, Springer-Verlag, New York.
McLACHLAN, G.J. and PEEL, D. (2000): *Finite mixture models*. Wiley, New York.
RAFTERY, A.E. (1995): Bayesian model selection in social research (with discussion). *Sociological Methodology, 25, 111-196* .
ROEDER, K. and WASSERMAN, L. (1997): Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association 92, 894-902*.
ROUSSEEUW, R.J. and LEROY, A.M. (2003): *Robust regression and outlier detection*. Wiley, New York.
SCHWARZ, G. (1978): Estimating the dimension of a model. *The Annals of Statistics 6, 461-464*.

A Regression Subset-Selection Strategy for Fat-Structure Data

Cristian Gatu[1,2], Marko Sysi-Aho[1], and Matej Orešič[1]

[1] VTT Technical Research Centre of Finland
Tietotie 2, Espoo, P.O. Box 1000, FI-02044 VTT, Finland,
cristian.gatu@unine.ch, marko.sysi-aho@vtt.fi, matej.oresic@vtt.fi.
[2] Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi
16, General Berthelot Str., 700483 Iasi, Romania.

Abstract. A strategy is proposed for finding the most significant linear regression submodel for fat-structure data, that is when the number of variables n exceeds the number of available observations m. The method consists of two stages. First, a heuristic is employed to preselect a number of variables n_S such that $n_S \leq m$. The second stage performs an exhaustive search on the reduced list of variables. It employs a regression tree structure that generates all possible subset models. Non-optimal subtrees are pruned using a branch-and-bound device. Cross validation experiments on a real biomedical dataset are presented and analyzed.

Keywords: regression tree, branch-and-bound, model selection, fat-structure data

1 Introduction

An important problem in statistical modeling is that of computing the best-subset regression models or, equivalently, finding the best regression equation (Hastie et al. (2001)). Given a list of possible variables to be included in the regression model, the aim is to identify a subset that optimizes some statistical criterion. Most of the criteria used to evaluate the subsets rely on the residual sum of squares (Searle (1971), Sen and Srivastava (1990)). In the case of the standard regression model with n parameters there are $2^n - 1$ possible submodels that have to be evaluated and compared.

Forward, backward and stepwise procedures based on adding and/or deleting variables with respect to specific criteria can be employed in regression subset selection. However, these methods search few combinations of variables and rarely succeed in finding the best model (Hocking (1976), Seber (1977)). More enhanced procedures include ridge regression, the non-negative garrote, the least angle regression and the least shrinkage and selection operator (Breiman (1995), Efron et al. (2004), Fan and Li (2001), Tibshirani (1996)). These methods employ a form of automatic variable selection and shrinkage with the aim of identifying a parsimonious submodel that has good prediction ability. Another common approach to subset selection is the exhaustive

computation of the best-subset regression models. This can be achieved by enumerating and evaluating all possible submodels (Edwards and Havranek (1987), Hocking (1983), Miller (2002), Sen and Srivastava (1990), Gatu and Kontoghiorghes (2006a)). The restriction of this approach is that the number of variables to choose from should not be too large. The advantage of such an exhaustive search is that it is guaranteed to yield the optimum solution.

Here, the special case where the number of variables exceeds the number of available observations (hereafter called fat-structure data), is considered. A two step optimization strategy is proposed. This combines the heuristic approach and the exhaustive search aiming at saving computational time and obtaining quality solutions, respectively.

The structure of the paper is as follows. In the next section the biomedical data motivating the work is described. Section 3 gives pointers to a previously introduced exhaustive search method for subset selection. It is based on a regression tree structure and employs a branch-and-bound device. The new strategy for subset selection for fat-structure data is described in Section 4 together with a cross validation procedure for choosing the best submodels. Experiments on the biomedical data are presented and analyzed in Section 5. Finally, Section 6 provides conclusions.

2 Biomedical data

The dataset used in this paper was generated using a mass spectrometry based lipidomics platform (Orešič et al. (2006)), applied to adipose tissue biopsies of 44 subjects with varying degrees of obesity, and the corresponding Body Mass Index (BMI). The BMI is defined as the individual's body weight divided by the square of their height. The formulas universally used in medicine produce a unit of measure of kg/m^2. The BMI provides a simple numeric measure of a person's "fatness" or "thinness", allowing health professionals to discuss over- and under-weight problems more objectively with their patients.

A total of 333 lipid metabolites were detected using the UPLC/MS lipidomics platform. Raw data was processed using MZmine v0.60 software (Katajamaa et al. (2006)), and the lipids were identified using an internal library as previously described (Yetukuri et al. (2007)). Thus, the design of the study enables the investigation of the changes in the adipose tissue lipid profile in the context of other clinical variables (*e.g.*, BMI). The regression analysis may therefore reveal which variables (*i.e.*, lipids) best describe the observed changes in clinical variables associated with obesity. Inferences on potential biological mechanisms associated with changes in clinical variables can be made by knowing the lipid changes and their identities.

3 Exhaustive methods for regression subset selection

Consider the standard regression model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim (0, \sigma^2 I_m), \tag{1}$$

where $y \in \Re^m$ is the dependent variable vector, $X \in \Re^{m \times n}$ is the exogenous data matrix, $\beta \in \Re^n$ is the coefficient vector and $\varepsilon \in \Re^n$ is the noise vector. The columns of X correspond to the exogenous variables $V = \{v_1, \ldots, v_n\}$. A submodel S of (1) comprises some of the variables in V. The problem of finding the best-subset regression model can be formulated as optimizing a statistical criterion which is a function of the residual sum of squares and the dimensions of the model. This is equivalent in finding the best submodels corresponding to each model size, *i.e.*, in computing

$$\underset{S \subseteq V}{\operatorname{argmin}} \operatorname{RSS}(S) \quad \text{subject to} \quad |S| = k, \qquad \text{for} \quad k = 1, \ldots, n, \tag{2}$$

where $\operatorname{RSS}(S)$ denotes the residual sum of squares of the submodel comprising only the variables in S.

One approach in solving (2) is the straight-forward method of generating all $2^n - 1$ possible submodels (Hocking (1976), LaMotte and Hocking (1970), Miller (1984)). As n increases the number of submodels to be computed increases exponentially. The dropping columns algorithm (DCA) derives all submodels by generating a regression tree (Clarke (1981), Smith and Bremner (1989), Gatu and Kontoghiorghes (2003)). The parallelization of the DCA moderately improves its practical value (Gatu and Kontoghiorghes (2003)).

A computationally efficient branch-and-bound algorithm (BBA) has been devised (Gatu and Kontoghiorghes (2006a), Gatu et al. (2007), Hofmann et al. (2007)). The BBA provides the solution of (2) and avoids the computation of the whole regression tree generated by the DCA. Specifically, it employs a cutting test to prune non-optimal subtrees when searching for the best submodels. The algorithm was built on the fundamental property

$$\text{if} \quad S_1 \subseteq S_2 \quad \text{then} \quad \operatorname{RSS}(S_1) \geq \operatorname{RSS}(S_2). \tag{3}$$

That is, deleting variables from a regression increases the residual sum of squares of the resulting submodel (Gatu and Kontoghiorghes (2006a)). The BBA-1 which is an improved version of the BBA, preorders the n variables according to their strength in the root node, *i.e.* prior the generation of the DCA regression tree. The ith and jth variables are sorted such that $\operatorname{RSS}(V_{-i}) \geq \operatorname{RSS}(V_{-j})$ for each $1 \leq i \leq j \leq n$, where V_{-i} is the set V from which the ith variable has been deleted. The BBA-1 has been shown to outperform computationally the previously introduced leaps-and-bounds algorithm (Furnival and Wilson (1974), Gatu and Kontoghiorghes (2006a), Hofmann et al. (2007)).

The BBA-1 will be used in the subsequent to perform exhaustive subset selection and will be referred to as best-subset search. Details on the BBA and the BBA-1 can be found in Gatu and Kontoghiorghes (2006a) and Hofmann et al. (2007).

4 New heuristic strategy for subset selection

The particular constraint considered here is the fat structure. Specifically, the data contains measurements on 333 variables (lipid metabolites), from which part of them (173) are discarded based on biological arguments. Still, 160 variables remain to be investigated while only 44 observations are available.

Here, a two-stage procedure that aims to identify the best-subset regression model is proposed. The aim is to find a good quality solution in a reasonable computing time. The first stage reduces the number of variables in the original set. Specifically, it employs a heuristic algorithm to select $n_S \leq m$ most promising variables out of n. The heuristic could be a know method such as forward, backward, stepwise, least angle regression, etc., or a new purposely designed procedure. The main constraint in choosing the heuristic is the available computing time. The selected variables are used in the second stage to perform an exhaustive best-subset search. The procedure is summarized in Algorithm 1.

Algorithm 1. Subset regression selection for fat-structure data

1: **procedure** select(y, X, n_S)
2: Use a heuristic to preselect n_S variables $V_S \subseteq V$.
3: Build X_S with the columns from X that correspond to V_S.
4: Perform best-subset search on the model (y, X_S).
5: **end procedure**

The input of the "select" procedure is the original data of form (y, X) and n_S, the number of variables to be preselected by the heuristic. The latter is performed in line 2 of Algorithm 1. Once the set of the most promising variables V_S are preselected a new submodel (y, X_S) is built. The matrix X_S contains the columns of X that corresponds to the elements of V_S. The best-subset search is performed on (y, X_S) in line 4 of Algorithm 1 by calling BBA-1 (Gatu and Kontoghiorghes (2006a)). The output of the procedure is a list containing the best submodels found for each model size from 1 to n_S, *i.e.*, the solution of problem (2), where now $n = n_S$ and $V = V_S$.

It is important to know how well a selected model does generalize to new data. Since the number of variables (160) is much larger than the number of observations (44), it is possible to select a linear model that perfectly fits to the sample, giving zero training error. That is, the residual error from the observations into which the model is fitted. However, the model with zero

training error is over-fitted to the training data and will typically generalize poorly. In general, if the complexity of the model is increased then its bias decreases and its variance increases. One widely used measure for the generalization error is

$$\mathrm{ERR} = E\Big[\sum_{i=1}^{m}\big(y_i - f(X_i)\big)^2\Big], \tag{4}$$

where method f is applied to an independent test sample from the joint distribution of y and X.

The generalization error provides a convenient criterion for comparing different models, even if they do not belong to the same functional class. If the generalization error of model f_1 is smaller than that of the model f_2, then it is clear that one should prefer using model f_1 for prediction purposes. Thus, it is reasonable for the model selection to search the minimum of the generalization error.

In practice, a major challenge with the use of generalization error is the difficulty of its estimation. Several methods have been proposed for this purpose (Hastie et al. (2001)), but probably the simplest and most widely used method for estimating prediction error is cross-validation (CV). This method directly estimates the generalization error (4). In cross-validation, part of the data is used to fit the model and a different part is used to test the fitted model. The selection of the training and testing data can be implemented in several ways. Random subset selection of the observations have been used in order to modulate the effect of chance on affecting the estimate. Specifically, from the m available observations, the m_1 are randomly selected for training and the remaining $m - m_1$ are used for testing. This random selection is then repeated T times and can be written as

$$(y \;\; X) \equiv \begin{pmatrix} y_1^{(i)} & X_1^{(i)} \\ y_2^{(i)} & X_2^{(i)} \end{pmatrix} \begin{matrix} m_1 \\ m - m_1 \end{matrix}, \qquad \text{where} \quad i = 1, \cdots, T. \tag{5}$$

The estimated generalization error is then computed by

$$\mathrm{CV}(k) = \frac{1}{T}\sum_{i=1}^{T}\big(y_2^{(i)} - f_k^{(i)}(X_2^{(i)})\big)^T\big(y_2^{(i)} - f_k^{(i)}(X_2^{(i)})\big), \tag{6}$$

where $k = 1, \cdots, m_1$ indexes the model size and $f_k^{(i)}$ denotes the best submodel of size k found in the ith step of the CV procedure. Algorithm 2 summarizes this method.

Algorithm 2. Cross-validation procedure for subset selection

1: **procedure** cv(y, X, T, m_1)
2: **for** $i = 1, \cdots, T$ **do**

3: Use (5) to select the training and the testing sample.
4: **call** select($y_1^{(i)}$, $X_1^{(i)}$, m_1).
5: Store $f_1^{(i)}, \cdots, f_{m_1}^{(i)}$, the best submodels of size $1, \cdots, m_1$.
6: **end for**
7: Use (6) to compute CV(k), for $k = 1, \cdots, m_1$.
8: **end procedure**

5 Experimental results

The algorithms were implemented as C++ shared libraries and subsequently used in the R statistical software environment (R Development Core Team (2005)). The GNU compiler collection was used to generate the libraries. The tests were run on a Pentium-class machine with 1 Gb RAM in a Linux environment. An intercept term was included in all submodels.

In the cross validation procedure 34 and 10 observations were used as training and testing data, respectively. The number of runs T was set to 1000. In the first stage 32 variables out of 160 are preselected using as heuristic the least angle regression method (Efron et al. (2004)). The best-subset search is performed on the model which has 32 variables and an intercept. A list of best submodels of size $1, \cdots, 32$ is obtained in each of the T runs.

The left-hand side of Figure 1 plots the 1000 run median value of the sum of squared errors (SSE) computed on the test data when using the best submodels corresponding to each model size $1, \cdots, 32$. This is given by the continuous curve. The dotted curves bound the 0.5 – 0.95 confidence interval. The plot suggests 2 as optimum model size.

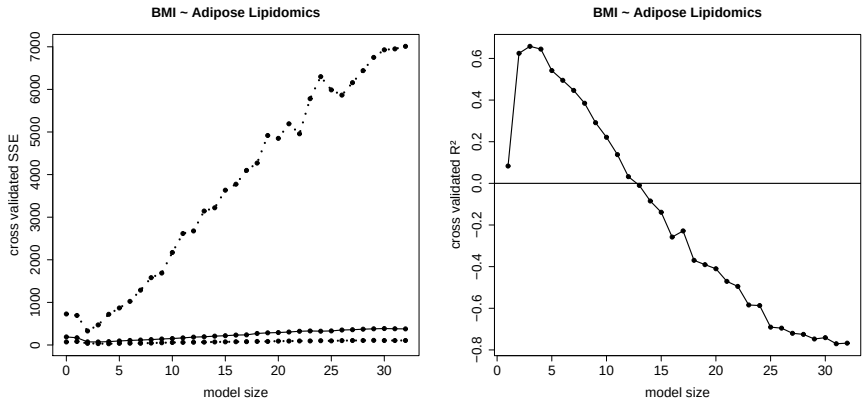

Fig. 1. The 1000 run median values of the cross validated SSE and R^2 for the (BMI, Adipose Lipidomics) data.

The right-hand side of Figure 1 plots the 1000 run median value of the cross validated R^2 (CV-R^2) for each model size. Its optimum (maximum) value is one, and can be negative if the submodel obtained using the training data fits the testing data worst than the model having only an intercept. The plot indicates that the best submodel has 2, 3 or 4 variables. The simplest model with 2 variables is retained.

The cross validation procedure outputs a list of 1000 best submodel for the retained optimum size 2. An important issue arising is the stability of the best model of size 2 over the runs. Figure 2 shows the 1000 run selection frequencies for the two variables. In 951 runs the variable 145 has been selected as one of the variables. As second variable, 30 and 27 have been selected in 480 and 351 runs, respectively. The latter variables are highly correlated and correspond to similar biological measurements.

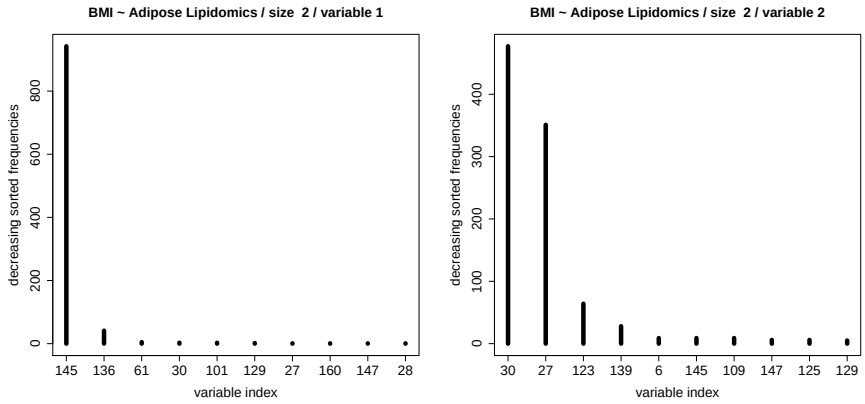

Fig. 2. The 1000 run selection frequencies of the variables of the size 2 best submodels for the (BMI, Adipose Lipidomics) data.

Figure 3 plots the values of the observed testing values of the BMI against the estimated ones, when using the submodels built with the variables (145, 30) and (145, 27). Over the 1000 runs these submodels have been selected as the best one in 480 and 351 runs, respectively. The median R^2 value of training data and the CV-R^2 of the testing data are also displayed.

6 Conclusions

An optimization strategy that aims to identify the best-subset regression models was proposed. The special case of fat-structure data was considered. The new approach consists of two stages that combine the heuristic and exhaustive search aiming to reduce the computation time and to obtain quality

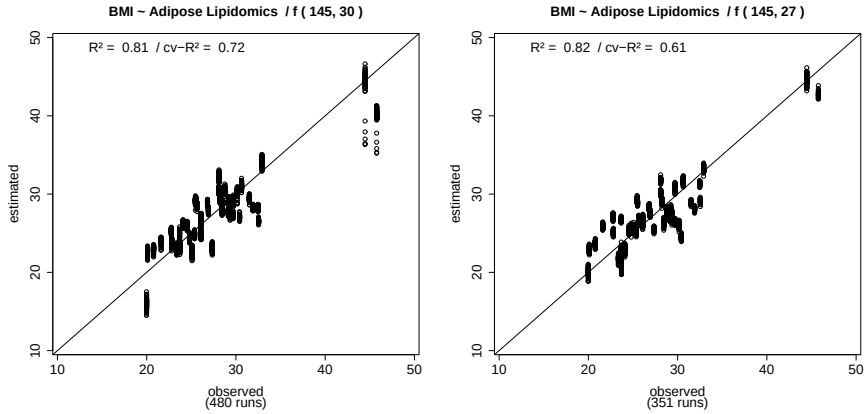

Fig. 3. The 1000 runs cross validation fit of the best submodels. The observed against the estimated values of the testing output variable.

solutions, respectively. In the first stage a heuristic was employed to pre-select the $n_S \leq m$ variables from the original list. The selected variables depend on the choice of the heuristic. The second stage — which is the core of the proposed strategy — performs an exhaustive best-subset search on the reduced list of variables. For this purpose a previously introduced branch-and-bound algorithm was employed. The algorithm is based on a regression tree strategy that generates all possible submodels of a given set. During the exhaustive search, non-optimal subtrees are pruned using a cutting test (Gatu and Kontoghiorghes (2006a), Gatu et al. (2007), Hofmann et al. (2007)). The new strategy outputs a list of best submodels corresponding to each model size $1, \cdots, n_S$.

In order to improve the selected submodels the cross validation method was employed. In each cross validation run a list of best subset models were generated. Experiments on real biomedical data showed that highly sparse submodels with a good prediction accuracy can be identified. The method was compared on the same data with other established methods such as LASSO, LARS and ElasticNet (Efron et al. (2004), Hui and Trevor (2005), Tibshirani (1996)). The same random splits of training and testing data were used in the cross-validation test. None of these methods obtained a stable solution. Over the 1000 runs, the obtained submodels were highly dependent on the data split. Furthermore, for a given split, the submodels had worse fit than the models obtained by our method.

The proposed strategy promises to be an effective statistical model selection method for fat-structure data, which may replace or complement the existing methods. Furthermore, it can be extended and adapted to deal with more complex models such as the general linear, the vector autoregressive

and seemingly unrelated regression models (Gatu and Kontoghiorghes (2005 and 2006b), Gatu et al. (2008)).

Acknowledgements

The work of the first author was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The authors thank Tuulikki Seppänen-Laakso, Kirsi Pietiläinen, Hannele Yki-Järvinen, Jaakko Kaprio and Aila Rissanen for collaboration in the biomedical project that led to the data utilized in this paper.

References

BREIMAN, L. (1995): Better subset regression using Nonnegative Garrote. *Technometrics, 37 (4), 373-383.*

CLARKE, M.R.B. (1981): Algorithm AS163. A Givens algorithm for moving from one linear model to another without going back to the data. *Applied Statistics, 30 (2), 198-203.*

EDWARDS, D. and HAVRANEK, T. (1987): A fast model selection procedure for large families of models. *Journal of the American Statistical Association, 2 (397), 205-213.*

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004): Least angle regression. *The Annals of Statistics, 32 (2), 407-499.*

FAN, J. and LI, R. (2001): Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96 (456), 1348-1360.*

FURNIVAL, G. and WILSON, R. (1974): Regression by leaps and bounds. *Technometrics, 16, 499-511.*

GATU, C. and KONTOGHIORGHES, E.J. (2003): Parallel algorithms for computing all possible subset regression models using the QR decomposition. *Parallel Computing, 29 (4), 505-521.*

GATU, C. and KONTOGHIORGHES, E.J. (2005): Efficient strategies for deriving the subset VAR models. *Computational Management Science, 4, 253-278.*

GATU, C. and KONTOGHIORGHES, E.J. (2006a): Branch-and-bound algorithms for computing the best subset regression models. *Journal of Computational and Graphical Statistics, 15, 139-156.*

GATU, C. and KONTOGHIORGHES, E.J. (2006b): Estimating all possible SUR models with permuted exogenous data matrices derived from a VAR process. *Journal of Economic Dynamics and Control, 30, 721-739.*

GATU, C. and KONTOGHIORGHES, E.J. (2008): An efficient branch-and-bound strategy for Subset Vector Autoregressive model selection. *Journal of Economic Dynamics and Control, Forthcoming.*

GATU, C., YANEV, P. I. and KONTOGHIORGHES, E.J. (2007): A graph approach to generate all possible regression submodels. *Computational Statistics and Data Analysis, 52, 799-815.*

HASTIE, T.J., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer-Verlag, New York.

HOCKING, R.R. (1976): The analysis and selection of variables in linear regression. *Biometrics, 32, 1-49*.

HOCKING, R.R. (1983): Developments in linear regression methodology: 1959-1982. *Technometrics, 25 (3), 219-230*.

HOFMANN, M., GATU, C. and KONTOGHIORGHES, E.J. (2007): Efficient algorithms for computing the best-subset regression models for large scale problems. *Computational Statistics and Data Analysis, 52, 16-29*.

HUI, Z. and TREVOR, H. (2005): Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67 (2), 301-320*.

KATAJAMAA, M., MIETTINEN, J. and OREŠIČ, M. (2006): MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics, 22, 634-636*.

LAMOTTE, L.R. and HOCKING, R.R. (1970): Computational efficiency in the selection of regression variables. *Technometrics, 12 (1), 83-93*.

MILLER, A.J. (1984): Selection of subsets of regression variables. *Journal of the Royal Statistical Society, 147, 389-425*.

MILLER, A.J. (2002): *Subset selection in regression*. Chapman and Hall, second edition.

OREŠIČ, M., VIDAL-PUIG, A. and HÄNNINEN, V. (2006): Metabolomics approaches to phenotype characterization and applications to complex diseases. *Expert Rev. Mol. Diagn., 6, 575-585*.

R DEVELOPMENT CORE TEAM (2005): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SEARLE, S.R. (1971): *Linear models*. John Wiley, New York.

SEBER, G.A.F. (1977): *Linear regression analysis*. John Wiley, New York.

SEN, A. and SRIVASTAVA, M. (1990): *Regression analysis. Theory, methods and applications*. Springer.

SMITH, D.M. and BREMNER, J.M. (1989): All possible subset regressions using the QR decomposition. *Computational Statistics and Data Analysis, 7, 217-235*.

TIBSHIRANI, R. (1996): Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B, 58 (1), 267-288*.

YETUKURI, L., KATAJAMAA, M., MEDINA-GOMEZ, G., SEPPÄNEN-LAAKSO, T., VIDAL-PUIG, A. and OREŠIČ, M. (2007): Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. *BMC Syst. Biol., 1, e12*.

Fast Robust Variable Selection

Stefan Van Aelst[1], Jafar A. Khan[2], and Ruben H. Zamar[3]

[1] Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 S9, B-9000 Gent, Belgium, *Stefan.VanAelst@UGent.be*
[2] Department of Statistics, University of Dhaka
Dhaka-1000, Bangladesh, *jkhan66@gmail.com*
[3] Department of Statistics, University of British Columbia
Vancouver, British Columbia, V6T-1Z2, Canada *ruben@stat.ubc.ca*

Abstract. We discuss some computationally efficient procedures for robust variable selection in linear regression. A key component in these procedures is the computation of robust correlations between pairs of variables. We show that the robust variable selection procedures can easily handle missing data under the assumption that data are missing completely at random.

Keywords: correlation, missing data, robustness, variable selection

1 Introduction

Variable selection for regression in large data sets with a large number of candidate predictors is a challenging task that requires computationally efficient algorithms. Two issues that often further complicate this task are (i) that large data sets are often of uneven quality and thus may contain outliers and other anomalies and (ii) large data sets often contain missing values. Deleting all observations with missing values (complete-case analysis) can lead to a huge loss of information in high dimensional data sets. Therefore, good variable selection procedures need to be computationally efficient and robust and be able to handle missing data such that observations with missing values do not need to be deleted.

Robust model selection for linear regression so far has mainly focused on the development of robust selection criteria that can be used to compare models. See e.g. Ronchetti (1997) for robust versions of the selection criteria AIC and C_p, respectively. Maronna et al. (2006, p. 151) proposed a robust Final Prediction Error (FPE) criterion. Ronchetti, Field, and Blanchard (1997) proposed robust model selection by cross-validation while Müller and Welsh (2005) and Salibian-Barrera and Van Aelst (2007) considered robust selection criteria based on bootstrap procedures. However, most of these papers do not propose any strategy to select the set of models that are to be compared, and often suggest using the time-consuming all-subsets approach. For least squares regression, time-efficient techniques for all-subsets selection exist (see e.g. Furnival and Wilson (1974), Gatu and Kontoghiorghes (2006),

Hofmann et al. (2007)). However, it does not seem to be straightforward to extend these techniques beyond least squares.

Two major drawbacks of most robust model selection methods are that (i) they make variable selection very time consuming, as they require the calculation of a time demanding robust fit for a large number of submodels and (ii) they have no way to handle missing values beyond deleting cases with missing values. To address the first drawback, Khan et al. (2007a,b) recently proposed computationally efficient procedures for robust variable selection that can handle a large number d of possible predictors - e.g. several hundreds of candidate predictors. Note that in such cases a robust fit of the 'full model' may not be feasible anymore due to the numerical complexity of robust estimates when d is very large (e.g. $d \geq 200$) or simply because d exceeds the number of cases, n.

The procedures of Khan et al. (2007a,b) focus on sequencing the candidate predictors in order of importance. Standard methods to sequence the candidate predictors are step-by-step algorithms such as forward (FS) or stepwise (SW) selection (see, e.g. Weisberg (1985), chap. 8). Note that these algorithms are agressive in the sense that they include or exclude predictors completely. Selected variables are included completely as their contribution is determined by the least squares estimate of their regression coefficient. This greedy approach may prevent important predictors that are correlated with predictors already included in the model to enter the model. Efron et al. (2004) recently proposed Least angle regression (LARS) which is a powerful and computationally very efficient technique to sequence candidate predictors that is less agressive than FS or SW. At each step LARS selects the predictor that has the largest correlation with the current residual. This predictor is included in the model with the size of its coefficient determined such that the updated residuals have a correlation with the predictor that is equal to the maximal correlation between the updated residuals and the not yet selected predictors. Note that also the coefficients of all other predictors already in the model are updated such that the correlation of the updated residual and each predictor in the model is equal to the maximal correlation between the updated residuals and the not yet selected predictors. At this moment the next predictor enters the model and all regression coefficients are updated again. The 'partial' including of predictors in the model in LARS allows important predictors correlated with already selected predictors to enter the model as well if they contribute sufficiently in explaining the response.

The sequencing procedures FS, SW and LARS are all based on classical correlations among variables. Unfortunately, this makes them extremely sensitive to outliers and thus these procedures yield poor results when the data are contaminated. These algorithms attempt to select the covariates that will fit well all the cases (including the outliers), and often fail to select the model that would have been chosen if those outliers were not present in the data. Moreover, even if it would be possible to detect outliers beforehand, similar

as for observations with missing values, aggressive deletion of outliers is not desirable, because we may end up deleting a lot of observations which are outliers only with respect to predictors that will not be in the model. Therefore, Khan et al. (2007a,b) developed robust alternatives to the standard FS, SW and LARS procedures that yield reliable results if the data are contaminated, but at the same time mostly preserve the computational efficiency of the standard procedures. These procedures are not confined to continuous predictors but can handle mixtures of continuous and categorical predictors.

Note that in the prediction of future cases, outliers can occur as well. However, we argue that it is not reasonable to attempt to predict the future outliers without knowledge of the underlying mechanism that produces them. Therefore, our robust algorithms will select important variables in the presence of outliers, and predict well the future non-outlying cases.

In the next Section we shortly review the key parts of the algorithms for robust forward (RFS) and stepwise (RSW) selection and robust least angle regression (RLARS) as developed in Khan et al. (2007a,b). We explain how missing values can easily be handled with these algorithms under the assumption of data missing completely at random. Section 3 shows the results of a simulation study to investigate the effect of missing data on the performance of the algorithms. The conclusions are given in Section 4.

2 Robust variable selection procedures

In Khan et al. (2007a,b) it is shown that, once the data are standardized using the sample means and sample standard deviations, the FS, SW, and LARS algorithms can be expressed completely in terms of correlations of the original variables. This explains why they are very fast to compute but yield poor results when the data is contaminated. To strengthen the robustness properties of these procedures without affecting their computational efficiency too much, Khan et al. (2007a,b) robustly standardized the data and then replaced the nonrobust sample correlations by robust counterparts.

Note that affine equivariance and regression equivariance are generally considered to be important properties for (robust) regression estimators. However, these properties are not required in the context of variable selection, because we do not consider general linear transformations of the given covariates. The only transformations that should not affect the selection result are linear transformations of individual variables, i.e., shifts and scale changes. Hence, variable selection methods such as FS, SW and LARS are based on correlations among the variables. Therefore, robust variable selection procedures need to be robust against correlation outliers, that is, outliers that affect the classical correlation estimates but can not be detected by looking at the individual variables separately. The RFS, RSW, and RLARS procedures are based on robust correlation estimates. Hence, they are robust against correlation outliers and thus suitable for robust variable selection.

Note that the robust variable selection algorithms RSF, RSW, and RLARS only handle the problem of selecting a list of important predictors, but we do not yet fit the selected model. The final model(s) resulting from the selection procedure usually contains only a small number of predictors compared to the initial dimension d, when d is large. Therefore, to robustly fit the final model, we can use a highly robust regression estimator such as an MM-estimator (Yohai (1987)) that is resistant to all types of outliers. Note that we always use models with intercept.

To robustly standardize the data, straightforward choices for the measures of center and scale are the highly robust and rapidly computable median (med) and median absolute deviation (mad). Note that to avoid singularities, categorical predictors are standardized using their sample mean and sample variance. Obtaining good robust estimates of correlation among the variables is less straightforward. For d-dimensional datasets, robust estimates of correlation are often derived from affine-equivariant, robust estimators of scatter. However, this is very time-consuming, particularly for large values of d. Moreover, the computation of such robust correlation matrices becomes unstable when the dimension d is large compared to the sample size n. On the other hand, only a few of the d covariates are typically included in the final model, and the computation of the whole d-dimensional correlation matrix at once will unnecessarily increase the numerical complexity of the otherwise computationally efficient algorithms. Therefore, we resort to robust approaches that calculate pairwise correlations one at the time. This pairwise approach for robust correlation estimation is not only computationally suitable and provides robustness against bivariate correlation outliers, but it is also more convenient (compared to the full d-dimensional approach) for robust variable selection because it allows us to compute only the required correlations at each step of the algorithm.

A computationally efficient, robust estimator of bivariate correlation can be derived from an affine-equivariant bivariate M-estimator of scatter as defined by Maronna (1976). For robustly standardized data $\mathbf{x}_i = (x_{i1}, x_{i2})^t$; $i = 1, \ldots, n$, the bivariate M-estimator of scatter is defined as the solution $\mathbf{V}$ of

$$\frac{1}{n} \sum_i u_2(d_i^2) \mathbf{x}_i \mathbf{x}_i^t = \mathbf{V},$$

where $d_i^2 = \mathbf{x}_i^t \mathbf{V}^{-1} \mathbf{x}_i$, and the function $u_2(t) = \min(c/t, 1)$. The constant c controls the robustness and efficiency of the M-estimator. Points at distance from the center (which is zero in our case) larger than c are downweighted. Therefore, a smaller value of c increases robustness in the sense that observations are downweighted sooner, but also the efficiency of the estimator is lower. A larger value of c yields a higher efficiency, but also less robustness because observations need to be further from the center before they get downweighted. Following Khan et al. (2007a), we use $c = 9.21$, the 99% quantile of a χ^2_2 distribution. The M-estimator of scatter is affine equivariant and has

breakdown point 1/3 in two dimensions (Maronna (1976)). Note that the breakdown point is the smallest fraction of contamination that can make the largest eigenvalue of the scatter estimator arbitrarily large or the smallest eigenvalue arbitrarily small. The M-estimator of scatter can be calculated using a standard iteratively reweighted least squares procedure starting from the identity matrix.

The stepwise procedure (SW) requires stopping rules to determine at each step of the algorithm whether a variable is added or deleted or the procedure stops. Also in forward selection a stopping rule is often used to decide how many variables will be selected. In the standard nonrobust procedures these stopping rules are often based on partial F-statistics and corresponding critical values of the corresponding F-distribution. Khan et al. (2007a) have shown that these partial F-statistics can also be completely written in terms of correlations. Hence, also robust values of these partial F-statistics can be obtained by using robust correlations in the calculation of these statistics. Moreover, in Khan et al. (2007a) it is argued that the critical values of the F-distribution are still reasonable critical values if the robust correlations are M-estimates of bivariate correlation with $c = 9.21$. To determine the number of important predictors when sequencing the variables with the RLARS procedure, Khan et al. (2007b) have proposed to use a 'learning curve'. Once the RLARS sequence has been obtained, the successive models (starting from a model with only one predictor) can be fitted using a robust estimator such as an MM-estimator. A robust measure of R^2 for each of these models can then be plotted against the size of the model. The size that corresponds with the point where the curve in this elbow plot levels off can then be used as an estimate of the number of important predictors.

For very large, high-dimensional data sets, Khan et al (2007b) introduced an even faster robust pairwise correlation estimator. This correlation estimator uses the principle of *bivariate Winsorization*, a generalization of the univariate Winsorization as introduced in Huber (1981), see also Alqallaf et al. (2002). A bivariate Winsorization of robustly standardized bivariate data is based on an initial robust bivariate correlation matrix $\mathbf{R}_0$ and corresponding tolerance ellipse. The outliers are shrunken to the border of this ellipse by using the bivariate transformation $\mathbf{u} = \min(\sqrt{c/d(\mathbf{x})},\ 1)\ \mathbf{x}$ with $\mathbf{x} = (x_1, x_2)^t$. Here $d(\mathbf{x}) = \sqrt{\mathbf{x}^t \mathbf{R}_0^{-1} \mathbf{x}}$ is the Mahalanobis distance of $\mathbf{x}$ based on the initial bivariate correlation matrix $\mathbf{R}_0$. The tuning constant c again controls the robustness of the procdure. Khan et al. (2007b) proposed to use $c = 5.99$, the 95% quantile of the χ^2_2 distribution. Figure 1 illustrates bivariate Winsorization. Bivariate Winsorization shrinks the outliers to the boundary of the ellipse. Correlation outliers are thus appropriately downweighted so that a robust correlation estimate is obtained. The bivariate Winsorized correlation estimate is the classical correlation estimate obtained from the bivariate Winsorized data.

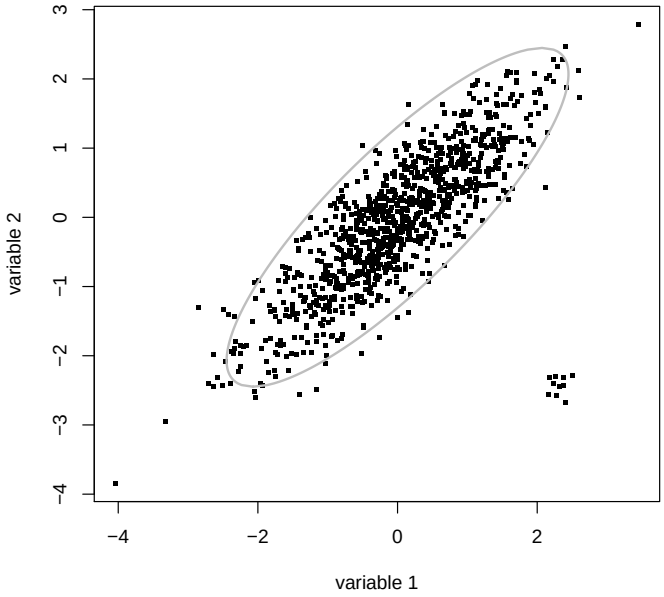

Fig. 1. Bivariate Winsorization tolerance ellipse which connects points of equal Mahalanobis distance (2.45) based on an initial correlation matrix $\mathbf{R}_0$.

Choosing an appropriate initial correlation matrix $\mathbf{R}_0$ is a crucial part of the bivariate Winsorization procedure. Khan et al (2007b) proposed a new method called adjusted Winsorization. This method applies univariate Winsorization to each of the two components of the robustly standardized bivariate data. That is, for each component, the observations $x_{1j}, \ldots, x_{nj}$, are transformated to $u_{ij} = \min(\max(-c, x_{ij}), c);\ i = 1, \ldots, n$, and $j = 1, 2$. However, two different values of the tuning constant c are used. The four quadrants relative to the center zero are considered. A larger tuning constant c_1 is used to Winsorize the points lying in the two diagonally opposed quadrants that contain the majority of the data points (called the "major quadrants") and a smaller tuning constant c_2 is used to Winsorize the remaining data (in the "minor quadrants"). Khan et al. (2007b) used $c_1 = 2$ and $c_2 = \sqrt{h}c_1$, where $h = n_2/n_1$ with n_1 the number of observations in the major quadrants and $n_2 = n - n_1$. The initial correlation matrix $\mathbf{R}_0$ is obtained by computing the classical correlation matrix of the adjusted Winsorized data. For the same data as in Figure 1, Figure 2 shows how adjusted Winsorization deals with bivariate outliers. The outliers are shrunken to the boundaries of the squares. By using a smaller tuning constant in the minor

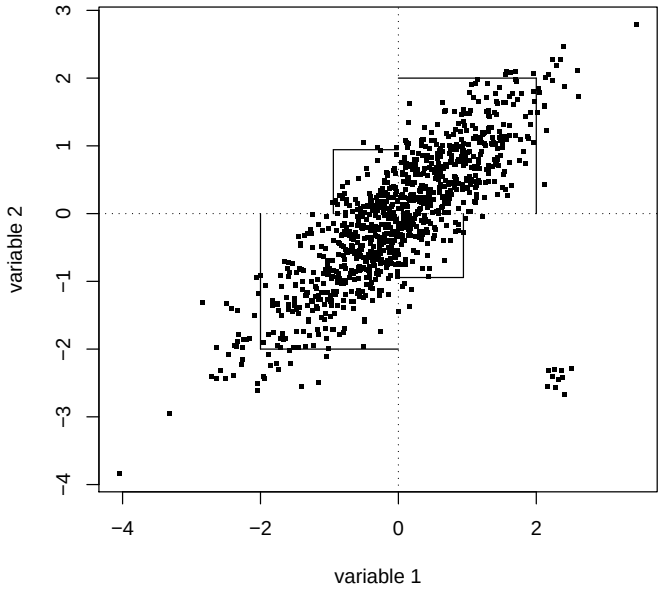

Fig. 2. Adjusted Winsorization to compute the initial robust correlation estimate $\mathbf{R}_0$ (with $c_1 = 2$ and $c_2 = \sqrt{h}c_1$). The outlying points are shrunken to the edges or corners of the squares.

constants, correlation outliers are sufficiently shrunken to obtain a robust correlation estimate.

If the data set contains missing values, then each robust pairwise correlation can be calculated from all observations that are complete for the variables under consideration (available-case analysis). This simple approach avoids the aggressive deletion of all observations with at least one missing component prior to the analysis. Available-case analysis is a valid approach under the assumption that the data are missing completely at random (MCAR) which means that the probability of missingness does not depend on the data (observed nor missing) (Little and Rubin (1987), Little (1992)). Hence, if the MCAR assumption is reasonable, by applying an available-case analysis we can avoid a huge loss of information if missing data are scattered throughout the data set.

To obtain more stable and reliable results the robust sequencing procedures can be combined with bootstrap (see Khan et al. (2007b)). This is a common approach in machine learning methods such as random forests (Breiman (2001), see also Hastie et al. (2001)). The bootstrap procedure works as follows. We generate B bootstrap samples from the original dataset,

and for each bootstrap sample we sequence the covariates using the robust sequencing procedure, i.e. RLARS, RFS, or RSW. This produces B sequences of the candidate predictors. For each covariate we then calculate its average rank in the B sequences to obtain the bootstrap sequence of the variables.

When dealing with high-dimensional datasets it may not be convenient or even possible to sequence all the covariates for each bootstrap sample. Note that the original sample would already be singular if the dimension d of the data exceeds the sample size. We easily overcome this problem by sequencing only the first $m < n$ covariates for each bootstrap sample. We then rank the covariates according to the number of times (out of B) they are actually sequenced. When ties occur, the order of the covariates is determined by their average rank in the sequences. The resulting procedures are denoted B-RLARS, B-RFS, and B-RSW, where the B naturally stands for 'bootstrap'.

3 Simulations

To investigate the behavior of our robust variable sequencing procedures and the effect of missing values, we consider the simulation setting of Khan et al. (2007b) which is based on the design of Frank and Friedman (1993). We first create a linear model

$$y = L_1 + L_2 + \cdots + L_k + \sigma\varepsilon,$$

with k latent variables, where $L_1, L_2, \ldots, L_k$ and ε are independent standard normal variables. The value of σ is chosen so that the signal to noise ratio is equal to 3. A set of d candidate predictors is created as follows. Let $e_1, ..., e_d$ be independent standard normal variables and let

$$\begin{aligned}
X_i &= L_i + \tau e_i, \qquad i = 1, ..., k \\
X_{k+1} &= L_1 + \delta e_{k+1} \\
X_{k+2} &= L_1 + \delta e_{k+2} \\
X_{k+3} &= L_2 + \delta e_{k+3} \\
X_{k+4} &= L_2 + \delta e_{k+3} \\
&\vdots \\
X_{3k-1} &= L_k + \delta e_{3k-1} \\
X_{3k} &= L_k + \delta e_{3k}
\end{aligned}$$

and

$$X_i = e_i \qquad i = 3k+1, ..., d$$

The constants $\delta = 5$ and $\tau = 0.3$ are chosen so that $\operatorname{corr}(X_1, X_{k+1}) = \operatorname{corr}(X_1, X_{k+2}) = \operatorname{corr}(X_2, X_{k+3}) = \cdots = \operatorname{corr}(X_k, X_{3k}) = 0.5$. Note that covariates $X_1, ..., X_k$ are "low noise" perturbations of the latent variables and constitute our "target covariates". Variables $X_{3k+1}, ..., X_d$ are independent

noise covariates and variables $X_{k+1}, ..., X_{3k}$ are noise covariates which are correlated with the target covariates.

To allow for a fraction ϵ of outliers we considered the following sampling distributions, listed in increasing order of difficulty:

(1) $\varepsilon \sim N(0,1)$, no contamination;
(2) $\varepsilon \sim (1-\epsilon)N(0,1) + \epsilon\, N(0,1)/\text{Uniform}(0,1)$, symmetric, slash contamination;
(3) $\varepsilon \sim (1-\epsilon)N(0,1) + \epsilon\, N(20,1)$, asymmetric, shifted normal contamination;
(4) same as (2), except that contaminated cases come along with high leverage X–values for all 50 candidate predictors (normal random variables with mean 50 and variance 1 in our simulation);
(5) same as (3), but with high leverage outliers, as described in (4).

We generated 100 independent samples of size $n = 150$ from the five simulation designs described above, with $k = 6$ latent variables and $d = 50$ candidate covariates and fraction of contamination $\epsilon = 10\%$. For each of these datasets we sequenced the variables using B-LARS and B-RFS (without stopping rule) using $B = 50$ bootstrap samples. We sequenced the first 25 predictors for each simulated data set. To study the effect of missing values on these procedures, we then introduced 10% of missing values (that is, 15 cases) at random in each of the 6 target predictors, as well as in the 12 correlated noise variables. Note that this implies that about 85% of the observations contains missing values, and thus a complete-case analysis would not be feasible due to the huge loss of information.

To summarize the simulation results, we determine for each sequence the number t_m of target variables included in the first m sequenced variables, with m ranging between 1 and 25. Figure 3 shows for each of the sampling situations the average (over the 100 datasets) of t_m for both B-RLARS and B-RFS and for data without and with missing values.

From Figure 3a we see that in this setting both procedures B-RLARS and B-RFS are equally effective in selecting the important predictors at the beginning of the sequence when there is no contamination in the data. Moreover, the large fraction of missing values does not affect the performance of the procedures. Figures 3b and c show that 10% of contamination in the response (symmetric or asymmetric) does not affect much the performance of B-RLARS. The missing values also do not have much inpact on its performance. For symmetric contamination (Figure 3b) B-RFS also performs well, but its performance is worse with asymmetric contamination (Figure 3c). Similarly as for B-RLARS, the missing values only have a small impact on the performance of B-RFS. From Figures 3d and e we see that 10% of leverage points (symmetric or asymmetric contamination) clearly affects the robust procedures to some extent. Asymmetric contamination has a larger effect than symmetric contamination. Note that Khan et al. (2007a,b) have shown

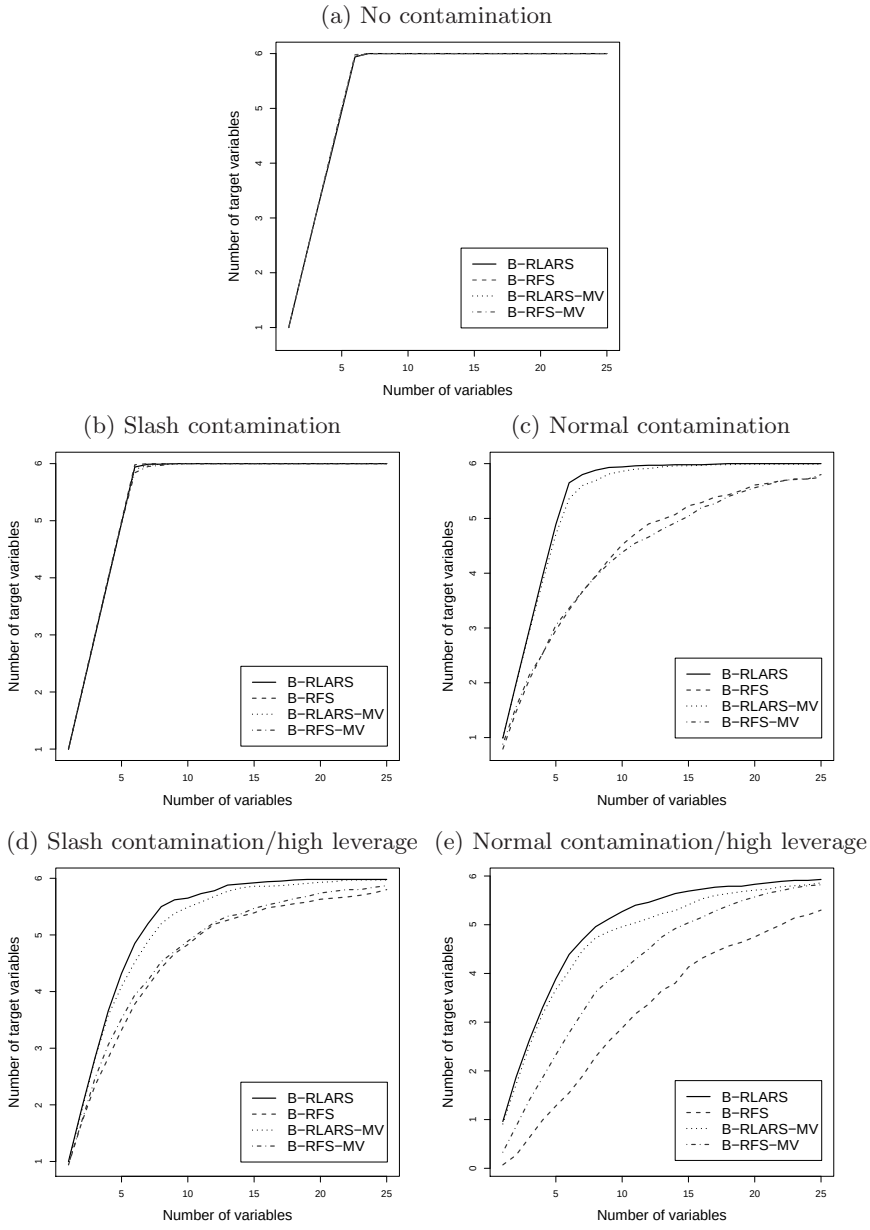

Fig. 3. Averages of the number of target variables t_m versus m for each of the methods and sampling situations considered. We generated datasets with $d = 50$ predictors, $k = 6$ latent variables, and 10% of contamination ($\epsilon = 0.1$).

that the effect on the robust procedures is much smaller than on the nonrobust selection procedures. Also with high leverage contamination, the effect of missing values on B-RLARS is small. Surprisingly, the performance of B-RFS is better for data with missing values than for data without missing values, the effect being larger for asymmetric leverage points. A possible explanation may be that with missing values the leverage points affect less the robust correlation estimates because part of the missing data are in the leverage points (and thus dropped when calculating robust correlations). Moreover the missing data occurs in the target variables and the correlated noise variables. Due to the missingness, the correlation estimates among these values might be lower, leading to a better performance for the greedy forward selection procedure.

4 Conclusions

We have shown that the robust sequencing procedures RLARS, RFS, and RSW can easily be extended to handle data with missing values under the assumption of data missing completely at random. For the design considered in the simulation study we can conclude that missing values have only a small effect on the performance of the bootstrapped version of RLARS while bootstrap RFS showed actually better performance for contaminated data with missing values than for contaminated data without missing values. Further research will focus on computationally efficient procedures to handle missing data in the context of variable selection under the more general assumption of data missing at random (that is, the missingness may depend on the observed data for the observation, but not on the missing data).

References

ALQALLAF, F.A., KONIS, K.P., MARTIN, R.D. and ZAMAR, R.H. (2002): Scalable robust covariance and correlation estimates for data mining. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, 14-23.

BREIMAN, L. (2001): Random Forests. *Machine Learning 24, 5-32*.

EFRON, B.E., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004): Least angle regression. *The Annals of Statistics 32 (2), 407-451*.

FRANK, I. and FRIEDMAN, J.H. (1993): A statistical view of some chemometrics regression tools. *Technometrics 35, 109-148*.

GATU, C. and KONTOGHIORGHES, E.J. (2006): Branch-and-bound algorithms for computing the best subset regression models. *Journal of Computational and Graphical Statistics 15, 139-156*.

FURNIVAL, G. and WILSON, R. (1974): Regression by leaps and bounds. *Technometrics 16, 499-511*.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The Elements of Statistical Learning*. Springer-Verlag, New York.

HOFMANN M., GATU C. and KONTOGHIORGHES, E.J. (2007): Efficient algorithms for computing the best subset regression models for large-scale problems. *Computational Statistics and Data Analysis 52 (1), 16-29.*

HUBER, P.J. (1981): *Robust Statistics.* John Wiley, New York.

KHAN, J.A., VAN AELST, S. and ZAMAR, R.H. (2007a): Building a robust linear model with forward selection and stepwise procedures. *Computational Statistics and Data Analysis 52 (1), 239-248.*

KHAN, J.A., VAN AELST, S. and ZAMAR, R.H. (2007b): Robust linear model selection based on least angle regression. *Journal of the American Statistical Association 102 (480), 1289-1299.*

LITTLE, R.J.A. (1992): Regression with missing X's: a review. *Journal of the American Statistical Association 87 (420), 1227-1237.*

LITTLE, R.J.A. and RUBIN, D.B. (1987): *Statistical Analysis with Missing Data.* John Wiley, New York.

MARONNA, R.A. (1976): Robust M-estimators of multivariate location and scatter. *The Annals of Statistics 4, 51-67.*

MARONNA, R.A., MARTIN, R.D. and YOHAI, V.J. (2006): *Robust Statistics: Theory and Methods.* John Wiley, Chichester.

MÜLLER, S. and WELSH, A.H. (2005): Outlier robust model selection in linear regression. *Journal of the American Statistical Association 100 (472), 1297-1310.*

RONCHETTI, E. (1997): Robustness aspects of model choice. *Statistica Sinica 7, 327-338.*

RONCHETTI, E., FIELD, C. and BLANCHARD, W. (1997): Robust linear model selection by cross-validation. *Journal of the American Statistical Association, 92, 1017-1023.*

SALIBIAN-BARRERA, M. and VAN AELST, S. (2007): Robust model selection using fast and robust bootstrap. submitted for publication.

WEISBERG, S. (1985): *Applied Linear Regression* (2nd ed.). Wiley-Interscience, New York.

YOHAI, V.J. (1987): High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics 15, 642-656.*

Part XI

Models for Latent Class Detection

Latent Classes of Objects and Variable Selection

Giuliano Galimberti, Angela Montanari, and Cinzia Viroli

Statistics Department, University of Bologna, Italy,
giuliano.galimberti@unibo.it, angela.montanari@unibo.it, cinzia.viroli@unibo.it

Abstract. In this paper we present a model based clustering approach which contextually performs dimension reduction and variable selection. In particular we assume that the data have been generated by a linear factor model with latent variables modeled as gaussian mixtures (thus obtaining dimension reduction) and we shrink the factor loadings, resorting to a penalized likelihood method, with an L1 penalty (thus realizing automatic variable selection). We derive an EM algorithm to obtain the penalized model estimates and a modified BIC criterion to select the penalization parameter. We evaluate the performance of the proposed method on simulated data.

Keywords: factor analysis, LASSO, finite Gaussian mixtures

1 Introduction

When a large number of variables is observed on a set of units, with the goal of detecting clusters of units, it may be extremely unlikely that natural groupings will exist based on all the attributes. Usually clustering, if it exists, occurs only within a relatively small unknown subset of variables.

It is well known that most clustering methods are strongly derailed by the presence of non informative variables.

Model based clustering assumes that the data come from a finite mixture model with each component corresponding to a cluster. For quantitative data each mixture component is usually modeled as a multivariate gaussian distribution.

When the number of observed variables is large, it is well known that gaussian mixture models represent an over-parameterized solution as, besides the mixing weights, it is required to estimate the mean vector and the variance covariance matrix for each component. This issue has been widely and variously addressed in the statistical literature.

There has also been an increasing interest in variable selection for model based clustering, mostly within the Bayesian framework (Liu et al., 2003, Hoff, 2005), but recently also in the frequentist one (Pan and Shen, 2007). The idea is to parameterize the mean of the k-th cluster $\mu_k = \mu + \delta_k$, where μ is the global mean. If some components of δ_k are 0, then the corresponding

attributes are non informative to clustering, at least as far as the cluster location is concerned.

In this paper we address both issues simultaneously by assuming that the data have been generated by a linear factor model with latent variables modeled as gaussian mixtures and by shrinking the factor loadings, resorting to a penalized likelihood method with an L1 penalty.

In the following we first briefly review the standard model based clustering and present our approach to dimension reduction; afterwards we propose an implementation with an L_1 penalty resulting in soft-thresholding on the estimated factor loadings and thus realizing automatic variable selection. We derive an EM algorithm to obtain the penalized model estimates and a modified BIC criterion to select the penalization parameter. We evaluate the performance of the proposed method on simulated data.

2 Model based clustering

Let $\mathbf{y}$ be a p-dimensional vector of continuous observed variables. According to the model based approach to clustering the density of $\mathbf{y}$ can be modelled by a mixture of a sufficiently large enough number k of multivariate normal component distributions each of which corresponds to a unit cluster

$$f(\mathbf{y};\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \phi(\mathbf{y};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i), \tag{1}$$

where the vector $\boldsymbol{\theta}$ of unknown parameters consists of the mixing proportions w_i, the component means $\boldsymbol{\mu}_i$, and the component-covariance matrices $\boldsymbol{\Sigma}_i$ and $\phi(\mathbf{y};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ denotes the p-variate normal density function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$.

Banfield and Raftery (1993) proposed a general structure for geometric cross-cluster constraints by parameterizing covariance matrices through eigenvalue decomposition in the form

$$\boldsymbol{\Sigma}_i = \lambda_i \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^{\top}, \tag{2}$$

where $\mathbf{D}_i$ is the orthogonal matrix of eigenvectors (controlling component orientation), $\mathbf{A}_i$ is a diagonal matrix whose elements are proportional to the eigenvalues (thus defining shape), and λ_i is an associated constant of proportionality (defining the volume). Those parameters could be the same for each cluster or might be allowed to vary across the different components.

However if p is large relative to the sample size, the recourse to this decomposition may not be problem-free. A different approach is given by the so called mixture of factor analyzers which assumes that within each component the data are generated according to an ordinary factor model, thus reducing the number of parameters from which the variance covariance matrices depend (McLachlan et al., 2003).

2.1 Dimension reduction

In this paper we assume, without loss of generality, that the mean centered p observed continuous variables have been generated according to the linear factor model

$$\mathbf{y} = \boldsymbol{\Lambda}\mathbf{z} + \mathbf{u}. \tag{3}$$

where $\mathbf{z}$ is a r-dimensional vector of latent variables, $\boldsymbol{\Lambda}$ is the factor loading matrix and $\mathbf{u}$ is a p-dimensional Gaussian term which includes the so called specific factors with zero mean and diagonal covariance matrix $\boldsymbol{\Psi}$. We further assume that the vector of latent variables $\mathbf{z}$ can be modeled according to a finite mixture of multivariate Gaussians

$$\mathbf{z} \sim \sum_{i=1}^{k} w_i \phi_i^{(r)}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{4}$$

where w_i are the unknown mixing proportions, $\phi_i^{(r)}$ is the r-dimensional Gaussian density with component mean and variance covariance matrix $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ respectively. The only requirements we impose on the factors are that they have zero mean and identity covariance matrix, thus the mixture parameters must satisfy the requirements:

$$E(\mathbf{z}) = \sum_{i=1}^{k} w_i \boldsymbol{\mu}_i = 0$$

$$Var(\mathbf{z}) = \sum_{i=1}^{k} w_i \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\top}\right) - \left(\sum_{i=1}^{k} w_i \boldsymbol{\mu}_i\right)\left(\sum_{i=1}^{k} w_i \boldsymbol{\mu}_i\right)^{\top} = \mathbf{I}_r$$

These assumptions mean that the latent variables are centered and uncorrelated but they are mutually dependent, since for non Gaussian random variables uncorrelatedness does not imply independence. This approach represents a generalization of the ordinary factor analysis model which is reproduced when the number of mixture components is equal to one, $k = 1$.

It is worth noting that modeling the factors as a multivariate Gaussian mixture amounts to model the observed variables as a particular multivariate Gaussian mixture model too:

$$\begin{aligned} f(\mathbf{y}) &= \int f(\mathbf{y}|\mathbf{z}) f(\mathbf{z}) d\mathbf{z} \\ &= \int \phi^{(p)}(\boldsymbol{\Lambda}\mathbf{z}, \boldsymbol{\Psi}) \sum_{i=1}^{k} w_i \phi_i^{(r)}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) d\mathbf{z} \\ &= \sum_{i=1}^{k} w_i \int \phi^{(p)}(\boldsymbol{\Lambda}\mathbf{z}, \boldsymbol{\Psi}) \phi_i^{(r)}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) d\mathbf{z}, \end{aligned}$$

that is

$$\mathbf{y} \sim \sum_{i=1}^{k} w_i \phi_i^{(p)}(\boldsymbol{\Lambda}\boldsymbol{\mu}_i, \boldsymbol{\Lambda}\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}) \tag{5}$$

which allows for heteroscedastic mixture components, sharing the same $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ matrices. From (5) it clearly turns out that the proposed approach yields a remarkable reduction in the number of free parameters either in the mean vectors and in the variance covariance matrices.

The log-likelihood of the proposed model is given by

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^{n} \log \sum_{i=1}^{k} w_i \phi_i^{(p)}(\mathbf{y}_h; \boldsymbol{\Lambda}\boldsymbol{\mu}_i, \boldsymbol{\Lambda}\boldsymbol{\Sigma}_i\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}) \tag{6}$$

where $\boldsymbol{\theta}$ collectively denotes the set of model parameters.

2.2 Variable selection

Within the regression context, variable selection has recently been addressed through maximum penalized likelihood. In particular the LASSO approach has turn out to be able to perform a soft thresholding on the estimated coefficients, thus realizing automatic variable selection (Tibshirani, 1996). Following this approach we propose a penalized model based clustering within model (5). Specifically the LASSO penalization is introduced on the factor loadings. The penalized log likelihood is

$$Q(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - n\gamma \sum_{j=1}^{p} \sum_{l=1}^{r} |\lambda_{jl}|, \tag{7}$$

where $\ell(\boldsymbol{\theta})$ is the log-likelihood of the model and γ is a trade-off parameter that needs to be properly selected. As pointed out by Fan and Li (2001), the derivation of penalized maximum likelihood estimates is a challenging task, due to the fact that the penalization term is not differentiable at $\lambda_{jl} = 0$. To overcome this difficulty they propose to locally approximate the penalty term $|\lambda_{jl}| \; \forall \; j, l$, by a quadratic function

$$|\lambda_{jl}| \approx |\lambda_{jl0}| + \frac{(\lambda_{jl}^2 - \lambda_{jl0}^2)}{2\,(|\lambda_{jl0}|)} \tag{8}$$

for $\lambda_{jl} \approx \lambda_{jl0}$. This local quadratic approximation allows the use of a Newton-Raphson algorithm for maximizing the penalized likelihood in (7). At each iteration of this algorithm, the value of λ_{jl0} in (8) represents the provisional estimate of λ_{jl}. However the drawback of this choice is that when $\lambda_{jl0} = 0$, the denominator $2\,(|\lambda_{jl0}|)$ makes (8) undefined. In this situation, Fan and Li (2001) suggest to set the estimate for λ_{jl} equal to zero and to stop the

algorithm when λ_{jl0} is very close to zero. As an alternative to this solution, Hunter and Li (2005) extend the local quadratic approximation idea by perturbing expression (8) as follows

$$p_\varepsilon(\lambda_{jl}) = |\lambda_{jl0}| + \frac{(\lambda_{jl}^2 - \lambda_{jl0}^2)}{2(\varepsilon + |\lambda_{jl0}|)} \tag{9}$$

for some $\varepsilon > 0$. Then maximum penalized likelihood estimates for the factor loadings can be obtained by replacing $|\lambda_{jl}|$ in (7) with (9) by resorting to a minorize-maximize (MM) procedure (see Hunter and Li (2005) for further details).

2.3 Model selection

The number of factors r, the number of groups k and the value of the penalty parameter γ can be chosen through an exhaustive search, using model selection criteria, which take into account both the fit and the complexity of the model. The fit of the model is measured by the optimal value of the log likelihood, while the complexity can be given by the number of free parameters denoted by v. The presence of a penalty term reduces the number of free parameters of the factor loading matrix. According to Hunter and Li (2005) and Khalili and Chen (2007), the effective degrees of freedom of $\boldsymbol{\Lambda}$ are computed as follows:

$$tr\left\{\left[\ell''(\boldsymbol{\theta}) - n\gamma E(\boldsymbol{\Lambda})\right]^{-1}\ell''(\boldsymbol{\theta})\right\}$$

where $\ell''(\boldsymbol{\theta})$ contains the second derivatives of the log likelihood function with respect to λ_{jl}, and $E(\boldsymbol{\Lambda})$ is a diagonal matrix with elements $\frac{1}{\varepsilon+|\lambda_{jl}|}$ (with $j = 1, \ldots, p$ and $l = 1, \ldots, r$) evaluated at the maximum penalized estimates.

In the literature, different criteria for combining fit and complexity have been proposed, each of which has its own advantages and limitations. In this work we consider the Bayesian Information Criterion (BIC) proposed by Schwarz (1978), which takes the form

$$-2\ell(\hat{\boldsymbol{\theta}}) + v\log n.$$

The BIC is one of the most widely used criterion, particularly in the context of model based clustering (see for example McLachlan and Peel, 2000, Fraley and Raftery, 2002a).

3 Maximum penalized likelihood estimation

In order to derive the maximum penalized likelihood estimates for the proposed model parameters, the penalized likelihood function has to be maximized but it is clear from expression (6) and (7) that its direct optimization

with respect the different parameters is intractable. The maximum likelihood estimation problem can be solved using the EM-algorithm (Dempster *et al.*, 1977), since the proposed model consists of two layers of missing data and the complete density of the observed and latent variables can be expressed in a simplified hierarchical form.

The two layers of missing data are given by the factors, $\mathbf{z}$, and by the so called allocation variable, which derives from modeling the factor as a mixture of Gaussians. In fact, in finite mixture models a sample of observations can be viewed as arising from k underlying populations of proportions w_i with $i = 1, \ldots, k$. The so called allocation latent variable, s, is a vector of dimension k which assumes value equal to 1 if the observation belongs to one of the k populations and 0 elsewhere. Without loss of generality, we imagine hereinafter that $s^{(i)} = 1$ where $s^{(i)}$ denotes the i^{th} element of s.

It is evident that s follows a multinomial distribution

$$f(s; \boldsymbol{\theta}) = \prod_{i=1}^{k} w_i^{s^{(i)}}, \tag{10}$$

and therefore $f(s^{(i)} = 1; \boldsymbol{\theta}) = w_i$.

The conditional density of the factors given the allocation variable is multivariate Gaussian

$$f(\mathbf{z}|s^{(i)} = 1; \boldsymbol{\theta}) = \phi_i^{(r)}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{11}$$

Given the two layers of latent variables the complete density $f(\mathbf{y}, \mathbf{z}, s; \boldsymbol{\theta})$ can be expressed in the hierarchical form:

$$f(\mathbf{y}, \mathbf{z}, s; \boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}) f(\mathbf{z}|s; \boldsymbol{\theta}) f(s; \boldsymbol{\theta}). \tag{12}$$

This hierarchical form allows to decompose the complete density as the product of three known densities; the first term is a p-dimensional Gaussian $f(\mathbf{y}|\mathbf{z}) = \phi^{(p)}(\boldsymbol{\Lambda}\mathbf{z}, \boldsymbol{\Psi})$ as a consequence of the model assumptions and the other two terms have been reported in (10) and (11).

As the complete density depends on unobservable variables we maximize its conditional expectation given the observed data, using a fixed set of parameters, $\boldsymbol{\theta}'$:

$$\arg\max_{\boldsymbol{\theta}} \left\{ E_{\mathbf{z},s|\mathbf{y},\boldsymbol{\theta}'} \left[\sum_{h=1}^{n} \log f\left(\mathbf{y}_h, \mathbf{z}_h, s_h|\boldsymbol{\theta}\right) \right] - n\gamma \sum_{j=1}^{p} \sum_{l=1}^{r} p_\varepsilon\left(\lambda_{jl}\right) \right\}.$$

The EM-algorithm alternates between two steps, the expectation and the maximization ones, until convergence in $Q(\boldsymbol{\theta})$. In the first step, the so called E-step, the expected value of the penalized log-likelihood given the observed data is calculated on the basis of provisional estimates of the parameters, denoted by $\boldsymbol{\theta}'$. In the second step, the M-step, the expectation of $Q(\boldsymbol{\theta})$ is

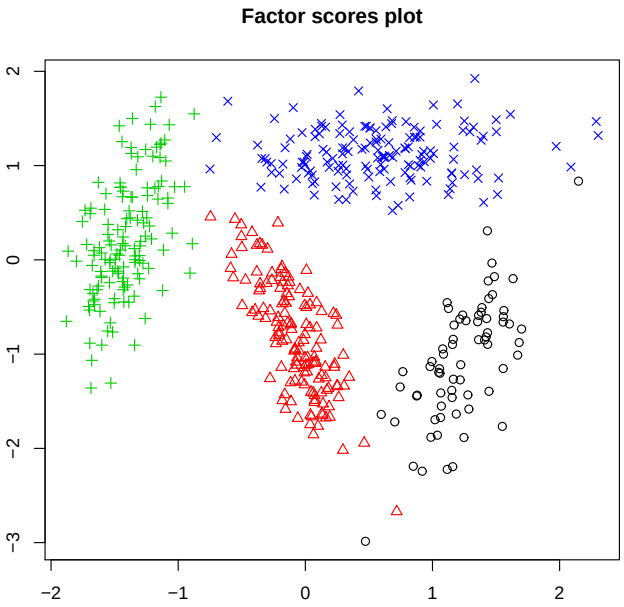

Fig. 1. Factor distribution.

maximized with respect to $\boldsymbol{\theta}$ to obtain new provisional estimates. Due to the presence of the penalty term, a Newton-Raphson procedure has been implemented to derive the factor loading estimates as they can not be expressed in closed form.

Even if the EM algorithm represents an elegant way to solve the estimation problem for the proposed model, it is well known that it suffers from some limitations, such as the dependence of the solution upon the starting values and the possibility of falling into a local optimum. In implementing the EM algorithm for the proposed model, we considered the ordinary unconstrained factor analysis estimates as starting points for $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ and multiple random starting points for the other model parameters.

4 Experimental results: a simulation study

The effectiveness of the proposed penalized model based clustering is evaluated on simulated data. A sample of $n = 500$ observations is generated from the noisy mixing process of two factors distributed according to a bivariate Gaussian mixture with $k = 4$ components.

In Figure 1 the scatterplot of the two factor scores is displayed. It is clear from this graph that the 4 groups are jointly well separated. The two

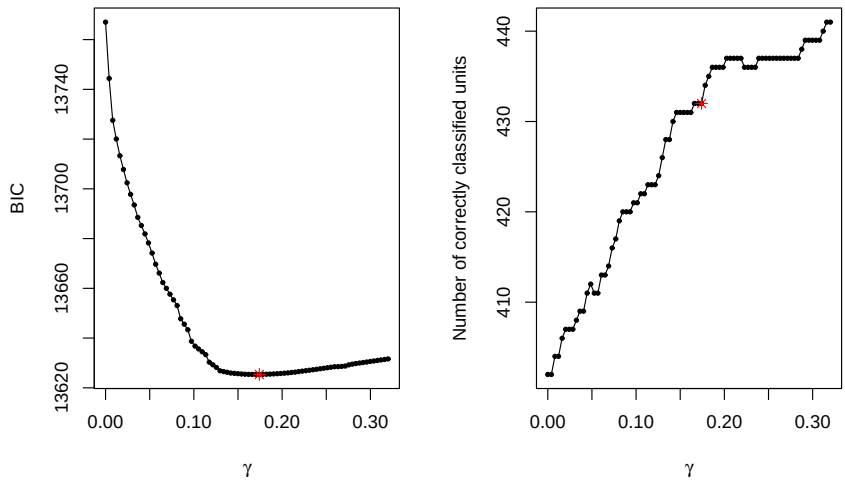

Fig. 2. Performances of the penalized mixture factor analysis model. The first graph shows the values of the BIC for estimated models with different levels of penalty term. The second graph represents the number of correctly classified units accordingly to the different estimated models.

dependent factors which represent the latent structure of the model have been linearly mixed by the coefficients of the factor loading matrix reported in the first two columns of table 1 and a reasonable level of Gaussian noise is added to the resulting $p = 20$ variables.

The factor loading coefficients have been chosen in order to identify a set of 10 relevant variables which are linear combination of the factors and a set of 10 not relevant variables which are only expression of some Gaussian noise, according to the assumption on the specific term.

On this data several penalized mixture models have been estimated with different values of penalty parameter ranging from 0 to 1. The different estimated models are then compared by evaluating the BIC criterion with the effective degrees of freedom of the penalized model. In the first graph of figure 2 the values of the BIC for estimated models with different values of the penalty parameter are shown. The curve exhibits its minimum in $\gamma = 0.174$, corresponding to a BIC value of 13625.38. The BIC value for the unrestricted model is 13767.02. For comparison purposes, we also fitted a 20−dimensional Gaussian mixture model with $k = 4$ components by Mclust software (Fraley and Raftery, 1999, 2002b, 2003). The BIC value for the model selected by Mclust is 14755.18.

Table 1. Factor loading coefficients.

	True parameters		Unrestricted estimates		Varimax rotation		Penalized estimates	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
y_1	0.00	0.50	−0.06	0.46	-0.13	0.45	−0.03	0.45
y_2	0.00	−0.40	0.04	−0.43	0.10	−0.42	0.02	−0.42
y_3	0.00	0.30	−0.04	0.28	−0.07	0.27	−0.02	0.27
y_4	0.40	0.60	0.32	0.61	0.24	0.65	0.35	0.58
y_5	-0.70	0.20	−0.72	0.14	−0.73	0.04	−0.70	0.17
y_6	0.50	-0.30	0.55	−0.27	0.58	−0.19	0.52	−0.29
y_7	0.30	0.00	0.29	0.04	0.28	0.07	0.28	0.00
y_8	-0.40	0.00	−0.43	−0.04	−0.42	−0.10	−0.42	0.00
y_9	0.70	0.00	0.70	0.05	0.69	0.15	0.69	0.00
y_{10}	-0.50	0.00	−0.50	−0.03	−0.49	−0.09	−0.48	0.00
y_{11}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
y_{12}	0.00	0.00	0.04	0.01	0.04	0.02	0.00	0.00
y_{13}	0.00	0.00	0.04	−0.01	0.04	−0.01	0.00	0.00
y_{14}	0.00	0.00	0.03	0.03	0.03	0.03	0.00	0.00
y_{15}	0.00	0.00	−0.01	−0.02	−0.01	−0.03	0.00	0.00
y_{16}	0.00	0.00	−0.01	0.00	−0.01	0.00	0.00	0.00
y_{17}	0.00	0.00	0.00	0.04	−0.01	0.04	0.00	0.01
y_{18}	0.00	0.00	−0.04	−0.04	−0.03	−0.05	−0.01	0.00
y_{19}	0.00	0.00	−0.04	−0.04	−0.03	−0.04	0.00	0.00
y_{20}	0.00	0.00	−0.04	0.00	−0.04	−0.01	−0.01	0.00

In the second graph the number of correctly classified units is computed for each estimated model. The BIC based optimal penalized model correctly classifies 432 units out of the total of 500 observations thus yielding an improvement of the classification performance of both the unrestricted mixture factor analysis (402 units) and the model selected by Mclust (406 units).

In order to be sure that the obtained results are not affected by overfitting, a second independent dataset has been generated with the same parameter setting. Units of this dataset have been classified in $k = 4$ clusters according to their posterior probabilities computed on the basis of the previously estimated models. Table 2 shows their confusion matrices. The number of correctly classified units in this second dataset is equal to 353 in the unrestricted mixture factor analysis (MFA), 432 in the penalized mixture factor analysis (PMFA) and 399 in Gaussian mixture model (MCLUST). Again, the BIC based optimal penalized model yields an improvement of the classification performance. It is interesting to note that the unrestricted mixture factor analysis is strongly affected by overfitting, since its classification performance dramatically worsens on the independent data set.

Columns 3 and 4 of table 1 contain the factor loading estimates of the unrestricted mixture factor analysis with $\gamma = 0$. It is evident from the table

Table 2. Confusion matrices on a test dataset.

	Estimated cluster membership											
True cluster	MFA				PMFA				MCLUST			
membership	1	2	3	4	1	2	3	4	1	2	3	4
1	38	33	0	1	46	25	0	1	69	3	0	0
2	68	51	13	1	16	105	12	0	48	1	75	9
3	14	0	104	13	0	0	131	0	0	3	25	103
4	2	0	14	148	2	0	12	150	0	152	1	11

that, although the factor loadings of the irrelevant variables are small, in most cases they are not equal to zero thus not allowing to objectively discard them. Even a varimax rotation of these coefficient estimates (columns 5 and 6 of table 1) does not solve the selection problem. In the last columns the factor loading estimates of the penalized model with $\gamma = 0.174$ are reported. In this case the introduction of the penalty term, not only allows to obtain a better classification of the units, but also leads to a sparse structure of the factor loading matrix, in which most of the irrelevant variables are correctly discarded.

References

BANFIELD, J.D. and RAFTERY, A.E. (1993): Model-based Gaussian and non-Gaussian Clustering, *Biometrics, 49, 803-821.*

DEMPSTER, N.M., LAIRD, A.P. and 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.*

FAN, J. and LI, R. (2001): Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association, 96, 1348-1360.*

FRALEY, C. and RAFTERY, A.E. (1999): MCLUST: Software for model-based cluster analysis, *Journal of Classification, 16, 297-206.*

FRALEY, C. and RAFTERY, A.E. (2002a): Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association, 97, 611-631.*

FRALEY, C. and RAFTERY, A.E. (2002b): MCLUST: Software for model-based clustering, discriminant analysis, and density estimation, *Technical Report No. 415, Department of Statistics, University of Washington.*

FRALEY, C. and RAFTERY, A.E. (2003): Enhanced Software for model-based clustering, discriminant analysis, and density estimation: MCLUST, *Journal of Classification, 20, 263-286.*

HOFF, P.D. (2005): Subset clustering of binary sequences, with an application to genomic abnormality data, *Biometrics, 61, 1027-1036.*

HUNTER, D.R. and LI, R. (2005): Variable selection using MM algorithms, *The Annals of Statistics, 33, 1617-1642.*

KHALILI, A. and CHEN, J. (2007): Variable selection in finite mixture of regression models, *Journal of the American Statistical Association, 102, 1025-1038*.

LIU, J.S., ZHANG, J.L., PALUMBO, M.J. and LAWRENCE, C.E. (2003): Bayesian clustering with variable and transformation selection (with discussion), *Bayesian Statistics, 7, 249-275*.

MCLACHLAN, G.J., PEEL, D. (2000): *Finite Mixture Models*, John Wiley & Sons INC, New York.

MCLACHLAN, G.J., PEEL, D. and BEAN, R.W. (2003): Modelling high-dimensional data by mixtures of factor analyzers, *Computational Statistics and Data Analysis, 41, 379-388*.

PAN, W., and SHEN, X. (2007): Penalized model-based clustering with application to variable selection *Journal fo machine learning research, 8, 1145-1164*.

SCHWARZ, G. (1978): Estimating the Dimension of a Model, *Annals of Statistics, 6, 461-464*.

TIBSHIRANI, R. (1996): Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B, 58, 267-288*.

Modelling Background Noise in Finite Mixtures of Generalized Linear Regression Models

Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstrasse 33, 80539 München, Germany,
Friedrich.Leisch@stat.uni-muenchen.de

Abstract. In this paper we show how only a few outliers can completely break down EM-estimation of mixtures of regression models. A simple, yet very effective way of dealing with this problem, is to use a component where all regression parameters are fixed to zero to model the background noise. This noise component can be easily defined for different types of generalized linear models, has a familiar interpretation as the empty regression model, and is not very sensitive with respect to its own parameters.

Keywords: mixture models, generalized linear models, robust statistics, R

1 Introduction

Finite mixture models have been used for more than 100 years, but have seen a real boost in popularity over the last decades due to the tremendous increase in available computing power. The areas of application of mixture models range from biology and medicine to physics, economics and marketing. On the one hand these models can be applied to data where observations originate from various groups and the group affiliations are not known, and on the other hand to provide approximations for multi-modal distributions (Everitt & Hand (1981), Titterington et al (1985); McLachlan & Peel (2000)).

In the 1990s finite mixture models have been extended by mixing standard linear regression models as well as generalized linear models (Wedel & DeSarbo (1995)). An important area of application of mixture models and also of these extensions are in market segmentation (Wedel & Kamakura (2001)), where finite mixture models replace more traditional cluster analysis and cluster-wise regression techniques as state of the art.

For mixtures without a regression part, i.e., model-based clustering, several authors have investigated the effect of outliers on parameter estimates, and how outliers can be treated to get more robust behaviour. A comprehensive theoretical analysis for breakdown points of ML-estimators of location-scale mixtures can be found in Hennig (2004). Suggested solutions for robustification against outliers include

1. to add a noise component which is either uniform over the convex hull of the complete data set (Banfield & Raftery (1993)), or an improper constant uniform (Hennig & Coretto (2007)),
2. replace Gaussian densities with t-densities (Mclachlan & Peel (2000)), and
3. trimming observations (Cuesta-Albertos et al (1997)).

In this paper we present a new noise component to model outliers and show that our approach combines several aspects of the above. In addition, it can be easily extended to mixtures of regression models and has a natural interpretation in this context as the null model of no interaction between predictors and response.

2 Mixtures of GLMs

Consider finite mixture models with K components of form

$$h(y|x,\psi) = \sum_{k=1}^{K} \pi_k f(y|x,\theta_k) \tag{1}$$

$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1$$

where y is a (possibly multivariate) dependent variable with conditional density h, x is a vector of independent variables, π_k is the prior probability of component k, θ_k is the component specific parameter vector for the density function f, and $\psi = (\pi_1, , \ldots, \pi_K, \theta_1', \ldots, \theta_K')'$ is the vector of all parameters.

If f is a univariate normal density with component-specific mean $\mu_k(x) = \alpha_k + \beta_k' x$ and variance σ_k^2, we have $\theta_k = (\alpha_k, \beta_k', \sigma_k^2)'$ and Equation (1) describes a mixture of standard linear regression models, also called *latent class regression*. If f is a member of the exponential family, we get a mixture of generalized linear models. For multivariate normal f and $x \equiv 1$ we get a mixture of Gaussians without a regression part (model-based clustering).

The posterior probability that observation (x, y) belongs to class j is given by

$$\mathbb{P}(j|x,y,\psi) = \frac{\pi_j f(y|x,\theta_j)}{\sum_k \pi_k f(y|x,\theta_k)} \tag{2}$$

The posterior probabilities can be used to segment data by assigning each observation to the class with maximum posterior probability. In the following we will refer to $f(\cdot|\cdot,\theta_k)$ as *mixture components* or *classes*, and the groups in the data induced by these components as *clusters*.

The log-likelihood of a sample of N observations $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ is given by

$$\log L = \sum_{n=1}^{N} \log h(y_n|x_n,\psi) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f(y_n|x_n,\theta_k) \right) \tag{3}$$

and can usually not be maximized directly. The most popular method for maximum likelihood estimation of the parameter vector ψ is the iterative expectation-maximization algorithm (EM, Dempster et al (1977)):

Estimate the posterior class probabilities for each observation

$$\hat{p}_{nk} = \mathbb{P}(k|x_n, y_n, \hat{\psi})$$

using Equation (3) and derive the prior class probabilities as

$$\hat{\pi}_k = \frac{1}{N} \sum_{n=1}^{N} \hat{p}_{nk}$$

Maximize the log-likelihood for each component separately using the posterior probabilities as weights

$$\max_{\theta_k} \sum_{n=1}^{N} \hat{p}_{nk} \log f(y_n|x_n, \theta_k) \tag{4}$$

The E- and M-steps are repeated until the likelihood improvement falls under a pre-specified threshold or a maximum number of iterations is reached.

Parameter estimates in standard linear models with Gaussian errors and most other GLMs are rather sensitive to outliers, because the maximum likelihood estimate is basically a mean value, which is not a robust statistic. For mixtures of regression models the problem is even more pronounced, because the variance is no longer a nuisance parameter, it needs to be estimated to compute likelihoods and posterior probabilities in each EM iteration.

One solution would be to use robust regression in the M-step, however this would violate the EM principle as the resulting estimates are no longer maximum likelihood estimates. Hence, convergence is no longer guaranteed even for clean data. In addition we run into the problem that robust estimates ususally themselves are computationally very demanding, we need estimates for every component in every EM-iteration, and convergence of EM is usually rather slow. Hence, we would need to compute expensive estimates very often.

3 Modelling background noise

Outliers or background noise can be modeled by adding a noise component f_0 to our mixture model from Equation 1:

$$h(y|x, \psi) = \pi_0 f_0(y|x, \theta_0) + \sum_{k=1}^{K} \pi_k f(y|x, \theta_k) \tag{5}$$

$$\pi_k \geq 0, \quad \sum_{k=0}^{K} \pi_k = 1$$

In the following we will call f_0 the *noise component*, and the remaining components for $k = 1, \ldots, K$ the *regular components*.

Banfield & Raftery (1993) and Hennig & Coretto (2007) use a uniform distribution for f_0, the main difference is that the former estimate the range of the uniform from the data, while the latter use either an improper uniform with pre-specified fixed value for the height of the density, or an ML estimate for the complete mixture including the noise component. Both consider only the case of model-based clustering, i.e., no regression.

3.1 Gaussian response

For mixtures of regression models there is a natural other candidate for the noise component, the null model which assumes no relationship between predictors x and response y. For notational simplicity, consider for the moment standard linear regression models with Gaussian noise, such that

$$f(y|x, \theta_k) = \phi\left(\frac{y - \mu_k(x)}{\sigma_k}\right) = \phi\left(\frac{y - \alpha_k - \beta_k' x}{\sigma_k}\right)$$

where $\phi(\cdot)$ denotes the density of the standard normal distribution. Using a noise component of form

$$f_0(y|x, \theta_0) = f_0(y|\theta_0) = \phi(\left(\frac{y - \mu_0}{\sigma_0}\right)$$

means we add a compontent corresponding to an empty regression model of form $y = \mu_0 + \epsilon$.

There are three possible ways to define the noise parameters μ_0 and σ_0:

NP1: set to fixed values in advance based on expert opinion,
NP2: estimate from data but hold fixed during EM iterations, e.g., to mean and standard deviation of y, or
NP3: treat f_0 as a regular mixture component and estimate its parameters by EM together with all other parameters of the model.

Obvously NP1 is the most robust variant, because it does not depend on the data at all. However, our simulations show that NP2 is also very robust, so we consider only the data-driven solutions NP2 and NP3 for the remainder of this paper.

Using an empty regression model as noise component has several attractive features: The noise component has the same functional form as the other components, so it is particularly easy to implement in software given the rest of the mixture model, see Section 4. There is also a natural interpretation of parameter π_0, which is the probability that an observation originated from the empty model. This is closely related to popular statistics of standalone regression models such as R^2 or analysis of variance F, which also compare a regression model with the empty model.

The effects of including the noise component can easily be seen by taking a look at the posterior probabilities (2). If we fix $\sigma_0^2 \equiv \text{var}(y)$, then

$$\sigma_0 \geq \sigma_k, \quad k = 1, \ldots, K$$

with (approximate) equality only for components where $\beta_k \approx 0$, and usually all σ_k are smaller than σ_0. Hence, the posterior probability of the noise component equals the ratio of a normal density with large density to the sum of several normal densities, see Equation 3.

Figure 1 shows examples for $\sigma_0 = 2\sigma_1$, $\sigma_0 = 4\sigma_1$, and $\sigma_0 = 8\sigma_1$. The posterior probabilities of the noise component are larger than 0.99 outside the interval $[-4, 4]$, and larger than 0.9 outside of $[-3, 3]$. Observations which are further than 4 standard deviations away from a regular mixture component have zero weight in the M step in Equation 2 of the EM-procedure.

Choosing a Gaussian noise component rather than a uniform makes no large difference in which observations are marked as outliers. If σ_0 is large (as intended), then the Gaussian is very flat and over the main part it is very similar to the uniform. The big advantage is that the support of the Gaussian is unbounded, although it will become very small outside of, say, $\mu_0 \pm 4\sigma_0$. However, the weights used in (2) are ratios of densities (3), and due to the larger variance the density of the Gaussian noise component will always be much larger than the densities of the regular components in regions far away from the center. Thus, we knock out outliers everywhere except for the main support regions of the regular components. For uniforms, we need to solve the ill-conditioned estimation problem of the boundaries of the uniform distribution, see Hennig & Coretto (2007) for a detailed discussion. For the Gaussians exact estimation of variance is not really critical (a rather unusual situation!), Figure 1 shows that the value of σ_0 has not much influence on which observations are marked as outliers. Preliminary simulations studies (not shown here) confirm this bahaviour.

3.2 Other GLMs

The same form of noise component can easily be used in other continous members of the exponential family, as well as in some discrete distributions like the Poisson. Due to the limited space of this conference paper we cannot give full formulas or examples. The basic principle is always to have the null model with no regression part as noise component, and estimate the parameters of the noise component from the complete data set.

E.g., an exponential distribution with a large and constant mean value gives a noise component with a rather flat density on $\mathbb{R}^+$, which downweights large outliers, similar for the gamma distribution. For Poisson responses one can use overdispersed quasi-Poisson noise components. It is not so clear how the concept can be used for GLMs for categorical data (binomial, multinomial), but in this case even the definition of "outliers" or "background noise" is problematic.

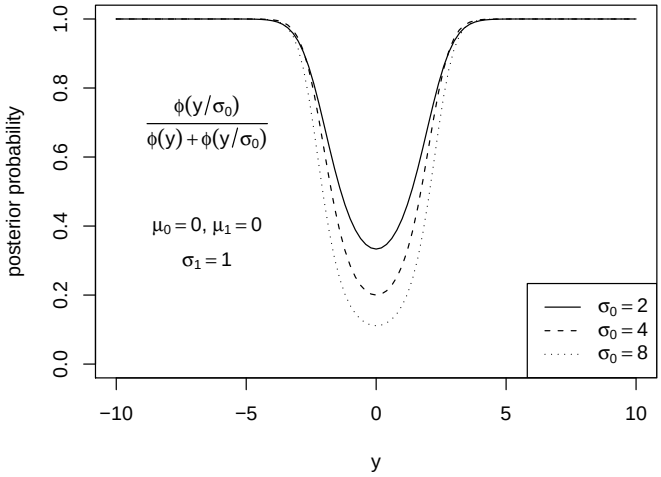

Fig. 1. Posterior probability of the noise component.

4 Software implementation

All simulation results shown below were computed using R (R Development Core Team (2007)) extension package `flexmix` (Leisch (2004), Gruen & Leisch (2007)). The standard driver for mixtures of GLMs in `flexmix` is `FLXMRglm`. The new extension fixes the first component to be the noise component, and dispatches to the standard driver for the rest. The current development version of the software can be obtained from the author upon request and will be released on CRAN (`http://cran.r-project.org`) as part of `flexmix` later this year.

It allows to estimate the parameters of the noise component either fixed from the complete data set, in which case only π_0 is estimated by maximum likelihood, or by weighted maximum likelood with weights proportional to the probability of being a member of the noise component. The latter approach has the advantage that the null model can be interpreted at par with the regular components, but is not robust against outliers which are located close to each other.

5 Artificial example

First we consider a simple example introduced by Leisch (2004) with two latent classes of size 100 each:

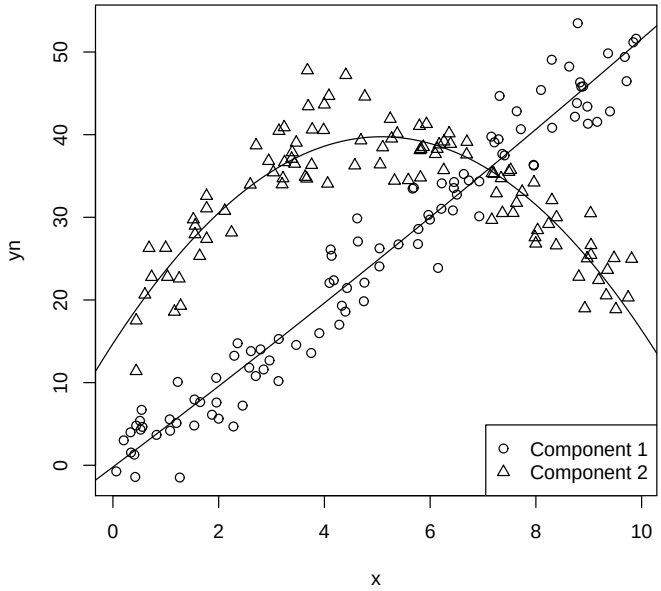

Fig. 2. A two component mixture regression example. The lines correspond to the fitted values of a model estimated with the EM algorithm.

$$\text{Class 1: } y = 5x + \epsilon$$
$$\text{Class 2: } y = 15 + 10x - x^2 + \epsilon$$

with $\epsilon \sim N(0, 9)$ and prior class probabilities $\pi_1 = \pi_2 = 0.5$. The data set can be loaded into R with the command `data("NPreg", package="flexmix")`. The result of fitting a mixture model of with two quadratic polynomial components to the data can be seen in Figure 2.

If we add three outliers on the top left corner to the data set, EM estimation breaks down and gives completely wrong results, see Figure 3. Note that this is the result with the best likelihood of 20 replications of the EM algorithm, and not simply a problem of convergence in a local minimum. Estimating the model with an additional noise component correctly identifies the three outliers with posterior probabilities numerically equal to 1. As a result, estimation of the two regular components is now correct again, see Figure 4.

Mean and variance of the noise component were fixed to the corresponding empirical estimates from the response variable. If we have only a few outliers in the same spot, we cannot reliably estimate the parameters μ_0 and σ_0 by EM. Another situation is shown in Figure 5, where 20 uniform noise

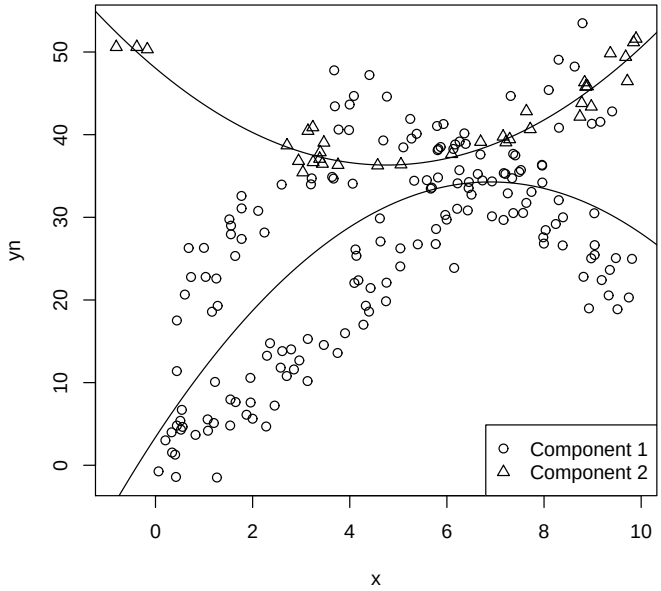

Fig. 3. The same data set as in Figure 2 with three outliers. The lines correspond to the best model found by EM, which is completely broken.

observations have been added on a rectangle that is larger than the original data range. Again, outliers not located in the main part of the original data set are correctly identified, and both the linear and the parabolic components were almost exactly identified. There is now a little bit more curvature in the fitted model for the linear class, but note that both components have a linear model with parameter estimates for intercept, x and x^2. It is impossible to distinguish original data points from background noise that is located close to the original data, so some effect is to be expected.

6 Simulation study

We also conducted several simulation studies to see whether it makes a huge difference if we estimate the parameters of the noise component by NP2 or NP3 for the case of uniform background noise. We fixed the data set described above and added $0, 5, 10, \ldots, 50$ noise observations from a uniform distribution on $[-5, 15] \times [-10, 60]$ in the same way as we did in Figure 5. For each number of noise points we drew 100 data sets, ran the EM algorithm 5 times on each and kept only the best model to avoid local minima. The

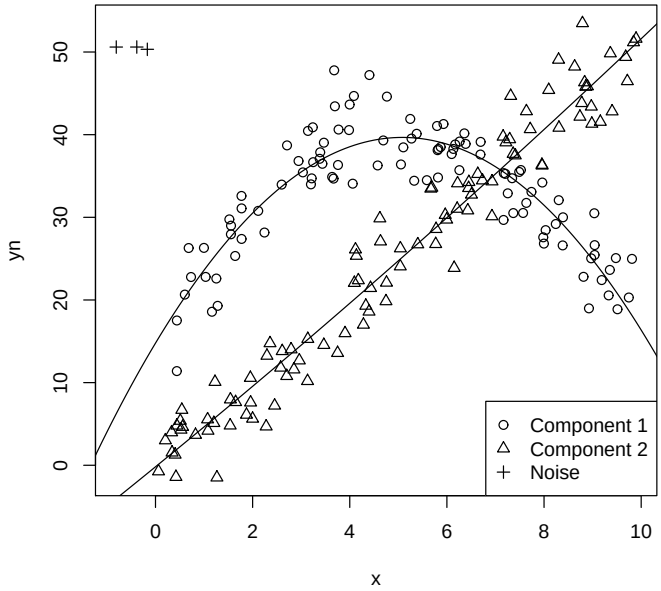

Fig. 4. The same data set as in Figure 3 using a model with a noise component. The three outliers are correctly identified.

estimated regression coefficients of the mixture models where then compared to the true parameter values.

Figure 6 shows boxplots of the Euclidean distance between estimated and true parameters. Without noise ("zero points added") EM converged to the same solution all the time, these values can be used as reference baseline. As expected, estimation error increases when more and more noise points are added, but there is no large difference between schemes NP2 and NP3. NP2 seems to be slightly better for fewer outliers, while NP3 is slightly better for more outliers. There are 2 components with 3 regression coefficents each, i.e., a total of 6 coefficients. Estimation errors range between almost zero to a median of about 7 for 50 noise points. If we divide this by the number of coefficients, we get an average error of $7/6 \approx 1.1$ per coefficient. This is not too bad, considering that 20% of the complete data set are noise and the sample size is not that large.

If we fit a mixture model without noise component, we get a median error of about 7 if we add only 5 noise points, and a median error of 15 for 10 noise points. In both cases variation is very large and EM often gets stuck in bad solutions like Figure 3. For more than 10 noise points EM estimation breaks

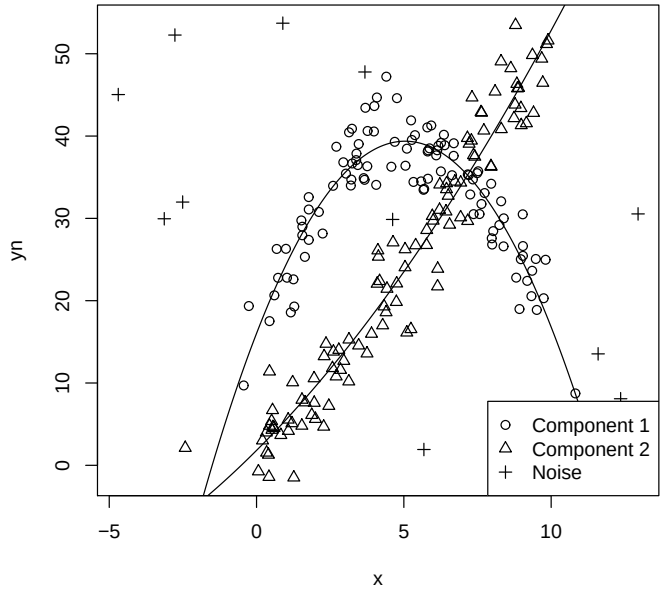

Fig. 5. The same data set as in Figure 2 with 20 outliers distributed uniformly on $[-5, 15] \times [-10, 60]$.

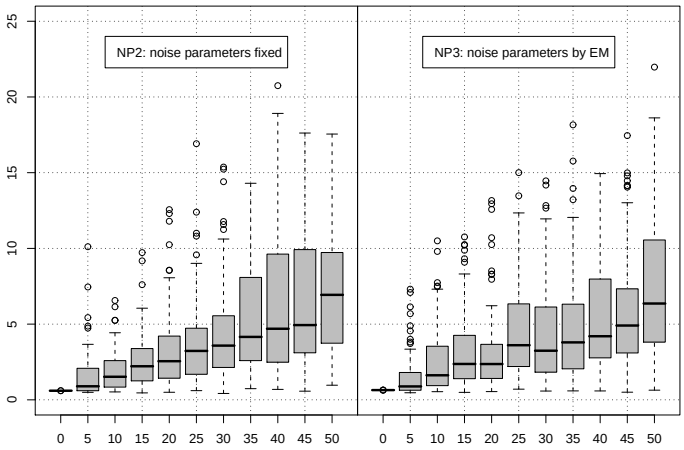

Fig. 6. Distance between estimated and true parameter values for data sets with 0–50 uniform background noise values.

completely down and yields only random results with median errors of 45 and larger. Thus, by using a noise component, we can add 10 times as many noise points for comparable increase in estimation error. Simulations with other data sets of different size, dimension and number of mixture components showed similar results.

7 Outlook

We have successfully applied the proposed methodology in a consulting project modelling customer satisfaction. The data are surveys of tourists rating Austrian alpine skiing resorts. Each respondent rated dozens of detailed aspects of the resort (quality of slopes, lifts, restaurants, entertainment, ...), the task was to identify which items had a strong impact on the overall satisfaction. A global model for all tourists makes no sense, as different subgroups of the tourist population will have different preferences. For most tourists it can be assumed, that only few items have a strong impact on overall satisfaction, the remainder being more or less noise.

We are currently working on a systematic benchmark study to confirm the findings of our preliminary simulations studies like the one presented above. This also includes GLMs with other response distributions, which were only discussed shortly in this paper due to space limitations. Another line of research is to see how other approaches presented in the literature for model based-clustering can be adapted to the case of mixtures of regression models. E.g., it should be rather straightforward to replace the normal distribution with a t-distribution if the degrees of freedom are fixed in advance.

Acknowledgements

Flexmix is joint work with Bettina Grün. This research was supported by the Austrian Science Foundation (FWF) under grant P17382.

References

BANFIELD, J.D. and RAFTERY, A.E. (1993): Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49(3), 803–821.*

CUESTA-ALBERTOS, J.A., GORDALIZA, A. and MATRAN, C. (1997): Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, *25(2), 553–576.*

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

EVERITT, B.S. and HAND, D.J. (1981): *Finite Mixture Distributions.* London: Chapman and Hall.

GRÜN, B. and LEISCH, F. (2007): Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, *51(11), 5247–5252*.

HENNIG, C. (2004): Breakdown points for maximum likelihood estimators of location-scale mixtures. *The Annals of Statistics*, *32(4), 1313–1340*.

HENNIG, C. and CORETTO, P. (2007): The noise component in model-based cluster analysis. In: *Proceedings of GfKl-2007*. Springer Verlag, Studies in Classification, Data Analysis, and Knowledge Organization.

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*.

MCLACHLAN, G. and PEEL, D. (2000): *Finite Mixture Models*. John Wiley and Sons Inc.

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.

TITTERINGTON, D., SMITH, A. and MAKOV, U. (1985): *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.

WEDEL, M. and DESARBO, W.S. (1995): A mixture likelihood approach for generalized linear models. *Journal of Classification*, *12, 21–55*.

WEDEL, M. and KAMAKURA, W.A. (2001): *Market Segmentation - Conceptual and Methodological Foundations*. Kluwer Academic Publishers, Boston, MA, USA, 2nd edition.

Clustering via Mixture Regression Models with Random Effects

Geoffrey J. McLachlan[1], Shu Kay (Angus) Ng[2], and Kui Wang[3]

[1] Department of Mathematics & Institute for Moelcular Bioscience, University of Queensland, Brisbane 4072, Australia, *gjm@maths.uq.edu.au*
[2] School of Medicine, Griffith University, University Drive, Meadowbrook QLD 4131, Australia, *s.ng@griffith.edu.au*
[3] Department of Mathematics, University of Queensland, Brisbane 4072, Australia, *kwang@maths.uq.edu.au*

Abstract. In this paper, we consider the use of mixtures of linear mixed models to cluster data which may be correlated and replicated and which may have covariates. For each cluster, a regression model is adopted to incorporate the covariates, and the correlation and replication structure in the data are specified by the inclusion of random effects terms. The procedure is illustrated in its application to the clustering of gene-expression profiles.

Keywords: correlated data, random effects, mixed linear models, gene profiles

1 Introduction

Finite mixture models are being commonly used in a wide range of applications in practice concerning density estimation and clustering. In the 1960s, the fitting of finite mixture models by maximum likelihood had been studied in a number of papers, including the seminal papers by Day (1969) and Wolfe (1965). However, it was the publication of the seminal paper of Dempster et al. (1977) on the EM algorithm that greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by maximum likelihood is a classic example of a problem that is simplified considerably by the EM's conceptual unification of maximum likelihood (ML) estimation from data that can be viewed as being incomplete; see, for example, Ganesalingam and McLachlan (1978), McLachlan (1982), McLachlan and Basford (1988), Banfield and Raftery (1994), Fraley and Raftery (1998, 2002), and McLachlan and Peel (2000).

We let $\boldsymbol{Y}$ denote a random vector consisting of p feature variables associated with the random phenomenon of interest. We let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ denote an observed random sample of size n on $\boldsymbol{Y}$. With the finite mixture model-based approach to density estimation and clustering, the density of $\boldsymbol{Y}$ is modelled as a mixture of a number (g) of component densities $f_i(\boldsymbol{y})$ in some unknown

proportions $\pi_1, \ldots, \pi_g$. That is, each data point is taken to be a realization of the mixture probability density function (p.d.f.),

$$f(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{y}), \tag{1}$$

where the mixing proportions π_i are nonnegative and sum to one. In density estimation, the number of components g can be taken sufficiently large for (1) to provide an arbitrarily accurate approximation to the underlying density function. For clustering purposes, each component in the mixture model (1) corresponds to a cluster. The posterior probability that an observation with feature vector $\boldsymbol{y}_j$ belongs to the ith component of the mixture is given by

$$\tau_i(\boldsymbol{y}_j) = \pi_i f_i(\boldsymbol{y}_j)/f(\boldsymbol{y}_j) \tag{2}$$

for $i = 1, \ldots, g$. A probabilistic clustering of the data into g clusters can be obtained in terms of the fitted posterior probabilities of component membership for the data.

An outright partitioning of the observations into g nonoverlapping clusters $C_1, \ldots, C_g$ is effected by assigning each observation to the component to which it has the highest estimated posterior probability of belonging. Thus the ith cluster C_i contains those observations $\boldsymbol{y}_j$

$$\hat{z}_{ij} = \arg\max_h \hat{\tau}_h(\boldsymbol{y}_j), \tag{3}$$

and $\hat{\tau}_i(\boldsymbol{y}_j)$ is an estimate of $\tau_i(\boldsymbol{y}_j)$. As the notation implies, $\hat{z}_{ij}$ can be viewed as an estimate of z_{ij} which, under the assumption that the observations come from a mixture of g groups $G_1, \ldots, G_g$, is defined to be one or zero according as the jth observation $\boldsymbol{y}_j$ does or does not come from G_i $(i = 1, \ldots, g; j = 1, \ldots, n)$.

For the clustering of continuous multivariate data, it is common to specify the component densities in (1) as belonging to the multivariate normal family with

$$f_i(\boldsymbol{y}) = \phi(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{4}$$

where $\phi(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ denotes the multivariate normal density function with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. We let $\boldsymbol{\Psi}$ be the vector of unknown parameters, consisting of the mixing proportions π_i, the component mean means $\boldsymbol{\mu}_i$, and the distinct elements of the component-covariance matrices $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. The maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}, \hat{\boldsymbol{\Psi}}$, is given by an appropriate root of the likelihood equation,

$$\partial \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}, \tag{5}$$

where $L(\boldsymbol{\Psi})$ denotes the likelihood function for $\boldsymbol{\Psi}$,

$$L(\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(\boldsymbol{y}_j; \boldsymbol{\Psi}),$$

and

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \quad (6)$$

Solutions of (5) corresponding to local maximizers of $\log L(\boldsymbol{\Psi})$ can be obtained via the expectation-maximization (EM) algorithm of Dempster et al. (1997); see also McLachlan and Krishnan (2008). Let $\hat{\boldsymbol{\Psi}}$ denote the estimate of $\boldsymbol{\Psi}$ so obtained.

It can be seen from (6) that the mixture model with unrestricted component-covariance matrices in its normal component distributions is a highly parameterized one with $\frac{1}{2}p(p+1)$ parameters for each component-covariance matrix $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, g)$. Thus some form of variable selection and/or regularization is needed with high-dimensional data as, for example, proposed in McLachlan et al. (2002).

In this paper, we wish to consider mixture model-based methods for clustering where there is known to be some structure on the data. In particular, we wish to focus on the approach proposed by Ng et al. (2006). Two examples will be given to illustrate its application. Although attention is focussed here solely on the clustering of the gene profiles that can be formed from the output from a series of microarray experiments, the procedure is widely applicable to the clustering of data from other experimental sources.

2 Mixtures of linear mixed models

In applying the normal mixture model (6) to cluster multivariate (continuous) data, it is assumed as in most typical cluster analyses using any other method that

(a) there are no replications on any particular entity specifically identified as such;
(b) all the observations on the entities are independent of one another.

These assumptions should hold for the clustering of, say, tissue samples consisting of the expression levels of many (possibly thousands) of genes, although the tissue samples have been known to be correlated for different tissues due to flawed experimental conditions. However, condition (b) will not hold for the clustering of gene profiles, since not all the genes are independently distributed, and condition (a) will generally not hold either as the gene profiles may be measured over time or on technical replicates. While this correlated structure can be incorporated into the normal mixture model (6) by appropriate specification of the component means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$, it can be difficult to fit the model under such specifications.

To illustrate this, we assume in the sequel that the observed data vector $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$ contains the expression levels of the jth gene obtained from a series of p microarray experiments; see, for example, McLachlan et

al. (2004). Typically in such problems, the number of genes n is very large relative to the number of microarray experiments p. In molecular biology, the $\boldsymbol{y}_j$ are referred to as the gene profiles. The underlying idea for clustering the gene profiles is that if coregulation indicates shared functionality, then clusters defined to this level of abstraction represent biological modules. If the microarray experiments were measured at p different time points, then the problem is one of clustering time-course data (that is, time series data).

Suppose, for example, the first p_1 tissue samples were obtained from p_1 healthy patients (Group G_1) and the remaining p_2=1-p_1 tissue samples were from unhealthy patients (Group G_2). Then we could cluster the gene profiles $\boldsymbol{y}_j$ by applying the normal mixture model (6) provided we impose some restrictions on the component means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ to represent this known structure. We would put

$$\boldsymbol{\mu}_i = \boldsymbol{X}\,(\boldsymbol{\mu}_{1i}^T,\, \boldsymbol{\mu}_{2i}^T)^T, \tag{7}$$

where the design matrix $\boldsymbol{X}$ is given by

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{I}_{p_1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{p_2} \end{pmatrix}, \tag{8}$$

and where

$$\boldsymbol{\mu}_{hi} = \mu_{hi}\mathbf{1}_{p_h} \quad (h = 1, 2).$$

Here $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix and and $\mathbf{1}_{p_h}$ denotes the p_h-dimensional vector with each element equal to one. Similarly, the ith component-covariance matrix $\boldsymbol{\Sigma}_i$ can be appropriately specified. Taking the microarray experiments to be independent of one another, then it would be reasonable to assume that

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_{1i} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_{2i} \end{pmatrix} \quad (i = 1, 2), \tag{9}$$

where $\boldsymbol{\Sigma}_{hi} = \sigma_{hi}^2 \boldsymbol{I}_{p_h}$ $(h = 1, 2)$.

In this simple case of cross-sectional data, the normal mixture model (6) can be fitted with minor modifications to incorporate the constraints (8) and (12). However, if the p_h tissue samples are measured over p_h time points in Group G_i, then it would not be reasonable to take $\boldsymbol{\Sigma}_{hi}$ to be diagonal. In such cases the M-step will most likely not exist in closed form even for simple models to specify the dependence over time of the tissue samples. The problems will be exacerbated if there is replication of the tissue samples (longitudinal data), in particular, technical replicates. Moreover, this approach is assuming that all the genes are independently distributed. Thus, we consider an approach to this problem that is based on mixtures of linear mixed models.

3 EMMIX-WIRE procedure

We consider the so-called EMMIX-WIRE (**EM**-based **MIX**ture analysis **W**ith **R**andom **E**ffects) procedure developed by Ng et al. (2006) to handle the clustering of correlated data that may be replicated. They adopted conditionally a mixture of linear mixed models to specify the correlation structure between the variables and to allow for correlations among the observations.

To formulate this procedure, we consider the clustering of n gene profiles $\boldsymbol{y}_j$ $(j = 1, \ldots, n)$, where we let $\boldsymbol{y}_j = (\boldsymbol{y}_{1j}^T, \ldots, \boldsymbol{y}_{mj}^T)^T$ contain the expression values for the jth gene profile and $\boldsymbol{y}_{tj} = (y_{1tj}, \ldots, y_{r_t tj})^T \quad (t = 1, \ldots, m)$ contains the r_t replicated values in the tth biological sample $(t = 1, \ldots, m)$ on the jth gene. The dimension d of $\boldsymbol{y}_j$ is given by $p = \sum_{t=1}^{m} r_t$. With the EMMIX-WIRE procedure, the observed d-dimensional vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are assumed to have come from a mixture of a finite number, say g, of components in some unknown proportions $\pi_1, \ldots, \pi_g$, which sum to one. Conditional on its membership of the ith component of the mixture, the profile vector $\boldsymbol{y}_j$ for the jth gene $(j = 1, \ldots, n)$ follows the model

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{U}\boldsymbol{b}_{ij} + \boldsymbol{V}\boldsymbol{c}_i + \boldsymbol{\epsilon}_{ij} \quad (i = 1, \ldots, g), \tag{10}$$

where the elements of the q_β-dimensional vector $\boldsymbol{\beta}_i$ are fixed effects (unknown constants) used in modelling the conditional mean of $\boldsymbol{y}_j$ in the ith component $(i = 1, \ldots, g)$. In (9), $\boldsymbol{b}_{ij}$ (a q_b-dimensional vector) and $\boldsymbol{c}_i$ (a q_c-dimensional vector) represent the unobservable gene- and cluster-specific random effects, respectively. These random effects represent the variation due to the heterogeneity of genes and samples (corresponding to $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}^T, \ldots, \boldsymbol{b}_{ip}^T)^T$ and $\boldsymbol{c}_i$, respectively). The random effects $\boldsymbol{b}_i$ and $\boldsymbol{c}_i$, and the measurement error vector $(\boldsymbol{\epsilon}_{i1}^T, \ldots, \boldsymbol{\epsilon}_{ip}^T)^T$ are assumed to be mutually independent, where $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are known design matrices of the corresponding fixed or random effects, respectively. If the covariance matrix $\boldsymbol{H}_i$ is taken to be diagonal, then the expression levels on the jth gene in different biological samples are taken to be independent. The presence of the random effect $\boldsymbol{c}_i$ for the expression levels of genes in the ith component induces a correlation between the profiles of genes within the same cluster. This is in contrast to the mixed-effects models approaches in Luan and Li (2003), McLachlan et al. (2004), Celeux et al. (2005), and Qin and Self (2006) that involve only gene-specific random effects. Their methods thus require the independence assumption for the genes which, however, will not hold in practice for all the genes. Recently, Booth et al. (2008) have adopted a Bayesian approach to this problem in which genes within the same cluster are taken to be correlated.

With the LMM, the distributions of $\boldsymbol{b}_{ij}$ and $\boldsymbol{c}_i$ are taken, respectively, to be multivariate normal $N_{q_b}(\boldsymbol{0}, \boldsymbol{H}_i)$ and $N_{q_c}(\boldsymbol{0}, \theta_{ci}\boldsymbol{I}_{q_c})$, where $\boldsymbol{H}_i$ is a $q_b \times q_b$ covariance matrix. The measurement error vector $\boldsymbol{\epsilon}_{ij}$ is also taken to be multivariate normal $N_p(\boldsymbol{0}, \boldsymbol{A}_i)$, where $\boldsymbol{A}_i = \mathrm{diag}(\boldsymbol{W}\boldsymbol{\xi}_i)$ is a diagonal matrix

constructed from the vector $(\boldsymbol{W}\boldsymbol{\xi}_i)$ with $\boldsymbol{\xi}_i = (\sigma^2_{i1}, \ldots, \sigma^2_{iq_e})^T$ and $\boldsymbol{W}$ a known $p \times q_e$ zero-one design matrix.

We now consider two examples in which we apply the EMMIX-WIRE procedure to a real data set and a simulated data set, as considered previously in the literature.

4 Example 1: Yeast cell data

In this example, we consider the CDC28 dataset, which contains more than 6000 genes measured at 17 time points (0, 10, 20,..., 160) over 160 minutes, which is about two periods of yeast cell under CDC28 condition. Cho et al. (2001) and Yeung et .al. (2001) identified and clustered some of the 6000 genes into different functional groups. For example, Yeung et al. (2001) presented 384 genes corresponding to five functional groups, among which there are 237 genes falling into four MIPS functional groups (DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins). Wong et al. (2007) reanalysed the 237 cell cycle data, using their two-stage clustering method and found that it outperformed the other methods that they tried. They were an hierarchical method, k-means, SOM, SOTA, and a normal mixture model-based procedure corresponding to (6), which were all used to cluster the 237 genes into $g = 4$ clusters. On comparing the latter with the four MIPS functional groups, they reported that the the Rand Index (RI) for their two-stage method was equal to 0.7087. In this paper, we shall compare the EMMIX-WIRE procedure with the two-stage clustering method.

In this example, the gene profile vector $\boldsymbol{y}_j$ for the jth gene is given by

$$\boldsymbol{y}_j = (y_{j1}, \ldots, y_{jp})^T,$$

where y_{jt} denotes the expression level of the jth gene at time $t\,(t = 1, \ldots, p)$ and p=17. Before proceeding to fit the model (9), we first estimated the period T in the linear regression model in which

$$y_{jt} = \beta_0 + \beta_1 cos(2\pi t/T) + \beta_2 sin(2\pi t/T) + e_{jt},$$

where $t_j = 0, 10, 20, , 160$, and T is the period, and where it is assumed that $e_{jt} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. To estimate the period T of the data, we first fixed T at its lower limit T_0, and then calculated the Least Squares (LS) estimate and its mean squared error. We then increased T_0 by 1 to get a new T, and then calculated the LS estimate and its MSE. This was repeated until a reasonable upper limit of T, $T_1(> T_0)$, was obtained. Comparing all the MSE's, the LS estimate of T corresponding to the minimum MSE is taken as our estimated period T. As these time course data consist of 17 time points of ten minute intervals (starting from zero), we took $T_0 = 10$.

Using the dataset of 384 genes posted by Yeung et al. (2001), we obtained an estimated cell cycle period of 73 min, assuming the initial phase to be zero. As a period of 73 min is about half of 160 min, it would seem to be a reasonable estimate. Also, since the 237 cell cycle data is a subset of the 384 cell cycle data, we assume here that it follows the same time cycle of 73 minutes.

The model (9) was fitted with $\beta_i = (\beta_{1i}, \beta_{2i})^T$ as the fixed-effects vector for the ith component and with the tth row of the design matrix $\boldsymbol{X}$, corresponding to the time point t, given by

$$(cos(2\pi t/T) \qquad sin(2\pi t/T)) \tag{11}$$

for $t = 1, \ldots, p$. The design matrix $\boldsymbol{U}$ was taken to be $\mathbf{1_p}$ (that is, $q_b = 1$) with $\boldsymbol{b}_{ij} = b_{ij}$, the common random effect for all time points shared by the jth gene, and $\boldsymbol{H}_i = \boldsymbol{I}_p$. The cluster-specific random effect $\boldsymbol{c}_i$ was specified as $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{ip})^T$ with $q_c = p$ and $\boldsymbol{V} = \boldsymbol{I}_p$. With respect to the error terms, we took $\boldsymbol{W} = \boldsymbol{I}_p$ with $q_e = p$.

Concerning the number of components, we report in Table 1 the values of BIC obtained for various levels of the number of components g. As we were unable to calculate the likelihood exactly under the model (9) in the case of nonzero cluster-specific random-effects terms $\boldsymbol{c}_i$, we approximated it by taking the gene-profile vectors to be independently distributed in forming the log likelihood in calculating the value of BIC. According to the tabulated values of BIC in Table 1, we should choose $g = 4$ components, which agrees with the number of MIPS functional groups in these genes.

Table 1. Values of BIC for Various Levels of the Number of Components g.

The	Number	of	Components		
2	3	4	5	6	7
10883	10848	10837	10865	10890	10918

For $g = 4$, we found that the estimated variance $\boldsymbol{\theta}_{ci}$ for the cluster-specific random-effects term was equal to 0.227, 0.280, 0.043, and 0.137, which indicates some level of correlation within at least three of the four clusters. The Rand Index and its adjusted value were equal to 0.7808 and 0.5455, which compare favourably to the corresponding values of 0.7087 and 0.3697, as obtained by Wong et al. (2007) for their method. On permuting the cluster labels to minimize the error rate of the clustering with respect to the four MIPS functional groups, we obtained an error rate of 0.291. We also clustered the genes into four clusters by not having cluster-specific random-effects terms $\boldsymbol{c}_i$ in (9), yielding lower values of 0.7152 and 0.4442 for the Rand Index and its adjustment. The estimated error rate was equal to 0.316. Hence in this example, the use of cluster-specific random-effects terms leads to a

clustering that corresponds more closely to the underlying functional groups than without their use.

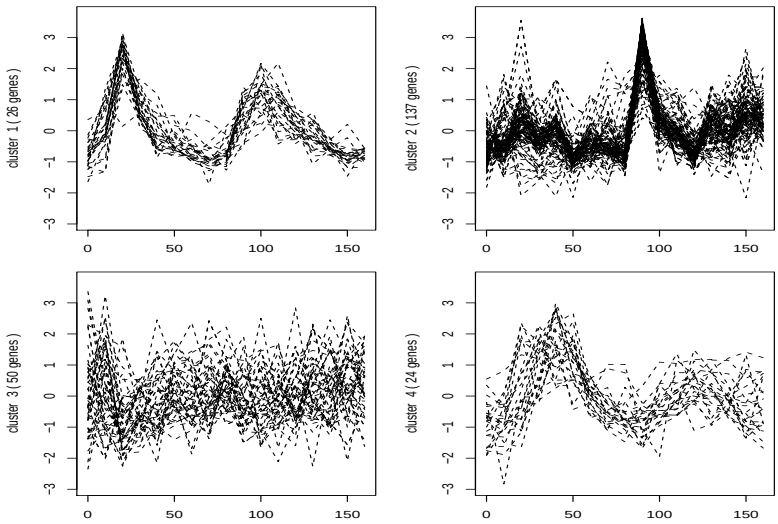

Fig. 1. Clusters of Gene-Profiles Obtained by Mixture of Linear Mixed Models with Cluster-Specific Random Effects.

The clustering obtained in the latter case, however, is still superior in terms of the Rand Index and its adjusted value for the two-stage method of Wong et al. (2007), which was the best on the basis of these criteria in their comparative analysis. We also fitted the mixed linear model mixture (9)

Table 2. Summary of Clustering Results for $g = 4$ Clusters.

Model	Rand Index	Adjusted Rand Index	Error Rate
1	0.7808	0.5455	0.291
2	0.7152	0.4442	0.316
3	0.7133	0.3792	0.4093
Wong	0.7087	0.3697	Not available

without the sine-cos regression model (10) for the mean, but with a separate (fixed effects) term at each of the $p = 17$ time points; that is, we set $\boldsymbol{X} = \boldsymbol{I}_p$ and took β_i to be a p-dimensional vector of fixed effects. We did not include cluster-specific random-effects terms $\boldsymbol{c}_i$ due to their nonidentifiability in this case. This nonregression model gave worse results for the Rand Index and the error rate than with the regression model (9) using the sine-cos curve to

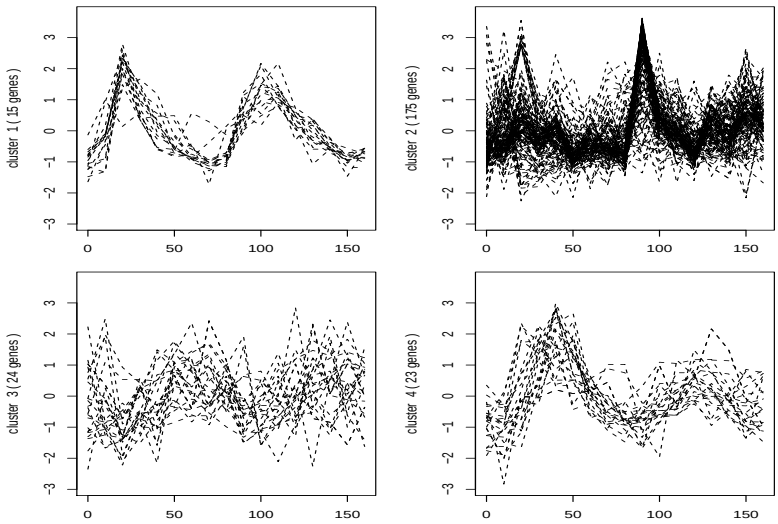

Fig. 2. Clusters of Gene-Profiles Obtained by Mixture of Linear Mixed Models without Cluster-Specific Random Effects.

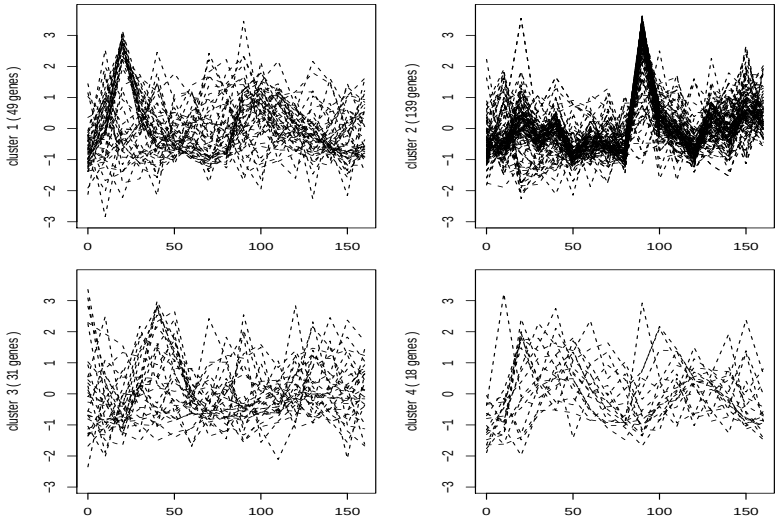

Fig. 3. Plots of Gene Profiles Grouped According to Their True Functional Grouping.

specify the mean at a given time point. The results for this nonregression version are listed under Model 3 in Table 2, where the clustering results have been summarized. In this table, Models 1 and 2 correspond to the use of the regression model (9) with and without cluster-specific random-effects terms.

In Figures 1 and 2, we give the plots of the gene profiles as clustered into $g = 4$ clusters as obtained by fitting the mixture of linear mixed models (9) with and without cluster-specific random-effects terms $\boldsymbol{c}_i$. In Figure 3, the plots of the gene profiles are grouped according to their actual functional grouping.

5 Example 2: Simulated data

Another example in Wong et al. (2007) is a simulated data from Michaud et al. (2003). Using EMMIX-WIRE, we have recognized all nine patterns perfectly correct (See Figure 4). Rand index is 1, error rate is zero. Wong et al.(2007) got their Rand Index as 0.9961 and Adjusted Rand Index as 0.9806.

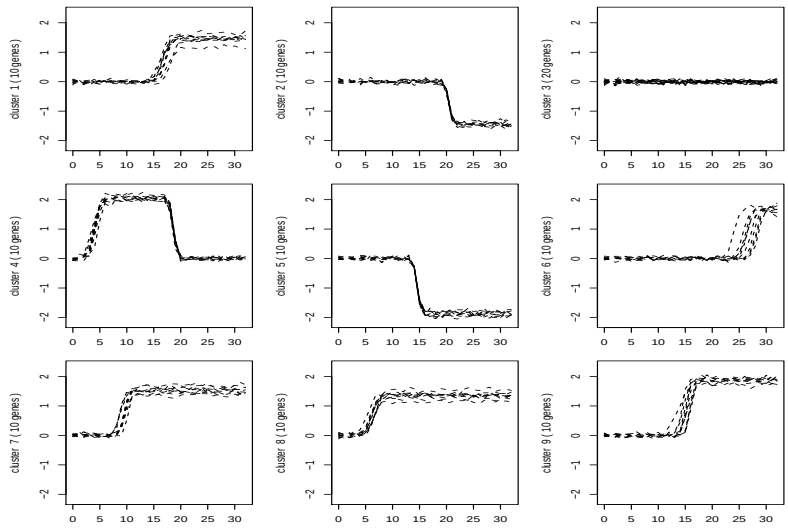

Fig. 4. Clusters of Simulated Gene-Profiles Obtained by Mixture of Linear Mixed Models with Cluster-Specific Random Effects.

References

BANFIELD, J.D. and RAFTERY, A.E. (1993): Model-based Gaussian and non-Gaussian clustering. *Biometrics 49 803–821.*

CHO, R.J., HUANG, M., CAMPBELL, M.J. et al. (2001): Transcriptional regulation and function during the human cell cycle. *Nature Genetics 27 48–54.*

DAY, N.E. (1969): Estimating the components of a mixture of two normal distributions. *Biometrika 56 463–474.*

BOOTH, J.G., CASELLA, G. and HOBERT, J.P. (2008): Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society B 70 119–139.*

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.*

DEMPSTER, A.P., LAIRD, N.M., and 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.*

FRALEY, C. and RAFTERY, A.E. (1998): How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal 41 578–588.*

FRALEY, C. and RAFTERY, A.E. (2002): Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97 611–631.*

GANESALINGAM, S. and McLACHLAN, G.J. (1978): The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika 65 658–662.*

LUAN, Y. and LI, H. (2003): Clustering of time-course gene expression data using a mixed-effects model with *B*-splines. Bioinformatics 19 474–482.

McLACHLAN, G.J. (1982): The classification and mixture maximum likelihood approaches to cluster analysis. In: P.R. Krishnaiah and L. Kanal (Eds.): *Handbook of Statistics Vol. 2.* North-Holland, Amsterdam, 199–208.

McLACHLAN, G.J. and BASFORD, K.E. (1988): *Mixture Models: Inference and Applications to Clustering.* Dekker, New York.

McLACHLAN, G.J., BEAN, R.W., and PEEL, D. (2002): A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics 18g 413–422.*

McLACHLAN, G.J., DO, K.-A., and AMBROISE, C. (2004): *Analyzing Microarray Gene Expression Data.* Wiley, Hoboken, New Jersey.

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

MICHAUD, D.J., Marsh, A.G., and Dhurjati, P.S. (2003): eXPatGen: generating dynamic expression patterns for the systematic evaluation of analytical methods. *Bioinformatics 19 1140–1146.*

NG, S.K., McLACHLAN, G.J., WANG, K., BEN-TOVIM JONES, L., and Ng, S.W. (2006): A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics 22 1745–1752.*

QIN, L.-X., and SELF, S.G. (2006): The clustering of regression models method with applications in gene expression data. *Biometrics 62 526–533.*

WOLFE, J.H. (1965): A computer program for the computation of maximum likelihood analysis of types. *U.S. Naval Personnel Research Activity, Technical Report, SRM 65-112, San Diego.*

WONG, D.S.V., WONG, F.K., and WOOD, G.R. (2007): A multi-stage approach to clustering and imputation of gene expression profiles. *Bioinformatics 23 998–1005.*

YEUNG, K.Y. , FRALEY, C., MURUA, A., RAFTERY, A.E., and RUZZO, W.L. (2001): Model-based clustering and data transformations for gene expression data. *Bioinformatics 17 977–987.*

Part XII

Multiple Testing Procedures

Testing Effects in ANOVA Experiments: Direct Combination of all Pair-Wise Comparisons Using Constrained Synchronized Permutations

Dario Basso[1], Fortunato Pesarin[2], and Luigi Salmaso[1]

[1] Department of Management and Engineering, University of Padova, Str.lla S. Nicola 14, 36100 Vicenza, Italy, *dario@stat.unipd.it*, *salmaso@gest.unipd.it*
[2] Department of Statistics, University of Padova, Via Cesare Barttisti 241-243, 35100 Padova, Italy, *fortunato.pesarin@unipd.it*

Abstract. Synchronized permutation tests have been introduced to test for main effects and interaction in two-way ANOVA problems (Pesarin, 2001, Salmaso, 2003, Basso et al., 2007). The permutation space is restricted in order the test statistic for each factor or interaction to depend only on the true effects (if any) under testing and on exchangeable errors. The proposed test statistics can be viewed as a direct combination (i.e. the sum) of partial statistics to perform all pair-wise comparisons between the effects of a main factor. Note that the tests for each comparison are dependent. Constrained synchronized permutations allow to test all pair-wise comparisons between pairs of effects of a main factor. Simultaneous and individual confidence intervals for each comparison can be provided through CSP tests.

Keywords: confidence intervals, dependent tests, multiple comparisons.

1 Introduction

In all ANOVA problems the main interest is on testing the null hypothesis of no treatment effects. In case the null hypothesis is rejected, the aim of the researcher is to find out which levels of the treatments caused the rejection of the null hypothesis. This is usually done *after* the global test, and it is commonly known as *post hoc comparisons*. Here on, by global test we mean a test which is suitable to test for the null hypothesis involving all treatment levels. The post hoc comparisons are carried out by considering all possible pair-wise comparisons between pairs of treatment effects. A correction accounting for multiplicity is usually adopted, as in the well known Tukey honest significant difference (Hsu, 1996, Dean and Voss, 1999), which is based on Student's t test.

As a matter of fact, the tests for each pair-wise comparison are dependent. In this work we introduce a permutation test which is suitable to test for

main effects and interaction in two-way ANOVA. The test for main effects can be decomposed into several partial tests which are suitable for pair-wise comparisons. A restricted kind of permutation, introduced by Pesarin (2001) and Salmaso (2003) is adopted in order the partial test to depend only on the specific effects under study. This kind of randomization, named *synchronized permutations*, allow us to account for dependence among $K \geq 2$ partial tests. Simultaneous confidence intervals can be provided for each pair-wise comparison in order to easily understand which effects led to the rejection of the global null hypothesis, as in Basso et al. (2007).
The problem when dealing with (exact) permutation tests under the two-way ANOVA model assumptions is that units from different blocks are not exchangeable, since they may differ in their expected values (which depend on main effects and interaction effects not being tested). However, a sufficient condition in order to apply a permutation test is that the test statistic (e.g. to test for factor A) has a discrete uniform null distribution over is support. Another requirement in order to obtain separate inferences (for each main factor and interaction effects) is that the test statistic depends only on the effects of interest. This can be done by adopting a restricted kind of permutation, and by applying the side-conditions on nuisance effects, as will be shown in Section 2.

2 Partial tests for each pair-wise comparison

Let y_{ijk} $(i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, n)$ denote the generic response element on a balanced $I \times J$ ANOVA model. That is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where μ is the population mean, α_i and β_j are main effects, γ_{ij} are interaction effects and ε_{ijk}'s are exchangeable errors with zero mean and finite variance σ^2. We also assume that side-conditions $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$ hold. Let us first consider the problem of testing $H_{0A}^{is} : \alpha_i = \alpha_s$ against the alternative $H_{1A}^{is} : \alpha_i \neq \alpha_s$, $1 \leq i < s \leq I$. Here on, we will refer to H_{0A}^{is} as a *partial hypothesis* and we define a *partial test* as a suitable test to assess H_{0A}^{is}. The response elements in blocks A_iB_j and A_sB_j depend on the effects of interest α_i and α_s and on nuisance effects β_j, γ_{ij} and γ_{sj}. In order to obtain a separate inference for H_{0A}^{is}, we need the partial test statistic to depend only on the effects of interest. To this end, note that the statistic:

$$\begin{aligned} {}^{A}T_{is|j} &= \sum_k y_{ijk} - \sum_k y_{sjk} \\ &= n\left(\alpha_i - \alpha_s + \gamma_{ij} - \gamma_{sj} + \bar{\varepsilon}_{ij\cdot} - \bar{\varepsilon}_{sj\cdot}\right), \end{aligned} \tag{1}$$

does not depend on the nuisance effect β_j. Here $\bar{\varepsilon}_{ij\cdot}$ and $\bar{\varepsilon}_{sj\cdot}$ are sampling means of n exchangeable errors. The side conditions on the interaction effects

allow us to obtain a test statistic to assess H_{0A}^{is} which does not depend on γ_{ij} and γ_{sj}. If we sum ${}^{A}T_{is|j}$ over index j:

$$\begin{aligned} {}^{A}T_{is} = \sum_j {}^{A}T_{is|j} &= nJ(\alpha_i - \alpha_s) + \sum_j \gamma_{ij} - \sum_j \gamma_{sj} + \sum_j \bar{\varepsilon}_{ij\cdot} - \sum_j \bar{\varepsilon}_{sj\cdot} \\ &= nJ[\alpha_i - \alpha_s] + nJ[\bar{\varepsilon}_{i\cdot\cdot} - \bar{\varepsilon}_{s\cdot\cdot}], \end{aligned} \tag{2}$$

we obtain a test statistic which only depends on the effects of interest and on a linear combination of exchangeable errors. Therefore ${}^{A}T_{is}$ is a suitable partial test statistic for H_{0A}^{is} within a permutation framework.
In order to obtain the permutation null distribution of ${}^{A}T_{is}$ some care is needed: the response elements are not exchangeable with respect to different treatments of factor B, because of potential active effects β_j, $j = 1, \ldots, J$. This means that only permutations within blocks sharing the same level of factor B are allowed. Moreover, consider a permutation between the response elements of blocks A_iB_j and A_sB_j, and denote as y_{ijk}^* and y_{sjk}^* the generic response elements of those blocks after a random permutation of $\mathbf{y}_{is|j} = [\mathbf{y}_{ij}, \mathbf{y}_{sj}]'$. This means that $0 \leq \nu_{is|j}^* \leq n$ units have been exchanged between blocks A_iB_j and A_sB_j. The permutation structure of ${}^{A}T_{is|j}$ (the statistic (1) computed on the permuted data) is:

$$\begin{aligned} {}^{A}T_{is|j}^* &= (n - \nu_{is|j}^*)(\alpha_i + \gamma_{ij}) + \nu_{is|j}^*(\alpha_s + \gamma_{sj}) + n[\beta_j + \bar{\varepsilon}_{ij\cdot}^*] \\ &\quad - (n - \nu_{is|j}^*)(\alpha_s + \gamma_{sj}) - \nu_{is|j}^*(\alpha_i + \gamma_{ij}) - n[\beta_j + \bar{\varepsilon}_{sj\cdot}^*] \\ &= (n - 2\nu_{is|j}^*)[(\alpha_i - \alpha_s) + (\gamma_{ij} - \gamma_{sj})] + n[\bar{\varepsilon}_{ij\cdot}^* - \bar{\varepsilon}_{sj\cdot}^*]. \end{aligned}$$

Let us consider J random permutations between blocks A_iB_j and A_sB_j for $j = 1, \ldots, J$ (i.e. $\nu_{is|j}^*$ depends on the index j). The partial test ${}^{A}T_{is}$ computed on the permuted observations is:

$$\sum_j {}^{A}T_{is|j}^* = K[\alpha_i - \alpha_s] + \sum_j (n - 2\nu_{is|j}^*)(\gamma_{ij} - \gamma_{sj}) + nJ[\bar{\varepsilon}_{i\cdot\cdot}^* - \bar{\varepsilon}_{s\cdot\cdot}^*], \tag{3}$$

where $K = nJ - 2\sum_j \nu_{is|j}^*$ and $\bar{\varepsilon}_{i\cdot\cdot}^*$ and $\bar{\varepsilon}_{s\cdot\cdot}^*$ are sampling means of nJ exchangeable errors. Clearly, (3) does not depend on nuisance effects γ_{ij} and γ_{sj} if and only if $\nu_{is|j}^*$ does not depend on index j. That is, if we let $\nu_{is|j}^* = \nu_{is}^*$ $\forall j$, then:

$${}^{A}T_{is}^* = (n - 2\nu_{is}^*)J[\alpha_i - \alpha_s] + nJ[\bar{\varepsilon}_{i\cdot\cdot}^* - \bar{\varepsilon}_{s\cdot\cdot}^*]. \tag{4}$$

This means that, if H_{0A}^{is} is true, ${}^{A}T_{is}$ and ${}^{A}T_{is}^*$ are realizations of the same random variable, since the errors are assumed to be exchangeable. Therefore, the null permutation distribution of the partial test statistic ${}^{A}T_{is}$ can be obtained by considering all possible permutations between pairs of blocks A_iB_j and A_sB_j ($j = 1, \ldots, J$) involving the same number of exchanges ν_{is}^* in each pair and by computing each time the test statistic (4). Because of that, this restricted kind of permutation is called 'synchronized', since they

require the same number of units to be exchanged in each pair of blocks. This is necessary in order to obtain a null distribution which does not depend on the nuisance effects.
The cardinality of the support of the partial test statistic depends on how the synchronized permutations are obtained, as will be discussed in next section. It is straightforward to see that the null distribution of (2) is symmetric with respect to zero. Let C be the cardinality of the support of (4), then define the partial p-value for two-sided alternatives as:

$$ {}^{A}p_{is} = 2\min\left\{\frac{1}{C}\sum_{b=1}^{B} I({}^{A}T^{*}_{is} \leq {}^{A}T_{is}), \frac{1}{C}\sum_{b=1}^{B} I({}^{A}T^{*}_{is} \geq {}^{A}T_{is})\right\}, $$

where $I(\cdot)$ is the indicator function. Of course, there are other test statistics leading to the same inference on $\alpha_i - \alpha_s$: a permutationally equivalent statistic is for instance ${}^{A}T_{is} = \bar{y}_{i\cdot\cdot} - \bar{y}_{s\cdot\cdot}$, where $\bar{y}_{i\cdot\cdot} = (nJ)^{-1}\sum_j \sum_k y_{ijk}$. We will consider this partial test statistic in order to determine permutation tests and confidence intervals for $\alpha_i - \alpha_s$.
Similarly we can define test statistics and permutation strategies to test for the partial hypothesis $H^{jh}_{0B} : \beta_j = \beta_h$. Here the same number of units ν^{*}_{jh} must be exhanged between pairs of blocks A_iB_j and A_iB_h, $i = 1, \ldots, I$. The partial test statistic for one pair of factor B effects is:

$$ {}^{B}T^{*}_{jh} = \sum_i \left[\sum_k y^{*}_{ijk} - \sum_k y^{*}_{ihk}\right] = (n - 2\nu^{*}_{jh})(\beta_j - \beta_h) + nI(\bar{\varepsilon}^{*}_{\cdot j\cdot} - \bar{\varepsilon}^{*}_{\cdot h\cdot}). $$

Suitable tests for interaction effects have also been developed, but this topic falls out the scope of this paper, and we refer to Pesarin (2001) for further discussion.

3 Global tests for main factor hypotheses

In the previous section we have introduced some partial tests which are suitable for testing partial null hypotheses H^{is}_{0A} (or H^{jh}_{0B}). Since the aim of ANOVA analysis is to determine whether there is at least one active (i.e. non null) effect, the partial tests should be combined togheter in order to obtain a global test to assess the global null hypothesis $H_{0A} : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$. Note that if the global null hypothesis holds, then H^{is}_{0A} is true for any couple of indeces $1 \leq i < s \leq I$. Therefore we may also write:

$$ H_{0A} : \bigcap_{1\leq i<s\leq I} H^{is}_{0A}. $$

On the contrary, H_{0A} must be rejected when at least one of the partial null hypothesis is rejected. A suitable combining test accounting for the significance of the partial test should therefore be applied in order to assess H_{0A}.

This can be done adopting the nonparametric combination of dependent tests (Pesarin, 2001). The nonparametric combination (NPC) is a continuous function whose arguments are monotonic functions of the partial test statistics, such as the test statistics themselves or their related p-values. In our testing problem, we have $K = I(I-1)/2$ partial hypotheses to be combined. Let λ_k, $k = 1, \ldots, K$ be a partial test statistic (or its related p-value): a suitable NPC function ψ should satisfy the following properties:

1. It should be continuous in all its K arguments;
2. It should be non-decreasing on all its arguments, i.e.:

$$\psi(\lambda_1, \lambda_2, \ldots, \lambda_k, \ldots, \lambda_I) \leq \psi(\lambda_1, \lambda_2, \ldots, \lambda'_k, \ldots, \lambda_I)$$

   whenever λ'_k is at least as significant as λ_k against the related partial null hypothesis;
3. It should reach its supremum (possibly not finite) when a least one of its arguments leads to reject the related partial null hypothesis almost surely.

There are many functions which satisfy the above requirements. For example, if $\lambda_k = {}^A T_{is}$, then some suitable combining functions are:

- the `maxT` combining function: $\psi = \max_{1 \leq i < s \leq I} {}^A T_{is}{}^2$;
- the `Direct` combining function: $\psi = \sum_{1 \leq i < s \leq I} {}^A T_{is}{}^2$.

Other suitable combining function, when the arguments are the partial p-values ${}^A p_{ij}$, are:

- the `minP` combining function: $\psi = \min_{1 \leq i < s \leq I} {}^A p_{is}$;
- the `Fisher` combining function: $\psi = -2 \sum_{1 \leq i < s \leq I} \log({}^A T_{is})$;
- the `Liptak` combining function: $\psi = \sum_{1 \leq i < s \leq I} [1 - \phi(\log {}^A T_{is})]$, where $\phi(\cdot)$ is the inverse cumulative distribution function of a standard normal random variable;

Focusing the attention on the direct combining function, we have that ${}^A T_{is}{}^2$ is significant against H_{0A}^{is} for large values. Let $I = 3$, then a suitable global test for H_{0A} is:

$$ {}^A T = {}^A T_{12}^2 + {}^A T_{13}^2 + {}^A T_{23}^2. \tag{5}$$

Clearly, (5) is continuous and non-decreasing on all its arguments (properties 1 and 2), and it reaches its maximum when ${}^A T_{is}{}^2 = \max {}^A T_{is}^{*\,2} \; \forall\, i,\, s$, i.e. when all partial tests are extremely significant against the related null hypothesis. Focusing the attention on the Fisher combining function, let $\lambda_k = {}^A p_{is}$, then we may also define as a global test:

$$ {}^A T = -2[\log({}^A p_{12}) + \log({}^A p_{13}) + \log({}^A p_{23})].$$

The above global test still satisfies the required properties $1 \rightarrow 3$. In this case, the combinig function reaches its supremum (not finite) when at least one

partial p-value tends to zero. In both the considered examples, large values of ${}^{A}T$ are significant against the global null hypothesis H_{0A}.
Note that the partial tests in (5) are dependent, even though the orignal observations are independent. To account for this dependence, a suitable permutation strategy is the following: let $\mathbf{y}^*_{is|1} = \pi_1[\mathbf{y}_{i1}, \mathbf{y}_{s1}]'$ be a random permutation of observations in blocks A_iB_1 and A_sB_1. If the same permutation is applied to the remaining $J-1$ pairs of blocks (i.e. if $\mathbf{y}^*_{is|j} = \pi_j[\mathbf{y}_{ij}, \mathbf{y}_{sj}]' \, \forall j$), then the dependence structure among the partial test statistics is maintained. This entails the cardinality of the permutation null distribution of each partial test statistic to be equal to:

$$C = \binom{2n}{n},$$

since the number of exchanges in each pair of blocks is entirely determined by π_1. This kind of permutations are called *Constrained Synchronized Permutations* (CSPs), since units in each pair of blocks are swapped according to their original positions within each block. Another possible choice is to swap the same number of units in each pair of blocks independently, provided that the same number of exchanges has to be made in each pair of blocks. This kind of permutations are called *Unconstrained Synchronized Permutations* (USPs). It is easy to understand that the dependence among the partial test statistics is not maintained by the USPs. Figure 3 represents the projections of the permutation space points $[{}^{A}T^*_{12}, {}^{A}T^*_{13}, {}^{A}T^*_{23}]$ when CSPs (black dots) and USPs (grey dots) are applied in a 3×2 ANOVA model with $n = 4$. The correlations between partial tests are evident when the CPSs are applied. USPs are suitable when only few replicates are available (say $n \leq 3$) because the number of available permutations rapidly increases with n, I and J. This fact is directly related to the minimum achivable significance level, which, in permutation tests, is equal to $1/C$ (for CSPs).
The algorithm to perform partial tests to assess each H^{is}_{0A} and a global test on the effects of factor A adopting the direct combining function and CSPs is as follows:

- Let $\mathit{\Pi} = [\pi_1, \pi_2, \ldots, \pi_C]$, $C = \binom{2n}{n}$, be the set of all possible rearrangements of $2n$ observations into groups of n, and let π_1 be the identity transformation: $\pi_1(1, \ldots, 2n) = [1, \ldots, 2n]$;
- for $c = 1, \ldots, C$, repeat the following steps:

  1. Obtain $\mathbf{y}^*_{is|j} = [\mathbf{y}^*_{ij}, \mathbf{y}^*_{sj}]'$, where $\mathbf{y}^*_{is|j} = \pi_c(\mathbf{y}_{is|j})$ and $\mathbf{y}_{is|j} = [\mathbf{y}_{ij}, \mathbf{y}_{sj}]'$; $\mathbf{y}_{is|j}$ is the pooled vector of observations from blocks A_iB_j and A_sB_j, $j = 1, \ldots, J$;
  2. Compute the permutation values of the partial test statistics:

$$ {}^{A}T^{*c}_{is} = \left(\bar{y}^*_{i\cdot\cdot} - \bar{y}^*_{s\cdot\cdot}\right)^2 \qquad 1 \leq i < s \leq I; $$

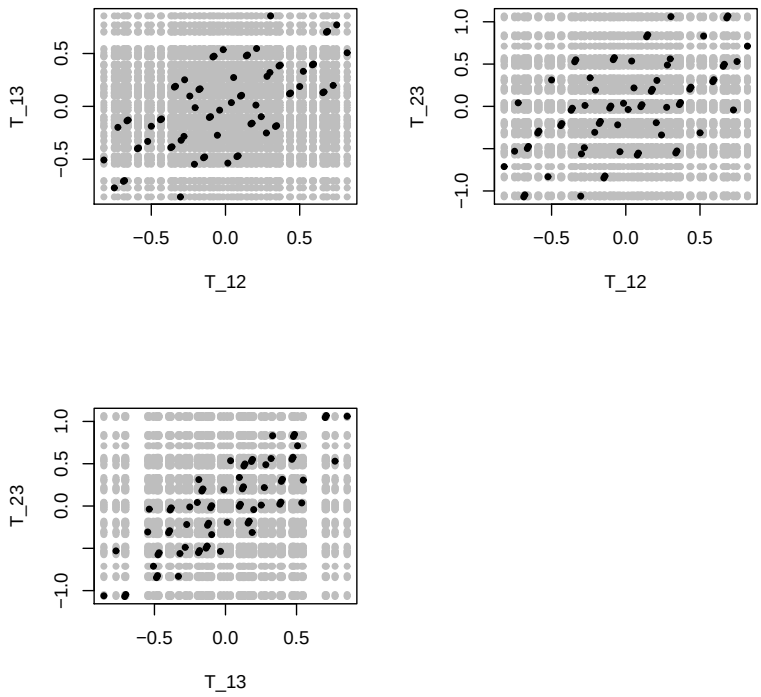

Fig. 1. Projections of the permutation points $[{}^{A}T^{*}_{12}, {}^{A}T^{*}_{13}, {}^{A}T^{*}_{23}]$ in a 3×2 design with $n = 4$. Black dots = CSPs; Grey dots = USPs.

3. Compute the permutation value of the global test statistic as:

$$ {}^{A}T^{*c} = \sum_{i<s} {}^{A}T^{*c}_{is} $$

- obtain the partial p-values to assess H^{is}_{0A} as:

$$ {}^{A}p_{is} = \frac{1}{C} \sum_{c=1}^{C} I({}^{A}T^{*c}_{is} \geq {}^{A}T^{*1}_{is}); $$

- obtain the global p-value to assess H_{0A} as:

$$ {}^{A}p = \frac{1}{C} \sum_{c=1}^{C} I({}^{A}T^{*c} \geq {}^{A}T^{*1}); $$

Be reminded that ${}^{A}T^{*1}_{is} = {}^{A}T_{is}$ and ${}^{A}T^{*1} = {}^{A}T$, i.e. the first elements of the (partial and global) permutation distributions are the observed values of the test statistics because they are obtained from the observed data. A

multiplicity problem arises since we are dealing with C partial tests. A suitable multiplicity correction should therefore be applied to the partial test in order to control the family-wise error rate (FWE). In this work we will consider the classic Bonferroni correction, although other corrections may be adopted. Partial and global tests are exact since, under partial and global null hypotheses, the related test statistics only depend on combinations of exchangeable errors. Global and partial tests can be viewed similarly to the F test and post-hoc comparison tests in the parametric framework. Moreover, they can be performed simoultaneously.

4 Individual and simultaneous pseudo confidence intervals

The nonparametric combination of dependent tests allow us to perform partial and global tests at the same time, and a quick analysis tool can be provided by drawing the confidence intervals for all pairwise differences of main effects. The upper bound of the CSP confidence intervals for $\alpha_i - \alpha_s$ can be provided by applying the following algorithm:

- Choose the desired confidence level $1 - \alpha$ and degree of precision ϵ; α should be multiple of $1/C$;
- Choose $\delta_{is} = \alpha_i - \alpha_s > 0$; let:

$$\begin{aligned} \tilde{y}_{ijk} &= y_{ijk} \\ \tilde{y}_{sjk} &= y_{sjk} - \bar{y}_{s\cdot\cdot} + \bar{y}_{i\cdot\cdot} + \delta_{is}; \end{aligned}$$

- Obtain the observed value of the test statistic:

$$^{A}\tilde{T}(\delta_{is}) = \bar{\tilde{y}}_{s\cdot\cdot} - \bar{\tilde{y}}_{i\cdot\cdot} = \delta_{is}$$

- repeat C times:
  1. Apply the CSPs to each pair of block A_iB_j and A_sB_j. Let $\tilde{y}^*_{ijk}$ and $\tilde{y}^*_{sjk}$ be the observations in blocks A_iB_j and A_sB_j after a CSP has been performed for $j = 1, \ldots, J$.
  2. Compute the value of the test statistic according to the constrained synchronized permutation methodology:

$$^{A}\tilde{T}^*(\delta_{is}) = \bar{\tilde{y}}^*_{s\cdot\cdot} - \bar{\tilde{y}}^*_{i\cdot\cdot}$$

- Obtain the p-value of $^{A}\tilde{T}(\delta_{is})$:

$$^{A}\tilde{p}(\delta_{is}) = \frac{\#[^{A}\tilde{T}^*(\delta_{is}) \geq {}^{A}\tilde{T}(\delta_{is})]}{C}$$

- if the condition:

$$|^{A}\tilde{p}(\delta_{is}) - \alpha/2| < \epsilon$$

  is satisfied, then $\bar{y}_{s\cdot\cdot} - \bar{y}_{i\cdot\cdot} + \delta_{is}$ is the upper bound of the permutation confidence interval. Otherwise, repeat the algorithm by increasing δ_{is} until the above condition is satisfied.

In order to obtain the lower bound of the confidence interval, just repeat the same algorithm by adding a negative δ_{is} and by computing the p-value as ${}^{A}\tilde{p}(\delta_{is}) = \#[{}^{A}\tilde{T}^{*}(\delta_{is}) \leq {}^{A}\tilde{T}(\delta_{is})]C^{-1}$. The previous algorithm is motivated by the fact that the permutation null distribution depends on δ_{is}.
Note that the confidence interval and the permutation test are neither obtained by the same permutation approach nor with the same data. This is because the null distribution of the test statistic (4) implicitly assumes that H_{0A}^{is} is true, otherwise exchangeability does not hold. Instead, the confidence interval must be derived for whatever δ_{is}, not necessarely when $\delta_{is} = 0$ (i.e., the necessary condition to apply the test). Data of one sample in the confidence interval algorithm are first centered in order the variance of the samples to depend not on δ_{is} (just like the pooled variance of the t-test for a two sample problem does not depend on δ_{is}), then δ_{is} is added to these data. This means that there is no one-to-one correspondence between the statistical test and the confidence interval, as usually happens when the parametric approach is applied.
However, it is possible to define a 'pseudo' confidence interval by obtaining the permutation distribution of the test statistic (4) and by centering it on the observed difference $\bar{y}_{i\cdot\cdot} - \bar{y}_{s\cdot\cdot}$. In this way, there is a one-to-one correspondence between the permutation test and the pseudo-CI. By that, we mean that the permutation test rejects the null hypothesis if and only if the 'pseudo' confidence interval does not contain zero.
Let ${}^{A}\tilde{C}I_{1-\alpha}^{is}$ be the pseudo CI obtained from the corresponding permutation test at a significance level α, then a (non randomized) global test can be also defined as a function of ${}^{A}\tilde{C}I_{1-\alpha}^{is}$'s:

$$ {}^{A}\tilde{T} = 1 - \prod_{i<s} I\left(0 \ni {}^{A}\tilde{C}I_{1-\alpha}^{is}\right), $$

where ${}^{A}\tilde{T} = 1$ means the rejection of the global null hypothesis. If all ${}^{A}\tilde{C}I_{1-\alpha}^{is}$'s are obtained with the same confidence level $1-\alpha$, then:

$$ P[{}^{A}\tilde{T} = 1] = 1 - (1-\alpha)^{K}, $$

therefore by applying Bonferroni's correction to the partial test significance levels, we can obtain a simultaneous acceptance region for the global null hypothesis. A suitable representation is shown in Figure 2, which is referred to a simulation under H_{0A} in a 3×2 balanced design with $n = 4$. This representation has been suggested by Hsu (1996). The segments represent the ${}^{A}\tilde{C}I_{1-\alpha}^{is}$'s: as long as every segment crosses the 45° line the global null hypothesis is not rejected at a significance level $\alpha^{G} = 1-(1-\alpha)^{K}$. The symbols m_1, m_2 and m_3 indicates the observed values of $\hat{\delta}_{is}$, $1 \leq i < s \leq 3$.
In order to control the FWE at an α^{G} significance level, a Bonferroni correction can be applied by performing each partial test at a significance level

$\alpha = 1 - (1 - \alpha^G)^{\frac{1}{K}}$. Individual pseudo CI's can be provided by not considering the correction for multiplicity, and by performing each partial test with a significance level equal to α.
The discrete nature of permutation null distributions should be recalled. Since not all significance levels are attainable, it is not possible to achieve the exact confidence level when n is too small (say $n = 4, 5$), however, with $n = 6$ the (partial) permutation distribution consists in 924 distinct values, therefore the usual nominal significance levels can be adopted (see next section).

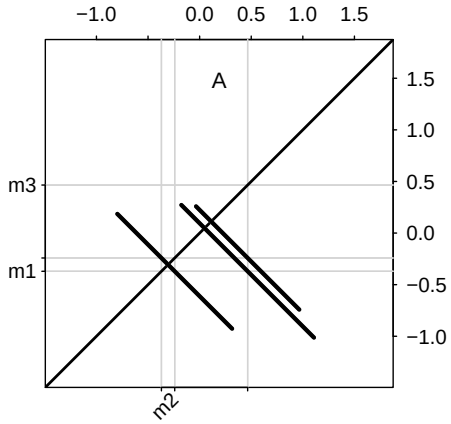

Fig. 2. Representation of the 'pseudo' - confidence intervals due to Hsu (1996).

5 Simulation study

In this section we investigate the behaviour of the CSP tests when $I = 3$, $J = 2$ and $n = 4, 6$. The proposed testing procedure is then compared to Tukey's well known honest significant method. All simulations were implemented with the `R` software and involved 1000 independent data generations from a standard normal distribution and from a Student's t distribution with 2 d.f. The last choice does not satisfy the assumption of finite error variance, which, however, is not necessary in order the permutation test to be unbiased.
When $n = 4$, the attainable confidence levels of the partial tests are multiple of $1/35$ (since the alternative is always two-sided). Thus, we have considered $34/35, 33/35$ and $31/35$ as nominal confidence levels, which are reported in the $1 - \alpha$ column of the result tables. The attainable significance levels of the global test are multiple of $1/35$ as well, therefore it is not possible to achieve the exact global theoretical confidence level $1 - \alpha^G = (1 - \alpha)^3$.
Instead, when $n = 6$, the cardinality of the permutation distribution elements

is high enough to achieve the nominal confidence levels almost exactly, therefore we have set the nominal significance levels for the global test equal to 0.01, 0.05, 0.1, and applied a Bonferroni correction to the significance levels of the partial tests.

Each table reports the results of the comparison between CSP and Tukey tests for a given sample size and error distribution. The true difference among pairs of effects are reported in the second line of each table (so that H_{0A}^G holds on the left side of the tables). For each nominal confidence level $1-\alpha$ (in bold), the proportion of times a partial ${}^A\tilde{C}I_{1-\alpha}^{is}$ strictly contained zero is reported. The same results are shown for the global test, which are the proportion of times that all the (pseudo) partial CIs (strictly) contained zero. Note that the displayed results are equal to $1-\hat{\pi}(\delta)$, where $\hat{\pi}(\delta)$ is the observed power of the related test.

$1-\alpha$	δ_{12}	δ_{13}	δ_{23}	$1-\alpha^G$		$1-\alpha$	δ_{12}	δ_{13}	δ_{23}	$1-\alpha^G$	
CSP	0	0	0	$\hat{P}[{}^A\tilde{T}=0]$		CSP	1	1	0	$\hat{P}[{}^A\tilde{T}=0]$	
0.971	0.962	0.972	0.970	**0.917**	0.917	**0.971**	0.757	0.747	0.975	**0.917**	0.607
0.943	0.933	0.943	0.916	**0.838**	0.844	**0.943**	0.601	0.597	0.931	**0.838**	0.423
0.886	0.888	0.879	0.867	**0.695**	0.747	**0.886**	0.413	0.410	0.888	**0.695**	0.246
Tukey	Partial Tests			Global Test		Tukey	Partial Tests			Global Test	
0.971	0.950	0.968	0.959	**0.917**	0.901	**0.971**	0.606	0.580	0.963	**0.917**	0.445
0.943	0.899	0.910	0.926	**0.838**	0.809	**0.943**	0.436	0.415	0.928	**0.838**	0.277
0.886	0.837	0.842	0.856	**0.695**	0.679	**0.886**	0.274	0.251	0.836	**0.695**	0.130

Table 1. Observed confidence levels of individual CI's and acceptance rates of H_{0A}^G with CSP and Tukey test. $\varepsilon \sim N(0,1)$, $n=4$.

$1-\alpha$	δ_{12}	δ_{13}	δ_{23}	$1-\alpha^G$		$1-\alpha$	δ_{12}	δ_{13}	δ_{23}	$1-\alpha^G$	
CSP	0	0	0	$\hat{P}[{}^A\tilde{T}=0]$		CSP	1	1	0	$\hat{P}[{}^A\tilde{T}=0]$	
0.996	0.996	0.993	1	**0.990**	0.989	**0.996**	0.819	0.844	0.996	**0.990**	0.721
0.983	0.987	0.977	0.976	**0.950**	0.950	**0.983**	0.615	0.610	0.983	**0.950**	0.454
0.966	0.955	0.95	0.958	**0.900**	0.892	**0.966**	0.481	0.519	0.968	**0.900**	0.332
Tukey	Partial Tests			Global Test		Tukey	Partial Tests			Global Test	
0.996	0.995	0.997	0.997	**0.990**	0.989	**0.996**	0.728	0.752	0.995	**0.990**	0.617
0.983	0.987	0.983	0.980	**0.950**	0.959	**0.983**	0.525	0.498	0.980	**0.950**	0.353
0.966	0.944	0.947	0.954	**0.900**	0.883	**0.966**	0.371	0.419	0.959	**0.900**	0.247

Table 2. Observed confidence levels of individual CI's and acceptance rates of H_{0A}^G with CSP and Tukey test. $\varepsilon \sim N(0,1)$, $n=6$.

$1-\alpha$	δ_{12}	δ_{13}	δ_{23}	$1-\alpha^G$		$1-\alpha$	δ_{12}	δ_{13}	δ_{23}	$1-\alpha^G$	
CSP	0	0	0	$\hat{P}[{}^A\tilde{T}=0]$		CSP	1	1	0	$\hat{P}[{}^A\tilde{T}=0]$	
0.971	0.966	0.970	0.972	**0.917**	0.921	**0.971**	0.871	0.873	0.978	**0.917**	0.777
0.943	0.949	0.949	0.928	**0.838**	0.858	**0.943**	0.793	0.793	0.936	**0.838**	0.646
0.886	0.892	0.892	0.884	**0.695**	0.749	**0.886**	0.682	0.705	0.890	**0.695**	0.526
Tukey	Partial Tests			Global Test		Tukey	Partial Tests			Global Test	
0.971	0.970	0.974	0.981	**0.917**	0.934	**0.971**	0.880	0.884	0.982	**0.917**	0.808
0.943	0.932	0.935	0.931	**0.838**	0.852	**0.943**	0.771	0.757	0.931	**0.838**	0.638
0.886	0.847	0.860	0.831	**0.695**	0.661	**0.886**	0.624	0.651	0.831	**0.695**	0.461

Table 3. Observed confidence levels of individual CI's and acceptance rates of H_{0A}^G with CSP and Tukey test. $\varepsilon \sim t_2$, $n = 4$.

$1-\alpha$	δ_{12}	δ_{13}	δ_{23}	$1-\alpha^G$		$1-\alpha$	δ_{12}	δ_{13}	δ_{23}	$1-\alpha^G$	
CSP	0	0	0	$\hat{P}[{}^A\tilde{T}=0]$		CSP	1	1	0	$\hat{P}[{}^A\tilde{T}=0]$	
0.996	0.996	0.996	0.998	**0.990**	0.991	**0.996**	0.930	0.932	0.995	**0.990**	0.876
0.983	0.979	0.988	0.98	**0.950**	0.954	**0.983**	0.850	0.835	0.988	**0.950**	0.738
0.966	0.963	0.964	0.959	**0.900**	0.903	**0.966**	0.780	0.779	0.967	**0.900**	0.652
Tukey	Partial Tests			Global Test		Tukey	Partial Tests			Global Test	
0.996	0.999	0.998	0.998	**0.990**	0.995	**0.996**	0.955	0.943	0.998	**0.990**	0.913
0.983	0.986	0.989	0.982	**0.950**	0.964	**0.983**	0.868	0.857	0.988	**0.950**	0.789
0.966	0.969	0.96	0.961	**0.900**	0.905	**0.966**	0.770	0.782	0.964	**0.900**	0.655

Table 4. Observed confidence levels of individual CI's and acceptance rates of H_{0A}^G with CSP and Tukey test. $\varepsilon \sim t_2$, $n = 6$.

The permutation pseudo - CI's and Tukey's CI's show similar performances in each scenario, although Tukey's test seems to be a little more powerful when normal errors are considered. Note how both procedures maintain the nominal confidence level $1-\alpha$ in the power comparison regarding ${}^AT_{23}$.

References

BASSO, D., CHIARANDINI, M., SALMASO, L. (2007): Synchronized permutation tests in I × J designs. *Journal of Statistical Planning and Inference, 137(8), 2564-2578.*

DEAN, A., VOSS, D. (1999). *Design and Analysis of Experiments.* Springer-Verlag, New York.

HSU, J.C. (1996). *Multiple Comparisons. Theory and Methods.* Chapman & Hall.

PESARIN, F. (2001): *Multivariate Permutation Tests with Applications in Biostatistics.* Wiley, Chichester.

SALMASO, L. (2003): Synchronized permutation tests in factorial designs. *Communication in Statistics - Theory and Methods, 32(7), 1419-1437.*

Multiple Comparison Procedures in Linear Models

Frank Bretz[1], Torsten Hothorn[2], and Peter Westfall[3]

[1] Statistical Methodology, Clinical Information Sciences, Novartis Pharma AG, 4002 Basel, Switzerland, *frank.bretz@novartis.com*
[2] Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstraße 33, D–80539 München, Germany, *Torsten.Hothorn@stat.uni-muenchen.de*
[3] Texas Tech University, Lubbock, TX 79409, U.S.A, *westfall@ba.ttu.edu*

Abstract. Multiplicity is a difficult and ubiquitous problem. The problem of evaluating multiple experimental questions occurs in many areas of applications, such as, for example, in clinical trials assessing more than one outcome variable, or in agricultural field experiments comparing several irrigation systems. If multiple null hypotheses are tested simultaneously, the probability of declaring effects when none exists increases beyond the nominal type I error level used for the individual comparisons. In this paper we review multiple comparison procedures in the linear model framework. We use the *multcomp* package from R to illustrate the methods with a linear regression example.

Keywords: multiplicity, multiple testing, multivariate t, *multcomp*, R

1 Introduction

Over the past decades, multiplicity has attracted more and more attention. Multiplicity, and thus the need for adequate multiple comparison procedures, arises in many areas of applications, such as, for example, in clinical trials assessing multiple outcome variables, or in agricultural field experiments comparing several irrigation systems. Potential sources of multiplicity include the comparison of several treatments or dose groups, multiple endpoints, multiple time points, interim analyses, multiple tests of the same hypothesis (for example, parametric and nonparametric), variable and model selection, and subgroup analysis. For a detailed theoretical treatise of multiple comparison procedures we refer to Hochberg and Tamhane (1987), Hsu (1996), and to Dmitrienko, Tamhane and Bretz (2009) in the context of clinical trials.

Assume for the purpose of illustration that two null hypotheses H_1 and H_2 are tested each at level $\alpha = 0.05$ using independent test statistics. If both H_1 and H_2 are true, the probability of incorrectly rejecting at least one of the two null hypothesis is $1-(1-\alpha)^2 = 0.0975$, which is substantially larger than the nominal level of 0.05. For increasing number of hypotheses the inflation in size becomes even larger. If, for example, 20 truly null hypotheses are tested, one incorrect rejection is to be expected.

The Bonferroni approach is a common multiple testing procedure, which compares the observed marginal p-values $p_1, \ldots, p_m$ with the common threshold α/m, where m is the number of hypotheses under investigation. Assuming that the m null hypotheses $H_1, \ldots, H_m$ are all true and that the p-values are identically distributed as uniform on $(0, 1)$, it follows from the Bonferroni inequality that the probability to reject at least one of the m null hypotheses is

$$P\left(\bigcup_{i=1}^{m}\left\{p_i \leq \frac{\alpha}{m}\right\}\right) \leq \sum_{i=1}^{m} P\left(p_i \leq \frac{\alpha}{m}\right) = \alpha.$$

The Bonferroni approach is thus a very general method, which is valid for any correlation structure among the test statistics. However, it is well known that the Bonferroni approach is conservative in the sense that other test procedures exist, which reject at least as many hypotheses as the Bonferroni approach.

In the following we describe a general approach to multiple comparison procedures in linear models, which includes additional assumptions on the joint distribution of the test statistics. In Section 2 we use a linear regression example to motivate the problem and to illustrate the use of the *multcomp* package in R, which implements the methods described in this paper. In Section 3 we give a general description of multiple comparison procedures in linear models. We conclude with some remarks in Section 8.

2 A linear regression example

2.1 Motivation

We consider the *cats* regression example from Venables and Ripley (1997) to motivate the subsequent discussion. The aim is to predict the heart weight Y from body weight X for 144 cats using the linear regression model

$$Y = \beta_1 + \beta_2 X + \varepsilon,$$

where β_1 denotes the intercept, β_2 denotes the slope, and ε denotes the Gaussian error term. A linear regression model can be fitted in R using the command

```
R> lm.cats <- lm(Hwt ~ Bwt, data = cats)
R> lm.cats
Call: lm(formula = Hwt ~ Bwt, data = cats)

Coefficients:
(Intercept)          Bwt
     -0.357        4.034
```

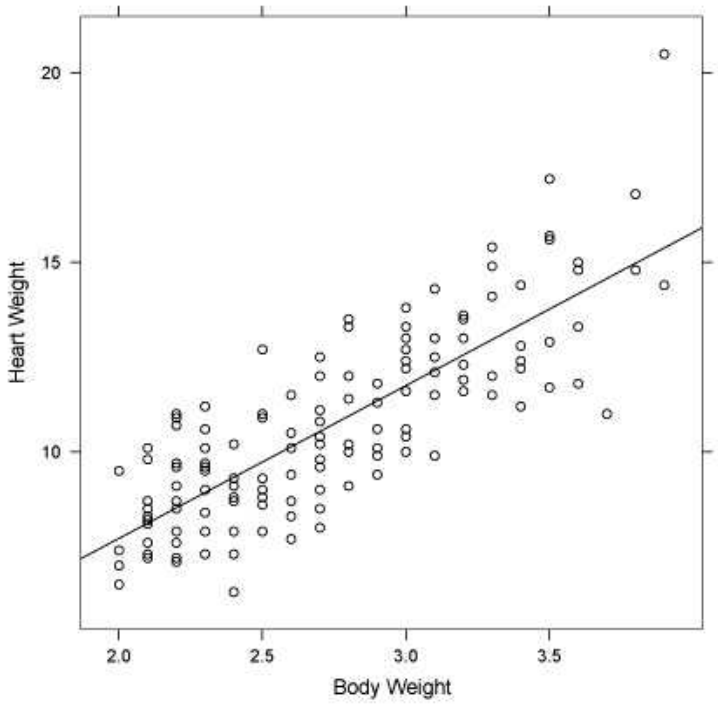

Fig. 1. A scatter plot of the *cats* data.

Figure 1 displays a scatter plot of the data including the regression line.

Assume that we are interested in testing whether the intercept or the slope equal zero. The two related null hypotheses

$$H_1 : \beta_1 = 0 \quad \text{and} \quad H_2 : \beta_2 = 0. \tag{1}$$

can be tested using correlated t tests. To simplify the notation, let

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2}$$

The estimates of the regression coefficients β and their covariance matrix can be extracted from the previously fitted model by calling

```
R> betahat <- coef(lm.cats)
R> Vbetahat <- vcov(lm.cats)
```

Given these numbers we can compute the vector $\mathbf{t}$ containing the two individual t test statistics and its associated correlation matrix as (see Section 3 for the theoretical background)

```
R> C <- diag(2)
R> Sigma <- diag(1/sqrt(diag(C %*% Vbetahat %*% t(C))))
R> t <- Sigma %*% C %*% betahat
R> Cor <- Sigma %*% (C %*% Vbetahat %*% t(C)) %*% t(Sigma)
```

Note that $\mathbf{t} = (-0.5152, 16.1194)^t$ with associated correlation matrix

```
        [,1]    [,2]
[1,]  1.0000 -0.9846
[2,] -0.9846  1.0000
```

Adjusted p-values are finally computed from the underlying bivariate t distribution using the *pmvt* function in the *mvtnorm* package (Hothorn et al., 2002; Genz and Bretz, 2002):

```
R> library("mvtnorm")
R> df.cats <- nrow(cats) - length(betahat)
R> q <- sapply(abs(t), function(x) 1 - pmvt(-rep(x,
+     2), rep(x, 2), corr = Cor, df = df.cats))
```

Note that by construction the adjusted p-values $q_i, i = 1, 2$, are corrected for multiplicity and can thus directly be compared with the pre-specified significance level α. In our example, $q_1 = 0.6562$ and $q_2 < 0.0001$, which indicates that the slope is significantly different from 0 but the intercept is not.

Alternatively, we can compute a critical value $u_{1-\alpha}$ derived from the bivariate t distribution and compare the test statistics $\mathbf{t} = (t_1, t_2)^t$ with it. Using the function

```
R> delta <- rep(0, 2)
R> myfct <- function(x, conf) {
+     lower <- rep(-x, 2)
+     upper <- rep(x, 2)
+     pmvt(lower, upper, df = df.cats, corr = Cor,
+         delta, abseps = 1e-04)[1] - conf
+ }
```

we can compute the critical value $u_{1-\alpha}$ with the *uniroot* function

```
R> u <- round(uniroot(myfct, lower = 1, upper = 5,
+     conf = 0.95)$root, 3)
R> u
[1] 2.043
```

In our example we set the confidence level as $1 - \alpha = 0.95$ and obtain $u_{1-\alpha} = 2.043$. Since $t_1 = -0.5152 < u_{1-\alpha}$ and $t_2 = 16.1194 > u_{1-\alpha}$, we thus obtain the same test decisions as before. In addition, the critical value

$u_{1-\alpha} = 2.043$ can be used to compute simultaneous confidence intervals for the parameters β_1 and β_2. Since the parameter estimates are so highly correlated, the critical value 2.043 from the bivariate t distribution is much smaller than the Bonferroni critical value $t_{0.9875,142} = 2.265$.

2.2 The *multcomp* package

As seen from the example in Section 2.1, implementing multiple comparisons involve a number of individual steps. The *multcomp* package provides a formal framework to replace the previous calculations by standardized function calls. In the following we apply it to the *cats* example to illustrate its use. The *glht* function from *multcomp* takes a fitted response model and a matrix $\mathbf{C}$ defining the hypotheses of interest to perform the multiple comparisons:

```
R> library("multcomp")
R> cats.ht <- glht(lm.cats, linfct = C)
R> summary(cats.ht)
     Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = Hwt ~ Bwt, data = cats)

Linear Hypotheses:
                 Estimate Std. Error t value p value
(Intercept) == 0   -0.357      0.692   -0.52    0.66
Bwt == 0            4.034      0.250   16.12  <1e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported)
```

For each parameter $\beta_i, i = 1, 2$, *multcomp* reports its estimate and standard error. Taking the ratio of these two values for each parameter results in the value of the test statistic. The adjusted p-values are reported in the last column. Note that they are the same as calculated Section 2.1.

In addition, simultaneous confidence intervals can be calculated for each parameter using the the *confint* function:

```
R> confint(cats.ht)
Simultaneous Confidence Intervals for General Linear Hypotheses

Fit: lm(formula = Hwt ~ Bwt, data = cats)

Estimated Quantile = 2.043

Linear Hypotheses:
                 Estimate lwr    upr
```

```
(Intercept) == 0 -0.357   -1.771  1.058
Bwt == 0          4.034    3.523  4.545

95% family-wise confidence level
```

The two-sided confidence interval for the intercept includes the 0, thus reflecting the previous test decision that we can not conclude the intercept to be statistically significant. We further conclude that the slope parameter lies roughly between 3.5 and 4.5.

3 Multiple comparisons in linear models

We now introduce a unified framework for multiple hypothesis testing in general linear models. We use the example from Section 2 to illustrate some of subsequent expressions. For details on linear model theory we refer to standard textbooks, such as Searle (1971), for example.

We consider the common general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3}$$

for a $n \times 1$ vector of observations $\mathbf{Y} = (Y_1, \ldots, Y_n)^t$. In model (3), $\mathbf{X}$ denotes $n \times p$ design matrix with fixed and known entries (x_{ij}) and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ denotes the fixed and unknown parameter vector. The $n \times 1$ error vector $\boldsymbol{\varepsilon}$ is assumed to follow a n-dimensional normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)^t$ and covariance matrix $\sigma^2\mathbf{I}_n$, where σ^2 denotes the residual error and $\mathbf{I}_n$ denotes the identity matrix of dimension n. Model (3) implies that each individual observation y_i follows the linear model

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Assume that we are interested in performing pre-specified comparisons among the parameters $\beta_1, \ldots, \beta_p$. To this end, define a $p \times 1$ vector $\mathbf{c} = (c_1, \ldots, c_p)^t$ of known constants. The vector $\mathbf{c}$ thus reflects a single experimental comparison of interest by considering the linear combination $\mathbf{c}^t\boldsymbol{\beta}$ with associated null hypothesis

$$H : \mathbf{c}^t\boldsymbol{\beta} = a \tag{4}$$

for a fixed and known constant a. In the following, we refer to $\mathbf{c}^t\boldsymbol{\beta}$ as the (linear) function of interest. If we have multiple experimental questions, m say, we obtain m vectors $\mathbf{c}_1, \ldots, \mathbf{c}_m$, which can be summarized by the matrix $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_m)$.

In the linear regression example from Section 2, we have measurements of heart and body weight for $n = 144$ cats which are assumed to follow the

linear model (3), where

$$\mathbf{Y} = \begin{pmatrix} 7.0 \\ 7.4 \\ 9.5 \\ \vdots \\ 16.8 \\ 14.4 \\ 20.5 \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1\ 2.0 \\ 1\ 2.0 \\ 1\ 2.0 \\ \vdots\ \vdots \\ 1\ 3.8 \\ 1\ 3.9 \\ 1\ 3.9 \end{pmatrix}.$$

The $m = 2$ hypotheses are given in equation (1) with $\mathbf{a} = (a_1, a_2)^t = \mathbf{0}$, which result in the matrix $\mathbf{C}$ specified in equation (2).

Standard linear model theory ensures that the usual least square estimates

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{Y} \tag{5}$$

and

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\nu} \tag{6}$$

are unbiased estimates of $\boldsymbol{\beta}$ and σ, respectively, where $\nu = n - \text{rank}(\mathbf{X})$ and $(\mathbf{X}^t\mathbf{X})^-$ denotes some generalized inverse of $\mathbf{X}^t\mathbf{X}$. We are interested in the pivotal quantities

$$t_j = \frac{\mathbf{c}_j^t\hat{\boldsymbol{\beta}} - a_j}{\hat{\sigma}\sqrt{\mathbf{c}_j^t(\mathbf{X}^t\mathbf{X})^-\mathbf{c}_j}}, \quad j = 1, \ldots, m, \tag{7}$$

one for each experimental question defined through $\mathbf{c}_j$. By construction, each test statistic $t_j, j = 1, \ldots, m$, follows under the null hypothesis (4) a central univariate t distribution with ν degrees of freedom. The joint distribution of $t_1, \ldots, t_m$ is multivariate t with ν degrees of freedom and correlation matrix

$$\mathbf{R} = \mathbf{D}\mathbf{C}^t(\mathbf{X}^t\mathbf{X})^-\mathbf{C}\mathbf{D},$$

where $\mathbf{D} = \text{diag}(\mathbf{c}_i^t(\mathbf{X}^t\mathbf{X})^-\mathbf{c}_i)^{-1/2}$. In the asymptotic case $\nu \to \infty$ or if σ is known, the corresponding limiting multivariate normal distribution holds. Confidence intervals for $\mathbf{c}_j^t\boldsymbol{\beta} - a_j$ with simultaneous coverage probability $1-\alpha$ are given by

$$\left[\mathbf{c}_j^t\hat{\boldsymbol{\beta}} - a_j - u_{1-\alpha}\hat{\sigma}\sqrt{\mathbf{c}_j^t(\mathbf{X}^t\mathbf{X})^-\mathbf{c}_j}; \mathbf{c}_j^t\hat{\boldsymbol{\beta}} - a_j + u_{1-\alpha}\hat{\sigma}\sqrt{\mathbf{c}_j^t(\mathbf{X}^t\mathbf{X})^-\mathbf{c}_j}\right],$$

$j = 1, \ldots, m$, where $u_{1-\alpha}$ denotes the critical value derived from the multivariate normal or t distribution. Numerical integration methods to calculate the multivariate normal and t probabilities required for the computation of adjusted p-values $q_i, i = 1, \ldots, m$, and critical values $u_{1-\alpha}$ are described by Genz and Bretz (2002). For a general overview about these distributions we

refer to the books of Kotz and Nadarajah (2004), Kotz, Balakrishnan and Johnson (2000) and Tong (1990).

In the *cats* example, we have $\hat{\boldsymbol{\beta}} = (-0.3567, 4.0341)^t$ and $\hat{\sigma} = 1.4524$. In Section 2 we extracted this information from the fitted linear model. Alternatively, one can compute these numbers using equations (5) and (6). One also checks that plugging these estimates into equation (7) gives the test statistics $t_1 = -0.5152$ and $t_2 = 16.1194$ with $\nu = 142$ degrees of freedom. Recalling the correlation -0.9846 between the two test statistics, one can then compute the required bivariate t probabilities for the multiplicity adjusted p-values, as shown in Section 2.1.

4 Conclusions

We have reviewed the general theory of multiple comparison procedures in the context of linear models. The framework outlined in Section 3 allows the inclusion of covariates and/or factorial treatment structures in classical regression and ANOVA applications. Many well-known multiple comparison procedures fit into this framework, such as the Tukey test for all-pairwise comparisons and the Dunnett test for many-to-one comparisons, see Hochberg and Tamhane (1987) and Hsu (1996) for details.

Hothorn et al. (2008) extended the canonical description from Section 3 to more general parametric and semi-parametric models, which allows a unified treatise of multiple comparisons for generalized linear models, mixed models, survival models, etc. The underlying methods are all implemented in the *multcomp* package, which in turn relies on the multivariate normal or t probabilities returned by *mvtnorm*. We refer to Hothorn et al. (2008) for further examples on its use.

As a matter of fact, the methods discussed in this paper can be used to construct more powerful closed test procedures, as first discussed by Westfall (1997). That is, for a given family of null hypotheses $H_1, \ldots, H_n$, an individual hypothesis H_i is rejected only if all intersection hypotheses $H_J = \bigcap_{j \in J} H_j$ with $i \in J \subseteq \{1, \ldots, n\}$ are rejected (Marcus et al., 1976). The *multcomp* package uses $\max t$ type statistics for each intersection hypothesis based on the methods from this paper, thus accounting for stochastic dependencies. Furthermore, the implementation of *multcomp* exploits logical constraints, leading to computationally efficient, yet powerful truncated closed test procedures, see Westfall and Tobias (2007).

References

DMITRIENKO, A., TAMHANE, A.C. and BRETZ, F. (2009): *Multiple Testing in Pharmaceutical Statistics*. Taylor and Francis, Boca Raton (in press)

GENZ, A. and BRETZ, F. (2002): Methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics, 11, 950-971.*

HOCHBERG, Y. and TAMHANE, A.C. (1987): *Multiple Comparison Procedures*. Wiley, New York.

HOTHORN, T., BRETZ, F. and GENZ, A. (2001): On multivariate t and Gauss probabilities in R. *R Newsletter, 1(2), 27-29*.

HOTHORN, T., BRETZ, F. and WESTFALL, P.H. (2008): Simultaneous inference in parametric models. *Biometrical Journal* (in press).

HSU, J.C. (1996): *Multiple Comparisons*. Chapman and Hall, London.

KOTZ, S., BALAKRISHNAN, N., and JOHNSON, N.L. (2000): *Continuous Multivariate Distributions*. Wiley, New York.

KOTZ, S. and NADARAJAH, S. (2004): *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge.

MARCUS, R., PERITZ, E. and GABRIEL, K.B. (1976): On closed testing procedures with special reference to ordered analysis of variance. *Biometrika, 63, 655-660*.

SEARLE, S. (1971): *Linear Models*. Wiley, New York

TONG, Y.L. (1990): *The Multivariate Normal Distribution*. Springer, New York.

VENABLES, W.N. and RIPLEY, B.D. (1997): *Modern Applied Statistics With S*. Springer, New York.

WESTFALL, P.H. (1997): Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association, 92, 299-306.*

WESTFALL, P.H. and TOBIAS, R.(2007): Multiple testing of general contrasts: Closure and the extended Shaffer-Royen methods. *Journal of the American Statistical Association, 102, 487-494.*

Inference for the Top-k Rank List Problem

Peter Hall[1] and Michael G. Schimek[2]

[1] Department of Mathematics and Statistics, The University of Melbourne
Parkville, VIC, 3010, Australia, *halpstat@ms.unimelb.edu.au*
[2] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz
Auenbruggerpl. 2, 8036 Graz, Austria, *michael.schimek@meduni-graz.at*

Abstract. Consider a problem where N items (objects or individuals) are judged by assessors using their perceptions of a set of performance criteria, or alternatively by technical devices. In particular, two assessors might rank the items between 1 and N on the basis of relative performance, independently of each other. We aggregate the rank lists in that we assign *one* if the two assessors agree, and *zero* otherwise. How far can we continue into this sequence of 0's and 1's before randomness takes over? In this paper we suggest methods and algorithms for addressing this problem.

Keywords: ordered list, moderate deviation bound, nonparametric inference, rank aggregation, random degeneration, top-k rank list

1 Introduction

In various fields of application we are confronted with lists of distinct objects in rank order. The ordering might be due to a measure of strength of evidence or to an assessment based on expert knowledge or a technical device. The ranking might also represent some measurement taken on the objects which might not be comparable across the lists, for instance, because of different assessment technologies or levels of measurement error.

In this paper our interest is to consolidate such lists of common objects, under the assumption of a general decrease of the probability for consensus of rankings with increasing distance from the top rank position. This assumption is reasonable for the rank aggregation applications we have in mind, and is equivalent to the notion of random degeneration of paired rank information (when our input consists of two lists). Applications include the combined analysis of gene expression measurements across experiments and array platforms, data integration of results from internet search engines, or the determination of consensus rankings in customer surveys. The first two examples involve an additional aspect we wish to consider, namely lists of extreme length, say, ten thousand and more (usually resulting from high-dimensional analysis problems). The longer the lists, the more likely we are to observe non-overlapping ranks.

Rank order problems are not new. They have been intensively studied in psychometrics in the nineteen twenties (e.g. Thurstone, 1931), and later on

in biometric experimental design problems (e.g. Bradley and Terry, 1952). A more recent account is given in the book by Kendall and Gibbson (1990), however, the classical statistical methodology cannot cope with very long lists.

The last few years have seen an increasing research interest in rank aggregation, focusing on computational approaches that allow calculation of a consolidated list that satisfies a suitable minimum total distance criterion with respect to two or more input lists. The goal of combining information across multiple large data sources, studies or experiments is highly challenging indeed. The naive approach would be combinatorial, however, even for moderately sized data sets it would be NP hard (see e.g. Fagin et al., 2003). Current attempts to overcome this difficulty follow two different strategies. Dwork et al. (2001), and Fagin, Kumar and Sivakumar (2003), developed Markov chain meta-search algorithms for the internet, summarizing majority preferences between pairs of objects across lists. DeConde et al. (2006) applied the Markov process framework to microarray findings obtained across different array platforms. Lin and Ding (2008) derived a cross-entropy Monte Carlo method for the integration of rank lists in genomic studies. What these new algorithms have in common is the fact that they are still computationally highly demanding. Our experience for the simple two-list integration case, with the cross-entropy Monte Carlo method, is that the number of rankings needs to be limited to about two hundred. Because of that, partial instead of full lists are analyzed all the time, and the list length k of a so-called *top-k list* is chosen arbitrarily. Such an ad-hoc approach is dissatisfying. This motivates us to propose a *moderate deviation*-based inference concept for identifying the k as the rank position where the consensus information of the two lists, representing the same objects, degenerates into noise. To cope with the above-mentioned demands, we need to provide a mathematical concept, as well as an algorithmic solution, that can cope with very long lists of the order of tens of thousands. In the problems that motivate this work, the total number of objects is much larger than the number of comparable rankings before noise prevails. In particular, assessors or technical devices often agree in the case of many of the first 100 or so objects, but can give extremely noisy rankings to the remaining objects, out of perhaps 10,000.

In Section 2 a sequence that represents the paired rank information of two lists is introduced. Then in Section 2, for the problem outlined in Section 1, a simplified mathematical model is proposed, and a suitable algorithm is derived. In Section 3 we study the numerical properties of the proposed algorithm via simulations, and give recommendations for the choice of technical and tuning parameters. The theoretical properties of our methodology are derived in Section 5, summarized in three theorems [1]. Finally we illustrate our top-k list inference procedure on real data from a study in molecular medicine.

[1] A journal paper also providing the proofs is in preparation

2 Definition of the problem

Assume data are observed as a sequence of 0's and 1's, say $I_1, \ldots, I_N$. Early entries in the sequence are predominantly 1's, but the sequence eventually degenerates into noise, at which point the 0's and 1's arise randomly and in approximately equal proportions. Our main interest in the sequence is this non-stationarity; we wish to define what is meant by, and to estimate, *the point j_0 at which the sequence degenerates into noise*, i.e. identify the proper length of a partial list, usually denoted a *top-k list*.

Taking a mathematical view of the problem for a moment, and defining $p_j = P(I_j = 1)$, we assume that there exists $j_0 \geq 2$ with the property that $p_j \geq \frac{1}{2}$ for $1 \leq j \leq j_0 - 1$, $p_{j_0-1} > \frac{1}{2}$, and $p_j = \frac{1}{2}$, for $j \geq j_0$. We seek to estimate j_0. We might interpret j_0 as the point at which an estimable signal in the sequence $I_1, \ldots, I_N$ ceases, and noise takes over.

Let us assume that there are two "assessors" that both rank, independently of each other, N distinct objects $o_1, \ldots, o_N$ according to the extent to which a particular attribute is present. The ranking is from 1 to N, without ties. We might take $I_j = 1$ if the ranking given by the second assessor to the object ranked j by the first assessor, is not distant more than d, say, from j, and $I_j = 0$ otherwise. We could symmetrise this definition with respect to the assessors, for example by asking that both, or at least one, of the two distances not exceed 1. If we take $d = 0$ then symmetry already prevails, but then we may need to adjust for "irregular" rankings; see Subsection 3.3.

Of course, the value of d represents a tuning parameter. In numerical practice it would generally be chosen by a mixture of experimentation with the real data, and simulation from a model which gave results reflecting those data.

3 Model and methodology

3.1 A simplified mathematical model

In the discussion below, and in theoretical work in Section 5, we shall assume that the Bernoulli random variables I_j are independent. Clearly, this is not exactly correct, but it is a reasonable working approximation. Given that our interest is in assessing distinctly nonstationary features of the process I_j; because it is all but simple to model dependence in a nonstationary time-series of 0's and 1's; and since, in the real-data examples that motivate our work, it is awkward to identify dependencies between assessors; it seems difficult to replace independence by a practical alternative assumption. Our independence assumption is justified in practice because $k \ll N$. Even in the top-k list many I_j's are expected to have negligible dependence because of irregular rankings.

As a consequence our simplified mathematical model is based on the following two assumptions:

1. *Independent Bernoulli random variables* $I_1, \ldots, I_N$ *are observed, with* $p_j \geq \frac{1}{2}$ *for each* $j \leq j_0 - 2$, $p_{j_0-1} > \frac{1}{2}$, *and* $p_j = \frac{1}{2}$ *for* $j \geq j_0$. *From this information we wish to estimate the value of* j_0, *the point of degeneration into noise.*
2. *The "general decrease" of* p_j *for increasing* j, *implied by this condition, need not be monotone.*

3.2 The algorithm for estimating j_0

The algorithm consists of an ordered sequence of "test stages," $s_1, s_2, \ldots$. Stage s_k terminates a distance J_{s_k} into the sequence $I_1, \ldots, I_N$. When k is odd, J_{s_k} is a potential lower bound to j_0. We use the word *potential* since random fluctuations can lead to errors, in which case the assertion that $J_{s_{2k-1}}$ is a lower bound may not be strictly correct. However, one can show that when $k = 1$, the probability that $J_{s_{2k-1}}$ is a lower bound for j_0 is approximately equal to 1 under our mathematical model (an analogous result could be derived for each $k \geq 1$).

Stage s_k starts by drawing a pilot sample of size ν, consisting of the set of values I_j for which j is among the first ν indices to the right of $J_{s_{k-1}} - r\nu$, if k is odd, or to the left of $J_{s_{k-1}} + r\nu$, if k is even. (We could use a different pilot sample size for each k, but considerations of simplicity lead us not to.) Here, $r > 1$ is fixed; if r is not an integer then we interpret $r\nu$ as the integer part of that quantity. We include $J_{s_{k-1}} \pm r\nu$ in the set of ν indices, although it could be excluded. The sequence of consecutive steps that leads from $J_{s_k} \pm r\nu$ to J_{s_k} is called the *test stream* for stage s_k.

More generally, we use pilot samples of size ν to construct

$$\hat{p}_j^+ = \frac{1}{\nu} \sum_{\ell=j}^{j+\nu-1} I_\ell \quad \text{and} \quad \hat{p}_j^- = \frac{1}{\nu} \sum_{\ell=j-\nu+1}^{j} I_\ell \,. \tag{1}$$

These quantities represent estimates of p_j computed from the ν data pairs I_ℓ for which ℓ lies immediately to the right of j, or immediately to the left, respectively. Pilot-sample size plays the role of a smoothing parameter in our algorithm; its practical choice will be discussed in Section 3.

Choose the constant $C > 0$ so that

$$z_\nu \equiv \bigl(C\,\nu^{-1}\,\log\nu\bigr)^{1/2} \tag{2}$$

is a moderate-deviation bound for testing the null hypothesis H_0 that $p_k = \frac{1}{2}$ for ν consecutive values of k, versus the alternative H_1 that $p_k > \frac{1}{2}$ for at least one of the values of k. In particular, assuming that H_0 applies to the ν consecutive values of k in the respective series at (1), we reject H_0 if and only if either $\hat{p}_j^+ - \frac{1}{2} > z_\nu$ or $\hat{p}_j^- - \frac{1}{2} > z_\nu$.

Under H_0, the variance of $\hat{p}_j^\pm$ equals $(4\nu)^{-1}$. Therefore we should take $C > \frac{1}{4}$ if we are to control moderate deviations; see Section 3 for aspects

of selection. These considerations are similar to the moderate-deviation arguments implicit in work of Donoho and Johnstone (1994), and Donoho et al. (1995), when constructing thresholds for wavelet expansions. See Amosova (1972), and Rubin and Sethuraman (1965), for discussion of bounds for moderate deviations.

Stage s_1 of the algorithm starts with $J_{s_0} = 1$, and takes the jth pilot sample equal to the set of values I_k, $j \le k \le j + \nu - 1$. From these respective samples we compute the estimators $\hat{p}_1^+, \hat{p}_2^+, \ldots$, defined at (1), and we conduct successive tests by asking whether it is true that $\hat{p}_j^+ - \frac{1}{2} > z_\nu$. We stop at the first integer $j = J_{s_1}$ for which this inequality fails, i.e. for which the null hypothesis H_0 is not rejected. The test stream corresponding to the first stage of the algorithm is the increasing sequence of integers $1, \ldots, J_{s_1}$.

Thus, J_{s_1} is our first potential lower bound to the value of j_0. Stage s_1, and any stage s_k for odd k, consists of a sequence of tests conducted by moving to the right; stage s_2, and any stage s_k for even k, consists of a sequence of tests conducted by moving to the left.

The second stage starts at $j = J_{s_1} + r\nu$. More generally, any even-numbered stage starts $r\nu$ to the right of the point where the previous odd-numbered stage ended, and any odd-numbered stage starts $r\nu$ units to the left of where the previous even-numbered stage concluded. (The algorithm terminates if the latter point would be less than 1.)

In stage s_2 we conduct successive tests by asking whether it is true that $\hat{p}_j^- - \frac{1}{2} > z_\nu$ for $j = J_{s_1} + r\nu, J_{s_1} + r\nu - 1, \ldots$. We stop at the first integer $j = J_{s_2}$ for which the inequality holds. By construction, $J_{s_2} \ge J_{s_1} + \nu - 2$. The decreasing sequence of integers $J_{s_1} + r\nu, \ldots, J_{s_2}$ is the test stream corresponding to stage s_2 of the algorithm.

More generally, in odd-numbered stages we move to the right in one-unit steps, and stop as soon as the inequality $\hat{p}_j^+ - \frac{1}{2} > z_\nu$ fails; in even-numbered stages we move to the left and stop as soon as $\hat{p}_j^- - \frac{1}{2} > z_\nu$. This see-sawing action helps to guard against the algorithm stopping too soon because of a chance cluster of "noisy" values of I_j. By jumping $r\nu$ units in the direction we had been moving during the previous test stream, and then working in the opposite direction, we make the algorithm more robust against such clusters.

Our rules for constructing the estimator $\hat{j}_0$ of j_0, and for terminating the algorithm, are as follows:

The algorithm is terminated and the results of stage s_{2k-1} are stored if one or more of the following three cases arise:

(i) The algorithm enters a loop where the two adjacent stages, s_{2k-1} and s_{2k}, are repeated ad infinitum;

(ii) for some k, $J_{s_{2k+1}} \le J_{s_{2k-1}}$;

(iii) $J_{s_{2k}} - 2\nu \le 1$.

In each of these cases our estimator of j_0 is $\hat{j}_0 = J_{s_{2k-1}} + \frac{1}{2}\nu$. If the algorithm has not terminated by stage s_{2k-1} then $J_{s_{2k+1}} > J_{s_{2k-1}}$ (see Theorem 1 in Section 5). Therefore, if the algorithm does not terminate then it ultimately

reaches the right-hand end of the sequence $1, \ldots, N$ *of integers, and in this case we define* $\hat{j}_0 = N$.

The suggestion that $\hat{j}_0 = J_{s_{2k-1}} + \frac{1}{2}\nu$ is based on the fact that there is moderately strong evidence that j_0 lies towards the middle of the set of ν indices immediately to the right of $J_{s_{2k-1}}$.

3.3 Adjustments to allow for irregular rankings

There are situations in which our idealised mathematical model can be upset by, among other things, one assessor's inability to rank an object which the other assessor can rank, and in fact ranks quite highly; or by the two assessors giving quite different rankings to an object, with one of these rankings being so high that it alters other high rankings that would otherwise be among those on which the assessors agreed. Let us give an extreme example of the last-mentioned difficulty. One assessor might give rank 1 to object o_1, which the other assessor ranks as N, and the assessors might concur on the relative rankings of each of the objects $o_2, \ldots, o_{N-1}$. In this case, and despite the latter concurrence, all rankings of all objects differ.

We may deal with these problems by allowing the experimenter to discard up to a predetermined number, m say, of ranked objects at each test stage, prior to carrying out the test. Other adjustments could also be made, for example by allowing interchanges of a given number of rankings, or by adjoining new "pseudo objects" with rankings determined by the experimenter. However, the number of possibilities that need to be checked in these cases is so large that the associated computational demand leads to difficulties. Our decision to remove the first m objects at each stage, rather than only at the start of the algorithm, is also made on grounds of simplicity. If the objects were removed only at the start then we would typically have to go back and adjust the choice of omitted objects during the sth stage, for each s.

The procedure in which a predetermined number, m say, of objects is omitted from the rankings can be incorporated into the algorithm as follows. Suppose we have just reached test stage s. We construct the subsequent test stream, including the pilot sample, as described in Subsection 3.2; and we also construct the analogue of that stream which is obtained by removing any ℓ, where $1 \le \ell \le m$, of the objects among $1, \ldots, N$. The point J_{s_k} at which we declare the stream to have concluded is taken to be the minimum, if k is odd, or the maximum, if k is even, of the conclusion points over all choices of k, for $1 \le k \le m$.

4 Numerical properties

Because our mathematical model in Subsection 3.1 cannot be more than an approximation to the complex decision problem we wish to address, an iterative algorithm, adjustable for irregularity, was developed. Its implementation

in the statistical and graphical computing environment R is highly efficient (the numerical operations are approximately linear in N).

To execute the algorithm several parameters need to be set. Some of them are technical parameters, one is a tuning (smoothing) parameter. The same is true for all the rank aggregation algorithms mentioned in Section 1, independently of what kind of methodology they use. In that sense there is no way to "automatically" consolidate lists of ranked objects for the reason of irregular rankings, and because of unknown probabilities p_j in the top-k rank list. In the aforementioned rank aggregation algorithms default values are used (usually based on ad hoc assumptions about the data).

To study the numerical properties of the algorithm we performed a simulation experiment and analyzed its sensitivity with respect to the estimate $\hat{j}_0$ for various parameter settings. Because we have assumed that the data (a top-k list plus the remainder part of the list where noise has taken over) are observed as a sequence of $I_1, \ldots, I_{j_0-1}, I_{j_0}, \ldots, I_N$, where $k = j_0 - 1$, we could generate the I_j's of the two list segments, separated by the index j_0, as follows: Construct all non-stationary sequences as combinations of two independent Bernoulli random variables, with the probabilities $p(\text{seg1}) \in [0.6, 0.7, 0.8, 0.9, 1]$ for segment 1, and $p(\text{seg2}) \in [0.1, 0.2, 0.3, 0.4, 0.5]$ for segment 2. For the top-k list it is assumed that $P(I_j = 1)$ is larger than 0.5, hence there is perfect or at least some consensus of rankings. For the remaining part of the list it is supposed that $P(I_j = 1)$ is 0.5 or smaller, which means no or low consensus. The R function `rbinom()` was used to construct these sequences. The length N of the full list was always 1,000. The range of indices j_0 we wished to estimate via our inference procedure was taken to be $j \in [10, 20, 30, 40, 50, 100, 150, 200, 250, \ldots, 500]$.

Let us next specify the range of technical parameters, i.e. r and C, and tuning parameters, i.e. ν, where the latter represents the pilot sample size. We require $r > 1$, a parameter necessary to define the test stream for stage s_k. It is connected to the pilot sample size ν, and was considered for $r \in [1.1, 1.2, 1.3, 1.4, 1.5, 1.6, \ldots, 2.0]$. The specification of the moderate deviation bound in equation (2) requires us to choose the constant C. We know from Subsection 3.2 that under H_0, the variance of $\hat{p}_j^{\pm}$ equals $(4\nu)^{-1}$. Therefore we should take $C > \frac{1}{4}$ if we are to control moderate deviations. Actually we studied the algorithm for $C \in [0.251, 0.3, 0.35, 0.4, 0.45, 0.50, \ldots, 1]$.

The pilot sample size ν plays the role of a smoothing parameter and is therefore critical with respect to the estimation of $\hat{j}_0$. We considered values $\nu \in [10, 20, 30, 40, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500]$ in our simulations.

Finally we comment on the parameter m. It allows the experimenter to discard a predetermined number of ranked objects at each test stage s_k. Its choice demands some prior knowledge of the original data that are subject to ranking, but it does not matter in the simulations as we constructed the data stream without that sort of irregularity.

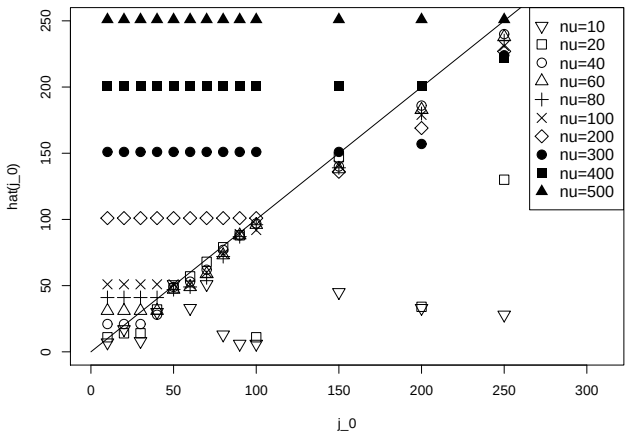

Fig. 1. Plot of estimates $\hat{j}_0$ against the true j_0 for selected choices of parameter ν.

To illustrate the results of our simulation study we display two typical plots. In Fig. 2 (assuming $C = 0.4$ and $r = 1.2$) for an input sequence characterised by $p(\text{sec1}) = 0.8$ and $p(\text{sec2}) = 0.2$, we compare the real j_0 with the estimates $\hat{j}_0$ for various values of ν. It can be seen clearly that longer top-k rank lists require a larger pilot sample size ν. In general, we observed the following approximate relationship: ν should be larger than j_0, not exceeding $2j_0$, when C is chosen appropriately. (Our theoretical work in section 5 will address the case where j_0/N is relatively small, in which case the ranking degenerates into noise relatively early. There, somewhat smaller values of ν/j_0 are generally appropriate. However, as j_0/N increases, a different theoretical argument can be used to show that improved performance occurs for larger values of ν/j_0.) The role of the constant C is illustrated by Fig. 3 ($p(\text{sec1}) = 0.8$ and $p(\text{sec2}) = 0.3$). There the point of degeneration into noise is fixed at 40. We study pilot sample sizes in the proximity of j_0, ranging between 10 and 100. The best estimation of j_0 is obtained for $C < 0.6$ as long as we do not grossly over- or undersmooth (i.e. avoid selecting $\nu = 10$ or $\nu = 100$).

Our main simulation results can be summarized in the following way. The technical constant r can be fixed at, say $r = 1.2$, without substantial influence on the estimate $\hat{j}_0$ in most scenarios. Indeed, our method seems to be robust against choice of r, provided we take care, as indicated above, when selecting ν. (For example, if we were to choose ν too small it would be advisable to attempt to correct for this by choosing r relatively large.)

We observe that the constant C has a quite different influence. The effect of its choice on the estimate is highly dependent on the combination of the probabilities $p(\text{seg1})$ and $p(\text{seg2})$. When $p(\text{seg1})$ is numerically close to $p(\text{seg2})$, i.e. the two segments are poorly separated, the value of C is important. Choices that allow us to compensate for poor segment separability are in the range of $0.25 < C < 0.6$. Larger values of C should not be con-

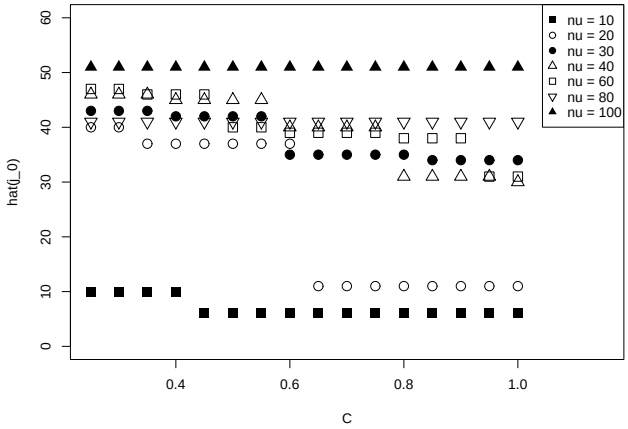

Fig. 2. Plot of estimates $\hat{j}_0$ in dependence of constant C and tuning parameter ν.

sidered (obviously the worst choice is $C = 1$). The pilot sample size ν is our smoothing parameter and seems to be empirically best approximated by values ranging from j_0 to $2j_0$ (of course, in reality we do not know j_0). This means in practice that a set of values related to the assumed index k of the top list (at least its order of magnitude should be known) should be considered. The choice of ν is certainly more critical than the selection of C.

5 Theoretical properties

Our first theorem justifies the claims made by rules about long-run properties of the algorithm, and in particular about termination.

Theorem 1. *For each integer k, $J_{s_{2k-1}} + \nu - 1 \le J_{s_{2k}}$. The values of $J_{s_{2k-1}}$, for $k \ge 1$, form a strictly monotone increasing sequence, until the algorithm terminates. That is, for each $k \ge 1$, if the algorithm has not terminated by the end of stage s_{2k-1} then $J_{s_{2k+1}} > J_{s_{2k-1}}$.*

Next we take an asymptotic view of properties of the estimator $\hat{j}_0$. We suppose that $j_0 = j_0(n)$ is on the scale of n, in particular that $C_1\, n \le j_0 \le C_2\, n$ where $0 < C_1 < C_2 < \infty$ are constants, and n is an integer that we shall permit to diverge to infinity. Furthermore, it is assumed that N is much larger than n, i.e. $n/N \to 0$ as $n \to \infty$. These assumptions reflect the sizes of j_0 and N in the simulated examples in Section 3. We take the pilot sample size, ν, to diverge with n; it will increase at a slower rate than n itself. Given $B > C^{1/2}$, put $\bar{p}_j = \sum_{j \le k \le j+\nu-1} p_k$ and

$$j_1 = j_1(B, \nu) = \sup\Big\{ j \in [1, j_0] : \bar{p}_j - \frac{1}{2} \ge B\big(\nu^{-1} \log \nu\big)^{1/2} \Big\}, \qquad (3)$$

with $j_1 = j_0$ if this set is empty.

Theorem 2. *Assume that the data $I_1, \ldots, I_N$ are generated by our mathematical model; that C, in the definition of z_ν at (2), satisfies $C > \frac{1}{2}$; that $\nu = \nu(n) = o(n)$; and that, for constants $0 < B_1 < 1$ and $B_2 > 0$, $\nu \geq B_2\, n^{B_1}$. Then, if B is so large that $2\, B_1\, (B - C^{1/2})^2 > 1$, and with $j_1 = j_1(B, \nu)$, we have $|\hat{j}_0 - j_0| = O_p\{\min(j_0 - j_1, \nu)\}$.*

To illustrate the implications of Theorem 2, let us suppose that p_j decreases to $\frac{1}{2}$ in a way that is bounded below by a linear function:

$$p_j \geq \min\left\{\tfrac{1}{2} + \alpha_n\,(j_0 - j), 1\right\} \text{ for } 1 \leq j \leq j_0\,, \text{ and } p_j = \tfrac{1}{2} \text{ for } j > j_0\,, \quad (4)$$

where $\alpha_n > 0$ denotes a small quantity depending on n and decreasing to zero. As α_n decreases more rapidly, the problem of estimating j_0 becomes, in general, more difficult. In particular, taking $\nu = \lfloor n^\beta \rfloor$ (the integer part of n^β) where $0 < \beta < 1$ is fixed, we see from Theorem 2 that

$$\alpha_n = n^{-3\beta/2}\,(\log n)^{1/2} \quad (5)$$

represents an order of magnitude of α_n for which approximation of j_0 by $\hat{j}_0$ at rate n^β is possible. As β decreases, the rate of approximation to j_0 by $\hat{j}_0$ improves, but to achieve the faster rate the difficulty of the problem must decrease; that is, the order of magnitude of α_n must increase.

It can be shown that, in the context of sequences $\{p_j\}$ satisfying (4), the order of magnitude of α_n at (5) is minimax-optimal, except for the logarithmic factor there. If it were known that p_j admitted the concise "no faster than linear" rate of decrease at (4), then an improved estimator of j_0 could be constructed, and the logarithmic factor could be removed from (5). Nevertheless, $\hat{j}_0$ is minimax-optimal in a different sense. We treat this problem next.

First we describe a class $\mathcal{P}$ of probability sequences $p = (p_1, \ldots, p_N)$. Given $0 < C_1 < C_2 < \infty$, $0 < C_3 < \frac{1}{2}$ and $0 < \beta < 1$, let $\mathcal{J}$ denote the set of integers $j_0 \in [C_1\, n, C_2\, n]$; let $\mathcal{P}(j_0)$ be the class of p for which $\bar{p}_j \geq \frac{1}{2} + C_3$ for all $j \leq j_0 - 2\,\lfloor n^\beta \rfloor$, $p_{j_0 - 1} > 0$, and $p_j = 0$ for $j \geq j_0$; and let $\mathcal{P}$ denote the union of $\mathcal{P}(j_0)$ over $j_0 \in \mathcal{J}$. Assume the data are generated by our simplified model in Subsection 3.1, let $\hat{j}_0$ denote the estimator of j_0 introduced in Subsection 3.2, with $\nu = \lfloor n^\beta \rfloor$, and let $\tilde{j}_0$ be any measurable function of the data. Property (6) below is a version of Theorem 2 holding uniformly in $p \in \mathcal{P}$, and (7) is a converse to that result.

Theorem 3. *If the data $I_1, \ldots, I_N$ are generated by our mathematical model, then there exist constants $B_1, B_2 > 0$ such that*

$$\lim_{n \to \infty} \sup_{p \in \mathcal{P}} P\big(\big|\hat{j}_0 - j_0\big| \geq B_1\, n^\beta\big) = 0\,, \quad (6)$$

and for all sufficiently large n,

$$\inf_{\tilde{j}_0} \sup_{p \in \mathcal{P}} P\big(\big|\tilde{j}_0 - j_0\big| > B_2\, n^\beta\big) = 1\,. \quad (7)$$

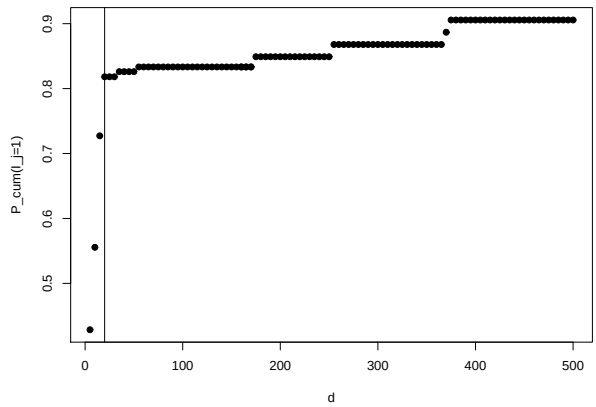

Fig. 3. Plot of cumulative probabilities $P_{\text{cum}}(I_j = 1)$ across a range of distances d.

6 An application in molecular medicine

To illustrate our approach we applied it to publicly available genomic data. They are part a large breast cancer study (Sørlie et al., 2001) with the aim of classifying carcinomas based on variations in gene expression patterns derived from cDNA microarrays (Stanford technology). We reanalyzed one set of genes, hybridized in two different laboratories, applying the R package `samr`, thus obtaining two rankings (rank lists) consisting of the same genes, based on differential expression (criterion: four-fold change). The list length was $N = 500$ (no missing values). Our goal was to estimate the point of degeneration into noise j_0, assuming $p = \frac{1}{2}$, and a top-k list of genes highly supported by the assessment of both laboratories.

To execute our algorithm we had to specify the distance parameter d. For that purpose we produced an exploratory plot (see Fig. 4) of the cumulative probabilities $P_{\text{cum}}(I_j = 1)$ for j and for $d \in \{5, 10, 15, 20, 25, \ldots, 500\}$, pointing at $d = 20$ (vertical line at that $j = d$ where the steps of $P_{\text{cum}}(I_j = 1)$ stabilise). Assuming $m = 0$ and that C is chosen from the interval $(0.25, 1]$, a pilot sample size of $\nu = 40$ resulting in $\hat{j}_0 = 22$ seems to be reasonable; see Fig. 5. Further we can conclude from this plot that $0.25 < C < 0.6$ robustifies the effect of the pilot sample size on the estimate $\hat{j}_0$ ($\nu = 10$ is too small and $\nu = 50$ is too large, the other values produce a $\hat{j}_0 \in [18, 24]$). The length of the top-k rank list is $k = \hat{j}_0 - 1 = 21$ (remember, the full lists consist of 500 objects). For this value of k, the overlap of objects (genes) between the two laboratories is about 80%. These are the differentially expressed genes that should for sure be further analysed with respect to patient survival.

References

AMOSOVA, N.N. (1972): Limit theorems for the probabilities of moderate deviations. (In Russian). *Vestnik Leningrad. Univ. 13, 5-14.*

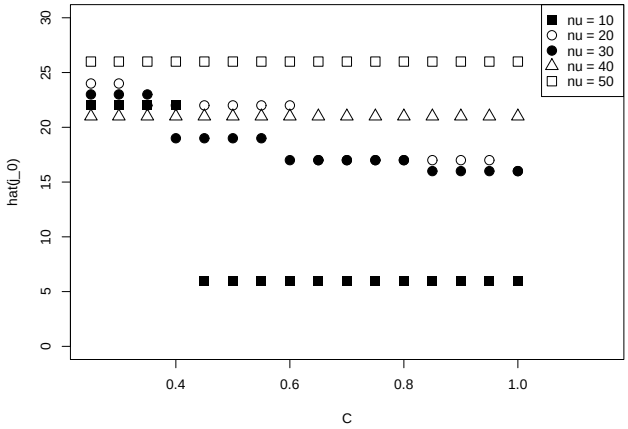

Fig. 4. Plot of estimates $\hat{j}_0$ in dependence of constant C and tuning parameter ν.

BRADLEY, R.A., TERRY, M.A. (1952): Rank analysis of incomplete block designs. *Biometrika 39, 324-345.*

DECONDE, R.P., HAWLEY, S., FALCON, S., CLEGG, N., KNUDSEN, B., and ETZIONI, R. (2006): Combined results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology 5, 1, article 15.*

DONOHO, D.L., JOHNSTONE, I. (1994): Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81, 425-455.*

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G., PICARD, D. (1995): Wavelet shrinkage: asymptopia? (With discussion). *J. Roy. Statist. Soc. B 57, 301-369.*

WORK, C., KUMAR, R., NAOR, M., SIVAKUMAR, D. (2001): Rank aggregation methods for the Web. http://www10.org/cdrom/papers/577/

FAGIN, R., KUMAR, R., SIVAKUMAR, D. (2003): Comparing top-k lists. *SIAM J. Discrete Math. 17, 134-160.*

KENDALL, M., GIBBONS, J.D. (1990): *Rank Correlation Methods.* Edward Arnold, London.

LIN, S., DING, J. (2008): Integration of ranked lists via Cross Entropy Monte Carlo with applications to mRNA and microRNA studies. *Biometrics*, to appear.

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

RUBIN, H., SETHURAMAN, J. (1965): Probabilities of moderate deviations. *Sankhyā A 27, 325-346.*

THURSTONE, L.L. (1931): Rank order as a psychological method. *J. Exp. Psychol. 14, 187-201.*

SØRLIE et al. (2001): Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS 98, 10869-10874.*

Acknowledgement: The second author would like to thank his PhD student Eva Budinská (Masaryk University) for help with the computations, and Gerhard Bachmaier (Medical University of Graz) for technical support.

Part XIII

Random Search Algorithms

Monitoring Random Start Forward Searches for Multivariate Data

Anthony C. Atkinson[1], Marco Riani[2], and Andrea Cerioli[2]

[1] Department of Statistics, London School of Economics
London WC2A 2AE, UK, *a.c.atkinson@lse.ac.uk*
[2] Dipartimento di Economia, Università di Parma
43100 Parma, Italy, *mriani@unipr.it*, *andrea.cerioli@unipr.it*

Abstract. During a forward search from a robustly chosen starting point the plot of maximum Mahalanobis distances of observations in the subset may provide a test for outliers. This is not the customary test. We obtain distributional results for this distance during the search and exemplify its use. However, if clusters are present in the data, searches from random starts are required for their detection. We show that our new statistic has the same distributional properties whether the searches have random or robustly chosen starting points.

Keywords: clustering, Mahalanobis distance, order statistic, outlier detection, robustness

1 Introduction

The forward search is a powerful general method for detecting systematic or random departures from statistical models, such as those caused by outliers and the presence of clusters. The forward search for multivariate data is given book-length treatment by Atkinson, Riani and Cerioli (2004). To detect outliers they study the evolution of Mahalanobis distances calculated during a search through the data that starts from a carefully selected subset of observations. More recently Atkinson and Riani (2007) suggested the use of many searches starting from random starting points as a tool in the detection of clusters. An important aspect of this work is the provision of bounds against which to judge the observed values of the distances. Atkinson and Riani (2007) use simulation for this purpose as well as providing approximate numerical values for the quantiles of the distribution.

These theoretical results are for the minimum Mahalanobis distance of observations not in the subset used for fitting when the starting point of the search is robustly selected. In this paper we consider instead the alternative statistic of the maximum Mahalanobis distance amongst observations in the subset. We derive good approximations to its distribution during the forward search and empirically compare its distribution to that of the minimum distance, both for random and robust starts. We find for the maximum distance,

but not for the minimum, that the distribution of the distance does not depend on how the search starts. Our ultimate purpose is a more automatic method of outlier and cluster identification.

We start in §2 with an introduction to the forward search that emphasises the importance of Mahalanobis distances in outlier detection. Some introductory theoretical results for the distributions of distances are in §3. Section 4 introduces the importance of random start searches in cluster detection. Our main theoretical results are in §5 where we use results on order statistics to derive good approximations to the distribution of the maximum distance during the search. Our methods are exemplified in §6 by the analysis of data on horse mussels. The comparison of distributions for random and elliptical starts to the search is conducted by simulation in §7.

2 Mahalanobis distances and the forward search

The tools that we use for outlier detection and cluster identification are plots of various Mahalanobis distances. The squared distances for the sample are defined as

$$d_i^2(\hat{\mu}, \hat{\Sigma}) = \{y_i - \hat{\mu}\}^T \hat{\Sigma}^{-1} \{y_i - \hat{\mu}\}, \tag{1}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are estimates of the mean and covariance matrix of the n observations.

In the forward search the parameters μ and Σ are replaced by their standard unbiased estimators from a subset of m observations, yielding estimates $\hat{\mu}(m)$ and $\hat{\Sigma}(m)$. From this subset we obtain n squared Mahalanobis distances

$$d_i^2(m) = \{y_i - \hat{\mu}(m)\}^T \hat{\Sigma}^{-1}(m) \{y_i - \hat{\mu}(m)\}, \qquad i = 1, \ldots, n. \tag{2}$$

We start with a subset of m_0 observations which grows in size during the search. When a subset $S(m)$ of m observations is used in fitting, we order the squared distances and take the observations corresponding to the $m + 1$ smallest as the new subset $S(m + 1)$. In what we call 'normal progression' this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave.

In our examples we look at forward plots of quantities derived from the distances $d_i(m)$. These distances tend to decrease as n increases. If interest is in the latter part of the search we may use **scaled** distances

$$d_i^{\mathrm{SC}}(m) = d_i(m) \times \left(|\hat{\Sigma}(m)| / |\hat{\Sigma}(n)| \right)^{1/2v}, \tag{3}$$

where v is the dimension of the observations y and $\hat{\Sigma}(n)$ is the estimate of Σ at the end of the search.

To detect outliers Atkinson et al. (2004) and Atkinson and Riani (2007) examined the minimum Mahalanobis distance amongst observations not in the subset

$$d_{\min}(m) = \min d_i(m) \quad i \notin S(m), \tag{4}$$

or its scaled version $d^{\text{sc}}_{\min}(m)$. In either case let this be observation $i_{\min}(m)$. If observation $i_{\min}(m)$ is an outlier relative to the other m observations, the distance (4) will be large compared to its reference distribution.

In this paper we investigate instead the properties of the maximum Mahalanobis distance amongst the m observations in the subset

$$d_{\max}(m) = \max d_i(m) \quad i \in S(m), \tag{5}$$

letting this be observation $i_{\max}(m)$. Whether we monitor $d_{\max}(m)$ or $d_{\min}(m)$ the search is the same, progressing through the ordering of $d_i^2(m)$.

3 Minimum and maximum Mahalanobis distances

We now consider the relationship between $d_{\min}(m)$ and $d_{\max}(m)$ as outlier tests. This relationship depends on the subsets $S(m)$ and $S(m+1)$.

Let the kth largest ordered Mahalanobis distance be $d_{[k]}(m)$ when estimation is based on the subset $S(m)$. In normal progression

$$d_{[m+1]}(m) = d_{\min}(m) \tag{6}$$

and $S(m+1)$ is formed from $S(m)$ by the addition of observation $i_{\min}(m)$. Likewise, in normal progression this will give rise to the largest distance within the new subset, that is

$$d_{[m+1]}(m+1) = d_{\max}(m+1) = d_{i\min(m)}(m+1). \tag{7}$$

The distance $d_{i\min(m)}(m+1)$ is that for the new observation $i_{\min}(m)$ when the parameters are estimated from $S(m+1)$. The consequence of (7) is that both $d_{\max}(m+1)$ and $d_{\min}(m)$ are tests of the outlyingness of observation $i_{\min}(m)$.

Although both statistics are testing the same hypothesis they do not have the same numerical value and should be referred to different null distributions. In §5 we discuss the effect of the ranking of the observations on these distributions as well as the consequence of estimating μ and Σ from a subset of the observations. For estimates using all n observations $d_{\max}(n)$ is one of the distances in (1). Standard distributional results in, for example, Atkinson et al. (2004, §2.6) show that

$$d_i^2(\hat{\mu}, \hat{\Sigma}) = d_i^2(n) \sim \frac{(n-1)^2}{n} \text{ Beta}\left(\frac{v}{2}, \frac{n-v-1}{2}\right). \tag{8}$$

On the other hand, $d^2_{\min}(n-1)$ is a deletion distance in which the parameters are estimated, in general, with the omission of observation i. The distribution of such distances is

$$d^2_{(i)} \sim \frac{n}{(n-1)} \frac{v(n-2)}{(n-v-1)} F_{v,n-v-1}, \tag{9}$$

although the distribution of $d^2_{\min}(n-1)$ depends on the order statistics from this distribution. For moderate n the range of the distribution of $d^2_i(n)$ in (8) is approximately $(0, n)$ rather than the unbounded range for the F distribution of the deletion distances. As we shall see, the consequence is that the distribution of $d^2_{\max}(m)$ has much shorter tails than that of $d^2_{\min}(m)$, particularly for small m.

Our argument has been derived assuming normal progression. This occurs under the null hypothesis of a single multivariate normal population when there are no outliers or clusters in the data, so that the ordering of the observations by closeness to the fitted model does not alter appreciably during the search. Then we obtain very similar forward plots of $d_{\max}(m)$ and $d_{\min}(m)$, even if they have to be interpreted against different null distributions. In fact, we do not need the order to remain unchanged, but only that $i_{\min}(m)$ and $i_{\max}(m+1)$ are the same observation and that the other observations in $S(m+1)$ are those that were in $S(m)$. Dispersed outliers likewise do not appreciably affect the ordering of the data. This is however affected by clusters of observations that cause appreciable changes in the parameter estimates as they enter $S(m)$. A discussion of the ordering of observations within and without $S(m)$ is on pp. 68-9 of Atkinson et al. (2004).

4 Elliptical and random starts

To find the starting subset for the search Atkinson et al. (2004) use the robust bivariate boxplots of Zani, Riani and Corbellini (1998) to pick a starting set $S^*(m_0)$ that excludes any two-dimensional outliers. The boxplots have elliptical contours, so we refer to this method as the elliptical start. However, if there are clusters in the data, the elliptical start may lead to a search in which observations from several clusters enter the subset in sequence in such a way that the clusters are not revealed. Searches from more than one starting point are then needed to reveal the clustering structure. If we start with an initial subset of observations from each cluster in turn, the other clusters are revealed as outliers. However, such a procedure is not suitable for automatic cluster detection. Atkinson and Riani (2007) therefore instead run many forward searches from randomly selected starting points, monitoring the evolution of the values of $d_{\min}(m)$ as the searches progress. Here we monitor $d_{\max}(m)$.

As the search progresses, the examples of Atkinson and Riani (2007) show that the effect of the starting point decreases. Once two searches have the same subsets $S(m)$ for some m, they will have the same subsets for all successive m. Typically, in the last third of the search the individual searches from random starts converge to that from the elliptical start. The implication is that the same envelopes can be used, except in the very early stages of the search, whether we use random or elliptical starts. If we are looking for a few

outliers, we will be looking at the end of the search. However, the envelopes for $d_{\min}(m)$ and $d_{\max}(m)$ will be different.

5 Envelopes from order statistics

For relatively small samples we can use simulation to obtain envelopes for $d_{\max}(m)$ during the search. For larger samples we adapt the method of Riani, Atkinson and Cerioli (2007) who find very good approximations to the envelopes for $d_{\min}(m)$ using order statistics and a result of Tallis (1963) on truncated multivariate normal distributions.

Let $Y_{[m]}$ be the mth order statistic from a sample of size n from a univariate distribution with c.d.f. $G(y)$. From, for example Lehmann (1991, p. 353) and Guenther (1977), the required quantile of order γ of the distribution of $Y_{[m]}$ say $y_{m,n;\gamma}$ can be obtained as

$$y_{m,n;\gamma} = G^{-1}\left(\frac{m}{m + (n-m+1)x_{2(n-m+1),2m;1-\gamma}}\right), \tag{10}$$

where $x_{2(n-m+1),2m;1-\gamma}$ is the quantile of order $1-\gamma$ of the F distribution with $2(n-m+1)$ and $2m$ degrees of freedom. Riani et al. (2007) comment that care needs to be taken to ensure that the numerical calculation of this inverse distribution is sufficiently accurate as $m \to n$, particularly for large n and extreme γ.

We now consider our choice of $G(x)$, which is different from that of Riani et al. (2007). We estimate Σ on $m-1$ degrees of freedom. The distribution of the m distances in the subset can, from (8), be written as

$$d_i^2(m) \sim \frac{(m-1)^2}{m} \text{ Beta}\left(\frac{v}{2}, \frac{m-v-1}{2}\right), \qquad i \in S(m). \tag{11}$$

The estimate of Σ that we use is biased since it is calculated from the m observations in the subset that have been chosen as having the m smallest distances. However, in the calculation of the **scaled** distances (3) we approximately correct for this effect by multiplication by a ratio derived from estimates of Σ. So the envelopes for the scaled Mahalanobis distances derived from $d_{\max}(m)$ are given by

$$V_{m,\gamma} = \sqrt{\frac{(m-1)^2}{m}}\sqrt{y_{m,n;\gamma}}, \tag{12}$$

with G the beta distribution in (11).

For **unscaled** distances we need to correct for the bias in the estimate of Σ. We follow Riani et al. (2007) and consider elliptical truncation in the multivariate normal distribution. From the results of Tallis (1963) they obtain the large-sample correction factor

$$c_{FS}(m) = \frac{m/n}{P(X^2_{v+2} < \chi^2_{v,m/n})}, \tag{13}$$

with $\chi^2_{v,m/n}$ the m/n quantile of χ^2_v and X^2_{v+2} a chi-squared random variable on $v+2$ degrees of freedom. Envelopes for unscaled distances are then obtained by scaling up the values of the order statistics

$$V^*_{m,\gamma} = c_{FS}(m)V_{m,\gamma}.$$

Figure 1 shows the agreement between simulated envelopes (continuous lines) and theoretical envelopes (dotted lines) for $d_{\max}(m)$ when $n = 1000$. Scaled distances are in the upper panel; agreement between the two sets of envelopes is excellent throughout virtually the whole range. Agreement for the unscaled distances in the lower panel of the figure is less good, but is certainly more than satisfactory for inferences about outliers at least in the last half of the search.

Unfortunately, the inclusion of $\hat{\Sigma}(n)$ in the expression for scaled distances (3) results in small distances in the presence of outliers, due to the inflation of the variance estimate and to consequent difficulties of interpretation. For practical data analysis we have to use the unscaled distances, which are less well approximated.

6 Horse mussels

As an example of the uses of elliptical and random starts in the analysis of multivariate data we look at measurements on horse mussels from New Zealand introduced by Cook and Weisberg (1994, p. 161) who treat them as regression with muscle mass, the edible portion of the mussel, as response. They focus on independent transformations of the response and of one of the explanatory variables. Atkinson et al. (2004, §4.9) consider multivariate normality obtained by joint transformation of all five variables.

There are 82 observations on five variables: shell length, width, height and mass and the mass of the mussels' muscle, which is the edible part.

We begin with an analysis of the untransformed data using a forward search with an elliptical start. The left-hand panel of Figure 2 monitors $d_{\min}(m)$, whereas the right-hand panel monitors $d_{\max}(m)$. The two sets of simulation envelopes were found by direct simulation of 5,000 forward searches. The figure shows how very different the two distributions are at the beginning of the search. That in the left-hand panel for $d_{\min}(m)$ is derived from the unbounded F distribution (9) whereas that for $d_{\max}(m)$ in the right-hand panel is derived from the beta distribution (11).

The two traces are very similar once they are calibrated by the envelopes. They both show appreciable departure from multivariate normality in the last one third of the search. Since we are selecting observations by their closeness to the multivariate normal model, we expect departure, if any, to be at the end of the search. Even allowing for the scaling of the two plots, the maximum distances seem to show less fluctuation at the beginning of the search. For

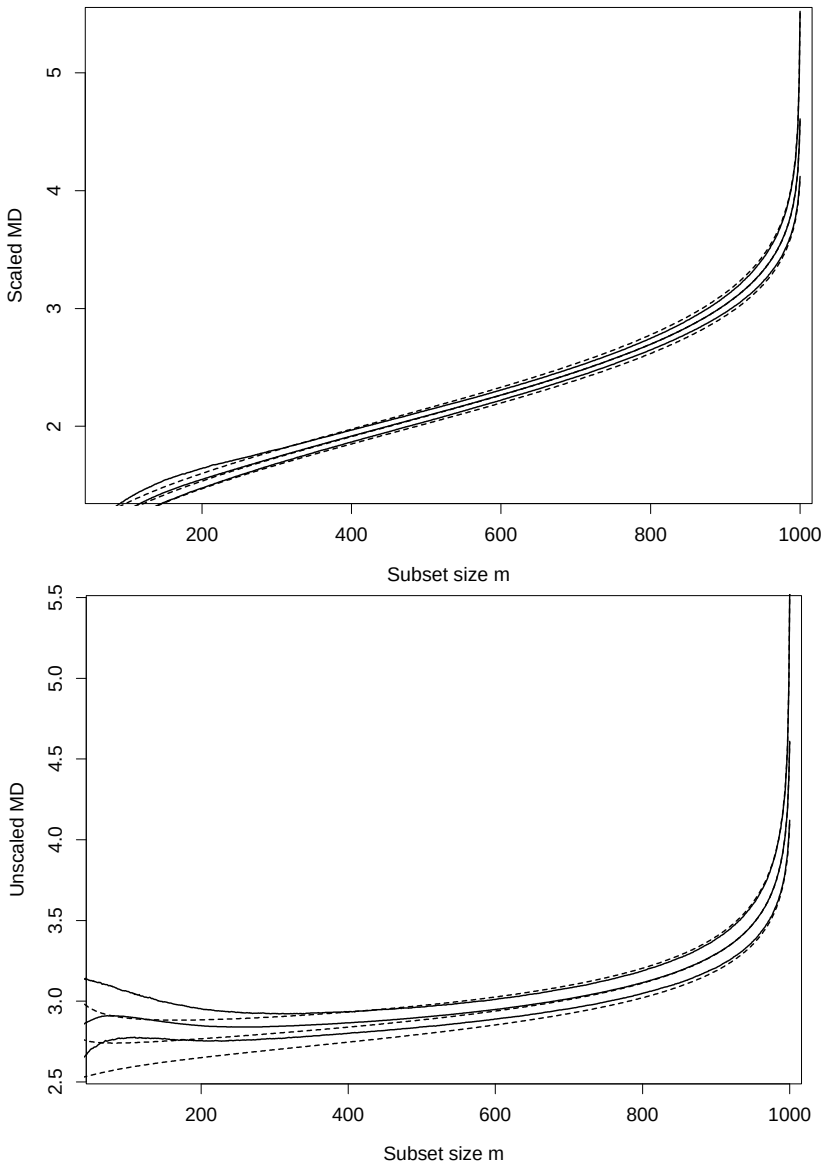

Fig. 1. Envelopes for Mahalanobis distances $d_{\max}(m)$ when $n = 1000$ and $v = 5$. Dotted lines from order statistics, continuous lines from 5,000 simulations. Upper panel scaled distances, lower panel unscaled distances. Elliptical starts. 1%, 50% and 99% points.

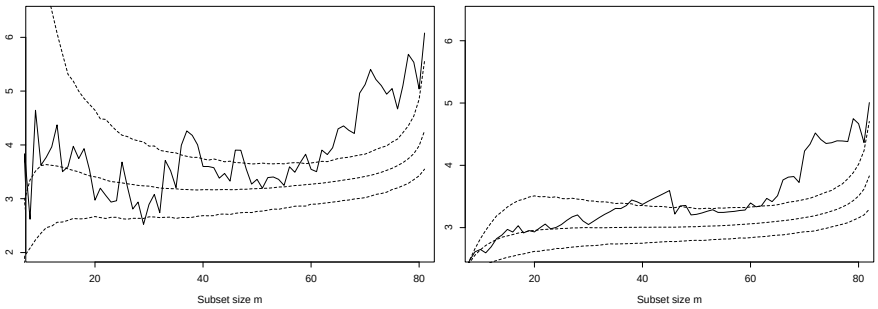

Fig. 2. Horse mussels: forward search on untransformed data. Left-hand panel $d_{\min}(m)$, right-hand panel $d_{\max}(m)$. Elliptical starts; 1%, 50% and 99% points from 5,000 simulations.

much of the rest of the paper we focus on plots of the maximum Mahalanobis distances $d_{\max}(m)$.

We now analyse the data using a multivariate version of the parametric transformation of Box and Cox (1964). As a result of their analysis Atkinson et al. (2004) suggest the vector of transformation parameters $\lambda = (0.5, 0, 0.5, 0, 0)^T$; that is, the square root transformation for y_1 and y_3 and the logarithmic transformation for the other three variables. We look at forward plots of $d_{\max}(m)$ to see whether this transformation yields multivariate normality.

The upper-left panel of Figure 3 shows the maximum distance for all $n = 82$ observations for the transformed data. The contrast with the right-hand panel of Figure 2 is informative. The plot still goes out of the 99% envelope at the end of the search, but the number of outliers is much smaller, now only around 5.

The last five units to enter are those numbered 37, 16, 78, 8 and finally 48. The plot of the maximum distance in the upper-right panel of Figure 3 shows that, with these five observations deleted, the last value just lies below the 99% point of the distribution. We have found a multivariate normal sample, after transformation, with five outliers. That there are five outliers, not four, is confirmed in the lower panel of Figure 3 where observation 37 has been reincluded. Now the plot of maximum distances goes outside the 99% envelope at the end of the search.

The limits in figures like 3 have been simulated to have the required pointwise level, that is they are correct for each m considered independently. However, the probability that the observed trace of values of $d_{\max}(m)$ exceeds a specific bound at least once during the search is much greater than the pointwise value. Atkinson and Riani (2006) evaluate such simultaneous probabilities; they are surprisingly high.

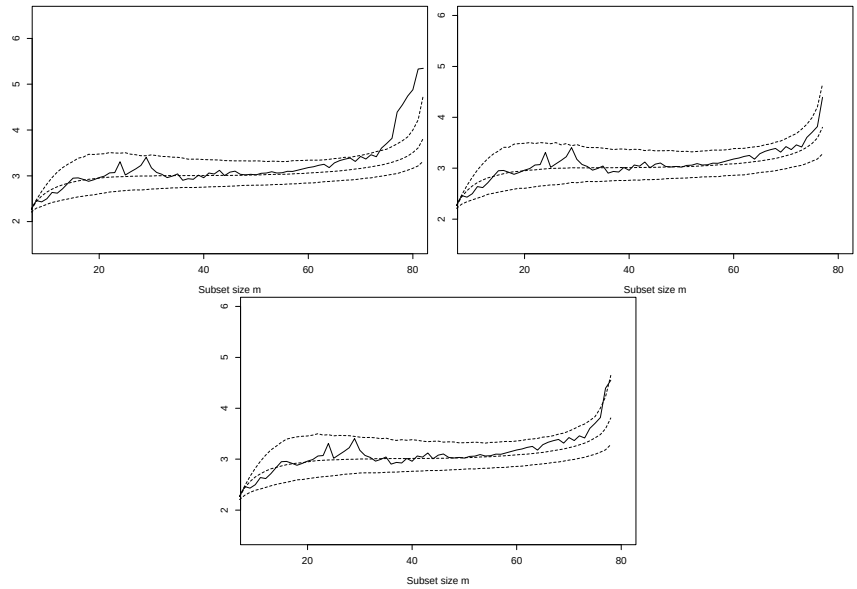

Fig. 3. Horse mussels: forward search on transformed data. Plots of $d_{\max}(m)$ for three different sample sizes. Upper-left panel $n = 82$; upper-right panel, observations 37, 16, 78, 8 and 48 removed ($n = 77$). Lower panel, observation 37 re-included ($n = 78$). Elliptical starts; 1%, 50% and 99% points from 5,000 simulations. There are five outliers.

7 Elliptical and random starts

We have analysed the mussels data and investigated the properties of $d_{\max}(m)$ and $d_{\min}(m)$ using searches with elliptical starts. Finally we look at the properties of plots of both distances when random starts are used, for example to aid in the identification of clusters.

The upper panel of Figure 4 presents a comparison of simulated envelopes for random and elliptical starts for $d_{\max}(m)$ from data with the same dimensions as the mussels data. For this small data set there is no operationally important difference between the two envelopes. The important conclusion is that, for larger data sets, we can use the approximations of §5 based on order statistics whether we are using random or elliptical starts.

The surprising conclusion that we obtain the same envelopes for searches from elliptical or random starts however does not hold when instead we monitor $d_{\min}(m)$. The lower panel of Figure 4 repeats the simulations for $d_{\min}(m)$. Now there is a noticeable difference, during the first half of the search, between the envelopes for random and those from elliptical starts.

We now consider the implications of this difference on the properties of individual searches. The left-hand panel of Figure 5 repeats the simulated envelopes for elliptical starts from the upper panel of Figure 4 and adds 250

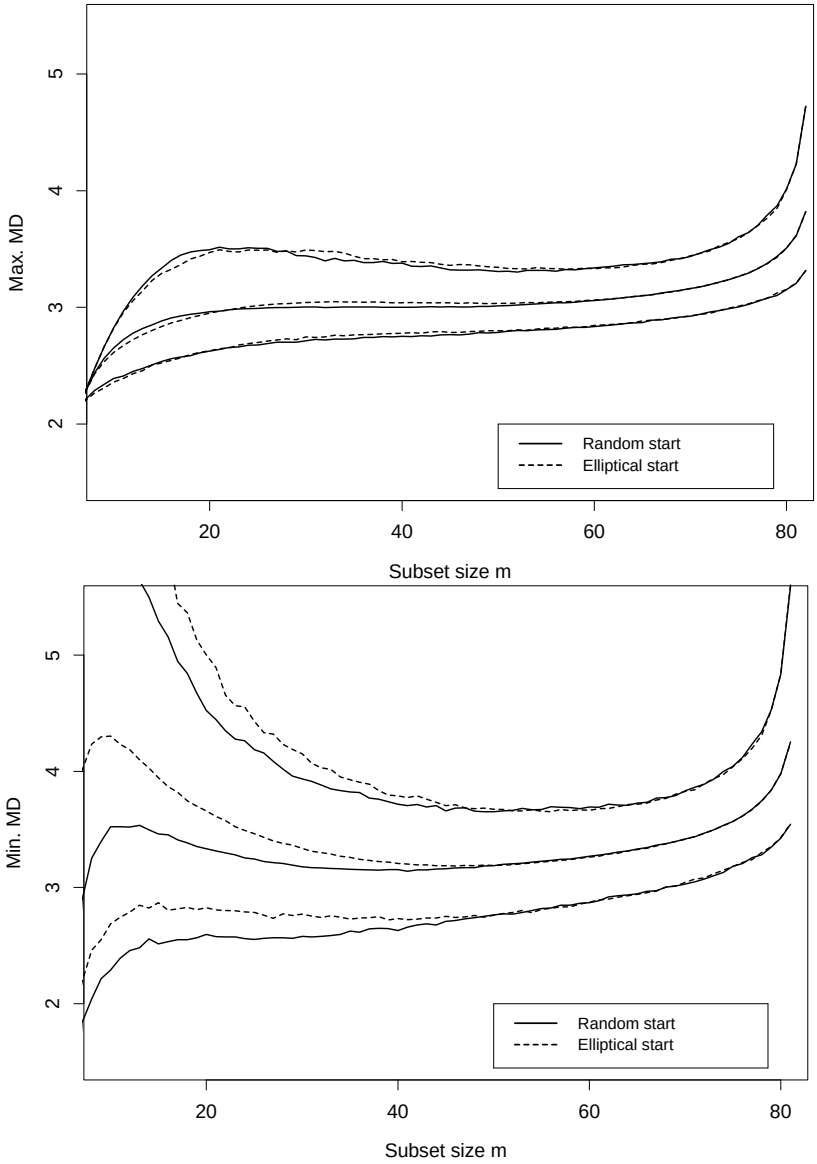

Fig. 4. Simulated envelopes for Mahalanobis distances when $n = 82$ and $v = 5$. Upper panel, $d_{\max}(m)$, lower panel $d_{\min}(m)$ Dotted lines from elliptical starts, continuous lines from random starts. 1%, 50% and 99% points from 5,000 simulations.

trajectories of distances for $d_{\max}(m)$ for simulations from random starting points. The simulated values nicely fill the envelope, although there are a surprising number of transient peaks above the envelope. The right-hand panel of the figure repeats this process of envelopes from elliptical starts and trajectories from random starts but for the minimum distances $d_{\min}(m)$. Now, as we would expect from Figure 4, the simulated values sit a little low in the envelopes. If a subset contains one or more outliers, these will give rise to a too large estimate of Σ. As a consequence, some of the distances of units not included in the subset will be too small and the smallest of these will be selected as $d_{\min}(m)$. On the contrary, if outliers are present in $S(m)$ when we calculate $d_{\max}(m)$, the distance that we look at will be that for one of the outliers and so will not be shrunken due to the too-large estimate of Σ.

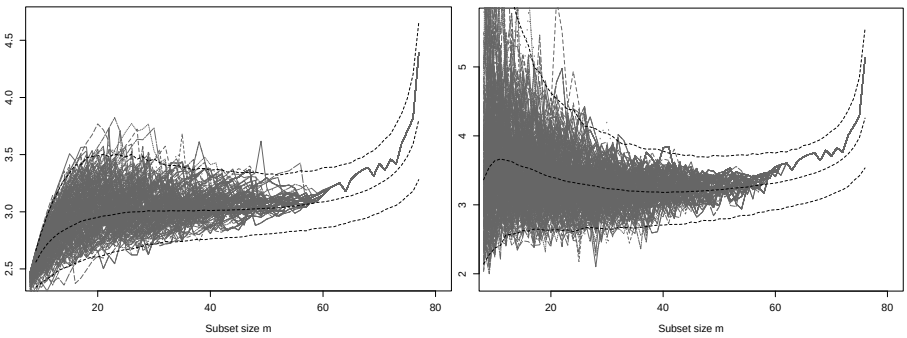

Fig. 5. Simulated envelopes from elliptical starts for Mahalanobis distances when $n = 82$ and $v = 5$ with 250 trajectories from random starts. Left-hand panel $d_{\max}(m)$, right-hand panel $d_{\min}(m)$.

Our justification for the use of random start forward searches was that searches from elliptical starts may not detect clusters in the data if these start from a subset of units in more than one cluster. We have however analysed the mussels data using values of $d_{\max}(m)$ from elliptical starts. Our conclusion, from Figure 3, was that after transformation there were 77 units from a multivariate normal population and five outliers. We checked this conclusion using random start forward searches with $d_{\max}(m)$ and failed to detect any clusters.

The purpose of this paper has been to explore the properties of the maximum distance $d_{\max}(m)$. We have found its null distribution and obtained good approximations to this distribution for use in the forward search.The lack of dependence of this distribution on the starting point of the search is an appealing feature. However, we need to investigate the properties of this measure when the null distribution does not hold. One particular question is whether use of $d_{\max}(m)$ provides tests for outliers and clusters that are as powerful as those using the customary minimum distance $d_{\min}(m)$.

References

ATKINSON, A.C. and RIANI, M. (2000): *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

ATKINSON, A.C. and RIANI, M. (2006): Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics 15, 460–476*.

ATKINSON, A.C. and RIANI, M. (2007): Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis 52, 272-285*. doi:10.1016/j.csda.2006.12.034.

ATKINSON, A.C., RIANI, M. and CERIOLI, A (2004): *Exploring Multivariate Data with the Forward Search*. Springer-Verlag, New York.

BOX, G.E.P. and COX, D.R. (1964): An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B 26, 211-246*.

COOK, R.D. and WEISBERG, S. (1994): *An Introduction to Regression Graphics*. Wiley, New York.

GUENTHER, W.C. (1977): An easy method for obtaining percentage points of order statistics. *Technometrics 19, 319-321*.

LEHMANN, E. (1991): *Point Estimation, 2nd edition*. Wiley, New York.

RIANI, M., ATKINSON, A.C. and CERIOLI, A. (2007): *Results in finding an unknown number of multivariate outliers in large data sets*. Research Report 140, Department of Statistics, London School of Economics.

TALLIS, G.M.(1963): Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics 34, 940-944*.

ZANI, S., RIANI, M. and CORBELLINI, A. (1998): Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis 28, 257-270*.

Generalized Differential Evolution for General Non-Linear Optimization

Saku Kukkonen[1] and Jouni Lampinen[2]

[1] Lappeenranta University of Technology
Laboratory of Applied Mathematics
P.O. Box 20, FI-53851 Lappeenranta, Finland,
Saku.Kukkonen@lut.fi
[2] University of Vaasa
Department of Computer Science
P.O. Box 700, FI-65101 Vaasa, Finland,
Jouni.Lampinen@uwasa.fi

Abstract. In this article different development phases and the current state of the Generalized Differential Evolution algorithm, that we have developed, are summarized. Generalized Differential Evolution is a general purpose solver for non-linear global optimization problems with multiple constraints and objectives. It is based on a relatively recent Evolutionary Algorithm, Differential Evolution, extending it for solving constrained multi-objective problems.

Keywords: evolutionary algorithms, differential evolution, multi-objective optimization, diversity maintenance, constraint handling

1 Introduction

During the last couple of decades Evolutionary Algorithms (EAs) have gained increasing popularity due to their capability of dealing with difficult objective functions, which are, *e.g.*, discontinuous, non-convex, multi-modal, and non-differentiable. Also multi-objective EAs (MOEAs) have gained popularity since they are capable of providing multiple solution candidates in a single run that is especially desirable with multi-objective optimization problems (MOPs).

Constrained multi-objective optimization is referring to simultaneous optimization of M objective functions subjected to K constraint functions:

$$\begin{aligned} &\text{minimize} \quad \{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})\} \\ &\text{subject to } g_k(\boldsymbol{x}) \leq 0, \ k = 1, 2, \ldots, K \ . \end{aligned} \tag{1}$$

Typically, MOPs are converted to single-objective optimization problems by predefining weighting factors for different objectives to express the relative importance of each. Optimizing several objectives simultaneously without articulating the relative importance of each objective *a priori* is called Pareto-optimization (Pareto (1896)). A solution is *Pareto-optimal* if none of the

objectives can be improved without impairing at least one other objective. If the solution can be improved so, that at least one objective improves and the other objectives do not decline, the new solution (Pareto-)dominates the original solution.

The objective of Pareto-optimization is to find a set of solutions that are not dominated by any other solution. A set of Pareto-optimal solutions form a Pareto front, and an approximation of the Pareto front is called a set of non-dominated solutions, because the solutions in this set are not dominating each other in the space of objective functions. From the set of non-dominated solutions the decision-maker picks a solution, which provides a suitable compromise between the objectives for his/her needs. This can be viewed as *a posteriori* articulation of the decision-makers preferences on the relative importance of each objective.

Weak dominance relation $\preceq$ between two vectors is defined such that $\boldsymbol{x}_1$ weakly dominates $\boldsymbol{x}_2$, *i.e.*, $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$ iff $\forall i : f_i(\boldsymbol{x}_1) \leq f_i(\boldsymbol{x}_2)$. Dominance relation $\prec$ between two vectors is defined such a way that $\boldsymbol{x}_1$ dominates $\boldsymbol{x}_2$, *i.e.*, $\boldsymbol{x}_1 \prec \boldsymbol{x}_2$ iff $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2 \ \wedge \ \exists i : f_i(\boldsymbol{x}_1) < f_i(\boldsymbol{x}_2)$. The dominance relationship can be extended to take into consideration constraint values besides objective values. *Constraint-domination* $\prec_c$ is defined here such a way that $\boldsymbol{x}_1$ constraint-dominates $\boldsymbol{x}_2$, *i.e.*, $\boldsymbol{x}_1 \prec_c \boldsymbol{x}_2$ iff any of the following conditions is true (Lampinen (2001)):

1. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are infeasible and $\boldsymbol{x}_1$ dominates $\boldsymbol{x}_2$ in the constraint function violation space.
2. $\boldsymbol{x}_1$ is feasible and $\boldsymbol{x}_2$ is not.
3. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are feasible and $\boldsymbol{x}_1$ dominates $\boldsymbol{x}_2$ in the objective function space.

The definition for weak constraint-domination $\preceq_c$ is analogous when the dominance relation is changed to weak dominance in the above definition. The weak constraint-domination relation can be formally defined as:

$$\boldsymbol{x}_1 \preceq_c \boldsymbol{x}_2 \quad \text{iff} \quad \begin{cases} \begin{cases} \exists k \in \{1,\ldots,K\} : g_k(\boldsymbol{x}_1) > 0 \\ \wedge \\ \forall k \in \{1,\ldots,K\} : g'_k(\boldsymbol{x}_1) \leq g'_k(\boldsymbol{x}_2) \end{cases} \\ \vee \\ \begin{cases} \forall k \in \{1,\ldots,K\} : g_k(\boldsymbol{x}_1) \leq 0 \\ \wedge \\ \exists k \in \{1,\ldots,K\} : g_k(\boldsymbol{x}_2) > 0 \end{cases} \\ \vee \\ \begin{cases} \forall k \in \{1,\ldots,K\} : g_k(\boldsymbol{x}_1) \leq 0 \wedge g_k(\boldsymbol{x}_2) \leq 0 \\ \wedge \\ \forall m \in \{1,\ldots,M\} : f_m(\boldsymbol{x}_1) \leq f_m(\boldsymbol{x}_2) \end{cases} \end{cases}, \tag{2}$$

where $g'_k(\boldsymbol{x}) = \max(g_k(\boldsymbol{x}), 0)$ represents a constraint violation.

2 Differential Evolution

Differential Evolution (DE) (Storn and Price (1995), Price *et al.* (2005)) is a relatively new EA that has gained considerable popularity during the previous years. Design principles in DE are simplicity, efficiency, and use of floating-point encoding instead of binary numbers that are often used in Genetic Algorithms (Goldberg (1989)). The original DE algorithm was designed for unconstrained single objective optimization over continuous spaces. This article describes Generalized Differential Evolution (GDE) algorithm, a DE extension for multi-objective and multi-constrained optimization, and its development phases.

In the following, the most popular DE variant, *DE/rand/1/bin*, is described in detail. Since *DE/rand/1/bin*, is designed for unconstrained single-objective optimization, the notations in this section are for single-objective optimization.

Like any typical EA, DE is starting with an *initial population* of candidate solutions, which is then improved by applying *selection*, *mutation*, and *crossover* operations until *stopping criterion*, *e.g.*, a predefined for the number of generations to be computed, is reached. Typically, the values for the initial population are generated randomly. Formally this can be presented as:

$$\begin{array}{l} \mathcal{P}_G = \{\boldsymbol{x}_{1,G}, \boldsymbol{x}_{2,G}, \ldots, \boldsymbol{x}_{NP,G}\}, \quad \boldsymbol{x}_{i,G} = (x_{1,i,G}, x_{2,i,G}, \ldots, x_{D,i,G}) \\ x_{j,i,G=0} = x_j^{(lo)} + rand_j[0,1] \cdot \left(x_j^{(hi)} - x_j^{(lo)}\right) \\ i = 1, 2, \ldots, NP, \quad NP \geq 4, \quad j = 1, 2, \ldots, D \end{array} , \quad (3)$$

where $\mathcal{P}_G$ denotes a population after G generations, $\boldsymbol{x}_{i,G}$ denotes a decision vector (or individual) of the population, and $rand_j[0,1]$ denotes a uniformly distributed random variable in the range $[0,1]$. Terms $x_j^{(lo)}$ and $x_j^{(hi)}$ are the lower and upper parameter bounds, respectively. The size of the population is denoted by NP and the dimension of decision vectors is denoted by D.

The new trial solutions are generated by mutation and crossover operations. DE goes through each decision vector $\boldsymbol{x}_{i,G}$ of the population and creates a corresponding trial vector $\boldsymbol{u}_{i,G}$ as follows:

$$\begin{array}{l} r_1, r_2, r_3 \in \{1, 2, \ldots, NP\}, \\ \text{(randomly selected,} \\ \quad \text{except mutually different and different from } i\text{)} \\ j_{rand} = \text{round}\,(rand_i[0,1) \cdot D) \\ \text{for}(j = 1; j \leq D; j = j + 1) \\ \{ \\ \quad \text{if}(rand_j[0,1) < CR \vee j == j_{rand}) \\ \qquad u_{j,i,G} = x_{j,r_3,G} + F \cdot (x_{j,r_1,G} - x_{j,r_2,G}) \\ \quad \text{else} \\ \qquad u_{j,i,G} = x_{j,i,G} \\ \} \end{array} . \quad (4)$$

Indices r_1, r_2, and r_3 are referring to three randomly selected population members. *CR* and F are user defined control parameters. *CR* is a probability value controlling the crossover operation. F is a scaling factor controlling mutation and its value is typically $(0, 1+]$ (1+ is expressing that there is no hard theoretical upper limit). The weighted difference between two randomly chosen vectors $F \cdot (\boldsymbol{x}_{r_1,G} - \boldsymbol{x}_{r_2,G})$ defines the magnitude and direction of mutation. The difference is then simply added to a third randomly chosen vector, $\boldsymbol{x}_{r_3,G}$, in order to mutate it. Thus, the weighted differential of two population members is applied to mutate the third one.

DE is a self-adaptive process in the same way as in Covariance Matrix Adaptation Evolutionary Strategies (Hansen and Ostermeier (1996)) but without such algorithmic complexity and the computational burden of covariance matrix calculations that are scaling unfavorably with the problem dimensionality. Other strengths of DE are simplicity and ability to perform a rotationally invariant search.

Finally, after each mutation and crossover operation the generated new trial vector $\boldsymbol{u}_{i,G}$ is compared to the old decision vector $\boldsymbol{x}_{i,G}$. If the trial vector has equal or lower objective value, then it replaces the old vector in the population:

$$\boldsymbol{x}_{i,G+1} = \begin{cases} \boldsymbol{u}_{i,G} \text{ if } \quad f(\boldsymbol{u}_{i,G}) \leq f(\boldsymbol{x}_{i,G}) \\ \boldsymbol{x}_{i,G} \text{ otherwise} \end{cases} . \tag{5}$$

Thereby, the overall presentation of basic DE version, *DE/rand/1/bin*, is:

$$\text{Input :} D, G_{max}, NP \geq 4, F \in (0, 1+], CR \in [0, 1], \text{and initial bounds: } \boldsymbol{x}^{(lo)}, \boldsymbol{x}^{(hi)}$$

$$\text{Initialize :} \begin{cases} \forall i \leq NP \wedge \forall j \leq D : x_{j,i,G=0} = x_j^{(lo)} + rand_j[0,1] \cdot \left(x_j^{(hi)} - x_j^{(lo)} \right) \\ i = \{1, 2, \ldots, NP\}, j = \{1, 2, \ldots, D\}, G = 0, rand_j[0,1] \in [0,1] \end{cases}$$

$$\begin{cases} \text{While } G < G_{max} \\ \quad \begin{cases} \text{Mutate and recombine:} \\ r_1, r_2, r_3 \in \{1, 2, \ldots, NP\}, \text{randomly selected,} \\ \qquad \text{except mutually different and different from } i \\ j_{rand} \in \{1, 2, \ldots, D\}, \text{randomly selected for each } i \\ \\ \forall i \leq NP \begin{cases} \forall j \leq D, u_{j,i,G} = \begin{cases} x_{j,r_3,G} + F \cdot (x_{j,r_1,G} - x_{j,r_2,G}) \\ \qquad \text{if} \quad rand_j[0,1) < CR \vee j == j_{rand} \\ x_{j,i,G} \quad \text{otherwise} \end{cases} \\ \text{Select :} \\ \boldsymbol{x}_{i,G+1} = \begin{cases} \boldsymbol{u}_{i,G} \text{ if} f(\boldsymbol{u}_{i,G}) \leq f(\boldsymbol{x}_{i,G}) \\ \boldsymbol{x}_{i,G} \text{ otherwise} \end{cases} \end{cases} \\ G = G + 1 \end{cases} \end{cases} . \tag{6}$$

3 Generalized Differential Evolution

Generalized Differential Evolution (GDE) is an extension of DE for an arbitrary number of objectives and constraints. In the following, different development versions of GDE are briefly described and their performances are compared with two well known test problems, ZDT2 problem taken from (Deb (2001), p. 357) and DTLZ4 taken from (Deb *et al.* (2005)). With the both problems control parameter values $CR = 0.2$, $F = 0.2$, and $G_{max} = 250$ were applied. The size of the population, NP, for the bi-objective ZDT2 problem was 100, and for the tri-objective DTLZ4 problem was 200. The results for these problems are shown in Figure 1.

3.1 The first version, GDE1

The first version, GDE1, was proposed by Lampinen (2001) as a further development from the constraint handling approach based on dominance relation (Lampinen (2002)). GDE1 extended basic DE for constrained multi-objective optimization simply by modifying the selection operation of DE to apply weak constraint-domination as the selection criteria:

$$\boldsymbol{x}_{i,G+1} = \begin{cases} \boldsymbol{u}_{i,G} \text{ if } \quad \boldsymbol{u}_{i,G} \preceq_c \boldsymbol{x}_{i,G} \\ \boldsymbol{x}_{i,G} \text{ otherwise} \end{cases} . \tag{7}$$

The weak constraint-domination relation is used to have a congruity with the selection operation of DE. Thus, in the case of equality, the trial vector is preferred.

The final populations of GDE1 for the ZDT2 and DTLZ4 problems are shown in Figure 1. Only edges of Pareto-front of DTLZ4 are found. In both cases solution points are rather unevenly distributed. Especially DTLZ4 demonstrates that the diversity of solutions found by GDE1 isn't particularly good, since there is no diversity maintenance mechanism. Such a mechanism was therefore added into the subsequent GDE versions. Still, GDE1 provided surprisingly good results with some problems (Kukkonen and Lampinen (2004a), Kukkonen and Lampinen (2004b)) but has been found rather sensitive to the control parameters values (Kukkonen and Lampinen (2005)).

3.2 The second version, GDE2

The second version, GDE2, added on a diversity maintenance operation. Again, only the selection operation of basic DE was modified. The selection is based on crowding in the objective space when the trial and old vector are feasible and non-dominating each other (Kukkonen and Lampinen (2004c)):

$$\boldsymbol{x}_{i,G+1} = \begin{cases} \boldsymbol{u}_{i,G} \text{ if } \begin{cases} \boldsymbol{u}_{i,G} \preceq_c \boldsymbol{x}_{i,G} \\ \vee \\ \begin{cases} \forall k \in \{1,\ldots,K\} : g_k(\boldsymbol{u}_{i,G}) \leq 0 \\ \wedge \\ \neg\left(\boldsymbol{x}_{i,G} \prec \boldsymbol{u}_{i,G}\right) \\ \wedge \\ d_{\boldsymbol{u}_{i,G}} \geq d_{\boldsymbol{x}_{i,G}} \end{cases} \end{cases} \\ \boldsymbol{x}_{i,G} \text{ otherwise} \end{cases}, \tag{8}$$

where d_i is a distance measure for measuring the distance from a particular solution i to its neighbor solutions. As a distance measure *crowding distance* (Deb (2001) pp. 248–249) was used. Since there is no non-dominated sorting (Deb (2001) pp. 33–44), crowding is measured among the whole population. This improves the extent and distribution of the obtained set of solutions but slows down the convergence of the overall population because it favors isolated solutions far from the Pareto-front until all the solutions are converged near the Pareto-front. GDE2 is also observed sensitive to the selection of the control parameter values.

The final populations of GDE2 for ZDT2 and DTLZ4 problems are shown in Figure 1. A better diversity especially for DTLZ4 was obtained in comparison to GDE1 can be observed.

3.3 The third version, GDE3

The third version, GDE3, is the latest version (Kukkonen and Lampinen (2005), Kukkonen and Deb (2006a)). Besides the selection, another part of basic DE has also been modified. Now, in the case of comparing feasible and non-dominating solutions, both vectors are saved. Therefore, at the end of a generation the population may become larger than originally. Before proceeding to the next generation, the population size is reduced using non-dominated sorting and pruning based on diversity preservation to select the best solution candidates to survive.

The GDE3 algorithm is presented in Equation 9. Parts that are new compared to previous GDE versions are framed in Equation 9. Without these parts, the algorithm is identical to GDE1. After a generation, the size of the population may be larger than originally. In this case, the population size is then reduced back to the original size based on a similar selection approach used in NSGA-II (Deb (2001) pp. 245–253). Population members are sorted based on non-dominance and crowding. Then the worst population members according to these measurements are removed. Non-dominated sorting is modified to take into consideration also constraints. The selection based on the crowding distance is improved over the original method of NSGA-II to provide a better distributed set of vectors (Kukkonen and Deb (2006a)).

Compared to the earlier GDE versions GDE3 improves the ability to handle MOPs by giving a better distributed set of solutions and being less sen-

sitive to the selection of control parameter values. GDE3 has been compared against NSGA-II and found at least comparable according to experimental results (Kukkonen and Lampinen (2005)).

GDE3 has been later improved with a pruning technique for problems with more than two objectives (Kukkonen and Deb (2006b)). The final populations of GDE3 with the improved pruning technique (based on distances to two nearest neighbors, 2-NN) for the ZDT2 and DTLZ4 problems are shown in Figure 1. The distributions are clearly better for both problems than those provided by GDE1 and GDE2. In fact, the distribution is approaching the ideal one, and can be distinguished from it only by numerical measurements, not anymore by visual observations. GDE3 with the described diversity preservation technique participated a multi-objective optimization competition arranged at the 2007 IEEE Congress on Evolutionary Computation (Kukkonen and Lampinen (2007)). GDE3 received a winning entry nomination in the competition.

4 Constrained multi-objective optimization example

All GDE versions include a constraint handling mechanism. This mechanism was first introduced for single-objective optimization with DE by Lampinen (2002) and later extended into multi-objective optimization with the GDE.

To provide a practical example of solving a constrained multi-objective problem, we applied each GDE version to a bi-objective spring design problem (Deb (2001) pp. 453–455). The problem is to design a helical compression spring, which has a minimum volume and minimal stress. Both objective functions are non-linear and the problem has three variables: the number of spring coils x_1 (integer), the wire diameter x_2 (discrete having 42 non-equispaced values), and the coil diameter x_3 (real). Besides the boundary constraints, the problem has eight inequality constraints from which most are non-linear. Formal description of the problem is:

$$\begin{aligned}
\text{Minimize} \quad & f_1(\boldsymbol{x}) = 0.25\pi^2 {x_2}^2 x_3(x_1+2), \\
\text{Minimize} \quad & f_2(\boldsymbol{x}) = \tfrac{8KP_{max}x_3}{\pi {x_2}^3}, \\
\text{subject to} \quad & g_1(\boldsymbol{x}) = l_{max} - \tfrac{P_{max}}{k} - 1.05(x_1+2)x_2 \geq 0, \\
& g_2(\boldsymbol{x}) = x_2 - d_{min} \geq 0, \\
& g_3(\boldsymbol{x}) = D_{max} - (x_2 + x_3) \geq 0, \\
& g_4(\boldsymbol{x}) = C - 3 \geq 0, \\
& g_5(\boldsymbol{x}) = \delta_{pm} - \delta_p \geq 0, \\
& g_6(\boldsymbol{x}) = \tfrac{P_{max}-P}{k} - \delta_w \geq 0, \\
& g_7(\boldsymbol{x}) = S - \tfrac{8KP_{max}x_3}{\pi {x_2}^3} \geq 0, \\
& g_8(\boldsymbol{x}) = V_{max} - 0.25\pi^2 {x_2}^2 x_3(x_1+2) \geq 0, \\
& x_1 \text{ is integer, } x_2 \text{ is discrete, } x_3 \text{ is continuous .}
\end{aligned} \tag{10}$$

The parameters used are as follows:

$$\begin{array}{llll} K = \frac{4C-1}{4C-4} + \frac{0.615x_2}{x_3}, & P = 300 \text{ lb}, & D_{max} = 3 \text{ in}, & k = \frac{G{x_2}^4}{8x_1{x_3}^3}, \\ P_{max} = 1000 \text{ lb}, & \delta_w = 1.25 \text{ in}, & \delta_p = \frac{P}{k}, & l_{max} = 14 \text{ in}, \\ S = 189000 \text{ psi}, & \delta_{pm} = 6 \text{ in}, & d_{min} = 0.12 \text{ in}, & C = D/d, \\ G = 11500000, & V_{max} = 30 \text{ in}^3 \ . & & \end{array} \tag{11}$$

Non-dominated points extracted from the final population of the different GDE versions after a single run are shown in Figure 2. The applied control parameter values for all GDE versions were $CR = 0.5$, $F = 0.3$, $NP = 100$, and $G_{max} = 100$.

The GDE versions can be implemented in such a way that the number of function evaluations is reduced. The reason for this is that the constraint-domination relation is used in the selection. Even comparison between single constraint values can reveal that the trial vector does not constraint-dominate the old vector, and therefore the old vector is preserved. The number of function evaluations needed for the GDE versions with the spring design problem are reported in Table 1. It can be observed that the constraint handling approach used in the GDE versions reduce the actual number of function evaluations.

Further examples on solving constrained multi-objective problems by GDE can be found from (Kukkonen and Lampinen (2004a), Kukkonen and Lampinen (2005), Kukkonen and Lampinen (2006)).

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	f_1	f_2
GDE1	10100	8619	8526	8458	7321	7320	4607	3862	3857	1881
GDE2	10100	8155	8104	7966	7300	7298	5184	4622	4604	4604
GDE3	10100	9036	8963	8912	8582	8582	4874	4361	4353	4353

Table 1. Number of needed constraint (g) and objective (f) function evaluations by GDE1, GDE2, and GDE3 for the spring design problem.

5 Conclusions

Generalized Differential Evolution (GDE) is a real-coded general purpose EA extended from DE to handle multiple objectives and constraints.

The first version, GDE1, extended DE for constrained multi-objective optimization in a simple way by applying a modified the selection rule of basic DE. The basic idea is that the trial vector is selected to replace the old vector in the next generation if the trial vector weakly constraint-dominates the old vector. There is neither explicit non-dominated sorting during the optimization process nor any mechanism for maintaining diversity. Also, there is no extra repository for non-dominated vectors. Still, GDE1 has been observed

to provide surprisingly good results but also found rather sensitive to the selection of the control parameter values.

The second version, GDE2, makes a decision based on crowding when the trial and the old vector are feasible and non-dominating each other in the objective function space. This improves the extent and distribution of an obtained set of solutions but slows down the convergence of the population because it favors isolated solutions far from the Pareto-front. This version, too, has been observed to be rather sensitive to the selection of the control parameters values.

The third and latest version is GDE3. Besides the selection, another part of basic DE has also been modified. Now, in the case of feasible and non-dominating solutions, both vectors are saved for the population of the next generation. Before starting the next generation, the size of the population is reduced using non-dominated sorting and pruning based on diversity preservation. GDE3 has been later improved with a pruning technique for problems with more than two objectives. GDE3 provides better distribution than the earlier GDE versions, and it seems to be also more robust in terms of the selection of the control parameter values.

In the end, it can be concluded that GDE3 is a potential alternative for constrained multi-objective global optimization.

References

DEB, K. (2001): *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons, Chichester, England.

DEB, K., THIELE, L., LAUMANNS, M., and ZITZLER, E. (2005): Scalable test problems for evolutionary multiobjective optimization. In: A. Abraham, R. Jain, and R. Goldberg (Eds.): *Evolutionary Multiobjective Optimization*. Springer-Verlag, London, 105–145.

GOLDBERG, D.E. (1989): Genetic Algorithms in Search, Optimization & Machine Learning. Addison-Wesley.

HANSEN, N. and OSTERMEIER, A. (1996): Adapting arbitrary normal mutation distributions in evolutionary strategies: the covariance matrix adaptation. In: *IEEE Proceedings of the 1996 International Conference on Evolutionary Computation (ICEC '96)*. Nayoya, Japan, IEEE Service Center, 312–317.

KUKKONEN, S. and DEB, K. (2006a): Improved pruning of non-dominated solutions based on crowding distance for bi-objective optimization problems. In: *IEEE Proceedings of the 2006 Congress on Evolutionary Computation (CEC 2006)*. Vancouver, BC, Canada, IEEE Service Center, 3995–4002.

KUKKONEN, S. and DEB, K. (2006b): A fast and effective method for pruning of non-dominated solutions in many-objective problems. In: T.P. Runarsson, H.-G. Beyer, E. Burke, J.J. Merelo-Guervós, L.D. Whitley, and X. Yao (Eds.): *Proceedings of the 9th International Conference on Parallel Problem Solving from Nature (PPSN IX)*. Reykjavik, Iceland, Springer, 553–562.

KUKKONEN, S. and LAMPINEN, J. (2004a): Mechanical component design for multiple objectives using Generalized Differential Evolution. In: I.C. Parmee

(Ed.): *Proceedings of the 6th International Conference on Adaptive Computing in Design and Manufacture (ACDM 2004)*. Bristol, United Kingdom, Springer, 261–272.

KUKKONEN, S. and LAMPINEN, J. (2004b): Comparison of Generalized Differential Evolution algorithm to other multi-objective evolutionary algorithms. In: *Proceedings of the 4th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2004)*. Jyväskylä, Finland, 20 pages.

KUKKONEN, S. and LAMPINEN, J. (2004c): An extension of Generalized Differential Evolution for multi-objective optimization with constraints. In: X. Yao *et al.* (Eds.): *Proceedings of the 8th International Conference on Parallel Problem Solving from Nature (PPSN VIII)*. Birmingham, England, Springer, 752–761.

KUKKONEN, S. and LAMPINEN, J. (2005): GDE3: The third evolution step of Generalized Differential Evolution. In: *IEEE Proceedings of the 2005 Congress on Evolutionary Computation (CEC 2005)*. Edinburgh, United Kingdom, IEEE Service Center, 443–450.

KUKKONEN, S. and LAMPINEN, J. (2006): Constrained real-parameter optimization with Generalized Differential Evolution. In: *IEEE Proceedings of the 2006 Congress on Evolutionary Computation (CEC 2006)*. Vancouver, BC, Canada: IEEE Service Center, 911–918.

KUKKONEN, S. and LAMPINEN, J. (2007): Performance assessment of Generalized Differential Evolution 3 (GDE3) with a given set of problems. In: *IEEE Proceedings of the 2007 Congress on Evolutionary Computation (CEC 2007)*. Singapore: IEEE Service Center, 3593–3600.

LAMPINEN, J. (2001): *DE's selection rule for multiobjective optimization*. Technical Report, Lappeenranta University of Technology, Department of Information Technology.

LAMPINEN, J. (2002): A constraint handling approach for the Differential Evolution algorithm. In: *IEEE Proceedings of the 2002 Congress on Evolutionary Computation (CEC 2002)*. Honolulu, HI, USA, IEEE Service Center, 1468–1473.

PARETO, V. (1896): *Cours d'Economie Politique*. Libraire Droz, Geneve.

PRICE, K.V., STORN, R.M., and LAMPINEN, J.A. (2005): *Differential Evolution: A Practical Approach to Global Optimization*. Springer-Verlag, Berlin.

STORN, R. and PRICE, K.V. (1995): *Differential Evolution – A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces*. Technical Report, ICSI, University of California, Berkeley.

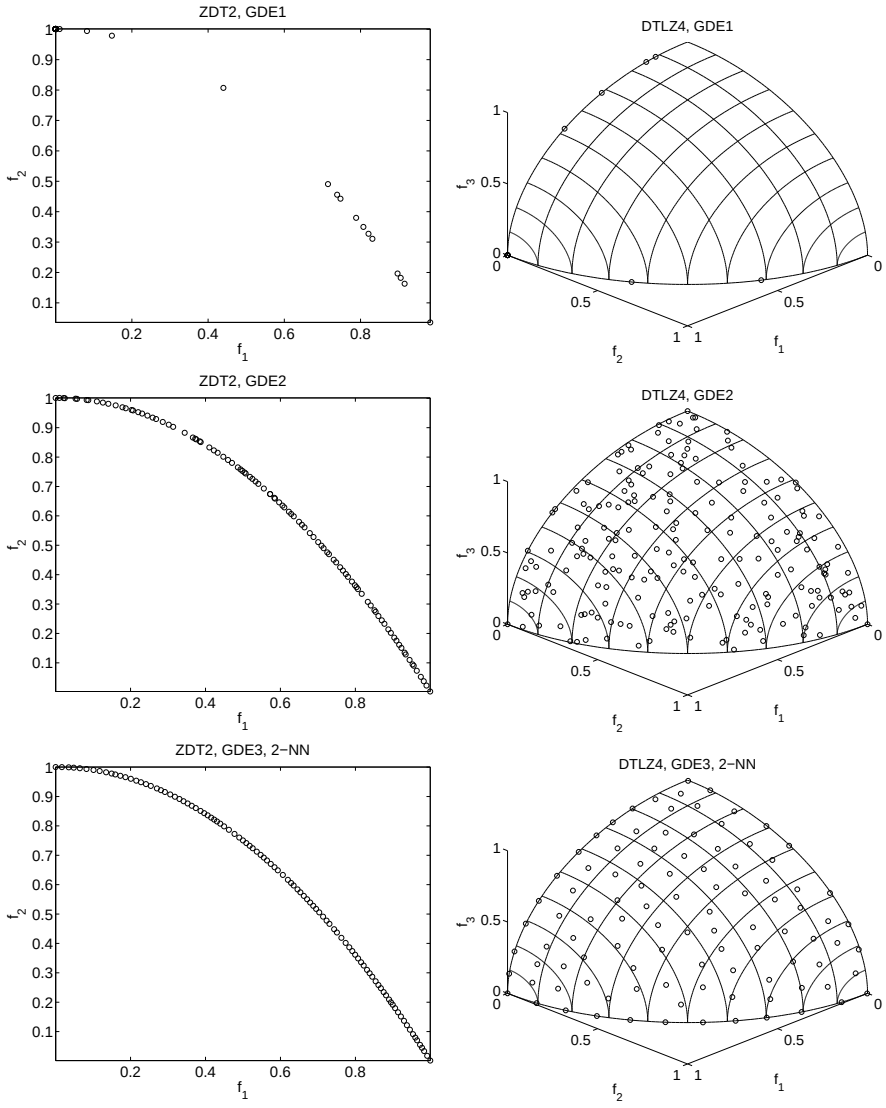

Fig. 1. The results for ZDT2 and DTLZ4 problems using GDE1, GDE2, and, GDE3 with the 2-NN diversity maintenance technique.

$$
\begin{aligned}
&\text{Input : } D, G_{max}, NP \geq 4, F \in (0, 1+], CR \in [0, 1], \text{and initial bounds: } \boldsymbol{x}^{(lo)}, \boldsymbol{x}^{(hi)} \\
&\text{Initialize : } \begin{cases} \forall i \leq NP \wedge \forall j \leq D : x_{j,i,G=0} = x_j^{(lo)} + rand_j[0,1] \cdot \left(x_j^{(hi)} - x_j^{(lo)} \right) \\ i = \{1, 2, \ldots, NP\}, j = \{1, 2, \ldots, D\}, G = 0, n = 0, rand_j[0,1] \in [0,1], \end{cases} \\
&\left\{ \begin{array}{l} \text{While } G < G_{max} \\ \quad \forall i \leq NP \left\{ \begin{array}{l} \text{Mutate and recombine:} \\ r_1, r_2, r_3 \in \{1, 2, \ldots, NP\}, \text{randomly selected,} \\ \qquad \text{except mutually different and different from } i \\ j_{rand} \in \{1, 2, \ldots, D\}, \text{randomly selected for each } i \\ \\ \forall j \leq D, u_{j,i,G} = \begin{cases} x_{j,r_3,G} + F \cdot (x_{j,r_1,G} - x_{j,r_2,G}) \\ \qquad \text{if} \quad rand_j[0,1) < CR \vee j == j_{rand} \\ x_{j,i,G} \quad \text{otherwise} \end{cases} \\ \text{Select :} \\ \boldsymbol{x}_{i,G+1} = \begin{cases} \boldsymbol{u}_{i,G} \text{ if } \quad \boldsymbol{u}_{i,G} \preceq_c \boldsymbol{x}_{i,G} \\ \boldsymbol{x}_{i,G} \text{ otherwise} \end{cases} \\ \\ \boxed{\begin{array}{l} \text{Set :} \\ \\ n = n + 1 \\ \boldsymbol{x}_{NP+n,G+1} = \boldsymbol{u}_{i,G} \end{array} \text{ if } \begin{cases} \forall k : g_k(\boldsymbol{u}_{i,G}) \leq 0 \\ \wedge \\ \boldsymbol{x}_{i,G+1} == \boldsymbol{x}_{i,G} \\ \wedge \\ \neg \left(\boldsymbol{x}_{i,G} \prec \boldsymbol{u}_{i,G} \right) \end{cases} } \end{array} \right. \\ \\ \boxed{\begin{array}{l} \text{While } n > 0 \\ \\ \quad \text{Select } \boldsymbol{x} \in \mathcal{P} = \{\boldsymbol{x}_{1,G+1}, \boldsymbol{x}_{2,G+1}, \ldots, \boldsymbol{x}_{NP+n,G+1}\} : \\ \qquad \begin{cases} \boldsymbol{x} \text{ belongs to the last non-dominated set of } \mathcal{P} \\ \wedge \\ \boldsymbol{x} \text{ is the most crowded in the last non-dominated set} \end{cases} \\ \quad \text{Remove } \boldsymbol{x} \\ \quad n = n - 1 \end{array}} \\ \\ G = G + 1 \end{array} \right. \qquad (9)
\end{aligned}
$$

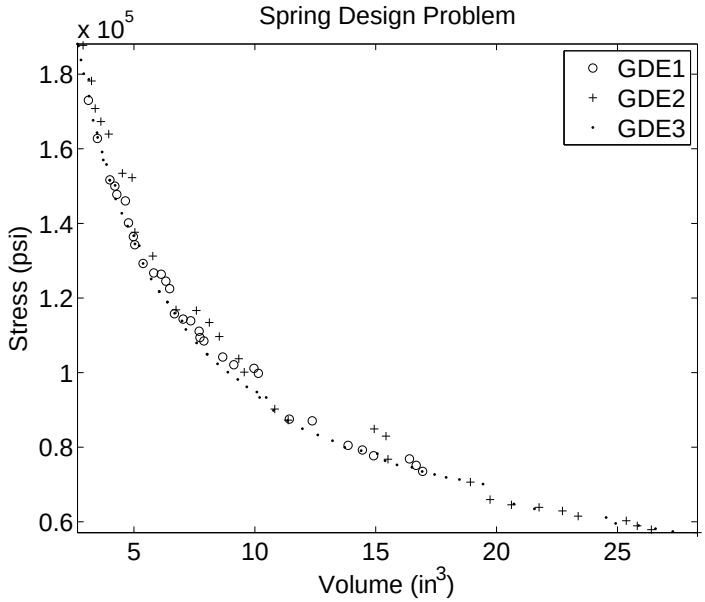

Fig. 2. The spring design problem solved with GDE1, GDE2, and GDE3.

Statistical Properties of Differential Evolution and Related Random Search Algorithms

Daniela Zaharie

Faculty of Mathematics and Computer Science, West University of Timişoara
bv. Vasile Pârvan, no. 4, 300223 Timişoara, Romania, *dzaharie@info.uvt.ro*

Abstract. The aim of this paper is to analyze the impact on the expected population mean and variance of several variants of mutation and crossover operators used in differential evolution algorithms. As a consequence of this analysis a simple variance based mutation operator which does not use differences but has the same impact on the population variance as classical differential evolution operators is proposed. A preliminary analysis of the distribution probability of the population in the case of a differential evolution algorithm for binary encoding is also presented.

Keywords: differential evolution, population variance, mutation and crossover operators, premature convergence, binary encoding

1 Introduction

The analysis of the population dynamics induced by evolutionary operators is an important issue in understanding the behavior of evolutionary algorithms and in inferring rules about choosing adequate operators and control parameters. There are two main approaches in analyzing the dynamics of an evolutionary algorithm (Okabe (2005)): a cumulants based approach which tries to describe the dynamics by using cumulants (e.g. mean, variance etc.) and a model based approach which tries to build a probability model of the population based on the properties of the operators. Most results were obtained in the case of mutation operators based on normally distributed additive perturbations (Beyer (1998)). In the case of other evolutionary operators similar studies are significantly fewer. This is also the case of Differential Evolution (DE), a successful stochastic heuristic for global optimization for which the theoretical results on the impact of operators on the population properties are still limited. DE was introduced in (Storn and Price (1995)) and is based on a particular way of constructing so-called mutant vectors by using differences between randomly selected elements from the current population. Unlike stochastic mutation, typical to evolution strategies, the DE mutation uses only information extracted from the current population. For each mutant vector, a trial vector is constructed through a crossover operation. This trial vector competes with the corresponding element of the current population and the best one, with respect to the objective function, is transferred into the next generation. In the following we shall consider objective functions,

$f : D \subset \mathbb{R}^n \to \mathbb{R}$, to be minimized thus we are dealing with minimization problems of size n. The overall structure of DE (see Algorithm 1) is typical for evolutionary algorithms, the particularities of the algorithm being related with the mutation and crossover operators. By combining different DE mutation and crossover operators various schemes have been designed. In the DE literature these schemes are specified by using the convention $DE/a/b/c$ where a denotes the manner of constructing the mutant vector, b denotes the number of differences involved in the construction of the mutant vector and c denotes the crossover type.

Population initialization: $X(0) \leftarrow \{x_1(0), \ldots, x_m(0)\}$
$g \leftarrow 0$
while the stopping condition is false **do**
  for $i = \overline{1, m}$
    $y_i \leftarrow$ generateMutant$(X(g))$
    $z_i \leftarrow$crossover$(x_i(g)$,$y_i)$
    if $f(z_i) < f(x_i(g))$ **then** $x_i(g+1) \leftarrow z_i$ **else** $x_i(g+1) \leftarrow x_i(g)$
  endfor
  $g \leftarrow g + 1$
endwhile

Fig. 1. The overall structure of a generational DE .

Previous work on analyzing DE behavior by using a model-based approach is presented in Xue et al. (2005) and in Ali and Fatti (2006). Xue et al. analyze the impact of mutation on the population distribution starting from the assumption that the population current has a normal distribution. On the other hand, Ali and Fatti derive a rather sophisticated distribution probability which corresponds to the offspring obtained by mutation starting from a population uniformly distributed in the search space. The cumulants based approach is used in Zaharie (2002) where the influence of DE mutation and binomial crossover on the expected population variance is analyzed.

The main aim of this paper is to extend the results presented in Zaharie (2002) and in Zaharie (2007) for other crossover variants and to analyze a simple variance based mutation. The next section presents the mutation and crossover operators involved in the analysis while the main theoretical results are presented in Section 3. A variance based mutation having a behavior similar to DE/rand/1/* with respect to the impact on the population variance is presented in Section 4. Section 5 presents some preliminary results on a DE for binary encoding and Section 6 concludes the paper.

2 Differential evolution operators

2.1 Mutation operators

Mutation in differential evolution algorithms has the role of constructing mutant vectors by perturbing elements of the current population. The main

particularity of DE mutation is the fact that the perturbation term is related with the difference between some randomly selected elements. Such a difference based mutation operator is more related to a recombination than to a classical mutation operator. Its main property is the fact that it acts as a self-referential mutation allowing a gradual exploration of the search space. The general form of the standard DE mutation is:

$$y_i = \lambda x_* + (1-\lambda)x_{I_i} + \sum_{l=1}^{L} F_l \cdot (x_{J_{il}} - x_{K_{il}}), \qquad i = \overline{1,m} \tag{1}$$

where x_* is the best element of the current population, $\lambda \in [0,1]$ is a coefficient which controls the influence of the best element, L is the number of differences, $F_l > 0$ is for each $l \in \{1,\ldots,L\}$ a scaling factor. I_i, J_{il} and K_{il} are random values uniformly selected from $\{1,\ldots,m\}$ and such that they are distinct. Most frequently used particular cases are when $L = 1$ and $\lambda \in \{0,1\}$. Thus for $\lambda = 0$ one obtains the DE/rand/1/* variant:

$$y_i = x_{I_i} + F \cdot (x_{J_i} - x_{K_i}), \qquad i = \overline{1,m} \tag{2}$$

and for $\lambda = 1$ one obtains the DE/best/1/* variant:

$$y_i = x_* + F \cdot (x_{J_i} - x_{K_i}), \qquad i = \overline{1,m}. \tag{3}$$

Other simple variants of these mutation operators are obtained by replacing the constant F with a random variable, ξ. Examples of such variants are when $\xi \sim N(0,F)$ (in Abbas (2001)), $\xi \sim N(F,\sigma)$ (in Ronkkonen (2003)) or even $\xi \sim U[F_{min}, F_{max}]$.

2.2 Crossover operators

In Evolutionary Algorithms the crossover operator usually combines features from different parents. In the case of DE algorithms, since the mutation operator is already based on a recombination of individuals, the role of crossover is somewhat different. It just allows the construction of an offspring by combining the current element and that generated by mutation. This can be ensured either by mixing the components (as in binomial and exponential DE crossover) or by an arithmetical recombination between the current and the mutant elements (as in the DE/current-to-rand variants). In the case of binomial crossover the components of the trial element z_i are obtained as:

$$z_i^j = \begin{cases} y_i^j & \text{if } U_j(0,1) < CR \text{ or } j = j_0 \\ x_i^j & \text{otherwise} \end{cases}, \qquad i = \overline{1,m}, j = \overline{1,n} \tag{4}$$

where $U_j(0,1)$ is a random value uniformly distributed in $[0,1]$, j_0 is a randomly selected value from $\{1,\ldots,n\}$ and $CR \in [0,1]$ is a parameter which controls the number of components taken from the mutant vector,

y_i, and is called crossover rate. The number of components taken form the mutant vector follows a binomial distribution of parameters n and $p_m = CR(1 - 1/n) + 1/n$. The value p_m can be interpreted as mutation probability as long as it specifies the probability for a component to be taken from the mutant vector.

In the exponential crossover the trial vector is constructed by taking consecutive components from the mutant:

$$z_i^j = \begin{cases} y_i^j & \text{for } j \in \{j_0, \langle j_0 + 1\rangle_n, \ldots, \langle j_0 + L - 1\rangle_n\} \\ x_i^j & \text{otherwise} \end{cases} \tag{5}$$

In eq. (5) j_0 is a randomly selected index, $\langle j\rangle_n$ denotes the remainder of the division of j by n plus 1 and L is a random variable which follows a truncated geometric distribution (Zaharie (2007)). In this case the mutation probability satisfies:

$$p_m = \frac{1 - CR^n}{n(1 - CR)} \tag{6}$$

The arithmetical recombination consists in a convex combination of the current and mutant vector. Thus

$$z_i = (1 - q)x_i + qy_i, \qquad i = \overline{1, m} \tag{7}$$

with $q \in [0, 1]$ controlling the relative weight of the mutant vector. In some implementations (see for instance Mezura et.al (2006)) the arithmetical crossover is used just as a second step in generating the mutant vector while the trial vector is obtained by mixing the components of the vector given in Eq. (7) with the current element using binomial or exponential crossover.

3 Influence of mutation and crossover on the population mean and variance

As a result of the application of evolutionary operators the population changes its distribution. The parameters of the population distribution, especially the mean and variance, can give information about the region in the search space where the population is concentrated and about its diversity. A population can be interpreted as a set of random vectors, but since all components are evolved based on the same rule the analysis can be conducted componentwise. In the following we shall analyze the impact of several DE mutation and crossover operators on the population mean and variance. Let us denote by $\{X_1, \ldots, X_m\}$ the current population and by $\{Z_1, \ldots, Z_m\}$ the population obtained after mutation and crossover. Each element of this population is a random variable $Z_i = Y_i \cdot \mathbf{1}_{M_i} + X_i \cdot \mathbf{1}_{\overline{M}_i}$, with $\mathbf{1}_{M_i}$ denoting the indicator function corresponding to the event that Z_i equals the mutant element, Y_i. Thus $\mathbf{1}_{M_i}$ is a random variable with the mean $E(\mathbf{1}_{M_i}) = p_m$.

Similarly $E(\mathbf{1}_{\overline{M}_i}) = 1 - p_m$. The difference between binomial and exponential crossover is given only by different means of $\mathbf{1}_{M_i}$ and $\mathbf{1}_{\overline{M}_i}$.

As mutation operators we shall analyze the following variants:

$$Y_i = \lambda X_* + (1-\lambda)X_{I_i} + \sum_{l=1}^{L} \xi_l \cdot (X_{J_{il}} - X_{K_{il}}) \tag{8}$$

and

$$Y_i = (1-\eta)X_i + \eta X_{I_i} + \xi \cdot (X_{J_i} - X_{K_i}) \tag{9}$$

In eq. (8) X_* denotes the best element of the current population, $\lambda \in [0,1]$ and ξ_l denote random variables independent with respect to all other variables. The most known case is when $L = 1$ and ξ is constant and equal to the scaling factor, F. If $\lambda = 0$ one obtains the DE/rand/1/* variant and when $\lambda = 1$ one have the DE/best/1/* variant.

In eq. (9), η is usually a random variable on $[0,1]$. This variant is related both to current-to-rand variants and to those which use arithmetical recombination (in the case when $\eta = q$ and $\xi = q \cdot F$). In both cases I_i, J_i and K_i are uniformly distributed on $\{1, ...m\}$ and have distinct values.

In the following we shall estimate the expected mean and variance of the population obtained by applying mutation and crossover. The expected mean, $E(\overline{Z})$, of a population $\{Z_1, \ldots, Z_m\}$ of identically distributed random variables equals $E(Z_i)$ for an arbitrary i. Thus $E(\overline{Z}) = E(Z_i) = p_m E(Y_i) + (1-p_m)E(X_i)$. Since for any random index I one have that $E(X_I) = E(\overline{X})$ it follows that in the case of eq. (8) one have that $E(\overline{Y_i}) = \lambda X_* + (1-\lambda)E(\overline{X})$, thus $E(\overline{Z}) = p_m \lambda X_* + (1 - p_m \lambda)E(\overline{X})$. Therefore when $\lambda = 0$ the expected population mean remains unchanged by mutation and crossover. When $\lambda > 0$ the population mean is biased toward the best element of the population. It is easy to check that the property of conserving the population mean is also true in the case of the mutation specified by eq. (9). The impact of selection depends on the objective function and is more difficult to be analyzed. However it is easy to see that after selection, the mean of the objective function values corresponding to the population elements decreases for all variants of mutation and crossover.

Let us turn now to the analysis of the expected population variance. Preserving the population diversity plays an important role in avoiding premature convergence and in stimulating the ability of differential evolution to follow dynamic optima. A natural measure of the diversity of a population of scalars, $X = \{X_1, \ldots, X_m\}$, is the population variance $Var(X) = \sum_{i<j}(X_i - X_j)^2/m^2$. In the case of populations of vectors the average of componentwise variances can be considered as a measure of diversity. In the following we shall analyze, in the one-dimensional case, the impact on the population variance of the mutation variants given by eqs. (8) and (9) combined with binomial, exponential and arithmetical crossover. In all cases we estimate the expected population variance, $E(Var(Z))$.

Proposition 3. *The expected population variance after mutation and crossover is:*

$$E(Var(Z)) = \left(1 + 2p_m \sum_{l=1}^{L} E(\xi_l^2) - \frac{p_m(2-p_m)}{m} - \lambda p_m^2 \frac{m-1}{m}\right) E(Var(X)) \\ +\lambda^2 p_m (1-p_m) \frac{m-1}{m} E((X_* - X_{I_i})^2) \tag{10}$$

in the case of mutation operator given by (8) and

$$E(Var(Z)) = \left(1 + 2p_m \left(E(\eta^2) - \frac{m-1}{m} E(\eta) + E(\xi^2)\right)\right. \\ \left. - \frac{p_m^2}{m}(2E(\eta) + E(\eta^2))\right) E(Var(X)) \tag{11}$$

in the case of mutation operator given by (9).

Proof. See Appendix.

In Proposition 1, p_m is given by $p_m = CR(1 - 1/n) + 1/n$ in the case of binomial crossover and by eq. (6) in the case of exponential crossover. Important particular cases of eq. (10) are when $\lambda = 0$ and $p_m = 1$. By denoting $F^2 = \sum_{l=1}^{L} E(\xi_l^2)$ we have in the first case:

$$E(Var(Z)) = \left(1 + 2p_m F^2 - \frac{p_m(2-p_m)}{m}\right) E(Var(X)) \tag{12}$$

and in the second

$$E(Var(Z)) = \left((1-\lambda)\frac{m-1}{m} + 2F^2\right) E(Var(X)). \tag{13}$$

When η is a constant q, $E(\xi^2) = F^2$ and $p_m = 1$ one obtains a simple current-to-rand version for which the eq. (11) becomes:

$$E(Var(Z)) = \left(1 + 2F^2 - 2q + \frac{2m-1}{m} q^2\right) E(Var(X)). \tag{14}$$

If F^2 is replaced with q^2F^2 then eq. (14) corresponds to the DE/rand/1 variant combined with arithmetical crossover. On the other hand, when η is uniformly distributed on $[0, 1]$ and $p_m \in [0, 1]$ then

$$E(Var(Z)) = \left(1 + 2p_m F^2 - \frac{1}{3m}(4p_m^2 + (m-3)p_m)\right) E(Var(X)). \tag{15}$$

In almost all cases (except for the case when $\lambda > 0$ and $p_m < 1$) there is a simple linear relationship between the expected variance of the population obtained by mutation and crossover and the variance of the current population: $E(Var(Z)) = c(CR, F, q, m, n)E(Var(X))$. The coefficient of this dependence involves all parameters which influence the algorithm behavior.

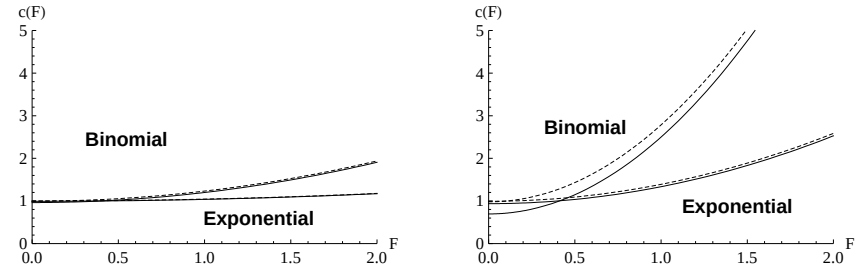

Fig. 2. Dependence between the variance factor, c, and F for DE/rand/1/* (dashed line) and DE/current-to-rand/1/* with $\eta \in U(0,1)$(continuous line). Parameters: $m = n = 50$, $CR = 0.1$ (left) and $CR = 0.9$ (right).

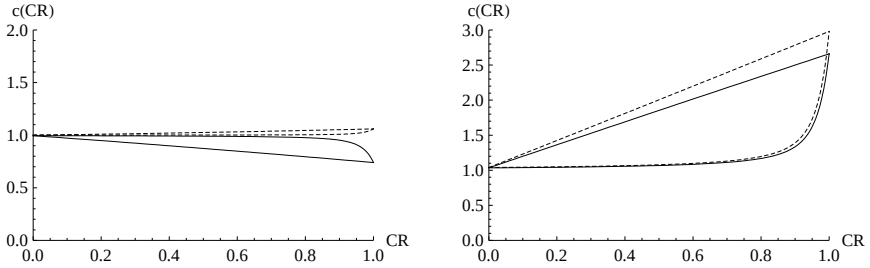

Fig. 3. Dependence between the variance factor,c, and CR for DE/rand/1/* (dashed line) and DE/current-to-rand/1/* with $\eta \in U(0,1)$ (continuous line). Binomial crossover: linear dependence, exponential crossover: nonlinear dependence. Parameters: $m = n = 50$, $F = 0.2$ (left) and $F = 1$ (right).

The advantage of such a property is the fact that one can control the impact which mutation and crossover have on the population variance by changing the values of the parameters involved in c. Figures 2,3,4 and 5 illustrate the dependence of the factor $c(CR, F, q, m, n)$ on the values of parameters and on the algorithm type. The main remarks are: *(i)* c usually increases with CR and F but in a different way in the case of binomial and exponential crossover; *(ii)* the DE/current-to-rand variant is characterized by values of c slightly smaller than DE/rand; moreover, for small values of F (e.g. $F = 0.2$) c decreases when CR increases; *(iii)* the ratio m/n does not significantly influence the factor c, meaning that using larger populations does not stimulate the population diversity; *(iv)* both in the case of DE/best and DE/current-to-rand the variance is significantly increasing with the value of F but it decreases with λ and has a non-monotonous behavior with respect to q.

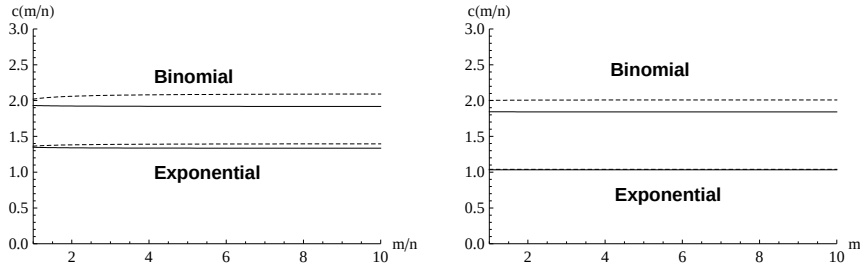

Fig. 4. Dependence of the variance factor, c, on the population size factor, s ($m = sn$) for DE/rand/1/* (dashed line) and DE/current-to-rand/1/* with $\eta \in U(0,1)$(continuous line). Parameters: $CR = 0.5$, $F = 1$, $n = 10$ (left) and $n = 100$ (right).

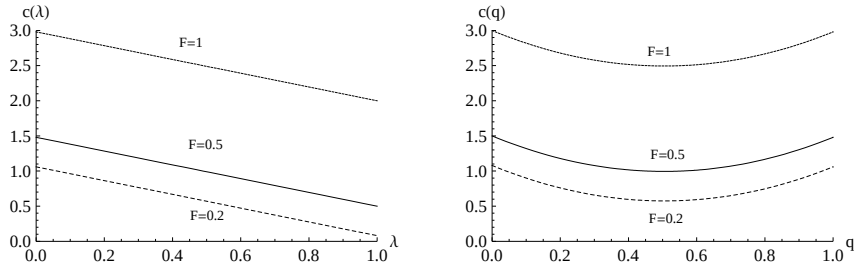

Fig. 5. Dependence of the variance factor, c, on the parameter λ in the case of DE/best-to-rand (left) and on parameter q in the case of DE/current-to-rand when $\eta = q$ (right). Parameters: $CR = 1$, $m = n = 50$.

4 A simple variance based variant

Classical mutation operators based on an additive perturbation also lead to a linear dependence but with a non zero free term, i.e. $E(Var(Z)) = aE(Var(X)) + b$. Let us consider the case when $Z_i = (X_{I_i} + \xi_i)1_{M_i} + X_i 1_{\overline{M_i}}$ with $E(M_i) = p_m$ and ξ_i independent random variables having $E(\xi_i) = 0$ and $E(\xi_i^2) = F^2$. In this case one obtains that $E(Var(Z)) = (1 + p_m^2/m - 2p_m/m)E(Var(X)) + 2p_m(m-1)/mF^2$. It is easy to see that if $E(\xi_i^2) = F^2 E(Var(X))m/(m-1)$ one obtains the same dependence between the expected variances as in the case of mutation given by eq. (8) with $\lambda = 0$ and $L = 1$. Thus using the mutation rule

$$y_i^j = x_{I_i^j} + F\sqrt{\frac{m}{m-1}Var(x^j)}N(0,1), \qquad j = \overline{1,n} \tag{16}$$

where $Var(x^j)$ is the variance of the current population corresponding to the jth component, one obtains a simple mutation rule which combined with a crossover strategy leads to the same behavior with respect to the evolution of the population variance as DE/rand/1/*. Empirical studies conducted for

classical test functions show that the mutation given by eq. (16) behaves better than a simple evolution strategy based on a normal perturbation of distribution $N(0, F)$. Table 1 presents results obtained by 30 independent runs of the algorithms on Griewank test function for $n = 100$. The maximal number of function evaluations (nfe) is set to 500000 and a run is considered successful if the best value of the population (f^*) is less than $\epsilon = 10^{-8}$. These results illustrate the fact that for some pairs of values (CR, F) the variance based mutation leads to a behavior similar to that of DE. Thus by just ensuring that the population variance have the same dynamics one can partially reproduce the behavior of DE. On the other hand, the empirical results show that the difference-based perturbation cannot just be replaced with a variance-based perturbation since for other values of parameters or for other test functions the difference-based mutation leads to better results than the variance-based one.

5 Binary differential evolution

Encouraged by the success of DE in continuous optimization several authors recently proposed variants of DE for binary encoding (Gong (2006)). A simple approach is just to use the classical operators in order to evolve trial vectors in $[0, 1]^n$ and transform their components in binary values, using a threshold function, only when the objective function is to be evaluated. The variant which we analyze uses a binary encoding and is based on the following mutation rule, inspired from (Gong (2006)):

$$Y_i^j = \begin{cases} x_{I_i}^j & \text{if } x_{J_i}^j = x_{K_i}^j \text{ or } U \geq F \\ 1 - x_{I_i}^j & \text{otherwise} \end{cases} \tag{17}$$

where $F \in [0, 1]$ and U is a random value uniformly generated in $[0, 1]$. The components of the trial vector, Z_i, are obtained by applying one of the DE crossover operators. Disregarding the type of crossover we shall denote by p_m the probability that a component in the trial vector is taken from the mutant vector. In the following we shall analyze the influence the mutation and crossover have on the distribution of a population of scalar elements.

Let (p_0, p_1) be the probability distribution of the current population (p_0 is the probability that a randomly selected element has the value 0). Thus for two randomly selected elements x_{K_i} and x_{J_i} we have that

$$Prob(x_{K_i} = x_{J_i}) = p_0^2 + p_1^2 \text{ and } Prob(x_{K_i} \neq x_{J_i}) = 2p_0p_1 \tag{18}$$

It follows that $Prob(Y_i = 0) = p_0(p_0^2 + p_1^2) + 2Fp_0p_1^2 + 2(1 - F)p_0p_1^2$ and $Prob(Y_i = 1) = p_1(p_0^2 + p_1^2) + 2Fp_0^2p_1 + 2(1 - F)p_0^2p_1$. Consequently, the probabilities corresponding to the trial element Z_i are $Prob(Z_i = 0) = p_0(1 + 2p_mFp_1(p_1 - p_0))$ and $Prob(Z_i = 1) = p_1(1 + 2p_mFp_0(p_0 - p_1))$. On the other hand, in the case of a simple binary mutation ($Z_i = 1 - X_i$ with probability

Table 1. Comparative results of DE/rand/1/bin, variance based mutation (var/bin) and normal mutation (norm/bin) combined with binomial crossover. Test function: Griewank. Parameters: $m = n = 100$.

CR	F	DE/rand/1/bin		var/bin		norm/bin	
		$\langle f^* \rangle$	Success	$\langle f^* \rangle$	Success	$\langle f^* \rangle$	Success
		$stdev(f^*)$	$\langle nfe \rangle$	$stdev(f^*)$	$\langle nfe \rangle$	$stdev(f^*)$	$\langle nfe \rangle$
0.1	0.5	$9 \cdot 10^{-9}$	30/30	$9 \cdot 10^{-9}$	30/30	0.3304	0/30
		$\pm 10^{-10}$	(380416)	$\pm 10^{-10}$	(190290)	± 0.3134	(500000)
0.5	0.5	10^{-4}	0/30	$9 \cdot 10^{-9}$	30/30	0.2890	0/30
		$\pm 10^{-5}$	(500000)	$\pm 10^{-10}$	(204703)	± 0.087	(500000)
0.9	0.5	0.0078	18/30	$1.27 \cdot 10^{-8}$	27/30	0.5523	0/30
		± 0.0125	(306933)	$\pm 10^{-8}$	(470792)	± 0.039	(500000)
0.1	0.2	$9 \cdot 10^{-9}$	30/30	0.0158	24/30	0.6352	0/30
		$\pm 2 \cdot 10^{-10}$	(137090)	0.0318	(131887)	± 0.365	(500000)
0.5	0.2	0.0959	18/30	1.3469	0/30	0.4322	0/30
		± 0.1657	(87666)	1.5373	(500000)	± 0.3319	(500000)

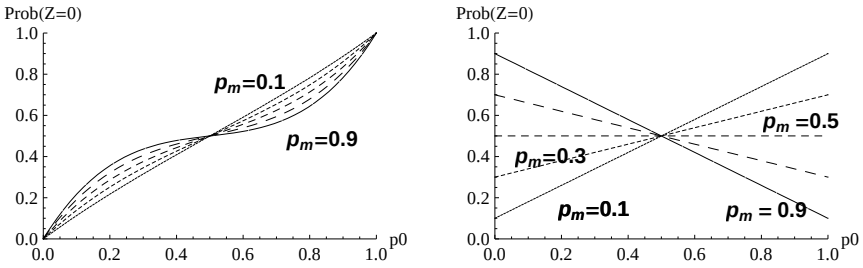

Fig. 6. Dependence of $Prob(Z_i = 0)$ on p_0 and p_m in the case of binary DE mutation (left) and classical binary mutation (right).

p_m) the corresponding probabilities are $Prob(Z_i = 0) = p_0 + p_m(p_1 - p_0)$ and $Prob(Z_i = 1) = p_1 + p_m(p_0 - p_1)$. The different impact on the population distribution of the DE binary mutation and of the classical binary mutation is illustrated in Figure 6. Unlike the classical binary mutation, the DE binary mutation leads to small changes in the population distribution for all values of p_m. On the other hand one have to remark that $Prob(Z_i = 0) - Prob(Z_i = 1) = (1-2p_m)(p_0-p_1)$ in the case of classical binary mutation and $Prob(Z_i = 0) - Prob(Z_i = 1) = (1-2p_mFp_0p_1)(p_0-p_1)$ in the case of DE mutation. Both variants tend to decrease the difference between p_0 and p_1 but the decrease is smaller in the case of DE variant.

6 Conclusions

Almost for all DE variation operators the expected population variance after mutation and crossover is related to the current population variance by a

simple linear dependence based on a coefficient $c(CR, F, q, m, n)$ which involves all parameters characterizing the algorithm. This allows us to control the evolution of the population diversity by just changing the algorithm's parameters. Significant differences have been identified between the behavior of binomial, exponential and arithmetical crossover. A simple mutation rule which does not involve differences but just the estimation of the current variance was proposed. It has the same behavior as DE/rand/1/* with respect to the population variance evolution. Numerical experiments show that for some sets of parameters the variance based mutation combined with binomial crossover behaves better than DE/rand/1/bin but for other ones worse. This suggests on one hand that the evolution of the population variance has a significant influence on the behavior of the algorithm and on the other hand that the difference based mutation induces a dynamics which cannot be entirely mimicked by using the population variance estimation instead. The analysis of the influence of the DE binary mutation on the population distribution shows that it leads to a dynamics different than that induced by classical binary mutation. However further analysis is needed to assess its effectiveness for real problems.

Appendix. *Proof of Proposition 1.* Since $E(Var(X)) = \frac{m-1}{2m} E((X_i - X_j)^2)$ for any pair of distinct indices (i, j) it follows that it is enough to find the relationship between $E((Z_i - Z_j)^2)$ and $E((X_i - X_j)^2)$ for an arbitrary pair (i, j) of distinct values. Based on the fact that $Z_i = Y_i \mathbf{1}_{M_i} + X_i \mathbf{1}_{\overline{M}_i}$ it follows:

$$\begin{aligned} E((Z_i - Z_j)^2) = p_m^2 E((Y_i - Y_j)^2) + 2p_m(1 - p_m)E((Y_i - X_j)^2) \\ +(1 - p_m)^2 E((X_i - X_j)^2) \end{aligned} \tag{19}$$

If I and J are two random indices from $\{1, ..., m\}$ then $E((X_I - X_J)^2) = 2E((Var(X))$ if I and J can be identical and $E((X_I - X_J)^2) = \frac{2m}{m-1} E((Var(X))$ if I and J take distinct values. Using these relations one can compute $E((Y_i - Y_j)^2)$ and $E((Y_i - X_j)^2)$ when Y_i is given by eqs. (8) and (9).

By taking into account that I_i and I_j can be identical but K_{il} and J_{il}, K_{jl} and J_{jl} are respectively distinct one obtains, in the case of eq.(8) that

$$E((Y_i - Y_j)^2) = \frac{2m}{m-1}((1 - \lambda)^2 + 2\sum_{l=1}^{L} E(\xi_l^2))E(Var(X))$$

and

$$E((Y_i - X_j)^2) = 2(1 + \frac{m}{m-1}\sum_{l=1}^{L} \xi_l^2)E(Var(x)) + \lambda^2 E((X_* - X_{I_i})^2)$$

By replacing these into eq. (19) one obtains eq. (10). In the case of eq.(9) one have that:

$$E((Y_i - Y_j)^2) = \frac{2m}{m-1}\left(E((1-\eta)^2) + 2E(\xi^2) + \frac{m-1}{m}E(\eta^2)\right)E(Var(X))$$

and

$$E((Y_i - X_j)^2) = \frac{2m}{m-1}\left(1 + E(\eta^2) + E(\xi^2) - \frac{m-1}{m}E(\eta)\right)E(Var(X))$$

By replacing these into eq. (19) one obtains eq. (11).

Acknowledgments. Thanks are due to Olivier François for valuable suggestions on the analysis of the variance-based variant. This work is supported by Romanian project 11-028/14.09.2007.

References

ABBAS, H.A., SARKER, R. and NEWTON, C. (2001): A Pareto Differential Evolution Approach to the Vector Optimisation Problems. In: *Proceedings of CEC 2001, IEEE Computer Press, 971-978.*

ALI, M.M. and FATTI, L.P. (2006): A Differential Free Point Generation Scheme in the Differential Evolution Algorithm. *Journal of Global Optimization 35, 551-572.*

BEYER, H.G. (1998): On the Dynamics of EAs without Selection. In: W. Banzhaf and C.R. Reeves, *Proceedings of the Fifth Workshop on Foundations of Genetic Algorithms - FOGA 1998, 5-26*

GONG, T. and TUSON, A. (2007): Differential Evolution for Binary Encoding. In: Soft Computing in Industrial Applications, (Springer-Verlag, 2007) vol. 39, 251-262.

MEZURA-MONTES, E., VELASQUEZ-REYES, J. and COELLO COELLO, C.A.(2006): A Comparative Study of Differential Evolution Variants for Global Optimization. In: Maarten Keijzer et al. (Eds.) *Genetic and Evolutionary Computation Conference (GECCO'2006), Vol. 1, (ACM Press,2006) 485-492.*

OKABE,T., JIN, Y. and SENDHOFF,B. (2005): Theoretical Comparisons of Search Dynamics of Genetic Algorithms and Evolution Strategies. In: *Proceedings of CEC 2005, 382–389.*

RONKKONEN, J., LAMPINEN, J. (2003): On using normally distributed mutation steplength for the differential evolution algorithm. In: R. Matoušek and P. Ošmera (Eds.): *Proc. of Mendel 2003, 9th International Conference on Soft Computing*, 11–18.

STORN, R. and PRICE, K. (1995): Differential Evolution a simple and efficient adaptive scheme for global optimization over continuous spaces. International Computer Science Institute, Berkeley, TR- 95-012.

XUE, F., SANDERSON, A.C. and GRAVES, R.J. (2005): Multi-objective differential evolution - algorithm, convergence analysis and applications. In: *Proceedings of CEC 2005, IEEE Computer Press, 743-750.*

ZAHARIE, D. (2002): Critical values for the control parameters of differential evolution algorithms. In: R. Matoušek and P. Ošmera, (Eds.), *Proceedings of 8th International Conference on Soft Computing, Mendel 2002, 62-67.*

ZAHARIE, D. (2007): A Comparative Analysis of Crossover Variants in Differential Evolution. In: *Proceedings of the IMCSIT, 2nd International Symposium Advances in Artificial Intelligence and Applications, 2007, 171-181.*

Part XIV

Robust Statistics

Robust Estimation of the Vector Autoregressive Model by a Least Trimmed Squares Procedure

Christophe Croux and Kristel Joossens

Faculty of Business and Economics , Katholieke Universiteit Leuven
Naamsestraat 69, B-3000 Leuven, Belgium, *christophe.croux@econ.kuleuven.be*

Abstract. The vector autoregressive model is very popular for modeling multiple time series. Estimation of its parameters is typically done by a least squares procedure. However, this estimation method is unreliable when outliers are present in the data, and therefore we propose to estimate the vector autoregressive model by using a multivariate least trimmed squares estimator. We also show how the order of the autoregressive model can be determined in a robust way. The robust procedure is illustrated on a real data set.

Keywords: robustness, multivariate time series, outliers, trimming, vector autoregressive models

1 Introduction

The use of autoregressive models for predicting and modelling univariate time series is standard and well known. In many applications, one does not observe a single time series, but several series, possibly interacting with each other. For these multiple time series the vector autoregressive model became very popular, and is described in standard textbooks on time series (e.g. Brockwell and Davis 2003, Chapter 7). In this paper we propose a robust procedure to estimate vector autoregressive models and to select their order.

Let $\{y_t \mid t \in \mathbb{Z}\}$ be a p-dimensional stationary time series. The vector autoregressive model of order k, denoted by VAR(k), is given by

$$y_t = \mathcal{B}_0' + \mathcal{B}_1' y_{t-1} + \ldots + \mathcal{B}_k' y_{t-k} + \varepsilon_t, \tag{1}$$

with y_t a p-dimensional vector, the intercept parameter $\mathcal{B}_0'$ a vector in $\mathbb{R}^p$ and the slope parameters $\mathcal{B}_1, \ldots, \mathcal{B}_k$ being matrices in $\mathbb{R}^{p \times p}$. Throughout the paper M' will stand for the transpose of a matrix M. The p-dimensional error terms ε_t are supposed to be independently and identically distributed with a density of the form

$$f_{\varepsilon_t}(u) = \frac{g(u'\Sigma^{-1}u)}{(\det\Sigma)^{1/2}}, \tag{2}$$

with Σ a positive definite matrix, called the *scatter matrix* and g a positive function. If the second moment of ε_t exists, Σ will be (proportional

to) the covariance matrix of the error terms. Existence of a second moment, however, will not be required for the robust estimator. We focus on the unrestricted VAR(k) model, where no restrictions are put on the parameters $\mathcal{B}_0, \mathcal{B}_1, \ldots, \mathcal{B}_k$.

Suppose that the multivariate time series y_t is observed for $t = 1, \ldots, T$. The vector autoregressive model (1) can be rewritten as a multivariate regression model

$$y_t = \mathcal{B}' x_t + \varepsilon_t, \tag{3}$$

for $t = k+1, \ldots, T$ and with $x_t = (1, y'_{t-1}, \ldots, y'_{t-k})' \in \mathbb{R}^q$, where $q = pk+1$. The matrix $\mathcal{B} = (\mathcal{B}'_0, \mathcal{B}'_1, \ldots, \mathcal{B}'_k)' \in \mathbb{R}^{q \times p}$ contains all unknown regression coefficients. In the language of regression, $X = (x_{k+1}, \ldots, x_T)' \in \mathbb{R}^{n \times q}$ is the matrix containing the values of the explanatory variables and $Y = (y_{k+1}, \ldots, y_T)' \in \mathbb{R}^{n \times p}$ the matrix of responses, where $n = T - k$. The classical least squares estimator for the regression parameter $\mathcal{B}$ in (3) is given by the well known formula

$$\hat{\mathcal{B}}_{\mathrm{OLS}} = (X'X)^{-1} X'Y,$$

and the scatter matrix Σ is estimated by

$$\hat{\Sigma}_{\mathrm{OLS}} = \frac{1}{n-p} (Y - X\hat{\mathcal{B}}_{\mathrm{OLS}})'(Y - X\hat{\mathcal{B}}_{\mathrm{OLS}}). \tag{4}$$

In applied time series research, one is aware of the fact that outliers can seriously affect parameter estimates, model specification and forecasts based on the selected model. Outliers in time series can be of different nature, the most well known types being additive outliers and innovational outliers. With respect to the autoregressive model (1), an observation y_t is an additive outlier if only its own value has been affected by contamination. On the other hand, an outlier is said to be innovational if the error term ε_t in (1) is contaminated. Innovational outliers will therefore have an effect on the next observations as well, due to the dynamic structure in the series. Additive outliers have an isolated effect on the time series, but they still may seriously affect the parameter estimates.

Several procedures to detect different types of outliers for univariate time series have been proposed. For a detailed treatment of robust univariate time series analysis we refer to Maronna, Martin and Yohai (2006, Chapter 8). While most previous studies focus on a single series, this paper deals with robust analysis of multivariate time series.

A common practice for handling outliers in a multivariate process is to first apply univariate techniques to the component series in order to remove the outliers, followed by treating the adjusted series as outlier-free and model them jointly. But this procedure encounters several difficulties. First, in a multivariate process, contamination in one component may be caused by

an outlier in the other components. Secondly, a multivariate outlier cannot always be detected by looking at the component series separately, since it can be an outlier for the correlation structure only. Therefore it is better to cope with outliers in a multivariate framework. Tsay, Peña and Pankratz (2000) discuss the problem of multivariate outliers in detail.

The aim of this paper is to propose a robust estimation procedure for the vector autoregressive model, the most popular model for multiple time series analysis. Not much work has been done for the robust estimation of multivariate time series. Franses, Kloek and Lucas (1999) used Generalized M-estimators, which are known to have low robustness in higher dimensions. Another approach was taken by García Ben, Martínez and Yohai (1999), using so-called *Residual Autocovariance* (RA)-estimators, being an affine equivariant version of the estimators of Li and Hui (1989). García Ben et al. (1999) showed, by means of a simulation study, that the RA-estimators are resistant to outliers. Using an appropriate starting value, the RA-estimators are iteratively computed as solutions of certain estimating equations.

Our proposal for obtaining a resistant estimator for the VAR model is to replace the multivariate least squares estimator for (3) by a highly robust estimator. We will use the *Multivariate Least Trimmed Squares* (MLTS) estimator, discussed in Agulló, Croux and Van Aelst (2008). This estimator is defined by minimizing a trimmed sum of squared Mahalanobis distances, and can be computed by a fast algorithm. The procedure also provides a natural estimator for the scatter matrix of the residuals, which can then be used for model selection criteria. This estimator is reviewed in Section 2. The robustness of the estimator is studied by means of several simulation experiments in Section 3, where a comparison with the RA-estimators is also made. In Section 4 it is explained how to select the autoregressive order of the model in a robust way. The robust VAR methodology is applied on real data sets in Section 5, while Section 6 concludes.

2 The multivariate least trimmed squares estimator

The unknown parameters of the VAR(k) will be estimated via the multivariate regression model (3). For this the Multivariate Least Trimmed Squares estimator (MLTS), based on the idea of the Minimum Covariance Determinant estimator (Rousseeuw and Van Driessen 1999), is used. The MLTS selects the subset of h observations having the property that the determinant of the covariance matrix of its residuals from a least squares fit, solely based on this subset, is minimal.

Consider the data set $Z = \{(x_t, y_t),\ t = k+1, \ldots, T\} \subset \mathbb{R}^{p+q}$. Let $\mathcal{H} = \{H \subset \{k+1, \ldots, T\} \mid \#H = h\}$ be the collection of all subsets of size h. For any subset $H \in \mathcal{H}$, let $\hat{\mathcal{B}}_{\mathrm{OLS}}(H)$ be the classical least squares fit based on the observations of the subset:

$$\hat{\mathcal{B}}_{\mathrm{OLS}}(H) = (X_H' X_H)^{-1} X_H' Y_H,$$

where X_H and Y_H are submatrices of X and Y, consisting of the rows of X, respectively Y, having an index in H. The corresponding scatter matrix estimator computed from this subset is then

$$\hat{\Sigma}_{\text{OLS}}(H) = \frac{1}{h-p}(Y_H - X_H\hat{\mathcal{B}}_{\text{OLS}}(H))'(Y_H - X_H\hat{\mathcal{B}}_{\text{OLS}}(H)).$$

The MLTS estimator is now defined as

$$\hat{\mathcal{B}}_{\text{MLTS}}(Z) = \hat{\mathcal{B}}_{\text{OLS}}(\hat{H}) \quad \text{where} \quad \hat{H} = \underset{H\in\mathcal{H}}{\text{argmin}} \det \hat{\Sigma}_{\text{OLS}}(H), \tag{5}$$

and the associated estimator of the scatter matrix of the error terms is given by

$$\hat{\Sigma}_{\text{MLTS}}(H) = c_\alpha \hat{\Sigma}_{\text{OLS}}(\hat{H}). \tag{6}$$

In definition (6), c_α is a correction factor to obtain consistent estimation of Σ at the model distribution (2) of the error terms, and α the trimming proportion for the MLTS estimator, i.e. $\alpha \approx 1 - h/n$. In the case of multivariate normal error terms it has been shown (e.g. Croux and Haesbroeck 1999) that $c_\alpha = (1-\alpha)/F_{\chi^2_{p+2}}(q_\alpha)$. Here $F_{\chi^2_q}$ is the cumulative distribution function of a χ^2 distribution with q degrees of freedom, and $q_\alpha = \chi^2_{q,1-\alpha}$ is the upper α-quantile of this distribution.

Equivalent characterizations of the MLTS estimator are given by Agulló, Croux and Van Aelst (2008). They prove that any $\tilde{\mathcal{B}} \in \mathbb{R}^{p\times q}$ minimizing the sum of the h smallest squared Mahalanobis distances of its residuals (subject to $\det \Sigma = 1$) is a solution of (5). In mathematical terms,

$$\hat{\mathcal{B}}_{\text{MLTS}} = \underset{\mathcal{B},\ \Sigma;\ |\Sigma|=1}{\text{argmin}} \sum_{s=1}^{h} d^2_{s:n}(\mathcal{B}, \Sigma).$$

Here $d_{1:n}(\mathcal{B}, \Sigma) \leq \ldots \leq d_{n:n}(\mathcal{B}, \Sigma)$ is the ordered sequence of the residual Mahalanobis distances

$$d_s(\mathcal{B}, \Sigma) = \left((y_t - \mathcal{B}'x_t)'\Sigma^{-1}(y_t - \mathcal{B}'x_t)\right)^{1/2}, \tag{7}$$

for $\mathcal{B} \in \mathbb{R}^{p\times q}$. We see that the MLTS-estimator minimizes the sum of the h smallest squared distances of its residuals, and is therefore the multivariate extension of the Least Trimmed Squares (LTS) estimator of Rousseeuw (1984).

Since the efficiency of the MLTS estimator is rather low, the reweighted version is used in this paper, to improve the performance of MLTS. The Reweighted Multivariate Least Trimmed Squares (RMLTS) estimates are defined as

$$\hat{\mathcal{B}}_{\text{RMLTS}} = \hat{\mathcal{B}}_{\text{OLS}}(J) \quad \text{and} \quad \hat{\Sigma}_{\text{RMLTS}} = c_\delta \hat{\Sigma}_{\text{OLS}}(J), \tag{8}$$

where $J = \{j \in \{1,\ldots,n\} \mid d^2_j(\hat{\mathcal{B}}_{\text{MLTS}}, \hat{\Sigma}_{\text{MLTS}}) \leq q_\delta\}$ and $q_\delta = \chi^2_{q,1-\delta}$. The idea is that outliers have large residuals with respect to the initial robust MLTS estimator, resulting in a large residual Mahalanobis distance

$d_j^2(\hat{\mathcal{B}}_{\text{MLTS}}, \hat{\Sigma}_{\text{MLTS}})$. If the latter is above the critical value q_δ, then the observation is flagged as an outlier. The final RMLTS is then based on those observations not having been detected as outliers. In this paper, we set $\delta = 0.01$ and take as trimming proportion for the initial MLTS estimator $\alpha = 25\%$.

3 Simulation experiments

In order to study the robustness of the estimators, we perform a simulation study comparing the OLS estimator with the robust RMLTS and the RA estimators. As in García Ben et al. (1999), RA estimators are computed as iteratively reweighted maximum likelihood estimates, with a Tukey Biweight weight function (tuned to have a 95% relative asymptotic efficiency for Gaussian innovations). Since this weight function is redescending, it is important to use a robust starting value to ensure convergence to the "right" solution. In our implementation, the RMLTS was used as starting value.

We generate bivariate time series according to the VAR(2) model

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} .10 \\ .02 \end{pmatrix} + \begin{pmatrix} .40 & .03 \\ .04 & .20 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} .100 & .005 \\ .010 & .080 \end{pmatrix} \begin{pmatrix} y_{1,t-2} \\ y_{2,t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}, \quad (9)$$

where $\varepsilon_t \sim N_2(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & .2 \\ .2 & 1 \end{pmatrix}. \quad (10)$$

The aim is to look at the effect of the outliers on the parameter estimates. There are 10 regression parameters to be estimated, and to summarize the performance of the estimators, we calculate the total Bias and total Mean Squared Error (MSE). The former is computed as

$$\text{Bias} = \sqrt{\sum_{i=1}^{q}\sum_{j=1}^{p}(\frac{1}{\text{nsim}}\sum_{s=1}^{\text{nsim}}\hat{\mathcal{B}}_{ij}^s - \mathcal{B}_{ij})^2} \approx \|E[\hat{\mathcal{B}} - \mathcal{B}]\|,$$

where $\hat{\mathcal{B}}^s$, for $s = 1, ..., \text{nsim}$, is the estimate obtained from the s-th generated series, $\mathcal{B}$ is the true parameter value and nsim= 1000 the number of simulations. The MSE is given by

$$\text{MSE} = \sum_{i=1}^{q}\sum_{j=1}^{p}[\frac{1}{\text{nsim}}\sum_{s=1}^{\text{nsim}}(\hat{\mathcal{B}}_{ij}^s - \mathcal{B}_{ij})^2].$$

After generating series of length T =500, according to model (9), m outliers will be introduced. The classical and robust estimators are used to estimate this VAR(2) model for the uncontaminated series ($m = 0$), and for the contaminated ones ($m > 0$), where several types of outliers are considered. Below

we look at the effect of additive, innovational, and correlation outliers on the different estimators. Note that other types of contamination do exist, like level shifts and patches of outliers.

Additive outliers are introduced by randomly selecting m bivariate observations, and contaminating them by adding the value 10 to all the components of the selected observations. We consider different contamination levels, ranging from one single outlier up to 5% of additive outliers, i.e. $m = 25$. The Bias and MSE for the OLS, RA and RMLTS estimator are given in Table 1, as a function of the number m of additive outliers.

Both Bias and MSE grow with the number of outliers, the increase being much faster for the non robust OLS. Using the robust estimators instead of OLS leads to a very small loss in efficiency when no outliers are present. When even only one outlier is present, the RA and RMLTS are already more efficient, and this decrease in MSE becomes very substantial for larger amounts of outliers. Comparing the robust procedures, RMLTS performs slightly better as RA in this simulation setting.

Table 1. Simulated Bias and Mean Squared Error for the OLS, and the robust RA and RMLTS estimator of a bivariate VAR(2) model, in presence of m additive outliers in a series of length 500.

	OLS		RA		RMLTS	
m	Bias	MSE	Bias	MSE	Bias	MSE
0	0.00	0.020	0.00	0.022	0.00	0.022
1	0.08	0.030	0.02	0.023	0.02	0.023
2	0.14	0.045	0.03	0.025	0.03	0.024
3	0.18	0.063	0.05	0.028	0.04	0.026
4	0.22	0.079	0.06	0.031	0.04	0.027
5	0.25	0.096	0.07	0.035	0.05	0.029
10	0.38	0.193	0.14	0.061	0.07	0.039
15	0.51	0.319	0.21	0.086	0.11	0.057
20	0.64	0.478	0.25	0.101	0.17	0.080
25	0.76	0.659	0.29	0.115	0.25	0.104

Innovational outliers are generated by first randomly selecting m innovation terms ε_t in (9). Then add the value 10 to the first component of the innovations, yielding the contaminated innovations series ε_t^C. Bivariate series are then simulated according to (9), but with ε_t replaced by ε_t^C. The Bias and MSE when estimating the uncontaminated ($m = 0$) and contaminated series are given in Table 2, for the classical as well as the robust estimation procedures.

The Bias and MSE for OLS grow for an increasing number of outliers, although at a smaller rate than for contamination with additive outliers. For

Table 2. Simulated Bias and Mean Squared error for the OLS, and the robust RA and RMLTS estimator of a bivariate VAR(2) model, in presence of m innovational outliers in a series of length 500.

	OLS		RA		RMLTS	
m	Bias	MSE	Bias	MSE	Bias	MSE
0	0.00	0.021	0.00	0.022	0.00	0.022
1	0.02	0.022	0.00	0.021	0.00	0.021
2	0.04	0.023	0.01	0.020	0.01	0.020
3	0.06	0.025	0.01	0.019	0.01	0.019
4	0.08	0.029	0.01	0.018	0.01	0.018
5	0.10	0.033	0.01	0.018	0.01	0.018
10	0.20	0.068	0.01	0.017	0.01	0.017
15	0.30	0.123	0.01	0.016	0.01	0.016
20	0.40	0.198	0.01	0.016	0.01	0.016
25	0.49	0.289	0.01	0.017	0.01	0.016

the robust estimator we see a small decrease of the MSE, implying that the robust procedure is precise in presence than in absence of innovational outliers! This is due to the fact that an innovational outlier in the time series results in a single vertical outlier, but also in several good leverage points when estimating the autoregressive model. The robust method can cope with the vertical outlier and takes profit of the good leverage points to decrease the MSE. The OLS estimator gets biased due to the vertical outliers, but the presence of the good leverage points explains why the effect of innovational outliers is less strong than for additive outliers. Finally, note that the difference between the two robust approaches is not significant here, showing again that RMLTS and RA perform very similarly. Hence, the RA method does neither improves, neither deteriorates the initial RMLTS estimate.

Correlation outliers are generated as innovational outliers, but instead of (10), we take

$$\Sigma = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix} \tag{11}$$

and place the innovation outliers all at the same position $(2, -2)'$. By placing the outliers in this way, they are only outlying for the correlation structure, and not with respect to the marginal distributions of the innovations. This type of outliers strongly influences results of a (robust) univariate analysis. To illustrate this, we will estimate the VAR model (9) equation by equation, applying twice a univariate reweighted least trimmed squares estimator (RLTS) instead of the RMLTS. Bias and MSE when estimating the uncontaminated and contaminated series by OLS, the univariate RLTS and the multivariate RMLTS, are given in Table 3.

Table 3. Simulated Bias and Mean Squared error for the OLS, robust univariate (RLTS) and multivariate (RMLTS) estimators of a bivariate VAR(2) model in presence of m correlation outliers in a series of length 500.

	OLS		RLTS		RMLTS	
m	Bias	MSE	Bias	MSE	Bias	MSE
0	0.01	0.084	0.01	0.098	0.01	0.093
1	0.01	0.074	0.01	0.088	0.01	0.083
2	0.02	0.069	0.02	0.083	0.01	0.076
3	0.02	0.056	0.02	0.074	0.01	0.069
4	0.02	0.054	0.03	0.067	0.01	0.062
5	0.03	0.046	0.03	0.065	0.01	0.059
10	0.06	0.046	0.06	0.054	0.01	0.044
15	0.08	0.043	0.08	0.049	0.01	0.037
20	0.11	0.044	0.11	0.048	0.01	0.032
25	0.14	0.049	0.14	0.053	0.01	0.030

When no outliers are present, there is hardly any difference between the different estimation procedures: the robust procedures show only a marginal loss in MSE. From Table 3 one can see that the univariate RLTS yields a comparable Bias as for OLS, growing for an increasing number of correlation outliers. On the other hand, the multivariate RMLTS approach offers protection against the correlation outliers, remaining almost without bias. As for the previous simulation scheme, the MSE tends to decrease with the number of outliers (because the latter introduce good leverage points). We conclude from this simulation experiment that a fully multivariate robust approach is necessary when estimating a VAR model.

4 Determining the autoregressive order

To select the order k of a vector autoregressive model, information criteria are computed for several values of k and an optimal order is selected by minimizing the criterion. Most information criteria are in terms of the value of the log likelihood l_k of the VAR(k) model. Using the model assumption (2) for the distribution of the error terms, we get

$$l_k = \sum_{t=k+1}^{T} g(\varepsilon_t' \Sigma^{-1} \varepsilon_t) - \frac{n}{2} \log\det \Sigma,$$

with $n = T - k$. When error terms are multivariate normal the above leads to

$$l_k = -\frac{n}{2} \log\det \Sigma - \frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{t=k+1}^{T} \varepsilon_t' \Sigma^{-1} \varepsilon_t. \tag{12}$$

The log likelihood will depend on the autoregressive order via the estimate of the covariance matrix of the residuals. For the ordinary least squares estimator we have

$$\hat{\Sigma}_{\mathrm{OLS}} = \frac{1}{n-p} \sum_{t=k+1}^{T} \hat{\varepsilon}_t(k)\hat{\varepsilon}_t'(k),$$

where the $\hat{\varepsilon}_t(k)$ are the residuals corresponding with the estimated VAR(k) model. Using trace properties, the last term in (12) equals the constant $-(n-p)p/2$ for the OLS estimator. To prevent that outliers might affect the optimal selection of the information criteria, we estimate Σ by the RMLTS estimator:

$$\hat{\Sigma}_{\mathrm{RMLTS}} = \frac{c_\delta}{m(k)-p} \sum_{t \in J(k)} \hat{\varepsilon}_t(k)\hat{\varepsilon}_t'(k),$$

with $J(k)$ as in (8) and $m(k)$ the number of elements in $J(k)$. The last term in (12) equals now $-(m(k)-p)p/(2c_\delta)$.

The most popular information criteria to select the order of the autoregressive model are of the form

$$\frac{-2}{n} l_k + h(n)\frac{(kp+1)p}{n},$$

where $(kp+1)p$ is the number of unknown parameters, which penalizes for model complexity, and where $h(n)$ can take different forms. We will consider the following three criteria: the popular Akaike information criterion, corresponding to $h(n) = 2$, the Hannan-Quinn criterion, corresponding to $h(n) = 2\log(\log(n))$ and the Schwarz criterion, also called the Bayesian Information Criterium), for which $h(n) = \log(n)$.

5 Example

As an example, we consider the bivariate time series of *maturity rates* (Tsay 2002, p. 324–325). The first series "GS1" is the 1-year Treasury constant maturity rate, and the second series "N3" is the 3-year Treasury constant maturity rate. The data are monthly and sampled from April 1953 to January 2001. As in the book of Tsay (2002), we work with the log-transformed version of both series. We consider the series as stationary. From the plot of the series (Figure 1), it can be seen that there might be some outliers around the years 1954 and 1958.

In Table 4 different lag length criteria, as discussed in Section 4, are presented, once based on the OLS estimator, and once based on the RMLTS. The information criteria clearly depend on the chosen estimator. For example, when using the AIC the classical method suggests a VAR(8) model while the robust indicates a VAR(6) model. On the other hand the Schwarz criterion selects an optimal order 3 for both estimators. Since it is well known that the

Fig. 1. Time plot of the "maturity rate" series. The solid line represents the 1-Year Treasury constant maturity rate and the dashed line the 3-Year Treasury constant maturity rate, both in logs.

Table 4. Lag length criteria using the OLS and RMLTS estimator for the "maturity rate" series.

k	1	2	3	4	5	6	7	8
Based on OLS estimation								
AIC	-7.35	-7.58	-7.61	-7.62	-7.61	-7.60	-7.62	**-7.62**
HQ	-7.33	-7.55	**-7.57**	-7.57	-7.54	-7.53	-7.53	-7.52
SC	-7.306	-7.50	**-7.51**	-7.48	-7.44	-7.40	-7.39	-7.36
Based on RMLTS estimation								
AIC	-7.43	-7.62	-7.67	-7.69	-7.69	**-7.74**	-7.69	-7.71
HQ	-7.42	-7.59	-7.63	-7.64	-7.62	**-7.67**	-7.60	-7.61
SC	-7.39	-7.55	**-7.57**	-7.564	-7.52	-7.55	-7.46	-7.45

latter criterion yields a consistent estimate of the optimal order we continue the analysis with $k = 3$.

After estimating the VAR(3) model with the robust RMLTS estimator, the corresponding robust residual distances $d_t(\hat{\mathcal{B}}_{\text{RMLTS}}, \hat{\Sigma}_{\text{RMLTS}})$ are computed as in (7), for $t = k+1, \ldots, T$. Figure 2 displays these distances with respect to the time index, and high residual distances indicate outlying observations. It is important to compute these distances based on the robust RMLTS, in order to avoid the well-known masking effect. Furthermore, it is common to compare these distances with a critical value from the chi-square distribution with p degrees of freedom, and we took $\chi_{p,0.99}$. Figure 2 reveals that several suspectable high residuals are detected, in particular around

the years 1954 and 1958. But there are also a couple of other, less extreme outliers, which are more difficult to retrieve from the time series plot in Figure 1. Due to the presence of outliers, it is appropriate to make use of robust methods for further analysis of this data set.

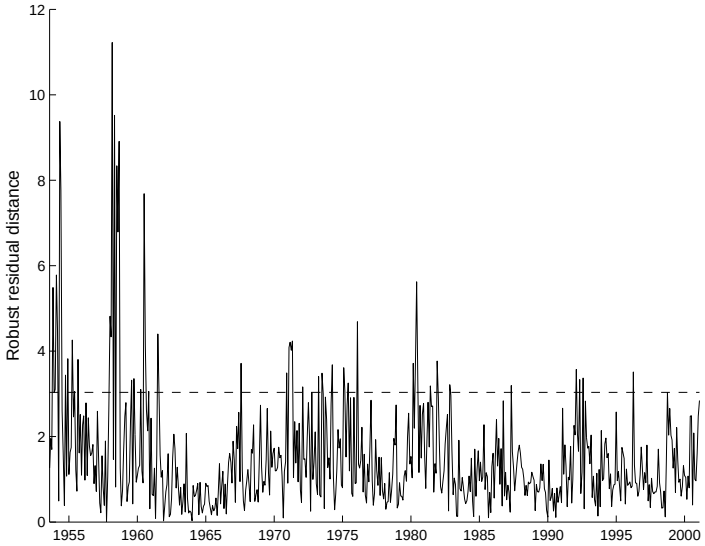

Fig. 2. Robust residual distances for the "maturity rate" series, based on RMLTS estimator of a VAR(3) model. The dashed line represents the critical value at the 1% level.

6 Conclusions

For multivariate time series correlation outliers can be present, which are not necessarily visible in plots of the single univariate series. Development of robust procedures for multiple time series analysis is therefore even more important than for univariate time series analysis.

In this paper we have shown how robust multivariate regression estimators can be used to estimate Vector Autoregressive models. We use the reweighted multivariate least trimmed squares estimator, but other robust multivariate regression estimators could be used as well. Software to compute the MLTS estimator is available at *http://www.econ.kuleuven.be/christophe.croux/public*.

The estimation of VAR models as multivariate regression models has one major disadvantage. A fraction ε of outliers in the original series can produce up to $k\varepsilon$ outliers for the regression model (1), due to the fact that k delayed versions of the time series are used as explanatory variables. Hence, if a

robust regression estimator has a breakdown point of, for example, 1/2, this reduces to $1/(2k)$ when estimating the VAR(k) model. To solve this problem of propagation of outliers, it has been proposed to first filter the series with a robust filter, and then to apply a robust estimator on the robustly filtered data (see Bianco et al. 2001, Maronna et al. 2006). Other types of robust filters were proposed by Davies et al. (2004) and Fried et al. (2006). However, while robust filters are available for univariate series, multivariate versions have not been developed yet, up to our best knowledge, and we leave this for future research.

In the simulation experiments the RMLTS estimators have been compared with the residual autocovariance (RA) estimators of García Ben et al. (1999). The RA estimates are computed iteratively, and we propose to use the RMLTS as a starting value for computing the RA estimators. It turned out that both robust estimators behave then similarly. If there are no outliers in the data set present, the robust estimators perform almost as good as the classical estimator. But if there are outliers, bias and MSE only remain under control when using the robust estimator.

References

AGULLÓ, J., CROUX, C., VAN AELST, S. (2008): The multivariate least trimmed squares estimator. *Journal of Multivariate Analysis 99(3), 311-318.*

BIANCO, A.M., GARCIA BEN, M., MARTINEZ, E.J., YOHAI, V. J. (2001): Outlier detection in regression models with ARIMA errors using robust estimates. *Journal of Forecasting 20, 565–579.*

BROCKWELL, P.J., DAVIS, R.A. (2003): Introduction to Time Series and Forecasting. Wiley, New York.

CROUX, C., HAESBROECK, G. (1999): Influence function and efficiency of the MCD-scatter matrix estimator. *Journal of Multivariate Analysis 71, 161–190.*

DAVIES, P.L., FRIED, R., AND GATHER, U. (2004): Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference 122, 65–78.*

FRANSES, H.P., KLOEK, T., LUCAS, A. (1999): Outlier robust analysis of long-run marketing effects for weekly scanning data. *Journal of Econometrics 89, 293–315.*

FRIED, R., BERNHOLT, T., GATHER, U. (2006): Repeated median and hybrid filters. *Computational Statistics and Data Analysis, 50, 2313–2338.*

GARCIA BEN, M., MARTINEZ, E.J., YOHAI, V.J. (1999): Robust estimation in vector autoregressive moving average models. *Journal of Time Series Analysis 20, 381–399.*

LI, W.K., HUI, Y.V. (1989): Robust multiple time series modelling. *Biometrika 76, 309–315.*

MARONNA, R.A., MARTIN, R.D., YOHAI, V.Y. (2006): *Robust Statistics: Theory and Practice.* Wiley, New York.

ROUSSEEUW, P.J. (1984): Least median of squares regression. *Journal of the American Statistical Association 79, 871–880.*

ROUSSEEUW, P.J., VAN DRIESSEN, K. (1999): A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41, 212–223.*
TSAY, R.S. (2002): *Analysis of Financial Time Series.* Wiley, New York.
TSAY, R.S., PEÑA, D., PANKRATZ, A.E. (2000): Outliers in multivariate time series. *Biometrika 87, 789–804.*

The Choice of the Initial Estimate for Computing MM-Estimates

Marcela Svarc[1] and Víctor J. Yohai[2]

[1] Departamento de Matemática y Ciencias, Universidad de San Andrés, Vito Dumas 284, 1644 Victoria, Pcia. de Buenos Aires, Argentina,
msvarc@udesa.edu.ar
[2] Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires
Ciudad Universitaria, Pabellón 1, 1426 Buenos Aires, Argentina,
vyohai@dm.uba.ar

Abstract. We show, using a Monte Carlo study, that MM-estimates with projection estimates as starting point of an iterative weighted least squares algorithm, behave more robustly than MM-estimates starting at an S-estimate and similar Gaussian efficiency. Moreover the former have a robustness behavior close to the P-estimates with an additional advantage: they are asymptotically normal making statistical inference possible.

Keywords: robust regression, S-estimates, P-estimates

1 Introduction

The most commonly used estimator for linear models is the least squares (LS) estimate. An observation is an atypical point or outlier if it does not fit the regression model which is followed by the large majority of the observations. It is well known that the LS estimate is extremely sensitive to outliers. In fact, a single outlier can have an unbounded effect on the LS estimate. Estimators which are not much influenced by outliers are called robust.

One measure of robustness of an estimate is its breakdown point. Heuristically, the breakdown point is the minimum fraction of arbitrary outliers that can take the estimate beyond any limit. Hampel (1971) introduced the breakdown point as an asymptotic concept, and Donoho and Huber (1983) gave the corresponding finite sample notion. The maximum possible asymptotic breakdown point of an equivariant regression estimate is 0.5.

Yohai (1987) introduced the class of MM-estimates which simultaneously have high breakdown point and high efficiency under normal errors. An MM-estimate requires an initial estimate with high breakdown point but not necessarily efficient. This initial estimate is used to compute an M-scale of the residuals Then using this scale and starting with the initial estimate, a redescending efficient M-estimate is computed using the iterated weighted least squares (IWLS) algorithm.

In general the MM-estimate computed with the IWLS algorithm corresponds to a local minimum of the M-estimate loss function, which is close to the initial estimate. Since, as we shall see later, this loss function may have more than one local minima, the degree of robustness of the MM-estimate is going to be related to the degree of robustness of the initial estimate used to start the IWLS algorithm. The most common implementation of the MM-estimate is to take as initial value an S-estimate with breakdown point 0.5.

Maronna and Yohai (1993) proposed the class of projection estimates (P-estimates) for linear models. They show that these estimates are highly robust. In fact, when the degree of robustness is measured by the maximum asymptotic bias (MAB), these estimates are much more robust than Least Median of Squares, Least Trimmed Squares and S-estimates. One shortcoming of P-estimates is that they are not asymptotically normal.

In this work we compare by Monte Carlo simulation the MM-estimate which uses the P-estimate as initial value with the MM-estimates that start with the S-estimate. We found that MM-estimates that use a P-estimate as initial value have better robustness properties than MM-estimates starting at an S-estimate. Moreover, the advantage of the MM-estimates starting at a P-estimate over the P-estimates is that they are asymptotically normal and highly efficient.

2 Robustness measures

Consider the linear model with p random regressors

$$y_i = \alpha_0 + \beta_0^{'}\mathbf{x}_i + u_i\,,\ i = 1, \ldots, n\,, \tag{1}$$

where $\beta_0 = (\beta_{01}...\beta_{0p})'$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$. We will assume that $(u_1, \mathbf{x}_1), ...,\ (u_n, \mathbf{x}_n)$ are i.i.d. random vectors and that u_i is independent of $\mathbf{x}_i$. We denote by F_0 the distribution of the errors u_i's, G_0 the distribution of the $\mathbf{x}_i$'s and H_0 the joint distribution of $(y, \mathbf{x})$. Throughout the paper we will assume

P1. F_0 has a density f_0 which is symmetric and unimodal.

One way to measure the degree of robustness for large samples is the maximum asymptotic bias which is defined as follows.

Define the contamination neighborhood of size ε of H_0 given by

$$V_\varepsilon(H_0) = \{H : H = (1-\varepsilon)H_0 + \varepsilon H^*,\ \text{where } H^* \text{ is arbitrary}\}.$$

Consider a sequence of estimates $\widehat{\gamma}_n = (\widehat{\alpha}_n, \widehat{\beta}_n)$ of $\gamma_0 = (\alpha_0, \beta_0)$ such that for any $H \in V_\varepsilon(H_0)$, if $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n,\ y_n)$ is a random sample of H, then

$$\lim_{n\to\infty} \widehat{\gamma}_n((x_1, y_1), ..., (\mathbf{x}_n, y_n)) = \widehat{\gamma}_\infty(H) = (\widehat{\alpha}_\infty(H), \widehat{\beta}_\infty(H)) \text{ a.s.},$$

where $(\widehat{\alpha}_\infty(H), \widehat{\beta}_\infty(H)) \in R^2$. Then, the maximum asymptotic biases (MAB) of $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ are defined by

$$\text{MAB}(\{\widehat{\alpha}_n\}, H_0, \varepsilon) = \max_{H \in V_\varepsilon(H_0)} |\widehat{\alpha}_\infty(H) - \alpha_0|,$$

and

$$\text{MAB}(\{\widehat{\beta}_n\}, H_0, \varepsilon) = \max_{H \in V_\varepsilon(H_0)} \widehat{(\beta}_\infty(H) - \beta_0)' \Sigma_{\mathbf{x}} \widehat{(\beta}_\infty(H) - \beta_0), \qquad (2)$$

where $\Sigma_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$. The reason why $\Sigma_{\mathbf{x}}$ is included in (2) is to make this definition affine equivariant.

For some estimates it is very complicated to compute MAB. In this cases, we can consider the pointwise MAB (PMAB)

$$\text{PMAB}(\{\widehat{\beta}_n\}, H_0, \varepsilon) = \max_{H \in V_\varepsilon^*(H_0)} (\widehat{\beta}_\infty(H) - \beta_0)' \Sigma_{\mathbf{x}} (\widehat{\beta}_\infty(H) - \beta_0),$$

where

$$V_\varepsilon^*(H_0) = \{H : H = (1-\varepsilon)H_0 + \varepsilon \delta_{(y^*, \mathbf{x}^*)},\ (y^*, \mathbf{x}^*) \in R^{p+1}\},$$

and where $\delta_{(y^*, \mathbf{x}^*)}$ is the point mass distribution at $(y^*, \mathbf{x}^*)$. In a similar way we can define $\text{PMAB}(\{\widehat{\alpha}_n\}, H_0, \varepsilon)$.

A measure of the robustness behavior for finite samples of an estimate $(\widehat{\alpha}_n, \widehat{\beta}_n)$, is the pointwise maximum mean square error (PMMSE) defined by

$$\text{PMMSE}(\widehat{\alpha}_n, H_0, \varepsilon) = \max_{H \in V_\varepsilon^*(H_0)} E((\widehat{\alpha}_n(H) - \alpha_0)^2)$$

and

$$\text{PMMSE}(\widehat{\beta}_n, H_0, \varepsilon) = \max_{H \in V_\varepsilon^*(H_0} (E((\widehat{\beta}_n(H) - \beta_0)' \mathbf{\Sigma}_{\mathbf{x}} (\widehat{\beta}_n(H) - \beta_0)).$$

3 S-estimates

Put $\gamma = (\alpha, \beta)$, $\gamma_0 = (\alpha_0, \beta_0)$. Define the residual vector $\mathbf{r}(\gamma) = (r_1(\gamma), ..., r_n(\gamma))$ by $r_i(\gamma) = y_i - \alpha - \beta' \mathbf{x}_i$. Consider a function ρ satisfying property **P2** below

P2. $\rho : R \to R$ satisfies: (i) ρ is even, (ii) $\rho(0) = 0$, (iii) $0 \leq u_1 < u_2$ implies $\rho(u_1) \leq \rho(u_2)$, (iv) ρ is bounded, (v) $\sup \rho > 0$.

The S-estimates introduced by Rousseeuw and Yohai (1984) are defined by

$$\widehat{\gamma} = \arg \min_{\gamma \in R^{p+1}} S(\gamma),$$

where $S(\gamma)$ is defined as the value s solving

$$\frac{1}{n-p-1}\sum_{i=1}^{n}\rho\left(\frac{r_i(\gamma)}{s}\right)=b,$$

where b is a given number.

Rousseeuw and Yohai (1984) and Davies (1990) proved that under general assumptions that include **P1** and **P2** we have

$$n^{1/2}(\widehat{\gamma}-\gamma_0)\rightarrow_D N(0,\sigma^2c(\psi,F_0,\sigma)E(\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}')),$$

where $\widetilde{\mathbf{x}}'=(1,\mathbf{x}')$, $\rightarrow^D$ denotes convergence in distribution, $\psi=\rho'$, σ is the asymptotic value of s which is given by the solution of

$$E_{F_0}\left(\rho\left(\frac{u}{\sigma}\right)\right)=b, \tag{3}$$

and where

$$c(\psi,F,\sigma)=\frac{E_F\left(\psi^2\left(\frac{u}{\sigma}\right)\right)}{E_F^2\left(\psi'\left(\frac{u}{\sigma}\right)\right)}. \tag{4}$$

Generally ρ is calibrated so that σ coincides with the standard deviation when F_0 is normal. A necessary and sufficient condition for this is that the solution of (3) be $\sigma=1$ when F_0 is the N(0,1) distribution function.

Rousseeuw and Yohai (1984) have also shown that if $P(\mathbf{a}'\mathbf{x}=b)=0$ for all $\mathbf{a}\in R^p$ and $b\in R$, the asymptotic breakdown point of an S-estimate is given by

$$\varepsilon^*=\min\left(\frac{b}{a},1-\frac{b}{a}\right),$$

where $a=\max_u\rho(u)$. Note that if $b=a/2$ then $\varepsilon^*=0.5$, which is the highest asymptotic breakdown point that a regression equivariant estimate can have.

One family of functions satisfying **P2** is the bisquare family ρ_c^B where $c>0$, given by

$$\rho_c^B(u)=\begin{cases}3\left(\frac{u}{c}\right)^2-3\left(\frac{u}{c}\right)^4+\left(\frac{u}{c}\right)^6 & \text{if } |u|\le c\\ 1 & \text{if } |u|>c.\end{cases} \tag{5}$$

If we take $c=1.56$ and $b=0.5$, the asymptotic breakdown point of the corresponding S-estimate is 0.5. Moreover with these choices, when F_0 is normal, the solution σ of (3) coincides with the standard deviation.

Hossjer (1992) showed that S-estimates can not be simultaneously highly efficient under a normal model and have a high breakdown point such as 0.5. The largest asymptotic efficiency of an S-estimate with breakdown point 0.5 is 0.33.

For S-estimates, it may be proved that the MAB and PMAB coincide, and closed expressions can be found in Martin, Yohai and Zamar (1989). For

MM-estimates there are no closed expressions for MAB and PMAB. However numerical calculations (see Chapter 5 of Maronna, Martin and Yohai (2006)) show that at least in the case that ρ_0 and ρ_1 are taken in the bisquare family, the PMAB of the MM-estimates with efficiencies 0.85 and 0.95 starting at the S-estimate coincide with the PMAB of the initial S-estimate.

4 P-estimates

Maronna and Yohai (1993) introduced the P-estimates which are defined as follows. For any $\gamma = (\alpha, \beta) \in R^{p+1}$, and $\eta = (\mu, \nu) \in R^{p+1}$ let

$$A(\gamma, \eta) = \text{median}_{1 \le i \le n} \frac{r_i(\gamma)}{\eta' \widetilde{\mathbf{x}}_i}.$$

Note that since u_i and $\mathbf{x}_i$ are independent, under **P1** we have

$$A(\gamma_0, \eta) = \text{median } \frac{u_i}{\eta' \widetilde{\mathbf{x}}_i} \to 0 \text{ a.s.. for all } \gamma \in R^{p+1}$$

Then it is natural to define the projection estimates by

$$\widehat{\gamma} = \arg \min_{\gamma \in R^{p+1}} B(\gamma),$$

where

$$B(\gamma) = \sup_{\eta \in R^{p+1}} s(\eta) |A(\gamma, \eta)|.$$

and where $s(\eta) = \text{MAD}(\eta' \widetilde{\mathbf{x}}_i)$

The main results on P-estimates that can be found in Maronna an Yohai (1993) are

- The P-estimates are regression, affine and scale equivariant.
- The rate of consistency of the P-estimates is $n^{1/2}$. However the asymptotic distribution is not normal.
- Assume that $P(\mathbf{a}'\mathbf{x} = b) = 0$ for all $\mathbf{a} \in R^p$ and $b \in R$. Then the asymptotic breakdown point of P-estimates is 0.5
- The maximum bias of the P-estimates satisfies $\text{MAB}(\widehat{\gamma}_n, \varepsilon, H) \le 2C(\varepsilon, H) + o(\varepsilon)$, where $C(\varepsilon, H)$ is a lower bound of MAB for equivariant regression estimates

In Table 1 we compare the MAB of several estimates: the S-estimate based on the bisquare function with breakdown point 0.5, the least median of squares (LMS) and least trimmed squares (LTS) proposed by Rousseeuw (1984) and the P-estimate. Note that the P-estimate has the smallest MAB.

Table 1. Maximum Asymptotic Bias of Robust Estimates.

Estimate	ε			
	0.05	0.10	0.15	0.20
LMS	0.53	0.83	1.13	1.52
LTS	0.73	1.02	1.46	2.02
S	0.56	0.88	1.23	1.65
P	0.16	0.36	0.56	0.82

5 MM-estimates

The class of MM-estimates proposed by Yohai (1987) combines high breakdown point with high asymptotic efficiency under normality. To define the MM-estimates we require two functions ρ_0 and ρ_1 satisfying **P2** and such that $\rho_1 \leq \rho_0$. Then the MM-estimates are defined as follows:

1-Start with a consistent estimate $\widehat{\gamma}_0$ with breakdown point 0.5. It is not necessary that this estimate has high efficiency.

2- Compute an M-scale s of $\mathbf{r}(\widehat{\gamma}_0)$ with breakdown point 0.5 by

$$\frac{1}{n-p-1}\sum_{i=1}^{n}\rho_0\left(\frac{r_i(\widehat{\gamma}_0)}{s}\right)=b,$$

where $b=\max_u \rho_0(u)/2$.

3- Compute a local minimum $\widehat{\gamma}_1$ of

$$M_n(\gamma)=\frac{1}{n}\sum_{i=1}^{n}\rho_1\left(\frac{r_i(\gamma)}{s}\right)$$

such that

$$M_n(\widehat{\gamma}_1)\leq M_n(\widehat{\gamma}_0).$$

Yohai (1987) proved that, under very general assumptions, $\widehat{\beta}_1$ conserves the breakdown of $\widehat{\gamma}_0$ independently of the choice of ρ_1. Moreover, under very general assumptions,

$$n^{1/2}(\widehat{\beta}_1-\beta)\rightarrow^{D} N(0,\sigma^2 c(\psi_1,F_0,\sigma)E(\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}'),$$

where $\psi_1=\rho_1'$, σ is the asymptotic value of s which is given by (3) and $c(\psi,F,\sigma)$ is given by (4). When ρ_0 is chosen so that σ coincides with the standard error when u is normal, the Gaussian asymptotic efficiency of the MM-estimate with respect to the LS-estimate is

$$\mathrm{EFF}=\frac{E_\phi^2\left(\psi_1'(u)\right)}{E_\phi\left(\psi_1^2(u)\right)},$$

where ϕ is the standard normal distribution.

Therefore, since this efficiency depends only on ρ_1, we can choose this function so that under Gaussian errors EFF be equal to any desired value. For example, we can choose ρ_0 and ρ_1 in the bisquare family; i.e., we can take $\rho_0 = \rho^B_{c_0}$ and $\rho_1(u) = \rho^B_{c_1}$. In that case it will be convenient to take $c_0 = 1.55$ so that the value of σ coincides with the standard deviation under normal errors. The value of c_1 should be chosen according to the desired Gaussian efficiency. For example for an efficiency of 0.95, $c_1 = 4.65$, and for an efficiency of 0.85 $c_1 = 3.46$.

One way to compute the MM-estimate $\widehat{\gamma}_1$ is by means of the iterative weighted least squares (IWLS) algorithm starting at $\gamma^{(1)} = \widehat{\gamma}_0$. The recursion step for this algorithm is as follows:

Given $\gamma^{(j)}$ we define the weights $w_i = w(r_i(\gamma^{(j)})/s)$, $1 \leq i \leq n$, where $w(u) = \psi(u)/u$. Then, $\gamma^{(j+1)}$ is the weighted least square estimate

$$\gamma^{(j+1)} = \arg\min_{\gamma} \sum_{i=1}^{n} w_i r_i^2(\gamma)$$

In general the function $M_n(\gamma)$ has several local minima. When $n \to \infty$, $M_n(\gamma)$ converges to

$$M_\infty(\gamma) = E_H\left(\rho_1\left(\frac{y - \gamma'\widetilde{\mathbf{x}}}{\sigma(H)}\right)\right), \tag{6}$$

where H is the joint distribution of $(y, \mathbf{x})$ and $\sigma(H)$ is the asymptotic scale defined by

$$E_H\left(\rho_0\left(\frac{y - \widehat{\alpha}_{0\infty} - \widehat{\beta}'_{0\infty}\mathbf{x}}{\sigma(H)}\right)\right) = b.$$

When the linear model is satisfied and **P1** holds for F_0, the only local minima of (6) is at $\gamma = \gamma_0$.

Suppose now for simplicity that the model does not have intercept, $\beta_0 = \mathbf{0}$ and that there is a fraction $(1 - \varepsilon)$ of outliers equals to $(y_0, \mathbf{x}_0)$,where $\mathbf{x}_0 = (x_0, 0, ..., 0)$. In this case M_∞ depends only on β_1, the first coordinate of β. The worst situation is when $|x_0| \to \infty$ an in this case $M_\infty(\beta_1)$ has two local minima, one at 0 and another close to the the contamination slope $m_0 = y_0/x_0$. There exists a value m^* such that that when $m_0 < m^*$ the global minimum is the local minimum closest to m_0, and when $m_0 > m^*$ the global minimum is at 0. As a consequence of this, if we choose as estimate the global minimum, the maximum asymptotic bias is m^*. We illustrate this behavior in Figure 1, where $M_\infty(\beta)$ is plotted for three values of the slope m_0. In this case we take $\mathbf{x} \sim N(0, I)$ and $u \sim N(0, 1)$ so that $\beta_0 = 0$.

The local minimum $\widehat{\gamma}_1$ to which the IWLS algorithm converges, depends on the initial estimate $\widehat{\gamma}_0$. In general we can state the following rule: if we start the IWLS algorithm sufficiently close to a local minimum, it will converge to that local minimum. Therefore the degree of robustness of $\widehat{\gamma}_1$ is going to be

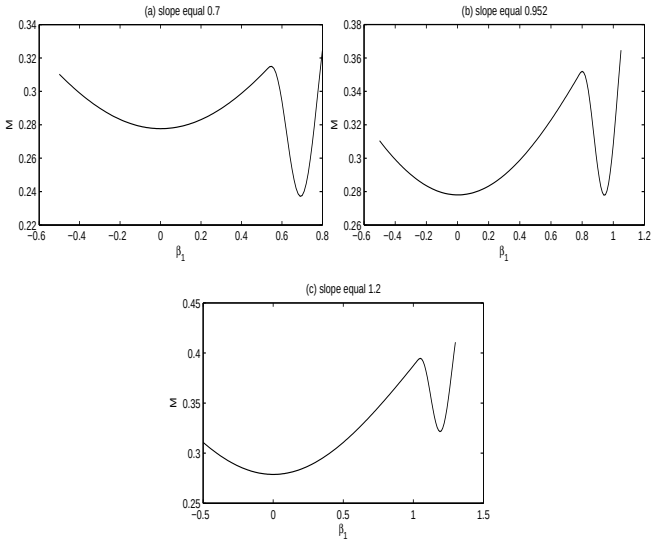

Fig. 1. Plot of $M_\infty(\beta)$ for three values of the contamination slope: (a) the global minimum is close to the contamination slope (b) there are two global minima (c) the global minimum is at 0.

related to the degree of robustness of $\widehat{\gamma}_0$. However as it is shown in Yohai (1987), the asymptotic efficiency of the MM-estimate is independent of the choice of $\widehat{\gamma}_0$.

The most popular choice of $\widehat{\gamma}_0$ (the one employed in the SPLUS, R and SAS. programs) is to start with an S-estimate based on ρ_0. However, as is shown in Table 1, the MAB of the P-estimate is smaller that the one of the S-estimate. For that reason we can expect that an MM-estimate that takes as $\widehat{\gamma}_0$ a P-estimate would be more robust than the MM-estimate that starts at an S-estimate. Simultaneously, since both MM-estimates have the same asymptotic efficiency we can expect a similar behavior under a linear model with Gaussian errors and no outliers.

6 Monte Carlo results

In this Section we report the results of a Monte Carlo study aimed to compare the performance of the MM-estimates starting at an S- and a P- estimate. The functions ρ_0 and ρ_1 were taken in the bisquare family (5). We took $\rho_0 = \rho^{(B)}_{1.55}$ and $b = 0.5$ so that (3) holds. Moreover, we chose $\rho_1 = \rho^{(B)}_{3.44}$ which corresponds to an MM-estimate having an asymptotic Gaussian efficiency of 0.85.

We consider samples of size $n = 100$, and $p = 2$, 5 and 10. In all cases a fraction $(1-\varepsilon)$ of the observations $(y_i, \mathbf{x}_i)$ were taken from a multivariate normal distribution, and the remaining observations are equal outliers $(y_0, \mathbf{x}_0)$. Because of the equivariance properties of all the estimates considered in this study, without loss of generality, the normal observations were taken with mean $\mathbf{0}$ and identity covariance matrix. This corresponds to $\beta_0 = \mathbf{0}$ and $\alpha_0 = 0$.

Because of the equivariance of the estimates considered in the study, without loss of generality, the values of $\mathbf{x}_0$ were taken of the form $(x_0, 0, .., 0)$ and $y_0 = mx_0$. We took two values of x_0: $x_0 = 1$ (low leverage outliers) and $x_0 = 10$ (high leverage outliers). We took a grid of values of m with step 0.1 and looked for the value that achieves the maximum mean square error. The number of the Monte Carlo replications was $N = 500$.

The S-estimate was computed with the fast algorithm for S-estimates proposed by Salibian-Barrera and Yohai (2006) and the P-estimate with the algorithm based on subsampling described in Maronna and Yohai (1993) taking the same set of candidates than for the S-estimate. The number of subsamples used for both estimates was 500.

In order to measure the performance of each estimate we compute the sample mean squared error (MSE) as follows. Suppose that $\gamma_n^{(1)} = (\widehat{\alpha}_n^{(1)}, \widehat{\beta}_n^{(1)})$, ..., $\widehat{\gamma}_n^{(N)} = (\widehat{\alpha}_n^{(N)}, \widehat{\beta}_n^{(N)})$ are N replications of an estimate $\widehat{\gamma}_n = (\widehat{\alpha}_n, \widehat{\beta}_n)$ of γ_0. Then we estimate the MSE of $\widehat{\alpha}_n$ and of $\widehat{\beta}_n$ by

$$\widehat{\mathrm{MSE}}\left(\widehat{\alpha}_n\right) = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{\alpha}_n^{(i)} - \alpha_0\right|^2 \tag{7}$$

and

$$\widehat{\mathrm{MSE}}\left(\widehat{\beta}_n\right) = \frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\beta}_n^{(i)} - \beta_0\right\|^2 \ , \tag{8}$$

where $\|\|$ denotes the Euclidean norm.

We compute the following estimates: the P-estimate (P), the S-estimate (S), the MM-estimate starting from an S-estimate (SMM) and the MM-estimate starting from a P-estimate (PMM).

Table 2 show the MSE 's when there are no outliers. Tables 3 and 4 report the maximum MSE's for the case that $\varepsilon = 0.10$ for low and high leverage outliers respectively. Finally, Tables 5 and 6 give the MSE 's for the case $\varepsilon = 0.20$. Table 7 contains the standard errors of the difference between the MSE's of the SMM- and PMM-estimates when there are no outliers. The standard errors under outlier contamination stay close to these values. From the analysis of Tables 2-7 we can draw the following conclusions

- When there are not outliers both MM-estimates have a similar behavior

- When $x_0 = 1$, the MMS- and MMP-estimates have similar maximum MSE's for p equal 2 and 5. When $p = 10$ the MMP-estimate outperforms the MMS-estimate.
- When $x_0 = 10$, the MMP-estimate behave much better than the MMS-estimate for all values of p.

Table 2. Mean Square Errors Without Outliers.

Estimates	$\widehat{\alpha}_n$			$\widehat{\beta}_n$		
	p					
	2	5	10	2	5	10
P	0.015	0.016	0.018	0.034	0.08	0.18
S	0.033	0.036	0.040	0.067	0.21	0.43
MMP	0.012	0.012	0.012	0.023	0.06	0.13
MMS	0.012	0.013	0.013	0.023	0.06	0.14

Table 3. Maximum Mean Square Errors when $\varepsilon = 0.10$ and $x_0 = 1$.

Estimates	$\widehat{\alpha}_n$			$\widehat{\beta}_n$		
	p					
	2	5	10	2	5	10
P	0.040	0.044	0.054	0.099	0.19	0.33
S	0.128	0.162	0.223	0.218	0.46	0.92
MMP	0.049	0.052	0.057	0.067	0.11	0.21
MMS	0.049	0.051	0.055	0.067	0.11	0.21

Table 4. Maximum Mean Square Errors when $\varepsilon = 0.20$ and $x_0 = 1$.

Estimates	$\widehat{\alpha}_n$			$\widehat{\beta}_n$		
	p					
	2	5	10	2	5	10
P	0.15	0.15	0.25	0.43.	0.84	1.92
S	0.55	0.73	1.29	1.20	2.31	4.82
MMP	0.29	0.33	0.40	0.35	0.48	0.72
MMS	0.26	0.30	0.44	0.33	0.48	1.01

Figure 2 shows the MSE as a function of the contamination slope for the two MM-estimates in the case of $x_0 = 10$, $\varepsilon = 0.10$ and $p = 10$. We observe

Table 5. Maximum Mean Square Errors when $\varepsilon = 0.10$ and $x_0 = 10$.

Estimates	$\widehat{\alpha}_n$			$\widehat{\beta}_n$		
			p			
	2	5	10	2	5	10
P	0.037	0.035	0.035	0.13	0.26	0.52
S	0.076	0.073	0.090	0.50	0.79	1.33
MMP	0.017	0.017	0.020	0.20	0.24	0.43
MMS	0.024	0.026	0.034	0.42	0.56	0.84

Table 6. Maximum Mean Square Errors when $\varepsilon = 0.20$ and $x_0 = 10$.

Estimates	$\widehat{\alpha}_n$			$\widehat{\beta}_n$		
			p			
	2	5	10	2	5	10
P	0.12	0.12	0.15	0.71	1.30	2.59
S	0.21	0.27	0.31	1.90	3.09	4.90
MMP	0.03	0.05	0.07	0.65	1.12	2.17
MMS	0.07	0.11	0.17	1.75	2.48	3.84

Table 7. Standard errors of the difference of the MSE's of the SMM- and PMM-estimates.

Coefficient	p		
	2	5	10
intercept	0.0005	0.002	0.004
slopes	0.001	0.004	0.013

that the MMP estimate behaves better than the MMS estimate uniformly on the contamination slope m. Similar behaviors occurs for all the other values of p and ε at $x_0 = 10$.

The computing time required to fit the PMM with 500 subsamples to a data set of 500 observations and 10 regressors is approximately 4 seconds using a MATLAB program and a PC computer with an AMD Athlon 1.8 GHz processor

7 Concluding remarks

A Monte Carlo study has shown that MM-estimates that use a P-estimate as starting value, have a degree of robustness comparable to that of the P-estimate. and much higher than that of the MM-estimate starting at an S-estimate. On the other hand both MM-estimates have comparable Gaussian efficiencies. An additional advantage of the MM-estimate starting at

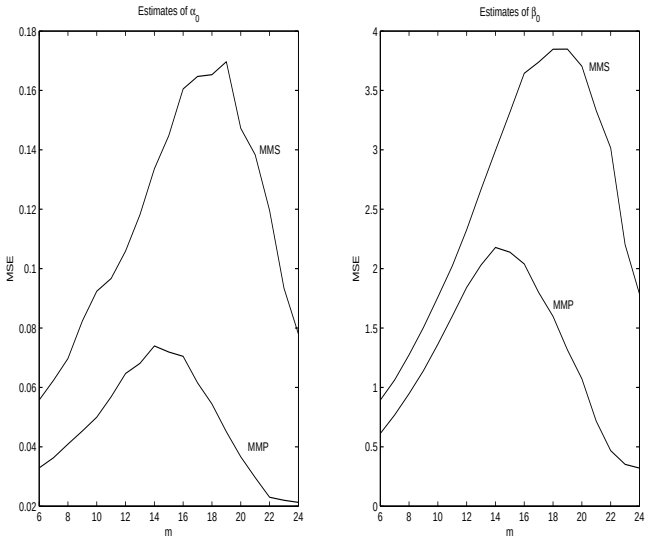

Fig. 2. MSE 's of MM-estimates for $x_0 = 10$..

a P-estimate is that, contrary to what happens with the P-estimate, it is asymptotically normal and thus allows statistical inference.

References

DAVIES, L. (1990): The asymptotics of S-estimators in the linear regression model. *The Annals of Statistics*, **18**, 1651-1675.

DONOHO, D.L. and HUBER, P.J. (1983): The notion of breakdown-point. In: P.J. Bickel, K.A. Doksum and J.L Hodges, Jr. (Eds.): *A Festschrift for Erich L. Lehmann*. Wadsworth, Belmont, California, 157-184.

HAMPEL, F.R. (1971): A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics, 42, 1887-1896.*

HOSSJER, 0. (1992): On the optimality of S-estimators, *Statistics & Probability Letters, 14, 413-419.*

MARONNA, R.A. and YOHAI, V.J. (1993): Bias-robust estimates of regression based on projections. *The Annals of Statistics, 21, 965-990.*

MARONNA, R.A., MARTIN, R.D. and YOHAI, V.J. (2006): *Robust Statistics: Theory and Methods*, Wiley, Chichister.

MARTIN, R.D., YOHAI, V.J. and ZAMAR, R. (1989). Min-max bias robust regression. *The Annals of Statistics, 17, 1608-1630.*

ROUSSEEUW, P.J. (1984): Least median of squares regression, *J. Am. Stat. Assoc., 79, 871-880.*

ROUSSEEUW, P.J. and YOHAI, V.J. (1984): Robust regression by means of S-estimators. In: J. Franke, W. Hardle and R.D. Martin (Eds): *Robust and Nonlinear Time Series, Lecture Notes in Statistics, 26, 256-272*. Springer-Verlag, Berlin.

SALIBIAN-BARRERA, M. and YOHAI, V.J. (2006): A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics, 15, 414-427*.

YOHAI, V.J. (1987): High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics, 15, 642-656*.

Metropolis Versus Simulated Annealing and the Black-Box-Complexity of Optimization Problems

Ingo Wegener*

Technische Universität Dortmund
44221 Dortmund, Germany, *ingo.wegener@uni-dortmund.de*

Abstract. Many real-world optimization problems cannot be modeled by a well-described objective function to apply methods from mathematical optimization theory. Then randomized search heuristics are applied - often with good success. Although heuristical by nature, they are algorithms and can be analyzed like all randomized algorithms, at least in principle. Two fundamental results of this kind are presented to show how such a theory can be developed.

Keywords: randomized search heuristics, minimum spanning tree, black-box-complexity

1 Introduction

One of the best-studied areas in computer science is the design and analysis of algorithms for optimization problems. This holds for deterministic algorithms as well as for randomized algorithms (see, e.g., Cormen, Leiserson, and Rivest (1996) and Motwani and Raghavan (1995)). The criterion of the analysis is the asymptotic (with respect to the problem dimension) worst-case (with respect to the problem instance) expected (with respect to the random bits used by the algorithm) run time of the algorithms. Up to now, large lower bounds need some complexity theoretical assumption like $NP \neq P$. For almost all well-known optimization problems the best algorithms in this scenario are problem-specific and use essentially the structure of the considered problem.

Therefore, randomized search heuristics like randomized local search, tabu search, simulated annealing, and all variants of evolutionary and genetic algorithms are typically not considered in this context. They do not beat the highly specialized algorithms in their domain. Nevertheless, practitioners repeat surprisingly good results for these heuristics. This makes

* Supported in part by the DFG collaborative research projects SFB 475 (Reduction of Complexity for Multivariate Data Structures) and SFB 531 (Computational Intelligence).

it necessary to develop a theory of randomized search heuristics. Some early approaches should be mentioned: Glover and Laguna (1993) for tabu search, Kirkpatrick, Gelett, and Vecchi (1983) and Sasaki and Hajek (1988) for simulated annealing, Wegener (2001) and Droste, Jansen, and Wegener (2002) for evolutionary algorithms. Here we present two fundamental results in this direction.

First, a problem known as open since more than ten years about the Metropolis algorithm and simulated annealing is solved (this is based on Wegener (2005)). Then we describe a complexity theory for randomized search heuristics excluding highly specialized algorithms and apply the theory of two-person zero-sum games to prove a large lower bound for a fundamental class of functions (this is based on Droste, Jansen, and Wegener (2006)).

2 The Metropolis algorithm and Simulated Annealing

We describe the Metropolis algorithm (CMA) with temperature T for minimization problems on $\{0,1\}^m$. The first search point x is chosen in some way discussed later. Each round of an infinite loop consists of local change and selection.

Local change: Let x be the current search point. Choose $i \in \{1,\ldots,m\}$ uniformly at random and flip x_i, i.e., let $x' = (x'_1,\ldots,x'_m)$ where $x'_j = x_j$, if $j \neq i$, and $x'_i = 1 - x_i$.
Selection of the new current search point with respect to a fitness function f:

if $f(x') \leq f(x)$: select x',
if $f(x') > f(x)$: select x' with probability $exp\{-(f(x') - f(x))/T\}$,
otherwise select x.

We have to discuss some details in order to ensure that our results are not based on too special choices. Randomized search heuristics do not produce a certificate that a search point is opimal. Therefore, the algorithm contains an infinite loop, but the run time is defined as the number of rounds until an optimal search point is produced. A round cannot be performed in time $\mathcal{O}(1)$ but quite efficiently and people have agreed to count the number of rounds.

We choose 1^m as starting point. This is for our problems the worst legal search point and similar to the choice 0^n for the maximum matching problem (Sasaki and Hajek (1988)) and the maximum clique problem (Jerrum (1992)). This choice excludes the lower bound technique of Sasaki (1991) which only ensures the existence of a bad starting point.

Finally, we introduce SA based on a cooling schedule $T(t)$. The initial temperature $T(1)$ may depend on m and the largest possible weight ω_{max}. The temperature $T(t)$ applied by the selection operator in step t equals α^{t-1}. $T(1)$, there $\alpha < 1$ is a constant which may depend on m and an upper bound on ω_{max}. This cooling schedule does not include any knowledge about the problem instance. We use a kind of "continuous cooling", other possibilities are longer phases with a constant temperature or dynamic cooling schedules that depend on the success rate (where a step is called successful if x' is selected) or the rate of f-improving steps.

Next we describe a challenge posed in the literature. For graphs on m edges a search point $x \in \{0,1\}^m$ chooses all edges e_i where $x_i = 1$. We investigate the minimum spanning tree problem (MSTP) on connected graphs.

We have chosen the fitness function f where $f(x) = \infty$ for search points x describing unconnected graphs and where $f(x)$ is the total weight of all chosen edges if x describes a connected graph. Unconnected graphs are never accepted as current search points. This again is in accordance with Sasaki and Hajek (1988) and Jerrum (1992). All search points are legal solutions in the graph bisection problem and therefore Jerrum and Sorkin (1993, 1998) start with randomly chosen search points.
We follow Sasaki and Hajek (1988) and Jerrum (1992) in allowing only 1-bit neighborhoods. Neumann and Wegener (2004) have analyzed RLS with 1-bit and 2-bit flips (RLS equals the frozen MA at temperature $T = 0$) and a simple EA for the MSTP. These algorithms do not select new search points which are worse than the old one. Hence, their search strategy is completely different from the strategy applied by MA and SA that have to accept sometimes worsenings to find the optimum. Flips of two bits allow to include an edge into a tree and to exclude simultaneously an edge of the newly created cycle. RLS and the simple EA find an MST in an expected number of $\mathcal{O}(m^2(\log m + \log \omega_{max}))$ steps, where ω_{max} denotes the maximal weight. Note that we are not looking for a "best" algorithm for the MSTP. The main idea of an elitist EA is to reject worsenings and to escape from local optima by non-local steps. The main idea of MA and SA is to work with very local steps and to escape from local optima by accepting worsenings. The situation here is similar to the case of maximum mathcings where also flips of 2 bits are helpful to shorten augmenting paths, compare Sasaki and Hajek (1988) who analyze SA with 1-bit flips only and Giel and Wegener (2003) who analyze RLS with 1-bit and 2-bit flips and a simple EA.

3 Efficiency measures

There are many well-known convergence results on MA and SA. We want to distinguish "efficient behavior" from non-efficient one. The first idea is to define efficiency as expected polynomial time. We think that this is not a good choice. There may be a small probability of missing a good event for temperatures in some interval $[T_1, T_2]$. For temperatures smaller then T_1 it may be very unlikely that the good event happens. This may cause a superpolynomial or even exponential expected run time although the run time is polynomially bounded with overwhelming probability.

Definition 3. Let A be a randomized search heuristic (RSH) running for a polynomial number of $p(m)$ rounds and let $s(m)$ be the success probability, i.e., the probability that A finds an optimal search point whithin this phase. A is called

- successful, if $s(m) \geq 1/q(m)$ for some polynomial $q(m)$,
- highly successful, if $s(m) \geq 1 - 1/q(m)$ for some polynomial $q(m)$, and
- successful with overwhelming probability, if $s(m) = 1 - e^{-\Omega(m^\epsilon)}$ for some $\epsilon > 0$.

One can be satisfied with successful RSHs, since then multistart variants not depending on p and q are successful with overwhelming probability and have an expected polynomial run time. An RSH is called unsuccessful if, for each polynomial p, the success probability within $p(m)$ steps is $o(m^{-k})$ for each constant k. This implies a superpolynomial expected optimization time. Moreover, multistart variants do not help.

4 Metropolis vs. Simulated Annealing

Here, we are interested in simulated annealing and the Metropolis algorithm (which can be defined as SA with a fixed temperature). Both algorithms are defined in Section 2. It is an obvious question how to use the freedom to choose a cooling schedule for SA and whether this option is essential. Little is known about this leading Jerrum and Sinclair (1996, page 516) to the following statement: "It remains an outstanding open problem to exhibit a natural example in which simulated annealing with any non-trivial cooling schedule provably outperforms the Metropolis algorithm at a carefully chosen fixed value of α." In their paper, α is the temperature. The notion of a "natural example" is vague, but the known examples are obviously artificial. Sorkin (1991) has proven the considered effect for a so-called fractal energy landscape. The chaotic behavior of this function asks for different temperatures in different phases of the search. The artificial example due to Droste, Jansen, and Wegener (2001) allows a simpler analysis.

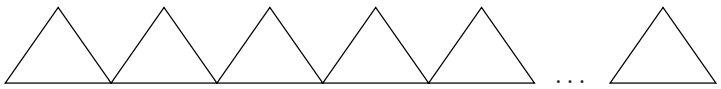

Fig. 1. Graphs called connected triangles.

Jerrum and Sorkin (1998) have analyzed the Metropolis algorithm for the graph bisection problem. They focus the interest on problems from combinatorial optimization: "Unfortunately no combinatorial optimization problem that has been subjected to rigorous theoretical analysis has been exhibited this phenomenon: those problems that can be solved efficiently by simulated annealing can be solved just as effectively by 'annealing' at a single carefully selected temperature. A rigorous demonstration that annealing is provably beneficial for some natural optimization problems would rate as a significant theoretical advance."

Our problem of choice is the minimum spanning tree problem (MSTP) which is contained in all textbooks on combinatorial optimization and should be accepted as "natural optimization problem." It should be obvious that SA cannot beat MA for each problem instance. E. g., for graphs where all edge weights equal 1 the frozen MA (at temperature 0) cannot be beaten by SA. In Section 3, we describe the notion of efficiency for randomized search heuristics and, in Section 5, we describe simple instances of the MSTP where SA outperforms MA. The underlying graphs will be so-called connected triangles (CT), see Figure 1.

The idea is to produce examples as simple as possible. This allows proofs which can be taught in introductory courses on randomized search heuristics. Afterwards, we try to understand which instances of the MSTP can be solved efficiently by SA and MA, only by SA, or by none of them. Weights $w_1, \ldots, w_m$ are called $(1+\varepsilon)$-separated if $w_i > w_j$ implies $w_i \geq (1+\varepsilon) \cdot w_j$. For each $\varepsilon(m) = o(1)$ there are graphs with $(1+\varepsilon(m))$-separated weights such that SA cannot attack them efficiently (Section 6). For each constant $\varepsilon > 0$, SA can attack all graphs with $(1+\varepsilon)$-separated weights efficiently. These results imply that SA outperforms MA on a much larger class of graphs than the connected triangles discussed in Section 5.

It should be obvious that we do not hope that SA or MA beats the well-known algorithms due to Kruskal and to Prim. Again we like to transfer a statement of Jerrum and Sorkin (1998) from minimum bisections to minimum spanning trees (MSTs): "Our main contribution is not, then, to provide a particularly effective algorithm for the minimum bisection problem ..., but to analyze the performance of a popular heuristic applied to a reasonably realistic problem in combinatorial optimization."

5 Simulated Annealing beats Metropolis on some simple graphs

Our plan is to present simple graphs where SA beats MA for each temperature. The graphs should allow proofs as simple as possible. The idea behind the chosen graphs is the following. The problem to compute an MST on graphs with many two-connected components is separable, i. e., an MST consists of MSTs on the two-connected components. We investigate graphs where each two-connected component can be handled easily by MA with a well-chosen temperature, but different components need different temperatures. To keep the analysis easy the components have constant size. This implies that, for high temperatures, each component can be optimized, but the solutions are not stable. They are destroyed from time to time and then reconstructed. Therefore, it is unlikely that all the components are optimized simultaneously. SA can handle these graphs efficiently.

As announced, we investigate connected triangles (CT), see Figure 1, with $m = 6n$ edges. The number of triangles equals $2n$ and the number of vertices equals $4n + 1$. The weight profile (w_1, w_2, w_3) of a triangle is simply the ordered vector of the three edge weights. We investigate CTs with n triangles with weight profile $(1, 1, m)$ and n triangles with weight profile (m^2, m^2, m^3). The unique MST consists of all edges of weight 1 or m^2.

Theorem 1. *The probability that the Metropolis algorithm applied to CTs with n triangles with weight profile* $(1, 1, m)$ *and n triangles with weight profile* (m^2, m^2, m^3) *computes the MST within* e^{cm} *steps (c a positive constant which is small enough) is bounded above by* $e^{-\Omega(m)}$*, i. e., MA is unsuccessful on these instances.*

Proof. We distinguish the cases of high temperature ($T \geq m$) and low temperature ($T < m$).

The low temperature case is easy. We do not care about the triangles with weight profile $(1, 1, m)$. For each other triangle, MA accepts the exclusion of the first flipping edge. By Chernoff bounds, with probability $1 - 2^{-\Omega(m)}$, we obtain $\Omega(m)$ triangles where the first spanning tree contains the heavy edge. In order to obtain the MST it is necessary to include the missing edge of weight m^2. If this edge is chosen to flip, the probability of selecting the new search point equals $e^{-m^2/T} \leq e^{-m}$. Hence, the success probability within $e^{m/2}$ steps is $e^{-\Omega(m)}$.

In the high temperature case, we do not care about the heavy triangles. For the light triangles, we distinguish between complete triangles (the search point chooses all three edges), optimal triangles (the two weight-1 edges are chosen), and bad triangles. The status of each triangle starts with "complete" and follows a Markov chain with the following transition probabilities:

	complete	optimal	bad
complete	$1-3/m$	$1/m$	$2/m$
optimal	$\frac{1}{m}\cdot e^{-m/T}$	$1-\frac{1}{m}\cdot e^{-m/T}$	0
bad	$\frac{1}{m}\cdot e^{-1/T}$	0	$1-\frac{1}{m}\cdot e^{-1/T}$

Let X_t be the number of optimal triangles after time step t, i. e., $X_0 = 0$. We are waiting for the first point of time t when $X_t = n$. Obviously, $|X_{t+1} - X_t| \leq 1$. Moreover,

$$\Pr(X_{t+1} = a+1 \mid X_t = a) \leq \frac{n-a}{m}$$

since it is necessary to flip the heaviest edge in one of the at most $n-a$ complete triangles, and

$$\Pr(X_{t+1} = a-1 \mid X_t = a) = \frac{a}{m}\cdot e^{-m/T} \geq \frac{a}{3m}$$

since $T \geq m$ and since it is necessary to flip the heaviest edge in one of the optimal triangles and to accept the new search point. Since we are interested in lower bounds, we use the upper bound for the probability of increasing a and the lower bound for the probability of decreasing a. Ignoring steps not changing a, we obtain the following transition probabilities for the new Markov chain Y_t:

$$\Pr(Y_{t+1} = a-1 | Y_t = a) = \frac{a/(3m)}{a/(3m)+(n-a)/m} = \frac{a}{3n-2a}.$$

There has to be a phase where the Y-value increases from $(10/11)n$ to n without reaching $(9/11)n$. In such a phase the probability of decreasing steps is bounded below by $\frac{(9/11)n}{3n-(18/11)n} = \frac{3}{5}$. Applying results on the gambler's ruin problem, the probability that one phase starting at $a = (10/11)n$ and finishing at $a = (9/11)n$ or $a = n$ stops at $a = n$ is bounded above by

$$((3/2)^{n/11}-1)/((3/2)^{2n/11}-1) = e^{-\Omega(m)}$$

since the probability of decreasing steps is at least by a factor of $3/2$ larger than the probability of increasing steps. Hence, the probability of finding the MST within e^{cm} steps, $c > 0$ small enough, is bounded by $e^{-\Omega(m)}$. □

Theorem 2. *Let p be a polynomial and let the cooling schedule be described by $T(1) = m^3$ and $\alpha = 1 - 1/(cm)$ for some constant $c > 0$. If c is large enough, the probability that simulated annealing applied to CTs with n $(1,1,m)$-triangles and n (m^2, m^2, m^3)-triangles computes the MST within $3cm\ln m$ steps is bounded below by $1 - 1/p(m)$.*

Proof. We only investigate the search until the temperature drops below 1. This phase has a length of at most $3cm\ln m$ steps and contains two subphases

where the temperature is in the interval $[m^2, m^{5/2}]$ or in the interval $[1, m^{1/2}]$. The length of each subphase is at least $(c/4)m \ln m$.

If $T \leq m^{5/2}$, the probability of including an edge of weight m^3 is bounded above by $e^{-m^{1/2}}$. Each run where such an event happens is considered as unsuccessful. If $T \in [m^2, m^{5/2}]$ and an (m^2, m^2, m^3)-triangle is optimal, this triangle remains optimal unless the event considered above happens. Applying Chernoff bounds to each edge and choosing c large enough, the probability of not flipping edges of each triangle at least $c'' \log m$ times is bounded by m^{-k}, $c'' > 0$ and k arbitrary constants. This is a second source of bad behavior. Now, we investigate one (m^2, m^2, m^3)-triangle and the steps flipping one of its edges. For each complete or bad triangle, there is a chance that it turns into optimal within the next two steps concerning this triangle. This happens if the right two edges flip in the right order (probability 1/9) and the inclusion of the edge with weight m^2 is accepted (probability $e^{-m^2/T} \geq e^{-1}$). The probability of not having a good pair among the at least $(c''/2) \log m$ step pairs, can be made much smaller than m^{-k} by choosing c'' large enough. Altogether, the probability that the first subphase does not finish with MSTs on all (m^2, m^2, m^3)-triangles can be made smaller than $1/(3p(m))$.

The same calculations for $T \in [1, m^{1/2}]$ and the $(1, 1, m)$-triangles show that the probability that the second subphase does not finish with MSTs on all $(1, 1, m)$-triangles can be made smaller than $1/(3p(m))$. Finally, the probability that an (m^2, m^2, m^3)-triangle has turned from optimal into non-optimal after the first subphase is smaller than $1/(3p(m))$. This proves the theorem. □

We have proved that SA is highly successful for the considered graph instances. It is easy to choose a cooling schedule such that SA is even successful with overwhelming probability, e. g., $T(1) = m^3$ and $\alpha = 1 - 1/m^2$.

This section contains the result announced in the title of the paper. In the following two sections, we investigate which graphs can be handled efficiently by MA and SA, only by SA, or by none of them.

6 Connected triangles with the same weight profile

It is interesting to understand how much different weights have to differ such that MA or SA are able to construct efficiently an MST. For this reason, we investigate graphs consisting of connected triangles in more detail. In this section, we consider the case of n CTs with the same weight profile $(w, w, (1 + \varepsilon(m)) \cdot w)$ where $\varepsilon(m) > 0$. We distinguish the cases where $\varepsilon(m)$ is bounded below by a positive constant ε and the case where $\varepsilon(m) = o(1)$.

Theorem 3. *If $\varepsilon(m) \geq \varepsilon > 0$, MA with an appropriate temperature finds the MST on CTs with n $(w, w, (1 + \varepsilon(m)) \cdot w)$-triangles in expected polynomial time and is successful with overwhelming probability.*

Proof. A good temperature has to fulfil two properties:

- It has to be low enough to distinguish w-edges effectively from $(1+\varepsilon)\cdot w$-edges.
- It has to be high enough to allow the inclusion of a w-edge in expected polynomial time.

We choose $\gamma := 3/\varepsilon$ and $T := w/(\gamma \cdot \ln m)$. The probability to accept the inclusion of a w-edge equals $e^{-w/T} = m^{-\gamma}$ while the corresponding probability for a $((1 + \varepsilon(m)) \cdot w)$-edge equals $m^{-\gamma\cdot(1+\varepsilon(m))} \leq m^{-\gamma-3}$. We analyze the success probability of a phase of length $m^{\gamma+2}$ starting with an arbitrary connected graph. The event to accept the inclusion of a heavy edge is considered as an unsuccessful phase. The probability of this event is bounded above by $1/m$. Following the lines of the proof of Theorem 2 we have for each triangle with overwhelming probability $\Omega(m^{\gamma+1})$ steps flipping an edge of this triangle which we partition into $\Omega(m^{\gamma+1})$ pairs of consecutive steps. The probability that a complete or bad triangle is turned within such two steps into an optimal one is $\Omega(m^{-\gamma})$. Hence, with overwhelming probability, all triangles turn into optimal during this phase and with probability at least $1 - 1/m$ none of them is turned into non-optimal. Hence, the expected number of phases is $O(1)$ and the probability that a sequence of m phases is unsuccessful is exponentially small. □

It is obvious how to tune the parameters in order to get improved run times. We omit such calculations which do not need new ideas. SA finds the MST in polynomial time with a probability exponentially close to 1 if it starts with $T(1) := w/(\gamma \cdot \ln m)$ and has a cooling schedule that cools down the temperature sufficiently slow. This follows in the same way as Theorem 3.

Theorem 4. *If $\varepsilon(m) = o(1)$, MA and SA are unsuccessful on CTs with n $(w, w, (1 + \varepsilon(m)) \cdot w)$-triangles.*

Proof. First, we investigate MA. The search starts with n complete triangles and each one has a probability of $2/3$ to be turned into a bad one before it is turned into an optimal one. With overwhelming probability, at least $n/2$ bad triangles are created where the missing w-edge has to be included in order to be able to turn it into an optimal triangle. The probability of including a w-edge within a polynomial number of $p(m)$ steps is bounded above by $p(m) \cdot e^{-w/T}$. This is bounded below by $\Omega(m^{-k})$ only if $e^{-w/T} = \Omega(m^{-\gamma})$ for some constant $\gamma > 0$. Hence, we can assume that $T \geq w/(\gamma \cdot \ln m)$ for some constant $\gamma > 0$.

Let $p^*(T)$ be the probability of accepting the inclusion of a w-edge and $p^{**}(T)$ the corresponding probability for a $((1 + \varepsilon(m)) \cdot w)$-edge. Since $T \geq$

$w/(\gamma \cdot \ln m)$ and $\varepsilon(m) = o(1)$,

$$\begin{aligned} p^*(T)/p^{**}(T) &= e^{-w/T} \cdot e^{(1+\varepsilon(m))\cdot w/T} \\ &= e^{\varepsilon(m)\cdot w/T} \\ &\leq e^{\varepsilon(m)\cdot\gamma\cdot\ln m} \\ &= m^{\varepsilon(m)\cdot\gamma}. \end{aligned}$$

Choosing m large enough, this gets smaller than any $m^\delta, \delta > 0$. It will turn out that this advantage of w-edges against $((1+\varepsilon(m))\cdot w)$-edges is too small. The stochastic process behind MA can be described by the parameters b (number of bad triangles) and c (number of complete triangles). We use the potential function $2b+c$ which starts with the value n and has the value 0 for the MST. The value of the potential function changes in the following way:

- It increases by 1 if a complete triangle turns into a bad one or an optimal one turns into a complete one. The probability of the first event equals $2c/m$, since we have to flip one of the two light edges of one of the complete triangles. The probability of the second event equals $p^{**}(T) \cdot (n-b-c)/m$ since we have to flip the heavy edge in one of the $n-b-c$ optimal triangles and to accept this flip.
- It decreases by 1 if a complete triangle turns into an optimal one (probability c/m) or a bad triangle turns into a complete one (probability $p^*(T) \cdot b/m$).
- It remains unchanged, otherwise.

Since we are interested in lower bounds on the optimization time, we can ignore all non-accepted steps, i. e., all steps not changing the potential. If $b \leq n^{1/2}$ and m is large enough, the probability that an accepted step increases the potential is at least $3/5$. This claim is equivalent to

$$\frac{2c/m + p^{**}(T)\cdot(n-b-c)/m}{2c/m + p^{**}(T)\cdot(n-b-c)/m + c/m + p^*(T)\cdot b/m} \geq \frac{3}{5}$$

which is equivalent to

$$2c + p^{**}(T)\cdot(n-b-c) \geq \frac{9}{5}c + \frac{3}{5}\cdot p^{**}(T)\cdot(n-b-c) + \frac{3}{5}p^*(T)\cdot b$$

and

$$\frac{1}{5}c + \frac{2}{5}p^{**}(T)\cdot(n-b-c) \geq \frac{3}{5}\cdot p^*(T)\cdot b.$$

This is obviously true if $c \geq 3\cdot b$. Otherwise, $n-b-c \geq n-4b \geq n-4n^{1/2}$ and it is sufficient to show that

$$2\cdot p^{**}(T)\cdot(n-4n^{1/2}) \geq 3\cdot p^*(T)\cdot n^{1/2}$$

or

$$p^*(T)/p^{**}(T) \leq \frac{2}{3}(n^{1/2}-4).$$

We have shown that this holds for large enough m, since $n = \Omega(m)$. The claim for MA follows now from results on the gambler's ruin problem. The probability to start with a potential of $n^{1/2}/2$ and to reach the value 0 before the value $n^{1/2}$ is exponentially small. Finally, we investigate a polynomial number of $p(m)$ steps of SA. Let d be chosen such that $p(m) \leq m^d$. We claim that it is unlikely that the potential drops below $n^{1/2}/4$ within m^d steps. With overwhelming probability, we produce a bad triangle. Therefore, it is necessary to accept the inclusion of a w-edge. Hence, as seen above, only steps where the temperature is at least $w/(\gamma \cdot \ln m)$ for some appropriate constant $\gamma > 0$ have to be considered. However, the analysis of MA treats all these temperatures in the same way. The probability to start with a potential of $n^{1/2}/2$ and to reach the value $n^{1/2}/4$ before $(3/4)n^{1/2}$ is still exponentially small. □

The proof also shows that SA with an arbitrary cooling schedule is unsuccessful in the considered situation.

References

CORMEN, T.H., LEISERON, C.E. and RIVEST, R.L. (1996): *Introduction to Algorithms*. MIT Press.

DROSTE, S., JANSEN, T., and WEGENER, I. (2001): Dynamic parameter control in simple evolutionary algorithms. FOGA'2000. In: W.N. Martin, and W.M. Spears (Eds.): *Foundations of Genetic Algorithms 6*. 275–294. Morgan Kaufmann.

DROSTE, S., JANSEN, T., and WEGENER, I. (2002): *On the analysis of the (1+1) evolutionary algorithm.* Theoretical Computer Science 276, 51–81.

DROSTE, S., JANSEN, T., and WEGENER, I. (2006): *Upper and lower bounds for randomized search heuristics in black-box optimization.* Theory of Computing Systems 4, 525-544.

GIEL, O. and WEGENER, I. (2003): Evolutionary algorithms and the maximum matching problem. *Proc. of 20th Symp. on Theoretical Aspects of Computer Science (STACS), LNCS 2607*, 415–426.

GLOVER, F. and LAGUNA, M. (1993): Tabu Search. In: C. Reeves, (Ed.): *Modern Heuristic Techniques for Combinatorial Problems*. Scientific Publishing, Blackwell, 71-140.

JERRUM, M. (1992): Large cliques elude the Metropolis process. *Random Structures and Algorithms 3, 347–359.*

JERRUM, M. and SINCLAIR, A. (1996): The Markov chain Monte Carlo method. An approach to approximate counting and integration. Ch. 12 of D. Hochbaum (Ed.). *Approximation Algorithms for NP-hard Problems*, 482–522. PWS Publishing Company.

JERRUM, M. and SORKIN, G. B. (1993): Simulated annealing for graph bisection. Proc. of 37th Symp. *Foundations of Computer Science (FOCS), 94–103.*

JERRUM, M. and SORKIN, G. B. (1998): The Metropolis algorithm for graph bisection. *Discrete Applied Mathematics 82, 155–175.*

KIRKPATRICK, S., GELETT JR. C. D., and VECCHI, M. P. (1983): Optimization by Simulated Annealing. *Science 220(4598), 671–680.*

MOTWANI R. and RAGHAVAN P. (1995): *Randomized Algorithms.* Cambridge University Press.

NEUMANN, F. and WEGENER, I. (2004): Randomized local search, evolutionary algorithms, and the minimum spanning tree problem. *Proc. of Genetic and Evolutionary Computation. GECCO 2004. LNCS 3102, 713–724.*

SASAKI, G. (1991): The effect of the density of states on the Metropolis algorithm. *Information Processing Letters 37, 159–163.*

SASAKI, G. and HAJEK, B. (1988): The time complexity of maximum matching by simulated annealing. *Journal of the ACM 35, 387–403.*

SORKIN, G. B. (1991): Efficient simulated annealing on fractal energy landscapes. *Algorithmica 6, 367–418.*

WEGENER, I. (2001): Theoretical aspects of evolutionary algorithms. *Proc. of ICALP 2001. LNCS 2076, 64-78.*

WEGENER, I. (2005): Simulated annealing beats Metropolis in combinatorial optimization. *Proc. of ICALP 2005. LNCS 3580, 589-601.*

Part XV

Signal Extraction and Filtering

Filters for Short Nonstationary Sequences: The Analysis of the Business Cycle

Stephen Pollock

Department of Economics, University of Leicester
Leicester, United Kingdom, *d.s.g.pollock@le.ac.uk*

Abstract. This paper gives an account of some techniques of linear filtering which can be used for extracting the business cycle from economic data sequences of limited duration. It is argued that there can be no definitive definition of the business cycle. Both the definition of the business cycle and the methods that are used to extract it must be adapted to the purposes of the analysis; and different definitions may be appropriate to different eras.

Keywords: linear filters, spectral analysis, business cycles

1 Introduction

In recent years, there has been a renewed interest amongst economists in the business cycle. However, compared with the economic fluctuations of the nineteenth century, the business cycle in modern western economies has been a tenuous affair. For many years, minor fluctuations have been carried on the backs of strongly rising trends in national income. Their amplitudes have been so small in relative terms that they have rarely resulted in absolute reductions in the levels of aggregate income. Usually, they have succeeded only in slowing its upward progress.

Faced with this tenuous phenomenon, modern analysts have also had difficulties in reaching a consensus on how to define the business cycle and in agreeing on which methods should be used to extract it from macroeconomic data sequences. Thus, the difficulties have been both methodological and technical. This paper will deal with both of these aspects, albeit that the emphasis will be on technical matters.

It seems that many of the methodological difficulties are rooted in the tendency of economists to objectify the business cycle. If there is no doubt concerning the objective reality of a phenomenon, then it seems that it must be capable of a precise and an unequivocal definition.

However, the opinion that is offered in this paper is that it is fruitless to seek a definitive definition of the business cycle. The definition needs to be adapted to the purposes of the analysis in question; and it is arguable that it should also be influenced by the behaviour of the economy in the era that is studied.

It is also argued that a clear understanding of the business cycle can be achieved only in the light of its spectral analysis. However, the spectral approach entails considerable technical difficulties. The classical theory of statistical Fourier analysis deals with stationary stochastic sequences of unlimited duration. This accords well with the nature of the trigonometrical functions on which spectral analysis is based. In business cycle analysis, one is faced, by contrast, with macroeconomic sequences that are of strictly limited durations and that are liable to be strongly trended.

In order to apply the methods of spectral analysis to the macroeconomic data, two problems must be addressed. First, the data must be reduced to stationarity by an appropriate method of detrending. There are various ways of proceeding; and a judicious choice must be made. Then, there is the short duration of the data, which poses the problem acutely of how one should treat the ends of the sample.

One way of dealing with the end-of-sample problem is to create a circular sequence from the detrended data. By travelling around the circle indefinitely, the infinite periodic extension of the data sequence is generated, which is the essential object of an analysis that employs the discrete Fourier transform.

Such an analysis is liable to be undermined whenever there are radical disjunctions in the periodic extension at the points where the end of one replication joins the beginning of the next. Therefore, a successful Fourier analysis depends upon a careful detrending of the data. It seems that it was the neglect of this fact that led one renowned analyst to declare that spectral analysis was inappropriate to economic data. (See Granger 1966.)

2 The interaction of the trend and the business cycle

The business cycle has no fixed duration. In a Fourier analysis, it can be represented as a composite of sinusoidal motions of various frequencies that fall within some bandwidth. We shall consider one modern convention that defines the exact extent of this bandwidth; but it seems more appropriate that it should be determined the light of the data.

If they are not allowed to overlap, it may be crucial to know where the low frequency range of the trend is deemed to end and where the higher range of the business cycle should begin. However, in this section, we shall avoid the issue by assuming that the business cycle is of a fixed frequency and that the trend is a simple exponential function.

In that case, the trend can be described by the function $T(t) = \exp\{rt\}$, where $r > 0$ is constant rate of growth. The business cycle, which serves to modulate the trend, is described by an exponentiated cosine function $C(t) = \exp\{\gamma \cos(\omega t)\}$. The product of the two functions, which can regarded as a model of the trajectory of aggregate income, is

$$Y(t) = \beta \exp\{rt + \gamma \cos(\omega t)\}. \tag{1}$$

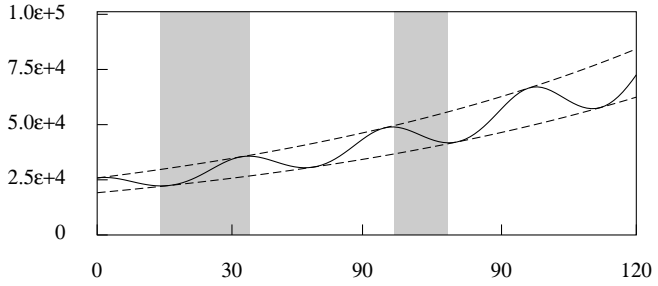

Fig. 1. The function $Y(t) = \beta \exp\{rt + \gamma \cos(\omega t)\}$ as a model of the business cycle. Observe that, when $r > 0$, the duration of an expansion exceeds the duration of a contraction.

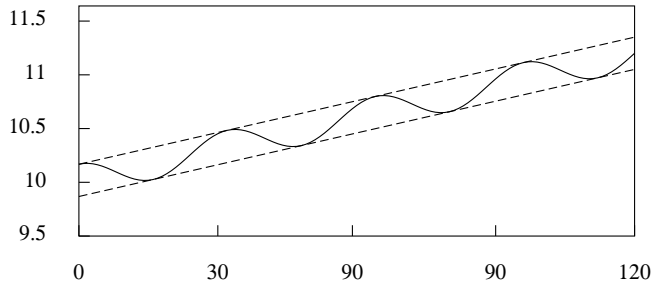

Fig. 2. The function $\ln\{Y(t)\} = \ln\{\beta\} + rt + \gamma \cos(\omega t)$ representing the logarithmic business cycle data. The duration of the expansions and the contractions are not affected by the transformation.

The resulting business cycles, which are depicted in Figure 1, have an asymmetric appearance. Their contractions are of lesser duration than their expansions; and they become shorter as the growth rate r increases.

Eventually, when the rate exceeds a certain value, the periods of contraction will disappear and, in place of the local minima, there will be only points of inflection. In fact, the condition for the existence of local minima is that $\omega\gamma > r$, which is to say that the product of the amplitude of the cycles and their angular velocity must exceed the growth rate of the trend.

Next, we take logarithms of the data to obtain a model, represented in Figure 2, that has additive trend and cyclical components. This gives

$$\ln\{Y(t)\} = y(t) = \mu + rt + \gamma \cos(\omega t), \tag{2}$$

where $\mu = \ln\{\beta\}$. Since logs effect a monotonic transformation, there is no displacement of the local maxima and minima. However, the amplitude of the fluctuations around the trend, which has become linear in the logs, is now constant.

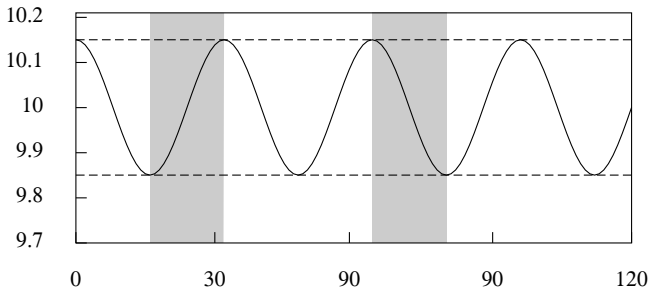

Fig. 3. The function $\mu + \gamma\cos(\omega t)$ representing the detrended business cycle. The duration of the expansions and the contractions are equal.

The final step is to create a stationary function by eliminating the trend. There are two equivalent ways of doing this in the context of the schematic model. On the one hand, the linear trend $\xi(t) = \mu + rt$ can be subtracted from $y(t)$ to create the pure business cycle $\gamma\cos(\omega t)$.

Alternatively, the function $y(t)$ can be differentiated to give $dy(t)/dt = r - \gamma\omega\sin(\omega t)$. When the latter is adjusted by subtracting the growth rate r, by dividing by ω and by displacing its phase by $-\pi/2$ radians—which entails replacing the argument t by $t - \pi/2$—we obtain the function $\gamma\cos(\omega t)$ again. Through the process of detrending, the phases of expansion and contraction acquire equal duration, and the asymmetry of the business cycle vanishes.

There is an enduring division of opinion, in the literature of economics, on whether we should be looking at the turning points and phase durations of the original data or at those of the detrended data. The task of finding the turning points is often a concern of analysts who wish to make international comparisons of the timing of the business cycle.

However, since the business cycle is a low-frequency component of the data, it is difficult to find the turning points with great accuracy. In fact, the pinnacles and pits that are declared to be the turning points often seem to be the products of whatever high-frequency components happen to remain in the data after it has been subjected to a process of seasonal adjustment.

If the objective is to compare the turning points of the cycles, then the trends should be eliminated from the data. The countries that might be compared are liable to be growing at differing rates. From the trended data, it will appear that those with higher rates of growth have shorter recessions with delayed onsets, and this can be misleading.

The various indices of an expanding economy will also grow at diverse rates. Unless they are reduced to a common basis by eliminating their trends, their fluctuations cannot be compared easily. Amongst such indices will be the percentage rate of unemployment, which constitutes a trend-stationary sequence. It would be difficult to collate the turning points in this index with those within a rapidly growing series of aggregate income, which might

not exhibit any absolute reductions in its level. A trenchant opinion to the contrary, which opposes the practice of detrending the data for the purposes of describing the business cycle, has been offered by Harding and Pagan (2003).

3 The bandpass definition of the business cycle

The modern definition of the business cycle that has been alluded to in the previous section is that of a quasi cyclical motion comprising sinusoidal elements that have durations of no less than one-and-a-half years and not exceeding eight years.

This definition has been proposed by Baxter and King (1999) who have declared that it was the definition adopted by Burns and Mitchell (1947) in their study of the economic fluctuations in the U.S. in the late nineteenth century and in the early twentieth century. However, it is doubtful whether Burns and Mitchell were so firm in their definition of what constitutes the business cycle. It seems, instead, that they were merely speaking of what they had discerned in their data.

The definition in question suggests that the data should be filtered in order to extract from it the components that fall within the stated range, which is described as the pass band. Given a doubly infinite data sequence, this objective would be fulfilled, in theory, by an ideal bandpass filter comprising a doubly infinite sequence of coefficients.

The ideal bandpass filter that transmits all elements within the frequency range $[\alpha, \beta]$ and blocks all others has the following frequency response:

$$\psi(\omega) = \begin{cases} 1 \ \text{ if } |\omega| \in (\alpha, \beta), \\ 0, \text{ otherwise.} \end{cases} \tag{3}$$

The coefficients of the corresponding time-domain filter are obtained by applying an inverse Fourier transform to this response to give

$$\psi_k = \int_{\alpha}^{\beta} e^{\mathrm{i}k\omega} d\omega = \frac{1}{\pi k}\{\sin(\beta k) - \sin(\alpha k)\}. \tag{4}$$

In practice, all data sequences are finite, and it is impossible to apply a filter that has an infinite number of coefficients. However, a practical filter may be obtained by selecting a limited number of the central coefficients of an ideal infinite-sample filter. In the case of a truncated filter based on $2q+1$ central coefficients, the elements of the filtered sequence are given by

$$\begin{aligned} x_t = \psi_q y_{t-q} + \psi_{q-1} y_{t-q+1} + \cdots + \psi_1 y_{t-1} + \psi_0 y_t \\ + \psi_1 y_{t+1} + \cdots + \psi_{q-1} y_{t+q-1} + \psi_q y_{t+q}. \end{aligned} \tag{5}$$

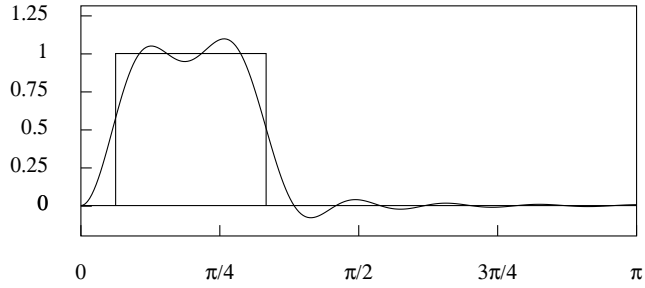

Fig. 4. The frequency response of the truncated bandpass filter of 25 coefficients superimposed upon the ideal frequency response. The lower cut-off point is at $\pi/15$ radians (11.25°), corresponding to a period of 6 quarters, and the upper cut-off point is at $\pi/3$ radians (60°), corresponding to a period of the 32 quarters.

Given a sample $y_0, y_1, \ldots, y_{T-1}$ of T data points, only $T-2q$ processed values $x_q, x_{q+1} \ldots, x_{T-q-1}$ are available, since the filter cannot reach the ends of the sample, unless it is extrapolated.

If the coefficients of the truncated bandpass or highpass filter are adjusted so that they sum to zero, then the z-transform polynomial $\psi(z)$ of the coefficient sequence will contain two roots of unit value. The adjustments may be made by subtracting $\sum_k \phi_k/(2q+1)$ from each coefficient. The sum of the adjusted coefficients is $\psi(1) = 0$, from which it follows that $1 - z$ is a factor of $\psi(z)$. The condition of symmetry, which is that $\psi(z) = \psi(z^{-1})$, implies that $1 - z^{-1}$ is also a factor. Thus the polynomial contains the factor

$$(1 - z)(1 - z^{-1}) = -z^{-1}(1 - z)^2, \tag{6}$$

within which $\nabla^2(z) = (1-z)^2$ corresponds to a twofold differencing operator.

Since it incorporates the factor $\nabla^2(z)$, the effect of applying the filter to a data sequence with a linear trend will be to produce an untrended sequence with a zero mean. The effect of applying it to a sequence with a quadratic trend will be to produce an untrended sequence with a nonzero mean.

The usual effect of the truncation will be to cause a considerable spectral leakage. Thus, if the filter is applied to trended data, then it is liable to transmit some powerful low-frequency elements that will give rise to cycles of high amplitudes within the filtered output. The divergence of the frequency response function from the ideal specification of (3) is illustrated in Figure 4.

An indication of the effect of the truncated filter is provided by its application to a quarterly sequence of the logarithms of consumption in the U.K. that is illustrated in Figure 5. The filtered sequence is in Figure 6, where the loss of the data from the ends is indicated by the vertical lines.

An alternative filter that is designed to reach the ends of the sample has been proposed by Christiano and Fitzgerald, (2003). The filter is described

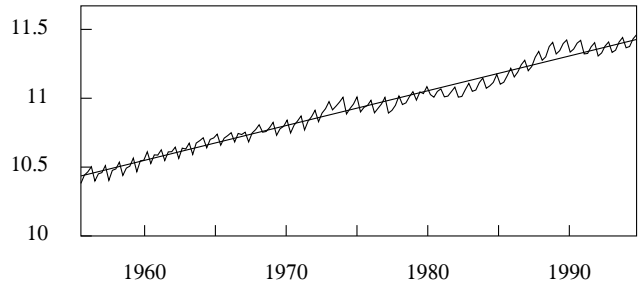

Fig. 5. The quarterly sequence of the logarithms of consumption in the U.K., for the years 1955 to 1994, together with a linear trend interpolated by least-squares regression.

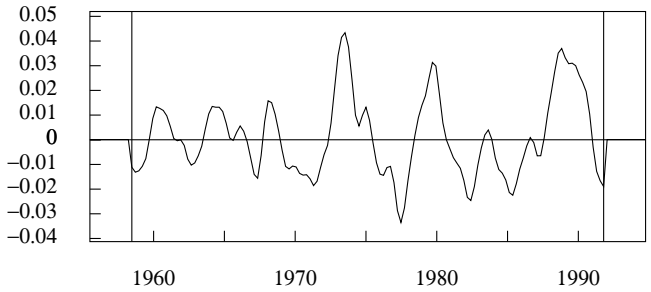

Fig. 6. The sequence derived by applying the truncated bandpass filter of 25 coefficients to the quarterly logarithmic data on U.K. Consumption.

by the equation

$$x_t = Ay_0 + \psi_t y_0 + \cdots + \psi_1 y_{t-1} + \psi_0 y_t + \psi_1 y_{t+1} + \cdots + \psi_{T-1-t} y_{T-1} + By_{T-1}. \tag{7}$$

This equation comprises the entire data sequence $y_0, \ldots, y_{T-1}$; and the value of t determines which of the coefficients of the infinite-sample filter are entailed in producing the current output. Thus, the value of x_0 is generated by looking forwards to the end of the sample, whereas the value of x_{T-1} is generated by looking backwards to the beginning of the sample.

If the process generating the data is stationary and of zero mean, then it is appropriate to set $A = B = 0$, which is tantamount to approximating the extra-sample elements by zeros. In the case of a data sequence that appears to follow a first-order random walk, it has been proposed to set A and B to the values of the sums of the coefficients that lie beyond the span of the data on either side. Since the filter coefficients must sum to zero, it follows that

$$A = -(\frac{1}{2}\psi_0 + \psi_1 + \cdots + \psi_t) \quad \text{and} \quad B = -(\frac{1}{2}\psi_0 + \psi_1 + \cdots + \psi_{T-t-1}). \tag{8}$$

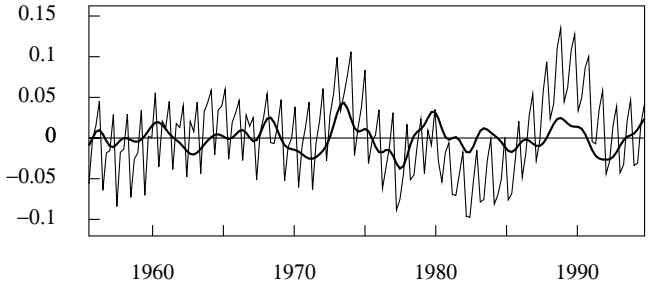

Fig. 7. The sequence derived by applying the bandpass filter of Christiano and Fitzgerald to the quarterly logarithmic data on U.K. Consumption.

The effect is tantamount to extending the sample at either end by constant sequences comprising the first and the last sample values respectively.

For data that have the appearance of having been generated by a first-order random walk with a constant drift, it is appropriate to extract a linear trend before filtering the residual sequence. In fact, this has proved to be the usual practice in most circumstances.

It has been proposed to subtract from the data a linear function $f(t) = \alpha + \beta t$ interpolated through the first and the final data points, such that $\alpha = y_0$ and $\beta = (y_{T-1} - y_0)/T$. In that case, there should be $A = B = 0$. This procedure is appropriate to seasonally adjusted data. For data that manifest strong seasonal fluctuations, such as the U.K. consumption data, a line can be fitted by least squares through the data points of the first and the final years. Figure 7, shows the effect of the application of the filter to the U.K. data adjusted in this manner.

The filtered sequence of Figure 7 has much the same profile in its middle section as does the sequence of Figure 6, which is derived by applying truncated bandpass filter. (The difference in the scale of the two diagrams tends to conceal this similarity.) However, in comparing filtered sequence to the adjusted data, it seems fair to say that it fails adequately to represent the prominent low-frequency fluctuations. It is also beset by some noisy high-frequency fluctuations that would not normally be regarded as part of the business cycle.

4 Polynomial detrending

The problems besetting the filtered sequence can be highlighted with reference to the periodogram of the residuals that are obtained by interpolating a polynomial trend line thorough the logarithmic data. Therefore, it is appropriate, at this juncture, to derive a formula for polynomial regression.

Therefore, let $L_T = [e_1, e_2, \ldots, e_{T-1}, 0]$ be the matrix version of the lag operator, which is formed from the identity matrix $I_T = [e_0, e_1, e_2, \ldots, e_{T-1}]$

of order T by deleting the leading column and by appending a column of zeros to the end of the array. The matrix that takes the p-th difference of a vector of order T is

$$\nabla_T^p = (I - L_T)^p. \tag{9}$$

We may partition this matrix so that $\nabla_T^p = [Q_*, Q]'$, where Q_*' has p rows. If y is a vector of T elements, then

$$\nabla_T^p y = \begin{bmatrix} Q_*' \\ Q' \end{bmatrix} y = \begin{bmatrix} g_* \\ g \end{bmatrix}; \tag{10}$$

and g_* is liable to be discarded, whereas g will be regarded as the vector of the p-th differences of the data.

The inverse matrix, which corresponds to the summation operator, is partitioned conformably to give $\nabla_T^{-p} = [S_*, S]$. It follows that

$$\begin{bmatrix} S_* & S \end{bmatrix} \begin{bmatrix} Q_*' \\ Q' \end{bmatrix} = S_* Q_*' + S Q' = I_T, \tag{11}$$

and that

$$\begin{bmatrix} Q_*' \\ Q' \end{bmatrix} \begin{bmatrix} S_* & S \end{bmatrix} = \begin{bmatrix} Q_*'S_* & Q_*'S \\ Q'S_* & Q'S \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & I_{T-p} \end{bmatrix}. \tag{12}$$

If g_* is available, then y can be recovered from g via $y = S_* g_* + Sg$.

The lower-triangular Toeplitz matrix $\nabla_T^{-p} = [S_*, S]$ is completely characterised by its leading column. The elements of that column are the ordinates of a polynomial of degree $p-1$, of which the argument is the row index $t = 0, 1, \ldots, T-1$. Moreover, the leading p columns of the matrix ∇_T^{-p}, which constitute the submatrix S_*, provide a basis for all polynomials of degree $p-1$ that are defined on the integer points $t = 0, 1, \ldots, T-1$.

A polynomial of degree $p-1$, represented by its ordinates in the vector f, can be interpolated through the data by minimising the criterion

$$(y - f)'(y - f) = (y - S_* f_*)'(y - S_* f_*) \tag{13}$$

with respect to f_*. The resulting values are

$$f_* = (S_*'S_*)^{-1} S_*' y \quad \text{and} \quad f = S_*(S_*'S_*)^{-1} S_*' y. \tag{14}$$

An alternative representation of the estimated polynomial is available, which is provided by the identity

$$S_*(S_*'S_*)^{-1} S_*' = I - Q(Q'Q)^{-1} Q'. \tag{15}$$

It follows that the polynomial fitted to the data by least-squares regression can be written as

$$f = y - Q(Q'Q)^{-1} Q' y. \tag{16}$$

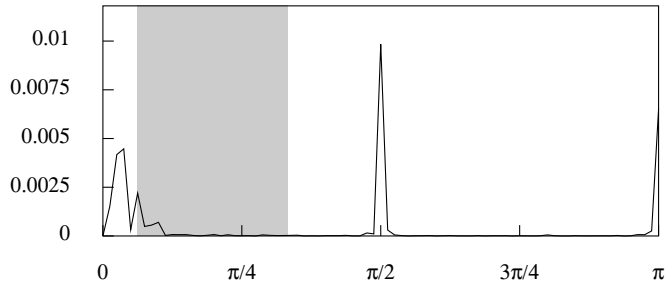

Fig. 8. The periodogram of the residual sequence obtained from the linear detrending of the logarithmic consumption data. A band, with a lower bound of $\pi/16$ radians and an upper bound of $\pi/3$ radians, is masking the periodogram.

A more general method of curve fitting, which embeds polynomial regression as a special case, is one that involves the minimisation of a combination of two sums of squares. Let f denote the vector of fitted values. Then, the criterion for finding the vector is to minimise

$$L = (y - f)'(y - f) + f'Q\Lambda Q'f. \tag{17}$$

The first term penalises departures of the resulting curve from the data, whereas the second term imposes a penalty for a lack of smoothness in the curve. The second term comprises $d = Q'f$, which is the vector of pth-order differences of f. The matrix Λ serves to generalise the overall measure of the curvature of the function that has the elements of f as its sampled ordinates, and it serves to regulate the penalty for roughness, which may vary over the sample.

Differentiating L with respect to f and setting the result to zero, in accordance with the first-order conditions for a minimum, gives

$$(y - f) = Q\Lambda Q'f = Q\Lambda d. \tag{18}$$

Multiplying the equation by Q' gives $Q'(y - f) = Q'y - d = Q'Q\Lambda d$, whence $\Lambda d = (\Lambda^{-1} + Q'Q)^{-1}Q'y$. Putting this into the equation $f = y - Q\Lambda d$ gives

$$f = y - Q(\Lambda^{-1} + Q'Q)^{-1}Q'y. \tag{19}$$

If $\Lambda^{-1} = 0$ in (19), and if Q' is the matrix version of the twofold difference operator, then the least-squares interpolator of a linear function is derived in the form equation (16). The sequence of regression residuals will be given by the vector $r = Q(Q'Q)^{-1}Q'y$; and it is notable that these residuals contain exactly the same information as the vector $g = Q'y$ of the twofold differences of the data. However, whereas the low-frequency structure would be barely visible in the periodogram of the differenced data, it will be fully evident in the periodogram of the residuals of a ploynomial regression.

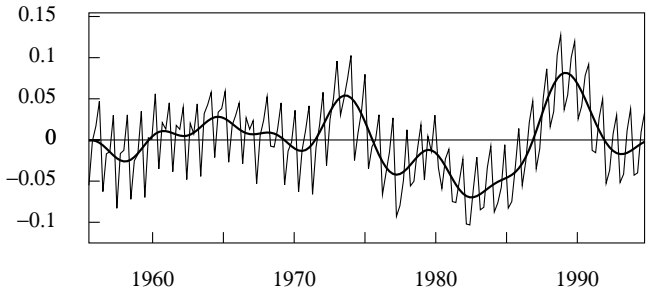

Fig. 9. The residual sequence from fitting a quadratic trend to the logarithmic consumption data. The interpolated line, which represents the business cycle, has been synthesised from the Fourier ordinates in the frequency interval $[0, \pi/8]$.

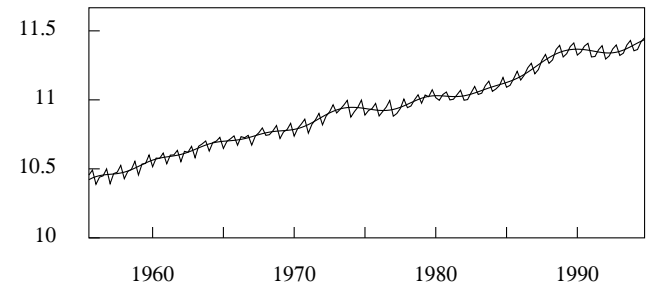

Fig. 10. The trend/cycle component of U.K. Consumption determined by the Fourier method, superimposed on the logarithmic data

The periodogram of the residual sequence obtained from a linear detrending of the logarithmic consumption data is presented in Figure 8. Superimposed upon the figure is a highlighted band that spans the interval $[\pi/16, \pi/3]$, which corresponds to the nominal pass band of the filters applied in the previous section.

Within this periodogram, the spectral structure extending from zero frequency up to $\pi/8$ belongs to the business cycle. The prominent spikes located at the frequency $\pi/2$ and at the limiting Nyquist frequency of π are property of the seasonal fluctuations. Elsewhere in the periodogram, there are wide dead spaces, which are punctuated by the spectral traces of minor elements of noise. The highlighted pass band omits much of the information that might be used in synthesising the business cycle.

5 The synthesis of the business cycle

To many economists, it seems implausible that the trend of a macroeconomic index, which is the product of events within the social realm, should be mod-

elled by polynomial, which may be described as a deterministic function. A contrary opinion is represented in this paper. We deny the objective reality of the trend. Instead, we consider it to be the product of our subjective perception of the data. From this point of view, a polynomial function can often serve as a firm benchmark against which to measure the fluctuations of the index. Thus, the linear trend that we have interpolated through the logarithms of the consumption data provides the benchmark of constant exponential growth.

It is from the residuals of a log-linear detrending of the consumption data that we wish to extract the business cycle. The appropriate method is to extract the Fourier components of the residual sequence that lie within the relevant frequency band. Reference to Figure 8 suggests that this band should stretch from zero up to the frequency of $\pi/8$ radians per quarter, which corresponds to a cycle with a duration of 4 years. In Figure 9, the sequence that is synthesised from these Fourier ordinates has been superimposed upon the sequence of the residuals of the linear detrending.

To provide a symbolic representation of the method, we may denote the matrix of the discrete Fourier transform and its inverse by

$$\begin{aligned} U &= T^{-1/2}[\exp\{-\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1], \\ \bar{U} &= T^{-1/2}[\exp\{\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1], \end{aligned} \tag{20}$$

Then, the residual vector $r = Q(Q'Q)^{-1}Q'y$ and its Fourier transform ρ are represented by

$$r = T^{1/2}\bar{U}\rho \quad \longleftrightarrow \quad \rho = T^{-1/2}Ur. \tag{21}$$

Let J be a matrix of which the elements are zeros apart from a string of units on the diagonal, which serve to select from ρ the requisite Fourier ordinates within the band $[0, \pi/8]$. Then, the filtered vector that represents the business cycle is given by

$$x = T^{1/2}\bar{U}J\rho = \{\bar{U}JU\}r = \Psi r. \tag{22}$$

Here, $\bar{U}JU = \Psi = [\psi^{\circ}_{|i-j|}; i, j = 0, \ldots, T-1]$ is a circulant matrix of the filter coefficients that would result from wrapping the infinite sequence of the ideal bandpass coefficients around a circle of circumference T and adding the overlying elements. Thus

$$\psi^{\circ}_k = \sum_{q=-\infty}^{\infty} \psi_{qT+k}. \tag{23}$$

Applying the wrapped filter to the finite data sequence via a circular convolution is equivalent to applying the original filter to an infinite periodic extension of the data sequence. In practice, the wrapped coefficients would be obtained from the Fourier transform of the vector of the diagonal elements of the matrix J.

The Fourier method can also be exploited to create a sequence that represents a combination of the trend and the business cycle. There are various ways of proceeding. One of them is to add the vector x to that of the linear or polynomial trend that has generated the sequence of residuals. An alternative method is to obtain the trend/cycle component by subtracting its complement from the data vector.

The complement of the trend/cycle component is a stationary component. Since a Fourier method can be applied only to a stationary vector, we are constrained to work with the vector $g = Q'y$, obtained by taking the twofold differences of the data.

Since the twofold diffencing entails the loss of two points, the vecotor g may be supplemented by a point at the beginning and a point at the end. The resulting vector may be denoted by q. The relevant Fourier ordinates are extracted by applying the selection matrix $I - J$ to the transformed vector $\gamma = Uq$. Thereafter, they need to be re-inflated to compensate for the differencing operation.

The frequency response of the twofold difference operator, which is obtained be setting $z = \exp\{-\mathrm{i}\omega\}$ in equation (6), is

$$f(\omega) = 2 - 2\cos(\omega), \tag{24}$$

and that of the anti-differencing operation is the inverse $1/f(\omega)$. The Fourier ordinates of a differenced vector will be reinflated by pre-multiplying their vector by the diagonal matrix $V = \mathrm{diag}\{v_0, v_1, \ldots, v_{T-1}\}$, which comprises the values $v_j = 1/f(\omega_j); j = 0, \ldots, T-1$, where $\omega_j = 2\pi j/T$.

The matrix that is to be applied to the Fourier ordinates of the differenced data is therefore $H = V(I - J)$. The resulting vector is transformed back to the time domain via the matrix $\bar{U}$ to produce the vector that is to be subtracted from the data vector y. The resulting estimate of the trend/cycle component is

$$z = y - \bar{U}HUq. \tag{25}$$

This is represented in Figure 10.

6 More flexible methods of detrending

Methods of detrending may be required that are more flexible than the polynomial interpolations that we have we considered so far. For a start, there is a need to minimise the disjunctions that occur in the periodic extension of the data sequence where the end of one replication joins the beginning of the next. This purpose can be served by a weighted version of a least-squares polynomial regression. If extra weight is given to the data points at the beginning and the end of the sample, then the interpolated line can be constrained pass through their midst; and, thereby, a major disjunction can be avoided.

The more general method of trend estimation that is represented by equation (19) can also be deployed. By setting $\Lambda^{-1} = \lambda^{-1}I$, a familiar filtering

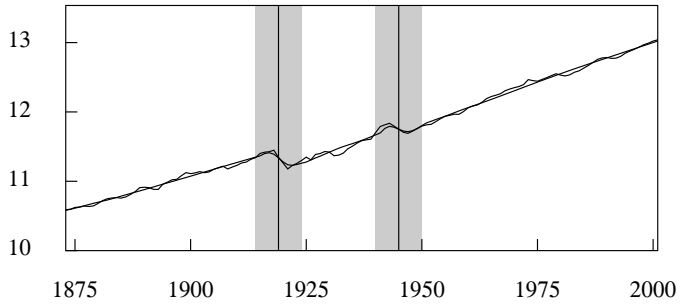

Fig. 11. The logarithms of annual U.K. real GDP from 1873 to 2001 with an interpolated trend. The trend is estimated via a filter with a variable smoothing parameter.

device is obtained that has been attributed by economists to Hodrick and Prescott (1980, 1997). In fact, an earlier exposition this filter was provided by Leser (1961), and its essential details can be found in a paper of Whittaker (1923).

The effect of the Hodrick–Prescott (H–P) filter depends upon the value of the smoothing parameter λ. As the value of the parameter increases, the vector f converges upon that of a linear trend. As the value of λ tends to zero, f converges to the data vector y. The effect of using the more flexible H–P trend in place of a linear trend is to generate estimates of the business cycle fluctuations that have lesser amplitudes and a greater regularity.

The enhanced regularity of the fluctuations is a consequence of the removal from the residual sequence of a substantial proportion of the fluctuations of lowest frequency, which can cause wide deviations from the line. This enhancement might be regarded as a spurious. However, it can be argued that such low-frequency fluctuations are liable to escape the attention of many economic agents, which is a reason for excluding them from a representation of the business cycle.

Whereas the H–P filter employs a globally constant value for the λ, it is possible to vary this parameter over the course of the sample. This will allow the trend to absorb the structural breaks or disturbances that might occasionally interrupt the steady progress of the economy. If it can be made to absorb the structural breaks, then the trend will not be thrown off course for long; and, therefore, it should serve as a benchmark against which to measure the cyclical variations when the economy resumes its normal progress. At best, the residual sequence will serve to indicate how the economy might have behaved in the absence of the break.

Figure 11 shows a trend function that has been fitted, using a variable smoothing parameter, to the logarithms of a sequence of annual data on real U.K. gross domestic product that runs from 1873 to 2001. Only the breaks

after the ends of the first and second world wars have been accommodated, leaving the disruptions of the 1929 recession to be expressed in the residual sequence. The effect has been achieved by attributing a greatly reduced value to the smoothing parameter in the vicinity of the post-war breaks. In the regions that are marked by shaded bands, the smoothing parameter has been given a value of 5. Elsewhere, it has been given a high value of 100,000, which results in trend segments that are virtually linear.

This example serves to illustrate the contention that the trend and the accompanying business cycle are best regarded as subjective concepts. The intention of the example is to remove from the residual sequence—and, therefore, from the representation of business cycle—the effects of two major economic disruptions. For the purpose of emphasising the extent of these disruptions, the contrary approach of fitting a stiff polynomial trend line through the data should be followed.

References

BAXTER, M. and KING, R.G. (1999), Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series, *Review of Economics and Statistics,* 81, 575–593.

BEVERIDGE, W.H. (1921), Weather and Harvest Cycles, *Economic Journal,* 31, 429–452.

BEVERIDGE, W.H. (1922), Wheat Prices and Rainfall in Western Europe, *Journal of the Royal Statistical Society,* 85, 412–478.

BURNS, A.M. and MITCHELL, W.C. (1947), *Measuring Business Cycles,* National Bureau of Economic Research, New York.

CHRISTIANO, L.J. and FITZGERALD, T.J.(2003), The Band-pass Filter, *International Economic Review,* 44 , 435–465.

GRANGER, C.W.J. (1966), The Typical Spectral Shape of an Economic Variable, *Econometrica,* 34, 150–161.

HARDING, D. and PAGAN , A. (2002), Dissecting the Cycle: A Methodological Investigation, *Journal of Monetary Economics.* 49, 365–381.

HODRICK, R.J. and PRESCOTT, E.C. (1980, 1997), *Postwar U.S. Business Cycles: An Empirical Investigation,* Working Paper, Carnegie–Mellon University, Pittsburgh, Pennsylvania. Published in 1997 *Journal of Money, Credit and Banking,* 29, 1–16.

LESER, C.E.V. (1961), A Simple Method of Trend Construction, *Journal of the Royal Statistical Society, Series B,* 23, 91–107.

WHITTAKER, E.T. (1923), On a New Method of Graduation, *Proceedings of the Royal Society of Edinburgh,* 44, 77–83.

Estimation of Common Factors Under Cross-Sectional and Temporal Aggregation Constraints: Nowcasting Monthly GDP and Its Main Components

Tommaso Proietti

SEFEMEQ, Università di Roma "Tor Vergata", Via Columbia 2, 00133 Roma, Italy, *tommaso.proietti@uniroma2.it*

Abstract. The paper estimates a large-scale mixed-frequency dynamic factor model for the euro area, using monthly series along with Gross Domestic Product (GDP) and its main components, obtained from the quarterly national accounts. Our model is a traditional dynamic factor model formulated at the monthly frequency in terms of the stationary representation of the variables, which however becomes nonlinear when the observational constraints are taken into account.

Keywords: non linear state space models, temporal disaggregation, nonlinear smoothing, monthly GDP, chain-linking

1 Introduction

Large scale factor models aim at extracting the main economic signals from a very large number of time series. The underlying idea is that the comovements among economic time series can be traced to a limited number of common factors. Factor models have been used in an increasing number of applications. The two most prominent areas are the construction of synthetic indicators, such as coincident indicators of real economic activity (Forni et al., 2000, 2001) and core inflation (Cristadoro et al., 2005), and forecasting macroeconomic variables (Stock and Watson, 2002a, Forni et al. 2005), in which case the information contained in large number of economic indicators is summarized in a few latent factors, which are then employed to forecast real output growth, or inflation. Other areas of applications are surveyed in Stock and Watson (2006).

The information set used for the estimation of factor models in typical applications is strongly unbalanced towards the series collected from the supply side of the economy, that is establishments surveys (e.g., industrial production and turnover, retail sales, financial statistics), and from administrative records (e.g. building permits and car registration). Important information from other institutional units and economic agents, namely households, is missed out, just because the underlying measurement process is more complex, with the consequence that the information becomes available with larger

delays. Notable examples are the Labor Force Survey and the Consumer Expenditure Surveys, which are carried out by the euro area member state and constitute essential sources for the labor market and consumption.

On the other hand, national accounts (NA) statistics provide a comprehensive and detailed record of the economic activities taking place within an economy, which are translated into a set of coherent and integrated measures of economic activity. The most comprehensive measure is provided by Gross Domestic Product (GDP); furthermore, the aggregates that arise from its decomposition according to the expenditure and the output approach (e.g. final consumption, gross capital formation, sectorial value added) are among the most relevant economic statistics for purposes of macroeconomic analysis and policy-making.

Hence, the NA aggregates can be considered as aggregate indicators of economic activity based on a set of definitions, concepts, classifications and accounting rules that are internationally agreed. The main problem is their observation frequency, which at present is quarterly for the euro area, and their timeliness, i.e., the fact that they are made available with considerable delay. A related point is that they are first released as preliminary estimates and then revised as new information accrues.

The aim of this paper is to estimate a large scale factor model of the euro area economy which combines the monthly information carried by a number of economic indicators (concerning industrial production, construction, retail sales, financial intermediation, employment and wages, exchange rates, external trade and business and consumer surveys) with the quarterly national accounts series. In particular, we consider a panel of 149 series, referring to the euro area for the period from January 1995 to June 2007, 17 of which are NA series and concern quarterly real GDP and its breakdown according to the expenditure and the output approaches. The presence of these series raises the fundamental issue of incorporating the observational constraints into the estimation process. The issue has two facets, the first being temporal aggregation and the second being contemporaneous aggregation. As far as the former is concerned, the factor model is specified in terms of the stationary representation of the series; our series can be taken to be stationary in terms of the logarithmic change with respect to the previous month (assuming that all are nonseasonal or seasonally adjusted). For the NA series the monthly changes are unobserved. What we observe are the quarterly totals, i.e. the sum of the levels of the three months making up the quarter. This simple fact renders the observational constraint nonlinear. Secondly, the NA series are subject to accounting identities that, due to chain linking, hold when the data are expressed at the prices of the previous year (see Eurostat, 1999, Bloem, et. al, 2001). This again makes the cross-sectional constraints nonlinear.

The introduction of the NA series in the model can be considered as the main contribution of this paper. Their consideration is essential to improve

the coverage of the economy and the representativeness of the factors. Secondly, as a by product our model produces nowcasts of monthly GDP and its components, along with measures of their reliability. Not only the factor estimates will benefit from the inclusion of GDP and its components, but also the disaggregate estimates of GDP will embody a large information set.

2 Description of the dataset

The available data consist of 132 monthly and 17 quarterly time series (i.e. a total of 149 series) for the period starting in January (1st quarter of) 1995 and ending in June (second quarter) of 2006, for a total of 150 monthly observation (38 quarterly observations). The series, extracted from the Europa database (http://epp.eurostat.ec.europa.eu/), can be grouped under the following main headings.

National accounts: 17 quarterly time series concerning the euro area GDP and its main components, the breakdown of total GDP by the output the expenditure approaches. All the series are expressed in millions of euro, chain-linked volumes, reference year 2000.
Industry: 53 monthly time series.
Construction: 7 monthly time series.
Retail Trade: 28 monthly time series.
Monetary and Financial indicators: 13 monthly time series.
Labour market: 5 monthly time series.
Business and consumer surveys: 22 monthly time series.

3 The Complete Data Factor Model

Let us suppose that the N time series are fully available and let us denote the individual time series in the original scale of measurement by $Y_{it}, i = 1, \ldots, N, t = 0, 1, \ldots, n$. We also assume that the series can be rendered stationary by the transformation $y_{it} - \varphi_i y_{i,t-1}, t = 1, \ldots, n$, where y_{it} is the Box-Cox transformation (Box and Cox, 1964) with parameter λ_i of the original series,

$$y_{it} = \begin{cases} \frac{Y_{it}^{\lambda_i} - 1}{\lambda_i}, \lambda_i \neq 0, \\ \ln Y_{it}, \quad \lambda_i = 0, \end{cases}$$

and $\varphi_i = 1$ if the series is difference stationary and 0 otherwise. For the series considered in our application, we can assume that the monthly logarithmic changes are stationary, so that $\lambda_i = 1$ and $\varphi_i = 1$, except for the Business and Consumer Survey series, for which $\lambda_i = 0$ and $\varphi_i = 1$.

The factor model that we formulate for the complete monthly series (i.e., the model that would be entertained if a complete set of N monthly time series were available) is a standard dynamic factor model, according to which

the series are conditionally independent, given a set of common factors. The common factors are generated by a stationary first order vector autoregressive process. The model for the i-th time series is formulated as follows:

$$\begin{aligned} y_{it} &= \varphi_i y_{i,t-1} + \mu_i + \sigma_i x_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, n, \\ x_{it} &= \boldsymbol{\theta}_i' \mathbf{f}_t + \xi_{it}, \qquad \xi_{it} \sim \mathrm{NID}(0, \psi_i), \\ \mathbf{f}_t &= \boldsymbol{\Phi} \mathbf{f}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathrm{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\eta); \end{aligned} \tag{1}$$

here μ_i represents the mean of the stationary transformation $y_{it} - \varphi_i y_{i,t-1}$, σ_i is its standard deviation, and x_{it} is the standardized stationary transformation of the original time series. The latter is expressed as a linear combination of K stationary common factors, $\mathbf{f}_t$, with zero mean, with weights collected in the $K \times 1$ vector $\boldsymbol{\theta}_i$ (factor loadings), plus an idiosyncratic component, ξ_{it}. The idiosyncratic component is orthogonal to the factors.

If we further let $\Delta y_{it} = y_{it} - \varphi_i y_{i,t-1}$ and $\Delta \mathbf{y}_t$ denote the stack of the stationary series, $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]'$, $\mathbf{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$, and similarly $\mathbf{x}_t = [x_{1t}, \ldots, x_{Nt}]'$, we can write $\Delta \mathbf{y}_t = \boldsymbol{\mu} + \mathbf{D}\mathbf{x}_t$, and the model for $\mathbf{x}_t$ has state space representation:

$$\begin{aligned} \mathbf{x}_t &= \boldsymbol{\Theta} \mathbf{f}_t + \xi_t, \qquad \xi_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}) \\ \mathbf{f}_t &= \boldsymbol{\Phi} \mathbf{f}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathrm{NID}(\mathbf{0}, \boldsymbol{\Sigma}_\eta) \end{aligned} \tag{2}$$

where $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N]'$ and $\boldsymbol{\Psi} = \mathrm{diag}\{\psi_1, \ldots, \psi_N\}$, $\mathrm{E}(\xi_t \boldsymbol{\eta}_t') = \mathbf{0}$ and $\mathbf{f}_0 \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_f)$, where $\boldsymbol{\Sigma}_f$ satisfies the matrix equation $\boldsymbol{\Sigma}_f = \boldsymbol{\Phi} \boldsymbol{\Sigma}_f \boldsymbol{\Phi}' + \boldsymbol{\Sigma}_\eta$.

As it is well known, the factor model is identified up to an invertible $K \times K$ matrix. A unique solution is obtained by imposing K^2 restrictions. We identify our factor model using the restriction that the upper $K \times K$ block of the loadings matrix is equal to the identity matrix, that is $\boldsymbol{\Theta} = [\mathbf{I}_K, \boldsymbol{\Theta}^{*'}]'$. The restriction exactly identifies the model; see Geweke and Singleton (1981), proposition 2.

Let us define the parameter vector $\boldsymbol{\Xi} = [(\boldsymbol{\Theta}^*)', (\boldsymbol{\Phi})', \mathrm{vech}(\boldsymbol{\Sigma}_\eta), \psi_1, \ldots, \psi_N]'$. For small N the parameters can be estimated by maximum likelihood, where the likelihood is evaluated by the Kalman filter (KF) via the prediction error decomposition, using a numerical quasi-Newton method. An application is Stock and Watson (1991). With large N, the evaluation of the likelihood is still efficiently performed by the KF; however the difficulty with maximising the likelihood via gradiend based methods is due to the high dimensionality of $\boldsymbol{\Xi}$, which has $NK + N + K^2$ unrestricted elements. In our application, in which $N = 149$ and $K = 6$, the number of unrestricted parameters is 1079.

A computationally viable alternative is to use the Expectation- Maximization (EM) algorithm of Dempster et al. (1977). The EM algorithm for state space models was introduced by Shumway and Stoffer (1982). For N large, an alternative asymptotically equivalent estimation strategy is to use principal components analysis, when we allow the number of time series N, or both N and n, to go to infinity.

4 Temporal aggregation

The N time series y_{it} are available at different frequencies of observation. In particular, the first block of $N_1 = 17$ time series, GDP and its main components, are quarterly. Since $Y_{it}, 1, \ldots, N_1$, is subject to temporal aggregation, we observe the quarterly totals:

$$Y_{i\tau} = \sum_{i=1}^{3} Y_{i,3\tau-i}, \quad \tau = 1, 2, \ldots, [(n+1)/3], \tag{3}$$

where $[\cdot]$ is the integer part of the argument.

For the statistical treatment it is useful to convert temporal aggregation into a systematic sampling problem; this can be done by constructing a cumulator variable, generated as a time-varying first order autoregression (see Harvey, 1989, and Harvey and Chung, 2000):

$$\begin{aligned} Y^c_{it} &= \rho_t Y^c_{i,t-1} + Y_{it}, \quad t = 0, \ldots, n \\ &= \rho_t Y^c_{i,t-1} + h_i(y_{it}) \end{aligned} \tag{4}$$

where $h_i(\cdot)$ is the Box-Cox inverse transformation,

$$h_i(y_{it}) = \begin{cases} (1 + \lambda_i y_{it})^{1/\lambda_i}, & \lambda_i \neq 0, \\ \exp(y_{it}), & \lambda_i = 0, \end{cases}$$

and ρ_t is the cumulator coefficient, equal to zero for t corresponding to the first month in the quarter and 1 otherwise:

$$\rho_t = \begin{cases} 0 \; t = 3(\tau - 1), & \tau = 1, \ldots, [(n+1)/3] \\ 1 \text{ otherwise} . \end{cases}$$

The cumulator (4) is nothing more than a recursive implementation of the temporal aggregation rule (3). Only a systematic sample of the cumulator variable Y^c_{it} is available; in particular, if the sample period starts with the first month of the quarter at $t = 0$, the observed end of quarter values occur at times $t = 3\tau - 1, \tau = 1, 2, \ldots, [(n+1)/3]$

In the case of the logarithmic transformation ($\lambda_i = 0$), $Y^c_{i0} = \exp y_{i0}$, $Y^c_{i1} = \exp(y_{i0}) + \exp(y_{i1})$, $Y^c_{i2} = \exp(y_{i0}) + \exp(y_{i1}) + \exp(y_{i2})$, $Y^c_{i3} = \exp(y_{i3})$, $Y^c_{i4} = \exp(y_{i3}) + \exp(y_{i4})$, $Y^c_{i5} = \exp(y_{i3}) + \exp(y_{i4}) + \exp(y_{i5})$, ... Only the values $Y^c_{i2}, Y^c_{i5}, \ldots$ are observed, while the intermediate ones will be missing. It it important to remark that in general, when the Box-Cox transformation parameter is different from one, the quarterly totals are a nonlinear function of the underlying (unobserved) monthly values y_{it} (e.g. the sum of the exponentials of three consecutive values). Now, since we postulate that the first differences Δy_{it} are stationary and they have a linear factor model representation, the temporal aggregation constraints are nonlinear. In other words, we observe $Y^c_{i\tau} = Y_{i,3\tau-1} + Y_{i,3\tau-2} + Y_{i,3\tau-3}$, but the linear model is formulated in terms of the unobserved $y_{i,3\tau-i}, i = 1, 2, 3$, which are the Box-Cox power transformation of $Y_{i,3\tau-i}$. Hence, temporal aggregation yields a nonlinear observational constraint.

5 Nonlinear smoothing

Conditional on $\boldsymbol{\Xi}$, we face the problem of estimating the factors $\mathbf{f}_t$ and the missing values $\mathbf{y}_{it}, i = 1, \ldots, N_1$, from the available information, which consists of $Y^c_{it}, i = 1, \ldots, N_1, t = 3\tau - 1, \tau = 1, 2, \ldots, [(n+1)/3]$, for the quarterly time series and y_{it} for $i = N_1 + 1, \ldots, N$. This is a nonlinear smoothing problem that can be solved by iterating the Kalman filter and smoother adapted to a sequentially linearized state space model.

The estimation is carried out by an iterative algorithm which is a *sequential linear constrained* method for solving a constrained nonlinear optimization problem; see Gill et al. (1989), section 7. This method has been applied to nonlinear aggregation in mixed models Proietti (2006).

Let us partition the vectors $\mathbf{Y}_t = [\mathbf{Y}'_{1t}, \mathbf{Y}'_{2t}]'$, $\mathbf{y}_t = [\mathbf{y}'_{1t}, \mathbf{y}'_{2t}]'$, such that $\mathbf{Y}_t = \mathbf{h}(\mathbf{y}_t)$ is the inverse Box-Cox transform of $\mathbf{y}_t$, $\Delta\mathbf{y}_t = [\Delta\mathbf{y}'_{1t}, \Delta\mathbf{y}'_{2t}]'$, $\mathbf{x}_t = [\mathbf{x}'_{1t}, \mathbf{x}'_{2t}]'$, $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2]'$, and the matrices $\mathbf{D} = \mathrm{diag}(\mathbf{D}_1, \mathbf{D}_2)$, $\boldsymbol{\Theta} = [\boldsymbol{\Theta}'_1, \boldsymbol{\Theta}'_2]'$, $\boldsymbol{\Psi} = \mathrm{diag}(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)$, where the subscript 1 indexes the national accounts series, and the dimension of the blocks are respectively N_1 and N_2. Further, define $\xi = [\xi'_1, \ldots, \xi'_n]'$, i.e. the stack of the idiosyncratic disturbances.

If $\mathbf{x}_t$ were fully observed and $\boldsymbol{\Xi}$ were known, the KFS would yield the values of $\mathbf{f}$ and ξ that maximise the complete data likelihood $g(\mathbf{x}, \mathbf{f}; \boldsymbol{\Xi}) = g(\mathbf{x}|\mathbf{f}; \boldsymbol{\Xi})g(\mathbf{f}; \boldsymbol{\Xi})$. Now, $\mathbf{x}_{1t}, t = 1, \ldots, n$, is not available, but we observe a systematic sample of the cumulator

$$\begin{aligned} \mathbf{Y}^c_{1t} &= \rho_t \mathbf{Y}^c_{1,t-1} + \mathbf{Y}_{1t}, \\ &= \rho_t \mathbf{Y}^c_{1,t-1} + \mathbf{h}(\mathbf{y}_{1t}), \end{aligned}$$

and $\mathbf{x}_{1t}$ is related to $\mathbf{y}_{1t}$ by $\mathbf{x}_{1t} = \mathbf{D}_1^{-1}(\Delta\mathbf{y}_{1t} - \boldsymbol{\mu}_1)$.

The smoothing problem is now to obtain the values $\mathbf{f}$ and ξ that maximise the complete data likelihood $g(\mathbf{x}, \mathbf{f}; \boldsymbol{\Xi})$, subject to the nonlinear observational constraints that we observe a systematic sample of $\mathbf{Y}^c_{1t} = \rho_t \mathbf{Y}^c_{1,t-1} + \mathbf{h}(\mathbf{y}_{1t})$, and $\mathbf{x}_{1t} = \mathbf{D}_1^{-1}(\Delta\mathbf{y}_{1t} - \boldsymbol{\mu}_1)$.

The optimisation problem is handled with the support of the KFS. Each time the observation constraint is linearised around a trial value by a first order Taylor series expansion; this operation yields a linear state space model and the corresponding KFS provides a new trial value for the disaggregate series. This sequence of linearisations is iterated until convergence and the end result is a set of disaggregate monthly estimates $\mathbf{Y}_1$ and factor scores which incorporate the temporal aggregation constraints. As a by-product, disaggregate (monthly) estimates of the missing values $\mathbf{x}_{1t}$ and thus of $\mathbf{y}_{1t}$ and $\mathbf{Y}_{it}$ will be made available.

5.1 Estimation of the factors and the disaggregated series

The factors and disaggregate values $\mathbf{Y}_{1t}$ are estimated by the following iterative scheme:

1. Start from a trial value $\mathbf{y}_{1t}^*, t = 0, \ldots, n$, (e.g. obtained from application of the univariate Chow-Lin disaggregation method, see Chow and Lin, 1971).
2. Form the linear state space approximating model, using the first-order Taylor expansion around $\mathbf{y}_{1t}^*$.
3. Use the Kalman filter and smoother to estimate the factors $\mathbf{f}_t$, the idiosyncratic components, and the disaggregate series $\mathbf{y}_{1t}$, and thus $\mathbf{Y}_{1t}$.
4. If $||\mathbf{y}_{1t}^* - \hat{\mathbf{y}}_{1t}^*||$ is greater than a specified tolerance value, set $\mathbf{y}_{1t}^* = \hat{\mathbf{y}}_{1t}^*$ and return to step 2; else, set $\mathbf{Y}_{1t}^* = \mathbf{h}(\mathbf{y}_{1t}^*)$.

At convergence, the estimated disaggregate values satisfy the aggregation constraints, that is the observed quarterly aggregate $\mathbf{Y}_{1\tau}$ equals $\mathbf{h}(\mathbf{y}_{1,3\tau-1}^*) + \mathbf{h}(\mathbf{y}_{1,3\tau-2}^*) + \mathbf{h}(\mathbf{y}_{1,3\tau-3}^*)$.

6 Chain-linking and contemporaneous aggregation constraints

The quarterly national accounts series are subject to a number of accounting deterministic constraints, when the aggregates are expressed at current prices and at the average prices of the previous year. In particular, the 17 series are bound together by the identities:

GDP at basic prices	$= \sum$ Value added of the 6 branches
GDP at market prices	= GDP at basic prices + Taxes less subsidies
GDP at market prices+IMP	= CONS+INV+EXP
Domestic demand	= CONS+INV
CONS	= CONS_H+CONS_G

where CONS = Final consumption expenditures, CONS_H = Household and NPISH final consumption expenditure, CONS_G = Final consumption expenditure: general government, INV = Gross Capital Formation, EXP = Exports of goods and services, IMP = Imports of goods and services.

The production of chained linked national accounts estimates has changed drastically the role of the contemporaneous aggregation constraints considered above. In particular, the constraints hold only when the series are expressed at the average prices of the previous year; loosely speaking, only in that case they are expressed genuinely at constant prices. Otherwise, chaining, which is a multiplicative operation, destroys the additivity of the constraints, and a nonzero discrepancy arises. GDP and its main components are expressed in chain-linked volumes (millions of euros), with reference year 2000, which implies that the constraints hold exactly for the four quarters of the year 2001. Interestingly, due to the application of the *annual overlap technique*, exposed below, the constraints are not entirely lost, but they continue to hold after a transformation of the data that we call "dechaining", which aims at expressing the chained values at the prices of the previous year.

The cross-sectional constraints can be enforced by a multistep procedure that de-chains the estimated monthly values, expressing them at the average prices of the previous year, and projects the estimates on the subspace of the constraints. The dechaining procedure is in line with that advocated by the IMF manual (see Bloem *et al.*, 2001).

7 Estimation results

Estimation of the unknown parameters and temporal disaggregation is carried out by an iterative algorithm which alternates two main steps until convergence. We start from a trial disaggregate time series $\mathbf{y}_{1t}^*, t = 0, \ldots, n,$, obtained from the temporal disaggregation of the quarterly national accounts series according to the univariate Chow-Lin procedure, using industrial production and retail sales (total) as monthly indicators. The disaggregate time series serve to construct the standardized stationary series $\mathbf{x}_t$, that form a balanced panel of monthly time series. The initial estimate of the parameter is computed by a principal component analysis of the covariance matrix of the $\mathbf{x}_t$'s.

The number of factors, K, is selected at this stage according to the information criteria proposed by Bai and Ng (2002).

Conditional on K, the estimation of the factor model involves the following steps:

1. Given a set of estimated disaggregate values $\hat{\mathbf{y}}_{1t}$, satisfying the temporal and contemporaneous aggregation constraints, we construct the pseudo complete balanced panel of time series $\mathbf{y}_t = [\hat{\mathbf{y}}_{1t}', \mathbf{y}_{2t}']'$, where $\mathbf{y}_{2t}$ are the observed monthly series. We then obtain the stationary transformation $\boldsymbol{\Delta}\mathbf{y}_t$ and estimate $\boldsymbol{\mu}$ and $\mathbf{D}$ by computing the sample average and the standard deviation of the individual time series. We construct the standardized stationary series $\mathbf{x}_t = \hat{\mathbf{D}}^{-1}(\boldsymbol{\Delta}\mathbf{y} - \hat{\boldsymbol{\mu}})$, and estimate the parameters of the factor model $\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_\eta, \boldsymbol{\Psi}$ by maximum likelihood using the EM algorithm or by principal component analysis.
2. Conditional on the parameter estimates, we estimate the disaggregate time series $\hat{\mathbf{y}}_{1t}$ (and thus $\hat{\mathbf{Y}}_{1t} = \mathbf{h}(\hat{\mathbf{y}}_{1t})$), consistent with the temporal and cross-sectional constraints. This step is carried out iteratively, with each iteration consisting of two steps:
   (a) estimate $\hat{\mathbf{y}}_{1t}$ enforcing the nonlinear temporal aggregation constraints, as detailed in (5.1);
   (b) enforce the cross-sectional temporal aggregation constraints by the de-chaining and chaining-back procedure outlined in section (6).

The estimated number of factors is $K = 6$: this can be considered as a conservative estimate. The share of the variance explained by the first three principal components is 34.13%, whereas that explained by the first six is 45.18%.

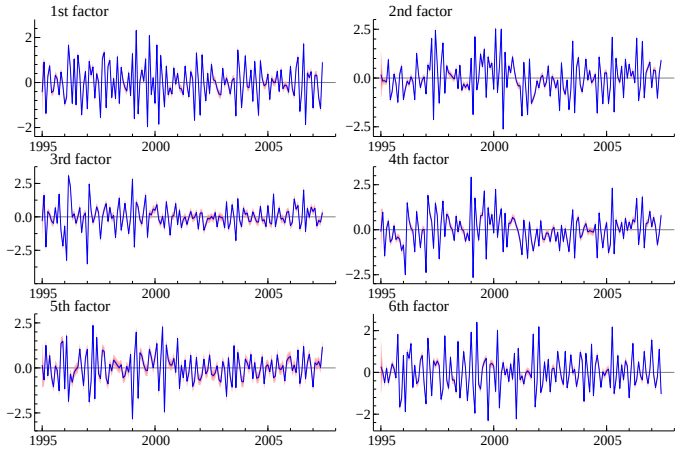

Fig. 1. Point and 95% interval estimates of the common factors.

The estimation of the factor model was carried out using both the EM algorithm and PCA, as far as the estimation of the parameter vector $\boldsymbol{\Xi}$ is concerned. Less than 200 iterations are required for convergence in both cases. The estimation results are very similar, both for $\boldsymbol{\Xi}$ and the disaggregate series; however, the estimated factors conditional on the PCA parameter estimates are slightly smoother than those obtained from the EM method. As a consequence, the disaggregate series $\hat{\mathbf{Y}}_{1t}$ have a smaller variation at the high frequencies. Since *ceteris paribus* we would prefer smoother estimates of monthly GDP and its components, the presentation of the results will henceforth concentrate on the PCA method. It should be recalled that PCA is used only for estimating the parameters in $\boldsymbol{\Xi}$; the factors are estimated along with the monthly GDP and its components according to the second step of our procedure (i.e. incorporating the temporal and cross-sectional aggregation constraints).

Figure 1 displays the point estimates of the six factors, $\tilde{\mathbf{f}}_{t|n}$, and the approximate 95% interval estimates, based on the assumption of normality. As the plot illustrates, the dynamic of the estimated factors is dominated by high frequency variation, resulting in a negative autocorrelation; also, the third factor captures the main economic shocks that affected the construction sectors. However, the factors capture also the dynamics of the euro area business cycle: in particular, this information is carried by the 2nd, 4th and 5th factors.

Figure 2 is a biplot of the estimated factor analysis. Series that load on the same factors will be represented by two close points; the labels "NA", "I", "C", "R", "F", "S" refer, respectively, to the national accounts series, industry, construction, retail, financial and monetary indicators, business and consumer surveys. A group of series with the same loadings pattern is hours

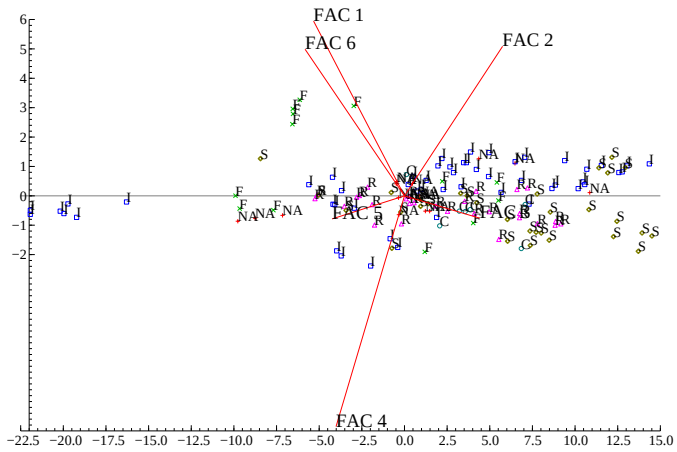

Fig. 2. Biplot of the factor loadings.

worked in industry, displayed to the left of the biplot. In general, series belonging to the same group tend to cluster together. The loading of a particular variable on a specific factor can be approximated by the orthogonal projection of the point representing the variable on the line representing the factor. The survey series are mostly related to the second and the third factors, whereas the financial variables are associated to the 4th and 6th factors. The monthly construction series are mostly associated to factor 3 (the loading of value added in the construction sector is 1).

A most important side output of our modeling effort is the estimation of monthly GDP and its main components. The estimates comply with the temporal aggregation constraints and the cross-sectional identities for the year 2001, and if the series are expressed at the prices of the previous year. Moreover, they are highly informative as they incorporate the information that is common to a large set of monthly indicators. Figure 3 displays monthly GDP at market prices, final consumption expenditures and gross capital formation, along with their monthly and yearly growth rates. It must be stressed that approximate measures of reliability of the estimates are directly available from the our methodology.

8 Conclusive remarks

The paper has proposed an iterative scheme for estimating a large scale factor model with data at different frequencies, providing an exact treatment of the temporal and cross-sectional aggregation constraints. The model is used to nowcast monthly GDP and its decomposition by expenditure type and by the output approach. The results are relevant not only because the estimated

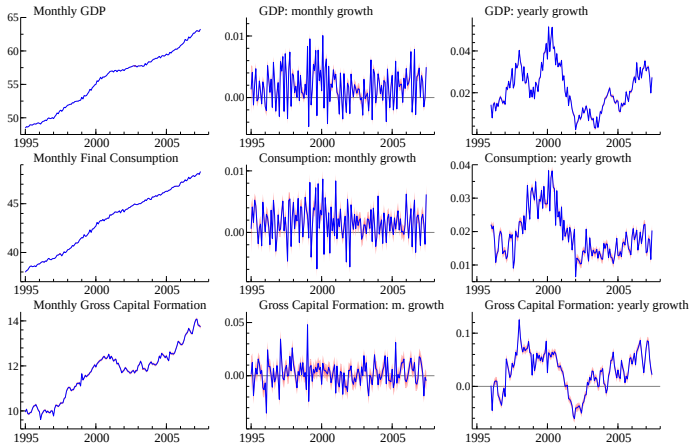

Fig. 3. Monthly estimates of GDP at market prices, Final Consumption and Gross Capital Formation (chained 2000 volumes), and monthly and yearly growth. Point and 95% interval estimates.

common factors embody the economic information contained in the national accounts macro variables, but also because the availability of monthly estimates of the national accounts series can be seen as a useful addition to the available published data.

References

AMENEGUAL, D. and WATSON, M.W. (2007): Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business and Economic Statistics*, 25, 91-96.

BAI, J. and NG, S. (2002): Determining the number of factors in approximate factor models. *Econometrica*, 70, 191-221.

BLOEM, A., DIPPELSMAN, R.J. and MAEHLE, N.O. (2001): *Quarterly National Accounts Manual Concepts, Data Sources, and Compilation*. International Monetary Fund.

BOX G.E.P., and COX, D.R. (1964): An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, B*, 26, 211–246.

CHOW, G., and LIN, A.L. (1971): Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series, *The Review of Economics and Statistics*, 53, 4, 372-375.

CRISTADORO, R., FORNI, M., REICHLIN, L. and VERONESE, G. (2005): A core inflation indicator for the euro area. *Journal of Money Credit and Banking*, 37, 539–560.

DEMPSTER, A., LAIRD, N. and RUBIN, D. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1):138.

DOORNIK, J.A. (2006): *Ox 4.0 - An Object-Oriented Matrix Programming Language*, Timberlake Consultants Ltd: London.

DURBIN J., and KOOPMAN, S.J. (2001): *Time Series Analysis by State Space Methods*, Oxford University Press: New York.

EUROSTAT (1999): *Handbook of quarterly national accounts*, Luxembourg, European Commission.

FORNI, M., HALLIN, M., LIPPI, M., and REICHLIN, L. (2000): The generalized dynamic factor model: identification and estimation. The *Review of Economics and Statistics*, 82, 540-554.

FORNI, M., HALLIN, M., LIPPI, M., and REICHLIN, L. (2001): Coincident and leading indicators for the euro area. *The Economic Journal*, 111, C62–C85.

FORNI, M., HALLIN, M., LIPPI, M., and REICHLIN, L. (2005): The generalized factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association*, 100, 830-40.

GEWEKE, J.F., and SINGLETON, K.J. (1981): Maximum likelihood confirmatory factor analysis of economic time series. *International Economic Review*, 22, 1980.

GILL, P.E., MURRAY, W., SAUNDERS, M.A., and WRIGHT, M.H. (1989): Constrained nonlinear programming. In: G.L. Nemhauser, A.H.G. Rinnooy Kan, M.J. Todd (Eds.), *Handbooks in Operations Research and Management Science*, Vol. 1, Optimization, pp. 171-210, Elsevier, Amsterdam, 1989.

HARVEY, A.C. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press: Cambridge.

HARVEY, A.C. and CHUNG, C.H. (2000): Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistics Society*, Series A, 163, 303-339.

PROIETTI, T. (2006): On the estimation of nonlinearly aggregated mixed models. *Journal of Computational and Graphical Statistics*, Vol. 15, 1–21.

SHUMWAY, R.H., and STOFFER, D.S. (1982): An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3, 253-264.

STOCK, J.H., and WATSON, M.W. (2002a): Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics*, 20, pp.147-162.

STOCK, J.H., and WATSON, M.W. (2002b): Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97, 1167-79.

STOCK, J.H., and WATSON, M.W. (2006): Forecasting with many predictors, in G. Elliott, C.W.J Granger and A. Timmermann (Eds.): *Handbook of Economic Forecasting, Volume 1*, Elsevier.

Part XVI

Index

Index