THOMAS KNEIB
GERHARD TUTZ
Editors

Statistical Modelling and Regression Structures

Festschrift in Honour of Ludwig Fahrmeir

Physica-Verlag
An Imprint of Springer Science+Business Media

Statistical Modelling and Regression Structures

Thomas Kneib · Gerhard Tutz
Editors

Statistical Modelling and Regression Structures

Festschrift in Honour of Ludwig Fahrmeir

Physica-Verlag

Editors
Prof. Dr. Thomas Kneib
Institut für Mathematik
Carl von Ossietzky Universität Oldenburg
26111 Oldenburg
Germany
thomas.kneib@uni-oldenburg.de

Prof. Dr. Gerhard Tutz
Institut für Statistik
Ludwig-Maximilians-Universität München
Akademiestraße 1
80799 München
Germany
gerhard.tutz@stat.uni-muenchen.de

ISBN 978-3-7908-2412-4 e-ISBN 978-3-7908-2413-1
DOI 10.1007/978-3-7908-2413-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009943264

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

Physica-Verlag is a brand of Springer-Verlag Berlin Heidelberg
Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The collected contributions contained within this book have been written by friends and colleagues to acknowledge Ludwig Fahrmeir's widespread and important impact on Statistics as a science, while celebrating his 65th Birthday.

As a young student, Ludwig started his career as a Mathematician, but he quickly turned into a rising and shining star within the German and international Statistics community. He soon obtained both his PhD and his Habilitation at the Technical University of Munich. After a short period as a visiting professor at the University of Dortmund, he returned to his homeland Bavaria and was appointed Full Professor of Statistics at the University of Regensburg, at the age of 32.

Some years later, he moved to the capital of Bavaria and became Professor at the Department of Statistics at the University of Munich. His appointment had significant impact on the Department since, soon after his arrival, Ludwig started an initiative to establish a collaborative research center on the "Statistical Analysis of Discrete Structures." After a successful application for initial funding, further funding was extended several times, until the research center reached the maximum period for funding in 2006. During the complete duration, Ludwig served as a speaker of the research center and – to cite one of the final referees – "managed it in an easy and efficient way and contributed several important results."

During the last forty years, Ludwig's work has had tremendous impact on the Statistics community. He was among the first researchers to recognize the importance of generalized linear models and contributed in a series of papers to the theoretical background of that model class. His interest in statistical modelling led to the organization of a workshop on "Statistical Modelling and Generalized Linear Models (GLIM)" in Munich in 1992 and culminated in the highly cited monograph on "Multivariate Statistical Modelling Based on Generalized Linear Models" that saw two printings and remains to be a key reference on applied statistical modelling utilizing generalized linear models. Ludwig also had great influence on the creation of the Statistical Modelling Society, and is currently on the advisory board of the corresponding journal on "Statistical Modelling." Both the society and journal emerged out of the early GLIM workshops and proceedings.

Of course, Ludwig's work is definitely not restricted to generalized linear models but – on the contrary – spans a wide range of modern Statistics. He co-authored or co-edited several monographs, e.g. on Multivariate Statistics, Stochastic Processes, Measurement of Credit Risks, as well as popular textbooks on Regression and an Introduction to Statistics. His recent research contributions are mostly concentrated in semiparametric regression and spatial statistics within a Bayesian framework.

When first circulating the idea of a Festschrift for the celebration of Ludwig's 65th birthday, all potential contributors were extremely positive, many immediately agreeing to contribute. These reactions atest to Ludwig's high personal and professional appreciation in the statistical community. The far reaching and variety of subjects covered within these contributions also represents Ludwig's broad interest and impact in many branches of modern Statistics.

Both editors of this Festschrift were lucky enough to work with Ludwig at several occasions and in particular early in their careers as PhD students and PostDocs. His personal and professional mentorship and his strong commitment made him a perfect supervisor and his patient, confident and encouraging working style will always be remembered by all of his students and colleagues. Ludwig always provided a friendly working environment that made it a pleasure and an honor to be a part of his working group. We are proud to be able to say that Ludwig is much more than a colleague but turned into a friend for both of us.

Oldenburg and Munich, January 2010 *Thomas Kneib, Gerhard Tutz*

Acknowledgements

The editors would like to express their gratitude to

- all authors of this volume for their agreement to contribute and their easy cooperation at several stages of putting together the final version of the Festschrift.
- Johanna Brandt, Jan Gertheiss, Andreas Groll, Felix Heinzl, Sebastian Petry, Jan Ulbricht and Stephanie Rubenbauer for their invaluable contributions in proofreading and correction of the papers, as well as in solving several LaTeX-related problems.
- the Springer Verlag for agreeing to publish this Festschrift and in particular Nils-Peter Thomas, Alice Blanck and Frank Holzwarth for the smooth collaboration in preparing the manuscript.

Contents

List of Contributors

Thomas Augustin
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: thomas.augustin@stat.uni-muenchen.de

Christiane Belitz
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany

Andreas Brezger
HypoVereinsbank, München, Germany, e-mail: andreas.brezger@hvb.de

Erik Brodin
Department of Mathematical Sciences, Chalmers University of Technology, Göteborg, Sweden, e-mail: ebrodin@math.chalmers.se

Raymond J. Carroll
Department of Statistics, Texas A& M University, College Station, Texas, USA, e-mail: carroll@stat.tamu.edu

Ciprian M. Crainiceanu
Department of Biostatistics, Johns-Hopkins-University Baltimore, USA, e-mail: ccrainic@jhsph.edu

Claudia Czado
Zentrum Mathematik, Technische Universität München, Garching, Germany, e-mail: cczado@ma.tum.de

Christian Dittrich
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany

Paul H. C. Eilers
Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands, e-mail: p.eilers@erasmusmc.nl

Konstantinos Fokianos
Department of Mathematics & Statistics, University of Cyprus, Nicosia, Cyprus, e-mail: fokianos@ucy.ac.cy

Ronja Foraita
Bremen Institute for Prevention Research and Social Medicine (BIPS), University of Bremen, Germany, e-mail: foraita@bips.uni-bremen.de

Sylvia Frühwirth-Schnatter
Institut für Angewandte Statistik, Johannes-Kepler-Universität Linz, Austria, e-mail: Sylvia.Fruehwirth-Schnatter@jku.at

Rudolf Frühwirth
Institut für Hochenergiephysik der Österreichischen Akademie der Wissenschaften, Wien, Austria, e-mail: fru@hephy.oeaw.ac.at

Moritz Grenke
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: moritz.grenke@campus.lmu.de

Andreas Groll
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: andreas.groll@stat.uni-muenchen.de

Sandra Hackensperger
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany

Alfred Hamerle
Lehrstuhl für Statistik, Wirtschaftswissenschaftliche Fakultät, Universität Regensburg, Germany, e-mail: alfred.hamerle@wiwi.uni-regensburg.de

Leonhard Held
Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Switzerland, e-mail: leonhard.held@ifspm.uzh.ch

Christian Heumann
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: christian.heumann@stat.uni-muenchen.de

Michael Höhle
Institut für Statistik, Ludwig-Maximilians-Universität München and Munich Center of Health Sciences, Germany,
e-mail: michael.hoehle@stat.uni-muenchen.de

Torsten Hothorn
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: torsten.hothorn@stat.uni-muenchen.de

Jianhua Z. Huang
Department of Statistics, Texas A&M University, College Station, Texas, USA, e-mail: jianhua@stat.tamu.edu

Judith Hübner
Institut für Mathematik, Technische Universität München, Germany

Vindi Jurinovic
Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE), Ludwig-Maximilians-Universität München, Germany

Göran Kauermann
Centre for Statistics, Bielefeld University, Dep of Business Administration and Economics, Bielefeld, Germany, e-mail: gkauermann@wiwi.uni-bielefeld.de

Stephan Klasen
Department of Economics, University of Göttingen, Germany, e-mail: sklasen@uni-goettingen.de

Claudia Klüppelberg
Center for Mathematical Sciences, Technische Universität München, Garching, Germany, e-mail: cklu@ma.tum.de

Thomas Kneib
Institut für Mathematik, Carl von Ossietzky Universität, Oldenburg, Germany, e-mail: thomas.kneib@uni-oldenburg.de

Stefan Lang
Department of Statistics, University of Innsbruck, Austria, e-mail: stefan.lang@uibk.ac.at

Jeffrey S. Lehman
JPMorganChase, Home Finance Marketing Analytics, Columbus, USA e-mail: jeff_lehman@bankone.com

Friedrich Leisch
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: friedrich.leisch@stat.uni-muenchen.de

Ulrich Mansmann
Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE), Ludwig-Maximilians-Universität München, Germany, e-mail: mansmann@ibe.med.uni-muenchen.de

Josue G. Martinez
Department of Statistics, Texas A&M University, College Station, Texas, USA, e-mail: jgmartinez@stat.tamu.edu

Brian D. Marx
Department of Experimental Statistics, Louisiana State University, Baton Rouge, USA, e-mail: bmarx@lsu.edu

Gernot Müller
Zentrum Mathematik, Technische Universität München, Garching, Germany, e-mail: mueller@ma.tum.de

Axel Munk
Institut für Mathematische Stochastik, Georg August Universität Göttingen, Germany, e-mail: munk@math.uni-goettingen.de

Thi-Ngoc-Giau Nguyen
Zentrum Mathematik, Technische Universität München, Garching, Germany, e-mail: ntngiau2002@yahoo.com

William I. Notz
Department of Statistics, The Ohio State University, USA, e-mail: win@stat.osu.edu

Iris Pigeot
Bremen Institute for Prevention Research and Social Medicine (BIPS), University of Bremen, Germany, e-mail: pigeot@bips.uni-bremen.de

Kilian Plank
Lehrstuhl für Statistik, Wirtschaftswissenschaftliche Fakultät, Universität Regensburg, Germany, e-mail: kilian.plank@wiwi.uni-regensburg.de

Mihaela Pricop
Institut für Mathematische Stochastik, Georg August Universität Göttingen, Germany, e-mail: pricop@math.uni-goettingen.de

Nikolay Robinzonov
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: nikolay.robinzonov@stat.uni-muenchen.de

Gerd Ronning
Wirtschaftswissenschaftliche Fakultät, Eberhard Karls Universität Tübingen, Germany, e-mail: gerd.ronning@uni-tuebingen.de

Håvard Rue
Department of Mathematical Science, Norwegian University of Science and Technology, Trondheim, Norway, e-mail: havard.rue@math.ntnu.no

Thomas J. Santner
Department of Statistics, The Ohio State University, USA, e-mail: tjs@stat.osu.edu

Markus Schmidberger
Institut für medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE), Ludwig-Maximilians-Universität München, Germany

Hans Schneeweiß
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: schneew@stat.uni-muenchen.de

Birgit Schrödle
Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Switzerland, e-mail: birgit.schroedle@ifspm.uzh.ch

Christian Seiler
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany

Michael Stanley Smith
Melbourne Business School, University of Melbourne, 200 Leicester Street, Carlton, Victoria 3053, Australia, e-mail: mike.smith@mbs.edu

Carolin Strobl
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: carolin.strobl@stat.uni-muenchen.de

Ralf Strobl
Institute for Health and Rehabilitation Sciences, Ludwig-Maximilians-Universität München, Germany

Gerhard Tutz
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: gerhard.tutz@stat.uni-muenchen.de

Gero Walter
Institut für Statistik, Ludwig-Maximilians-Universität München, Germany, e-mail: gero.walter@stat.uni-muenchen.de

Brian J. Williams
Los Alamos National Laboratory, Los Alamos, USA, e-mail: brianw@lanl.gov

The Smooth Complex Logarithm and Quasi-Periodic Models

Paul H. C. Eilers

Abstract Quasi-periodic signals, which look like sine waves with variable frequency and amplitude, are common in nature and society. Examples that will be analyzed in this paper are sounds of crickets, counts of sunspots, movements of ocean currents, and brightness of variable stars. Euler's formula for the complex logarithm, combined with smoothly changing real and imaginary components, provides a powerful model. It is highly non-linear and special care is needed to get starting values for an iterative estimating algorithm. The model is extended with a trend and harmonics. A cascaded link function allows modeling of quasi-periodic series of counts. The model and real-world applications are described in an expository style.

1 Foreword

Ludwig Fahrmeir has studied generalized linear models extensively and he has had a strong influence on their development and practical application. But I'm quite sure that he never worked on models with a complex (in the sense of imaginary numbers) link function, although he came near, in the study of seasonal generalized linear models (Fahrmeir & Tutz 2001). In this chapter we will see that a complex logarithmic link is natural start for models with variable periodicity.

2 Introduction

Many signals in nature and society are quasi-periodical: they show approximately a sinusoidal shape, but frequency and signal strength vary over time. Examples that will

Paul H. C. Eilers
Department of Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands,
e-mail: p.eilers@erasmusmc.nl

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_1,

be analyzed in this paper are sounds of crickets, counts of sunspots, movements of ocean currents, and brightness of variable stars. Many other examples can be found: sounds of musical instruments, business cycles and radio waves. In all cases it is of interest to get a parsimonious semi-parametric description of the instantaneous frequency and amplitude. I propose a model that achieves this goal, inspired by a fundamental formula from complex analysis: $\exp(\alpha + i\phi) = e^{\alpha}(\cos\phi + i\sin\phi)$. It follows that α can be interpreted as the logarithm of the amplitude and ϕ as the phase. Modeling both as smooth functions of time, $\alpha(t)$ and $\phi(t)$, we get the desired model.

Given a (complex or real) signal, we can fit the model, using variants of P-splines (Eilers & Marx 1996) for the semi-parametric description. The model is highly non-linear. With proper starting values iteratively reweighed least squares estimation quickly leads to the solution. However, finding good starting values is far from trivial, so it will be discussed at some length. When the data are rich enough, zero-crossings can be used. For sparse data, initial smooth interpolation is needed. Interpolation with standard P-splines will not always work. Harmonic smoothing, using a modified penalty, presents a solution.

Not always is a sinusoidal shape enough to fit observed data well. In such cases additional harmonics, having phase functions $k\phi(t)$, for small integer k, like 2 and 3, either in fixed proportions, or modulated by their own amplitude functions, will be added to the model.

When the data are patently non-normal, a cascade of two link functions, with a Poisson or binomial conditional data distribution, works well. This will be illustrated with an extensive data of daily sunspot counts.

This is an expository chapter and its structure is somewhat unusual in that there are no separate sections on Theory and Applications. In the next section, observed signals of increasing complexity are introduced sequentially, along with the tools for modeling them. I believe this improves readability. In Section 3 I sketch many opportunities for more complex models, to seduce other statisticians to enter this field. The chapter ends with a discussion.

3 Data and Models

Figure 1 shows a prototypical quasi-periodic signal. It is a segment of about 0.4 s of the song of the Confused Ground Cricket *Eunemobius confusus* (Elliott & Hershberger 2006). It is clear that the amplitude changes over time, but it is harder to judge whether the frequency changes and by how much.

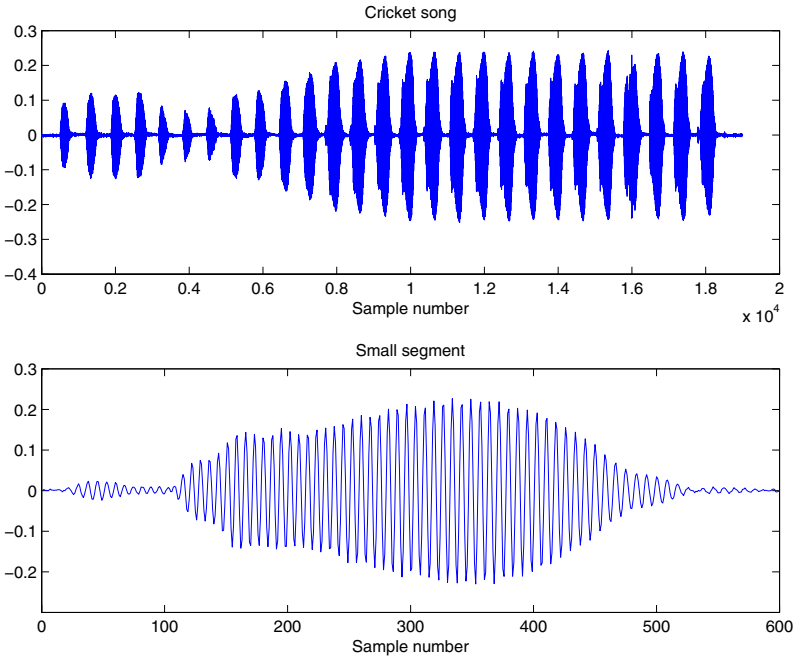


Fig. 1 Song of a cricket. Top: 19000 samples (approximately 0.4 s). Bottom: enlargement of the last burst.

3.1 The Basic Model

We are going to develop a smooth model for the complex logarithm of the signal, which gives us two very useful time series, one for the logarithm of the amplitude, the other for the phase, which is the integral of the local frequency.

Euler's formula from complex analysis states that

$$\exp(\alpha + i\phi) = e^{\alpha}(\cos\phi + i\sin\phi). \tag{1}$$

This is what we need for a quasi-periodic model, because when we let α and ϕ change smoothly over time, the real part of the exponential of $\alpha(t) + i\phi(t)$ has a cosine shape. Its local frequency is given by $d\phi/dt$ and the local amplitude by $\exp\alpha$.

3.2 Splines and Penalties

To obtain smooth curves for $\alpha(t)$ and $\phi(t)$ we model them as a sum of a generous number of scaled B-splines, allowing more flexibility than needed. A roughness

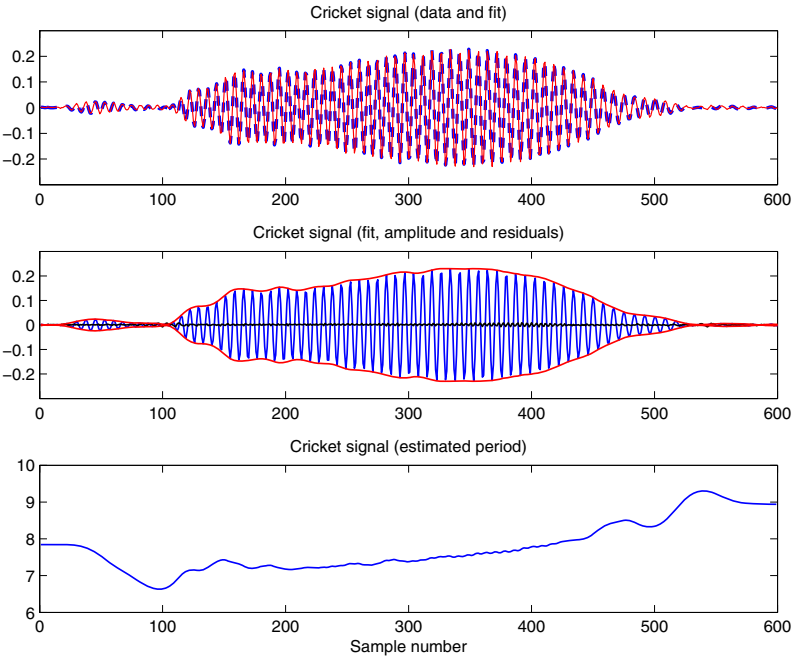


Fig. 2 Fit to the cricket song. Top: model fit (broken line) and signal (full line). Middle: model fit, plus and minus the amplitude, shown as an envelope, and residuals (very small). Bottom: the estimated local period.

penalty on the B-spline coefficients gives further continuous control over smoothness. This the P-spline idea, advocated by Eilers & Marx (1996), simplifying an earlier proposal by O'Sullivan (1986); recently Wand & Ormerod (2008) revisited the latter approach in more detail. So we have

$$\alpha(t) = \sum_{k=1}^{K} B_k(t) a_k; \quad \phi(t) = \sum_{k=1}^{K} B_k(t) c_k, \tag{2}$$

where $B_k(t)$ indicate the kth B-spline, evaluated at t.

The two-part penalty is

$$\text{Pen} = \lambda_\alpha \sum_k (\Delta^2 a_k)^2 + \lambda_\phi \sum_k (\Delta^2 c_k)^2. \tag{3}$$

If we observe a complex time series $x_j + iy_j$, $j = 1,\ldots,J$, at time points t_j, the following objective function is proposed:

$$S = \sum_{j=1}^{J} v_j (x_j - \mu_j)^2 + \sum_{j=1}^{J} w_j (y_j - \nu_j)^2 + \text{Pen}, \tag{4}$$

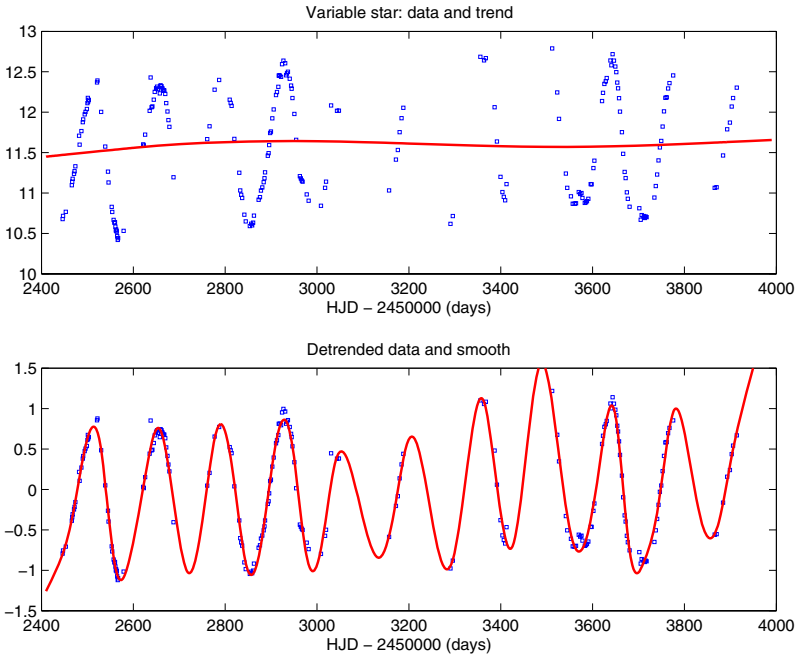


Fig. 3 Interpolating scattered data to find zero-crossing reliably. Top: magnitude of a variable star, with estimated trend. Bottom: light P-spline smoothing of the de-trended series.

with

$$\mu_j = \exp\alpha(t_j)\cos\phi(t_j); \quad \nu_j = \exp\alpha(t_j)\sin\phi(t_j). \tag{5}$$

We have two weighted sums of squares, one for the real, the other for the imaginary part. The weights are introduced for easy handling of time series without a complex component. In that case we fill in $y_j \equiv w_j \equiv 0$. Most observed time series only have a real part, but in physics and chemistry complex data are common. An example is NMR (nuclear magnetic resonance). Later in this paper we will encounter a time series of spatial positions and study it as a complex signal. In addition, selective use of zero weights allows us to exclude undesired data points. A very useful property of P-splines is that they smoothly and automatically interpolate (or extrapolate) wherever weights are zero.

The objective function in (4) is highly non-linear in the coefficients a and c. Essentially the cosine and sine are link functions in the sense of a generalized linear model (with Ba and Bc as linear predictors. Because quasi-periodicity is our game, these links functions are not monotone. Fitting the model is far from trivial; finding good starting values is a serious bottleneck. It will be discussed later. For the moment we assume that we do have good starting values. Then the real parts can be approximated by

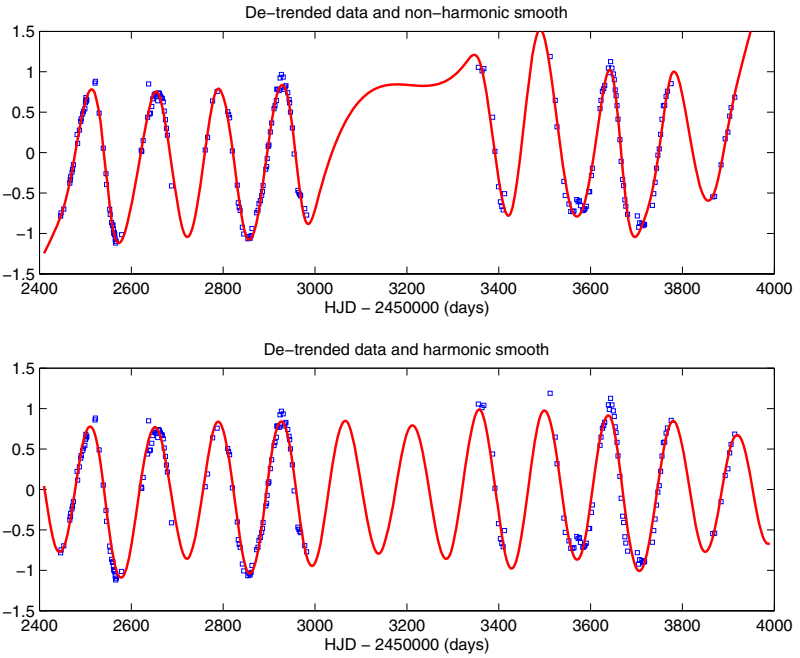


Fig. 4 Interpolating scattered data with different standard and harmonic penalties. De-trended data as in previous figure. Top: P-spline smoothing with a standard (second order difference) penalty. Bottom: P-spline smoothing with an harmonic penalty, based on a period of 110.

$$\mu(t) \approx \tilde{\mu}(t) + \sum_{k=1}^{K} \delta a_k \partial\tilde{\mu}(t)/\partial a_k + \sum_{k=1}^{K} \delta c_k \partial\tilde{\mu}(t)/\partial c_k, \tag{6}$$

where δa and δc indicate small corrections. A similar formula applies to the imaginary part $\nu(t)$. Formulas for the partial derivatives are easy to derive. Given an approximate solution we apply the Gauss-Newton algorithm: weighted regression of the residuals on the partial derivatives, to compute corrections δa and δc.

It is assumed that the number of B-splines in the basis B is large enough to over-fit the data. Then only the parameters λ_α and λ_ϕ tune smoothness.

When the observations are equally spaced on the time axis, one can drop the B-spline basis and replace it by the identity matrix. Ten $\alpha = a$ and $\phi = c$. Large systems of equations results, but they are extremely sparse, consisting of combinations of quindiagonal sub-matrices. All computations were done in Matlab, which can handle sparse matrices very efficiently and almost transparently. We are back then at the Whittaker smoother (Eilers 2004). In what follows this approach will be used, unless otherwise indicated. To simplify the presentation, we will speak of penalties on α and ϕ, implicating that this means a and c whenever a B-splines basis is included.

If we apply the model to the cricket data, we get an extremely good fit, as Figure 2 shows. The pattern of the amplitude $\exp\hat{\alpha}$ closely follows peaks and values of the signal. It is hard to judge length of period changes by eye, but the small size of the residuals shows that the model does track the variable frequency well. I have chosen to present the period instead of the frequency, because the former is more easily recognized in the data, from the spacing of zero-crossings.

3.3 Starting Values

Good starting values are crucial, especially for ϕ. The link function is extremely non-linear. If ϕ is too far off, the resulting quasi-periodic signal will have too few or too many periods, and we will get stuck in one of many local minima of the objective function.

The following procedure appears to be quite reliable for well-behaved data like the cricket sound. First, light smoothing is applied to the signal, to eliminate noise. The zero-crossings are located, giving the series u_k, for $k = 1, \ldots, p$, where p is the the number of periods of the signal. Because the cosine has zero-crossings there, at position u_k, ϕ has the value $v_k = (u_0 + k + 1/2)\pi$, with $u_0 = 0$ when the first crossing is downward, and $u_0 = 1$ otherwise. A smooth P-spline curve fit through the series (u, v) gives the desired starting estimate for ϕ.

In some unfavorable cases, when the signal locally is weak and noise is strong, zero-crossings may be missed. If the signal is strong enough on both sides, one can use zero-crossings separately there. The curves at both sides have to connect smoothly after shifting the right one up by an integer multiple of π. Because phase is changing smoothly, only a few possible integers are indicated. One adds $k\pi$ to v for the right segment and estimates a smooth curve to all (u, v) pairs and keeps the results for the integer k that gives the best fit (and is consistent with alternating downward and upward crossings).

If all else fails, one can fill in values for u and v by hand, based on visual inspection of the data.

A silent assumption was that data are complete. This is not always the case and then we have to interpolate first. Luckily, P-splines are very good at this. Figure 3 shows data on the magnitude (the brightness on a logarithmic scale) of a variable star, obtained from `http://www.astrouw.edu.pl/asas`, the All Sky Automated Survey (ASAS) database. A basis with 50 cubic B-splines was used for interpolation, with $\lambda = 0.01$. We are in luck, because there are some observations around day 3200 that steer the P-spline fit in the right direction. Figure 4 shows what would happen if these data were missing. We certainly cannot use the result directly for computing zero-crossings. We could introduce a gap and connect phase curve estimates for the left and right segments as described above.

An alternative approach is to use a harmonic penalty instead of second order differences (Eilers & Marx 1996). In the latter case we would use $\Delta^2 a_j = a_j - 2a_{j-1} + a_{j-2}$, but the harmonic penalty uses $a_j - 2qa_{j-1} + a_{j-2}$, with $q = \cos(2\pi d/P)$, where

d is the distance between knots, and P a desired period. Figure 4 shows the effect. As long as the weight of the penalty is not too large, the period does not have to be very precisely chosen. A little experimenting, guided by visual inspection of the interpolated signal is recommended. Here 50 B-splines were used, with $P = 110$ and $\lambda = 0.1$.

Once starting values for ϕ have been found, it is relatively simple to find them for α. In many cases it is enough to use a constant α, the logarithm of the average absolute value of the signal, divided by the average of the absolute value of $\cos\phi$. A more refined approach is to fit a P-spline-based varying-coefficient model (Eilers & Marx 2002).

3.4 Simple Trend Correction and Prior Transformation

The cricket song has the pleasant property that it has as symmetric, sine-like shape, without a trend. This is not always the case. Figure 5 shows monthly sunspot numbers as given by `data(sunspots)` in R. To remove the asymmetry to a large degree, a square root transformation works well, as the same figure shows. But now we see a prominent slowly fluctuating trend. A simple solution is to apply the discrete smoother with a large penalty ($\lambda = 1000$). The trend in the lower panel of Figure 5 has been computed this way. As Figure 6 shows, a quite good result is obtained for the sunspot data. Notice the peak in in the interval around 1795. Period length increased quite sharply there. There has been a long debate in astronomical circles about this decade (Usoskin et al. 2009). It seems that the original compilation of historic data filled in gaps in such a way that two relatively short periods have merged into a very long one. See also Section 3.6.

The quasi-periodic signal that results from the sunspots after transformation and trend detection shows an asymmetric shape: a fast rise and a slow fall. In Section 3.7 harmonics will be added to improve the model.

The proposed simple trend removal can be improved by adding an explicit trend to the model:

$$\mu = \exp(\alpha)\cos(\phi) + \gamma, \tag{7}$$

with roughness penalties on α, ϕ and γ. Fitting this model is only slightly more complicated than before. The simple trend estimate provides good starting values for γ.

3.5 A Complex Signal

Figure 7 shows an interesting data set. The location of a free-drifting sub-surface buoy in the Atlantic Ocean has been monitored for six months in 2001 (Lilly & Gascard 2006). The data can be found in the JLAB toolbox for Matlab, written by

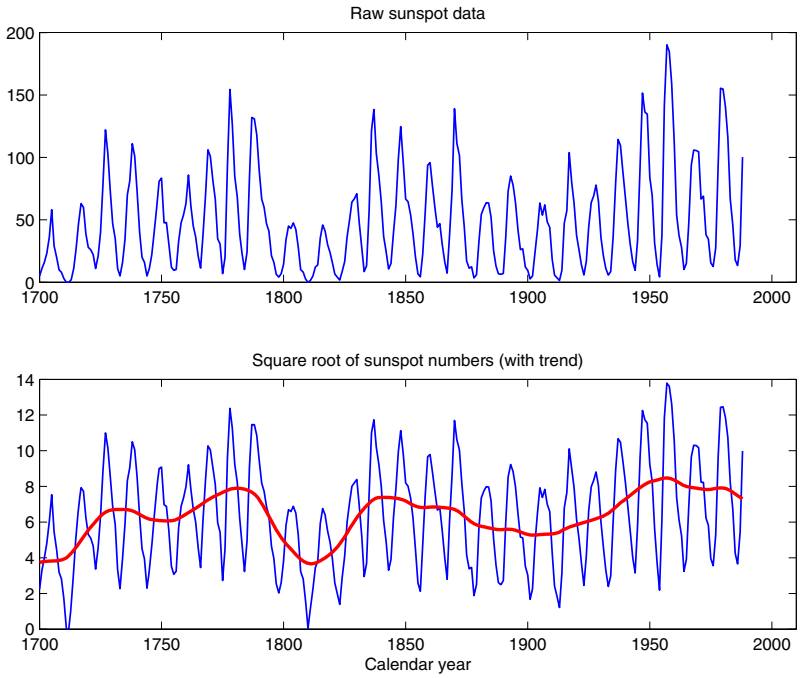


Fig. 5 Yearly averages of sunspot counts. Top: original time series. Bottom: after taking square roots (thin line), with estimated trend (thick line).

John Lilly (`www.jmlilly.net`). Almost circular movements on top of a large-scale trend are clearly visible. Figure 8 shows the time series of longitude and latitude, which each closely resemble a quasi-periodic signal on top of slow trends. Figure 9 shows results for the quasi-periodic model with trend, as applied to latitude.

If we interpret the trend-corrected longitude x as the real part of a complex signal $\exp(\alpha + i\phi)$, and y, the trend-corrected latitude, as the imaginary part, then necessarily we have that $y = \exp\alpha \sin\phi$. This means that $\hat{y} = \exp\hat{\alpha} \sin\hat{\phi}$ should be close to the observed latitude. This is indeed the case, although the amplitude of the longitude is larger, as Figure10 shows. An interpretation is that the buoy moves in orthogonal ellipses of variable size, with variable speed.

Several variants of the model can be envisioned. It might be that the amplitude x has a constant ratio to that of y. Or there might be a constant phase shift of y, relative to x. With the joint model $E(x) = \exp\alpha\cos\phi + \gamma$ and $E(y) = \exp(\alpha+\theta)\sin(\phi+\psi) + \gamma^*$ we can cater for that. In principle there might also be shared patterns in the trends γ and γ^*.

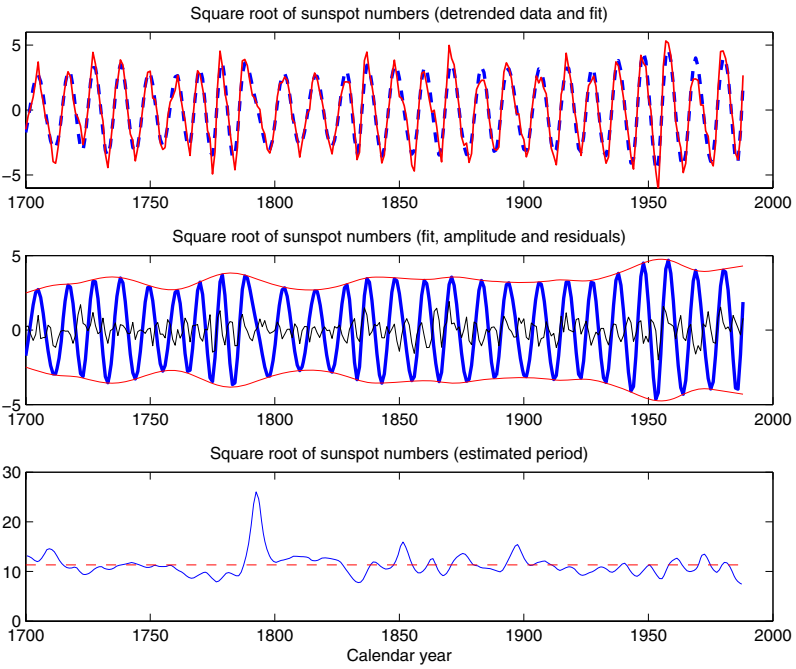


Fig. 6 Model fit to square roots of sunspot counts. Top: model fit (broken line) and signal (full line). Middle: model fit, plus and minus the amplitude, shown as an envelope, and residuals (very small). Bottom: the estimated local period.

3.6 Non-normal Data and Cascaded Links

Up till now least squares were a very reasonable objective function for model fitting, possibly after prior transformation of the data, like for the sunspots. This will not work if we are dealing with counts or binomial data. Generalized linear model technology (Fahrmeir & Tutz 2001) gives us the solution: 1) choose an appropriate data distribution; 2) introduce a link function between model components (the linear predictor) and expected values. In the case of observed counts the Poisson distribution and the logarithmic link function are the obvious choice.

Hoyt & Schatten (1998) proposed the group sunspot number (GSN). They compiled an extensive data set (`http://www.ngdc.noaa.gov/stp/SOLAR/`). It spans the period from 1610 tot 1995 and essentially it contains every observation that was available at the time of compilation. Although the data are organized differently, they can be considered as a series of triples: date, observer ID, sunspot group count. For many days there are multiple observation, especially in modern times, while for others there might be none. Figure 11 shows a short interval. The counts have been jittered vertically to give a better impression of the density of the data points.

The proposed model is

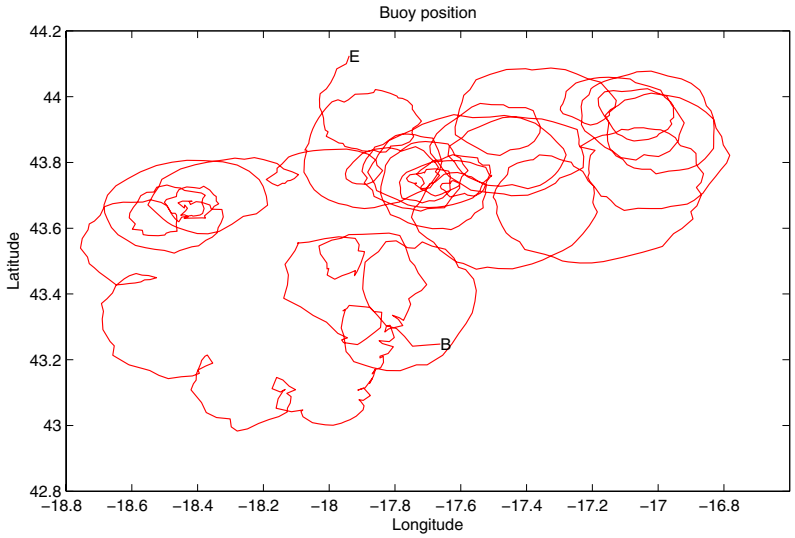


Fig. 7 Trajectory of a free-floating sub-surface buoy during six months. The symbol B marks the beginning and the symbol E the end of the trajectory.

$$\mu = \exp(\exp \alpha_j \cos \phi) + \gamma), \tag{8}$$

or $\mu = \exp(\exp(\alpha) \cos \phi + \gamma)$ if we ignore the imaginary parts. Here we have an unusual cascade of link functions, one of them being complex. It turns out that, with proper starting values, iterative re-weighted regression works well. Convergence seems to be linear. Starting values were obtained by first applying light smoothing with P-splines and interpolating the linear predictor on a dense grid. To this series α, ϕ and γ were fitted with least-squares.

The time interval has been chosen to shed more light on the short cycles around 1793. Only the zero-crossings below 1790 were used to get a starting estimate for ϕ. Beyond 1790 it was linearly extrapolated. This was done because light smoothing (to get starting values) did miss the dip near 1793, and so there was no zero-crossing there. With this special approach a dip was found, as Figure 11 shows. Judging the fit visually, it does not look too impressive around 1790. This is not the right place for an extensive discussion: the example serves as a proof of concept.

3.7 Adding Harmonics

In Section 3.4 we saw an asymmetry in the periods of the de-trended square root of the sunspot numbers. Figure 12 shows this more clearly for a smaller part of the data. The rises quickly and falls slowly. To improve the model, we add harmonics,

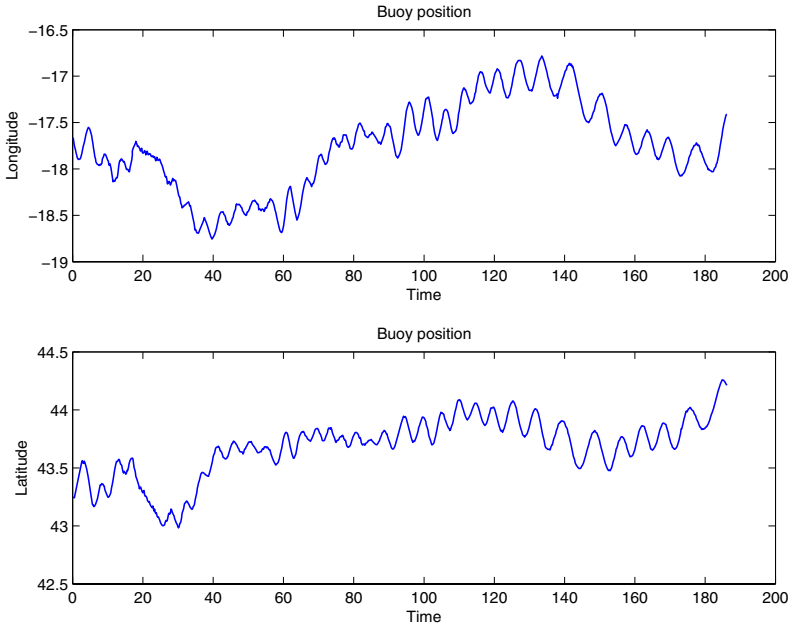


Fig. 8 Buoy data. Latitude and longitude time series.

in the spirit of Fourier analysis. We assume that the amplitude function is shared by all harmonics, so the model becomes

$$\mu = \exp \alpha [\cos \phi + \sum_h \delta_h \cos(h\phi) + \sum_h \varepsilon_h \sin(h\phi)] + \gamma. \tag{9}$$

The range of h generally will be small, say from 2 to 4. Unfortunately the beauty of the complex logarithm is now gone, but the model is effective, as will be seen.

For given α and ϕ, linear regression gives us the parameters δ and ε. The additional trigonometric terms complicate the partial derivatives with respect to α and ϕ, but in a well-structured way. To get started we estimate α and ϕ for the model without harmonics. Then we estimate δ and ε and start a number of iterations to improve amplitude/phase and harmonics parameters in turn. Figure 12 shows results, using second and third harmonics.

4 More to Explore

The complex logarithm is a natural candidate for modeling quasi-periodic data series. A number of model variants and useful application illustrates this claim. Yet I believe

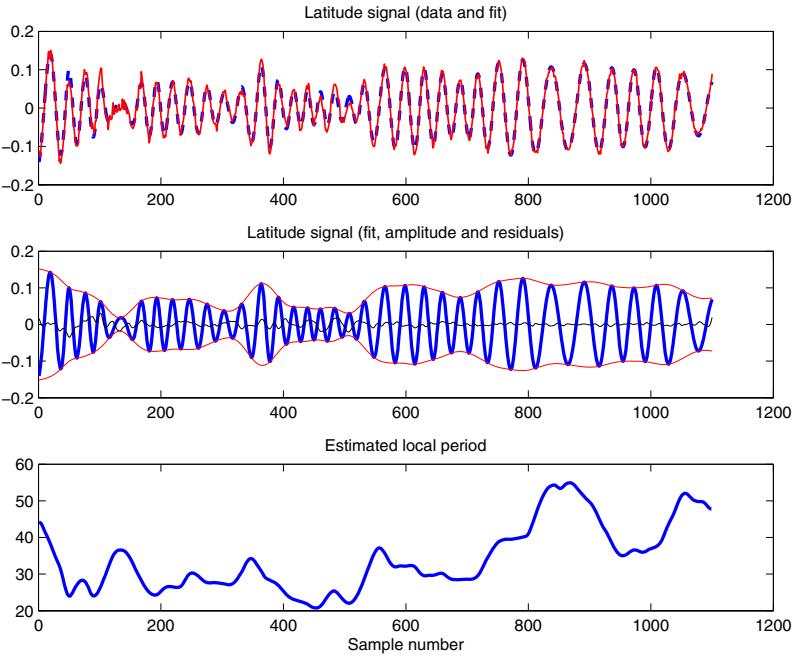


Fig. 9 Fit to the buoy data. Top: model fit (dotted line) and signal (full line). Middle: model fit, plus and minus the amplitude, shown as an envelope, and residuals (very small). Bottom: the estimated local period.

that I only scratched the surface. In this section I suggest a number of more or less obvious extensions and applications.

The examples that were presented before have in common that essentially only one quasi-periodic signal is present. When we study songs of animals like birds and whales, or echo-location signals of bats, we often encounter multiple signals. In special cases these are harmonic of the basic signal, with a different pattern in the amplitude. In other cases the individual signals might be unrelated. We can envision an additive model with several complex logarithm components. A serious problem then is to find starting estimates, because the zero-crossing method will not work anymore. A possible approach is to use time-frequency spectra, derived from wavelet, or other transforms. See also the Discussion.

We have seen examples of interpolation, but extrapolation will work essentially the same: just add “missing data" with zero weights at one or to boundaries and estimate the model. One might expect that much better extrapolated values will be obtained, compared to a standard smoother/extrapolator that does not take the quasi-periodic character of the data into account. It will be very interesting to compare results from the present model with official sunspot number predictions (see `http://solarscience.msfc.nasa.gov/predict.shtml`).

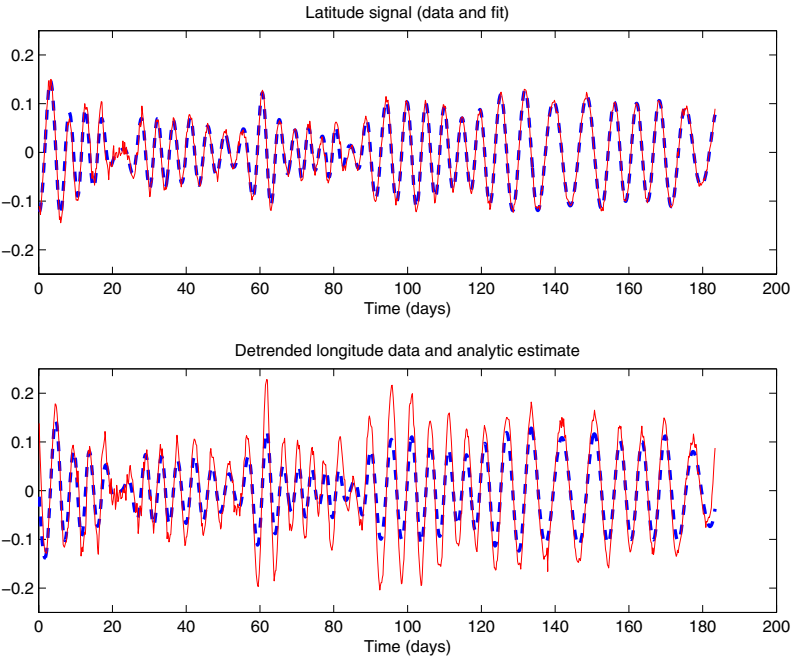


Fig. 10 Illustration of the complex character of the buoy data. Top: latitude data (full line) and model fit $e^{\alpha}\cos\phi$ (broken line). Bottom: longitude data (full line) and analytic estimate $e^{\alpha}\sin\phi$ (broken line).

The Group Sunspot Numbers (GSN) data set poses several fascinating challenges. First there is possible (local) over-dispersion. Visual inspection of the data suggests the existence of "bursts" in which very high counts are observed over several years. The Poisson model might not be adequate for parts or all of the data. Over-dispersion may be explained partially by inter-observer variability. In the GSN data set of almost half a million observations, by 463 different observers identifications occur. Hoyt & Schatten (1998) correct the counts each observer by comparing his/her total with the total of a highly trusted observer on the corresponding days. A more sophisticated approach would be to fit a large scale generalized linear mixed model to the counts, adding observers as a factor to the quasi-periodic component. This way one expects to get more stable correction parameters, because of shrinkage, and a more reliable estimate of GSN time series.

To model a more general shape than the cosine, harmonics were added in the model for the Wolf sunspot counts. Another approach would be to estimate a completely general periodic waveform. Instead of $\exp\alpha\cos\phi$ comes $\exp\alpha f(\lfloor\phi\rceil)$ where $f(.)$ is an arbitrary smooth periodic curve, defined over the domain from 0 to 2π, and $\lfloor\phi\rceil$ is the reduced phase. The latter is a number between 0 and 2π, computed as

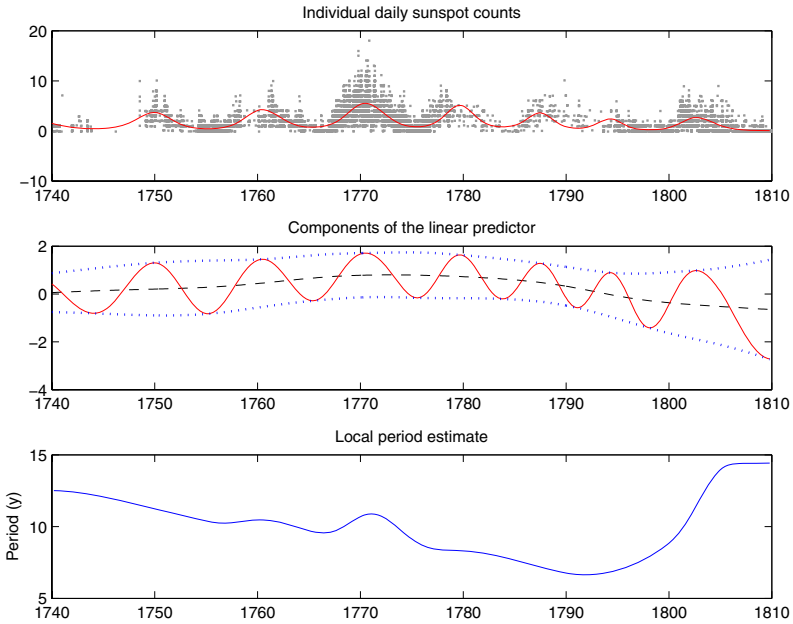


Fig. 11 The cascaded-link quasi-harmonic model applied to daily sunspot counts. Top: 9418 individual counts from 94 observers (dots) and expected values (line). Middle: trend (broken line) plus quasi-harmonic component (full line) and amplitude (dotted line). Bottom: estimated local period.

$\lfloor\phi\rceil = \phi - 2\pi(\lfloor\phi/2\pi\rfloor$, where $\lfloor x\rfloor$ indicates largest (possibly negative) integer less than x.

P-splines can be used to model $f(.)$, with special modifications to define both the basis functions and the penalty on the circle, so that they warp around from 2π to zero. Like with the harmonics, the model has a bilinear flavor and alternating between updating of f and α and ϕ is the logical choice. A model with this flavor, in which f is called the "carrier" wave, was described by Marx et al. (2009), in the simpler setting of seasonal models for monthly counts.

5 Discussion

When I presented parts of this material at a meeting, one reaction from the audience was: "but we already have wavelets!". Indeed it is the case that in a time-frequency presentation based on wavelets one immediately sees changes in frequency as tracks of ridges in a spectrum. But that is only a visual impression: to get the time course of the frequency, ridge-seeking algorithms are needed. Once a ridge is found, its height gives the time course of the amplitude (Carmona et al. 1998). This is far from trivial.

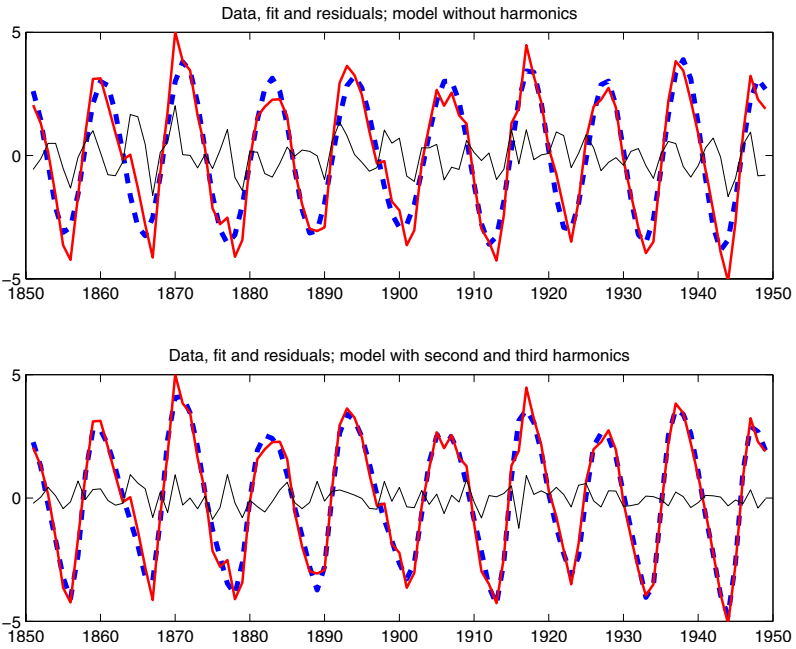


Fig. 12 The quasi-periodic model with harmonics, applied to square roots of sunspot counts (detrended). Top: data (full line), model fit without harmonics (broken line) and residuals (thin line). Bottom: dito, with second and third harmonics.

Wavelet transforms have also problems handling missing or unequally spaced data, and they don't deal well with non-normal data, like counts or binomial observations.

Having said that, it should be noted that wavelet transforms are very good at separately showing multiple quasi-periodic components in a signal. They might be a good tool to find starting estimates for a model with multiple smooth complex logarithm components.

I have not tried (yet) to automatically optimize the values of the penalty parameters (λ_α, λ_ϕ and λ_γ). Instead I played with their values and visually judged results. In an exploratory setting this is not much of a problem. In the applied sciences I expect that researchers will find it useful to play with smoothness to see results in different lights. Of course, all the tools of optimal smoothing (cross-validation, AIC, BIC, mixed model approaches) can be used, in principle. Because we are dealing with time series, correlated noise might be present, in which case automatic methods often lead to under-smoothing. A possible way out then is to model the noise too, e.g. by an autoregressive process (Currie & Durbán 2002).

There is a massive literature on time-frequency representations in physics and engineering; Carmona et al. (1998) is an example. The complex logarithm is a familiar tool in this literature, but I believe that the idea of smooth phase estimation has not received the attention it deserves.

References

Carmona, R., Hwang, W.-L. & Torrésani B. (1998) *Practical Time-Frequency Analysis*. Academic Press.

Currie, I.D. & Durbán, M. (2002) Flexible smoothing with P-splines: a unified approach. *Statistical Modelling* **2**: 333–349.

Eilers, P.H.C. (2003) A Perfect Smoother. *Analytical Chemistry* **75**: 3631–3636.

Eilers, P.H.C. & Marx, B.D. (1996) Flexible Smoothing with Splines and Penalties (with Discussion). *Statistical Science* **11**: 89–121.

Eilers, P.H.C & Marx, B.D. (2002) Generalized Linear Additive Smooth Structures. *Journal of Computational and Graphical Statistics* **11**: 735–751.

Elliott, L. & Hershberger, W. (2006) *The Songs of Insects*. Houghton Mifflin.

Fahrmeir, L. & Tutz, G. (2001) *Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd ed*. Springer.

Hoyt, D.V. & Schatten, K.H. (1998) Group Sunspot Numbers: A New Solar Activity Reconstruction. *Solar Physics* **181**: 491–512.

Lilly, J.M. & Gascard, J.-C. (2006) Wavelet ridge diagnosis of time-varying elliptical signals with application to an oceanic eddy. *Nonlinear Processes in Geophysics* **13**: 467–483.

Marx, B.D., Eilers, P.H.C., Gampe J. & Rau R. (2002) Bilinear Varying-Coefficient Models for Seasonal Time Series and Tables. *Computational Staistics* Published online July 24, 2009.

O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science* **1**: 605–527.

Usoskin, I.G., Mursula, K, Arlt, R & Kovaltsov, G.A. (2009) A solar cycle lost in 1793-1800: Early sunspot observations resolve the old mystery. *Astrophysical Journal Letters* **700**: L154–L157.

Wand, M.P. & Ormerod, J.T. (2008) On Semiparametric Regression with O'Sullivan Penalised Splines. *Australian and New Zealand Journal of Statistics* **50**: 179–198.

P-spline Varying Coefficient Models for Complex Data

Brian D. Marx

Abstract Although the literature on varying coefficient models (VCMs) is vast, we believe that there remains room to make these models more widely accessible and provide a unified and practical implementation for a variety of complex data settings. The adaptive nature and strength of P-spline VCMs allow a full range of models: from simple to additive structures, from standard to generalized linear models, from one-dimensional coefficient curves to two-dimensional (or higher) coefficient surfaces, among others, including bilinear models and signal regression. As P-spline VCMs are grounded in classical or generalized (penalized) regression, fitting is swift and desirable diagnostics are available. We will see that in higher dimensions, tractability is only ensured if efficient array regression approaches are implemented. We also motivate our approaches through several examples, most notably the German deep drill data, to highlight the breadth and utility of our approach.

1 Introduction

The varying coefficient model (VCM) was first introduced by Hastie & Tibshirani (1993). The main idea of the VCM is to allow regression coefficients to vary smoothly (interact) with another variable, thus generating *coefficient curves*. Such coefficient curves can, for example, reflect slow changes in time, depth, or any other indexing regressor. Hence regression coefficients are no longer necessarily constant. Typically estimation for the varying coefficients usually requires the backfitting algorithm, i.e. cycling through and updating each smooth term successively, until convergence. But backfitting also has drawbacks: no information matrix is being computed, so the computation of standard errors and effective model dimension, or efficient leave-one-out (LOOCV) cross-validation is not available. Also convergence can be slow.

Brian D. Marx
Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803 USA
e-mail: bmarx@lsu.edu

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_2,

We have published an efficient fitting algorithm for VCM, based on P-splines (Eilers & Marx, 2002), abbreviated as GLASS (Generalized Linear Additive Smooth Structures). GLASS directly fits all smooths simultaneously of the VCM, without backfitting. In the linear case it converges in one step, and in the generalized linear case it needs only a handful of iterations, similar to the iterative weighted regression for generalized linear models. Standard errors, LOOCV and effective dimension, and diagnostics are readily available at little extra cost. Further optimization is relatively easy and is based on data-driven techniques.

Our GLASS algorithm only considers coefficients that are smooth curves along one dimension (although it allows several of those components). However VCMs can be applied to problems with even richer structure, e.g. coefficients that vary in two or more dimensions and with other additive components in the model. Such data can be generated from modern image or spectral instrument, but can arise naturally from simple tabulations. In principle, using tensor-product P-splines, VCMs can be extended to higher-dimensions, allowing the estimation of (higher dimensional) coefficient surfaces. In theory this is allowed, but in practice one often encounters severe limitations in memory use and computation time. The reason is that large-scale multi-dimensional VCMs need a large basis of tensor products of B-splines. In combination with many observations this can lead to inefficiencies. Consider, as an example, an image of 500×500 pixels, to which one likes to fit a VCM, using a 10 by 10 grid of tensor products of B-splines. The regression basis has 250 thousand rows and 100 columns, or 25 million elements, each taking 200 Mb of memory. With several VCM components storing just the basis can already take on Gigabyte of memory. Computation times to compute inner products will be long. Note that the final system of penalized normal equations is not large with a few hundreds of coefficients. Recently very efficient algorithms have been published for smoothing of multidimensional data arrays with P-splines (Currie, Durbán, & Eilers 2006). They offer improvements of several orders of magnitude in memory use and computation time. With small adaptations, these algorithms can be used for multi-dimensional VCM fitting.

We do not attempt to survey all of the VCM developments. Rather, the major goal of this paper is to provide a unified, accessible, and practical implementation of VCMs using P-splines; one that is conducive to generalizations and tractable in a variety of relatively complex settings, such as two and three-dimensional space-varying GLM extensions, all while avoiding backfitting.

Warm-up: An Intuitive Example

We first illustrate the basic structure and mechanics of a VCM through a simple example. Consider the disk data with the triplets (y_i, x_i, t_i), $i = 1, \ldots, m$, where the response y_i is the *price* (Euro) of an IBM hard drive, the regressor x_i is its *size* (GB), and t_i is the indexing variable *month* (ranging from February 1999 through January 2000). Figure 1 displays the (x, y) scatterplot separately for four selected

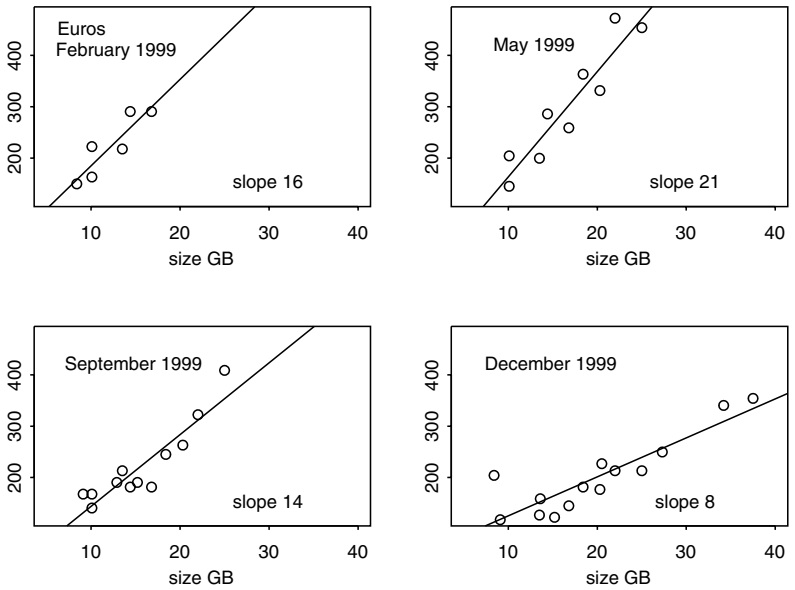


Fig. 1 IBM hard drives: price (Euro) vs. size (GB), at four different months.

months yielding some evidence of a varying (estimated) slope. The VCM combines the monthly data into one model, allowing the slope coefficient to vary smoothly in t. Consider modeling the mean response

$$\mu = x(t)f(t),$$

where $f(t)$ is a smooth slope function. Figure 2 displays the estimated $\hat{f}(t)$ (with twice standard error bands) which strongly suggests that the estimated Euro/GB is decreasing with time. The data points in Figure 2 represent the estimated slopes using the individual monthly data. Note that we are not simply smoothing the points on this graph, but rather borrowing strength from all the monthly data to produce a smooth *coefficient curve*. Such a VCM approach allows for interpolation of Euro/GB for months with missing data (e.g. March and August) or for months with only one observation where slope cannot be estimated. Further we can extrapolate Euro/GB into future months. The details for estimation follow in the coming sections.

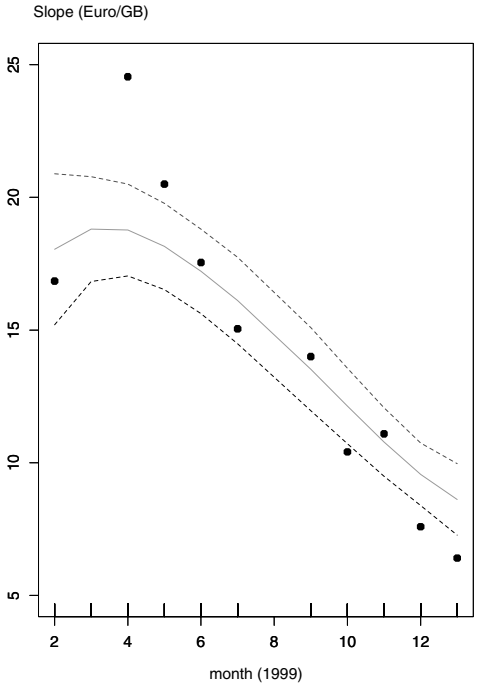


Fig. 2 IBM: Estimated varying slope, combining monthly data. The individual data points represent the estimated slopes using the data month by month. Note that March and August do not have estimate slopes since they have missing data or one observation.

2 "Large Scale" VCM, without Backfitting

The German Continental Deep Drill Program (KTB) was an ambitious project with its aim to study the properties and processes of the upper 10 km of the continental crust (`www.icdp-online.de/sites/ktb/`). The actual drill cuttings comprise of 68 variables measured at each of 5922 depth points (having a 1 m median spacing) down to a final depth of 9.1 km.

We primarily motivate varying coefficient models through the characterization of cataclastic fault zones, and relating the amount of cataclastic rocks (*CATR*), along varying depth, to other variables. Our response is mean amount of *CATR* (which in previous research has been transformed in either units of natural logarithm (log) or log-odds (logit) transformed volume percent), and our central explanatory variables include: Structural water (H_2O), graphite (C), Al_2O_3, Na_2O (all in units weight percent), and *Thermal Conductivity* (in units $Wm^{-1}K^{-1}$).

The KTB statistical models to date only used a subset of depth range. However we find the P-spline VCM is adaptive enough to incorporate the entire range of

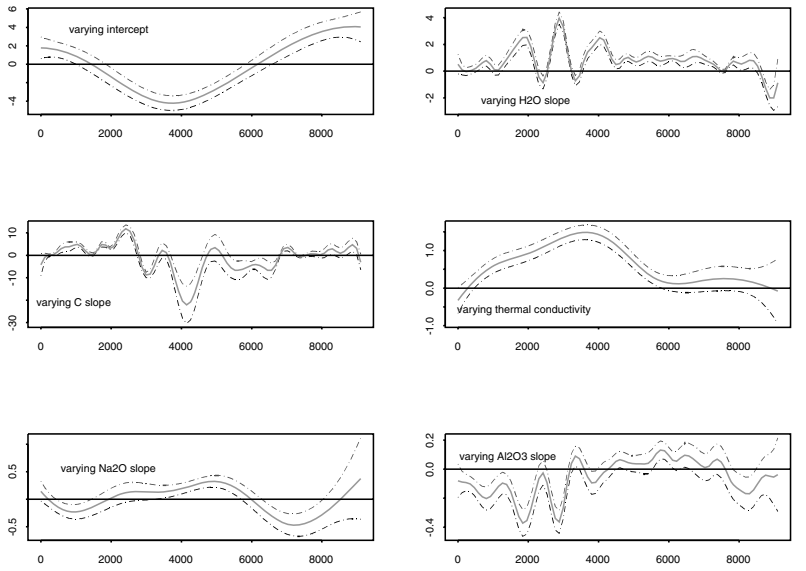


Fig. 3 Using log(*CATR)* as response, varying intercept and varying slopes for H_2O, *C*, *Thermal Conductivity*, Na_2O, Al_2O_3 using cubic ($q = 3$) P-splines with 40 equally-spaced knots, $d = 3$. Optimal tuning parameters chosen by EM. Twice standard bands are provided.

9.1km depth, thereby modelling all data zones simultaneously. The choice of these regressors comes, in part, from existing successful statistical analyses of the KTB data, by e.g. Kauermann & Küchenhoff (2003). These authors modelled the mean and dispersion structure of the amount of cataclastic rocks by focusing on a subset of drill samples ranging from 1000 to 5000 meters, which led to the identification of possible depth breakpoints and potential outliers. Further, Winter et al. (2002) investigated the relationship between the amount of cataclastic rocks to several geological variables using standard regression methods for two specific cataclastic zones within two lithologies: gneiss (1738-2380m) and metabasite (4524-4908m).

It is unrealistic to assume *constant* regression coefficients, along $0 - 9101$m (e.g. associated with H_2O, *C*, Al_2O_3, Na_2O, and *Thermal Conductivity*), and a VCM approach can be a reasonable model choice, thus allowing variables to have depth dependent flexible influence on the response.

Section 5 will provide the details, but to give an idea of how slope coefficients can vary positively and negatively along depth, consider Figure 3 that uses a P-spline VCM. The panels also present twice-standard error bands associated with the varying coefficients. Relative to the zero line we see evidence of reversals or moderating impacts of regressors on *CATR* as depth varies, e.g. *C* appears to have

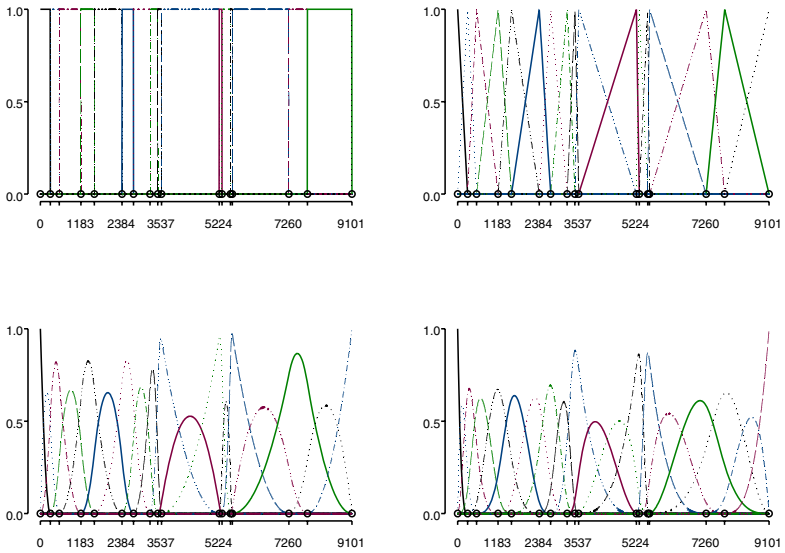


Fig. 4 *B*-spline bases with knots at specific depths: degrees $q = 0, 1, 2, 3$.

positive, negative, and a near zero impact on *CATR*, e.g., at depths of 2300m, 4000m and greater than 7000m, respectively.

The goodness-of-fit measures associated with P-spline VCM shows promise for applications to the KTB data. For example, the models of Winter et al. (2002) that target specific zones, only using a depth range of several hundred meters, reported R^2 values between 0.57– 0.60. Our VCM approach initially show a 12% – 21% improvement, while using the entire 9.1 km range over all data zones. A more thorough presentation of results is given in Section 7.

3 Notation and Snapshot of a Smoothing Tool: B-splines

We will see in the sections that follow that we initially approach smoothness of the coefficient vector (*not* the explanatory variables), in two ways: (a) by modelling coefficients with a B-splines at predetermined depths (knots), and (b) when the number and position of knots is assumed not to be known, by using penalized B-splines or P-splines (Eilers & Marx 1996).

3.1 General Knot Placement

We start with the building block of a complete B-spline basis. The shape of any one B-spline function depends on its degree q. For example, a B-spline takes a constant value (degree $q = 0$), has the form of a triangular density (degree $q = 1$), or can even resemble bell-shaped curves similar to the Gaussian density (e.g. higher degrees $q = 2,\ 3$). A B-spline function has only local support (e.g. in contrast to a Gaussian density). In fact it is constructed from smoothly joining polynomial segments. The positions on the indexing axis, t, where the segments come together, are called the knots. Some general properties of a degree q B-spline include: it consists of $q+1$ polynomial pieces of degree q; the derivatives at the joining points are continuous up to degree $q-1$; the B-spline is positive on the domain spanned by $q+2$ knots, and it is zero elsewhere.

A full B-spline basis is a sequence of B-splines functions along t, each shifted over one knot. Each B-spline is usually indexed by a unique knot, say the leftmost where the B-spline has support. Additional knots must be placed at the boundaries so that each B-spline spans the same number of knots. The knot placement may be general, allowing for unequal spacing. We denote the number of B-splines used in the regression as K, and at any given value of t there are exactly $q+1$ non-zero B-splines, and these values are used to construct the basis matrix B. Given m depths, a $m \times K$ regressor matrix can be constructed. B-spline smoothing is essentially multiple regression. Let $b_{ij} = B_j(t_i)$, $j = 1,\ldots,K$ indicates the value of the jth B-spline function at index t_i, and $B = [b_{ij}]$. The B-spline regressors (and their corresponding parameters) are anonymous in that they do not really have any scientific interpretation: rather predicted values are produced through linear combinations of the basis. We recommend the text by Dierckx (1993) for a nice overview.

Such a basis is well-suited for smoothing of a scatterplot of points (t_i, y_i), $i = 1,\ldots,m$. A smooth mean function can be expressed as $\mu = f(t) = B\alpha$, where B is a $m \times (K+q)$ regressor matrix and α is the unknown B-spline parameters. We minimize

$$S = ||y - B\alpha||^2, \tag{1}$$

with the explicit solution

$$\hat{\alpha} = (B'B)^{-1}B'y. \tag{2}$$

Given $\hat{\alpha}$, the estimated point on the curve at any (new) depth $t^\star$ is $\sum_{j=1}^{K} B_j(t^\star)\hat{\alpha}_j$.

3.2 Smoothing the KTB Data

For the KTB data, $K = 17$ specific knots locations (at depths in meters) are chosen based on prior knowledge of lithologies (Winter et al. 2002), with values: 0, 290, 552, 1183, 1573, 2384, 2718, 3200, 3427, 3537, 5224, 5306, 5554, 5606, 7260, 7800, and 9101 m. The complete B-spline basis (for $q = 0,\ 1,\ 2,\ 3$) using the above knots locations is provided in Figure 4. Using the B-spline bases displayed in

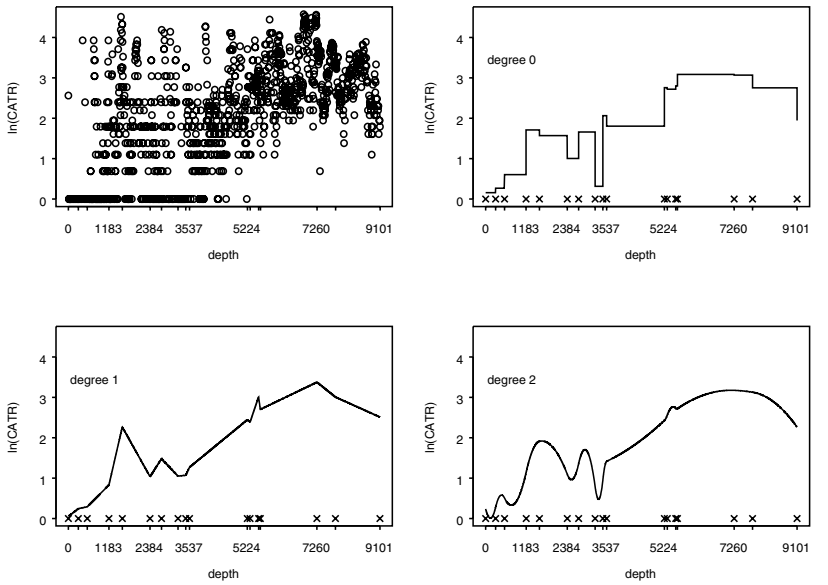


Fig. 5 Scatterplot of log(*CATR)* vs. *depth* and smooth estimated mean functions using B-splines of degree 0, 1, and 2. The "X" symbol indicates knot locations.

Figure 4, Figure 5 displays the estimated smooth mean function for the scatterplot of log(*CATR*) as a function of depth, for various bases degree and the specified $K = 17$ knots.

4 Using B-splines for Varying Coefficient Models

In addition to using smoothing techniques to estimate the mean response, consider broadening the model to control for another regressor, e.g. $x = H_2O$, which itself may also have a varying influence as a function of depth,

$$\mu(t) = \beta_0(t) + x(t)\beta_1(t). \tag{3}$$

This model is a generalization of the simple linear regression model ($\mu = \beta_0 + \beta_1 x$), where the static intercept and slope coefficients (β_0, β_1) are now replaced with coefficients that vary, and thus the regressor has a modified effect, for example depending on depth.

With B-spline smoothing and predetermined knots (along t), backfitting can be avoided and a varying coefficient model can be fit directly. This is clearly illustrated

in matrix notation by modelling the mean response in (3),

$$\begin{aligned}\mu &= B\alpha_0 + \text{diag}\{x(t)\}B\alpha_1 \\ &= (B|U)(\alpha_0', \alpha_1')' = Q\alpha,\end{aligned}$$

where the matrix diag$\{x(t)\}$ aligns the regressors with the appropriate slope value that is also smooth in t, i.e. $\beta_1(t) = B\alpha_1$. Note that the same B basis, built on the t axis, is used for both smooth components. This can be done with data having one natural indexing variable, e.g. as with depth in the KTB data. In general, there can be a different indexing variable for each varying coefficient, thus requiring differing B-spline bases for each additive term. We see that the effective regressors are $Q = (B|U)$, where $U = \text{diag}\{x(t)\}B$, which results in essentially a modest sized "multiple regression" problem. Notice that U boils down to nothing more than a simple row scaling of B. Straightforward least squares techniques similar to (2) are used to estimate the unknown B-spline parameters $\alpha = (\alpha_0, \alpha_1)'$ associated with the smooth intercept and slope. We minimize

$$S = ||y - Q\alpha||^2, \tag{4}$$

with the explicit solution

$$\hat{\alpha} = (Q'Q)^{-1}Q'y. \tag{5}$$

Thus estimated smooth coefficients can be constructed using $B\hat{\alpha}_j$ $(j = 0,\ 1)$, and $\hat{\mu} = Hy = Q\hat{\alpha}$, where the "hat" matrix is $H = Q(Q'Q)^{-1}Q'$.

Additive B-spline VCMs

The generalization to (3) follows for p regressors, each having varying slopes,

$$\mu(t) = \beta_0(t_0) + \sum_{j=1}^{p} \beta_j(t_j)x_j(t_j) \tag{6}$$

In matrix notation,

$$\begin{aligned}\mu &= B\alpha_0 + \sum_{j=1}^{p} \text{diag}\{x_j(t_j)\}B_j\alpha_j \\ &= (B|U_1|\ldots|\,U_p)(\alpha_0', \alpha_1', \ldots \alpha_p')' = R\theta,\end{aligned} \tag{7}$$

where generalizations of (4) and (5) follow naturally using R and θ. Notice that B_j is used in (6) to allow the differing indexing variables (t_j) for each regressor, $j = 1, \ldots,\ p$.

For illustration, Figures 6 and 7 display the fixed knot KTB varying coefficients using B-splines of degree 0 and 3, respectively.

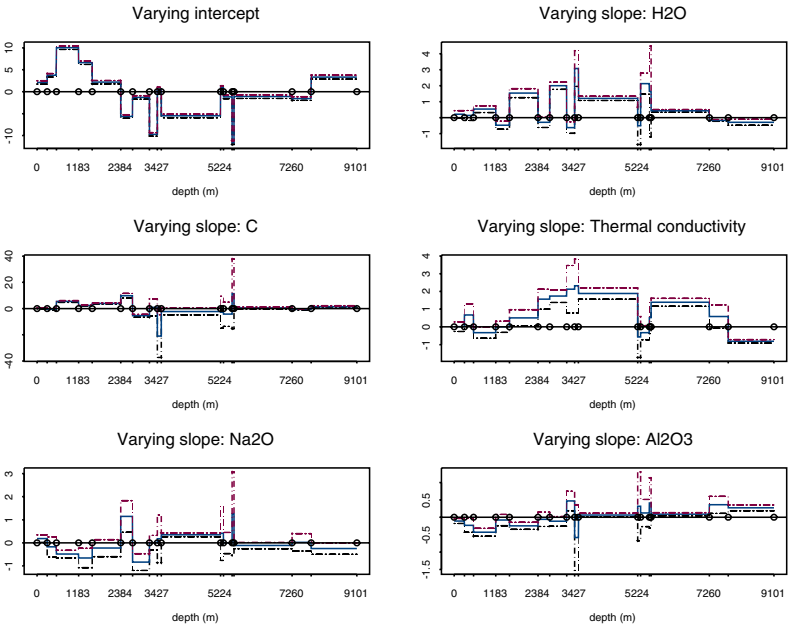


Fig. 6 Using log(*CATR*) as response, varying intercept and varying slopes for H_2O, *C*, *Thermal Conductivity*, Na_2O, Al_2O_3 using B-spline bases of degree 0. Twice standard bands are provided. Knots locations are indicated by both ticks and circles.

5 P-spline Snapshot: Equally-Spaced Knots & Penalization

The B-spline approach in the previous section required knowledge of the location and number of knots. In general, this information may not be known, and the placement of the proper number of knots is a complex nonlinear optimization problem. Circumventing these decisions, Eilers & Marx (1996) proposed an alternative P-spline smoothing approach, which has two steps to achieve smoothness: (i) Use a rich regression basis to purposely overfit the smooth coefficient vector with a modest number of (equally-spaced) B-splines. (ii) Ensuring further and the proper amount of smoothness through a difference penalty on adjacent B-spline coefficients. The main idea is that smoothness is driven by the amplitudes of α, and discouraging estimates of α that have erratic adjacent (neighboring) behavior can be sensible. A non-negative tuning parameter regularizes the influence of the penalty, with large (small) values leading to heavy (light) smoothing. For one smooth term, we now minimize

$$S^\star = ||y - B\alpha||^2 + \lambda ||D_d \alpha||^2. \tag{8}$$

The matrix D constructs dth order differences of α:

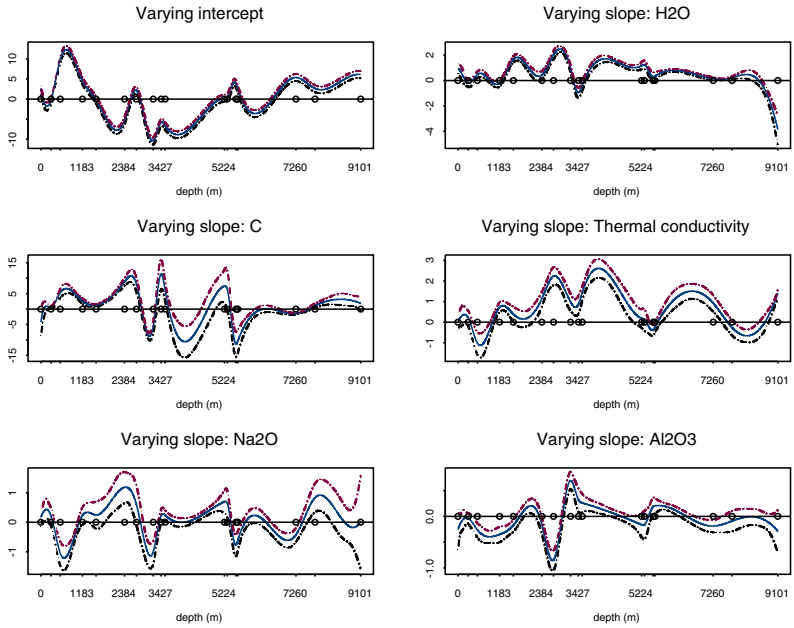


Fig. 7 Using log(*CATR*) as response, varying intercept and varying slopes for H_2O, *C*, *Thermal Conductivity*, Na_2O, Al_2O_3 using B-spline bases of degree 3. Twice standard bands are provided. Knots locations are indicated by both ticks and circles.

$$D_d\alpha = \Delta^d\alpha. \tag{9}$$

The first difference of α, $\Delta^1\alpha$ is the vector with elements $\alpha_{j+1} - \alpha_j$, for $j = 1,\ldots, K-1$. By repeating this computation on $\Delta\alpha$, we arrive at higher differences like $\Delta^2\alpha = \{(\alpha_{j+2} - \alpha_{j+1}) - (\alpha_{j+1} - \alpha_j)\}$ and $\Delta^3\alpha$. The $(n-1)\times n$ matrix D_1 is sparse, with $d_{j,j} = -1$ and $d_{j,j+1} = 1$ and all other elements zero. Examples of D_1 and D_2 of small dimension look like

$$D_1 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}; \quad D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{bmatrix}.$$

Actually, the number of equally-spaced knots does not matter much provided that enough are chosen to ensure more flexibility than needed: the penalty further smoothes with continuous control. The solution of (8) is

$$\hat{\alpha}_\lambda = (B'B + \lambda D_d' D_d)^{-1} B'y, \tag{10}$$

and the “effective" hat matrix is now given by

$$H_\lambda = B(B'B + \lambda D_d' D_d)^{-1} B'. \tag{11}$$

5.1 P-splines for Additive VCMs

When considering more regressor terms and in a VCM context, the model is as outlined in (6) with $\mu(t) = R\theta$, but now B is a rich basis using equally-spaced knots. The P-spline objective function in (8) must be further modified to allow differing flexibility across the p regressors, i.e. a separate λ is allowed for each term. We now have

$$S^\star = ||y - R\theta||^2 + \sum_{j=0}^{p} \lambda_j ||D_d \alpha_j||^2, \tag{12}$$

with a solution

$$\hat{\theta} = (R'R + P)^{-1} R'y,$$

where the penalty takes on the form $P = \text{block diag}(\lambda_0 D_d' D_d, \ldots, \lambda_p D_d' D_d)$. The block diagonal structure breaks linkage of penalization from one smooth term to the next one. Note that (12) uses a common penalty order d, but there is nothing prohibitive from allowing some terms to have different d. Thus

$$\hat{\mu} = R\hat{\theta} = Hy,$$

where $H = R(R'R + P)^{-1} R'$. Borrowing from Hastie & Tibshirani (1990), the effective dimension of the fitted smooth P-spline model is approximately trace(H). By noting the lower dimension and invariance of the trace of cyclical permutated matrices, effective dimension (ED) can be found efficiently using

$$\text{ED}(\lambda) = \text{trace}\{(R'R + P)^{-1} R'R\}. \tag{13}$$

The effective dimension of each smooth term is the trace of the portion of diagonal terms of H corresponding to each term.

5.2 Standard Error Bands

For fixed λ, twice standard error bands can be constructed relatively easily, and can be used as an approximate inferential tool, for example to identify potentially important depth windows that may relate each regressor to the response. We have

$$\text{var}(\hat{\theta}) = (R'R + P)^{-1} R' \sigma^2 I R (R'R + P)^{-1} = \sigma^2 (R'R + P)^{-1} R'R (R'R + P)^{-1}.$$

Thus the covariance matrix associated with the jth smooth component is

$$C_j = \sigma^2 B_j \{(R'R + P)^{-1} R'R (R'R + P)^{-1}\}_j B_j',$$

where $\{\cdot\}_j$ denotes the diagonal block associated with the jth component. The square root of the diagonal elements of C_j are used for error bands, as used in Figure 3. Setting $\lambda = 0$ yields the standard error bands for unpenalized B-splines, as presented in Figures 6 and 7.

6 Optimally Tuning P-splines

For B-spline models, apriori information is essential: The amount of smoothing is determined by the size of the B-spline basis and thus implicitly by the number and position of knots. The smaller the number of knots, the smoother the curve. For P-spline models where R only contains a few smooth terms, cross-validation measures or information criteria can be monitored by varying λ in a systematic way over a grid, and the "optimal" values for the λ vector can be chosen as the one that minimizes, e.g., LOOCV. Although this prediction oriented approach for choosing λ is tractable for low dimensions, it can become computationally taxing and unwieldy, e.g. in our KTB application with six smooth terms. We investigate an alternative estimation-maximization (E-M) approach based on viewing P-splines as mixed models, based on the work of Schall (1991), which appears very promising.

First we consider only one smooth term and then propose a generalized algorithm. Using a mixed model with random α, the log-likelihood, l, can be expressed as

$$-2l = m\log\sigma + n\log\tau + \frac{\|y - B\alpha\|^2}{\sigma^2} + \frac{\|D\alpha\|^2}{\tau^2}, \tag{14}$$

where the $\text{var}(\alpha) = \tau^2$ is the variance of the random effects and $\text{var}(\varepsilon) = \sigma^2$ is the variance of the random error. Maximizing (14) results in the system of equations

$$\left(B'B + \frac{\sigma^2}{\tau^2}D'D\right)\alpha = B'y,$$

and hence we can view $\lambda = \sigma^2/\tau^2$ as a ratio of variances. We also have, under expectation, that

$$\begin{aligned} E(\|y - B\hat{\alpha}\|^2) &\approx (m - ED) \times \sigma^2 \\ E(\|D\hat{\alpha}\|^2) &\approx ED \times \tau^2, \end{aligned} \tag{15}$$

where ED is the approximate effective dimension of the fit. Using (15), we can get a current estimate $\hat{\sigma}^2$ and $\hat{\tau}^2$ from fit. An updated fit can be made using updated $\hat{\sigma}^2/\hat{\tau}^2$, until convergence. We propose a generalized estimation-maximization (E-M) algorithm for the p-dimensional varying coefficient model $\mu = R\theta$:

Algorithm E-M P-spline to optimize $\boldsymbol{\lambda}$

1. Initializations:
   - Generously choose knots K (use 40 as default).
   - Initialize λ, $j = 1, \ldots, p$ (use 10^{-5} as default)
   - Choose B-spline basis degree q (cubic as default)
   - Choose penalty order d (use 3 as default)
   - Construct Penalty $P = \text{blockdiag}(\lambda_0 D'D, \ldots, \lambda_p D'D)$
   - $\hat{\theta} = (R'R + P)^{-1} R'y$
2. Cycle until $\Delta\lambda$ small
3. For $j = 0$ to p
   a. Compute the $ED_j = \text{trace}\{H\}_j$ (jth smooth diagonals in H)
   b. Estimation (E-step):
      i. $\hat{\sigma}^2 = \frac{\|y - R\hat{\theta}\|^2}{m - \sum_{j=0}^{p} \text{ED}_j}$
      ii. $\hat{\tau}_j^2 = \frac{\|D\hat{\theta}\|^2}{\text{ED}_j}$
      iii. $\hat{\lambda}_j = \frac{\hat{\sigma}^2}{\hat{\tau}_j^2}$
   c. Maximization (M-step):
      i. $P = \text{blockdiag}(\hat{\lambda}_0 D'D, \ldots \hat{\lambda}_p D'D)$
      ii. $\hat{\theta} = (R'R + P)^{-1} R'y$
4. Fit with converged vector $\hat{\lambda}$

end algorithm

Cross-validation Prediction Performance

A leave-one-out cross-validation measure can be computed swiftly, only requiring the diagonal elements of the "hat" matrix, h_{ii}, and the residuals $y - \hat{\mu} = y - R\hat{\theta}$. Note

$$y_i - \hat{\mu}_{-i} = (y_i - \hat{\mu}_i)/(1 - h_{ii}), \tag{16}$$

$\hat{\mu} = B(B'B)^{-1}B'y = Hy$, and $\hat{\mu}_{-i}$ is the fitted value for y_i that would be obtained if the model were estimated with y_i left out. It follows that $h_{ii} = r_i'(R'R + P)^{-1} r_i$, where r_i' indicates the ith row of R. Hence the diagonal elements of H and the cross-validation residuals can be computed with little additional work. We can define

$$\text{LOOCV} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{\mu}_{-i})^2} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y_i - \hat{\mu}_i}{1 - h_{ii}}\right)^2},$$

and this result holds in the unpenalized setting by simply setting all $\lambda = 0$ in H.

Table 1 Preliminary goodness-of-fit and cross-validation, by VCM degree.

Method	Basis q	Penalty d	Knots K	Eff. Dim	LOOCV	R^2
E-M P-spline	3	3	equally 40	155.8	0.737	0.704
E-M P-spline	0	1	equally 40	206.1	0.755	0.691
B-spline	3	-	fixed 17	120	0.773	0.683
B-spline	2	-	fixed 17	120	0.765	0.685
B-spline	1	-	fixed 17	120	0.779	0.670
B-spline	0	-	fixed 17	120	0.800	0.647

7 More KTB Results

The P-spline approach was fit using $K = 40$ knots for each of the six smooth components and corresponding difference penalties of order $d = 3$. Summary results are presented in Table 1, for various approaches. For the case $d = 3$, the optimal tuning parameters were chosen using the E-M algorithm above, which converged in 69 cycles, and yielded optimal $\lambda = (570, 0.038, 0.0023, 908, 1252, 6.63)$, respectively. Figure 3 presents the corresponding E-M based estimated smooth coefficients. The convergence criterion was $\max_j\{\Delta\lambda_j/\lambda_j\} < 10^{-8}$. The overall effective dimension of the P-spline fit was ED $= 155.8$. Notice that as λ increases, then ED decreases. When comparing P-splines (Figure 3) to the B-spline approach with unequally-spaced $K = 17$ knots (Figures 6 and 7), we find some general differences. First, the optimal overall ED is higher with P-splines (155.8), when compared to that of each B-spline ED (120), since each B-spline term has an ED=20. Further, some of the P-spline smooth terms need much less ED, e.g. intercept (11.8), *Thermal Conductivity* (15.0), and Na_2O (14.5), whereas other P-spline terms require considerably more ED, e.g. H_2O (40.6), C (34.4), and Al_2O_3 (39.6). We find that the general patterns of negative, positive, and moderate smooth coefficients is preserved from Figures 6 and 7, as a function of depth. However, the P-spline coefficients are smoothed in some cases, and sharpened in others. This P-spline approach required no prior knowledge of depth knots, and yields a very competitive model with an $R^2 = 0.704$– a considerable improvement over previously reported models. The CV value is 0.737, which is the lowest among models presented. Thus the P-VCM model for the KTB data experiences both increase in R^2 and reduction in CV error.

8 Extending P-VCM into the Generalized Linear Model

When responses are non-Normal, e.g. binary outcomes or Poisson counts, the P-spline varying coefficient model extends naturally into the generalized linear model (GLM) framework,

$$g(\mu(t)) = \beta_0(t_0) + \sum_{j=1}^{p} \beta_j(t_j)x_j(t_j)$$

In matrix notation,

$$\begin{aligned} g(\mu) &= B\alpha_0 + \sum_{j=1}^{p} \text{diag}\{x_j(t_j)\}B_j\alpha_j \\ &= (B|U_1|\ldots|U_p)(\alpha_0', \alpha_1', \ldots \alpha_p')' = R\theta, \end{aligned} \tag{17}$$

where the subscript j (on both t and B) highlights that differing indexing variables are allowed for each regressor. The GLM allows a (monotone) link function $g(\cdot)$ and requires independent observations from any member of the exponential family of distribution with $\mu = E(Y)$. The specifics of the GLM are well documented and tabulated, e.g. in Fahrmeir & Tutz (2001, Chapter 2).

The penalized objective function for the GLM is now

$$l^\star = l(\theta) - \sum_{j=0}^{p} \lambda_j ||D_d\alpha_j||^2, \tag{18}$$

where $l(\theta)$ is the log-likelihood function, which is a function of θ since $\mu = h(R\theta)$. The inverse link function is denoted as $h(\cdot)$ (with derivative $h'(\cdot)$). We now maximize $l^\star$ and find above that the penalty terms are now subtracted from $l(\theta)$, thus discouraging roughness of any varying coefficient vector. Fisher's scoring algorithm results in the iterative solution

$$\tilde{\theta}_{c+1} = (R'\tilde{V}_cR + P)^{-1}R'\tilde{V}_c\tilde{z}_c,$$

where again the penalty takes on the form $P = \text{block diag}(\lambda_0 D_d'D_d, \ldots, \lambda_p D_d'D_d)$, and $V = \text{diag}\{h'(R\theta)/\text{var}(y)\}$, $z = (y-\mu)/h'(R\theta) + R\theta$ are the usual GLM diagonal weight matrix and "working" dependent variable, respectively, at the current iteration c. Upon convergence, $\hat{\mu} = h(R\hat{\theta}) = h(\hat{H}y)$, with $\hat{H} = R(R'\hat{V}R + P)^{-1}R'\hat{V}$, and approximate effective dimension $ED \approx \text{trace}\{R'\hat{V}R(R'\hat{V}R + P)^{-1}\}$.

Polio Example with Poisson Counts

We apply P-VCM models to the discrete count time series data of monthly polio incidences in the United States (reported to the U.S. Center of Disease Control) during the years 1970 through 1987. The data are taken from Zeger (1988) and further analyzed by Eilers & Marx (2002). The monthly mean count is modeled with a penalized GLM with a Poisson response and log link function. We choose a model that allows a varying intercept, as well as varying slopes for the cosine and sine regressors (each with both annual and semi-annual harmonics),

$$\log(\mu(t)) = f_0(t) + \sum_{k=1}^{2} \{f_{1k}(t)\cos(k\omega t) + f_{2k}(t)\sin(k\omega t)\}, \tag{19}$$

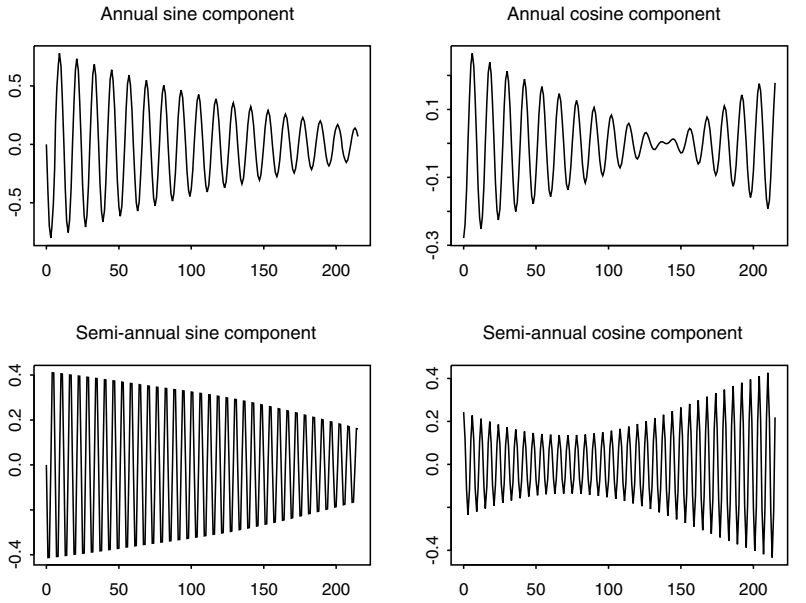


Fig. 8 Polio example: the annual and semi-annual varying cosine and sine effects.

where $\omega = 2\pi/12$ for the index $t = 1, \ldots, 216$. In matrix notation, we have

$$\log(\mu) = B\alpha_0 + \sum_{k=1}^{2} \{C_k B\alpha_{ck} + S_k B\alpha_{sk}\} = R\theta, \tag{20}$$

where $R = (B \mid C_1 B \mid C_2 B \mid S_1 B \mid S_2 B)$ and θ is the corresponding vector of augmented α's. The C and S are diagonal cosine and sine matrices that repeated cycle through the months (1 through 12) using the appropriate harmonic. Since the index is common for all regressors, we conveniently choose to use a common (cubic) basis B. Figure 8 displays the varying harmonic effects. We used 13 equally-spaced knots and a second order penalty for each term. Related to the work of Schall (1991), the optimal values of λ are also found using the E-M algorithm found in Section 6 (with small modification): 1. The estimation of the scale parameter is fixed to be one (step 3.b.i), and 2. Although no backfitting is performed, the maximization step is now the iterative method of scoring (step 3.c.ii). The estimate effective dimension is approximately 6.5 for the intercept, and 1.5 for each of the sine and cosine terms.

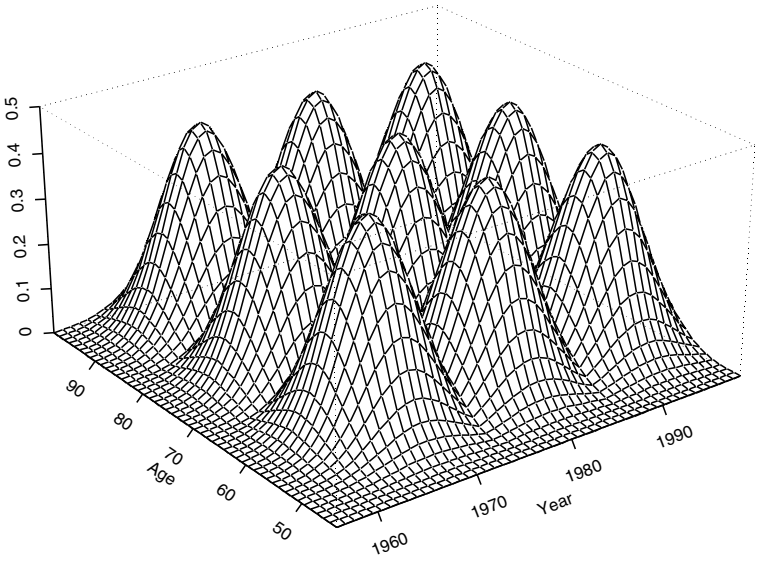


Fig. 9 A sparse portion of a tensor product B-spline basis.

9 Two-dimensional Varying Coefficient Models

An advantage of the P-spline approach to varying coefficient modeling is its ability to adapt to a variety of extensions with relatively little complication. We will see that it is rather straightforward to extend to an additive two-dimensional varying coefficient model in a generalized linear model setting. Such an approach requires P-VCM to use a tensor product B-spline basis and to use some care in constructing a sensible penalty scheme for the coefficients of this basis. In this way P-VCM remains nothing more than a moderately (generalized) penalized regression problem. Consider the tensor product basis provided in Figure 9. The basis is sparsely presented to give an impression of its structure; a full basis would have severe overlapping "mountains". Corresponding to each basis, there is an array of coefficients $\Theta = [\theta_{kl}], k = 1, \ldots, K$ and $l = 1, \ldots, L$ (one for each mountain), and these are the drivers of the two-dimensional varying coefficient surfaces. To avoid the difficult issue of optimal knot placement, P-VCM again takes two steps: (i) Use a rich $K \times L (< 1000)$ gridded tensor product basis that provides more flexibility than needed. (ii) Attach difference penalties on each row and on each column of θ with only one tuning parameter for rows and another one for columns. Figure 10 gives an idea of strong penalization of the coefficients.

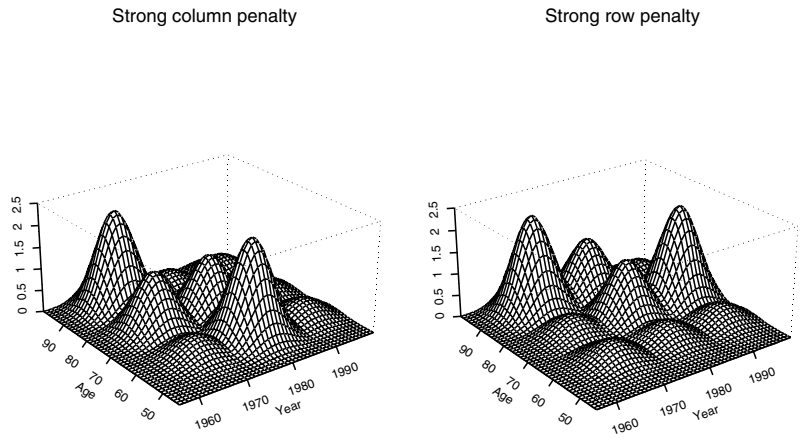


Fig. 10 A sparse portion of a strongly penalized tensor product B-spline basis.

9.1 Mechanics of 2D-VCM through Example

Figure 11 (top panel) displays log death counts resulting from respiratory disease for U.S. females. The image plot is actually 25,440 cells resulting from the cross-classification of age by monthly time intervals. Details of the data, as well as a thorough modeling presentation can be found in Eilers et al. (2008). The lower panel of Figure 11 display the marginal death count over time, which exhibits a strong and varying seasonal cyclical behavior. Consider the Poisson regression with a log link function

$$\log(\mu_{at}) = v_{at} + f_{at}\cos(\omega t) + g_{at}\sin(\omega t) = \eta_{at}, \tag{21}$$

with counts Y_{at} and $\mu_{at} = E(Y_{at})$. For simplicity, we suppress any offset term. The index $a = 1,\ldots,A$ refers to regressor *age* $(44-96)$, whereas year and month are combined to create a variable *time*, indexed by $t = 1,\ldots,T$ $(1-480)$. Annual cyclical behavior in the counts is modeled using the periodic sine and cosine regressors, with period 2π $(\omega = 2\pi/12)$. More harmonics can be added as needed. The two regressors are only indexed with t since the cyclical behavior is only assumed to be associated with *time*. The parameters $v,\ f,\ g$ are indexed by both $(a,\ t)$ and are the smooth (two-dimensional) varying coefficient surfaces for the intercept and slopes for the sine and cosine regressors, respectively.

To express each of the intercept, sine, and cosine varying coefficients smoothly, it is perhaps natural to work with a vectorized form of Θ denoted as $\theta_u = \text{vec}(\Theta_u)$, $u = 0,1,2$. A "flattened" tensor product B-spline basis $\mathbf{B}$ can be formed of dimension $AT \times KL$, such that $\text{vec}(s) = \mathbf{B}\theta_0$, $\text{vec}(f) = \mathbf{B}\theta_1$, and $\text{vec}(g) = \mathbf{B}\theta_2$. Each row of $\mathbf{B}$ designates one of the AT cell counts, and the columns contain the evaluations of each of the KL basis at that cell location. In matrix terms, (21) can be reexpressed as

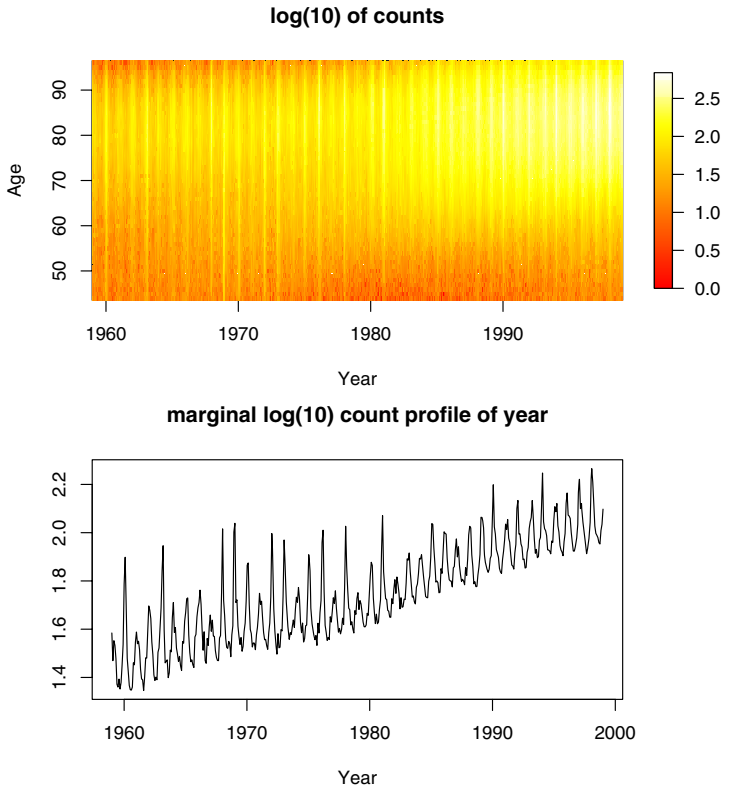


Fig. 11 Raw counts of female respiratory deaths in U.S. for ages 44–96 during 1959-1999 (top) and the marginal plot of time trend (bottom)

$$\begin{aligned} \text{vec}\{\log(\mu)\} &= \mathbf{B}\theta_0 + \text{diag}\left[\cos(\omega t)\right]\mathbf{B}\theta_1 + \text{diag}\left[\sin(\omega t)\right]\mathbf{B}\theta_2 \\ &= \mathbf{B}\theta_0 + \mathbf{U}_1\theta_1 + \mathbf{U}_2\theta_2 \\ &= \mathbf{M}\theta, \end{aligned} \tag{22}$$

where $\mathbf{M} = [\mathbf{B}|\mathbf{U}_1|\mathbf{U}_2]$ and $\theta' = (\theta_0', \theta_1', \theta_2')$ are the augmented bases and tensor product coefficients, respectively. The diagonalization of the regressors in (22) ensures that the each level of the regressor is weighted by its proper level of the varying coefficient. We now find (22) to be a standard Poisson regression model with effective regressors M of dimension $AT \times KL$ and unknown coefficients θ. The dimension of estimation is now reduced from initially $3 \times AT$ to $3 \times KL$.

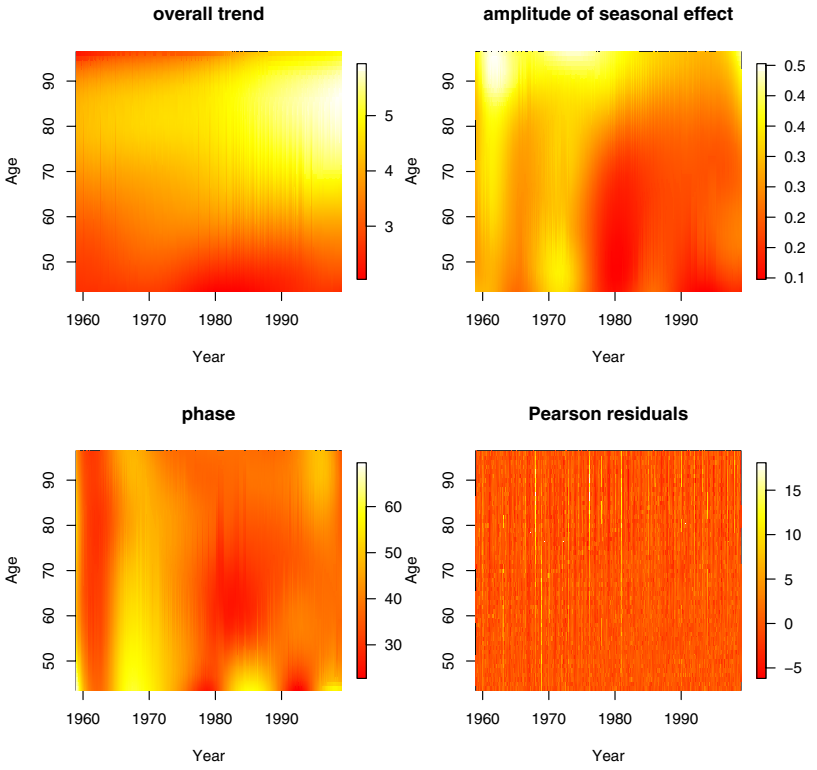


Fig. 12 Fit for female respiratory deaths: varying intercept (trend) (top, left); varying amplitude of (co)sine (top, right); varying phase in months of (co)sine (bottom, left); Pearson residuals (bottom, right)

9.2 VCMs and Penalties as Arrays

Consider the univariate basis: Let $B = [b_{tk}]$ ($\breve{B} = [\breve{b}_{al}]$) be the $T \times K$ ($A \times L$) B-spline basis on the time (age) domain. Denote $\mathscr{A}$, $\mathscr{B}$, and $\mathscr{C}$ as the $K \times L$ matrices of the tensor product coefficients for $V = [v_{ta}]$, $F = [f_{ta}]$, and $G = [g_{ta}]$ respectively. We can rewrite (22) as

$$\begin{aligned} \log(M) &= V + CF + SG \\ &= B\mathscr{A}\breve{B}' + CB\mathscr{B}\breve{B}' + SB\mathscr{C}\breve{B}', \end{aligned} \tag{23}$$

where $M = [\mu_{ta}]$ and C and S represent the (co)sine diagonal matrices defined in (22), and again any offset term is suppressed.

Penalties are now applied to both rows and columns of $\mathscr{A}$ and $\mathscr{B}$. Denote the (second order) difference penalty matrices D and $\breve{D}$ with dimensions $(K-2) \times K$ and $(L-2) \times L$, respectively. Recall Figure 10 that provides a visualization of strong

row and column penalization.The penalty is defined as $P = P_{\mathscr{A}} + P_{\mathscr{B}} + P_{\mathscr{C}}$, with the first term having the form $P_{\mathscr{A}} = \{\lambda_1 ||D\mathscr{A}||_F + \check{\lambda}_1 ||\mathscr{A}\check{D}'||_F\}$ with the other naturally following for $\mathscr{B}$ and $\mathscr{C}$. We denote $||\cdot||_F$ as the Frobenius norm, or the sum of the squares of all elements. The first portion of the penalty is equivalently

$$P_{\mathscr{A}} = \text{vec}(\mathscr{A})'[\lambda_1(I_L \otimes D'D) + \check{\lambda}_1(\check{D}'\check{D} \otimes I_K)]\text{vec}(\mathscr{A}),$$

where I is the identity matrix. The tensor product coefficients, $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{C}$ are found by maximizing the penalized Poisson log-likelihood function

$$l^{\star}(\mathscr{A}, \mathscr{B}) = l(\mathscr{A}, \mathscr{B}) - \frac{1}{2}P. \tag{24}$$

Optimization of the tuning parameters (six in this case) can be found using efficient clever searches, in a greedy way, over the λ space to minimize, e.g. AIC or QIC. Also an extension to the E-M algorithm is possible. Figure 12 presents optimal results based on QIC for the respiratory data using 13×13 equally-spaced tensor products and a second order penalty on rows and columns for each component.

9.3 Efficient Computation Using Array Regression

The array algorithm can be found in Currie et al. (2006). Without loss of generality, using only the first term in (23), the normal equations can be expresses as

$$(\check{B} \otimes B)'W(\check{B} \otimes B)\hat{\alpha} = \mathbf{Q}\hat{\alpha} = (\check{B} \otimes B)'Wy, \tag{25}$$

where W is a diagonal weight matrix and $y = \text{vec}(Y)$. With the dimension of $\check{B} \otimes B$ is $AT \times KL$, and can require much of memory space. Also, but perhaps less obvious, the multiplications and sums that lead to the elements of Q are rather fine-grained and waste an enormous amount of processing time. The problem is compounded when considering all terms in (23). Both problems are eliminated with by rearranging the computations.

Let $R = B \square B$ indicate the row-wise tensor product of B with itself. Hence R has T rows and K^2 columns and each row of R is the tensor product of the corresponding row of B with itself. One can show that the elements of

$$\mathbf{G} = (B \square B)'W(\check{B} \square \check{B})$$

have a one-to-one correspondence to the elements of $\mathbf{Q}$. Of course they are arranged differently, because $\mathbf{Q}$ has dimensions $KL \times KL$ and $\mathbf{G}$ dimensions $K^2 \times L^2$. However, it is easy to rearrange the elements of $\mathbf{G}$ to get $\mathbf{Q}$. Three steps are needed: 1) re-dimension $\mathbf{G}$ to a four-dimensional $K \times K \times L \times L$ array; 2) permute the second and third dimension; 3) re-dimension to a $KL \times KL$ matrix.

A similar, but simpler computation finds the right side of (25) by computing and rearranging $B'(W \cdot Y)\check{B}$, where $W \cdot Y$ indicates the element-wise product of W and

Y. In a generalized additive model or varying-coefficient model with multiple tensor product bases, weighted inner products of the different bases have to be computed using the same scheme as outlined above. Array regression offers very efficient computation with increases in fitting speed (of far more than 10-fold in most cases) when compared to the following unfolded representation. Typically array regression is used when the data are on a regular grid, however it is possible to include a mix of array and other standard regressors.

10 Discussion Toward More Complex VCMs

The adaptive nature and strength of P-splines allows extensions to even more complex models. We have already seen such evidence in this paper by moving from simple to additive P-VCMs, from standard to generalized settings, and from one-dimensional coefficient curves to two-dimensional coefficient surfaces. P-VCMs can also be extended into bilinear models, as presented in Marx et al. (2010). In all cases, P-VCMs further remain grounded in classical or generalized (penalized) regression, allowing swift fitting and desirable diagnostics, e.g. LOOCV.

P-VCMs can be broadened into higher dimensions, e.g. to have three-dimensional varying coefficient surfaces, and with several additive components. Heim et al. (2007) have successfully applied these models to brain imaging applications. Such a model is primarily achieved by broadening the tensor product basis from two to three dimensions and projecting the smooth three-dimensional coefficients onto this lower dimensional space. An additional penalty is needed for the third dimension or layer. In this setting, array regression is of utmost importance due to the formidable dimension of the unfolded design matrix and the number of computations to obtain, e.g., the information matrix. There is nothing prohibitive in P-VCM to consider even higher, e.g. four, dimensional VCM surfaces.

The P-VCM approach also lends itself nicely to high dimensional regression commonly present in chemometric applications, often referred to the multivariate calibration problem. In this setting, the scalar response has digitized “signal" regressors, ones that are ordered and actually ensemble a curve. Marx & Eilers (1999) used P-splines to estimate smooth coefficient curves, but tensor product P-VCMs can allow these smooth coefficient curves to vary over another dimensions. Figure 13 provides an example of how smooth high dimensional coefficient curves can vary over a third variable, temperature. Eilers & Marx (2003) show how to construct such special varying coefficient surfaces, while drawing connections to lower dimensional ribbon models and additive models.

There are various details that will need further investigation. Although it is not always easy to make complete and thorough comparisons across a wide range of other methods under exhaustive settings, it would be interesting to compare the P-VCM approach to Bayesian counterparts (Lang & Brezger 2004), mixed model counterparts (Ruppert, Wand & Carroll 2003) and structural regression approaches (Fahrmeir et al. 2004). Further, we only dampen any effects of serial correlation

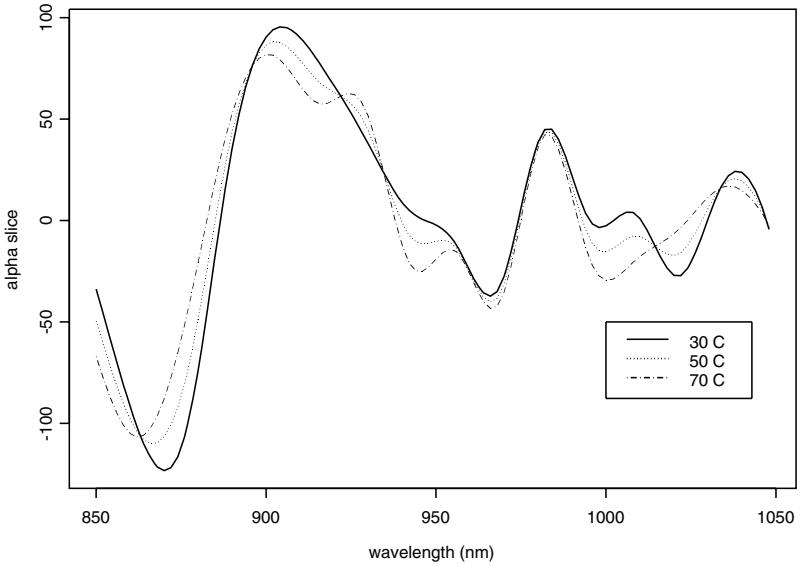


Fig. 13 Various slices of smooth signal regressors that vary over a third variable, temperature

in data through the use of a varying intercept in the model. In fairness, a more formal investigation of any possible auto-regressive (AR) error structure should be made, e.g. addressing deep drill depth varying covariance similarly to Kauermann & Küchenhoff (2003), $\text{corr}(y_i, y_{i+1}) = \rho(\tilde{t})^{|t_i - t_{i+1}|}$, where ρ is a smooth function in depth and $\tilde{t} = (t_i + t_{i+1})/2$. Additionally, although we extended E-M algorithm of Schall (1991) to optimize tuning parameters in the standard and generalized settings, the theory of this approach could be more formally grounded, and the stability of the algorithm should be investigated.

Acknowledgements I would like to thank Paul H.C. Eilers for his generous time and his numerous thought provoking conversations with me that led to a significantly improved presentation.

References

Currie, I. D., Durbán, M. & Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B*, **68**(2): 259–280.

Dierckx, P. (1995). *Curve and Surface Fitting with Splines*. Clarendon Press, Oxford.

Eilers, P. H. C., Gampe, J., Marx, B. D. & Rau, R. (2008). Modulation models for seasonal life tables. *Statistics in Medicine*, **27**(17): 3430–3441.

Eilers, P. H. C. & Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, **66**: 159–174.

Eilers, P. H. C. & Marx, B. D. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**(4): 758–783.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science*, **11**: 89–121.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**: 731–761.

Farhmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (2nd Edition)*. Springer, New York.

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

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, B*, **55**: 757–796.

Heim, S., Fahrmeir, L., Eilers, P. H. C., & Marx, B. D.(2007). Space-varying coefficient models for brain imaging. *Computational Statistics and Data Analysis*, **51**: 6212–6228.

Kauermann, G. & Küchenhoff, K. (2003). Modelling data from inside the Earth: local smoothing of mean and dispersion structure in deep drill data. *Statistical Modelling*, **3**: 43–64.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**(1): 183–212.

Marx, B. D. & Eilers, P. H. C. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, **41**: 1–13.

Marx, B. D., Eilers, P. H. C., Gampe, J. & Rau, R. (2010). Bilinear modulation models for seasonal tables of counts. *Statistics and Computing*. In press.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, New York.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**: 719–727.

Winter, H., Adelhardt, S., Jerak, A. & Küchenhoff, H. (2002). Characteristics of cataclastic shear zones of the ktb deep drill hole by regression analysis of drill cuttings data. *Geophysics Journal International*, **150**: 1–9.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, **75**: 621–629.

Penalized Splines, Mixed Models and Bayesian Ideas

Göran Kauermann

Abstract The paper describes the link between penalized spline smoothing and Linear Mixed Models and how these two models form a practical and theoretically interesting partnership. As offspring of this partnership one can not only estimate the smoothing parameter in a Maximum Likelihood framework but also utilize the Mixed Model technology to derive numerically handy solutions to more general questions and problems. Two particular examples are discussed in this paper. The first contribution demonstrates penalized splines and Linear Mixed Models in a classification context. Secondly, an even broader framework is pursued, mirroring the Bayesian paradigm combined with simple approximate numerical solutions for model selection.

1 Introduction

Since the seminal paper by Eilers & Marx (1996) the use of penalized spline smoothing, or P-spline smoothing as Eilers & Marx coined it, has become more and more popular in applied and recently also in theoretical statistics. The original idea traces back to O'Sullivan (1986) but the real breakthrough occurred with the book by Ruppert, Wand & Carroll (2003) who linked the idea of penalized spline smoothing to Linear Mixed Models (see also Wand 2003). The underlying principle and simple idea is as follows. An unknown smooth function is estimated by replacing the function by a high dimensional basis representation. For estimation a penalty is imposed on the spline coefficients, or to be more precisely on the variation of the spline coefficients, which induces a smooth fit. Making use of quadratic penalties and comprehending the penalty as *a priori* distribution yields a Linear Mixed Model in the classical sense (see Searle, Casella & McCulloch 1992). With this consoli-

Göran Kauermann
Centre for Statistics, Bielefeld University, Dep of Business Administration and Economics, POB 100131, 33501 Bielefeld, Germany, e-mail: gkauermann@wiwi.uni-bielefeld.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_3, © Springer-Verlag Berlin Heidelberg 2010

dation an interesting statistical link occurs. One important practical and compelling benefit of this link is that the smoothing (or penalty) parameter plays the role of the *a priori* (inverse) variance of the spline coefficients, which can now be estimated from the data using Maximum Likelihood. This is implemented for regression smoothing models in **R** (www.r-project.org) in the `Semipar` package accompanying the book of Ruppert, Wand & Carroll (2003) as well as in the newest `mgcv` package by Wood (2006), see also Ngo & Wand (2004).

The practicability and feasibility of penalized spline smoothing is probably one of the reasons why it has been investigated and applied in numerous papers in the recent years. A comprehensive and commendable survey of the last years' research activities in the field of penalized spline smoothing has been composed by Ruppert, Wand & Carroll (2009). We contribute to this work by pursuing a view focusing on the advantages and further possibilities of linking penalized spline smoothing to Linear Mixed Models. The benefit of making use of a Linear Mixed Model for smoothing goes well beyond of just getting an estimate for the smoothing parameter. A wide field of possibilities occurs when testing parametric regression functions against smooth alternatives. In a number of papers it has been shown how the likelihood ratio can be used as test statistics, see for instance Crainiceanu & Ruppert (2004) or Crainiceanu et al. (2005). Moreover the link to Linear Mixed Models shows advantages for instance in model selection (Vaida & Blanchard 2005) or when smoothing is applied to correlated errors (Krivobokova & Kauermann 2007) to just list two examples. The list of advantages is much longer and we refer to the survey article by Ruppert, Wand & Carroll (2009). We here give two further ideas supporting the approach of linking smoothing with Linear Mixed Models. First, we show how penalized splines can be used for classification and secondly we extend the idea by imposing a prior distribution also on the regression parameters. The latter idea is currently more on an experimental level, the first is based on a recent publication (Kauermann, Ormerod & Wand 2009).

2 Notation and Penalized Splines as Linear Mixed Models

As remarked above, the principal idea of penalized spline estimation is simple. Let y be the response variable, which is for the purpose of presentation assumed to be normally distributed with mean $m(x)$ and homoscedastic residual error ε, where $m(.)$ is an unknown smooth function and x a metrically scaled covariate. We replace $m(x)$ by $B(x)b$ with $B(x)$ as K dimensional spline basis located at knots $\tau_1, \ldots, \tau_K$, say. Treating b as parameter vector we impose the penalty $\lambda b^T \tilde{D} b$, where $\tilde{D}$ is an appropriately chosen penalty matrix and λ is the penalty parameter. A convenient choice for $B(.)$ is to use B-splines (see de Boor 1972) and to penalize the variation of coefficients b by taking differences of neighbouring spline coefficients (see Eilers & Marx 1996). Wand & Ormerod (2008) show that this (and other spline settings) can be rewritten to $B(x)b = X(x)\beta + Z(x)u$ with bases matrices $X(.)$ and $Z(.)$ resulting by simple matrix algebra. The penalty term $b^T \tilde{D} b$ is then equivalently formulated

on coefficients u only in the form $u^T Du$, where D is now of full rank and in fact $X(.)$ and $Z(.)$ can be chosen such that D is the identity matrix. Comprehending the quadratic penalty as (proper) *a priori* distribution allows to derive the likelihood for independent observations $(x_i, y_i),\ i = 1, \ldots, n$ from the Linear Mixed Model

$$Y|u \sim N(X\beta + Zu, \sigma_\varepsilon^2 I_n), \quad u \sim N(0, \sigma_u^2 D^{-1}), \tag{1}$$

where $Y = (y_1, \ldots, y_n)^T$ and X and Z are matrices with rows $X(x_i)$ and $Z(x_i)$, respectively, and $\sigma_u^2 = \sigma_\varepsilon^2/\lambda$. We denote the likelihood resulting from (1) by $l(\beta, \sigma_\varepsilon^2, \sigma_u^2)$. This likelihood is also called the *marginal likelihood* since it results by integrating out the random spline effects in (1) yielding the marginal model

$$Y \sim N(X\beta, \sigma_\varepsilon^2 V_\lambda),$$

with $V_\lambda = I + \lambda^{-1} ZD^{-1}Z^T$. It should be clear that the likelihood does also depend on the spline basis and, in particular, on the spline dimension K. It is thereby practical convention that the spline basis is set up before estimation and its dimension K is considered to be chosen generously but fixed and small compared to the sample size and also (to a practical amount) independent of the sample size. This is or has been the main criticism towards penalized spline smoothing, originating particularly from the classical spline smoothing community. In the recent years, the asymptotic properties and how the dimension of the spline basis should grow with the sample size has been under fruitful investigation, see Hall & Opsomer (2005), Li & Ruppert (2008), Kauermann, Krivobokova & Fahrmeir (2009) and Claeskens, Krivobokova & Opsomer (2009). Though these papers shed some light on the $n \to \infty$ scenario, they yield little practical impact on how to select the number of splines for $n < \infty$. The central paper in this respect is Ruppert (2002) who gives a rule of thumb on how to select K, the dimension or number of knots of a spline basis, respectively. Kauermann & Opsomer (2009) argue in this line but utilize the link to Linear Mixed Models.

Most results carry over to generalized response models by assuming that the response y now comes from an exponential family distribution with mean structure

$$E(y|x) = h\big\{m(x)\big\},$$

where $h(.)$ is the known link function. In fact replacing the smooth structure by splines and imposing the penalty on the spline coefficients as prior normality leads to the Generalized Linear Mixed Model (GLMM)

$$\begin{aligned} E(y|x,u) &= h\big\{X(x)\beta + Z(x)u\big\}, \\ u &\sim N\big(0, \sigma_u^2 D^{-1}\big). \end{aligned}$$

The marginal likelihood obtained by integrating out coefficients u is now not any longer analytical, unless y is normal. It can however be shown (see Kauermann, Krivobokova & Fahrmeir 2009) that simple Laplace approximation is justified so that the GLMM approach combined with Laplace integration (see Breslow & Clay-

ton 1993) yields the penalized likelihood fit of the smoother, see also Kauermann & Opsomer (2009). The approach is easily extended to more complex and more structured smoothing models, some of which are considered in the next sections.

3 Classification with Mixed Models

While the intention of penalized spline smoothing is to fit smooth functions we look subsequently at the question of selecting complete smooth functions in the style of model selection. The question is thereby first tackled in the context of classification. In classification the task is to predict a discrete valued variable y given a set of potential classifier variables x. In the simplest scenario, variable y is binary, indicating two groups of observations and x may be metrical or categorical. As example we later make use of the spam email data set provided by Hastie, Tibshirani & Friedman (2001). Here y indicates whether an email is spam $(y = 1)$ or not $(y = 0)$ and x is a high dimensional vector with each component giving the percentage of particular word or character combinations in the email. The intention is to predict $\hat{y}$, that is to classify an incoming email as spam or not spam based on quantities x. The field of classification is thoroughly well developed with numerous successful and competing methods, such as linear or quadratic discriminant analysis, neural networks, support vector machines, classification trees, to just mention a few. A detailed overview is provided in Hastie, Tibshirani & Friedman (2001). We contribute to this field by rewriting the classification problem as Generalized Additive (regression) Model considering y as response and x as covariates. The problem tackled now is to select a parsimonious model with smooth and parametric components. The field of model selection in Generalized Additive Models is thereby well developed, see for instance Wood (2006). We here focus on numerical feasibility and variable selection in the presence of a large number of potential covariates, see also Tutz & Binder (2006).

Assume that vector $x = (x_1, \ldots, x_p)$ contains p metrical covariates and assume that we also have q factorial (here binary) explanatory variables $v = (v_1, \ldots, v_q)$, say. A full generalized additive model for the probability of $y = 1$ would write as

$$\text{logit}\{P(y = 1|x, v)\} = \beta_0 + \sum_{j=1}^{p} m_j(x_j) + \sum_{j=1}^{q} v_j \beta_{vj}, \tag{2}$$

where $m_j(.)$ are smooth but unknown functions. Note that we need additional constraints on $m_j(.)$ to achieve identifiability (see Hastie & Tibshirani 1990) which are omitted here and subsequently for ease of presentation. Replacing the unknown functions by a high dimensional basis allows to replace (2) by

$$\text{logit}\{P(y = 1|x, v)\} = \beta_0 + \sum_{j=1}^{p} X(x_j)\beta_{xj} + \sum_{j=1}^{p} Z(x_j)u_j + \sum_{j=1}^{q} v_j \beta_{vj}. \tag{3}$$

We can fit the model after imposing a penalty on coefficients u_j in the form

$$u_j \sim N\left(0, \sigma_j^2 D_j\right), \quad j = 1, \ldots, p. \tag{4}$$

Note that with (3) and (4) we have constructed a Generalized Linear Mixed Model (GLMM).

There are now two regularizations necessary to make use of (3) in practice. First, spurious coefficients β_{xj} and β_{vj} need to be taken out of the model by setting $\beta_{xj} \equiv 0$ or $\beta_{vj} \equiv 0$. In the same way we need to set $u_j = 0$ if there is no evidence for a functional effect $m_j(x_j)$ in the data. While the first task can be handled in a classical parametric style, e.g. looking at p-values or information criteria, the second task is handled by setting $\sigma_j^2 \equiv 0$ to impose $u_j \equiv 0$. This suggests to select the model in a coherent way by running a forward selection in the following style. The parameters of the full model are $\theta = (\beta_{x1}, \ldots, \beta_{xp}, \sigma_1^2, \ldots, \sigma_p^2, \beta_{v1}, \ldots, \beta_{vq})$. One starts with the null model by setting all parameters to zero and we denote this parameter as $\theta^{(0)}$. Letting $\theta^{(t)}$ denote the parameter after the t-th step in the iteration we calculate for the j-th component of θ with $\theta_j^{(t)} = 0$ the score

$$U_j\left(\theta^{(t)}\right) = \left.\frac{\partial l(\theta)}{\partial \theta_j}\right|_{\theta=\theta^{(t)}}, \tag{5}$$

where $l(\theta)$ is the approximated *marginal* likelihood, that is after integrating out coefficients u_j using a Laplace approximation. Our proposal is to use $U_j\left(\theta^{(t)}\right)$ as selection criterion for the potential parameters in the model. If j refers to an index relating to β_x or β_v then high absolute values of $U_j\left(\theta^{(t)}\right)$ (assuming standardized covariates) indicate that component θ_j should be in the model. Similarly, if j refers to a component out of $\sigma_x^2 = (\sigma_1^2, \ldots, \sigma_p^2)$ then $U_j\left(\theta^{(t)}\right) < 0$ suggest that $\sigma_j^2 = 0$ while large positive values of $U_j\left(\theta^{(t)}\right)$ proposes to allow for a smooth effect of variable x_j in the model. It can be shown (see Kauermann, Ormerod & Wand 2009) that (5) is easily calculated since it is either a standard Wald statistics for index j referring to β_x or β_v or for index j referring to $\theta_j = \sigma_{\tilde{j}}^2$ one gets

$$U_j\left(\theta^{(t)}\right) = -\frac{1}{2}\,\mathrm{tr}\left(Z_{\tilde{j}}^T \hat{W}^{(t)} Z_{\tilde{j}} D_{\tilde{j}}^{-1}\right) + \hat{\varepsilon}^{(t)^T} Z_{\tilde{j}}^T D_{\tilde{j}}^{-1} Z_{\tilde{j}} \hat{\varepsilon}^{(t)}, \tag{6}$$

where $\hat{\varepsilon}^{(t)}$ is the fitted residual vector based on the current parameter estimate $\hat{\theta}^{(t)}$ and $\hat{W}^{(t)}$ is the diagonal weight matrix containing binomial variances. We can now successively include the covariates or smooth functions dependent on the absolute (for β_x and β_v) or positive (for σ_x^2) values of $U_j\left(\theta^{(t)}\right)$. After inclusion of a component we check with an information criterion whether the component should in fact be in or not. To maintain coherence, we propose to make use of the marginal Akaike criterion suggested in Wager, Vaida & Kauermann (2007), see also Vaida & Blanchard (2005). This is defined as

$$\mathrm{mAIC}\left(\hat{\theta}^{(t)}\right) = -2\, l\left(\hat{\theta}^{(t)}\right) + 2\left|\hat{\theta}^{(t)}\right|, \tag{7}$$

with $\left|\hat{\theta}^{(t)}\right|$ referring to the number of elements not set equal to zero. Hence smooth and parametric components are penalized by its numbers of parameters in the marginal likelihood.

The procedure attracts by its coherent style, available due to linking penalized spline smoothing and Generalized Linear Mixed Models. Moreover, and possibly more importantly, the procedure performs promisingly well in practice when being compared to available routines like standard generalized additive models or BRUTO (see Hastie & Tibshirani 1990). It beats these methods, both, in computing time and prediction error, details are provided in Kauermann, Ormerod & Wand (2009).

Data Example

We demonstrate the use of the algorithm with a classical data example in the field of classification. We make use of the 'spam' dataset (see Hastie, Tibshirani & Friedman 2001) which contains data on 4601 emails with 57 metrically scaled potential predictor variables. We could, in principle, fit an generalized additive model for all 57 variables using the `gam(.)` procedure in **R** (see Wood 2006). Alternatively, we can use the suggested forward selection routine combined with the marginal Akaike criterion. Even though the latter is a stepwise routine, it reduces the computing time compared to fitting a generalized additive model with all 57 covariates to about $1/20$. We select 36 covariates out of the 57 and by doing so we can also reduce the classification error from 5.89% for the full additive model to 5.38% for our selected model. The fitted curves $m_j(x_j)$ are shown in Figure 1. More details and studies about the performance of the routine are provided in Kauermann, Ormerod & Wand (2009).

4 Variable Selection with Simple Priors

We now extend the idea of the previous section to pursue a more general model selection by putting prior distributions not only on spline coefficients but also on parametric components.

4.1 Marginal Akaike Information Criterion

Following the notation of the previous chapters we notate the continuous covariates as $x_i = (x_{i1}, \ldots, x_{ip})$ and the (binary) factors as $v_i = (v_{i1}, \ldots, v_{iq})$, $i = 1, \ldots, n$. We assume for simplicity a normal response variable y_i with

$$y_i \sim N\big(m(x_i, v_i), \sigma^2\big),$$

where $m(.,.)$ is the unknown mean structure which needs to be determined. The intention is to find a parsimonious model for $m(x_i, v_i)$ such that

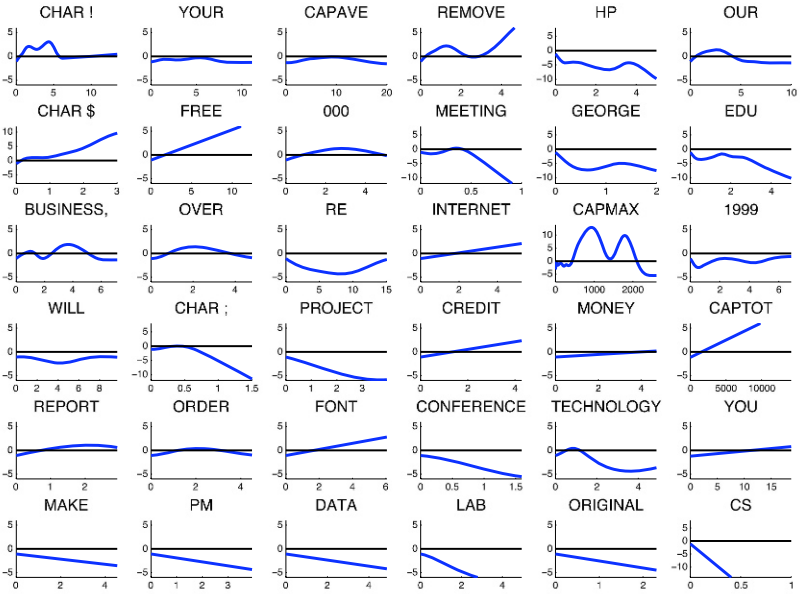


Fig. 1 Effect of percentage of occurrence of words on classification of emails in spam or non-spam.

a) it includes only relevant (significant) variables and
b) continuous covariates x may have functional, non-linear influence if required by the data.

We pursue the model selection exercise by imposing a Bayesian structure on the parameters in the functional estimate of $m(x,v)$. To demonstrate the model and its fitting strategy let us exemplary simplify $m(x_i,v_i)$ to the following structure

$$m(x_i,v_i) = \beta_0 + m_1(x_{i1}) + x_{i2}\beta_{x_2} + v_{i1}\beta_{v_1}, \tag{8}$$

that is covariate x_{i1} has smooth influence while covariate x_{i2} enters the model linearly and so does the factor v_{i1}. As nonparametric estimate of $m_1(x_{i1})$ we make use of penalized spline smoothing with unpenalized slope that is $m_1(x_{i1}) = X_1(x_{i1})\beta_{x_1} + Z_1(x_{i1})u_{x_1}$ with $X_1(x_{i1}) = x_{i1}$ and $Z(\cdot)$ as high dimensional basis. For fitting we now impose prior distributions on all regression coefficients, that is on spline coefficient u_{x_1} as well as on $\beta = (\beta_{x_1}, \beta_{x_2}, \beta_{v_1})$. Using normal priors we get the entire Linear Mixed Model

$$\begin{aligned}
\beta_{x_l} &\sim N(0,\sigma^2_{x_l}), \quad l=1,2,\\
\beta_{v_1} &\sim N(0,\sigma^2_{v_1}),\\
u_1 &\sim N(0,\sigma^2_{u_1} I),\\
Y_i \mid \beta_{x_1},\beta_{x_2},\beta_{v_1},u_1 &\sim N\big(m(x_i,v_i),\sigma^2_\varepsilon\big),
\end{aligned} \tag{9}$$

with $m(x_i,z_i)=\beta_0+x_{i1}\beta_{x_1}+x_{i2}\beta_{x_2}+v_{i1}\beta_{v_1}+Z(x_{i1})u_1$. Note that the prior structure mirrors the Linear Mixed Model but the idea is different to traditional Bayesian priors, like for instance Zellner's (1986) g-prior. We will see, however, that with the random prior structure above we can now run a model selection in a coherent style by checking whether particular components of the *a priori* variances are equal to zero. Apparently, if $\sigma^2_{u_1}\equiv 0$, for instance, then the function $m_1(x_{i1})$ simplifies to the linear component and setting $\sigma^2_{x_2}=0$ corresponds to $\beta_{x_2}\equiv 0$, that is covariate x_2 does not have an effect and should be excluded from the model. We should also impose adequate model hierarchy, for instance by imposing $\sigma_{u_1}\equiv 0$ if $\sigma^2_{x_1}\equiv 0$, that is if the smooth component is included in the model it requires the linear trend to be in the model as well. Let $\mathscr{I}=(\mathscr{I}_x,\mathscr{I}_v)$ be the index set of available metrically scaled covariates and binary factors, respectively and define with $\mathscr{L}=(\mathscr{L}_x,\mathscr{L}_v)\subset\mathscr{I}$ the index set of covariates included in the model. Finally let $\mathscr{S}\subset\mathscr{L}_x$ be the index set of smooth components in the model. For instance, for model (8) we have $\mathscr{L}=\big\{\{x_1,x_2\},\{v_1\}\big\}$ and $\mathscr{S}=\{x_1\}$. We may also include interaction effects as well as smooth interaction effects, but to maintain the presentation notationally simple we restrict the presentation here to additive models only. Let $\mathscr{M}=(\mathscr{L},\mathscr{S})$ denote the resulting model and for an index $j\in\mathscr{M}$ let B_j denote the corresponding basis function and b_j the corresponding coefficient, respectively. For index $j\in\mathscr{L}$ this refers to a single column of observed covariates $B_j=(x_{1j},\ldots,x_{nj})^T$ for $j\in\mathscr{I}_x$ or $B_j=(v_{1j},\ldots,v_{nj})^T$ for $j\in\mathscr{I}_v$, respectively. Accordingly $b_j=\beta_{xj}$ or $b_j=\beta_{vj}$, respectively. For $j\in\mathscr{S}$ we have $B_j=Z(x_j)$ and $b_j=u_j$. The idea is now to choose model $\mathscr{M}$ based on a forward selection routine. We therefore start with a simple model, like in the previous section, for instance the null model $\mathscr{M}=\emptyset$. In the r-th step of the algorithm we denote the current model with $\mathscr{M}^{(r)}$. Note that the parameters of the model are $\beta_0,\sigma^2_\varepsilon$ and the *a priori* variances σ^2_j, $j\in\mathscr{M}^{(r)}$, subsequently denoted by $\sigma^2_{\mathscr{M}^{(r)}}$. We denote with $\hat\beta_0,\hat\sigma^2_\varepsilon,\hat\sigma_{\mathscr{M}^{(r)}}$ the Maximum Likelihood estimates in the current model $\mathscr{M}^{(r)}$. Let now $j\notin\mathscr{M}^{(r)}$ be a potential candidate variable for inclusion in the model. Extending the model by including component j refers to extending $\sigma^2_{\mathscr{M}^{(r)}}$ to $\big(\sigma^2_{\mathscr{M}^{(r)}},\sigma^2_j\big)$ and corresponding likelihood denoted by $l(\beta_0,\sigma^2_\varepsilon,\sigma_{\mathscr{M}^{(r)}},\sigma^2_j)$. With simple matrix algebra it can be shown that

$$\begin{aligned}
l'_{(j)}\big(\hat\sigma^2_{\mathscr{M}^{(r)}}\big) &= \frac{\partial l\big(\hat\beta_0,\hat\sigma^2_\varepsilon,\hat\sigma^2_{\mathscr{M}^{(r)}},\sigma^2_j\big)}{\partial\sigma^2_j}\bigg|_{\sigma^2_j=0}\\
&= \left(-\mathrm{tr}\left(\hat\Sigma^{-1}_{\mathscr{M}^{(r)}}B_j\,B_j^T\right)+\frac{1}{\hat\sigma^2_\varepsilon}\hat E^T_{\mathscr{M}^{(r)}}B_j\,B_j^T\,\hat E_{\mathscr{M}^{(r)}}\right),
\end{aligned} \tag{10}$$

with $\hat{E}_{\mathscr{M}^{(r)}} = Y - \hat{\beta}_0 - \sum_{i \in \mathscr{M}} B_i \hat{b}_i$ as fitted residual vector in model $\mathscr{M}^{(r)}$ and $\hat{\Sigma}_{\mathscr{M}^{(r)}}$ as marginal variance defined through $\left(I + \sum_{i \in \mathscr{M}^{(r)}} \hat{\sigma}_i^2 B_i B_i^T / \hat{\sigma}_\varepsilon^2\right)$. Note that if the derivative in (10) is negative it indicates that the Maximum Likelihood estimator for σ_j^2 is at the boundary of the parameter space and we get $\hat{\sigma}_j^2 = 0$. Hence, we do not want to include component j in the model. If $l'_{(j)}\left(\hat{\sigma}^2_{\mathscr{M}^{(r)}}\right)$ on the other hand is positive we can include component j and calculate the (marginal) Akaike Information Criterion

$$\mathrm{mAIC}\left(\mathscr{M}^{(r)} \cup \{j\}\right) = -2l\left(\hat{\beta}_0, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_{\mathscr{M}^{(r)}}, \hat{\sigma}_j^2\right) - 2|\mathscr{M}^{(r)} \cup \{j\}|, \tag{11}$$

where $|\mathscr{M}|$ denotes the number of variance components in $\sigma^{(r)}_{\mathscr{M}}$. The procedure attracts in so far as model selection is carried out in a completely coherent framework which in fact is numerically quite easy and fast.

4.2 Comparison in Linear Models

To shed more light on the proposed model selection criterion we look at it in a classical scenario. Consider the simple linear model $y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i$ and assume for simplicity that $\beta_0 = 0$. Using the scenario and notation from above we set $b = \beta$ with $b = (b_1, \ldots, b_p)$ and design matrix $X = (X_1, \ldots, X_p)$. The Linear Mixed Model formulation is then

$$b_j \sim N(0, \sigma_j^2),\ j = 1, \ldots, p \ \text{ and } \ Y|b \sim N(Xb, \sigma^2 I_n). \tag{12}$$

After simple calculations we get the maximal marginal log likelihood

$$l\left(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_b^2\right) = -\frac{n}{2}\log\left(\hat{\sigma}_\varepsilon^2\right) - \frac{1}{2}\log\left|\hat{\Sigma}\right|, \tag{13}$$

with $\hat{\Sigma} = I + \sum_{j=1}^p \hat{\sigma}_j^2 X_j X_j^T / \hat{\sigma}_\varepsilon^2$ and

$$\hat{\sigma}_\varepsilon^2 = \frac{(Y - X\hat{b})^T (Y - X\hat{b})}{n - \mathrm{df}}, \quad \hat{\sigma}_j^2 = \frac{\hat{b}_j^2}{\mathrm{df}_j},$$

where df denotes the degree of the model defined through $\mathrm{df} = \sum_{j=1}^p \mathrm{df}_j$ and $\mathrm{df}_j = \mathrm{tr}\left(X_j^T X (X^T X + \hat{\Lambda})^{-1} e_j\right)$ with e_j as j-th dimensional unit vector and $\hat{\Lambda}$ as diagonal matrix having $\hat{\sigma}_\varepsilon^2 / \hat{\sigma}_j^2$ on its diagonal. Note that as $n \to \infty$ and if $\sigma_j^2 > 0$ for all j then $\mathrm{df}_j \to 1$ and hence $\mathrm{df} \to p$. In fact $\mathrm{df} = p + O_p(n^{-1})$. This asymptotic scenario will now be focused in more depth: Let $\hat{\varepsilon} = Y - X\hat{b}$ denote the residual in the Linear Mixed Model and let $\hat{E} = Y - X\hat{\beta}$ be the residual based on the ordinary least squares $\hat{\beta} = (X^T X)^{-1} X^T Y$. Note that $\hat{b} = \hat{\beta} - (X^T X)^{-1} \hat{\Lambda} \hat{\beta}$ so that

$$\hat{\varepsilon} = \hat{E} + X(X^T X)^{-1}\hat{\Lambda}\hat{\beta} + O_p(n^{-2}).$$

Hence, the residual variance estimate $\hat{\sigma}^2_\varepsilon$ can be rewritten as

$$\hat{\sigma}^2_\varepsilon = \frac{RSS}{n-p}(1 + O_p(n^{-1})) = \frac{RSS}{n-p} + O_p(n^{-1}), \tag{14}$$

where $RSS = \hat{E}^T\hat{E}$. This means $\hat{\sigma}^2_\varepsilon$ is asymptotically equal to the bias corrected residual variance estimate in the classical linear model. We denote with ρ_j the j-th eigenvalue of $X^T X\Lambda,\ j = 1,\ldots,p$. If the covariates X are linear independent then $\rho_j = O(n)$ so that $\log|\Sigma| \approx \sum_{j=1}^p \log(\rho_j + 1) \approx p\log(n)$. Looking at (14) and (13) this implies that (13) is asymptotically equal to the traditional Bayesian Information Criterion (BIC). However, the degree of the model itself depends on the design of covariates. To see this, let F be the limiting design matrix in that

$$X^T X/n \xrightarrow[n\to\infty]{} F,$$

and let $\tilde{\rho}_j$ be the eigenvalue of $F\Lambda$ with $\Lambda = \mathrm{diag}(\beta_j^2,\ j = 1,\ldots,p)$. Apparently, $\rho_j = n\tilde{\rho}_j\{1 + O_p(n^{-1})\}$. Employing the derived results we get that the likelihood (13) is asymptotically proportional to

$$l\big(\hat{\sigma}^2_\varepsilon, \hat{\sigma}^2_j, j = 1,\ldots,p\big) = -n\log(RSS) - p - \sum_{j=1}^p \log(n\tilde{\rho}_j + 1).$$

The marginal Akaike Information Criterion (11) results now as

$$\begin{aligned}\mathrm{mAIC}(\mathscr{M}) &= n\log(RSS) + 3p + \sum_{j=1}^p \log(n\tilde{\rho}_j + 1) \\ &\approx n\log(RSS) + 3p + \log(n)\sum_{j=1}^p \log\left(\tilde{\rho}_j + \frac{1}{n}\right).\end{aligned} \tag{15}$$

We can see, that the criterion stands somewhat between the classical AIC and the Bayesian Information Criterion. This becomes clear by assuming the following simplifications. First, assume that covariates are independent with mean zero and variance 1 which yields $\tilde{\rho}_j = \beta_j^2$. Apparently, if $\beta_j^2 = 0$, that is component j is not in the model, we have $\log(n\tilde{\rho}_j + 1) = 0$. In contrast, if all $\beta_j^2 \neq 0$ for $j = 1,\ldots,p$ we have the last component in (15) to depend on the sample size, which mirrors the Bayesian Information Criterion. We exemplify this property in more depth with the following simulation.

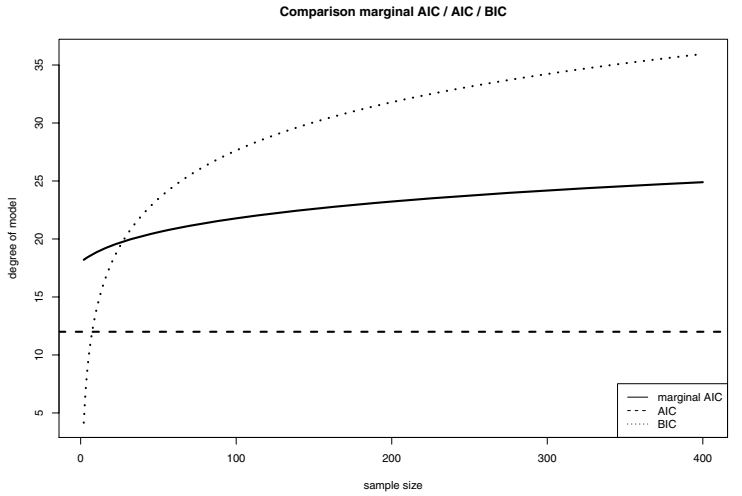


Fig. 2 Degree of the model dependent on the sample size.

4.3 Simulation

To demonstrate the performance of the routine we run two simulation studies. First, we simulate from the following simple linear model $y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2 = 0.5^2)$ and $p = 6$ indendent uniform covariates. We set $\beta = (\beta_1, \ldots, \beta_p) = (1, 0.5, 0.25, 0, 0, 0)$ that is the first three components of $x_i = (x_{i1}, \ldots, x_{i6})$ are in the model and the remaining three covariates are spurious. We run our model selection and compare this with available routines. Note that the classical marginal as well as the usual Akaike Information Criterion (AIC) as well as the Bayesian Information Criterion (BIC) can be written as

$$n \log(RSS) + \text{df},$$

where df stands for degree of the model which is calculated as

- $\text{df} = 3p + \sum_{j=1}^{p} \log(n\tilde{\rho}_j + 1)$ following (15) using the marginal AIC
- $\text{df} = 2p$ for the classical AIC and
- $\text{df} = \log(n)p$ for the standard BIC.

The three terms are shown in Figure 2 for the full model (p=6) and different sample sizes. The marginal Akaike lies between the AIC and the BIC once the sample size is large enough. This can also be seen in the inclusion probabilities of a variable. We therefore run 100 simulations with sample sizes $n = 50, 100, 200$ and 400, respectively and record the empirical inclusion probabilities shown in Table 1. For $n = 50$ the marginal AIC behaves similar to the BIC, but for larger sample

Table 1 Proportion of selecting the components with the different criteria.

	x_1	x_2	x_3	x_4	x_5	x_6
n=50						
marginal AIC	0.97	0.53	0.26	0.07	0.07	0.06
AIC	1.00	0.78	0.38	0.16	0.19	0.21
BIC	0.99	0.57	0.22	0.09	0.06	0.05
n=100						
marginal AIC	1	0.85	0.32	0.06	0.09	0.06
AIC	1	0.96	0.55	0.16	0.24	0.20
BIC	1	0.84	0.25	0.04	0.06	0.02
n=200						
marginal AIC	1	0.97	0.54	0.06	0.11	0.07
AIC	1	1.00	0.73	0.18	0.18	0.17
BIC	1	0.94	0.32	0.01	0.04	0.02
n=400						
marginal AIC	1	1	0.85	0.03	0.12	0.08
AIC	1	1	0.92	0.12	0.22	0.11
BIC	1	1	0.69	0.01	0.03	0.02

sizes it gives a compromise between the AIC and the BIC in that it selects the true functions more often than BIC and in the same way falsely selects spurious functions less frequently than AIC.

For the next simulation we investigate the selection of smooth functions by simulating from the model

$$y_i = m_1(x_{i1}) + m_2(x_{i2}) + \sum_{j=3}^{6} x_{ij}\beta_j + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, \sigma^2 = 0.25)$, $\beta_3 = 2, \beta_4 = 1, \beta_5 = 0, \beta_6 = 0$ and functions $m_1(x_{i1})$ and $m_2(x_{i2})$ as shown in Figure 3. Utilizing the notation from above the true model writes as $\mathscr{L} = \{x_1, x_2, x_3, x_4\}$ and $\mathscr{S} = \{x_1, x_2\}$. We simulate 100 simulations with sample size $n = 100$ and $n = 400$ and report the relative frequencies that a component is selected. The results are shown in Table 2 where we also included the true model indicated as 1 for components which are in the model and 0 otherwise. Apparently the procedure works promissing.

5 Discussion and Extensions

The two ideas described above may be just seen as a tip of an iceberg. In fact, the machinery which becomes available by linking penalized spline smoothing to Linear Mixed Models is not fully exploited yet. The partnership also opens the door to Bayesian modelling and merges the two principles to a coherent data analysis framework. In this respect we suggested to go one step ahead by making use of penalization

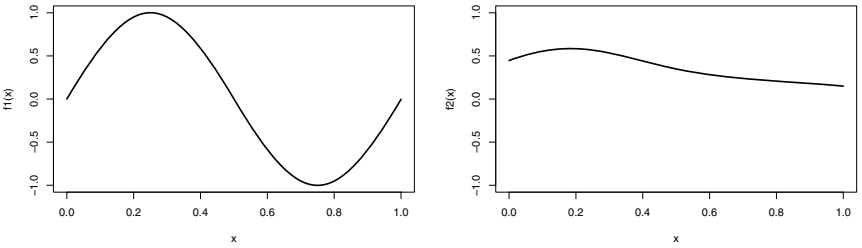


Fig. 3 Simulation function $m_1(x_1)$ and $m_2(x_2)$.

Table 2 Proportion of selecting a covariate linearly or as smooth function.

	x_1	x_2	x_3	x_4	x_5	x_6
linear components $\mathscr{L}$						
true	1	1	1	1	0	0
n=100	1.00	1.00	1.00	0.54	0.06	0.05
n=400	1.00	1.00	1.00	1.00	0.03	0.04
smooth components $\mathscr{S}$						
true	1	1	0	0	0	0
n=100	1.00	0.17	0.04	0.08	0.03	0.04
n=400	1.00	0.73	0.01	0.08	0.05	0.04

concepts for "normal" parameters as well, mirroring just a prior distribution on the parameter itself. It is also interesting to note that the Laplace approximation used quite centrally in the ideas discussed above deserves a better reputation than it used to have. Instead of a "poor man's" computation for those who want to avoid computationally more complex routines like MCMC, it appears to be numerically simple but still accurate. This view has also been recently proposed by Rue, Martino & Chopin (2009) coming from the Bayesian world. All in all, the partnership between penalized spline smoothing and Linear Mixed Models looks in fact flourishing.

References

Breslow, N. E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association* **88**: 9–25.

Claeskens, G., Krivobokova,T. & Opsomer, J. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, to appear.

Crainiceanu, M. & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society B* **66**: 165–185.

Crainiceanu, C., Ruppert, D., Claeskens, G. & Wand, M.P. (2005). Likelihood ratio tests of polynomial regression against a general nonparametric alternative. *Biometrika* **92**: 91–103.

de Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory* **6**: 50–62.

Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**: 89–121.

Hall, P. & Opsomer, J. (2005). Theory for penalised spline regression. *Biometrika* **92**: 105–118.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

Kauermann, G., Krivobokova, T. & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B* **71**: 487–503.

Kauermann, G. & Opsomer, J. (2009). Data-driven selection of the spline dimension in penalized spline regression. *Technical Report.*

Kauermann, G., Ormerod, J. & Wand, M. (2009). Parsimonious classification via generalized linear mixed models. *Journal of Classification*, to appear.

Krivobokova, T. & Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association* **102**: 1328–1337.

Li, Y. & Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, to appear.

Ngo, L. & Wand, M.P. (2004). Smoothing with mixed model software. *Journal of Statistical Software* **9**: 1–54.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science* **1**: 502–518.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B* **71**: 319–392.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**: 735–757.

Ruppert, R., Wand, M. & Carroll, R. (2003). *Semiparametric Regression.* Cambridge: Cambridge University Press.

Ruppert, D., Wand, M. & Carroll, R. (2009). Semiparametric regression during 2003–2007. *Technical Report.*

Searle, S., Casella, G. & McCulloch, C. (1992). *Variance Components.* Wiley, New York.

Tutz, G. & Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics* **62**: 961–971.

Vaida, F. & Blanchard, S. (2005). Conditional akaike information for mixed effects models. *Biometrika* **92**: 351–370.

Wager, C., Vaida, F. & Kauermann, G. (2007). Model selection for P-spline smoothing using Akaike information criteria. *Australian and New Zealand Journal of Statistics* **49**: 173–190.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics* **18**: 223–249.

Wand, M. & Ormerod, J. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics* **50**: 179–198.

Wood, S. (2006). *Generalized Additive Models.* London: Chapman & Hall.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, (eds. P. K. Goel & A. Zellner), pp. 233–243. North-Holland/Elsevier.

Bayesian Linear Regression — Different Conjugate Models and Their (In)Sensitivity to Prior-Data Conflict

Gero Walter and Thomas Augustin

Abstract The paper is concerned with Bayesian analysis under prior-data conflict, i.e. the situation when observed data are rather unexpected under the prior (and the sample size is not large enough to eliminate the influence of the prior). Two approaches for Bayesian linear regression modeling based on conjugate priors are considered in detail, namely the standard approach also described in Fahrmeir et al. (2007) and an alternative adoption of the general construction procedure for exponential family sampling models. We recognize that – in contrast to some standard i.i.d. models like the scaled normal model and the Beta-Binomial / Dirichlet-Multinomial model, where prior-data conflict is completely ignored – the models may show some reaction to prior-data conflict, however in a rather unspecific way. Finally we briefly sketch the extension to a corresponding imprecise probability model, where, by considering sets of prior distributions instead of a single prior, prior-data conflict can be handled in a very appealing and intuitive way.

Key words: Linear regression; conjugate analysis; prior-data conflict; imprecise probability

1 Introduction

Regression analysis is a central tool in applied statistics that aims to answer the omnipresent question how certain variables (called covariates / confounders, regressors, stimulus or independent variables, here denoted by x) influence a certain outcome (called response or dependent variable, here denoted by z). Due to the complexity of real-life data situations, basic linear regression models, where the

Gero Walter and Thomas Augustin
Institut für Statistik, Ludwig-Maximilians-Universität München, D-80539 München, Germany, e-mail: gero.walter@stat.uni-muenchen.de, thomas.augustin@stat.uni-muenchen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_4,

expectation of the outcome z_i simply equals the linear predictor $x_i^T\beta$, have been generalized in numerous ways, ranging from generalized linear models (Fahrmeir & Tutz (2001), see also Fahrmeir & Kaufmann (1985) for classical work on asymptotics) for non-normal distributions of $z_i \mid x_i$, or linear mixed models allowing the inclusion of clustered observations, over semi- and nonparametric models (Kauermann et al. 2009, Fahrmeir & Raach 2007, Scheipl & Kneib 2009), up to generalized additive (mixed) models and structured additive regression (Fahrmeir & Kneib 2009, Fahrmeir & Kneib 2006, Kneib & Fahrmeir 2007).

Estimation in such highly complex models may be based on different estimation techniques such as (quasi-) likelihood, general estimation equations (GEE) or Bayesian methods. Especially the latter offer in some cases the only way to attain a reasonable estimate of the model parameters, due to the possibility to include some sort of prior knowledge about these parameters, for instance by "borrowing strength" (e.g., Higgins & Whitehead 1996).

The tractability of large scale models with their ever increasing complexity of the underlying models and data sets should not obscure that still many methodological issues are a matter of debate. Since the early days of modern Bayesian inference one central issue has, of course, been the potentially strong dependence of the inferences on the prior. In particular in situations where data is scarce or unreliable, the actual estimate obtained by Bayesian techniques may rely heavily on the shape of prior knowledge, expressed as prior probability distributions on the model parameters. Recently, new arguments came into this debate by new methods for detecting and investigating *prior-data conflict* (Evans & Moshonov 2006, Bousquet 2008), i.e. situations where "... the observed data is surprising in the light of the sampling model and the prior, [so that] ... we must be at least suspicious about the validity of inferences drawn." (Evans & Moshonov 2006, p. 893)

The present contribution investigates the sensitivity of inferences on potential prior-data conflict: What happens in detail to the posterior distribution and the estimates derived from it if prior knowledge and what the data indicates are severely conflicting? If the sample size n is not sufficiently large to discard the possibly erroneous prior knowledge and thus to rely on data only, prior-data conflict should affect the inference and should – intuitively and informally – result in an increased degree of uncertainty in posterior inference. Probably most statisticians would thus expect a higher variance of the posterior distribution in situations of prior-data conflict.

However, this is by no means automatically the case, in particular when adopting conjugate prior models, which are often used when data are scarce, where only strong prior beliefs allow for a reasonably precise answer in inference. Two simple and prominent examples of complete insensitivity to prior-data conflict are recalled in Section 2: i.i.d. inferences on the mean of a scaled normal distribution and on the probability distribution of a categorical variable by the Dirichlet-Multinomial model.

Sections 3 and 4 extend the question of (in)sensitivity to prior-data to regression models. We confine attention to linear regression analysis with conjugate priors, because – contrary to the more advanced regression model classes – the linear model still allows a fully analytical access, making it possible to understand potential re-

strictions imposed by the model in detail. We discuss and compare two different conjugate models:

(i) the standard conjugate prior (SCP, Section 3) as described in Fahrmeir et al. (2007) or, in more detail, in O'Hagan (1994); and

(ii) a conjugate prior, called "canonically constructed conjugate prior" (CCCP, Section 4) in the following, which is derived by a general method used to construct conjugate priors to sample distributions that belong to a certain class of exponential families, described, e.g., in Bernardo & Smith (1994).

Whereas the former is the more general prior model, allowing for a very flexible modeling of prior information (which might be welcome or not), the latter allows only a strongly restricted covariance structure for β, however offering a clearer insight in some aspects of the update process.

In a nutshell, the result is that both conjugate models do react to prior-data conflict by an enlarged factor to the variance-covariance matrix of the distribution on the regression coefficients β; however, this reaction is unspecific, as it affects the variance and covariances of all components of β in a uniform way – even if the conflict occurs only in one single component.

Probably such an unspecific reaction of the variance is the most a (classical) Bayesian statistician can hope for, and traditional probability theory based on precise probabilities can offer. Indeed, Kyburg (1987) notes, that

> [...] there appears to be no way, within the theory, of distinguishing between the cases in which there are good statistical grounds for accepting a prior distribution, and cases in which the prior distribution reflects merely ungrounded personal opinion.

and the same applies, in essence, to the posterior distribution.

A more sophisticated modeling would need a more elaborated concept of imprecision than is actually provided by looking at the variance (or other characteristics) of a (precise) probability distribution. Indeed, recently the theory of imprecise probabilities (Walley 1991, Weichselberger 2001) is gaining strong momentum. It emerged as a general methodology to cope with the multidimensional character of uncertainty, also reacting to recent insights and developments in decision theory (see Hsu et al. (2005) for a neuro science corroboration of the constitutive difference of stochastic and non-stochastic aspects of uncertainty in human decision making, in the tradition of Ellsberg's (1961) seminal experiments) and artificial intelligence, where the exclusive role of probability as a methodology for handling uncertainty has eloquently been rejected (Klir & Wierman 1999):

> For three hundred years [...] uncertainty was conceived solely in terms of probability theory. This seemingly unique connection between uncertainty and probability is now challenged [... by several other] theories, which are demonstrably capable of characterizing situations under uncertainty. [...]
>
> [...] it has become clear that there are several distinct types of uncertainty. That is, it was realized that uncertainty is a multidimensional concept. [... That] multidimensional nature of uncertainty was obscured when uncertainty was conceived solely in terms of probability theory, in which it is manifested by only one of its dimensions.

Current applications include, among many other, risk analysis, reliability modeling and decision theory, see de Cooman et al. (2007), Augustin et al. (2009) and

Coolen-Schrijner et al. (2009) for recent collections on the subject. As a welcome byproduct imprecise probability models also provide a formal superstructure on models considered in robust Bayesian analysis (Ríos Insua & Ruggeri 2000) and frequentist robust statistic in the tradition of Huber & Strassen (1973), see also Augustin & Hable (2009) for a review.

By considering *sets* of distributions, and corresponding interval-valued probabilities for events, imprecise probability models allow to express the quality of the underlying knowledge in an elegant way. The higher the ambiguity, the larger c.p. the sets. The traditional concept of probability is contained as a special case, appropriate if and only if there is perfect stochastic information. This methodology allows also for a natural handling of prior-data conflict. If prior and data are in conflict, the set of posterior distributions are enlarged, and inferences become more cautious.

In Section 5 we briefly report that the CCCP model has a structure that allows a direct extension to an imprecise probability model along the lines of Quaeghebeur & de Cooman's (2005) imprecise probability models for i.i.d. exponential family models. Extending the models further by applying arguments from Walter & Augustin (2009) yields a powerful generalization of the linear regression model that is also capable of a component-specific reaction to prior-data conflict.

2 Prior-data Conflict in the i.i.d. Case

As a simple demonstration that conjugate models might not react to prior-data conflict reasonably, inference on the mean of data from a scaled normal distribution and inference on the category probabilities in multinomial sampling will be described in the following examples 1 and 2.

Example 1 (Samples from a scaled Normal distribution $\mathrm{N}(\mu,1)$*).* The conjugate distribution to an i.i.d.-sample x of size n from a scaled normal distribution with mean μ, denoted by $\mathrm{N}(\mu,1)$ is a normal distribution with mean $\mu^{(0)}$ and variance ${\sigma^{(0)}}^2$[1]. The posterior is then again a normal distribution with the following updated parameters:

$$\mu^{(1)} = \frac{\frac{1}{n}}{\frac{1}{n}+{\sigma^{(0)}}^2}\mu^{(0)} + \frac{{\sigma^{(0)}}^2}{\frac{1}{n}+{\sigma^{(0)}}^2}\bar{x} = \frac{\frac{1}{{\sigma^{(0)}}^2}}{\frac{1}{{\sigma^{(0)}}^2}+n}\mu^{(0)} + \frac{n}{\frac{1}{{\sigma^{(0)}}^2}+n}\bar{x} \qquad (1)$$

$${\sigma^{(1)}}^2 = \frac{{\sigma^{(0)}}^2 \cdot \frac{1}{n}}{{\sigma^{(0)}}^2+\frac{1}{n}} = \frac{1}{\frac{1}{{\sigma^{(0)}}^2}+n}. \qquad (2)$$

[1] Here, and in the following, parameters of a prior distribution will be denoted by an upper index $^{(0)}$, whereas parameters of the respective posterior distribution by an upper index $^{(1)}$.

The posterior expectation (and mode) is thus a simple weighted average of the prior mean $\mu^{(0)}$ and the estimation from data $\bar{x}$, with weights $\frac{1}{{\sigma^{(0)}}^2}$ and n, respectively.[2] The variance of the posterior distribution is getting smaller automatically.

Now, in a situation where data is scarce but with prior information one is very confident about, one would choose a low value for ${\sigma^{(0)}}^2$, thus resulting in a high weight for the prior mean $\mu^{(0)}$ in the calculation of $\mu^{(1)}$. The posterior distribution will be centered around a mean between $\mu^{(0)}$ and $\bar{x}$, and it will be even more pointed as the prior, because ${\sigma^{(1)}}^2$ is considerably smaller than ${\sigma^{(0)}}^2$ as the factor to ${\sigma^{(0)}}^2$ in (2) is quite smaller than one.

The posterior basically would thus say that one can be quite sure that the mean μ is around $\mu^{(1)}$, regardless if $\mu^{(0)}$ and $\bar{x}$ were near to each other or not, where the latter would be a strong hint on prior-data conflict. The posterior variance does not depend on this; the posterior distribution is thus insensitive to prior-data conflict.

Even if one is not so confident about one's prior knowledge and thus assigning a relatively large variance to the prior, the posterior mean is less strongly influenced by the prior mean, but the posterior variance still is getting smaller no matter if the data support the prior information or not.

The same insensitivity appears also in the widely used Dirichlet-Multinomial model:

Example 2 (Samples from a Multinomial distribution $M(\theta)$). Given a sample of size n from a multinomial distribution with probabilities θ_j for categories / classes $j = 1,\ldots,k$, subsumed in the vectorial parameter θ (with $\sum_{j=1}^{k}\theta_j = 1$), the conjugate prior on θ is a Dirichlet distribution $\mathrm{Dir}(\alpha^{(0)})$. Written in terms of a reparameterization used e.g. in Walley (1996), $\alpha_j^{(0)} = s^{(0)} \cdot t_j^{(0)}$ such that $\sum_{j=1}^{k} t_j^{(0)} = 1$, $(t_1^{(0)},\ldots,t_k^{(0)})^\top =: t^{(0)}$, it holds that the components of $t^{(0)}$ have a direct interpretation as prior class probabilities, whereas $s^{(0)}$ is a parameter indicating the confidence in the values of $t^{(0)}$, similar to the inverse variance as in Example 1, and the quantity $n^{(0)}$ in Section 4.[3]

The posterior distribution, obtained after updating via Bayes' rule with a sample vector $(n_1,\ldots,n_k)$, $\sum_{j=1}^{k} n_j = n$ collecting the observed counts in each category, is a Dirichlet distribution with parameters

$$t_j^{(1)} = \frac{s^{(0)}}{s^{(0)}+n} t_j^{(0)} + \frac{n}{s^{(0)}+n} \cdot \frac{n_j}{n}, \qquad s^{(1)} = s^{(0)} + n.$$

The posterior class probabilities $t^{(1)}$ are calculated as a weighted mean of the prior class probabilities and $\frac{n_j}{n}$, the proportion in the sample, with weights $s^{(0)}$ and n, respectively; the confidence parameter is incremented by the sample size n.

Also here, there is no systematic reaction to prior-data conflict. The posterior variance for each class probability θ_j calculates as

[2] The reason for using these seemingly strange weights will become clear later.

[3] If $\theta \sim \mathrm{Dir}(s,t)$, then $\mathbb{V}(\theta_j) = \frac{t_j(1-t_j)}{s+1}$. If s is high, then the variances of θ will become low, thus indication high confidence in the chosen values of t.

$$\mathbb{V}(\theta_j \mid n) = \frac{t_j^{(1)}(1-t_j^{(1)})}{s^{(1)}+1} = \frac{t_j^{(1)}(1-t_j^{(1)})}{s^{(0)}+n+1}\,.$$

The posterior variance depends heavily on $t_j^{(1)}(1-t_j^{(1)})$, having values between 0 and $\frac{1}{4}$, which do not change specifically to prior data conflict. The denominator increases from $s^{(0)}+1$ to $s^{(0)}+n+1$. Imagine a situation with strong prior information suggesting a value of $t_j^{(0)} = 0.25$, so one could choose $s^{(0)} = 5$, resulting in a prior class variance of $\frac{1}{32}$. When observing a sample of size $n = 10$ all belonging to class j (thus $n_j = 10$), being in clear contrast to the prior information, the posterior class probability is $t_j^{(1)} = 0.75$, resulting the enumerator value of the class variance to remain constant. Therefore, due to the increasing denominator, the variance decreases to $\frac{3}{256}$, in spite of the clear conflict between prior and sample information. Of course, one can construct situations where the variance increases, but this happens only in case of an update of $t_j^{(0)}$ towards $\frac{1}{2}$. If $t_j^{(0)} = \frac{1}{2}$, the variance will decrease for any degree of prior-data conflict.

3 The Standard Approach for Bayesian Linear Regression (SCP)

The regression model is noted as follows:

$$z_i = x_i^{\mathsf{T}}\beta + \varepsilon_i\,, \quad x_i \in \mathbb{R}^p\,,\ \beta \in \mathbb{R}^p\,,\ \varepsilon_i \sim \mathrm{N}(0,\sigma^2)\,,$$

where z_i is the response, x_i the vector of the p covariates for observation i, and β is the p-dimensional vector of adjacent regression coefficients.

The vector of regressors x_i for each observation i is generally considered to be non-stochastic, thus it holds that $z_i \sim \mathrm{N}(x_i^{\mathsf{T}}\beta,\sigma^2)$, or, for n i.i.d. samples, $z \sim \mathrm{N}(\mathbf{X}\beta,\sigma^2\mathbf{I})$, where $z \in \mathbb{R}^n$ is the column vector of the responses z_i, and $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the *design matrix*. Without loss of generality, one can either assume $x_{i1} = 1\ \forall i$ such that the first component of β is the intercept parameter[4], or consider only centered responses z and standardized covariates to make the estimation of an intercept unnecessary.

In Bayesian linear regression analysis, the distribution of the response z is interpreted as a distribution of z given the parameters β and σ^2, and prior distributions on β and σ^2 must be considered. For this, it is convenient to split the joint prior on β and σ^2 as $p(\beta,\sigma^2) = p(\beta \mid \sigma^2)p(\sigma^2)$ and to consider conjugate distributions for both parts, respectively.

In the literature, the proposed conjugate prior for $\beta \mid \sigma^2$ is a normal distribution with expectation vector $m^{(0)} \in \mathbb{R}^p$ and variance-covariance matrix $\sigma^2\boldsymbol{M}^{(0)}$, where $\boldsymbol{M}^{(0)}$ is a symmetric positive definite matrix of size $p \times p$. The prior on σ^2 is an inverse gamma distribution (i.e., $1/\sigma^2$ is gamma distributed) with parameters $a^{(0)}$ and $b^{(0)}$, in the sense that $p(\sigma^{-2}) \propto (\sigma^{-2})^{a^{(0)}+1}\exp\{-b^{(0)}\sigma^{-2}\}$. The joint prior on

[4] usually denoted by β_0; however, we stay with the numbering $1,\ldots,p$ for the components of β.

$\theta = (\beta, \sigma^2)^\mathsf{T}$ is then denoted as a normal – inverse gamma (NIG) distribution. The derivation of this prior and the proof of its conjugacy can be found, e.g., in Fahrmeir et al. (2007) or in O'Hagan (1994), the latter using a different parameterization of the inverse gamma part, where $a^{(0)} = \frac{d}{2}$ and $b^{(0)} = \frac{a}{2}$.

For the prior model, it holds thus that (if $a^{(0)} > 1$ resp. $a^{(0)} > 2$)

$$\begin{aligned} \mathbb{E}[\beta \mid \sigma^2] &= m^{(0)}\,, \qquad & \mathbb{V}(\beta \mid \sigma^2) &= \sigma^2 \boldsymbol{M}^{(0)}\,, \\ \mathbb{E}[\sigma^2] &= \frac{b^{(0)}}{a^{(0)}-1}\,, \qquad & \mathbb{V}(\sigma^2) &= \frac{(b^{(0)})^2}{(a^{(0)}-1)^2(a^{(0)}-2)}\,. \end{aligned} \tag{3}$$

As σ^2 is considered as nuisance parameter, the unconditional distribution on β is of central interest because it subsumes the shape of prior knowledge on β as expressed by the choice of parameters $m^{(0)}$, $\boldsymbol{M}^{(0)}$, $a^{(0)}$ and $b^{(0)}$. It can be shown that $p(\beta)$ is a multivariate noncentral t distribution with $2a^{(0)}$ degrees of freedom, location parameter $m^{(0)}$ and dispersion parameter $\frac{b^{(0)}}{a^{(0)}}\boldsymbol{M}^{(0)}$, such that

$$\mathbb{E}[\beta] = m^{(0)}\,, \qquad \mathbb{V}(\beta) = \frac{b^{(0)}}{a^{(0)}-1}\boldsymbol{M}^{(0)} = \mathbb{E}[\sigma^2]\boldsymbol{M}^{(0)}\,. \tag{4}$$

The joint posterior distribution $p(\theta \mid z)$, due to conjugacy, is then again a normal – inverse gamma distribution with the updated parameters

$$\begin{aligned} m^{(1)} &= \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\left(\boldsymbol{M}^{(0)^{-1}} m^{(0)} + \mathbf{X}^\mathsf{T} z\right), \\ \boldsymbol{M}^{(1)} &= \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}, \\ a^{(1)} &= a^{(0)} + \frac{n}{2}\,, \qquad b^{(1)} = b^{(0)} + \frac{1}{2}\left(z^\mathsf{T} z + m^{(0)^\mathsf{T}} \boldsymbol{M}^{(0)^{-1}} m^{(0)} - m^{(1)^\mathsf{T}} \boldsymbol{M}^{(1)^{-1}} m^{(1)}\right). \end{aligned}$$

The properties of the posterior distributions can thus be analyzed by inserting the updated parameters into (3) and (4).

3.1 Update of $\beta \mid \sigma^2$

The normal distribution part of the joint prior is updated as follows:

$$\mathbb{E}[\beta \mid \sigma^2, z] = m^{(1)} = \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\left(\boldsymbol{M}^{(0)^{-1}} m^{(0)} + \mathbf{X}^\mathsf{T} z\right) = (\mathbf{I} - \boldsymbol{A})\, m^{(0)} + \boldsymbol{A}\,\hat{\beta}_{LS}\,,$$

where $\boldsymbol{A} = \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}$. The posterior estimate of $\beta \mid \sigma^2$ thus calculates as a matrix-weighted mean of the prior guess and the least-squares estimate. The larger the diagonal elements of $\boldsymbol{M}^{(0)}$ (i.e., the weaker the prior information), the smaller the elements of $\boldsymbol{M}^{(0)^{-1}}$ and thus the 'nearer' is A to the identity matrix, so that the posterior estimate is nearer to the least-squares estimate.

The posterior variance of $\beta \mid \sigma^2$ calculates as

$$\mathbb{V}(\beta \mid \sigma^2, z) = \sigma^2 \boldsymbol{M}^{(1)} = \sigma^2 \left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\intercal \mathbf{X} \right)^{-1}.$$

As the elements of $\boldsymbol{M}^{(1)^{-1}}$ get larger with respect to $\boldsymbol{M}^{(0)^{-1}}$, the elements of $\boldsymbol{M}^{(1)}$ will, roughly speaking, become smaller than those of $\boldsymbol{M}^{(0)}$, so that the variance of $\beta \mid \sigma^2$ decreases.

Therefore, the updating of $\beta \mid \sigma^2$ is obviously insensitive to prior-data conflict, because the posterior distribution will not become flatter in case of a large distance between $\mathbb{E}[\beta]$ and $\hat{\beta}_{LS}$. Actually, as O'Hagan (1994) derives, for any $\phi = a^\intercal \beta$, i.e., any linear combination of elements of β, it holds that $\mathbb{V}(\phi \mid \sigma^2, z) \leq \mathbb{V}(\phi \mid \sigma^2)$, becoming a strict inequality if $\mathbf{X}$ has full rank. In particular, the variance of each β_i decreases automatically with the update step.

3.2 Update of σ^2

It can be shown (O'Hagan 1994) that

$$\mathbb{E}[\sigma^2 \mid z] = \frac{2a^{(0)} - 2}{2a^{(0)} + n - 2} \mathbb{E}[\sigma^2] + \frac{n-p}{2a^{(0)} + n - 2} \hat{\sigma}^2_{LS} + \frac{p}{2a^{(0)} + n - 2} \hat{\sigma}^2_{PDC}, \qquad (5)$$

where $\hat{\sigma}^2_{LS} = \frac{1}{n-p}(z - \mathbf{X}\hat{\beta}_{LS})^\intercal (z - \mathbf{X}\hat{\beta}_{LS})$ is the least-squares based estimate for σ^2, and $\hat{\sigma}^2_{PDC} = \frac{1}{p}(m^{(0)} - \hat{\beta}_{LS})^\intercal \left(\boldsymbol{M}^{(0)} + (\mathbf{X}^\intercal \mathbf{X})^{-1} \right)^{-1} (m^{(0)} - \hat{\beta}_{LS})$. For the latter it holds that $\mathbb{E}[\hat{\sigma}^2_{PDC} \mid \sigma^2] = \sigma^2$; the posterior expectation of σ^2 calculates thus as a weighted mean of three estimates:

(i) the prior expectation for σ^2,

(ii) the least-squares estimate, and

(iii) an estimate based on a weighted squared difference of the prior mean and the least-squares estimate for β.

The weights depend on $a^{(0)}$ (one prior parameter for the inverse gamma part), the sample size n, and the dimension of β, respectively. The role of the first weight gets more plausible when remembering the formula for the prior variance of σ^2 in (3), where $a^{(0)}$ appears in the denominator. A larger value of $a^{(0)}$ means thus smaller prior variance, in turn giving a higher weight for $\mathbb{E}[\sigma^2]$ in the calculation of $\mathbb{E}[\sigma^2 \mid z]$. The weight to $\hat{\sigma}^2_{LS}$ corresponds to the classical degrees of freedom, $n - p$. With the the sample size approaching infinity, this weight will dominate the others, such that $\mathbb{E}[\sigma^2 \mid z]$ approaches $\hat{\sigma}^2_{LS}$.

Similar results hold for the posterior mode instead of the posterior expectation.

Here, the estimate $\hat{\sigma}^2_{PDC}$ allows some reaction to prior-data conflict: it measures the distance between $m^{(0)}$ (prior) and $\hat{\beta}_{LS}$ (data) estimates for β, with a large distance resulting basically in a large value of $\hat{\sigma}^2_{PDC}$ and thus an enlarged posterior estimate for σ^2. The weighting matrix for the distances is playing an important role as well. The influence of $\boldsymbol{M}^{(0)}$ is as follows: for components of β one is quite certain about the assignment of $m^{(0)}$, the respective diagonal elements of $\boldsymbol{M}^{(0)}$ will be low, so that these

diagonal elements of the weighting matrix will be high. Therefore, large distances in these dimensions will increase t strongly. An erroneously high confidence in the prior assumptions on β is thus penalized by an increasing posterior estimate for σ^2. The influence of $\mathbf{X}^\mathsf{T}\mathbf{X}$ interprets as follows: covariates with a low spread in x-values, giving an unstable base for the estimate $\hat{\beta}_{LS}$, will result in low diagonal elements of $\mathbf{X}^\mathsf{T}\mathbf{X}$. Via the double inverting, those diagonal elements of the weighting matrix will remain low and thus give the difference a low weight. Therefore, $\hat{\sigma}^2_{PDC}$ will not excessively increase due to a large difference in dimensions where the location of $\hat{\beta}_{LS}$ is to be taken cum grano salis. As to be seen in the following subsection, the behavior of $\mathbb{E}[\sigma \mid z]$ is of high importance for posterior inferences on β.

3.3 Update of β

The posterior distribution of β is again a multivariate t, with expectation $\mathbb{E}[\beta \mid z] = \mathbb{E}\big[\mathbb{E}[\beta \mid \sigma^2, z] \mid z\big] = m^{(1)}$ (as described in Section 3.1) and variance

$$\mathbb{V}[\beta \mid z] = \frac{b^{(1)}}{a^{(1)}-1}\boldsymbol{M}^{(1)} = \mathbb{E}[\sigma^2 \mid z]\,\boldsymbol{M}^{(1)} \tag{6}$$

$$= \left(\frac{2a^{(0)}-2}{2a^{(0)}+n-2}\mathbb{E}[\sigma^2] + \frac{n-p}{2a^{(0)}+n-2}\hat{\sigma}^2_{LS} + \frac{p}{2a^{(0)}+n-2}\hat{\sigma}^2_{PDC}\right)\left(\boldsymbol{M}^{(0)^{-1}} + \mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$$

$$= \left(\frac{2a^{(0)}-2}{2a^{(0)}+n-2}\mathbb{E}[\sigma^2] + \frac{n-p}{2a^{(0)}+n-2}\hat{\sigma}^2_{LS} + \frac{p}{2a^{(0)}+n-2}\hat{\sigma}^2_{PDC}\right)$$
$$\cdot \left(\boldsymbol{M}^{(0)} - \boldsymbol{M}^{(0)}\mathbf{X}^\mathsf{T}(\mathbf{I} + \mathbf{X}\boldsymbol{M}^{(0)}\mathbf{X}^\mathsf{T})^{-1}\mathbf{X}\boldsymbol{M}^{(0)}\right),$$

not being directly expressible as a function of $\mathbb{E}[\sigma^2]\boldsymbol{M}^{(0)}$, the prior variance of β.

Due to the effect of $\mathbb{E}[\sigma^2 \mid z]$, the posterior variance-covariance matrix of β can increase in case of prior data conflict, if the rise of $\mathbb{E}[\beta \mid z]$ (due to an even stronger rise of t) can overcompensate the decrease in the elements of $\boldsymbol{M}^{(1)}$. However, we see that the effect of prior-data conflict on the posterior variance of β is *globally* and not component-specific; it influences the variances for *all* components of β to the same amount even if the conflict was confined only to some or even just one single component. Taking it to the extremes, if the prior assignment $m^{(0)}$ was (more or less) correct in all but one component, with that one being far out, the posterior variances will increase for all components, also for the ones with prior assignments that have turned out to be basically correct.

4 An Alternative Approach for Conjugate Priors in Bayesian Linear Regression (CCCP)

In this section, a prior model for $\theta = (\beta, \sigma^2)$ will be constructed along the general construction method for sample distributions that form a linear, canonical exponential family (see, e.g., Bernardo & Smith 1994). The method is typically used for the i.i.d. case, but the likelihood arising from $z \sim \mathrm{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$ will be shown to follow the specific exponential family form as well. The canonically constructed conjugate prior (CCCP) model will also result in a normal - inverse gamma distribution, but with a fixed variance - covariance structure. The CCCP model is thus a special case of the SCP model, which – as will be detailed in this section – offers some interesting further insights into the structure of the update step.

The likelihood arising from the distribution of z,

$$
\begin{aligned}
f(z \mid \beta, \sigma^2) &= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2)^{\frac{n}{2}}} \exp\Big\{ -\frac{1}{2\sigma^2}(z-\mathbf{X}\beta)^{\mathsf{T}}(z-\mathbf{X}\beta)\Big\} \\
&= \underbrace{\frac{1}{(2\pi)^{\frac{n}{2}}}}_{\mathbf{a}(z)} \exp\Big\{ \underbrace{\Big(\frac{\beta}{\sigma^2}\Big)}_{\psi_1}^{\mathsf{T}} \underbrace{\mathbf{X}^{\mathsf{T}}z}_{\tau_1(z)} \underbrace{-\frac{1}{\sigma^2}}_{\psi_2} \underbrace{\frac{1}{2}z^{\mathsf{T}}z}_{\tau_2(z)} \underbrace{-\Big(\frac{1}{2\sigma^2}\beta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\beta + \frac{n}{2}\log(\sigma^2)\Big)}_{n\mathbf{b}(\psi)} \Big\},
\end{aligned}
$$

indeed corresponds to the linear, canonical exponential family form

$$
f(z \mid \psi) = \mathbf{a}(z) \cdot \exp\{\langle \psi, \tau(z)\rangle - n \cdot \mathbf{b}(\psi)\},
$$

where $\psi = \psi(\beta, \sigma^2)$ is a certain function of β and σ^2, the parameters on which one wishes to learn. $\tau(z)$ is a sufficient statistic of z used in the update step. Here, we have

$$
\psi = \begin{pmatrix} \frac{\beta}{\sigma^2} \\ -\frac{1}{\sigma^2} \end{pmatrix}, \quad \tau(z) = \begin{pmatrix} \mathbf{X}^{\mathsf{T}}z \\ \frac{1}{2}z^{\mathsf{T}}z \end{pmatrix}, \quad \mathbf{b}(\psi) = \frac{1}{2n\sigma^2}\beta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\beta + \frac{1}{2}\log(\sigma^2). \tag{7}
$$

According to the general construction method, a conjugate prior for ψ can be obtained from these ingredients by the following equation:

$$
p(\psi) = \mathbf{c}(n^{(0)}, y^{(0)}) \cdot \exp\Big\{ n^{(0)} \cdot [\langle \psi, y^{(0)}\rangle - \mathbf{b}(\psi)] \Big\},
$$

where $n^{(0)}$ and $y^{(0)}$ are the parameters that define the concrete prior distribution of its distribution family; whereas ψ and $\mathbf{b}(\psi)$ were identified in (7). $\mathbf{c}$ corresponds to a normalization factor for the prior. When applying the general construction method to the two examples from Section 2, the very same priors as presented there will result, where $y^{(0)} = \mu^{(0)}$ and $n^{(0)} = 1/{\sigma^{(0)}}^2$ for the prior to the scaled normal model, and $y^{(0)} = t^{(0)}$ and $n^{(0)} = s^{(0)}$ for the prior to the multinomial model.

Here, the conjugate prior writes as

$$p(\psi)d\psi = \mathbf{c}(n^{(0)}, y^{(0)}) \exp\left\{ n^{(0)} \left[y^{(0)\mathsf{T}} \begin{pmatrix} \frac{\beta}{\sigma^2} \\ -\frac{1}{\sigma^2} \end{pmatrix} - \frac{1}{2n\sigma^2} \beta^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} \beta - \frac{1}{2} \log(\sigma^2) \right] \right\} d\psi .$$

As this is a prior on ψ, but we want to arrive at a prior on $\theta = (\beta, \sigma^2)^{\mathsf{T}}$, we must transform the density $p(\psi)$. For the transformation, we need the determinant of the Jacobian matrix $\frac{d\psi}{d\theta}$:

$$\left| \det\left(\frac{d\psi}{d\theta} \right) \right| = \left| \det \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{I}_p & -\frac{\beta}{(\sigma^2)^2} \\ \mathbf{0} & \frac{1}{(\sigma^2)^2} \end{pmatrix} \right| = \frac{1}{(\sigma^2)^{p+2}} .$$

Therefore, the prior on $\theta = (\beta, \sigma^2)^{\mathsf{T}}$ is

$$p(\theta)d\theta = p(\psi)d\psi \cdot \left| \det\left(\frac{d\psi}{d\theta} \right) \right| = \mathbf{c}(n^{(0)}, y^{(0)}) \cdot \tag{8}$$

$$\exp\left\{ n^{(0)} y_1^{(0)\mathsf{T}} \frac{\beta}{\sigma^2} - n^{(0)} y_2^{(0)} \frac{1}{\sigma^2} - \frac{n^{(0)}}{2n\sigma^2} \beta^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} \beta - \frac{n^{(0)}}{2} \log(\sigma^2) - (p+2) \log(\sigma^2) \right\} .$$

θ can now be shown to follow a normal – inverse gamma distribution by comparing coefficients. In doing that, some attention must be paid to the terms proportional to $-1/\sigma^2$ (appearing as $-\log(\sigma^2)$ in the exponent) because the normal $p(\beta \mid \sigma^2)$ and the inverse gamma $p(\sigma^2)$ will have to 'share' it. Furthermore, it is necessary to complete the square for the normal part, resulting in an additional term for the inverse gamma part.

The density of a normal distribution on $\beta \mid \sigma^2$ with a mean vector $\overline{m}^{(0)} = \overline{m}(n^{(0)}, y^{(0)})$ and a variance-covariance matrix $\sigma^2 \overline{\boldsymbol{M}}^{(0)} = \sigma^2 \overline{\boldsymbol{M}}(n^{(0)}, y^{(0)})$, both to be seen as functions of the canonical parameters $n^{(0)}$ and $y^{(0)}$, has the following form:

$$\begin{aligned} p(\beta \mid \sigma^2) &= \frac{1}{(2\pi)^{\frac{p}{2}} (\sigma^2)^{\frac{p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\beta - \overline{m}^{(0)} \right)^{\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \left(\beta - \overline{m}^{(0)} \right) \right\} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left\{ \overline{m}^{(0)\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \frac{\beta}{\sigma^2} - \frac{1}{2\sigma^2} \beta^{\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \beta \right. \\ &\qquad \left. - \frac{1}{2\sigma^2} \overline{m}^{(0)\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \overline{m}^{(0)} - \frac{p}{2} \log(\sigma^2) \right\} . \end{aligned}$$

Comparing coefficients with the terms from (8) depending on β, we get

$$\overline{\boldsymbol{M}}^{(0)^{-1}} = \overline{\boldsymbol{M}}(n^{(0)})^{-1} = \frac{n^{(0)}}{n} \mathbf{X}^{\mathsf{T}} \mathbf{X} , \qquad \overline{m}^{(0)} = \overline{m}(y^{(0)}) = n \left(\mathbf{X}^{\mathsf{T}} \mathbf{X} \right)^{-1} y^{(0)} .$$

With the square completed, the joint density of β and σ^2 reads as

$$p(\beta,\sigma^2) = \mathbf{c}(n^{(0)},y^{(0)})\cdot$$
$$\exp\Big\{\underbrace{n^{(0)}y_1^{(0)\mathsf{T}}\frac{\beta}{\sigma^2} - \frac{n^{(0)}}{2n\sigma^2}\beta^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\beta - \frac{1}{2\sigma^2}\Big(n\cdot n^{(0)}y_1^{(0)\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}y_1^{(0)}\Big) - \frac{p}{2}\log(\sigma^2)}_{\text{to } p(\beta|\sigma^2) \text{ (normal distribution)}}$$
$$\underbrace{-\frac{1}{\sigma^2}\Big(-\frac{n^{(0)}n}{2}y_1^{(0)\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}y_1^{(0)}\Big) - n^{(0)}y_2^{(0)}\frac{1}{\sigma^2} - \Big(\frac{n^{(0)}+p}{2}+2\Big)\log(\sigma^2)}_{\text{to } p(\sigma^2) \text{ (inverse gamma distribution)}}\Big\}. \quad (9)$$

Therefore, one part of the conjugate prior (9) reveals as a multivariate normal distribution with mean vector $\overline{m}^{(0)} = \overline{m}(y_1^{(0)}) = n\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\,y_1^{(0)}$ and covariance matrix $\sigma^2\overline{\boldsymbol{M}}^{(0)} = \sigma^2\overline{\boldsymbol{M}}(n^{(0)}) = \frac{n\sigma^2}{n^{(0)}}\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, i.e.

$$\beta \mid \sigma^2 \sim \mathrm{N}_p\left(n\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\,y_1^{(0)},\ \frac{n\sigma^2}{n^{(0)}}\,(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\right). \quad (10)$$

The other terms of (9) can be directly identified with the core of an inverse gamma distribution with parameters

$$\overline{a}^{(0)} = \frac{n^{(0)}+p}{2}+1 \qquad \text{and}$$
$$\overline{b}^{(0)} = n^{(0)}y_2^{(0)} - \frac{n^{(0)}}{2}y_1^{(0)\mathsf{T}}n(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}y_1^{(0)} = n^{(0)}y_2^{(0)} - \frac{1}{2}\overline{m}^{(0)\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)},$$

$$\text{i.e.,}\quad \sigma^2 \sim \mathrm{IG}\left(\frac{n^{(0)}+p+2}{2},\ n^{(0)}y_2^{(0)} - \frac{n^{(0)}}{2}y_1^{(0)\mathsf{T}}n(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}y_1^{(0)}\right). \quad (11)$$

We have thus derived the CCCP distribution on (β,σ^2), which can be expressed either in terms of the canonical prior parameters $n^{(0)}$ and $y^{(0)}$ or in terms of the prior parameters from Section 3, $\overline{m}^{(0)}$, $\overline{\boldsymbol{M}}^{(0)}$, $\overline{a}^{(0)}$ and $\overline{b}^{(0)}$. As already noted, $\overline{\boldsymbol{M}}^{(0)} = \frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ can be seen as a restricted version of $\boldsymbol{M}^{(0)}$. $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$ is known as a variance-covariance structure from the least squares estimate $\mathbb{V}(\beta) = \hat{\sigma}^2_{LS}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, and is here the fixed prior variance-covariance structure for $\beta \mid \sigma^2$. Confidence in the prior assignment is expressed by the choice of $n^{(0)}$: With $n^{(0)}$ chosen large relative to n, strong confidence in the prior assignment of $\overline{m}^{(0)}$ can be expressed, whereas a low value of $n^{(0)}$ will result in a less pointed prior distribution on $\beta \mid \sigma^2$.

The update step for a canonically constructed prior, expressed in terms of $n^{(0)}$ and $y^{(0)}$, possesses a convenient form: In the prior, the parameters $n^{(0)}$ and $y^{(0)}$ must simply be replaced by their updated versions $n^{(1)}$ and $y^{(1)}$, which calculate as

$$y^{(1)} = \frac{n^{(0)}y^{(0)} + \tau(z)}{n^{(0)}+n}, \qquad\qquad n^{(1)} = n^{(0)}+n.$$

4.1 Update of $\beta \mid \sigma^2$

As $y^{(0)}$ and $y^{(1)}$ are not directly interpretable, it is certainly easier to express prior beliefs on β via the mean vector $\overline{m}^{(0)}$ of the prior distribution of $\beta \mid \sigma^2$ just as in the SCP model. As the transformation $\overline{m}^{(0)} \mapsto y^{(0)}$ is linear, this poses no problem:

$$\begin{aligned}\mathbb{E}[\beta \mid \sigma^2, z] = \overline{m}^{(1)} &= n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} y_1^{(1)} = n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\left(\frac{n^{(0)}}{n^{(0)}+n} y_1^{(0)} + \frac{n}{n^{(0)}+n} \cdot \frac{1}{n}(\mathbf{X}^\mathsf{T} z)\right) \\ &= n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \frac{n^{(0)}}{n^{(0)}+n} \cdot \frac{1}{n}(\mathbf{X}^\mathsf{T}\mathbf{X})\overline{m}^{(0)} + n(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \frac{n}{n^{(0)}+n} \cdot \frac{1}{n}(\mathbf{X}^\mathsf{T} z) \\ &= \frac{n^{(0)}}{n^{(0)}+n}\mathbb{E}[\beta \mid \sigma^2] + \frac{n}{n^{(0)}+n}\hat{\beta}_{LS}\,. \end{aligned} \tag{12}$$

The posterior expectation for $\beta \mid \sigma^2$ is here a scalar-weighted mean of the prior expectation and the least squares estimate, with weights $n^{(0)}$ and n, respectively. The role of $n^{(0)}$ in the prior variance of $\beta \mid \sigma^2$ is directly mirrored here. As described for the generalized setting in Walter & Augustin (2009, p. 258) in more detail, $n^{(0)}$ can be seen as a parameter describing the "prior strength" or expressing "pseudocounts". In line with this interpretation, high values of $n^{(0)}$ as compared to n result here in a strong influence of $\overline{m}^{(0)}$ for the calculation of $\overline{m}^{(1)}$, whereas for small values of $n^{(0)}$, $\mathbb{E}[\beta \mid \sigma^2, z]$ will be dominated by the value of $\hat{\beta}_{LS}$.

The variance of $\beta \mid \sigma^2$ is updated as follows:

$$\mathbb{V}(\beta \mid \sigma^2, z) = \frac{n\sigma^2}{n^{(1)}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \frac{n\sigma^2}{n^{(0)}+n}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\,.$$

Here, $n^{(0)}$ is updated to $n^{(1)}$, and thus the posterior variances are automatically smaller than the prior variances, just as in the SCP model.

4.2 Update of σ^2

For the assignment of the parameters $\overline{a}^{(0)}$ and $\overline{b}^{(0)}$ to define the inverse gamma part of the joint prior, only $y_2^{(0)}$ is left to choose, as $n^{(0)}$ and $y_1^{(0)}$ are already assigned via the choice of $\overline{m}^{(0)}$ and $\overline{\boldsymbol{M}}^{(0)}$. To choose $y_2^{(0)}$, it is convenient to consider the prior expectation of σ^2 (alternatively, the prior mode of σ^2 could be considered as well):

$$\mathbb{E}[\sigma^2] = \frac{\overline{b}^{(0)}}{\overline{a}^{(0)}-1} = \frac{2n^{(0)}}{n^{(0)}+p} y_2^{(0)} - \frac{1}{n^{(0)}+p}\overline{m}^{(0)\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)}\,.$$

A value of $y_2^{(0)}$ dependent on the value of $\mathbb{E}[\sigma^2]$ can thus be chosen by the linear mapping

$$y_2^{(0)} = \frac{n^{(0)}+p}{2n^{(0)}}\mathbb{E}[\sigma^2] + \frac{1}{2n^{(0)}}\overline{m}^{(0)\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)}\,.$$

For the posterior expected value of σ^2, there is a similar decomposition as for the SCP model, and furthermore two other possible decompositions offering interesting interpretations of the update step of σ^2. The three decompositions are presented in the following.

4.2.1 Decomposition Including an Estimate of σ^2 Through the Null Model

The posterior variance of σ^2 calculates firstly as:

$$\begin{aligned}
\mathbb{E}[\sigma^2 \mid z] &= \frac{\overline{b}^{(1)}}{\overline{a}^{(1)} - 1} = \frac{2n^{(1)}}{n^{(1)} + p} y_2^{(1)} - \frac{1}{n^{(1)} + p} \overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)} \\
&= \frac{2n^{(0)}}{n^{(0)} + n + p} y_2^{(0)} + \frac{1}{n^{(0)} + n + p} z^{\mathsf{T}} z - \frac{1}{n^{(0)} + n + p} \overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)} \\
&= \frac{n^{(0)} + p}{n^{(0)} + n + p} \mathbb{E}[\sigma^2] + \frac{n - 1}{n^{(0)} + n + p} \frac{1}{n - 1} z^{\mathsf{T}} z \\
&\quad + \frac{1}{n^{(0)} + n + p} \left(\overline{m}^{(0)\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \overline{m}^{(0)} - \overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)} \right), \qquad (13)
\end{aligned}$$

and so displays as a weighted average of the prior expected value, $\frac{1}{n-1} z^{\mathsf{T}} z$, and a term depending on prior and posterior estimates for β, with weights $n^{(0)} + p$, $n - 1$ and 1, respectively. When adopting the centered z, standardized $\mathbf{X}$ approach, $\frac{1}{n-1} z^{\mathsf{T}} z$ is the estimate for σ^2 under the null model, that is, if $\beta = 0$. Contrary to what a cursory inspection might suggest, the third term's influence, having the constant weight of 1, will not vanish for $n \to \infty$, as the third term does not approach a constant.[5]

The third term reflects the change in information about β:

If we are very uncertain about the prior beliefs on β expressed in $\overline{m}^{(0)}$ and thus assign a small value for $n^{(0)}$ with respect to n, we will get relatively large variances and covariances in $\overline{\boldsymbol{M}}^{(0)}$ by a factor $\frac{n}{n^{(0)}} > 1$ to $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, resulting in a small term $\overline{m}^{(0)\mathsf{T}} \overline{\boldsymbol{M}}^{(0)^{-1}} \overline{m}^{(0)}$. After updating, the elements in $\overline{\boldsymbol{M}}^{(1)}$ become smaller automatically due to the updated factor $\frac{n}{n^{(0)}+n}$ to $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$. If the values of $\overline{m}^{(1)}$ do not differ much from the values in $\overline{m}^{(0)}$, the term $\overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)}$ would be larger than its prior counterpart, ultimately reducing the posterior expectation for σ^2 through the third term being negative. If $\overline{m}^{(1)}$ does significantly differ from $\overline{m}^{(0)}$, then the term $\overline{m}^{(1)\mathsf{T}} \overline{\boldsymbol{M}}^{(1)^{-1}} \overline{m}^{(1)}$ can actually result smaller than its prior counterpart and thus give a larger value of $\mathbb{E}[\sigma^2 \mid z]$ as compared with the situation $\overline{m}^{(1)} \approx \overline{m}^{(0)}$.

On the contrary, large values for $n^{(0)}$ with respect to n indicating high trust in prior beliefs on β lead to small variances and covariances in $\overline{\boldsymbol{M}}^{(0)}$ by the factor $\frac{n}{n^{(0)}} < 1$

[5] Although $\overline{m}^{(1)}$ approaches $\hat{\beta}_{LS}$, and $\overline{m}^{(0)}$ is a constant, $\overline{\boldsymbol{M}}^{(0)^{-1}}$ and $\overline{\boldsymbol{M}}^{(1)^{-1}}$ are increasing for growing n, with $\overline{\boldsymbol{M}}^{(1)^{-1}}$ increasing faster than $\overline{\boldsymbol{M}}^{(0)^{-1}}$. The third term will thus eventually turn negative, reducing the null model variance that has weight $n - 1$.

to $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, resulting in a larger term $\overline{m}^{(0)^\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)}$ as compared to the case with low $n^{(0)}$. After updating, variances and covariances in $\overline{\boldsymbol{M}}^{(1)}$ will become even smaller, amplifying the term $\overline{m}^{(1)^\mathsf{T}}\overline{\boldsymbol{M}}^{(1)^{-1}}\overline{m}^{(1)}$ even more if $\overline{m}^{(1)} \approx \overline{m}^{(0)}$, ultimately reducing the posterior expectation for σ^2 more than in the situation with low $n^{(0)}$. If, however, the values of $\overline{m}^{(1)}$ do differ significantly from the values in $\overline{m}^{(0)}$, the term $\overline{m}^{(1)^\mathsf{T}}\overline{\boldsymbol{M}}^{(1)^{-1}}\overline{m}^{(1)}$ can result smaller than its prior counterpart also here and even more so as compared to the situation with low $n^{(0)}$, giving eventually an even larger posterior expectation for σ^2.

4.2.2 Decomposition Similar to the SCP Model

A decomposition similar to the one in Section 3.2 can be derived by considering the third term from (13) in more detail:

$$
\begin{aligned}
&\overline{m}^{(0)^\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)} - \overline{m}^{(1)^\mathsf{T}}\overline{\boldsymbol{M}}^{(1)^{-1}}\overline{m}^{(1)} \\
&\quad = n^{(0)} \cdot n \cdot y_1^{(0)^\mathsf{T}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}y_1^{(0)} - (n^{(0)}+n)\cdot n\frac{n^{(0)}y^{(0)^\mathsf{T}} + z^\mathsf{T}\mathbf{X}}{n^{(0)}+n}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\frac{n^{(0)}y^{(0)} + \mathbf{X}^\mathsf{T}z}{n^{(0)}+n} \\
&\quad = \frac{n}{n^{(0)}+n}\left[\overline{m}^{(0)^\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)} - 2\overline{m}^{(0)^\mathsf{T}}\overline{\boldsymbol{M}}^{(0)^{-1}}\hat{\beta}_{LS} - \frac{n}{n^{(0)}}\hat{\beta}_{LS}^\mathsf{T}\overline{\boldsymbol{M}}^{(0)^{-1}}\hat{\beta}_{LS}\right] \\
&\quad = \frac{n}{n^{(0)}+n}\left(\overline{m}^{(0)} - \hat{\beta}_{LS}\right)^\mathsf{T}\overline{\boldsymbol{M}}^{(0)^{-1}}\left(\overline{m}^{(0)} - \hat{\beta}_{LS}\right) - z^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}z.
\end{aligned}
$$

Thus, we get

$$
\begin{aligned}
\mathbb{E}[\sigma^2 \mid \boldsymbol{z}] &= \frac{n^{(0)}+p}{n^{(0)}+n+p}\mathbb{E}[\sigma^2] + \frac{1}{n^{(0)}+n+p}\left(z^\mathsf{T}z - z^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}z\right) \\
&\quad + \frac{1}{n^{(0)}+n+p}\cdot\frac{n}{n^{(0)}+n}(\overline{m}^{(0)} - \hat{\beta}_{LS})^\mathsf{T}\overline{\boldsymbol{M}}^{(0)^{-1}}(\overline{m}^{(0)} - \hat{\beta}_{LS}) \\
&= \frac{n^{(0)}+p}{n^{(0)}+n+p}\mathbb{E}[\sigma^2] + \frac{n-p}{n^{(0)}+n+p}\cdot\underbrace{\frac{1}{n-p}(z - \mathbf{X}\hat{\beta}_{LS})^\mathsf{T}(z - \mathbf{X}\hat{\beta}_{LS})}_{\hat{\sigma}_{LS}^2} \\
&\quad + \frac{p}{n^{(0)}+n+p}\cdot\underbrace{\frac{n}{n^{(0)}+n}\frac{1}{p}(\overline{m}^{(0)} - \hat{\beta}_{LS})^\mathsf{T}\overline{\boldsymbol{M}}^{(0)^{-1}}(\overline{m}^{(0)} - \hat{\beta}_{LS})}_{=:\overline{\sigma}_{PDC}^2}. \qquad (14)
\end{aligned}
$$

The posterior expectation for σ^2 can therefore be seen also here as a weighted average of the prior expected value, the estimation $\hat{\sigma}_{LS}^2$ resulting from least squares methods, and $\overline{\sigma}_{PDC}^2$,[6] with weights $n^{(0)}+p$, $n-p$ and p, respectively. As in the update step

[6] $\mathbb{E}[\overline{\sigma}_{PDC}^2 \mid \sigma^2] = \sigma^2$ computes very similar to the calculations given in O'Hagan (1994, p. 249).

for $\beta \mid \sigma^2$, $n^{(0)}$ is guarding the influence of the prior expectation on the posterior expectation. Just as in the decomposition for the SCP model, the weight for $\hat{\sigma}^2_{LS}$ will dominate the others when the sample size approaches infinity. Also for the CCCP model, $\overline{\sigma}^2_{PDC}$ is getting large if prior beliefs on β are skewed with respect to "what the data says", eventually inflating the posterior expectation of σ^2. The weighting of the differences is similar as well: High prior confidence in the chosen value of $\overline{m}^{(0)}$ as expressed by a high value of $n^{(0)}$ will give a large $\overline{\boldsymbol{M}}^{(0)^{-1}}$ and thus penalizing erroneous assignments stronger as compared to a lower value of $n^{(0)}$. Again, $\mathbf{X}^\intercal\mathbf{X}$ weighs the differences for components with covariates having a low spread weaker due to the instability of the respective component of $\hat{\beta}_{LS}$ under such conditions.

4.2.3 Decomposition with Estimates of σ^2 Through Prior and Posterior Residuals

A third interpretation of $\mathbb{E}[\sigma^2 \mid z]$ can be derived by another reformulation of the third term in (13):

$$\begin{aligned}
&\overline{m}^{(0)\intercal}\overline{\boldsymbol{M}}^{(0)^{-1}}\overline{m}^{(0)} - \overline{m}^{(1)\intercal}\overline{\boldsymbol{M}}^{(1)^{-1}}\overline{m}^{(1)} = \frac{n^{(0)}}{n}\overline{m}^{(0)\intercal}\mathbf{X}^\intercal\mathbf{X}\overline{m}^{(0)} - \frac{n^{(1)}}{n}\overline{m}^{(1)\intercal}\mathbf{X}^\intercal\mathbf{X}\overline{m}^{(1)} \\
&= \frac{n^{(0)}}{n}(z-\mathbf{X}\overline{m}^{(0)})^\intercal(z-\mathbf{X}\overline{m}^{(0)}) - \frac{n^{(1)}}{n}(z-\mathbf{X}\overline{m}^{(1)})^\intercal(z-\mathbf{X}\overline{m}^{(1)}) \\
&\quad + \frac{n^{(1)}}{n}z^\intercal z - \frac{n^{(0)}}{n}z^\intercal z + \frac{n^{(0)}}{n}2z^\intercal\mathbf{X}\overline{m}^{(0)} - \frac{n^{(1)}}{n}2z^\intercal\mathbf{X}\overline{m}^{(1)} \\
&= \frac{n^{(0)}}{n}(z-\mathbf{X}\overline{m}^{(0)})^\intercal(z-\mathbf{X}\overline{m}^{(0)}) - \frac{n^{(1)}}{n}(z-\mathbf{X}\overline{m}^{(1)})^\intercal(z-\mathbf{X}\overline{m}^{(1)}) + z^\intercal z - 2z^\intercal\mathbf{X}\hat{\beta}_{LS}\,.
\end{aligned}$$

With this, we get

$$\begin{aligned}
\mathbb{E}[\sigma^2 \mid z] = \frac{n^{(0)}+p}{n^{(0)}+n+p}\mathbb{E}[\sigma^2] + \frac{n^{(0)}+p}{n^{(0)}+n+p}\underbrace{\frac{n^{(0)}}{n}\cdot\frac{1}{n^{(0)}+p}(z-\mathbf{X}\overline{m}^{(0)})^\intercal(z-\mathbf{X}\overline{m}^{(0)})}_{=:\sigma^{(0)^2},\ \text{as}\ \mathbb{E}[\sigma^{(0)^2}\mid\sigma^2]=\sigma^2} \\
+ \frac{2(n-p)}{n^{(0)}+n+p}\hat{\sigma}^2_{LS} - \frac{n^{(1)}+p}{n^{(0)}+n+p}\underbrace{\frac{n^{(1)}}{n}\cdot\frac{1}{n^{(1)}+p}(z-\mathbf{X}\overline{m}^{(1)})^\intercal(z-\mathbf{X}\overline{m}^{(1)})}_{=:\sigma^{(1)^2},\ \text{as}\ \mathbb{E}[\sigma^{(1)^2}\mid\sigma^2,z]=\mathbb{E}[\sigma^{(1)^2}\mid\sigma^2]=\sigma^2}\,.
\end{aligned} \tag{15}$$

Here, the calculation of $\mathbb{E}[\sigma^2 \mid z]$ is based again on $\mathbb{E}[\sigma^2]$ and $\hat{\sigma}^2_{LS}$, but nowe complemented with two special estimates: $\sigma^{(0)^2}$, an estimate based on the prior residuals $(z-\mathbf{X}\overline{m}^{(0)})^\intercal(z-\mathbf{X}\overline{m}^{(0)})$, and a respective posterior version $\sigma^{(1)^2}$, based on $(z-\mathbf{X}\overline{m}^{(1)})^\intercal(z-\mathbf{X}\overline{m}^{(1)})$. However, $\mathbb{E}[\sigma^2 \mid z]$ is only "almost" a weighted average of these ingredients, as the weights sum up to $n^{(0)}-p+n$ instead of $n^{(0)}+p+n$.

Especially strange is the negative weight for ${\sigma^{(1)}}^2$, actually making the factor to ${\sigma^{(1)}}^2$ result to -1. A possible interpretation would be to group $\mathbb{E}[\sigma^2]$ and ${\sigma^{(0)}}^2$ as prior-based estimations with joint weight $2(n^{(0)}+p)$, and $\hat{\sigma}^2_{LS}$ as data-based estimation with weight $2(n-p)$. Together, these estimations have a weight of $2(n^{(0)}+n)$, being almost (neglecting the missing $2p$) a "double estimate" that is corrected back to a "single" estimate with the posterior-based estimate ${\sigma^{(1)}}^2$.

4.3 Update of β

As for the SCP model, the posterior on β, being the most important distribution for inference, is a multivariate t with expectation $\overline{m}^{(1)}$ as described in Section 4.1. For $\mathbb{V}(\beta \mid z)$, one gets different formulations depending on the formula for $\mathbb{E}[\sigma^2 \mid z]$:

$$
\begin{aligned}
\mathbb{V}(\beta \mid z) &= \frac{\overline{b}^{(1)}}{\overline{a}^{(1)}-1}\overline{\boldsymbol{M}}^{(1)} = \mathbb{E}[\sigma^2 \mid z]\,\frac{n}{n^{(1)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \qquad (16)\\
&\overset{(13)}{=} \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(1)}}\underbrace{\mathbb{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\beta)} + \frac{n-1}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\frac{1}{n-1}z^{\mathsf{T}}z(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\\
&\quad + \frac{1}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\left(\overline{m}^{(0)\mathsf{T}}{\overline{\boldsymbol{M}}^{(0)}}^{-1}\overline{m}^{(0)} - \overline{m}^{(1)\mathsf{T}}{\overline{\boldsymbol{M}}^{(1)}}^{-1}\overline{m}^{(1)}\right)(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\\
&\overset{(14)}{=} \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(1)}}\underbrace{\mathbb{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\beta)} + \frac{n-p}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\underbrace{\hat{\sigma}^2_{LS}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\hat{\beta}_{LS})}\\
&\quad + \frac{p}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\overline{\sigma}^2_{PDC}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\\
&\overset{(15)}{=} \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(1)}}\underbrace{\mathbb{E}[\sigma^2]\frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\beta)} + \frac{n^{(0)}+p}{n^{(0)}+n+p}\frac{n^{(0)}}{n^{(1)}}\underbrace{{\sigma^{(0)}}^2\frac{n}{n^{(0)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{=:\mathbb{V}^{(0)}(\beta)}\\
&\quad + \frac{2(n-p)}{n^{(0)}+n+p}\frac{n}{n^{(1)}}\underbrace{\hat{\sigma}^2_{LS}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{\mathbb{V}(\hat{\beta}_{LS})} - \frac{n^{(1)}+p}{n^{(0)}+n+p}\underbrace{{\sigma^{(1)}}^2\frac{n}{n^{(1)}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}}_{=:\mathbb{V}^{(1)}(\beta)}.
\end{aligned}
$$

In these equations, it is possible to isolate $\mathbb{V}(\beta)$, $\mathbb{V}(\hat{\beta}_{LS})$ and, in the formulation with (15), the newly defined $\mathbb{V}^{(0)}(\beta)$ and $\mathbb{V}^{(1)}(\beta)$. However, all three versions do not constitute a weighted average, even when the formula for $\mathbb{E}[\sigma^2 \mid z]$ did have this property. Just as in the SCP model, $\mathbb{V}(\beta \mid z)$ can increase if the automatic abatement of the elements in $\overline{\boldsymbol{M}}^{(1)}$ is overcompensated by a strong increase of $\mathbb{E}[\sigma^2]$. Again, this reaction to prior-data conflict is unspecific because it depends on $\mathbb{E}[\sigma^2 \mid z]$ alone.

5 Discussion and Outlook

For both the SCP and CCCP model, $\mathbb{E}[\beta \mid z]$ results as a weighted average of $\mathbb{E}[\beta]$ and $\hat{\beta}_{LS}$, such that the posterior distribution on β will be centered around a mean somewhere between $\mathbb{E}[\beta]$ and $\hat{\beta}_{LS}$, with the location depending on the respective weights. The weights for the CCCP model appear especially intuitive: $\hat{\beta}_{LS}$ is weighted with the sample size n, whereas $\mathbb{E}[\beta]$ has the weight $n^{(0)}$ reflecting the "prior strength" or "pseudocounts". Due to this, prior-data conflict may at most affect the variances only. Indeed, for both prior models, $\mathbb{E}[\sigma^2 \mid z]$ can increase in the presence of prior-data conflict, as shown by the decompositions in Sections 3.2 and 4.2. Through the formulations (6) and (16) for $\mathbb{V}(\beta \mid z)$, respectively, it can be seen that the posterior distribution on β can in fact become less pointed than the prior when prior-data conflict is at hand. Nevertheless, the effect might be not be as strong as desired: In the formulations (5) and (14), respectively, the effect is based only on one term of the decomposition, and furthermore may be foiled through the automatic decrease of $\boldsymbol{M}^{(1)}$ and $\overline{\boldsymbol{M}}^{(1)}$.

Probably the most problematic finding is that this (possibly weak) reaction affects the whole variance-covariance matrix uniformally, and thus, in both models, the reaction to prior-data conflict is by no means component-specific.

Therefore, the prior models lack the capability to mirror the appropriateness of the prior assignments for each covariate separately. As the SCP model is already the most general approach in the class of conjugate priors, this non-specificity feature seems inevitable in Bayesian linear regression based on precise conjugate priors.

In fact, as argued in Section 1, a more sophisticated and specific reaction to prior-data conflict is only possible by extending considerations beyond the traditional concept of probability. Imprecise probabilities, as a general methodology to cope with the multidimensional nature of uncertainty, appears promising here. For generalized Bayesian approaches, the possibility to mirror the quality of prior knowledge is one of the main reasons for the paradigmatic skip from classical probability to interval / imprecise probability. In this framework ambiguity in the prior specification can be modeled by considering sets $\mathscr{M}_\vartheta$ of prior distributions. In the most common approach based on Walley's Generalized Bayes Rule (Walley 1991), posterior inference is then based on a set of posterior distributions $\mathscr{M}_{\vartheta|z}$, resulting from updating the distributions in the prior set element by element.

Of particular computational convenience are again models based on conjugate priors, as developed for the Dirichlet-Multinomial model by Walley (1996), see also Bernard (2009), and for i.i.d. exponential family sampling models by Quaeghebeur & de Cooman (2005), which were extended by Walter & Augustin (2009) to allow an elegant handling of prior-data conflict: With the magnitude of the set $\mathscr{M}_{\vartheta|z}$ mapping the posterior ambiguity, high prior-data conflict leads, ceteris paribus, to a large $\mathscr{M}_{\vartheta|z}$, resulting in high imprecision in the posterior probabilities, and cautious inferences based on it, while in the case of no prior-data conflict $\mathscr{M}_{\vartheta|x}$, and thus the imprecision, is much smaller.

The essential technical ingredient to derive this class of models is the general construction principle also underlying the CCCP model from Section 4, and thus that model can be extended directly to a powerful corresponding imprecise probability model.[7] A detailed development is beyond the scope of this contribution.

Acknowledgements We are very grateful to Erik Quaeghebeur and Frank Coolen for intensive discussions on foundations of generalized Bayesian inference, and to Thomas Kneib for help at several stages of writing this paper.

References

Augustin, T., Coolen, F. P., Moral, S. & Troffaes, M. C. (eds) (2009). *ISIPTA'09: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, Durham University, Durham, UK, July 2009*, SIPTA.

Augustin, T. & Hable, R. (2009). On the impact of robust statistics on imprecise probability models: a review, *ICOSSAR'09: The 10th International Conference on Structural Safety and Reliability, Osaka*. To appear.

Bernard, J.-M. (2009). Special issue on the Imprecise Dirichlet Model. *International Journal of Approximate Reasoning*.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*, Wiley, Chichester.

Bousquet, N. (2008). Diagnostic of prior-data agreement in applied bayesian analysis, **35**: 1011–1029.

Coolen-Schrijner, P., Coolen, F., Troffaes, M. & Augustin, T. (2009). Special Issue on Statistical Theory and Practice with Imprecision, *Journal of Statistical Theory and Practice* **3**.

de Cooman, G., Vejnarová, J. & Zaffalon, M. (eds) (2007). *ISIPTA'07: Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications, Charles University, Prague, Czech Republic, July 2007*, SIPTA.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms, *Q. J. Econ.* pp. 643–669.

Evans, M. & Moshonov, H. (2006). Checking for prior-data conflict, *Bayesian Analysis* **1**: 893–914.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum-likelihood estimator in generalized linear-models, *Annals of Statistics* **13**: 342–368.

Fahrmeir, L. & Kneib, T. (2006). Structured additive regression for categorial space-time data: A mixed model approach, *Biometrics* **62**: 109–118.

Fahrmeir, L. & Kneib, T. (2009). Propriety of posteriors in structured additive regression models: Theory and empirical evidence, *Journal of Statistical Planning and Inference* **139**: 843–859.

Fahrmeir, L., Kneib, T. & Lang, S. (2007). *Regression. Modelle, Methoden und Anwendungen*, Springer, New York.

Fahrmeir, L. & Raach, A. (2007). A Bayesian semiparametric latent variable model für mixed responses, *Psychometrika* **72**: 327–346.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.

Higgins, J. P. T. & Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis, *Statistics in Medicine* **15**: 2733–2749.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making, *Science* **310**: 1680–1683.

Huber, P. J. & Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities, *The Annals of Statistics* **1**: 251–263.

[7] For σ^2 fixed, the model from Section 3 can be comprised under a more general structure that also can be extended to imprecise probabilities, see Walter et al. (2007) and Walter (2006) for details.

Kauermann, R., Krivobokova, T. & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smooting, *J. Roy. Statist. Soc. Ser. B* **71**: 487–503.

Klir, G. J. & Wierman, M. J. (1999). *Uncertainty-based Information. Elements of Generalized Information Theory*, Physika, Heidelberg.

Kneib, T. & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression for interval-censored survival times, **34**: 207–228.

Kyburg, H. (1987). Logic of statistical reasoning, *in* S. Kotz, N. L. Johnson & C. B. Read (eds), *Encyclopedia of Statistical Sciences*, Vol. 5, Wiley-Interscience, New York, pp. 117–122.

O'Hagan, A. (1994). *Bayesian Inference, Vol. 2B of Kendall´s Advanced Theory of Statistics*, Arnold, London.

Quaeghebeur, E. & de Cooman, G. (2005). Imprecise probability models for inference in exponential families, *in* F. G. Cozman, R. Nau & T. Seidenfeld (eds), *ISIPTA '05: Proc. 4th Int. Symp. on Imprecise Probabilities and Their Applications*, pp. 287–296.

Ríos Insua, D. & Ruggeri, F. (eds) (2000). *Robust Bayesian Analysis*, Springer, New York.

Scheipl, F. & Kneib, T. (2009). Locally adaptive Bayesian P-splines with a normal-exponential-gamma prior, *Computational Statistics & Data Analysis* **53**: 3533–3552.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London.

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles, *Journal of the Royal Statistical Society. Series B. Methodological* **58**: 3–57.

Walter, G. (2006). *Robuste Bayes-Regression mit Mengen von Prioris — Ein Beitrag zur Statistik unter komplexer Unsicherheit*, Master's thesis, Department of Statistics, LMU Munich. Diploma thesis. `http://www.stat.uni-muenchen.de/~walter`.

Walter, G. & Augustin, T. (2009). Imprecision and prior-data conflict in generalized Bayesian inference., *Journal of Statistical Theory and Practice* **3**: 255–271.

Walter, G., Augustin, T. & Peters, A. (2007). Linear regression analysis under sets of conjugate priors, *in* G. de Cooman, J. Vejnarová & M. Zaffalon (eds), *ISIPTA '07: Proc. 5th Int. Symp. on Imprecise Probabilities and Their Applications*, pp. 445–455.

Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept*, Physika, Heidelberg.

An Efficient Model Averaging Procedure for Logistic Regression Models Using a Bayesian Estimator with Laplace Prior

Christian Heumann and Moritz Grenke

Abstract Modern statistics has developed numerous methods for linear and nonlinear regression models, but the correct treatment of model uncertainty is still a difficult task. One approach is model selection, where, usually in a stepwise procedure, an optimal model is searched with respect to some (asymptotic) criterion such as AIC or BIC. A draw back of this approach is, that the reported post model selection estimates, especially for the standard errors of the parameter estimates, are too optimistic. A second approach is model averaging, either frequentist (FMA) or Bayesian (BMA). Here, not an optimal model is searched for, but all possible models are combined by some weighting procedure. Although conceptually easy, the approach has mainly one drawback: the number of potential models can be so large that it is infeasible to calculate the estimates for every possible model. In our paper we extend an idea of Magnus et al. (2009), called WALS, to the case of logistic regression. In principal, the method is not restricted to logistic regression but can be applied to any generalized linear model. In the final stage it uses a Bayesian esimator using a Laplace prior with a special hyperparameter.

1 Introduction

Model selection using criterions such as AIC or BIC is nowadays routinely used for any type of regression models. However, the reported standard errors after model selection are usually too optimistic because the uncertainty introduced by the selection process is not taken into account. Leeb & Pötscher (2003, 2005b, 2005a, 2006, 2008) showed in a series of papers that the problems can be very serious. As an alternative to model selection, one can use model averaging strategies. Here, mainly two proce-

Christian Heumann and Moritz Grenke
Institut für Statistik, Ludwigstraße 33, Ludwig-Maximilians-Universität München, Germany, e-mail: christian.heumann@stat.uni-muenchen.de, moritz.grenke@campus.lmu.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_5,

dures have been proposed. Bayesian model averaging (BMA) was used by Raftery et al. (1997) and Hoeting et al. (1999). Also conceptually easy, the approach suffers from the fact that a (too) large number of models has to be fitted and there are also a number of problems associated with the computation of the marginal likelihood. Frequentist model averaging (FMA) as proposed by Hjort & Claeskens (2003) suffers from the same principal infeasibility as BMA. Usually, it is therefore proposed to concentrate on and average a relatively small number of plausible models. As an alternative to BMA, Magnus et al. (2009) proposed a method called WALS (weighted average least squares) for the linear regression model, when a number of predictors is required to be in the model, e.g. by substantial considerations or theory, and it is not clear whether additional predictors should be added to the model or not. This is a very practical view, e.g. in epidemiological studies. The epidemiologists often force the statistician to include confounder variables such as age, sex, smoking behaviour and social background into the model while it is only a hypothesis that expositions such as dust, air pollution, etc. have an (additional) influence on the disease state. In our short communication we try to generalize the approach of Magnus et al. (2009) to a logistic regression model as an example of a generalized linear model. Since this is still work in progress, the method is not fully developed. But we can show by a small simulation study that the method works better than the usual maximum likelihood method with respect to the mean squared error of the parameter estimates in the finite sample situation.

2 Model Averaging

Let us assume a logistic regression model in matrix form

$$\log\left(\frac{\pi}{1-\pi}\right) = X_1\beta_1 + X_2\beta_2 \,, \tag{1}$$

where π is $n \times 1$, X_1 is $n \times p_1$, β_1 is $p_1 \times 1$, X_2 is $n \times p_2$ and β_2 is $p_2 \times 1$. We assume that X_1 contains an intercept term, therefore $p_1 \geq 1$, and that $p = p_1 + p_2 < n$, i.e. we do not allow $p > n$.

We distinguish between X_1, the regressors we want in the model, i.e. the variables we force to be included in the model, and X_2, the additional regressors which may be important to be included in the model or not. In general there are then 2^{p_2} models to consider in a model averaging procedure, generated by setting different subsets of β_2 equal to zero. If, e.g., $p_2 = 3$, we have the following possible models: the model containing none of the predictors of X_2, three models containing one of the predictors of X_2, three models containing two of the predictors of X_2 and the model containing all three predictors of X_2. A specific model M_j, $j = 1, \ldots, 2^{p_2}$ can then be written as

$$\log\left(\frac{\pi}{1-\pi}\right) = X_1\beta_1 + X_{2j}\beta_{2j} \,, \tag{2}$$

where X_{2j} and β_{2j} are the corresponding subsets of predictors and parameters which are not a priori set to zero. Then model averaging uses two steps: in the first step, the parameter estimates are computed conditionally on the selected model and in the second step these estimates are combined (averaged) by some weighting procedure and we have already mentioned two approaches, BMA and FMA.

2.1 Orthogonalization

Now let us consider the matrix

$$M = W - WX_1(X_1WX_1)^{-1}X_1'W \ , \tag{3}$$

where $W = \text{diag}[\pi(1-\pi)]$ is $n \times n$. Note, that in the following, we constantly work with this true variance–covariance matrix of the response vector. Furthermore, the spectral decomposition of

$$X_2'MX_2 = P\Lambda P' \tag{4}$$

leads to

$$P'X_2'MX_2P = \Lambda \ , \tag{5}$$

where Λ is diagonal and P is orthogonal, such that $P'P = PP' = I$. This can be used to define a new matrix of covariates, X_2^*,

$$X_2^* = X_2P\Lambda^{-\frac{1}{2}} \ , \tag{6}$$

and a new parameter vector β_2^*,

$$\beta_2^* = \Lambda^{\frac{1}{2}}P'\beta_2 \ . \tag{7}$$

The original parameter vector β_2 can be computed by the reverse transformation

$$\beta_2 = P\Lambda^{-\frac{1}{2}}\beta_2^* \ . \tag{8}$$

We note that

$$X_2^*\beta_2^* = X_2P\Lambda^{-\frac{1}{2}}\Lambda^{\frac{1}{2}}P'\beta_2 = X_2\beta_2 \ , \tag{9}$$

and with (5)

$$\begin{aligned} X_2^{*'}MX_2^* &= \Lambda^{-\frac{1}{2}}P'X_2'MX_2P\Lambda^{-\frac{1}{2}} \\ &= \Lambda^{-\frac{1}{2}}\Lambda\Lambda^{-\frac{1}{2}} \\ &= I \ . \end{aligned} \tag{10}$$

2.2 Unrestricted Maximum Likelihood Estimation

Now assume, that we find the maximum likelihood estimator with the modified design matrix $X = X_1 : X_2^*$ where : means the column by column concatenation of X_1 and X_2^*. Then the asymptotic covariance matrix of $(\hat{\beta}_1', \hat{\beta}_2^{*'})'$ is

$$\begin{pmatrix} X_1'WX_1 & X_1'WX_2^* \\ X_2^{*'}WX_1 & X_2^{*'}WX_2^* \end{pmatrix}^{-1} . \tag{11}$$

Further let

$$Q = (X_1'WX_1)^{-1}X_1'WX_2^* . \tag{12}$$

The inverse can be computed explicitly using the formula for a partitioned inverse. Especially, the covariance of $\hat{\beta}_2^*$ can then be written as

$$\begin{aligned} &\operatorname{cov}(\hat{\beta}_2^*) = \\ &= \left\{ X_2^{*'}WX_2^* - X_2^{*'}WX_1(X_1'WX_1)^{-1}X_1'WX_2^* \right\}^{-1} \\ &= \left\{ X_2^{*'}\left[W - WX_1(X_1'WX_1)^{-1}X_1'W \right] X_2^* \right\}^{-1} \\ &= \left\{ X_2^{*'}MX_2^* \right\}^{-1} \\ &= I , \end{aligned} \tag{13}$$

see (10). Furthermore, because of (13), the covariance of $\hat{\beta}_1$ is

$$\begin{aligned} &\operatorname{cov}(\hat{\beta}_1) = \\ &= (X_1'WX_1)^{-1}\left(I + X_1'WX_2^* I X_2^{*'}WX_1(X_1'WX_1)^{-1} \right) \\ &= (X_1'WX_1)^{-1} + QQ' . \end{aligned} \tag{14}$$

Thus, asymptotically, $\hat{\beta}_2^*$ is normally distributed with covariance matrix I. Therefore, the components of $\hat{\beta}_2^*$ are also asymptotically independent. We note that the covariance of $\hat{\beta}_1$ and $\hat{\beta}_2^*$ is

$$\begin{aligned} \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2^*) &= -(X_1'WX_1)^{-1}X_1'WX_2^* I \\ &= -(X_1'WX_1)^{-1}X_1'WX_2^* \\ &= -Q . \end{aligned} \tag{15}$$

In summary, we can conclude that, asymptotically,

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2^* \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 \\ \beta_2^* \end{pmatrix}, \begin{pmatrix} (X_1'WX_1)^{-1} + QQ' & -Q \\ -Q' & I \end{pmatrix} \right) . \tag{16}$$

2.3 Restricted Approximate Maximum Likelihood Estimation

Fahrmeir & Tutz (2001, p. 141) give a first-order approximation for the restricted MLE $\hat{\beta}_{1r}$ under the hypothesis $\beta_2^* = 0$. Using former results, we get

$$\hat{\beta}_{1r} = \hat{\beta}_1 + Q\hat{\beta}_2^* \,. \tag{17}$$

Rearranging this equation leads to

$$\hat{\beta}_1 = \hat{\beta}_{1r} - Q\hat{\beta}_2^* \,. \tag{18}$$

Therefore we can even approximate $\hat{\beta}_1$ by first fitting the restricted model and then the unrestricted model. Why we do this will become clear in the following. Taking expectations at both sides leads to

$$\mathrm{E}(\hat{\beta}_{1r}) = \beta_1 + Q\beta_2^* \,. \tag{19}$$

Now consider the case that we only use a subset of the variables in X_2^*, say X_{2j}^* for model M_j. Then it still holds that $X_{2j}^{*\prime} M X_{2j}^* = I$. In a similar manner we define Q_j as in (12). Using only a subset means that a number of parameters of $\hat{\beta}_2^*$ is set equal to zero. Assume the parameters not set equal to zero as β_{2j}^*. Then proceeding as in the previous subsection 2.2, we get

$$\mathrm{cov}(\hat{\beta}_{2j}^*) = I \,, \tag{20}$$

$$\mathrm{cov}(\hat{\beta}_{1j}) = (X_1' W X_1)^{-1} + Q_j Q_j' \,, \tag{21}$$

and

$$\mathrm{cov}(\hat{\beta}_{1j}, \hat{\beta}_{2j}^*) = -Q_j \,. \tag{22}$$

Using again the first-order approximation, we get

$$\hat{\beta}_{1r} = \hat{\beta}_{1j} + Q_j \hat{\beta}_{2j}^* \,, \tag{23}$$

or

$$\hat{\beta}_{1j} = \hat{\beta}_{1r} - Q_j \hat{\beta}_{2j}^* \,. \tag{24}$$

Taking now expectation at both sides, using (19) leads to

$$\mathrm{E}(\hat{\beta}_{1j}) = \beta_1 + Q\beta_2^* - Q_j \beta_{2j}^* \,. \tag{25}$$

Now, recall, that the components of $\hat{\beta}_2^*$ are (asymptotically) independent. Defining a matrix L_j as a diagonal matrix with a one on the diagonal if the component of $\hat{\beta}_2^*$ is included in the model and zero elsewhere, we can write

$$X_{2j}^* = X_2^* L_j \tag{26}$$

$$Q_j = QL_j \ , \tag{27}$$

where obviously null components are added. Similarly,

$$\hat{\beta}^*_{2j} = L_j \hat{\beta}^*_2 \ . \tag{28}$$

Then equations (24) and (25) become

$$\hat{\beta}_{1j} = \hat{\beta}_{1r} - QL_jL_j\hat{\beta}^*_2 = \hat{\beta}_{1r} - QL_j\hat{\beta}^*_2 \tag{29}$$

and

$$\begin{aligned} \mathrm{E}(\hat{\beta}_{1j}) &= \beta_1 + Q\beta^*_2 - QL_j\beta^*_2 \\ &= \beta_1 + Q(I - L_j)\beta^*_2 \ , \end{aligned} \tag{30}$$

since $L_j^2 = L_j$. In summary, we can conclude that, asymptotically, for model M_j, we have

$$\begin{pmatrix} \hat{\beta}_{1j} \\ \hat{\beta}^*_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_1 + Q(I - L_j)\beta^*_2 \\ L_j\beta^*_2 \end{pmatrix} , \begin{pmatrix} (X_1'WX_1)^{-1} + QL_jQ' & -QL_j \\ -L_jQ' & L_j \end{pmatrix} \right) \ . \tag{31}$$

2.4 Model Averaging

Since the components of $\hat{\beta}^*_2$ are (asymptotically) independent, model averaging has only to take place for estimating β_1. The estimator is

$$b_1 = \sum_{j=1}^{2^{p_2}} \lambda_j \hat{\beta}_{1j} \ , \tag{32}$$

where the sum is taken over all 2^{p_2} different models by setting a subset of β^*_2 equal to zero, and the λ_j are weights, satisfying minimal conditions, such as $\sum_j \lambda_j = 1$, $\lambda_j \geq 0$. They usually depend on the data in some way. Now, from results of the previous section, we get

$$\begin{aligned} b_1 &= \hat{\beta}_{1r} - Q(\sum_{j=1}^{2^{p_2}} \lambda_j L_j)\hat{\beta}^*_2 \\ &= \hat{\beta}_{1r} - QL\hat{\beta}^*_2 \ , \end{aligned} \tag{33}$$

with $L = \sum_j \lambda_j L_j$. As Magnus et al. (2009) remark, while the L_j are deterministic, L is random, since the weights may be data dependent. But, as shown by Magnus et al. (2009), L is a (full rank) diagonal matrix. What we can derive is that, using (19),

$$\mathrm{E}(b_1) = \beta_1 + Q\beta^*_2 - Q\mathrm{E}(L\hat{\beta}^*_2) = \beta_1 - Q\mathrm{E}(L\hat{\beta}^*_2 - \beta^*_2) \ . \tag{34}$$

The next step in Magnus et al. (2009) uses the fact that, in a linear regression model, $\hat{\beta}_{1r}$ and $\hat{\beta}_2^*$ are independent. We conjecture that this is not the case here, especially if we, now for the first time, state, that W has to be estimated to make the estimators operational. Nevertheless, if the correlation is weak, we can assume, at least as an approximation, that

$$\mathrm{var}(b_1) \approx (X_1' W X_1)^{-1} + Q\mathrm{var}(L\hat{\beta}_2^*)Q' \tag{35}$$

and

$$\mathrm{mse}(b_1) \approx (X_1' W X_1)^{-1} + Q\mathrm{mse}(L\hat{\beta}_2^*)Q' \,. \tag{36}$$

Now, we conjecture, that b_1 is a good estimator in the mean squared error sense, if $L\hat{\beta}_2^*$ is a good estimator of β_2^*. Since L is diagonal, and if one chooses each diagonal entry of L, say l_j, to be dependent only on the corresponding element $\hat{\beta}_{2(j)}^*$, then, since the components of $\hat{\beta}_2^*$ are (asymptotically) independent, one can reduce the problem of finding the best estimator (in the mse sense) to finding a best estimator for each component of β_2^* separately. Thus, we can reduce the problem from dimension 2^{p_2} to p_2. Magnus et al. (2009) propose to use the Laplace estimator, element by element, using the fact that the elements of $\hat{\beta}_2^*$ are (asymptotically, independently) normal distributed with variance one. The Laplace estimator has been derived by these authors using a single normal observation (here a single component of $\hat{\beta}_2^*$) with variance one to estimate the mean (here a single component of β_2^*). Especially, they search a solution for estimating the mean η of a normal distribution from one single observation $x \sim N(\eta, 1)$. They found that, compared to four other estimators, the Laplace estimator using a Laplace prior distribution

$$p(\eta) = \frac{c}{2}\exp(-c|\eta|) \tag{37}$$

with hyperparameter $c = \log(2)$, implying that the prior median of η is zero and the prior median of η^2 is one, leads to an estimator with bounded risk, good properties around $|\eta| = 1$, and which is near-optimal in terms of minimax regret. Magnus et al. (2009) additionally state that it also comes closest to the prior idea of ignorance.

In Theorem 1 in Magnus et al. (2009), the posterior moments are derived and are given by

$$\mathrm{E}(\eta|x) = \frac{1+h(x)}{2}(x-c) + \frac{1-h(x)}{2}(x+c) \tag{38}$$

$$\mathrm{Var}(\eta|x) = 1 + c^2(1-h^2(x)) - \frac{c(1+h(x))\phi(x-c)}{\Phi(x-c)} \,, \tag{39}$$

where

$$h(x) = \frac{\exp(-cx)\Phi(x-c) - \exp(cx)\Phi(-x-c)}{\exp(-cx)\Phi(x-c) + \exp(cx)\Phi(-x-c)} \,, \tag{40}$$

and ϕ and Φ denote the density and cumulative distribution function of the standard normal distribution.

2.5 Algorithm

Now, we give a short description of the model averaging algorithm. Note that, at this stage, we use an estimator for the matrix W, derived from the unrestricted maximum likelihood estimator.

- Determine the regressors X_1 and X_2
- Compute the unrestricted maximum likelihood estimator $(\hat{\beta}_1, \hat{\beta}_2)$, using the original design matrices X_1 and X_2. From this model, get the predicted values $\hat{\pi}$ and the estimator for W as $W = \text{diag}[\hat{\pi}(1-\hat{\pi})]$.
- Using W of the previous step, compute the matrix M and the spectral decomposition of $X_2' M X_2 = P \Lambda P'$ and $X_2^* = X_2 P \Lambda^{-\frac{1}{2}}$.
- Compute the unrestricted maximum likelihood estimator $(\hat{\beta}_1, \hat{\beta}_2^*)$, using the original design matrix X_1 and the modified design matrix X_2^*.
- Using the Laplace prior with $c = \log(2)$, compute element by element the posterior moments $\tilde{\eta}_j = \text{E}(\eta_j | \hat{\beta}_{2(j)}^*)$ and $\omega_j = \text{var}(\eta_j | \hat{\beta}_{2(j)}^*)$, for $j = 1, \ldots, p_2$. Set $\tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_{p_2})$ and $\Omega = \text{diag}(\omega_1, \ldots, \omega_{p_2})$.
- Compute the model averaging estimators. First, compute the retransformed estimator for β_2, $b_2 = P \Lambda^{-\frac{1}{2}} \tilde{\eta}$. Then compute b_1. Option 1 is to use the restricted estimator $\hat{\beta}_{1r}$ from the restricted model using X_1 only and to compute $b_1 = \hat{\beta}_{1r} - Q b_2$. Option 2 (which we used in the simulation studies) is to compute b_1 using only X_1 in the model, but with an offset $X_2^* \tilde{\eta}$ or, equivalently, $X_2 b_2$ in the fitting procedure.
- The variance of b_2 can be computed as $\text{var}(b_2) = P \Lambda^{-\frac{1}{2}} \Omega \Lambda^{-\frac{1}{2}} P'$. We have several options for computing the variance of b_1, but this is left to future research. One possible option is to use $\text{var}(b_1) \approx F^{-1} + Q \text{var}(b_2) Q'$, where F^{-1} is computed from the offset model and $Q = F^{-1} X_1' \text{diag}[\tilde{\pi}(1-\tilde{\pi})] X_2$, where $\tilde{\pi}$ are the predicted values of the offset model.

Note that we have not used any subsequent iterations which would also be an option.

3 Simulation Study

We have set up a small simulation study showing the performance of our proposed procedure. In fact, we plan to extend the study to include Bayesian Model Averaging using programs discussed in Hoeting et al. (1999) which were only recently made available in the programming environment R. But for now, we simply compare the proposed method to the usual maximum likelihood estimate. The parameters of the simulation study were as follows:

- Sample size $n = 200$
- Number of covariates: $p = 10$ (including the intercept). The covariates were generated as multivariate normal using 5 different equi-correlations: 0, 0.2, 0.4, 0.6, 0.8.

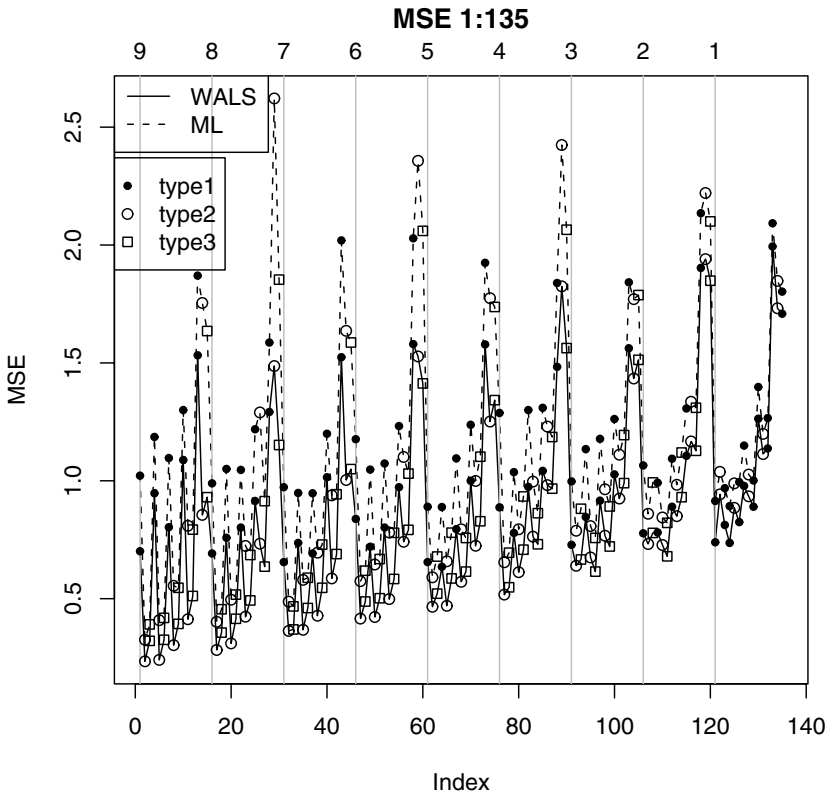


Fig. 1 All 135 parameter settings in one view. The vertical lines denote where p_2 changes to 9, 8, 7, 6, 5, 4, 3, 2 and 1, $p_1 = 10 - p_2$

- The number of columns p_1 of X_1 was ranging from $1, \ldots, 9$. p_2 was set to $10 - p_1$.
- Three different types of parameter settings for the true parameters β_1 and β_2 were used: $\beta_1 = (1, -1, 1, -1, \ldots,)'$, where the cut was set after p_1 components. For β_2 we used 3 different types:

  { Type 1: $\beta_2 = (-1, 1, -1, 1, \ldots)'$
  { Type 2: $\beta_2 = (-\frac{1}{2}, -\frac{1}{3}, -\frac{1}{4}, \ldots)'$
  { Type 3: $\beta_2 = (1, 0, 0, \ldots, 0)'$

  For each type, the cut was set after p_2 components.
- The number of simulations of each setting was $S = 500$.

That leads in summary to 135 different (9 p_1's $\times$ 5 correlations $\times$ 3 types for β) parameter combinations. For each of these combinations, the design matrix X was generated only once and the $S = 500$ simulations differ because of generating each

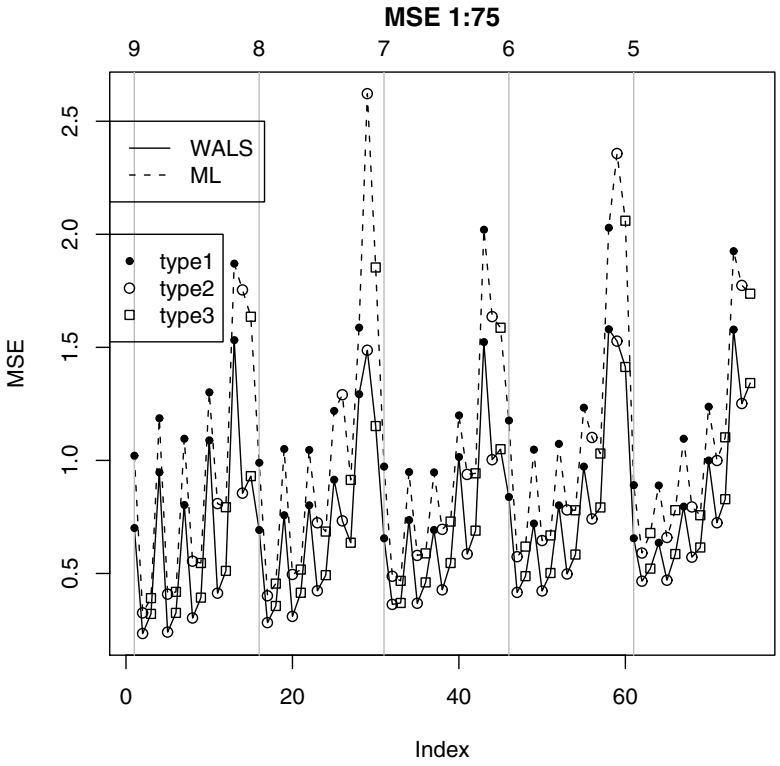


Fig. 2 Parameter settings 1:75 in one view. The vertical lines denote where p_2 changes to 9, 8, 7, 6 and 5, $p_1 = 10 - p_2$

time a new response vector. We then computed the empirical mean squared error for the maximum likelihood estimate and our model averaging estimate. The results are shown in figure 1 and give the trend involved in the results. Higher correlations in the covariates e.g. lead to higher MSEs. For improving the viewing of the results, we give two additional figures, one for the settings 1 to 75 (Figure 2), and one figure for the settings 76:135 (Figure 2). Remarkably, there is not even one situation, where the usual maximum likelihood estimator is better than the proposed estimator, although the sample size is reasonable compared to the number of covariates.

4 Conclusion and Outlook

We have developed a new method for model averaging following ideas of Magnus et al. (2009). Substantial modifications were necessary since Magnus et al. (2009)

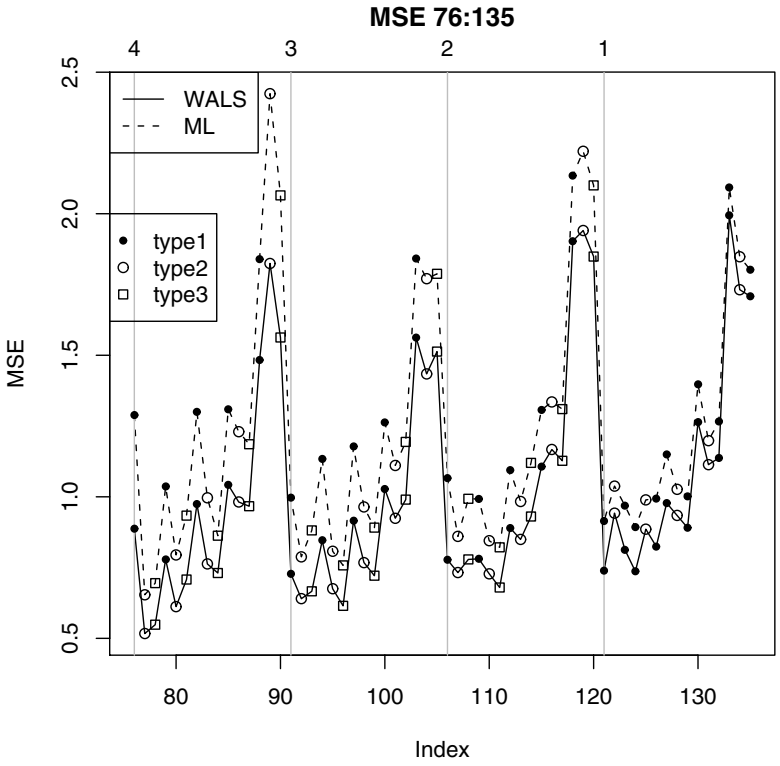


Fig. 3 Parameter settings 76:135 in one view. The vertical lines denote where p_2 changes to 4, 3, 2 and 1, $p_1 = 10 - p_2$

only treat the linear regression case with homoscedastic variance. It seems that the method has some potential as shown in the simulation study. We plan further to work on the formulas for the standard errors (we have tried different versions and it seems that some will work but further investigation is necessary) and to include BMA into the simulation study for comparison.

References

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer; 2nd edition, New York.

Hjort, N. & Claeskens, G. (2003). Frequentist model average estimators, *Journal of the American Statistical Association* **98**: 879–899.

Hoeting, J. A., D., M., Raftery, A. E. & Volinsky, C. (1999). Bayesian model averaging: a tutorial, *Statistical Science* **14**: 382–401.

Leeb, H. & Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations, *Econometric Theory* **19**: 100–142.

Leeb, H. & Pötscher, B. M. (2005a). The distribution of a linear predictor after model selection: Conditional finite-sample distributions and asymptotic approximations, *Journal of Statistical Planning and Inference* **134**: 64–89.

Leeb, H. & Pötscher, B. M. (2005b). Model selection and inference: Facts and fiction, *Econometric Theory* **21**: 21–59.

Leeb, H. & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators?, *Annals of Statistics* **34**: 2554–2591.

Leeb, H. & Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators?, *Econometric Theory* **24**: 338–376.

Magnus, J. R., Powell, O. & Prüfer, P. (2009). A comparison of two model averaging techniques with an application to growth empirics, *Journal of Econometrics, to appear* .

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models, *Journal of the American Statistical Association* **92**: 179–191.

Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA

Leonhard Held, Birgit Schrödle and Håvard Rue

Abstract Model criticism and comparison of Bayesian hierarchical models is often based on posterior or leave-one-out cross-validatory predictive checks. Cross-validatory checks are usually preferred because posterior predictive checks are difficult to assess and tend to be too conservative. However, techniques for statistical inference in such models often try to avoid full (manual) leave-one-out cross-validation, since it is very time-consuming. In this paper we will compare two approaches for estimating Bayesian hierarchical models: Markov chain Monte Carlo (MCMC) and integrated nested Laplace approximations (INLA). We review how both approaches allow for the computation of leave-one-out cross-validatory checks without re-running the model for each observation in turn. We then empirically compare the two approaches in an extensive case study analysing the spatial distribution of bovine viral diarrhoe (BVD) among cows in Switzerland.

Key words: Bayesian hierarchical models; INLA; Leave-one-out cross-validation; MCMC; Posterior predictive model checks

1 Introduction

Bayesian hierarchical models are widely used in applied statistics. Inference is typically based on Markov chain Monte Carlo (MCMC), a computer-intensive simulation-based approach. However, integrated nested Laplace approximations

Leonhard Held and Birgit Schrödle
Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland,
e-mail: leonhard.held@ifspm.uzh.ch, birgit.schroedle@ifspm.uzh.ch

Håvard Rue
Department of Mathematical Science, Norwegian University of Science and Technology, Trondheim, Norway, e-mail: havard.rue@math.ntnu.no

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_6,

(INLA) are a promising alternative to inference via MCMC in latent Gaussian models (Rue et al. 2009). The methodology is particularly attractive if the latent Gaussian model is a Gaussian Markov random field (GMRF) (Rue & Held 2005). In contrast to empirical Bayes approaches (Fahrmeir et al. 2004), the INLA approach incorporates posterior uncertainty with respect to hyperparameters. Examples where INLA is applicable include generalized linear mixed models (Breslow & Clayton 1993), disease mapping (Besag et al. 1991) including ecological regression (Clayton & Bernardinelli 1992, Natário & Knorr-Held 2003), spatial and spatio-temporal GMRF models (Gössl et al. 2001), dynamic (generalized) linear models (Fahrmeir 1992) and structured additive regression (Fahrmeir & Lang 2001).

A particularly interesting feature of INLA is that it provides leave-one-out cross-validatory model checks without re-running the model for each observation in turn. In this paper we review the computation of the conditional predictive ordinate (CPO) and the probability integral transform (PIT) in INLA and compare it with computation of the corresponding quantities using MCMC. We also consider posterior predictive model checks based on the whole data as an alternative to cross-validation. Section 2 reviews INLA and gives a detailed description how cross-validatory model checks are computed with INLA. Section 3 describes how these quantities are computed with MCMC. An extensive case study using an example from spatial epidemiology is described in Section 4 to compare the two approaches. We close with some discussion in Section 5.

2 The INLA Approach

The following section reviews INLA as an approach for approximate Bayesian inference in latent Gaussian models and shows how posterior and cross-validatory predictive checks can be computed using INLA.

2.1 Parameter Estimation with INLA

Consider a three-stage Bayesian hierarchical model based on an observation model $\pi(y|x) = \prod_i \pi(y_i|x_i)$, a parameter model $\pi(x|\theta)$, and a hyperprior $\pi(\theta)$. Here $y = (y_1, \ldots, y_n)$ denotes the observed data, x are unknown parameters which typically follow a GMRF, and θ are unknown hyperparameters. Note that reparametrization and parameter augmentation can be used to achieve $\pi(y_i|x) = \pi(y_i|x_i)$. The dimension of x will often be larger than n and we assume in the following that only the first n components of x are directly linked to the observations y.

Consider now the marginal posterior density

$$\pi(x_i|y) = \int_\theta \pi(x_i|\theta, y)\, \pi(\theta|y)\, d\theta$$

of the i-th component x_i of x. INLA approximates this by

$$\widetilde{\pi}(x_i|y) = \sum_k \widetilde{\pi}(x_i|\theta_k, y)\ \widetilde{\pi}(\theta_k|y)\ \Delta_k$$

using an approximation $\widetilde{\pi}(x_i|\theta, y)$ of $\pi(x_i|\theta, y)$ and an additional approximation $\widetilde{\pi}(\theta|y)$ of the marginal posterior density $\pi(\theta|y)$ of the hyperparameters θ. The weights Δ_k are chosen appropriately.

We first describe how $\pi(\theta|y)$ is approximated. Clearly,

$$\pi(x, \theta, y) = \pi(x|\theta, y) \times \pi(\theta|y) \times \pi(y), \tag{1}$$

so it follows that

$$\pi(\theta|y) \propto \frac{\pi(x, \theta, y)}{\pi(x|\theta, y)} \text{ for all } x.$$

INLA approximates $\pi(\theta|y)$ using a Laplace approximation (Tierney & Kadane 1986):

$$\widetilde{\pi}(\theta|y) \propto \frac{\pi(x, \theta, y)}{\widetilde{\pi}_G(x|\theta, y)}|_{x=x^*(\theta)}.$$

The numerator can be easily evaluated based on (1). The denominator $\widetilde{\pi}_G(x|\theta, y)$ is the Gaussian approximation (Rue et al. 2009, Section 2.2) of $\pi(x|\theta, y)$ and $x^*(\theta)$ is the mode of the full conditional $\pi(x|\theta, y)$, obtained through a suitable iterative algorithm. The approximate posterior density $\widetilde{\pi}(\theta|y)$ is "numerically explored" to obtain suitable support points θ_k and the respective weights Δ_k.

For approximating the first component $\pi(x_i|\theta, y)$, a Gaussian approximation (Rue & Martino 2007), easily extractable from $\widetilde{\pi}_G(x|\theta, y)$,

$$\tilde{\pi}_G(x_i|\theta, y) = \mathrm{N}(x_i; \mu_i(\theta), \sigma_i^2(\theta))$$

can be used. The approximation can be improved using a Laplace approximation

$$\tilde{\pi}_{LA}(x_i|\theta, y) \propto \mathrm{N}(x_i; \mu_i(\theta), \sigma_i^2(\theta)) \times \exp(\text{cubic spline}(x_i)),$$

or a simplified Laplace approximation based on the skew-normal distribution (Azzalini & Capitano 1999), for details see Rue et al. (2009).

As suggested in Fahrmeir & Kneib (2009), it is instructive to compare the INLA approach with a REML/Empirical Bayes estimation in mixed models. In the empirical Bayes approach no hyperprior $\pi(\theta)$ is necessary, so the (RE)ML marginal likelihood corresponds to the marginal posterior $\pi(\theta|y)$. The (RE)ML marginal likelihood is maximized and only the (RE)ML estimate of θ is used, so no uncertainty with respect to θ is taken into account. The empirical Bayes estimate of x_i corresponds to the Gaussian approximation of $\pi(x_i|\theta, y)$ with θ fixed at the (RE)ML estimate. Hierarchical likelihood (Lee et al. 2006) is a variation of this.

2.2 Posterior Predictive Model Checks with INLA

In order to check the fit of a Bayesian model posterior predictive checks were proposed by Gelman et al. (1996). The underlying concept of such checks is the posterior predictive distribution of a replicate observation Y_i which has density

$$\pi(y_i|y) = \int \pi(y_i|x_i, y) \cdot \pi(x_i|y)\, dx_i. \tag{2}$$

In Stern & Cressie (2000) it is suggested to use the posterior predictive p-value

$$\text{Prob}(Y_i \leq y_i^{obs}|y)$$

as a measure of model fit, here y_i^{obs} denotes the actually observed count. If data are discrete, the posterior predictive mid-p-value (Berry & Armitage 1995, Marshall & Spiegelhalter 2003)

$$\text{Prob}(Y_i < y_i^{obs}|y) + \frac{1}{2}\text{Prob}(Y_i = y_i^{obs}|y)$$

can be used instead. An alternative quantity that may be of interest is the posterior predictive ordinate $\pi(y_i^{obs}|y)$. Small values of $\pi(y_i^{obs}|y)$ will indicate an outlying observation.

Extreme posterior predictive (mid-)p-values can be used to identify observations that diverge from the assumed model. However, one drawback concerning the interpretation of posterior predictive p-values is that they do not have a uniform distribution even if the data come from the assumed model. See Hjort et al. (2006), Marshall & Spiegelhalter (2007) and references therein for further details.

We will now explain how posterior p-values can be computed with INLA (Rue et al. 2009). INLA returns an estimate of the posterior marginal of x_i in a discretised way: For $j = 1, \ldots, J$ support points $x_i^{(j)}$ an estimate $\widetilde{\pi}(x_i^{(j)}|y)$ of the posterior density $\pi(x_i^{(j)}|y)$ is given. The support points are chosen such that they cover all areas with non-negligible posterior density. The value of the posterior predictive density (2) can then be approximated using the trapezoidal rule:

$$\hat{\pi}(y_i|y) \approx \sum_{j=2}^{J} \pi(y_i|\frac{1}{2}(x_i^{(j-1)} + x_i^{(j)})) \cdot \frac{1}{2}(x_i^{(j)} - x_i^{(j-1)})(\widetilde{\pi}(x_i^{(j)}|y) + \widetilde{\pi}(x_i^{(j-1)}|y)). \tag{3}$$

Of course, alternative techniques such as Simpson's rule can also be used. For discrete data, the posterior predictive (mid-)p-value can easily be derived as the sum of such probabilities. For $y_i = y_i^{obs}$ we obtain an estimate of the posterior predictive ordinate.

2.3 Leave-one-out Cross-validation with INLA

INLA routinely computes the DIC (Spiegelhalter et al. 2002), a commonly used Bayesian model choice criterion. However, DIC may underpenalize complex models with many random effects (Plummer 2008, Riebler & Held 2009). Alternatively, the conditional predictive ordinate (CPO) (Pettit 1990, Geisser 1993) and the cross-validated probability integral transform (PIT) (Dawid 1984) are available in INLA:

$$\begin{aligned} \text{CPO}_i &= \pi(y_i^{obs}|y_{-i}), \\ \text{PIT}_i &= \text{Prob}(Y_i \leq y_i^{obs}|y_{-i}). \end{aligned}$$

Here y_{-i} denotes the observations y with the i-th component omitted. This facilitates the computation of the cross-validated log-score (Gneiting & Raftery 2007) for model choice. Similarly, PIT histograms (Czado et al. 2009) can be computed to assess calibration of out-of-sample predictions.

We will now describe how these quantities are computed in INLA without re-running the model. Throughout we assume that $y_{-i} = y_{-i}^{obs}$. However, we keep the explicit notation y_i^{obs} for the i-th observation to avoid confusion with other possible realisations of the corresponding random variable Y_i. As before, the vector y will always contain the observed data including y_i^{obs}.

First note that

$$\text{CPO}_i = \int \pi(y_i^{obs}|y_{-i},\theta)\pi(\theta|y_{-i})d\theta, \tag{4}$$

$$\text{PIT}_i = \int \text{Prob}(Y_i \leq y_i^{obs}|y_{-i},\theta)\pi(\theta|y_{-i})d\theta. \tag{5}$$

The first term in the integral in (4) now equals

$$\pi(y_i^{obs}|y_{-i},\theta) = 1 \Big/ \int \frac{\pi(x_i|y,\theta)}{\pi(y_i^{obs}|x_i,\theta)} dx_i. \tag{6}$$

To see this, first note that

$$\pi(x_i|y_{-i},\theta) = \frac{\pi(x_i|y,\theta)\pi(y_i^{obs}|y_{-i},\theta)}{\pi(y_i^{obs}|x_i,\theta)}. \tag{7}$$

Integration with respect to x_i gives (6).

In practice, (6) is computed using numerical integration. The denominator of the ratio in the integral in (6) is the likelihood contribution of the i-th observation and known. However, only an approximation $\widetilde{\pi}(x_i|y,\theta)$ of the numerator $\pi(x_i|y,\theta)$ is known using INLA, as described in Section 2.1. It depends on the accuracy of this approximation how accurate the numerical integration is. In particular, it may happen that the ratio $\widetilde{\pi}(x_i|y,\theta)/\pi(y_i^{obs}|x_i,\theta)$ is multimodal or tends to infinity for extreme values of x_i. It may also be difficult to locate the region of interest, i.e. the region with non-negligible contributions of $\pi(x_i|y,\theta)/\pi(y_i^{obs}|x_i,\theta)$. Such features are an artefact

and a consequence of an imprecise approximation of the numerator $\pi(x_i|y,\theta)$ in the tails. Fortunately, INLA flags such problematic cases, for details see Section 4.

The first term in the integral in (5) can be written as

$$\text{Prob}(Y_i \leq y_i^{obs}|y_{-i},\theta) = \int \text{Prob}(Y_i \leq y_i^{obs}|x_i,\theta)\pi(x_i|y_{-i},\theta)dx_i.$$

The first term in this integral can be computed easily from the likelihood. The second term is available from (7) using $\pi(y_i^{obs}|y_{-i},\theta)$ as computed in (6). As before, $\pi(x_i|y,\theta)$ is available approximately through INLA.

Finally, we need to compute

$$\pi(\theta|y_{-i}) = \frac{\pi(\theta|y)\pi(y_i^{obs}|y_{-i})}{\pi(y_i^{obs}|y_{-i},\theta)}. \tag{8}$$

The denominator $\pi(y_i^{obs}|y_{-i},\theta)$ is known from (6). An approximation to $\pi(\theta|y)$ is available from Section 2.1. Therefore, the normalizing constant

$$\pi(y_i^{obs}|y_{-i}) = 1\Big/\int \frac{\pi(\theta|y)}{\pi(y_i^{obs}|y_{-i},\theta)}d\theta \tag{9}$$

of (8) can be approximately calculated as

$$\widetilde{\pi}(y_i^{obs}|y_{-i}) = 1\Big/\sum_k \frac{\widetilde{\pi}(\theta_k|y)}{\widetilde{\pi}(y_i^{obs}|y_{-i},\theta_k)}\Delta_k. \tag{10}$$

Here the θ_k's are support points of the approximate marginal posterior density $\widetilde{\pi}(\theta|y)$, which has been obtained in the first step of the INLA fitting procedure as described in Section 2.1. So the estimate $\widetilde{\pi}(y_i^{obs}|y_{-i})$ is the *weighted harmonic mean* of the $\widetilde{\pi}(y_i^{obs}|y_{-i},\theta_k)$'s, $k = 1,\ldots,K$, with weights $w_k = \widetilde{\pi}(\theta_k|y)\Delta_k$.

All terms appearing in (4) and (5) are now computed. Final approximation of PIT_i using (5) is based on support points θ_k as in (10) by replacement of the integral with a finite sum. Concerning CPO_i, note that (4) has been approximated already in (10), so the additional integration is not necessary.

3 Predictive Model Checks with MCMC

MCMC delivers samples $x^{(1)},\ldots,x^{(S)}$ from the posterior distribution $\pi(x|y)$. Similarly, samples $\theta^{(1)},\ldots,\theta^{(S)}$ from the posterior distribution $\pi(\theta|y)$ of the hyperparameters can be obtained on a routine basis. These samples are typically dependent, but suitable "thinning" can be be applied to obtain approximately independent samples.

3.1 Posterior Predictive Model Checks with MCMC

Within MCMC the posterior predictive p-values can be derived by drawing a replicate observation $Y_i^{(s)}$ for each of the $s = 1, ..., S$ samples $x_i^{(s)}$ of the MCMC run and counting, how many replicated observations are less than or equal to the actually observed count y_i^{obs}. For discrete data, the posterior predictive mid-p-value and the posterior predictive ordinate can be computed analogously.

If the likelihood $\pi(y_i|x_i)$ is available in closed form, an alternative approach is to average the likelihood across all samples $x_i^{(s)}$ from $\pi(x_i|y)$:

$$\widehat{\pi}(y_i|y) = \frac{1}{S}\sum_{s=1}^{S} \pi(y_i|x_i^{(s)}).$$

This technique is known as Rao-Blackwellization (Gelfand & Smith 1990, Robert & Casella 2004, Casella & Robert 1996) and is typically more accurate than the approach based on replicates $Y_i^{(s)}$ from the predictive density. However, the Monte-Carlo error of the sample-based version is easier to assess so we have used this estimate in Section 4.

3.2 Leave-one-out Cross-validation with MCMC

Omitting the dependence on θ in (6) we obtain

$$\pi(y_i^{obs}|y_{-i}) = 1 \Big/ \int \frac{\pi(x_i|y)}{\pi(y_i^{obs}|x_i)} dx_i. \tag{11}$$

The immediate Monte-Carlo estimate of (11) is simply the harmonic mean of the likelihood values $\pi(y_i^{obs}|x_i)$,

$$\hat{\pi}(y_i^{obs}|y_{-i}) = 1 \Big/ \frac{1}{S}\sum_{s=1}^{S} \frac{1}{\pi(y_i^{obs}|x_i^{(s)})}, \tag{12}$$

evaluated at samples $x_i^{(1)}, \ldots, x_i^{(S)}$ from $\pi(x_i|y)$. This estimate goes back at least to Gelfand (1996) and is very easy to use in MCMC applications. However, the harmonic mean can be numerically unstable and may not even follow a central-limit theorem (Newton & Raftery 1994). This manifests itself by the occasional occurrence of a value $x_i^{(s)}$ with small likelihood $\pi(y_i^{obs}|x_i^{(s)})$ and hence large effect on the estimate (12). Indeed, Raftery (1996) has noted that the reciprocal of (12) may not even have finite variance.

However, for the computation of (mid-)p-values the value of $\pi(y_i|y_{-i})$ needs to be known for all $y_i \leq y_i^{obs}$. An importance sampling approach (Robert & Casella 2004) can be adopted to compute $\pi(y_i|y_{-i})$ for any y_i, not necessarily equal to y_i^{obs}. First

rewrite $\pi(y_i|y_{-i})$ as

$$\begin{aligned}\pi(y_i|y_{-i}) &= \int \pi(y_i|x_i)\pi(x_i|y_{-i})dx_i \\ &= \int \pi(y_i|x_i)\frac{\pi(x_i|y_{-i})}{\pi(x_i|y)}\pi(x_i|y)dx_i.\end{aligned}$$

The importance sampling estimate of $\pi(y_i|y_{-i})$ based on samples $x_i^{(1)},\ldots,x_i^{(S)}$ from $\pi(x_i|y)$ is hence

$$\hat{\pi}(y_i|y_{-i}) = \frac{\sum_{s=1}^{S}\pi(y_i|x_i^{(s)})w_i^{(s)}}{\sum_{s=1}^{S}w_i^{(s)}} \tag{13}$$

with importance weights

$$w_i^{(s)} = \frac{\pi(x_i^{(s)}|y_{-i})}{\pi(x_i^{(s)}|y)} \propto \frac{1}{\pi(y_i^{obs}|x_i^{(s)})},$$

compare Robert & Casella (2004, Equation (3.10)). For count data, the computation of cross-validatory (mid-)p-values reduces then to summing up the estimates $\hat{\pi}(y_i|y_{-i})$ for $y_i = 0,\ldots,y_i^{obs}$ (Marshall & Spiegelhalter 2003). Note that the importance sampling estimate (13) reduces to the harmonic mean (12), if $y_i = y_i^{obs}$.

The variance of importance sampling estimators is difficult to assess; in fact the estimate may not even have finite variance. In particular, if the weights $w_i^{(s)}$ vary widely, they will give too much importance to only a few values of $\pi(y_i|x_i^{(s)})$ and the estimator (13) will be quite unstable, even for large S. However, we have investigated the weights $w_i^{(s)}$ in Section 4 and have found no weight particularly large relative to the others.

3.3 Approximate Cross-validation with MCMC

We now describe an alternative approach, based on an idea originally presented by Marshall & Spiegelhalter (2003) for approximate cross-validation in disease mapping models via MCMC. The method is based on the assumption that

$$\pi(\theta|y_{-i}) \approx \pi(\theta|y).$$

This assumption is plausible for moderate to large dimension of y, since θ is a *global* hyperparameter. Its posterior distribution based on all observations y should not change much if a single observation y_i is omitted.

The Marshall & Spiegelhalter (2003) *mixed predictive approach* is to generate additional samples

$$\tilde{x}_i^{(s)} \sim \pi(x_i|\theta^{(s)}, y_{-i})$$

$s = 1, \ldots, S$, where $\theta^{(s)}$ is a sample from $\pi(\theta|y)$. The samples $\tilde{x}_i^{(s)}$ do not directly depend on y_i, only indirectly because $\theta^{(s)} \sim \pi(\theta|y)$ does depend on y_i. The $\tilde{x}_i^{(s)}$'s are therefore approximately cross-validated and can be used in various ways to compute the predictive model checks discussed earlier.

A straightforward approach to compute PIT values is to draw additional samples $\tilde{y}_i^{(s)}$ from the pseudo-cross-validated predictive distribution and to compute the proportion of samples which are not larger than the observed value y_i^{obs}. Similarly, CPO values can be estimated based on the proportion of samples equal to y_i^{obs}. Alternatively a Rao-Blackwell approach as described in Section 3.1 can be used. In our application the PIT and CPO values resulting from the sampling strategy and the Rao-Blackwellization were almost identical. Mixed predictive PIT and CPO values shown in the following section are computed using Rao-Blackwellization.

4 Application

In our application we consider a typical example from spatial epidemiology. The data considered are cases of bovine viral diarrhoe (BVD) among cows in Switzerland collected during the year 2008. On behalf of an eradication program each cow in Switzerland was tested and the herd was marked as infected, if one or more diseased cows within this herd were detected. As Switzerland is divided in 184 administrative regions, the number of cases is available aggregated on regional level. Additionally, the Principality of Liechtenstein was included in the analysis. A number of 7164 cases was detected in total. For one region the number of cases is missing.

Under the rare disease assumption the usual starting point is to assume that the number of disease cases y_i in region $i = 1, \ldots, 185$ is Poisson distributed with parameter λ_i, which can be interpreted as the relative risk of the disease in the respective region. Additionally, the number of herds m_i is included in the model as an offset to adjust for the different number of herds living in each region. Using a standard formulation with Poisson observation model and a logarithmic link the relative risk parameter λ_i is modelled using the specification

$$\eta_i = \log(\lambda_i) = \log(m_i) + \psi_i + v_i. \tag{14}$$

The spatially unstructured component v_i is assumed to be i.i.d. normally distributed with zero mean and unknown precision τ_v whereas ψ_i is assumed to be structured in space. To account for the assumption that geographically close areas have similar incidence rates the spatially structured component ψ_i is modelled as an intrinsic Gaussian Markov random field with unknown precision τ_ψ (Rue & Held 2005). This model was proposed by Besag et al. (1991), an extension to include covariates has been considered in Clayton & Bernardinelli (1992). The hyperpriors are chosen as

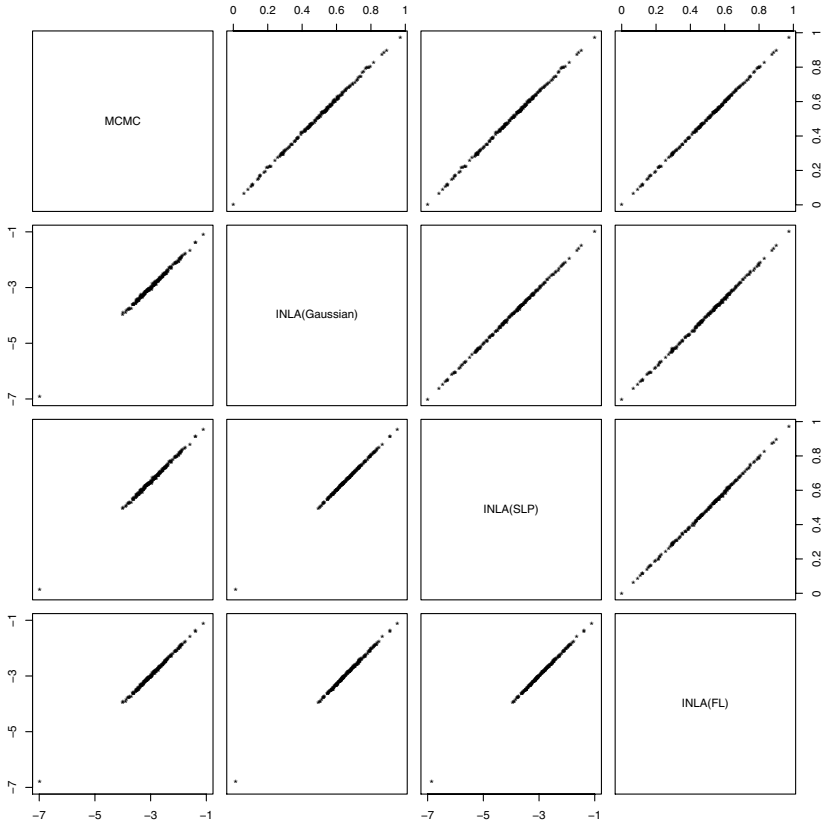


Fig. 1 Scatterplots of posterior predictive mid-p-values (above diagonal) and log posterior predictive ordinates (below diagonal) computed by MCMC and INLA using the Gaussian (Gaussian), simplified Laplace (SLP) and full Laplace (FL) approximation

$\tau_\psi \sim \text{Ga}(1, 0.018)$ and $\tau_v \sim \text{Ga}(1, 0.01)$, compare Bernardinelli et al. (1995) and Schrödle & Held (2009) for some motivation.

For the following analyses an MCMC run of length 930 000 was performed. Using every 30th iteration and a burn-in of 30 000 iterations, 30 000 MCMC samples have been stored. We also tested all three approximation methods available within INLA, as they are known to be differently accurate (Rue & Martino 2007, Rue et al. 2009). All calculations were done using the `inla` program version number 1.526.

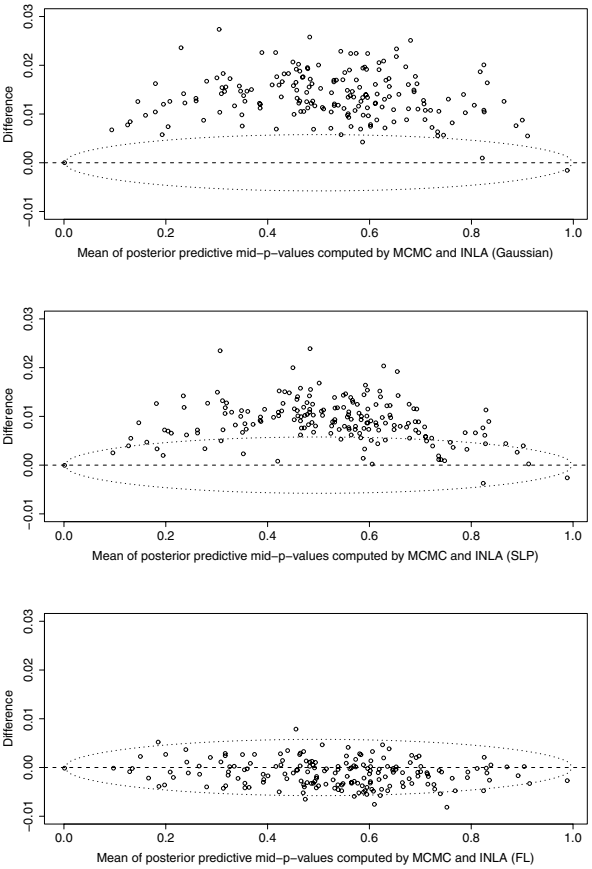


Fig. 2 Bland-Altman plot to investigate the agreement between posterior predictive mid-p-values computed by MCMC vs. INLA using the Gaussian, simplified Laplace and full Laplace approximation. The dotted lines indicate pointwise 95%-confidence intervals based on the Monte-Carlo error attached to the MCMC estimates

4.1 A Comparison of Posterior Predictive Model Checks

In the following the difference between the posterior predictive ordinates and posterior predictive mid-p-values computed by MCMC and INLA using three different approximation methods for the latent Gaussian field will be assessed.

Pairwise scatterplots are shown in Figure 1. The distribution of the posterior predictive ordinates is quite skewed and therefore shown on the log-scale. As can be seen from the plot, the estimates obtained with the four different methods look virtually identical.

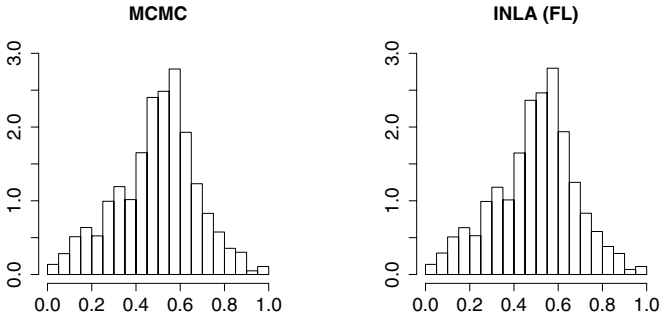


Fig. 3 Adjusted histograms of posterior predictive *p*-values computed by MCMC and INLA using the full Laplace approximation

The extent of agreement between any two methods can be visually examined in more detail using a plot suggested in Bland & Altman (1986), see also Kirkwood & Sterne (2003). The difference between two estimates is plotted on the vertical axis against the mean of each pair on the horizontal axis, see Figure 2. Also shown are 95%-confidence intervals indicating the Monte Carlo error attached to the MCMC estimates. The Monte Carlo standard error has been computed based on the assumption that the MCMC samples are independent. This assumption has been checked by visually inspecting the corresponding empirical autocorrelation functions.

Using this plot systematic bias can be detected and it can be examined if the differences between pairs of estimates depend on the actual value of the estimate. Posterior predictive mid-*p*-values obtained using the Gaussian and simplified Laplace approximation are slightly biased and typically smaller than the corresponding MCMC estimates. The bias is largest for mid-*p*-values around 0.5. For the full Laplace approximation the differences are close to zero and do not show any specific pattern. In fact, nearly all differences are now within the Monte Carlo confidence limits, i.e. the differences can be explained solely by the Monte Carlo error attached to the MCMC estimates. The MCMC estimates based on Rao-Blackwell were even closer to the INLA estimates.

Histograms of posterior predictive mid-*p*-values can be computed in analogy to the PIT histogram (Czado et al. 2009), which was recently proposed for count data. The results are shown in Figure 3 based on MCMC and INLA using the full Laplace approximation. There is virtually no difference to see.

The histograms can be compared with histograms of the cross-validated PIT values in Figure 6. As mentioned in Stern & Cressie (2000) and Marshall & Spiegelhalter (2007) posterior predictive *p*-values are not uniformly distributed and tend to be too conservative as the data are used twice. Indeed, the histograms in Figure 3 are far from uniformity with too many observations having mid-*p*-values around 0.5.

Table 1 Number of unreliable CPO/PIT values for the Gaussian, simplified Laplace and full Laplace approximation

Gaussian	56 unreliable CPO/PIT values
Simplified Laplace	18 unreliable CPO/PIT values
Full Laplace	13 unreliable CPO/PIT values

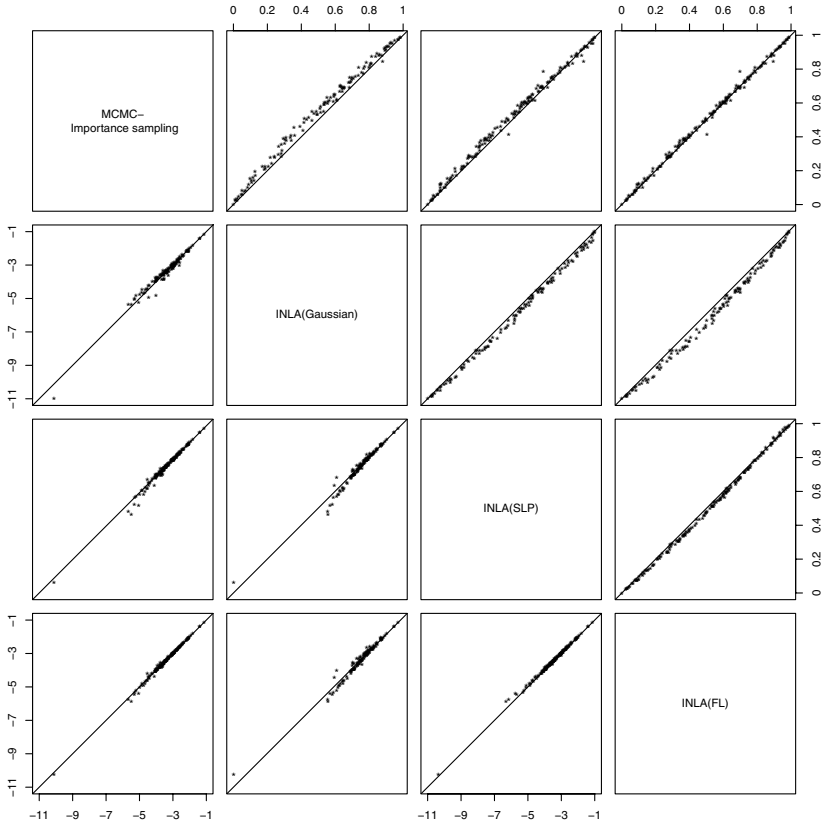


Fig. 4 Scatterplots of leave-one-out cross-validatory predictive mid-p-values (above diagonal) and log conditional predictive ordinates (below diagonal) computed by MCMC vs. INLA using the Gaussian (Gaussian), simplified Laplace (SLP) and full Laplace (FL) approximation

4.2 A Comparison of Leave-one-out Cross-validated Predictive Checks

Leave-one-out cross-validated predictive checks overcome the difficulties of posterior predictive checks mentioned in Section 4.1 and can be used to assess the predictive quality of a model (Marshall & Spiegelhalter 2003, Czado et al. 2009).

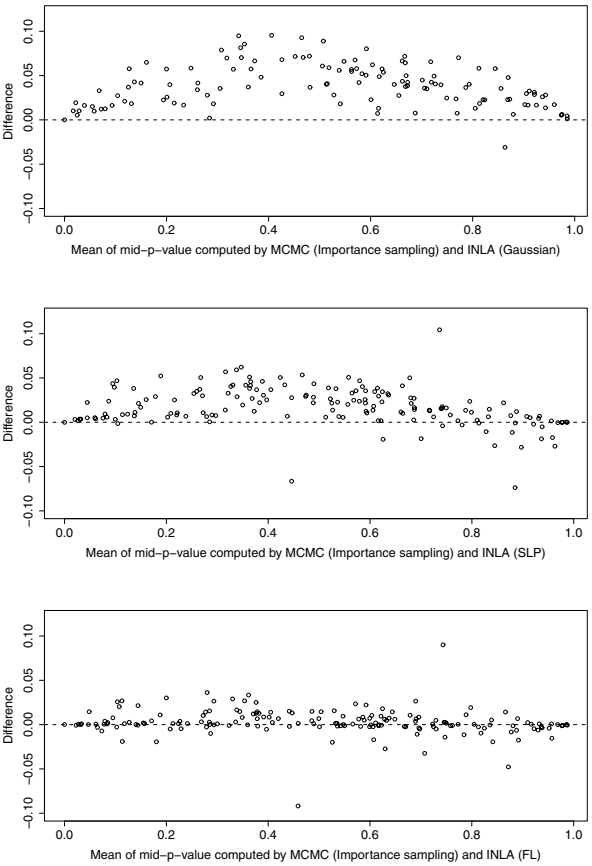


Fig. 5 Bland-Altman plot to investigate the agreement between leave-one-out cross-validated mid-*p*-values computed by MCMC (importance sampling) vs. INLA using the Gaussian, simplified Laplace and full Laplace approximation

Histograms of the PIT values have been proposed to assess the calibration of a model (Czado et al. 2009), the logarithmic score (Gneiting & Raftery 2007), the sum of the log CPO values, can be used for model choice.

INLA returns the CPO and PIT values, as described in Section 2.3. Since the approximation methods for the latent Gaussian field are known to be differently accurate (Rue & Martino 2007, Rue et al. 2009), an empirical comparison is conducted. However, numerical problems may occur when CPO and PIT values are computed in INLA. Some of the CPO and PIT values might not be reliable due to numerical problems in evaluating the integral in (6). INLA automatically stores a file `failure.dat` which contains failure flags for each observation. We considered

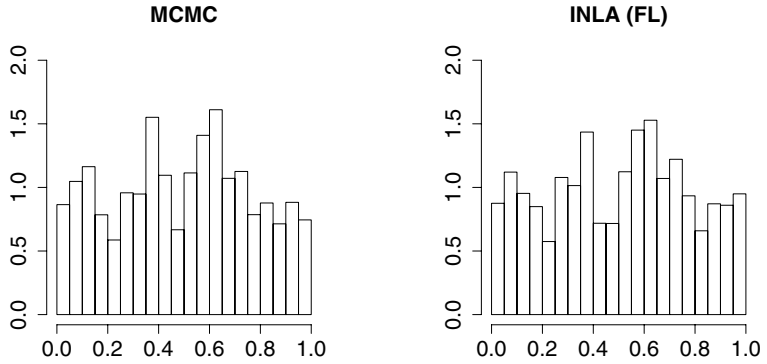


Fig. 6 Adjusted histogram of PIT values computed by MCMC and INLA using the full Laplace approximation

CPO/PIT values with flag equal to 1 as unreliable. Further details on this issue can be found in Martino & Rue (2009).

In Table 1 it is listed for how many observations the computation failed. Most failures occur based on the Gaussian approximation, the full Laplace approximation performs best.

In order to assess the performance of INLA the output will be compared with results from a MCMC analysis based on the estimates (12) and (13). Mid-p- and log CPO values calculated with INLA and MCMC are shown in Figure 4. Each sub-figure is based on all those observations where CPO and PIT values could be computed without failure with the corresponding INLA approximation technique(s) considered.

Figure 4 reveals that the full Laplace approximation is closest to MCMC concerning bias and the differences between the full Laplace and the MCMC output do not show any specific pattern. More details can be seen on the corresponding Bland-Altman plots of the leave-one-out cross-validated mid-p-values, see Figure 5. First of all, a comparison with the corresponding plot showing the posterior predictive mid-p-values (Figure 2) reveals that the differences between MCMC and INLA have increased. However, a similar pattern as in Figure 2 can be seen, with mainly positive differences for the Gaussian and simplified Laplace approximation. In contrast, the mid-p-values computed with the full Laplace approach are closest to the MCMC estimates and do not exhibit a systematic bias. The corresponding PIT histograms are shown in Figure 6 and are quite similar. Note that the PIT histograms are much closer to a uniform distribution than the corresponding posterior predictive histograms shown in Figure 3.

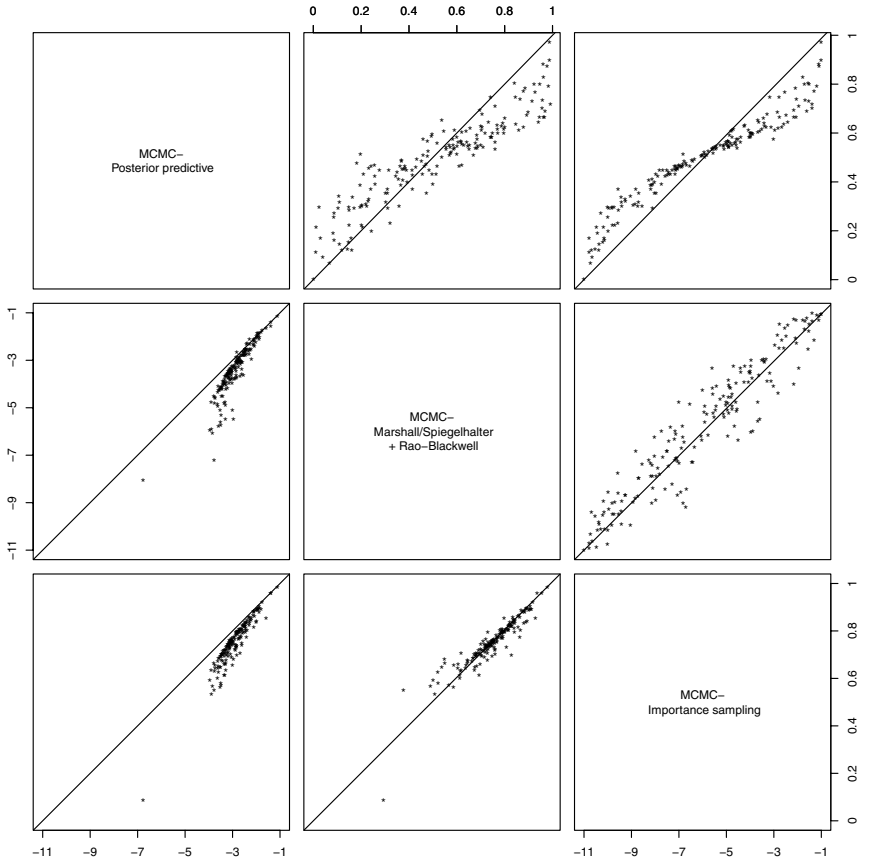


Fig. 7 Scatterplots of mid-p- (above diagonal) and log CPO-values (below diagonal) computed by MCMC using three different approaches: The posterior predictive approach, the mixed predictive approach proposed by Marshall and Spiegelhalter in combination with a Rao-Blackwellization, and importance sampling

4.3 A Comparison of Approximate Cross-validation with Posterior and Leave-one-out Predictive Checks using MCMC

CPO and mid-p-values resulting from a MCMC analysis have also been computed using the mixed predictive approach by Marshall & Spiegelhalter (2003) as described in Section 3.3. The approach is based on posterior samples of the precisions $\tau_v^{(s)}$ and $\tau_\psi^{(s)}$ based on the full data.

Approximately cross-validated samples of η_i and ψ_i are generated in a two-stage procedure based on a reparametrization of model (14) described in Knorr-Held & Rue (2002): First, $\tilde{\psi}_i^{(s)}$ is drawn from the conditional density

$$\tilde{\psi}_i^{(s)} | \psi_{-i}^{(s)}, \tau_\psi^{(s)} \sim N(\frac{1}{n_i} \sum_{j:j\sim i} \psi_j^{(s)}, \frac{1}{n_i \cdot \tau_\psi^{(s)}}).$$

Here n_i denotes the number of neighbours of region i. In a second step, a sample $\tilde{\eta}_i^{(s)}$ of the linear predictor is drawn using

$$\tilde{\eta}_i^{(s)} | \tilde{\psi}_i^{(s)}, \tau_v^{(s)} \sim N(\tilde{\psi}_i^{(s)}, \frac{1}{\tau_v^{(s)}}).$$

This gives pseudo-cross-validated samples $\tilde{\eta}_i^{(s)}$ of the linear predictor, as proposed in Marshall & Spiegelhalter (2003).

Figure 7 compares the mixed predictive approach with the posterior predictive and the cross-validatory approach based on importance sampling. Compared with the importance sampling and the mixed predictive estimates, the posterior predictive estimates are systematically biased. As suspected, the mid-p-values are shrunk towards 0.5. Interestingly, the mixed predictive approach is closer to the ("exact") cross-validatory approach based on importance sampling. There is no systematic bias, although there is some variation in the estimates. This is in contrast to Marshall & Spiegelhalter (2003), who report that the mixed predictive approach performs better than the importance sampling approach in a similar disease mapping model using the well-known Scotland lip cancer data.

5 Discussion

The case study revealed that the cross-validatory checks provided by INLA are close to "exact" importance sampling estimates based on MCMC. The agreement is best if the full Laplace approximation is used. However, the relatively large number of failures is a drawback. Fortunately, these failures are flagged by INLA and it is straightforward to "manually" remove such an observation and to compute the desired leave-one-out quantities directly. The predictive distribution for the observation removed can be calculated in exactly the same way as the posterior predictive distribution, see Section 2.2. For illustration, Figure 8 compares manually computed mid-p-values with the mid-p-values calculated based on the techniques described in Section 2.3 using the full Laplace approximation. The amount of agreement is remarkable.

We finally illustrate how the cross-validated log-score can be used for model comparison. To do so, we have considered two alternative models with either the unstructured or the structured component removed. The logarithmic score in the full model is -3.459, while in the reduced model with no unstructured component the score is even slightly larger (-3.454). However, the score of the model with only an unstructured component is considerably smaller (-3.779). This indicates that the structured component in the model is important, whereas the unstructured

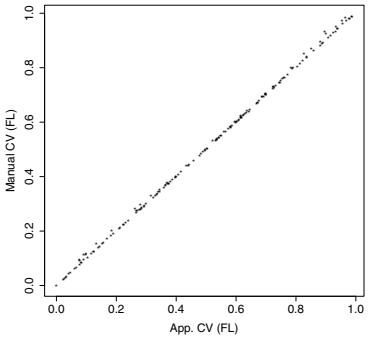


Fig. 8 Scatterplot of manually computed mid-*p*-values using INLA vs. approximate mid-*p*-values obtained from the standard INLA output; the comparison was conducted for the full Laplace approximation

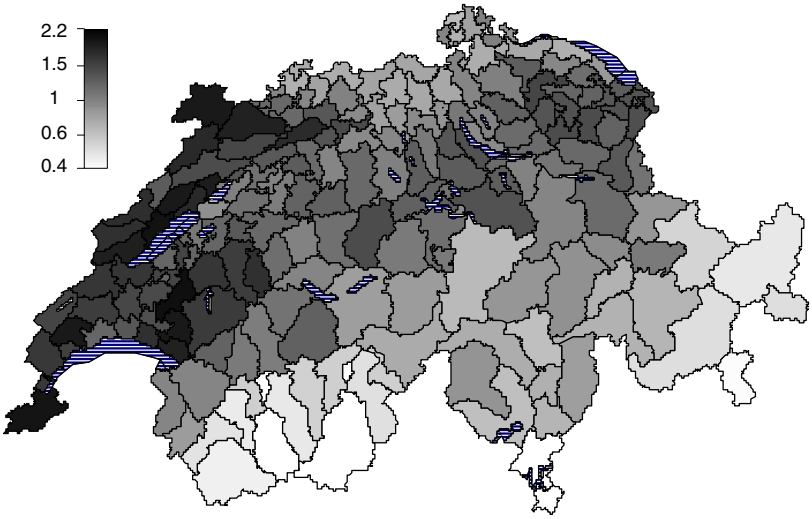


Fig. 9 Fitted relative incidence of BVD in Switzerland, 2008

component can be omitted. The estimated relative incidence obtained from the best model without unstructured component is finally shown in Figure 9.

References

Azzalini, A. & Capitano, A. (1999). Statistical applications of the multivariate skew normal distribution., *Journal of the Royal Statistical Society: Series B* **61**: 579–602.

Bernardinelli, L., Clayton, D. & Montomoli, C. (1995). Bayesian estimates of disease maps: How important are priors?, *Statistics in Medicine* **14**: 2411–2431.

Berry, G. & Armitage, P. (1995). Mid-p confidence intervals: a brief review, *Statistician* **44**: 417–423.

Besag, J., York, J. & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43**(1): 1–59.

Bland, J. & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* **i**: 307–310.

Breslow, N. & Clayton, D. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**: 9–25.

Casella, G. & Robert, C. (1996). Rao-Blackwellisation of sampling schemes, *Biometrika* **83**: 81–94.

Clayton, D. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risk, *in* J. Cuzick et al. (eds), *Geographical and Environmental Epidemiology. Methods for Small Area Studies*, Oxford University Press, pp. 205–220.

Czado, C., Gneiting, T. & Held, L. (2009). Predictive model assessment for count data, *Biometrics* . In press.

Dawid, A. P. (1984). Statistical theory: The prequential approach, *Journal of the Royal Statistical Society. Series A (General)* **147**: 278–292.

Fahrmeir, L. (1992). Posterior mode estimation by extended kalman filtering for multivariate dynamic generalized linear models, *Journal of the American Statistical Association* **87**: 501–509.

Fahrmeir, L. & Kneib, T. (2009). Discussion of Rue et al. (2009), *Journal of the Royal Statistical Society: Series B* **71**: 367.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective, *Statistica Sinica* **14**(3): 731–761.

Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors, *Journal of the Royal Statistical Society. Series C. Applied Statistics* **50**(2): 201–220.

Geisser, S. (1993). *Predictive Inference: An Introduction*, Chapman & Hall, London.

Gelfand, A. E. (1996). Model determination using sampling-based methods, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov chain Monte Carlo in Practice*, Chapman & Hall, pp. 145–161.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**(410): 398–409.

Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica* **6**: 733–807.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**: 359–378.

Gössl, C., Auer, D. P. & Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging, *Biometrics* **57**(2): 554–562.

Hjort, N. L., Dahl, F. A. & Steinbakk, G. H. (2006). Post-processing posterior predictive p-values, *Journal of the American Statistical Association* **101**(475): 1157–1174.

Kirkwood, B. & Sterne, J. (2003). *Medical Statistics*, 2nd edn, Blackwell Publishing, Oxford.

Knorr-Held, L. & Rue, H. (2002). On block updating in Markov random field models for disease mapping, *Scandinavian Journal of Statistics* **29**(4): 597–614.

Lee, Y., Nelder, J. A. & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects - Unified Analysis via H-likelihood*, Chapman & Hall/CRC.

Marshall, E. C. & Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease-mapping methods, *Statistics in Medicine* **22**(4): 1649–1660.

Marshall, E. C. & Spiegelhalter, D. J. (2007). Identifying outliers in Bayesian hierarchical models: a simulation-based approach, *Bayesian Analysis* **2**(2): 409–444.

Martino, S. & Rue, H. (2009). Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the INLA program. `http://www.math.ntnu.no/~hrue/GMRFLib`.

Natário, I. & Knorr-Held, L. (2003). Non-parametric ecological regression and spatial variation, *Biometrical Journal* **45**(6): 670–688.

Newton, M. A. & Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion), *Journal of the Royal Statistical Society: Series B* **56**: 3–48.

Pettit, L. I. (1990). The conditional predictive ordinate for the normal distribution, *Journal of the Royal Statistical Society: Series B* **52**: 175–184.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison, *Biostatistics* **9**(3): 523–539.

Raftery, A. E. (1996). Hypothesis testing and model selection, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov chain Monte Carlo in Practice*, Chapman & Hall, pp. 163–187.

Riebler, A. & Held, L. (2009). The analysis of heterogeneous time trends in multivariate age-period-cohort models, *Technical report*, University of Zurich, Biostatistics Unit. Conditionally accepted for *Biostatistics*.

Robert, C. & Casella, G. (2004). *Monte Carlo Statistical Methods*, Springer.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC Press, London.

Rue, H. & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models, *Journal of Statistical Planning and Inference* **137**: 3177–3192.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion), *Journal of the Royal Statistical Society: Series B* **71**: 319–392.

Schrödle, B. & Held, L. (2009). Evaluation of case reporting data from Switzerland: Spatio-temporal disease mapping using INLA, *Technical report*, University of Zurich, Biostatistics Unit.

Spiegelhalter, D. J., Best, N. G., Carlin, B. R. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society: Series B* **64**: 583–639.

Stern, H. & Cressie, N. (2000). Posterior predictive model checks for disease mapping models, *Statistics in Medicine* **19**: 2377–2397.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *J. Amer. Statist. Assoc.* **81**(393): 82–86.

Data Augmentation and MCMC for Binary and Multinomial Logit Models

Sylvia Frühwirth-Schnatter and Rudolf Frühwirth

Abstract The paper introduces two new data augmentation algorithms for sampling the parameters of a binary or multinomial logit model from their posterior distribution within a Bayesian framework. The new samplers are based on rewriting the underlying random utility model in such a way that only differences of utilities are involved. As a consequence, the error term in the logit model has a logistic distribution. If the logistic distribution is approximated by a finite scale mixture of normal distributions, auxiliary mixture sampling can be implemented to sample from the posterior of the regression parameters. Alternatively, a data augmented Metropolis–Hastings algorithm can be formulated by approximating the logistic distribution by a single normal distribution. A comparative study on five binomial and multinomial data sets shows that the new samplers are superior to other data augmentation samplers and to Metropolis–Hastings sampling without data augmentation.

Key words: Binomial data; multinomial data; data augmentation; Markov chain Monte Carlo; logit model; random utility model

1 Introduction

Applied statisticians and econometricians commonly have to deal with modelling a binary or multinomial response variable in terms of covariates. Examples include modelling the probability of unemployment in terms of risk factors, and modelling choice probabilities in marketing in terms of product attributes. A widely used tool

Sylvia Frühwirth-Schnatter
Institut für Angewandte Statistik, Johannes-Kepler-Universität Linz, Altenbergerstr. 69, 4040 Linz, Austria, e-mail: Sylvia.Fruehwirth-Schnatter@jku.at

Rudolf Frühwirth
Institut für Hochenergiephysik der Österreichischen Akademie der Wissenschaften, Nikolsdorfer Gasse 18, 1050 Wien, Austria, e-mail: fru@hephy.oeaw.ac.at

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_7,

for analyzing such data are binary or multinomial regression techniques using generalized linear models.

Estimation of these models is quite challenging, in particular if latent components are present, such as in random-effects modelling or in state space modelling of discrete data. Fahrmeir & Tutz (2001) provide a review of likelihood-based estimation methods; see also Fahrmeir & Kaufmann (1986a) and Fahrmeir & Kaufmann (1986b) for a rigorous mathematical treatment.

Zellner & Rossi (1984) were the first to perform Bayesian inference for a logit model using importance sampling based on a multivariate Student-t distribution, with mean and scale matrix being equal to the posterior mode and the asymptotic covariance matrix. Starting with Zeger & Karim (1991), many Markov chain Monte Carlo (MCMC) methods have been developed for the Bayesian estimation of the binary and the multinomial logit model. MCMC estimation has been based on single-move adaptive rejection Gibbs sampling (Dellaportas & Smith 1993), Metropolis–Hastings (MH) sampling (Gamerman 1997, Lenk & DeSarbo 2000, Rossi et al. 2005), data augmentation and Gibbs sampling (Holmes & Held 2006, Frühwirth-Schnatter & Frühwirth 2007), and data augmented Metropolis–Hastings sampling (Scott 2009).

In the present article we focus on practical Bayesian inference for binary and multinomial logit models using data augmentation methods. For these models, data augmentation relies on the interpretation of the logit model as a random utility model (McFadden 1974). Frühwirth-Schnatter & Frühwirth (2007) and Scott (2009) base data augmentation directly on this random utility model (RUM) by introducing the utilities as latent variables. Holmes & Held (2006) choose the differences of utilities as latent variables, which is the standard data augmentation method underlying MCMC estimation of probit models, see e.g. Albert & Chib (1993) and McCulloch et al. (2000). We call this interpretation the difference random utility model (dRUM).

In the following we show how to implement data augmentation based on the dRUM representation for the binary and the multinomial logit model. We introduce yet two other data augmentation MCMC samplers by extending the ideas underlying Frühwirth-Schnatter & Frühwirth (2007) and Scott (2009) to the dRUM representation. The extension of the data augmented MH algorithm of Scott (2009) is straightforward, while the extension of the auxiliary mixture sampling approach of Frühwirth-Schnatter & Frühwirth (2007) involves approximating the logistic distribution by a finite scale mixture of normal distributions (Monahan & Stefanski 1992).

We compare the two new data augmentation samplers with the three existing ones for several well-known case studies. This exercise reveals that data augmentation samplers based on the dRUM representation are considerably more efficient in terms of reducing autocorrelation in the resulting MCMC draws than data augmentation based on the RUM. Under the dRUM representation, both auxiliary mixture sampling and data augmented MH sampling are considerably faster than the sampler suggested by Holmes & Held (2006), making the two new samplers an attractive alternative to other data augmentation methods.

Since it is often believed that MCMC sampling without data augmentation can be even more efficient than MCMC sampling with data augmentation, we include several MH algorithms into our comparison, namely the independence MH sampler

suggested in Rossi et al. (2005), a multivariate random walk MH with asymptotically optimal scaling chosen as in Roberts & Rosenthal (2001), and the DAFE-R MH algorithm suggested by Scott (2009). While the independence MH sampler of Rossi et al. (2005) turns out to be superior to any other MH sampler without data augmentation, we find for all but one (very well-behaved) case study that our two new dRUM data augmentation samplers are superior to the independence MH sampler both in terms of efficiency and in terms of the effective sampling rate.

2 MCMC Estimation Based on Data Augmentation for Binary Logit Regression Models

Given a sequence $y_1, \ldots, y_N$ of binary data, the binary logit regression model reads:

$$\Pr(y_i = 1|\boldsymbol{\beta}) = \pi_i(\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta})}, \tag{1}$$

where $\mathbf{x}_i$ is a row vector of regressors, including 1 for the intercept, and $\boldsymbol{\beta}$ is an unknown regression parameter of dimension d. Furthermore we assume that, conditional on knowing $\boldsymbol{\beta}$, the observations are mutually independent.

To pursue a Bayesian approach, we assume that the prior distribution $p(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ is a normal distribution, $\mathrm{No}_d(\mathbf{b}_0, \mathbf{B}_0)$ with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$. The posterior density $p(\boldsymbol{\beta}|\mathbf{y})$ of $\boldsymbol{\beta}$ given all observations $\mathbf{y} = (y_1, \ldots, y_N)$ does not have a closed form:

$$p(\boldsymbol{\beta}|\mathbf{y}) \propto p(\boldsymbol{\beta}) \prod_{i=1}^{N} \frac{[\exp(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}}{1+\exp(\mathbf{x}_i\boldsymbol{\beta})}.$$

Hence Bayesian estimation relies either on data augmentation, to be discussed in this section, or on MH sampling, as in Section 4.

2.1 Writing the Logit Model as a Random Utility Model

The interpretation of a logit model as a random utility (RUM) model was introduced by (McFadden 1974). Two representations of the logit model as a RUM are common.

Let y_{ki}^u be the utility of choosing category k, which is assumed to depend on covariates $\mathbf{x}_i$. The RUM representation corresponding to the logit model reads:

$$y_{0i}^u = \mathbf{x}_i\boldsymbol{\beta}_0 + \delta_{0i}, \qquad \delta_{0i} \sim \mathrm{EV}, \tag{2}$$

$$y_{1i}^u = \mathbf{x}_i\boldsymbol{\beta}_1 + \delta_{1i}, \qquad \delta_{1i} \sim \mathrm{EV}, \tag{3}$$

$$y_i = I\{y_{1i}^u > y_{0i}^u\},$$

where $I\{\cdot\}$ is the indicator function and δ_{0i} and δ_{1i} are i.i.d. random variables following a type I extreme value (EV) distribution with density:

$$f_{\mathrm{EV}}(\delta) = \exp\left(-\delta - e^{-\delta}\right), \tag{4}$$

with expectation $\mathbb{E}(\delta) = \gamma$ and variance $\mathbb{V}(\delta) = \pi^2/6$, where $\gamma = 0.5772$ is Euler's constant.

Thus category 1 is observed, i.e. $y_i = 1$, iff $y_{1i}^u > y_{0i}^u$; otherwise $y_i = 0$. To achieve identifiability, it is assumed that $\boldsymbol{\beta}_0 = \mathbf{0}$, i.e. $\boldsymbol{\beta} = \boldsymbol{\beta}_1$, because only the difference $\boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ can be identified.

An alternative way to write the logit model as an augmented model involving random utilities is the difference random utility model (dRUM), which is obtained by choosing a baseline category, typically 0, and to consider the model involving the differences of the utilities:

$$\begin{aligned} z_i &= \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{Lo}, \\ y_i &= I\{z_i > 0\}, \end{aligned} \tag{5}$$

where $z_i = y_{1i}^u - y_{0i}^u$. The error term $\varepsilon_i = \delta_{1i} - \delta_{0i}$, being the difference of two i.i.d. EV random variables, follows a logistic (Lo) distribution, with density:

$$f_{\mathrm{Lo}}(\varepsilon) = \frac{e^{\varepsilon}}{(1+e^{\varepsilon})^2},$$

with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \pi^2/3$.

For both representations the binary logit regression model (1) results as the marginal distribution of y_i.

2.2 Data Augmentation Based on the Random Utility Model

Several data augmentation algorithms have been suggested for the logit model, all of which are based on the interpretation of a logit model as a random utility model. However, depending on whether the RUM or the dRUM is considered, different data augmentation algorithms result.

Frühwirth-Schnatter & Frühwirth (2007) and Scott (2009) consider the RUM representation (2) for data augmentation and introduce for each i, $i = 1, \ldots, N$, the latent utility of choosing category 1, i.e. $\mathbf{z} = (y_{11}^u, \ldots, y_{1N}^u)$, as missing data. Holmes & Held (2006) use the dRUM representation (5) and introduce the differences in utilities, i.e. $\mathbf{z} = (z_1, \ldots, z_N)$, as missing data. For both representations, data augmentation leads to a two-step MCMC sampler which draws from the conditional densities $p(\mathbf{z}|\boldsymbol{\beta}, \mathbf{y})$ and $p(\boldsymbol{\beta}|\mathbf{z}, \mathbf{y})$, respectively.

For both representations it is possible to sample all components of $\mathbf{z}|\boldsymbol{\beta}, \mathbf{y}$ simultaneously in a simple manner. For the RUM this step reads:

$$y_{1i}^{u} = -\log\left(\mathrm{Ex}\left(1+\lambda_i\right)+\mathrm{Ex}\left(\lambda_i\right)\left(1-y_i\right)\right), \tag{6}$$

where $\mathrm{Ex}(\lambda)$ denotes a random variable from an exponential distribution with density equal to $\lambda\exp(-\lambda y)$. For the dRUM this step reads:

$$z_i = \log(\lambda_i U_i + y_i) - \log(1 - U_i + \lambda_i(1-y_i)), \tag{7}$$

where $U_i \sim \mathrm{Un}[0,1]$. In both cases $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$.

In contrast to sampling from $p(\mathbf{z}|\boldsymbol{\beta},\mathbf{y})$, sampling from $p(\boldsymbol{\beta}|\mathbf{z},\mathbf{y})$ is not possible in closed form, regardless of the underlying representation. Conditional on $\mathbf{z}$, the posterior of $\boldsymbol{\beta}$ is independent of $\mathbf{y}$ and can be derived from regression models (3) or (5), respectively, which are linear in $\boldsymbol{\beta}$, but have a non-normal error term. Various methods have been suggested to cope with this non-normality when sampling the regression parameter $\boldsymbol{\beta}$.

Scott (2009) uses an independence MH algorithm where a normal proposal distribution $\mathrm{No}_d(\mathbf{b}_N,\mathbf{B}_N)$ for $\boldsymbol{\beta}$ is constructed by approximating the non-normal error δ_{1i} appearing in (3) by a normal error with same mean and variance:

$$\mathbf{b}_N = \mathbf{B}_N\left(\mathbf{B}_0{}^{-1}\mathbf{b}_0 + \frac{6}{\pi^2}\mathbf{X}'(\mathbf{z}-\gamma)\right), \qquad \mathbf{B}_N = \left(\mathbf{B}_0{}^{-1} + \frac{6}{\pi^2}\mathbf{X}'\mathbf{X}\right)^{-1}, \tag{8}$$

where row i of the $(N\times d)$ matrix $\mathbf{X}$ is equal to the regressor $\mathbf{x}_i$ of the logit model (1). This leads to a very fast sampler, because $\mathbf{B}_N$ is fixed while running MCMC; however, the acceptance rate might be low in higher dimensional problems.

Frühwirth-Schnatter & Frühwirth (2007) approximate the density of the EV distribution in (3) by the density of a finite normal mixture distribution with 10 components with optimized, but fixed parameters (m_r, s_r^2, w_r) in component r:

$$y_{1i}^{u} = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i|r_i \sim \mathrm{No}\left(m_{r_i}, s_{r_i}^2\right), \quad r_i \sim \mathrm{MulNom}(w_1,\ldots,w_{10}). \tag{9}$$

To perform MCMC estimation they add the latent indicators $\mathbf{r} = (r_1,\ldots,r_N)$ as missing data. The advantage of this additional data augmentation is that conditional on $\mathbf{z}$ and $\mathbf{r}$, the regression parameter $\boldsymbol{\beta}$ may be sampled from regression model (9), leading to a normal conditional posterior. To complete MCMC, each indicator r_i has to be sampled from the discrete posterior $r_i|z_i,\boldsymbol{\beta}$ which is a standard step in finite mixture modelling.

Holmes & Held (2006) represent the logistic distribution appearing in (5) as an infinite scale mixture of normals (Andrews & Mallows 1974):

$$z_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i|\omega_i \sim \mathrm{No}\left(0,\omega_i\right), \quad \sqrt{\omega_i}/2 \sim \mathrm{KS}, \tag{10}$$

where KS is the Kolmogorov–Smirnov distribution. To perform MCMC estimation they add the latent scaling factors $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_N)$ as missing data. Conditional on $\mathbf{z}$ and $\boldsymbol{\omega}$, the regression parameter $\boldsymbol{\beta}$ is sampled from regression model (10), leading to a normal conditional posterior. To complete MCMC, each scaling factor ω_i has to be sampled from the posterior $\omega_i|\boldsymbol{\beta},z_i$ which has no closed form, the density of

the KS distribution having no closed form either, but only a representation involving an infinite series. To sample ω_i, Holmes & Held (2006) implement a single move rejection sampling method based on deriving upper and lower squeezing functions from a truncated series representation of the density of the KS distribution. However, as will be illustrated by the case studies in Section 5, this rejection sampling step makes the algorithm computationally intensive and therefore quite slow.

2.3 Two New Samplers Based on the dRUM Representation

The case studies to be discussed in Section 5 demonstrate a remarkable advantage of Holmes & Held (2006) compared to Frühwirth-Schnatter & Frühwirth (2007), namely that the autocorrelations of the MCMC draws are in general much smaller, making the sampler more efficient. This increase in efficiency turns out to be closely related to using the dRUM rather than the RUM representation of the logit model.

In this paper, we propose two new samplers based on the dRUM representation of the logit model. They are constructed by applying the ideas underlying Frühwirth-Schnatter & Frühwirth (2007) and Scott (2009). As will be illustrated by the case studies, these samplers are much faster than the approach of Holmes & Held (2006), while the efficiency is about the same. Both are much more efficient than the corresponding ones in the RUM representation.

To apply the ideas underlying Scott (2009) to the dRUM representation, we construct a proposal density for $\boldsymbol{\beta}$ by approximating the error term in (5) by a normal error with zero mean and variance equal to $\pi^2/3$. Because a logistic error is closer to the normal distribution than an error following the EV distribution, it is to be expected that the acceptance rate for the resulting independence MH algorithm is much higher than in the RUM model. This expectation is confirmed by our case studies. Details of this sampler are given in Algorithm 1.

Algorithm 1 *Independence Metropolis–Hastings algorithm in the dRUM representation of a logit model.*

Choose starting values for $\boldsymbol{\beta}$ and $\mathbf{z} = (z_1, \ldots, z_N)$ and repeat the following steps:

(a) Propose $\boldsymbol{\beta}^{\text{new}}$ from the proposal $q(\boldsymbol{\beta}^{\text{new}}|\mathbf{z}) = \text{No}_d(\mathbf{b}_N, \mathbf{B}_N)$ with moments:

$$\mathbf{b}_N = \mathbf{B}_N \left(\mathbf{B}_0{}^{-1}\mathbf{b}_0 + \frac{3}{\pi^2}\mathbf{X}'\mathbf{z} \right), \qquad \mathbf{B}_N = \left(\mathbf{B}_0{}^{-1} + \frac{3}{\pi^2}\mathbf{X}'\mathbf{X} \right)^{-1}.$$

Accept $\boldsymbol{\beta}^{\text{new}}$ with probability $\min(\alpha, 1)$, where:

$$\alpha = \frac{p(\mathbf{z}|\boldsymbol{\beta}^{\text{new}})p(\boldsymbol{\beta}^{\text{new}})q(\boldsymbol{\beta}|\mathbf{z})}{p(\mathbf{z}|\boldsymbol{\beta})p(\boldsymbol{\beta})q(\boldsymbol{\beta}^{\text{new}}|\mathbf{z})},$$

and $p(\mathbf{z}|\boldsymbol{\beta})$ is the likelihood of model (5):

$$p(\mathbf{z}|\boldsymbol{\beta}) = \prod_{i=1}^{N} f_{\text{Lo}}(z_i - \mathbf{x}_i\boldsymbol{\beta}).$$

(b) Sample from $z_i|\boldsymbol{\beta}, \mathbf{y}$ for $i = 1, \ldots, N$ as in (7). □

To apply the ideas underlying Frühwirth-Schnatter & Frühwirth (2007) to the dRUM representation, we approximate in (5) the density of the logistic distribution $f_{\text{Lo}}(\varepsilon_i)$ by the density of a normal mixture distribution. As $f_{\text{Lo}}(\varepsilon_i)$ is symmetric around 0, it is sensible to use a finite scale mixture of normal distributions with all component means being equal to 0. For a fixed number H of components this mixture is characterized by component specific variances s_r^2 and weights w_r:

$$f_{\text{Lo}}(\varepsilon_i) \approx \sum_{r=1}^{H} w_r f_{\text{No}}(\varepsilon_i; 0, s_r^2). \tag{11}$$

The contribution of Monahan & Stefanski (1992) to the handbook of the logistic distribution (Balakrishnan 1992) contains such an approximation. As they use a different parameterization, the correct weights and variances in (11) are given by $w_r = p_r$ and $s_r^2 = 1/(s_r^\star)^2$, where p_r and $s_r^\star$ are the values published in Monahan & Stefanski (1992, Table 18.4.1). The corresponding parameters are reproduced in Table 1. We investigate the accuracy of this approximation as well as an alternative approximation in Subsection 2.4.

In general, we expect the number of components necessary to approximate the logistic distribution to be smaller than in Frühwirth-Schnatter & Frühwirth (2007), because the logistic distribution is much closer to the normal distribution than the EV distribution. In fact, the results in Subsection 2.4 show that the 3-component approximation of Monahan & Stefanski (1992) gives about the same acceptance rates as the 10-component approximation in the RUM representation, see Frühwirth-Schnatter & Frühwirth (2007, Table 2), while choosing $H = 6$ leads to an extremely accurate approximation. Thus we recommend choosing $H = 3$ in larger applications, where computing time matters, and to work with $H = 6$ whenever possible.

Having approximated the density of the logistic distribution by a scale mixture of H normal densities, we obtain a representation of the dRUM similar to (10), but ω_i is drawn with fixed probabilities $w_1, \ldots, w_H$ from the set $\{s_1^2, \ldots, s_H^2\}$:

$$\begin{aligned} z_i &= \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i|\omega_i \sim \text{No}(0, \omega_i), \\ \omega_i &= s_{r_i}^2, \qquad r_i \sim \text{MulNom}(w_1, \ldots, w_H). \end{aligned} \tag{12}$$

Note that in this way we approximate the logit model by a very accurate finite scale mixture of probit models.

Like in Holmes & Held (2006), we add the scaling factors $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ as missing data. However, an advantage compared to Holmes & Held (2006) is that instead of sampling ω_i directly, we sample an indicator r_i from the discrete posterior $r_i|z_i, \boldsymbol{\beta}$, which can be done in a very efficient manner, and define $\omega_i = s_{r_i}^2$. Details of this sampler are given in Algorithm 2.

Algorithm 2 *Auxiliary mixture sampling in the dRUM representation of a logit model.*

Choose starting values for $\mathbf{z} = (z_1, \ldots, z_N)$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ and repeat the following steps:

(a) Sample the regression coefficient $\boldsymbol{\beta}$ conditional on $\mathbf{z}$ and $\boldsymbol{\omega}$ based on the normal regression model (12) from $\mathrm{No}_d(\mathbf{b}_N, \mathbf{B}_N)$ with moments:

$$\mathbf{b}_N = \mathbf{B}_N \left({\mathbf{B}_0}^{-1} \mathbf{b}_0 + \sum_{i=1}^{n} \mathbf{x}_i' z_i / \omega_i \right), \qquad \mathbf{B}_N = \left({\mathbf{B}_0}^{-1} + \sum_{i=1}^{n} \mathbf{x}_i' \mathbf{x}_i / \omega_i \right)^{-1}.$$

(b) For $i = 1, \ldots, N$, sample from $z_i | \boldsymbol{\beta}, \mathbf{y}$ as in (7). Sample the indicator r_i conditional on z_i from the discrete density:

$$\Pr(r_i = j | z_i, \boldsymbol{\beta}) \propto \frac{w_j}{s_j} \exp\left[-\frac{1}{2} \left(\frac{z_i - \log \lambda_i}{s_j} \right)^2 \right],$$

and set $\omega_i = s_{r_i}^2$. The quantities $(w_j, s_j^2), j = 1, \ldots, H$ are the parameters of the H component finite mixture approximation tabulated in Table 1. □

2.4 Finite Mixture Approximations to the Logistic Distribution

Monahan & Stefanski (1992) obtained their finite scale mixture approximation by minimizing the KS-distance between the true and the approximate distribution function. The results are given in Table 1.

Because the approximation in Frühwirth-Schnatter & Frühwirth (2007) is based on minimizing the Kullback–Leibler distance between the densities, we redid a related analysis for the logistic distribution. The fitted components are reported in Table 2.

Similarly as in Frühwirth-Schnatter & Frühwirth (2007), we evaluate the effect of using different distance measures and different numbers of mixture components for a simple example, namely Bayesian inference for N i.i.d. binary observations $y_1, \ldots, y_N$, drawn with $\Pr(y_i = 1 | \beta) = \pi = e^\beta / (1 + e^\beta)$.

First we run the data augmented MH algorithm as in Algorithm 2, which corresponds to approximating the logistic distribution by the single normal distribution $\mathrm{No}\left(0, \pi^2/3\right)$, i.e. $H = 1$. Then the data augmented MH algorithm is refined by proposing β from an approximate model, where the logistic distribution is approximated by a scale mixture of H normal distributions with H ranging from 2 to 6. Similarly as in Frühwirth-Schnatter & Frühwirth (2007), we use numerical integration methods to compute the corresponding expected acceptance rate for various values of π and N. Table 3 and Table 4 report, respectively, the expected acceptance rate for the mixture approximation based on Monahan & Stefanski (1992) and the mixture approximation based on the Kullback–Leibler distance.

Table 1 Approximation of the density of the logistic distribution by finite scale mixtures of normal distributions with H components, based on Monahan & Stefanski (1992).

	$H=2$		$H=3$		$H=4$		$H=5$		$H=6$	
r	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$
1	1.6927	56.442	1.2131	25.22	0.95529	10.65	0.79334	4.4333	0.68159	1.8446
2	5.2785	43.558	2.9955	58.523	2.048	45.836	1.5474	29.497	1.2419	17.268
3			7.5458	16.257	4.4298	37.419	3.012	42.981	2.2388	37.393
4					9.701	6.0951	5.9224	20.759	4.0724	31.697
5							11.77	2.3291	7.4371	10.89
6									13.772	0.90745

Table 2 Approximation of the density of the logistic distribution by finite scale mixtures of normal distributions with H components, based on minimizing the K-L distance.

	$H=2$		$H=3$		$H=4$		$H=5$		$H=6$	
r	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$	s_r^2	$100w_r$
1	1.9658	68.966	1.4418	38.834	1.1509	20.638	0.95132	10.159	0.84678	5.8726
2	6.2324	31.034	3.7181	52.719	2.6072	52.008	1.9567	40.842	1.61	28.74
3			9.1139	8.4469	5.6748	25.032	3.8969	36.99	2.8904	36.756
4					11.884	2.3212	7.5025	11.233	5.0772	22.427
5							14.163	0.7753	8.9109	5.8701
6									15.923	0.33466

Table 3 Expected acceptance rate in percent for a Metropolis–Hastings algorithm, based for $H=1$ on the normal distribution $\mathrm{No}\left(0,\pi^2/3\right)$ and for $H>1$ on the scale mixture approximations of Monahan & Stefanski (1992). N is the number of i.i.d. binary observations, and π is the probability of observing 1.

π	N	H					
		1	2	3	4	5	6
0.05	1	90.990	99.165	99.889	99.984	99.998	100.00
	10	89.508	98.628	99.778	99.961	99.994	99.999
	100	88.562	97.956	99.630	99.932	99.986	99.997
	1000	88.267	97.850	99.549	99.906	99.980	99.996
0.20	1	90.787	99.188	99.889	99.980	99.997	100.00
	10	88.491	98.408	99.740	99.957	99.992	99.999
	100	88.273	97.831	99.611	99.927	99.986	99.997
	1000	88.139	97.697	99.518	99.900	99.979	99.995
0.50	1	90.966	99.207	99.897	99.984	99.998	100.00
	10	88.748	98.321	99.724	99.950	99.991	99.998
	100	88.289	97.883	99.630	99.929	99.986	99.997
	1000	88.236	97.678	99.520	99.899	99.978	99.995

As expected, by increasing the number of components the expected acceptance rate approaches 100% for both distances. The expected acceptance rates are rather similar for both distance measure; however, the approximations obtained by Monahan

Table 4 Expected acceptance rate in percent for a Metropolis–Hastings algorithm, based on a mixture approximation with H components minimizing the Kullback–Leibler distance. N is the number of i.i.d. binary observations, and π is the probability of observing 1.

π	N	H				
		2	3	4	5	6
0.05	1	98.788	99.786	99.958	99.992	99.992
	10	97.996	99.580	99.908	99.981	99.988
	100	97.745	99.499	99.879	99.973	99.987
	1000	97.732	99.470	99.875	99.972	99.986
0.20	1	98.750	99.791	99.958	99.992	99.992
	10	97.909	99.548	99.903	99.979	99.988
	100	97.696	99.475	99.875	99.973	99.986
	1000	97.618	99.457	99.873	99.972	99.986
0.50	1	98.818	99.798	99.960	99.992	99.992
	10	97.846	99.534	99.896	99.979	99.988
	100	97.654	99.477	99.873	99.971	99.986
	1000	97.625	99.463	99.873	99.971	99.986

& Stefanski (1992) are slightly better than the approximations based on the Kullback–Leibler distance. Both approximations are already very good for H as small as 3 and are extremely accurate for $H = 6$.

Note that the mixture approximation is applied not only once, but N times. Both tables show how the approximation error accumulates with increasing N. Again, we find that the mixture approximations derived by Monahan & Stefanski (1992) are slightly more reliable in this respect than the mixture approximations based on the Kullback–Leibler distance.

3 MCMC Estimation Based on Data Augmentation for the Multinomial Logit Regression Model

Let $\{y_i\}$ be a sequence of categorical data, $i = 1,\ldots,N$, where y_i is equal to one of $m+1$ unordered categories. The categories are labeled by $L = \{0,\ldots,m\}$, and for any k the set of all categories but k is denoted by $L_{-k} = L \setminus \{k\}$.

We assume that the observations are mutually independent and that for each $k \in L$ the probability of y_i taking the value k depends on covariates $\mathbf{x}_i$ in the following way:

$$\Pr(y_i = k|\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_m) = \pi_{ki}(\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_m) = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta}_k)}{\sum_{l=0}^{m}\exp(\mathbf{x}_i\boldsymbol{\beta}_l)}, \tag{13}$$

where $\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_m$ are category specific unknown parameters of dimension d. To make the model identifiable, the parameter $\boldsymbol{\beta}_{k_0}$ of a baseline category k_0 is set equal to $\mathbf{0}$: $\boldsymbol{\beta}_{k_0} = \mathbf{0}$. Thus the parameter $\boldsymbol{\beta}_k$ is in terms of the change in log-odds relative to

the baseline category k_0. In the following, we assume without loss of generality that $k_0 = 0$. To pursue a Bayesian approach, we assume that the prior distribution $p(\boldsymbol{\beta}_k)$ of each $\boldsymbol{\beta}_k$ is a normal distribution $\mathrm{No}_d(\mathbf{b}_0, \mathbf{B}_0)$ with known hyperparameters $\mathbf{b}_0$ and $\mathbf{B}_0$.

3.1 Data Augmentation in the RUM

As for the binary model, data augmentation is based on writing the multinomial logit model as a random utility model (McFadden 1974):

$$y_{ki}^u = \mathbf{x}_i \boldsymbol{\beta}_k + \delta_{ki}, \quad k = 0, \ldots, m, \tag{14}$$

$$y_i = k \Leftrightarrow y_{ki}^u = \max_{l \in L} y_{li}^u. \tag{15}$$

Thus the observed category is equal to the category with maximal utility. If the random variables $\delta_{0i}, \ldots, \delta_{mi}$ appearing in (14) are i.i.d. following an EV distribution, then the multinomial logit model (13) results as the marginal distribution of y_i.

Frühwirth-Schnatter & Frühwirth (2007) and Scott (2009) use this RUM formulation of the multinomial logit model to carry out data augmentation based on introducing the latent utilities as missing data, i.e. $\mathbf{z} = ((y_{k1}^u, \ldots, y_{kN}^u), k = 1, \ldots, m)$. As for the binary RUM it is possible to sample the latent utilities $\mathbf{z}|\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \mathbf{y}$ simultaneously:

$$y_{ki}^u = -\log\left(-\frac{\log(U_i)}{1 + \sum_{l=1}^m \lambda_{li}} - \frac{\log(V_{ki})}{\lambda_{ki}} I\{y_i \neq k\}\right), \tag{16}$$

where U_i and $V_{1i}, \ldots, V_{mi}$ are $m+1$ independent uniform random numbers in $[0,1]$, and $\lambda_{li} = \exp(\mathbf{x}_i \boldsymbol{\beta}_l)$ for $l = 1, \ldots, m$.

3.2 Data Augmentation in the dRUM

An alternative way to write a multinomial model is as a difference random utility model (dRUM) which is obtained by choosing a baseline category k_0 and considering the model involving the differences of the utilities. This representation is the standard choice in the MCMC literature on the multinomial probit model, see e.g. McCulloch et al. (2000) and Imai & van Dyk (2005).

If we write the multinomial logit model as a dRUM, we obtain the following representation:

$$z_{ki} = \mathbf{x}_i \boldsymbol{\beta}_k + \varepsilon_{ki}, \qquad \varepsilon_{ki} \sim \mathrm{Lo}, \qquad k = 1, \ldots, m, \tag{17}$$

$$y_i = \begin{cases} 0, & \text{if } \max_{l \in L_{-0}} z_{li} < 0, \\ k > 0, & \text{if } z_{ki} = \max_{l \in L_{-0}} z_{li} > 0, \end{cases}$$

where $z_{ki} = y^u_{ki} - y^u_{0i}$ and $\varepsilon_{ki} = \delta_{ki} - \delta_{0i}$. The regression parameters appearing in (17) are identical to the ones appearing in (13), because $\boldsymbol{\beta}_{k_0} = \boldsymbol{\beta}_0 = \mathbf{0}$.

In contrast to the multinomial probit model, where $\boldsymbol{\varepsilon}_i = (\varepsilon_{1i}, \ldots, \varepsilon_{mi})'$ follows a multivariate normal distribution, the vector $\boldsymbol{\varepsilon}_i$ appearing in the dRUM representation of the multinomial logit model has a multivariate logistic distribution with logistic marginals (Balakrishnan 1992, Section 11.2). While the errors in the RUM representation (14) are i.i.d., the errors ε_{ki} in the dRUM representation (17) are no longer independent across categories.

This complicates MCMC sampling to a certain degree. Following the MCMC literature on the multinomial probit model, we could introduce $\mathbf{z} = ((z_{k1}, \ldots, z_{kN}), k = 1, \ldots, m)$ as missing data and sample $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m | \mathbf{z}$ and $\mathbf{z} | \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \mathbf{y}$. However, while sampling $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m | \mathbf{z}$ is trivial in the multinomial probit model because $\boldsymbol{\varepsilon}_i$ is multivariate normal, this step is non-standard in the multinomial logit model because $\boldsymbol{\varepsilon}_i$ is multivariate logistic.

In the present paper we consider a different way of representing a multinomial model by differences in utilities. Note that equation (15) may be written as

$$y_i = k \Leftrightarrow y^u_{ki} > y^u_{-k,i}, \qquad y^u_{-k,i} = \max_{l \in L_{-k}} y^u_{li}. \tag{18}$$

Thus category k is observed iff y^u_{ki} is bigger than the maximum of all other utilities. Now we define for each (fixed) value of $k \in L_{-0}$ the latent variables w_{ki} as the difference between y^u_{ki} and $y^u_{-k,i}$ and construct binary observations $d_{ki} = I\{y_i = k\}$. Then it is possible to rewrite (18) as a binary model in the dRUM representation:

$$w_{ki} = y^u_{ki} - y^u_{-k,i}, \qquad d_{ki} = I\{w_{ki} > 0\}. \tag{19}$$

We term (19) the partial dRUM representation, because d_{ki} uses only partial information from the original data, namely whether y_i is equal to k or not.

It should be mentioned that the partial dRUM representation is not restricted to the multinomial logit model, but holds for arbitrary error distributions in the RUM representation (14). However, while the distribution of w_{ki} is in general unfeasible, it has an explicit form for the multinomial logit model. First of all,

$$\exp\left(-y^u_{-k,i}\right) \sim \mathrm{Ex}\left(\lambda_{-k,i}\right), \qquad \lambda_{-k,i} = \sum_{l \in L_{-k}} \lambda_{li}, \tag{20}$$

because $\exp(-y^u_{-k,i}) = \min_{l \in L_{-k}} \exp(-y^u_{li})$, and $\exp(-y^u_{li}) \sim \mathrm{Ex}(\lambda_{li})$. We recall that $\lambda_{li} = \exp(\mathbf{x}_i \boldsymbol{\beta}_l)$. (20) may be rewritten as $y^u_{-k,i} = \log(\lambda_{-k,i}) + \delta_{-k,i}$, where $\delta_{-k,i}$ follows an EV distribution. Therefore

$$w_{ki} = y^u_{ki} - y^u_{-k,i} = \mathbf{x}_i \boldsymbol{\beta}_k - \log\left(\lambda_{-k,i}\right) + \delta_{ki} - \delta_{-k,i},$$

where $\delta_{-k,i}$ and $\delta_{k,i}$ are i.i.d. following an EV distribution. Thus the multinomial logit model has the following partial dRUM representation:

$$w_{ki} = \mathbf{x}_i \boldsymbol{\beta}_k - \log\left(\lambda_{-k,i}\right) + \varepsilon_{ki}, \qquad d_{ki} = I\{w_{ki} > 0\}, \tag{21}$$

where $\varepsilon_{ki} \sim \text{Lo}$. Evidently, for $m = 1$, (21) reduces to the dRUM given by (5).

The constant $\log(\lambda_{-k,i})$ appearing in (21) is independent of $\boldsymbol{\beta}_k$ and depends only on the regression parameters $\boldsymbol{\beta}_{-k}$ of the remaining categories. Thus given $\mathbf{z}_k = (w_{k1}, \ldots, w_{kN})$ and $\boldsymbol{\beta}_{-k}$, the regression parameter $\boldsymbol{\beta}_k$ corresponding to category k appears only in a linear regression model with logistic errors, given by (21).

Thus the partial dRUM is very useful when implementing MCMC for a multinomial model. At each MCMC draw we iterate over the categories for $k = 1, \ldots, m$. For each k, the partial dRUM actually is a binary dRUM and we may proceed as in Subsection 2.3 to sample $\mathbf{z}_k|\boldsymbol{\beta}_k, \mathbf{y}$ and $\boldsymbol{\beta}_k|\boldsymbol{\beta}_{-k}, \mathbf{z}_k$.

Evidently, $w_{ki}|\boldsymbol{\beta}_k, y_i$ is distributed according to a logistic distribution, truncated to $[0, \infty)$ if $y_i = k$, and truncated to $(-\infty, 0]$ otherwise. Thus w_{ki} is sampled as:

$$w_{ki} = \log\left(\lambda_{ki}^{\star} U_{ki} + I\{y_i = k\}\right) - \log\left(1 - U_{ki} + \lambda_{ki}^{\star}\, I\{y_i \neq k\}\right),$$

where $U_{ki} \sim \text{Un}[0,1]$ and $\lambda_{ki}^{\star} = \lambda_{ki}/\lambda_{-k,i}$.

Then $\boldsymbol{\beta}_k$ is sampled from the non-normal regression model (21), where the constant $\log(\lambda_{-k,i})$ is added to both sides of equation (21) to obtain a zero mean error. To deal with the non-normality of ε_{ki}, one can apply any of the sampling strategies discussed in Subsection 2.3 for the dRUM representation of the logit model.

Actually, Holmes & Held (2006) sample $\boldsymbol{\beta}_k$ for a multinomial logit model using the partial dRUM representation, but do not provide a rigorous derivation from a random utility model as we did above. They represent the logistic distribution of ε_{ki} in (21) as an infinite scale mixture of normals and introduce and sample scaling factors ω_{ki}, $i = 1, \ldots, N$, for all $k = 1, \ldots, m$. As for the logit model, this sampler is rather demanding from a computational point of view.

Alternatively, we can apply the ideas underlying Scott (2009) to the partial dRUM representation (21). This involves sampling $\boldsymbol{\beta}_k$ by an independence MH algorithm, where the proposal is constructed from regression model (21) by replacing the logistic error term ε_{ki} by a normal error with the same variance, i.e. $\pi^2/3$.

Finally, the finite scale mixture approximation of the logistic distribution introduced in Subsection 2.3 may be applied to (21). This involves introducing and sampling indicators r_{ki}, $i = 1, \ldots, N$, for all $k = 1, \ldots, m$. Because this sampling step can be implemented in a very efficient way, auxiliary mixture sampling in the partial dRUM representation turns out to be much more efficient than the related sampler of Holmes & Held (2006).

4 MCMC Sampling without Data Augmentation

It is generally believed that MCMC samplers based on data augmentation are less efficient than MCMC samplers without data augmentation. However, we will demonstrate in Section 5 that the new data augmentation samplers introduced in this paper are more efficient than commonly used MH algorithms.

For our comparison we consider the two MH algorithms suggested in Rossi et al. (2005) and the DAFE-R MH algorithm suggested by Scott (2009). We assume without loss of generality that the baseline is chosen equal to 0 and provide details for the multinomial model. We use $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ to denote the vector of all unknown regression parameters. The binary model results with $m = 1$.

Rossi et al. (2005, Section 3.11) discuss various MH algorithms based on the expected Hessian of the negative log-posterior $-\log p(\boldsymbol{\beta}|\mathbf{y})$. The elements of this matrix read:

$$-\mathrm{E}\left(\frac{\partial^2 \log p(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}_k^2}\right) = {\mathbf{B}_0}^{-1} + \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \pi_{ki}(\boldsymbol{\beta})(1 - \pi_{ki}(\boldsymbol{\beta})), \tag{22}$$

$$-\mathrm{E}\left(\frac{\partial^2 \log p(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\beta}_l}\right) = -\sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \pi_{ki}(\boldsymbol{\beta}) \pi_{li}(\boldsymbol{\beta}).$$

An alternative approach uses the expected Hessian of the negative log-likelihood $-\log p(\mathbf{y}|\boldsymbol{\beta})$; however, this matrix is rank deficient if for a certain category k, $\pi_{ki} = 0$ for all $i = 1, \ldots, N$. Thus, adding the prior information matrix ${\mathbf{B}_0}^{-1}$ in (22) helps to stabilize the inverse of the expected Hessian in cases where for a certain k the probabilities π_{ki} are equal or close to 0 for most of the observations.

To obtain a proposal variance-covariance matrix that is independent of $\boldsymbol{\beta}$, the probabilities $\pi_{ki}(\boldsymbol{\beta})$ are substituted by some estimator, for instance $\hat{\pi}_{ki} = \pi_{ki}(\hat{\boldsymbol{\beta}})$, with $\hat{\boldsymbol{\beta}}$ being the posterior mode. It is useful to write the expected Hessian matrix as:

$$\mathbf{H} = \mathbf{I}_m \otimes {\mathbf{B}_0}^{-1} + \sum_{i=1}^{N} (\mathrm{Diag}\,(\hat{\boldsymbol{\pi}}_i) - \hat{\boldsymbol{\pi}}_i \hat{\boldsymbol{\pi}}_i') \otimes \mathbf{x}_i' \mathbf{x}_i,$$

where $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{1i} \cdots \hat{\pi}_{mi})'$.

Rossi et al. (2005) construct two kinds of MH algorithms based on the matrix $\mathbf{H}$, namely an independence MH algorithm with a multivariate Student-t proposal $t_\nu(\hat{\boldsymbol{\beta}}, \mathbf{H}^{-1})$ with a small number of degrees of freedom ν, and a random walk MH algorithm with proposal $\boldsymbol{\beta}^{\mathrm{new}}|\boldsymbol{\beta}^{\mathrm{old}} \sim \mathrm{No}_{md}(\boldsymbol{\beta}^{\mathrm{old}}, s^2\mathbf{H}^{-1})$ with scaling factor s^2. Roberts & Rosenthal (2001) prove that for a (md)-variate normal posterior distribution with variance-covariance equal to the identity matrix an asymptotically optimal scaling is given by $s^2 = 2.38^2/(md)$, with the corresponding optimal acceptance rate being equal to 0.234. Since the posterior $p(\boldsymbol{\beta}|\mathbf{y})$ is asymptotically normal with variance-covariance matrix equal to $\mathbf{H}^{-1}$, we use the following random walk proposal for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{\mathrm{new}}|\boldsymbol{\beta}^{\mathrm{old}} \sim \mathrm{No}_{md}\left(\boldsymbol{\beta}^{\mathrm{old}}, \frac{2.38^2}{md}\mathbf{H}^{-1}\right). \tag{23}$$

Rossi et al. (2005, p.95) suggest to use the scaling factor $s^2 = 2.93^2/(md)$; however, it turns out that this scaling is inferior to the asymptotically optimal scaling.

Scott (2009) introduces the so-called DAFE-R MH algorithm which is based on computing the asymptotic variance-covariance matrix of the augmented posterior $p(\boldsymbol{\beta}|\mathbf{y},\mathbf{z})$ from the augmented random utility model (3). This variance-covariance matrix is used as a proposal in a multivariate random walk MH algorithm for the *marginal* model. For the binary model this proposal reads:

$$\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}} \sim \text{No}_d\left(\boldsymbol{\beta}^{\text{old}}, \left(\mathbf{B}_0{}^{-1} + \frac{6}{\pi^2}\mathbf{X}'\mathbf{X}\right)^{-1}\right). \tag{24}$$

The DAFE-R MH algorithm is applied to a multinomial logit model by using the proposal $\boldsymbol{\beta}_k^{\text{new}}|\boldsymbol{\beta}_k^{\text{old}} \sim \text{No}_d(\boldsymbol{\beta}_k^{\text{old}}, (\mathbf{B}_0{}^{-1} + 6/\pi^2\mathbf{X}'\mathbf{X})^{-1})$ for single-move sampling of $\boldsymbol{\beta}_k$ from $p(\boldsymbol{\beta}_k|\boldsymbol{\beta}_{-k},\mathbf{y})$.

The proposal used in the DAFE-R MH algorithm has the advantage that the variance-covariance matrix depends only on $\mathbf{X}$ and consequently is very easily computed prior to MCMC sampling, while determining the Hessian $\mathbf{H}$ requires estimators of all unknown probabilities π_{ki}. However, since the DAFE-R is a random walk MH algorithm, it is likely to be inferior to the asymptotically optimal random walk (23), which is confirmed by the case studies in Section 5.

For a binary model, for instance, the proposal of the asymptotically optimal random walk simplifies to:

$$\boldsymbol{\beta}^{\text{new}}|\boldsymbol{\beta}^{\text{old}} \sim \text{No}_d\left(\boldsymbol{\beta}^{\text{old}}, \left(\frac{d}{2.38^2}\mathbf{B}_0{}^{-1} + \mathbf{X}'\text{Diag}\,(a_1,\ldots,a_N)\,\mathbf{X}\right)^{-1}\right),$$

where $a_i = \hat{\pi}_i(1-\hat{\pi}_i)d/2.38^2$. This proposal looks rather similar to the DAFE-R proposal (24), the main difference being the weight attached to $\mathbf{x}_i{}'\mathbf{x}_i$, which is equal to $6/\pi^2 = 0.6079$ rather than a_i for the DAFE-R algorithm. Thus if, on average, $6/\pi^2 > a_i$, the scaling of the DAFE-R algorithm is too small, causing the acceptance rate to be too high. For instance, if $\hat{\pi}_i = 0.5$ this happens if $d < 14$, while for $\hat{\pi}_i = 0.1$ this happens if $d < 38$. Thus we expect that the acceptance rate of the DAFE-R algorithm is too high in small regression models.

5 Comparison of the Various MCMC Algorithms

We apply nine different MCMC samplers to five well-known data sets. The (binary) nodal involvement data (Chib 1995) is a small data set ($N = 53$) with a small set of regressors ($d = 5$). The (binary) heart data (Holmes & Held 2006) is a medium sized data set ($N = 270$) with a larger set of regressors ($d = 14$). The (binary) German credit card data (Holmes & Held 2006) is a large data set ($N = 1000$) with a large number of regressors ($d = 25$). The (multinomial) car data (Scott 2009) is a medium sized data set ($N = 263$) with 3 categories and a small set of regressors ($d = 4$).

Finally, we consider the (multinomial) Caesarean birth data of Fahrmeir & Tutz (2001, Table 1.1), where the outcome variable has 3 categories (no infection and two

type of infections) and $N = 251$. The data are organized as a three-way contingency table with eight factor combinations. The table is very unbalanced with a few cells containing a large fraction of the data, while other cells are empty. This makes statistical inference quite a challenge, and for illustration we fit a saturated logit model, i.e. $d = 8$.

For all examples, we take an independent standard normal prior for each regression coefficient and use each MCMC method to produce $M = 10000$ draws from the posterior distribution after running burn-in for 2000 iterations. All implementations are carried out using MATLAB (Version 7.3.0) on a notebook with a 2.0 GHz processor.

Naturally, we prefer fast samplers being nearly as efficient as i.i.d. sampling from the posterior $p(\boldsymbol{\beta}|\mathbf{y})$. Thus in Tables 5– 9 we summarize for each data set the performance of the various samplers in CPU time T_{CPU} (in seconds) needed to obtain the M draws (excluding burn-in) and the efficiency compared to i.i.d. sampling.

To evaluate the loss of efficiency, we compute for each regression coefficient β_{kj}, $k = 1,\ldots,m$, $j = 1,\ldots,d$ the inefficiency factor

$$\tau = 1 + 2 \cdot \sum_{h=1}^{K} \rho(h),$$

where $\rho(h)$ is the empirical autocorrelation of the MCMC draws of that particular regression parameter at lag h. The initial monotone sequence estimator of Geyer (1992) is used to determine K, based on the sum of adjacent pairs of empirical autocorrelations $\Phi(s) = \rho(2s) + \rho(2s+1)$. If n is the largest integer so that $\Phi(s) > 0$ and $\Phi(s)$ is monotone for $s = 1,\ldots,n$, then K is defined by $K = 2n+1$. We determine for each regression coefficient the effective sample size ESS (Kass et al. 1998) according to $\mathrm{ESS} = M/\tau$. The closer ESS is to M, the smaller is the loss of efficiency. In Tables 5– 9 we report the median ESS for all regression coefficients, as well as the minimum and the maximum.

To compare a slow, but efficient sampler with a fast, but inefficient sampler, we consider for each regression coefficient the effective sampling rate ESR (per second), defined as $\mathrm{ESR} = \mathrm{ESS}/T_{\mathrm{CPU}}$, and report the median ESR for all regression coefficients, as well as the minimum and the maximum. The median ESR is the most significant number in comparing the different MCMC samplers: the higher the median, the better the sampler.

We analyze three samplers using data augmentation in the dRUM, namely the sampler of Holmes & Held (2006) (dRUM-HH), our new auxiliary mixture sampler which substitutes the logistic distribution by the finite scale mixture approximation of Monahan & Stefanski (1992) with $H = 3$ and $H = 6$ (dRUM-FSF), and the new data augmented MH sampler which uses the posterior of the approximate standard linear regression model as proposal in the spirit of Scott (2009) (dRUM-Scott). We consider two samplers using data augmentation in the RUM, namely the auxiliary mixture sampler of Frühwirth-Schnatter & Frühwirth (2007) (RUM-FSF) and the original data augmented MH sampler of Scott (2009) (RUM-Scott). Finally, we consider the various random walk MH algorithms discussed in Section 4, namely

Table 5 Comparing MCMC samplers for the nodal involvement data ($N = 53, d = 5, m = 1$); based on $M = 10000$ draws after burn-in of 2000 draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		25.4	3459.0	3883.5	4948.7	136.2	152.9	194.8
dRUM-FSF ($H = 3$)		5.1	3616.2	4025.1	4162.7	707.8	787.8	814.8
dRUM-FSF ($H = 6$)		5.5	3862.3	3986.1	4298.3	708.3	731.0	788.2
dRUM-Scott	71.5	2.9	3035.6	3156.4	3229.4	1061.8	1104.0	1129.6
RUM-FSF		8.7	213.6	233.4	305.7	24.6	26.9	35.3
RUM-Scott	32.9	3.6	459.6	533.8	593.5	126.2	146.6	163.0
MH-Rossi	14.5	3.7	837.8	884.8	1042.5	225.3	237.9	280.3
MH-RR	29.8	3.1	552.5	652.9	754.6	181.3	214.3	247.7
MH-Scott	54.4	3.0	339.5	450.6	477.4	111.5	147.9	156.7

Table 6 Comparing MCMC samplers for the heart data ($N = 270$, $d = 14$, $m = 1$); based on $M = 10000$ draws after burn-in of 2000 draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		94.3	863.2	1379.6	6225.6	9.2	14.6	66.0
dRUM-FSF ($H = 3$)		12.1	808.8	1432.4	5569.0	66.7	118.1	459.3
dRUM-FSF ($H = 6$)		14.7	931.6	1432.0	6196.7	63.4	97.4	421.5
dRUM-Scott	43.7	6.2	446.0	778.6	2037.7	72.3	126.2	330.2
RUM-FSF		31.0	57.2	94.0	868.5	1.84	3.03	28.0
RUM-Scott	5.6	7.5	17.0	30.7	156.1	2.3	4.1	20.8
MH-Rossi	18.0	5.5	320.6	421.2	588.1	58.6	77.0	107.5
MH-RR	27.0	4.8	212.1	255.2	300.7	44.5	53.6	63.1
MH-Scott	43.1	4.6	129.8	194.9	500.5	28.1	42.2	108.2

the independence MH sampler of Rossi et al. (2005) (MH-Rossi), the asymptotically optimal random walk MH sampler of Roberts & Rosenthal (2001) (MH-RR), and the DAFE-R algorithm of Scott (2009) (MH-Scott).

We start the various MCMC samplers in the following way. All MH algorithms (with and without data augmentation) as well as all partial dRUM samplers for the multinomial logit model need a starting value for $\boldsymbol{\beta}_k$, $k = 1, \ldots, m$, which is set to $\mathbf{0}$. All data augmentation samplers need starting values for $\mathbf{z}$. For binary models starting values for $\mathbf{z}$ are sampled under the RUM representation from (6) and under the dRUM representation from (7) using $\lambda_i = \log \hat{\pi} - \log(1 - \hat{\pi})$, where $\hat{\pi} = \min(\max(\sum_{i=1}^{N} y_i/N, 0.05), 0.95)$. For multinomial models starting values for $\mathbf{z}$ are sampled from (16) with $\lambda_{0i} = 1$ and $\lambda_{li} = \log \hat{\pi}_l - \log(1 - \hat{\pi}_l)$, where $\hat{\pi}_l = \min(\max(\sum_{i=1}^{N} I\{y_i = l\}/N, 0.05), 0.95)$ for $l = 1, \ldots, m$. These values are transformed according to (19) to obtain starting values for w_{ki} in the partial dRUM representation. Finally, all elements of the latent scaling factors $\boldsymbol{\omega}$ are initialized with 1 for dRUM-HH, with $\pi^2/3$ for dRUM-FSF, and with $\pi^2/6$ for RUM-FSF.

Not surprisingly, we find for all data sets that MH sampling without data augmentation is faster than any data augmentation sampler in terms of CPU time T_{CPU}.

Table 7 Comparing MCMC samplers for the German credit card data ($N = 1000$, $d = 25$, $m = 1$); based on $M = 10000$ draws after burn-in of 2000 draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		333.4	1556.0	2325.7	3494.4	4.7	7.0	10.5
dRUM-FSF ($H = 3$)		41.8	1573.5	2313.5	3780.1	37.7	55.4	90.4
dRUM-FSF ($H = 6$)		61.5	1666.9	2268.3	3872.2	27.1	36.9	63.0
dRUM-Scott	30.4	21.6	592.6	824.4	1090.8	27.4	38.2	50.5
RUM-FSF		134.2	91.5	133.7	261.5	0.68	1.00	1.95
RUM-Scott	0.8	25.0	9.7	11.8	26.1	0.39	0.47	1.0
MH-Rossi	7.1	11.2	117.3	178.5	290.9	10.4	15.9	25.9
MH-RR	25.0	11.1	92.5	138.2	188.0	8.3	12.5	17.0
MH-Scott	22.0	10.4	103.9	149.8	189.2	10.0	14.4	18.1

Table 8 Comparing MCMC samplers for the car data ($m = 2$, $N = 263$, $d = 3$); based on $M = 10000$ draws after burn-in of 2000 draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		182.5	1716.8	2558.2	3020.5	9.4	14.0	16.6
dRUM-FSF ($H = 3$)		20.9	1831.7	2535.8	3200.5	87.6	121.2	153.0
dRUM-FSF ($H = 6$)		27.20	1570.9	2307.6	2942.4	57.8	84.8	108.2
dRUM-Scott	70.2	13.0	1468.3	2101.1	2662.8	113.4	162.2	205.6
RUM-FSF		46.6	111.3	171.8	253.2	2.4	3.7	5.4
RUM-Scott	33.8	9.5	289.8	388.7	472.8	30.6	41.1	49.9
MH-Rossi	57.4	6.0	3158.0	3323.7	3899.0	526.3	554.0	649.8
MH-RR	27.2	5.3	366.2	397.2	499.1	69.3	75.2	94.5
MH-Scott	61.7	10.2	269.4	365.1	527.0	26.5	35.8	51.7

To evaluate any MH sampler (with or without data augmentation) we report additionally the acceptance rate a, which is averaged over the categories for MH-Scott for multinomial models. For both random walk MH samplers a should be close to the asymptotically optimal rate of 0.234, which is actually the case for MH-RR with the exception of the Caesarean birth data in Table 9. The acceptance rate of MH-Scott deviates from the asymptotically optimal rate for all examples but the German credit card data in Table 7, which causes the effective sample size and the effective sampling rate to be smaller than for MH-RR.

With the exception of the Caesarean birth data, MH-Rossi outperforms the other MH samplers without data augmentation in terms of effective sample size and effective sampling rate. This is true even for the German credit data where the acceptance rate is as low as 7.1%. In general, the acceptance rate of MH-Rossi varies considerably across the various case studies, being pretty high for the car data in Table 8 and being extremely small for the Caesarean birth data in Table 9.

For the Caesarean birth data the Hessian matrix is very ill-conditioned due to the unbalanced data structure mentioned earlier, leading to a very low acceptance rate both for MH-Rossi and MH-RR. For this particular data set MH-Scott outperforms

Table 9 Comparing MCMC samplers for the Caesarean birth data ($m = 2$, $N = 251$, $d = 8$); based on $M = 10\,000$ draws after burn-in of 2 000 draws.

	a	T_{CPU}	ESS (total draws)			ESR (draws/s)		
Sampler	(%)	(s)	min	med	max	min	med	max
dRUM-HH		177.8	1153.4	2643.1	4553.1	6.5	14.9	25.6
dRUM-FSF ($H = 3$)		21.0	1195.2	2587.8	4621.8	56.8	123.0	219.8
dRUM-FSF ($H = 6$)		26.4	1125.5	2777.4	4765.0	42.6	105.1	180.2
dRUM-Scott	63.9	12.3	714.9	1790.8	3084.8	58.4	146.2	251.8
RUM-FSF		42.1	148.6	344.1	899.6	3.5	8.2	21.4
RUM-Scott	23.4	10.2	213.5	389.9	729.2	21.0	38.3	71.6
MH-Rossi	2.0	5.7	37.0	89.1	120.0	6.5	15.5	20.9
MH-RR	3.9	4.9	22.8	50.0	83.5	4.7	10.3	17.1
MH-Scott	39.8	9.7	254.7	354.1	486.7	26.3	36.6	50.3

the other MH samplers, because it avoids the Hessian when constructing the variance-covariance matrix of the proposal density.

When comparing the various data augmentation samplers in the RUM and in the dRUM representation, we find for all case studies that both the effective sample size and the effective sampling rate are considerably higher for the dRUM representation than for the RUM representation, leading to the conclusion that data augmentation in the RUM should be avoided.

Among the data augmentation samplers in the dRUM representation, data augmented MH based on the approximate normal proposal (dRUM-Scott) is the fastest. As expected, the acceptance rate a, which should be as high as possible, is considerably larger for dRUM-Scott than under the RUM representation (RUM-Scott), because the logistic distribution underlying the dRUM is much closer to a normal distribution than the extreme value distribution underling the RUM. For the German credit data in Table 7, for instance, the acceptance rate increases from 0.8% for RUM-Scott to 30.4% for dRUM-Scott.

Compared to dRUM-Scott, the other two dRUM data augmentation samplers are slower, because both dRUM-HH and dRUM-FSF introduce the latent scaling factors $\boldsymbol{\omega}$ as a second set of auxiliary variables. We find that dRUM-HH requires much more computation time than dRUM-FSF, even if the latter uses the very accurate mixture approximation with six components, while the efficiency in terms of effective sample size is more or less the same. This makes our new dRUM auxiliary mixture sampler much more efficient in terms of effective sampling rate than the sampler of Holmes & Held (2006).

Interestingly the effective sample size of dRUM-HH and dRUM-FSF is larger than dRUM-Scott. Introducing the latent scaling factors $\boldsymbol{\omega}$ allows dRUM-HH and dRUM-FSF to accept $\boldsymbol{\beta}$ at each sweep of the MCMC sampler, because a conditional Gibbs step is implemented. In contrast to that dRUM-Scott uses an MH update for $\boldsymbol{\beta}$, meaning that the sampler is stuck at the current value with probability $1 - a$, which increases the autocorrelation in the MCMC sample.

When we compare our new dRUM data augmentation samplers, we find that they outperform any other data augmentation sampler in terms of the effective sampling rate. With the exception of the car data in Table 8, the samplers even outperform the independence MH sampler of Rossi et al. (2005). The relatively high acceptance rate of MH-Rossi for the car data explains its superiority for this particular example.

Finally, we discuss the performance of our new samplers in relation to each other. While dRUM-Scott is faster, dRUM-FSF has a higher effective sample size. The effective sampling rate is higher for dRUM-Scott with the exception of the German credit card data in Table 7, where the acceptance rate of dRUM-Scott is smaller than in the other examples. It appears from the various tables that an acceptance rate of dRUM-Scott above 40% makes the sampler more efficient in terms of the effective sampling rate than dRUM-FSF.

Because the coding of dRUM-Scott is extremely simple, we recommend to make this new data augmented MH sampler the first choice. However, while there is no tuning in the proposal of dRUM-Scott — which makes it easy to implement — there is, on the other hand, no control over the acceptance rate. Thus the acceptance rate may be arbitrarily small, depending on the particular application. Thus, if the acceptance rate turns out to be considerably smaller than say 40%, it is to be expected that dRUM-FSF is more efficient and should be the method of choice.

6 Concluding Remarks

In this paper we have introduced yet two other data augmentation algorithms for sampling the parameters of a binary or a multinomial logit model from their posterior distribution within a Bayesian framework. They are based on rewriting the underlying random utility model in such a way that only differences of utilities appear in the model. Applications to five case studies reveal that these samplers are superior to other data augmentation samplers and to Metropolis–Hastings sampling without data augmentation.

We have confined our investigations to the standard binary and multinomial logit regression model; however, we are confident that our new samplers will be of use for the MCMC estimation of more general latent variable models such as analyzing discrete-valued panel data using random-effects models, or analyzing discrete-valued time series using state space models. For latent variable models, auxiliary mixture sampling in the dRUM representation is of particular relevance, because introducing the auxiliary latent variables $\mathbf{z}$ and $\boldsymbol{\omega}$ leads to a conditionally Gaussian model, which allows efficient sampling of the random effects or the state vector.

Furthermore, dRUM auxiliary mixture sampling could be useful for Bayesian variable selection in binary data analysis simply by replacing less efficient samplers such as the Holmes & Held (2006) sampler, which was used in the same paper for variable selection in logistic regression models, and the RUM auxiliary mixture sampling, which was used in Tüchler (2008) for covariance selection in panel data models with random effects. Furthermore, it could be applied to the stochastic variable se-

lection approach of Frühwirth-Schnatter & Wagner (2009) for state space modelling of binary time series.

It remains an open issue whether representations comparable to the dRUM exist for more general discrete-valued distributions. Frühwirth-Schnatter et al. (2009) improve auxiliary mixture sampling for data from a binomial or multinomial distribution by using an aggregated RUM representation instead of the RUM representation of the underlying individual binary experiments. It seems worth investigating whether auxiliary mixture sampling for such data can be improved further using an aggregated version of the dRUM representation; however, we leave this issue for further research.

Acknowledgements The first author's research is supported by the Austrian Science Foundation (FWF) under the grant S 10309-G14 (NRN "The Austrian Center for Labor Economics and the Analysis of the Welfare State", Subproject "Bayesian Econometrics").

References

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**: 669–679.

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions, *Journal of the Royal Statistical Society, Ser. B* **36**: 99–102.

Balakrishnan, N. (ed.) (1992). *Handbook of the Logistic Distribution*, Marcel Dekker, New York.

Chib, S. (1995). Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association* **90**: 1313–1321.

Dellaportas, P. & Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling, *Applied Statistics* **42**: 443–459.

Fahrmeir, L. & Kaufmann, H. (1986a). Asymptotic inference in discrete response models, *Statistical Papers* **27**: 179–205.

Fahrmeir, L. & Kaufmann, H. (1986b). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, *The Annals of Statistics* **13**: 342–368.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer Series in Statistics, 2nd ed., Springer, New York/Berlin/Heidelberg.

Frühwirth-Schnatter, S. & Frühwirth, R. (2007). Auxiliary mixture sampling with applications to logistic models, *Computational Statistics and Data Analysis* **51**: 3509–3528.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L. & Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data, *Statistics and Computing* **19**, forthcoming.

Frühwirth-Schnatter, S. & Wagner, H. (2009). Stochastic model specification search for Gaussian and partially non-Gaussian state space models, *Journal of Econometrics*, forthcoming.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models, *Statistics and Computing* **7**: 57–68.

Geyer, C. (1992). Practical Markov chain Monte Carlo, *Statistical Science* **7**: 473–511.

Holmes, C. C. & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression, *Bayesian Analysis* **1**: 145–168.

Imai, K. & van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation, *Journal of Econometrics* **124**: 311–334.

Kass, R. E., Carlin, B., Gelman, A. & Neal, R. (1998). Markov chain Monte Carlo in practice: A roundtable discussion, *The American Statistician* **52**: 93–100.

Lenk, P. J. & DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects, *Psychometrika* **65**: 93–119.

McCulloch, R. E., Polson, N. G. & Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters, *Journal of Econometrics* **99**: 173–193.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour, *in* P. Zarembka (ed.), *Frontiers of Econometrics*, Academic, New York, pp. 105–142.

Monahan, J. F. & Stefanski, L. A. (1992). Normal scale mixture approximations to $F^\star(z)$ and computation of the logistics normal integral, *in* N. Balakrishnan (ed.), *Handbook of the Logistic Distribution*, Marcel Dekker, New York, pp. 529–549.

Roberts, G. O. & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms, *Statistical Science* **16**: 351–367.

Rossi, P. E., Allenby, G. M. & McCulloch, R. (2005). *Bayesian Statistics and Marketing*, Wiley, Chichester.

Scott, S. L. (2009). Data augmentation and the Bayesian analysis of multinomial logit models, *Statistical Papers*, forthcoming.

Tüchler, R. (2008). Bayesian variable selection for logistic models using auxiliary mixture sampling, *Journal of Computational and Graphical Statistics* **17**: 76–94.

Zeger, S. L. & Karim, M. (1991). Generalized linear models with random effects: a Gibbs sampling approach, *Journal of the American Statistical Association* **86**: 79–86.

Zellner, A. & Rossi, P. E. (1984). Bayesian analysis of dichotomous quantal response models, *Journal of Econometrics* **25**: 365–393.

Generalized Semiparametric Regression with Covariates Measured with Error

Thomas Kneib, Andreas Brezger and Ciprian M. Crainiceanu

Abstract We develop generalized semiparametric regression models for exponential family and hazard regression where multiple covariates are measured with error and the functional form of their effects remains unspecified. The main building blocks in our approach are Bayesian penalized splines and Markov chain Monte Carlo simulation techniques. These enable a modular and numerically efficient implementation of Bayesian measurement error correction based on the imputation of true, unobserved covariate values. We investigate the performance of the proposed correction in simulations and an epidemiological study where the duration time to detection of heart failure is related to kidney function and systolic blood pressure.

Key words: additive hazard regression; generalized additive models; MCMC; measurement error correction; penalized splines

1 Introduction

The presence of covariates measured with error in regression models can have severe impact on inferential conclusions drawn from naive estimates. This is particularly true for semiparametric regression models where the relation between responses and covariates is specified flexibly and therefore also more prone to disturbances induced by measurement error. A common phenomenon in naive analyses are estimates that

Thomas Kneib
Institut für Mathematik, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany, e-mail: thomas.kneib@uni-oldenburg.de

Andreas Brezger
HypoVereinsbank Munich, e-mail: andreas.brezger@hvb.de

Ciprian M. Crainiceanu
Department of Biostatistics, Johns-Hopkins-University Baltimore, e-mail: ccrainic@jhsph.edu

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_8,

are biased towards zero and therefore underestimate effects. In particular, in semiparametric regression models it will be more difficult to detect local extrema of a functional relationship and curvature will be underestimated. In general, the effect of measurement error is insidious and leads to biased estimates, misspecified variability and feature masking (Carroll et al. 2006). Hence, it is likely to erroneously conclude that covariates are not associated with the response variable or to obtain false conclusions about the precise functional form of relationships.

Based on work by Berry et al. (2002) for Gaussian scatterplot smoothing, we develop a flexible Bayesian correction procedure based on Markov chain Monte Carlo (MCMC) simulations for general semiparametric exponential family and hazard regression models. The key ingredient is the imputation of the unobserved, true covariate values in an additional sampling step, an idea dating back to Stephens & Dellaportas (1992) and Richardson & Gilks (1993), see also Gustafson (2004). The Bayesian approach considered in this paper combines a number of distinct advantages:

Flexibility in terms of the response type: A wide range of response types is supported, including exponential family regression (e.g. Binomial or Poisson responses) as well as right-censored continuous-time survival times. This is made possibly by the consideration of a iteratively weighted least squares proposals for the regression coefficients (Gamerman 1997, Brezger & Lang 2006), a proposal scheme that relies on Gaussian approximations of the full conditionals.

Flexibility in terms of the model equation: All nonparametric model components are specified flexibly in terms of Bayesian penalized splines (Brezger & Lang 2006, Jullion & Lambert 2007). The modular structure of Bayesian computations based on MCMC enables the consideration of models where several covariates are measured with error in combination with further nonparametric effects of covariates observed exactly. Spatial effects, varying coefficient terms, or random effects are readily available as additional model components and are also included in our software.

Flexibility in terms of the measurement error equation: Based on the classical model of uncorrelated additive Gaussian measurement error, longitudinally correlated repeated observations on the measurement error equation or other extended measurement error models could easily be included.

Numerically efficient implementation: Sparse matrix computations and efficient storage schemes in combination with data compression based on rounding provide a rather fast estimation procedure. This, in particular, allows us to consider more complex applications with large sample size and extensive simulation setups.

In the application that motivated our research, the duration until detection of heart failure is analyzed in a hazard regression model that includes nonlinear effects of kidney function measured by the glomerular filtration rate (GFR) and systolic pressure (SP). Both covariates are inherently subject to measurement error due to different reasons: while SP is measured with error due to the relatively imprecise instruments involved in standard hemodynamometry, GFR can only be obtained accurately based on a time-consuming, awkward procedure; thus, in practice, this procedure is replaced with an estimate (eGFR) predicted from creatinine, gender and age (Hsu et al. 2005).

The prediction equation has been derived from a regression model and an estimate of the measurement error variance is also available from a replication study. In case of SP, the measurement error variance is available from previous studies. The sample size of 15,000 observations and two covariates measured with error make this application challenging, since we are faced with the imputation of 30,000 true covariate values and the re-evaluation of the corresponding parts of the design matrix in each iteration.

2 Semiparametric Regression Models with Measurement Error

2.1 Observation Model

In a generalized semiparametric regression models (see for example Ruppert et al. (2003), Fahrmeir et al. (2004) or Wood (2006)), the expectation of (conditionally) independent responses y_i, $i,\ldots,n$, from univariate exponential families is related to an additive predictor

$$\eta_i = f_1(x_{i1}) + \ldots + f_r(x_{ir}) + v_i'\gamma \tag{1}$$

based on a response function h, i.e. $\mu_i = E(y_i|\eta_i) = h(\eta_i)$. The predictor is additively composed of smooth functions $f_1,\ldots,f_r$ of continuous covariates $x_1,\ldots,x_r$ in combination with parametric effects γ of further, typically categorical covariates v. For hazard regression models employed in survival analysis, data are given in the form of (conditionally) independent survival data (t_i,δ_i), $i=1,\ldots,n$ where t_i is the (right-censored) observed survival time and δ_i is the censoring indicator. Extending the classical Cox model, semiparametric hazard regression models (Hennerfeind et al. 2006, Kneib & Fahrmeir 2007) can then be specified as $\lambda_i(t) = \exp(\eta_i(t))$ where

$$\eta_i(t) = g_0(t) + f_1(x_{i1}) + \ldots + f_r(x_{ir}) + v_i'\gamma$$

is a semiparametric predictor consisting of the log-baseline hazard rate $g_0(t)$, r smooth functions of continuous covariates, and linear effects summarized in $v'\gamma$. The time-dependent function $g_0(t)$ relates to the baseline hazard rate $\lambda_0(t)$ in the Cox model via $\lambda_0(t) = \exp(g_0(t))$. In contrast to usual partial likelihood estimation, determination of the baseline hazard rate will be an integral part of model estimation in our framework. In particular, estimation will be based on the full instead of the partial likelihood.

Estimation of the nonlinear functions $f_j(x_j)$ is frequently complicated by the fact that in applications the corresponding covariates x_j are not observed exactly so that only contaminated surrogate variables are available. Naive estimates based on these surrogate variables will then be oversmoothed leading to estimates that are biased towards "no effect" models. In the following, we assume that the first r_1 covariates $x_1,\ldots,x_{r_1}$ are subject to measurement error while the remaining $r_2 = r - r_1$ covariates

$x_{r_1+1},\ldots,x_r$ are observed exactly. In particular, we allow for several covariates x_j measured with error.

2.2 Measurement Error Model

In the classical measurement error model (Carroll et al. 2006), the true measurements of the covariates are contaminated by i.i.d. Gaussian noise, leading to the measurement of proxy variables

$$w_{ij}^{(m)} = x_{ij} + u_{ij}^{(m)}, \quad m = 1,\ldots,M$$

where $u_{ij}^{(m)} \sim \mathrm{N}(0,\tau_{u,j}^2)$. In our modeling framework, we allow for the possibility of repeated measurements (indexed by $m = 1,\ldots,M$) on a covariate. For simplicity, we assume that the measurement error contaminations are mean zero and independent, i.e. $u_{ij} = (u_{ij}^{(1)},\ldots,u_{ij}^{(M)})' \sim \mathrm{N}(\mathbf{0},\tau_{u,j}^2 I_M)$. However, the MCMC sampling mechanism presented in Section 3 can straightforwardly be extended to more general situations where $u_{ij} \sim \mathrm{N}(\mu,\Sigma)$. Inclusion of covariances in Σ could for example be useful in combination with a longitudinal collection of the repeated measurements where Σ contains an equicorrelation or autoregressive correlation structure (see Wang & Pepe (2000) for such an example). Non-zero expectations μ can, for example, be employed to adjust for measurement bias in the repeated observations.

2.3 Prior Distributions

To complete the Bayesian specification, suitable priors have to be assigned to all model parameters. In the Bayesian perspective on the model, the unknown true covariate values x_{ij} are treated as additional unknowns and imputation becomes a part of the MCMC algorithm. For the fixed effects γ, we assume standard noninformative priors, i.e. $p(\gamma) \propto \text{const}$. In contrast, we assign informative priors to the smooth function to enforce smoothness of the corresponding estimates.

2.3.1 P-spline Priors

A parsimonious yet flexible modelling possibility for nonparametric function estimation are penalized splines as popularized by Eilers & Marx (1996) and extensively discussed in Ruppert et al. (2003). In our Bayesian framework, we employ the Bayesian analogue developed by Brezger & Lang (2006). For the sake of simplicity, we drop the function index j in the following description. To represent $f(x)$ (or $g_0(t)$ in case of hazard regression models) in terms of a flexible but finite dimensional class of functions, we assume that it can be expanded in B-splines of leading to the basis

function representation

$$f(x) = \sum_{k=1}^{K} \beta_k B_k^l(x)$$

where $B_k^l(x)$ are B-spline basis functions of degree l defined upon a set of knots $\kappa_1 < \ldots < \kappa_K$, and β_k are the corresponding regression coefficients. In the classical frequentist formulation of P-splines, smoothness of the functions $f(x)$ is enforced by adding a squared difference penalty of order d to the likelihood that essentially penalizes large variation in terms of the d-th derivative. In a Bayesian formulation, d-th order differences are replaced by d-th order random walks, e.g.

$$\beta_k - \beta_{k-1} \sim \mathrm{N}(0, \tau_\beta^2 \omega_k)$$

for first order random walks in the most simple case. This prior specification corresponds to local increments in the coefficient sequence with expectation zero and deviations controlled by the variance τ_β^2 and the distance between the corresponding knots $\omega_k = \kappa_k - \kappa_{k-1}$. The underlying rationale of the latter choice is that larger steps between two knots should also be reflected in the prior in allowing for larger variation. In contrast, the variance parameter τ_β^2 controls the overall variability of the function estimate with small values corresponding to very flat estimates whereas large values yield very flexible estimates. The weighted first order random walk can also be interpreted as a discrete approximation to continuous Brownian motion that yields a similar structure of the variance. Weighted second order random walks are also available (see Fahrmeir & Lang (2001)) but are less suitable in the context of measurement error correction since they enforce too smooth function estimates.

In combination with flat priors on the initial parameters, the joint distribution of the vector of regression coefficients $\beta = (\beta_1, \ldots, \beta_K)'$ can be deduced from the random walk specifications as the multivariate Gaussian distribution

$$p(\beta|\tau_\beta^2) \propto \exp\left(-\frac{1}{2\tau_\beta^2}\beta' K \beta\right).$$

The precision matrix K is also derived from the univariate random walk priors. For a first order random walk it can be represented as $K = D'\Omega D$, where D is a first order difference matrix and $\Omega = \mathrm{diag}(\omega_2, \ldots, \omega_K)$ contains the knot distances as weights.

In case of smoothing without measurement error, cubic P-splines (i.e. splines of degree $l = 3$) with approximately 20 equidistant knots and second order random walk prior have proven to be a useful standard choice (Brezger & Lang 2006). However, exploratory simulations showed that this claim no longer holds when measurement error is present. In particular, the high degree of the spline basis and the second order random walk enforce smoothness of the function estimates. Since measurement error in general leads to an attenuation of functional relationships, i.e. functions appear smoother than under the true relationship, a suitable prior in measurement error correction has to allow for more flexibility. In addition, choosing equidistant knots has the disadvantage that the prior variance of the random walk remains constant over

the whole domain of the covariate. When correcting for measurement error, adaptive priors with more variability in areas where a lot of observations have been collected mostly showed a better performance. In summary, we found linear splines with 20 quantile-based knots and (weighted) first order random walk prior to be a suitable default choice for nonparametric smoothing of covariates with measurement error. Cheng & Crainiceanu (2009) also support the choice of linear splines in showing that the full conditionals both for the regression coefficients and the true covariate values are then log-concave.

On a further stage of the hierarchy, a hyperprior is assigned to the variance parameter τ^2_β to allow for a data-driven amount of smoothness. Since the random walk prior is multivariate Gaussian, a computationally attractive choice is the conjugate inverse gamma prior $\tau^2_\beta \sim \text{IG}(a,b)$ that leads to a simple Gibbs update for the variance parameter.

A further generalisation of the model can be achieved by allowing for prior uncertainty in the knot positions as in the adaptive spline smoothing approaches by Denison et al. (1998) or Biller (2000). However, in most situations it will be sufficient to either assign a smoothness prior to the regression coefficients (provided that the basis is sufficiently rich) or to allow for data-driven determination of the knot placements (see also the supporting simulation results in Brezger & Lang (2006)). In combination with measurement error correction we found it advantageous to fix the knot positions since this avoids additional re-evaluations when imputing the unobserved covariate values.

2.3.2 Measurement Error Priors

For the covariates with measurement error, a prior for the true covariate values has to be specified, since they will be treated as additional unknowns in the Bayesian inferential procedure. A flexible default choice is given by the Gaussian distribution

$$x_{ij} \sim \text{N}(\mu_{x,j}, \tau^2_{x,j}).$$

Assigning hyperpriors to the parameters such as $\mu_{x,j} \sim \text{N}(0, \tau^2_\mu)$ with τ^2_μ fixed at a large value and $\tau^2_{x,j} \sim \text{IG}(a,b)$ allows the prior to accommodate to a variety of data-generating processes. In particular, the prior for the expectation is essentially noninformative when assuming a large value for the hypervariance τ^2_μ.

Note that treating the true covariate values as unknown parameters is not only a computational trick to obtain a fully specified model within the MCMC sampler, but allows inferences to be drawn about the true covariate values. In particular, we obtain a sample from the posterior of the true covariate values allowing to investigate for example the precision of the correction or whether the true covariate value exceeds a certain threshold.

Finally, a prior may be assigned to the measurement error variances, if uncertainty about the $\tau^2_{u,j}$ has to be incorporated. In combination with the Gaussian contamination error, again an inverse gamma prior $\tau^2_{u,j} \sim \text{IG}(a,b)$ is a suitable default choice.

Note, however, that reliable estimation of $\tau^2_{u,j}$ will typically require a larger number of repeated measurements on the covariates, in particular in non-Gaussian observation models where the likelihood carries less information on the variability in measurement error than in Gaussian models. In our application, the measurement error variances are available from replication experiments. Therefore we will also restrict our attention to the case of known measurement error variances in our simulations.

3 Bayesian Inference

3.1 Posterior & Full Conditionals

Summarizing all unknown quantities in the vector θ and assuming conditional independence of the prior distributions, the joint posterior in our class of semiparametric models can be summarized as

$$\begin{aligned}
p(\theta|\text{data}) \propto\ & p(\text{data}|\beta_1,\ldots,\beta_r,\gamma,x_1,\ldots,x_r) && \text{observation model likelihood}\\
& \prod_{j=1}^{r_1} p(w_j|x_j,\tau^2_{u,j}) && \text{measurement error likelihood}\\
& \prod_{j=1}^{r_1} p(\tau^2_{u,j}) && \text{measurement error variance priors}\\
& \prod_{j=1}^{r_1} p(x_j|\mu_{x,j},\tau^2_{x,j})p(\mu_{x,j})p(\tau^2_{x,j}) && \text{true covariate value priors}\\
& \prod_{j=1}^{r} p(\beta_j|\tau^2_{\beta,j})p(\tau^2_{\beta,j}) && \text{nonparametric effect priors.}
\end{aligned}$$

The likelihood is derived under the assumption of conditional independence such that the complete data likelihood factorises to individual likelihood contributions. In case of exponential family regression, the likelihood contributions equal the corresponding exponential family densities evaluated at the predictor η_i. Assuming non-informative, random right censored survival times, the complete data likelihood contributions in hazard regression models with individual hazard rates $\lambda_i(t)$ are given by

$$L_i(\eta_i) = \lambda_i(t_i)^{\delta_i} \exp\left(-\int_0^{t_i} \lambda_i(u)du\right),$$

see Hennerfeind et al. (2006).

From the posterior, we can now derive the full conditional distributions for all unknowns to construct a Markov Chain Monte Carlo simulation algorithm. While Gibbs updates can be derived for several parameters, Metropolis-Hastings steps are necessary for the regression coefficients and the true covariate values. Since some of the priors involved in the model specification are (partially) improper, it is not obvious that the joint posterior will be proper (see for example Hobert & Casella (1996)). Fahrmeir & Kneib (2009) provide conditions for the propriety of the posterior in

semiparametric Bayesian regression models without measurement error that will be fulfilled in most practically situations and we expect these results to carry over to the case with measurement error.

The full conditional for a true covariate value x_{ij} depends only on the i-th likelihood contribution $L_i(\eta_i)$ to the observation model and the i-th contribution to the measurement error model. Combining this likelihood information with the relevant priors yields (up to an additive constant) the log-full conditional

$$\log(p(x_{ij}|\cdot)) = l_i(\eta_i) - \frac{1}{2\tau^2_{u,j}} \sum_{m=1}^{M} (w_{ij}^{(m)} - x_{ij})^2 - \frac{1}{2\tau^2_{x,j}} (x_{ij} - \mu_{x,j})^2$$

where $l_i(\eta_i) = \log(L_i(\eta_i))$ is the i-th log-likelihood contribution. Obviously this full conditional does not correspond to a known distribution since both the log-likelihood contributions and the B-spline basis functions are non-linear in the covariate values. Following Berry et al. (2002) we consider a random walk proposal for imputing the covariate values where, based on the current value x_{ij}^{curr}, a new value is proposed as

$$x_{ij}^{prop} = x_{ij}^{curr} + \varepsilon, \qquad \varepsilon \sim \mathrm{N}\left(0, \frac{4\tau^2_{u,j}}{M}\right).$$

The choice of the random walk variance as being proportional to the measurement error variance but inverse proportional to the number of replicated measurements balances between the more precise knowledge about the true value that can be gathered from repeated measurements on the one hand and uncertainty introduced by large measurement error variance. The constant factor 4 has proven to work well in practice, according to our experience, but can be adjusted by the user to adapt the acceptance probabilities if needed.

The Gaussian measurement error model in combination with the conjugate inverse gamma priors for the measurement error variances, yields full conditionals that are also inverse gamma, i.e.

$$\tau^2_{u,j}|\cdot \sim \mathrm{IG}\left(a + \frac{nM}{2}, b + \frac{1}{2}\sum_{i=1}^{n}\sum_{m=1}^{M}(x_{ij}^{(m)} - x_{ij})^2\right).$$

Similarly, we obtain closed form full conditionals for the true covariate value hyperparameters:

$$\mu_{x,j}|\cdot \sim \mathrm{N}\left(\frac{n\bar{x}_j\tau^2_\mu}{n\tau^2_\mu + \tau^2_{x,j}}, \frac{\tau^2_{x,j}\tau^2_\mu}{n\tau^2_\mu + \tau^2_{x,j}}\right)$$

$$\tau^2_{x,j}|\cdot \sim \mathrm{IG}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(x_{ij} - \mu_{x,j})^2\right)$$

where $\bar{x}_j$ is the empirical mean of the currently imputed true covariate values.

Finally, the full conditionals for the regression coefficients have to be derived. Again, these are not available in closed form since the likelihood is non-linear in the parameters (for non-Gaussian responses). Based on work by Gamerman (1997) in the context of random effects, Brezger & Lang (2006) propose to construct a Gaussian approximation to the full conditional by performing one-step of a Fisher scoring algorithm based on the current sample for β_j. More precisely, this leads to an iteratively weighted least squares (IWLS) proposal for β_j based on a Gaussian distribution with precision matrix and mean

$$P_j = X_j' W X_j + \frac{1}{\tau^2_{\beta,j}} K_j \qquad \text{and} \qquad m_j = P_j^{-1} X_j' W (\tilde{y} - \eta_{-j}),$$

where the diagonal matrix W and the vector of working observations $\tilde{y}$ are constructed in complete analogy to the usual GLM case (compare Fahrmeir & Tutz (2001)) and $\eta_{-j} = \eta - X_j \beta_j$ is the j-th partial residual. Similar expressions are obtained for the vector of fixed effects, compare Brezger & Lang (2006) for details. The rationale for the IWLS proposal mechanism is that it automatically adapts to the location and the curvature of the corresponding full conditional thereby avoiding the necessity of manually tuning the MCMC sampler. Hennerfeind et al. (2006) describe similar proposal schemes for hazard regression models. The full conditional of the smoothing parameters $\tau^2_{\beta,j}$ is again inverse Gamma with updated parameters, i.e.

$$\tau^2_{\beta,j} | \cdot \sim \text{IG}\left(a + \frac{1}{2}\text{rank}(K_j), b + \frac{1}{2}\beta_j' K_j \beta_j\right).$$

3.2 Implementational Details & Software

Though a Metropolis-Hastings sampler can immediately be set up based on the full conditional distributions and proposals described in the previous section, an efficient implementation requires careful fine-tuning at several places. This is particularly the case for nonparametric function estimation involving a large number of regression coefficients and the measurement error correction problem, where the data, and therefore also the design matrices, change in each iteration. A naive implementation in a general specification language for Bayesian modeling such as WinBUGS or in a high-level interpreted programming language such as R would therefore be inefficient. As a consequence, we implemented our methodology as a part of the software package BayesX (`http://www.stat.uni-muenchen.de/~bayesx`, Brezger et al. (2005)), which has been specifically designed for the estimation of semiparametric regression models. The computational kernel is implemented in C++, allowing for an efficient treatment of loop-intensive MCMC simulations. A graphical user interface provides convenient access to the methodology and allows for a flexible model specification.

Table 1 Impact of rounding on computing times (in minutes) in the different simulation scenarios.

Digits	1	2	3	4	5
scenario (a)	3:11	7:11	9:56	10:03	10:28
scenario (b)	3:25	7:47	10:22	10:41	10:53
scenario (c)	6:12	14:24	19:41	20:14	20:21
scenario (c')	5:24	13:50	19:12	20:17	20:25

The computational bottleneck is simulating the regression coefficients of the penalized splines, in particular for the covariates measured with error. The first difficulty arises from the fact that for simulating from a K-dimensional multivariate Gaussian distribution, a K-dimensional system of equations has to be solved in each iteration. Replacing the simultaneous update with a single move algorithm would speed up computation but comes at the price of deteriorated mixing and convergence due to the ignored correlation of the elements in β_j. We therefore make use of sparse matrix computations, since the precision matrix P_j is a band matrix, see Rue (2001) and Brezger & Lang (2006) for details. This approach has the advantage to provide fast computations while keeping the correlation information included in the proposals.

The second difficulty is specific to the imputation of true covariate values: In each iteration new values are sampled, requiring the re-evaluation of the design matrix X_j. To shorten computation times, we consider two tricks: Firstly, instead of storing the complete design matrix, we only store the relevant part of it. Note that B-splines form a local basis such that in each row of X_j there are only $l+2$ non-zero entries (where l denotes the degree of the spline). Since we chose $l = 1$ as the standard in measurement error correction, there are actually only three values to be stored instead of K which is typically in the range of 20 to 40. Furthermore, only rows of the design matrix corresponding to distinct observed values of x_j have to be stored in combination with an index vector associating the observations with the different values for x_j. This storage scheme allows for a further reduction of computing times in a second step: Instead of storing the exact covariate values in double precision, we round them to a user-specified number of decimal places. As a consequence, several formerly distinct covariate values now coincide so that only a smaller number of rows of X_j has to be stored and re-computed in each iteration. In our simulations and applications we used two decimal places, a choice that lead to only negligible changes in the results while making a significant change in computing times in exploratory analyses. Table 1 provides some exemplary results for different decimal places and the simulation scenarios considered in the following section. There obviously is a tremendous gain in computing times for small decimal places, while computing times level off when using a large precision corresponding to almost no rounding.

Note also that due to the modular structure of MCMC algorithms, computing time only grows linearly when, for example, increasing the number of covariates subject to measurement error. Hence, computations with two covariates measured with error take approximately twice as long as computations with one covariate, which is in contrast to approaches where a decomposition of the correction problem in separate sub-problems is not feasible.

4 Simulations

4.1 Simulation Setup

To assess the properties of the proposed measurement error correction scheme and the validity of our implementation, we performed an extensive simulation study investigating model scenarios of increasing complexity:

(a) One covariate with measurement error:

$$\begin{array}{ll} \text{Observation model:} & \eta_i = \sin(x_i) + v_i\gamma, \\ \text{Measurement error model:} & w_i|x_i \sim \mathrm{N}(x_i, 1), \\ \text{Further settings:} & x_i \sim \mathrm{N}(0,1),\, v_i \sim \mathrm{N}(0,1),\, \gamma = 1. \end{array}$$

(b) One covariate measured with error in combination with a further nonparametric effect:

$$\begin{array}{ll} \text{Observation model:} & \eta_i = \sin(x_{i1}) + x_{i2}^2 + v_i\gamma, \\ \text{Measurement error model:} & w_i|x_i \sim \mathrm{N}(x_i, 1), \\ \text{Further settings:} & x_{i1} \sim \mathrm{N}(0,1),\, x_{i2} \sim \mathrm{U}(-1,1),\, v_i \sim \mathrm{N}(0,1),\, \gamma = 1. \end{array}$$

(c) Two covariates measured with error

$$\begin{array}{ll} \text{Observation model:} & \eta_i = \sin(x_{i1}) + 0.2x_{i2}^2 + v_i\gamma \\ \text{Measurement error model:} & w_{i1}|x_{i1} \sim \mathrm{N}(x_{i1}, 1),\, w_{i2}|x_{i2} \sim \mathrm{N}(x_{i2}, 0.64), \\ \text{Further settings:} & x_{i1} \sim \mathrm{N}(0,1),\, x_{i2} \sim \mathrm{N}(0,1),\, v_i \sim \mathrm{N}(0,1),\, \gamma = 1. \end{array}$$

Model (a) is the most simple one, where only one covariate is measured with error and the predictor contains only one single additional parametric covariate. In model (b), a second nonparametric effect is added to the predictor, but the corresponding covariate is observed exactly. Finally, in scenario (c), the covariate associated with the second nonparametric effect is also measured with error. Since scenario (c) is the most demanding one, we re-ran it with two replicated measurements on each of the covariates x_1 and x_2 to get an idea of the performance improvement by repeated observations on the measurement equation:

(c') Two covariates measured with error in two replications

$$\begin{array}{ll} \text{Observation model:} & \eta_i = \sin(x_{i1}) + 0.2x_{i2}^2 + v_i\gamma \\ \text{Measurement error model:} & w_{i1}^{(m)} \sim \mathrm{N}(x_{i1}, 1),\, w_{i2}^{(m)} \sim \mathrm{N}(x_{i2}, 0.64),\, m = 1,2, \\ \text{Further settings:} & x_{i1} \sim \mathrm{N}(0,1),\, x_{i2} \sim \mathrm{N}(0,1),\, v_i \sim \mathrm{N}(0,1),\, \gamma = 1. \end{array}$$

For each of the scenarios, we simulated data sets with responses from the following four types of responses:

(a) Binomial distribution with three replicated binary observations, i.e. $y_i \sim \mathrm{B}(3, \pi_i)$, $\pi_i = \exp(\eta_i)/(1+\exp(\eta_i))$.

(b) Binomial distribution with ten replicated binary observations, i.e. $y_i \sim \mathrm{B}(10, \pi_i)$, $\pi_i = \exp(\eta_i)/(1+\exp(\eta_i))$.
(c) Poisson distribution, i.e. $y_i \sim \mathrm{Po}(\lambda_i)$, $\lambda_i = \exp(\eta_i)$.
(d) Exponentially distributed duration times $T_i \sim \mathrm{Exp}(\lambda_i)$, $\lambda_i = \exp(\eta_i)$ subject to independent uniform censoring $C_i \sim \mathrm{U}(0, 50)$ resulting in an average censoring rate of 10%. The observed data is given by $t_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}(T_i \leq C_i)$.

For each response and each scenario, the sample size was fixed at $n = 500$ and the number of simulation replications was given by 100.

To benchmark the performance of the correction method, we did not only consider estimates from the imputation scheme, but also estimates based on the true covariate values and naive estimation based on the average of the measurements with error:

(a) Exact estimation: Use the true covariate values x_{ij} in the estimation procedure.
(b) Naive estimation: Use the average of repeated measurements $\bar{w}_{ij} = \sum w_{ij}^{(m)}/M$ as covariate.
(c) Corrected estimation: Impute the estimated true covariate values with MCMC.

The results from the exact and the naive estimation approach can serve as an upper and a lower bound for the performance of the corrected results.

4.2 Simulation Results

Figures 1 visualizes average estimates for the sine curve in scenario (a). As expected, the estimated curve in the naive approach is far too flat and almost equals a linear fit. In contrast, using the true covariate values leads to a satisfactory reproduction of the curve over a large part of the covariate domain. Note that only a very small number of observations is located outside the interval $[-2, 2]$ and therefore the deterioration of the average estimates in this area is simply due to a lack of data. MCMC-based measurement error correction falls in between the naive and the exact estimation results but indeed shows considerable correction. This becomes even more obvious from considering the MSEs (Figure 2), where the corrected results clearly outperform results from naive estimation. The improvement is smallest in the case of a binomial response with only three replications, where not too much information from the likelihood is available. For all other types of responses with increased likelihood information, the correction improves and the MSEs are closer to exact information than when using naive estimation.

When including an additional nonparametric effect to the model, results for the sine curve actually remain practically the same and are therefore not presented. To assess the impact on the effects without measurement error, Figure 3 shows boxplots of the MSE for binomial responses with ten replications and for survival times. Obviously there is some impact of measurement error also on the effects of covariates observed exactly but the change is much smaller compared to the effect on the sine curve. The most significant change is observed for survival times, and in this case MCMC-based imputation also yields more correction than for Binomial responses.

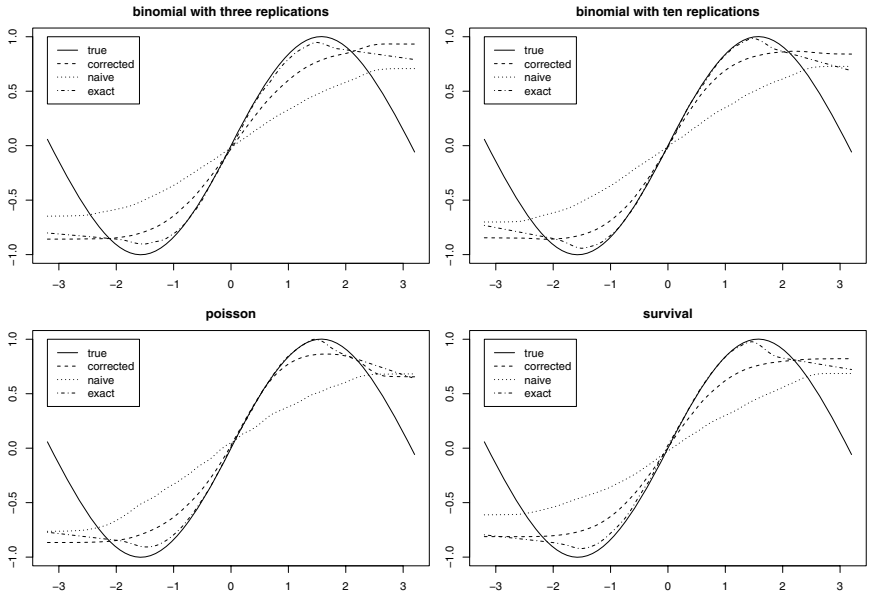


Fig. 1 Average function estimates for $\sin(x)$ for all four response types in scenario (a).

When considering two covariates with measurement error (Figure 4), results remain qualitatively the same as with one covariate: Quality of the estimates considerably increases when applying the proposed correction scheme with larger impact in case of response types with more information. Note, that the signal to noise ratio is smaller for the quadratic functions than for the sine curve and therefore correction is generally smaller for x_2 in terms of the bias although comparable improvements are achieved in terms of MSE. When including a second replication on the covariates measured with error, results improve even further (although of course also the results from the naive approach improve). In this case (Figure 5), the corrected estimates even start to indicate the local minimum and maximum of the sine curve, although the data in this area already get quite sparse. Similarly, the reproduction of the square function is now very close to the true function. In addition, the boxplots indicate that the corrected estimates perform almost as well as the estimates obtained with the true covariate values.

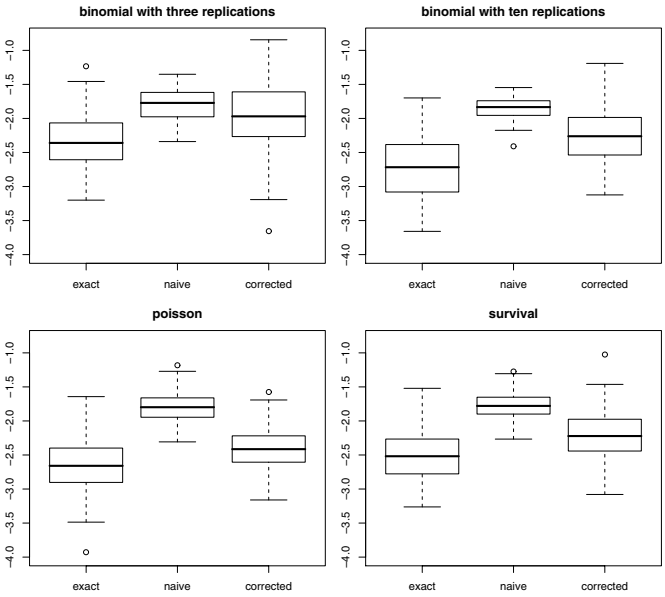


Fig. 2 Boxplots of log(MSE) for $\sin(x)$ for all four response types in scenario (a)

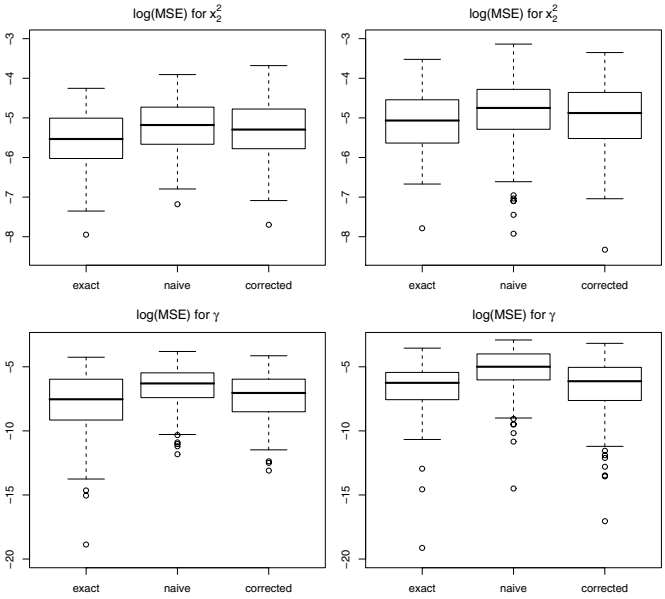


Fig. 3 Boxplots of log(MSE) for x_2^2 and γ for two response types in scenario (b) (binomial with ten replications in the left panel, survival in the right panel).

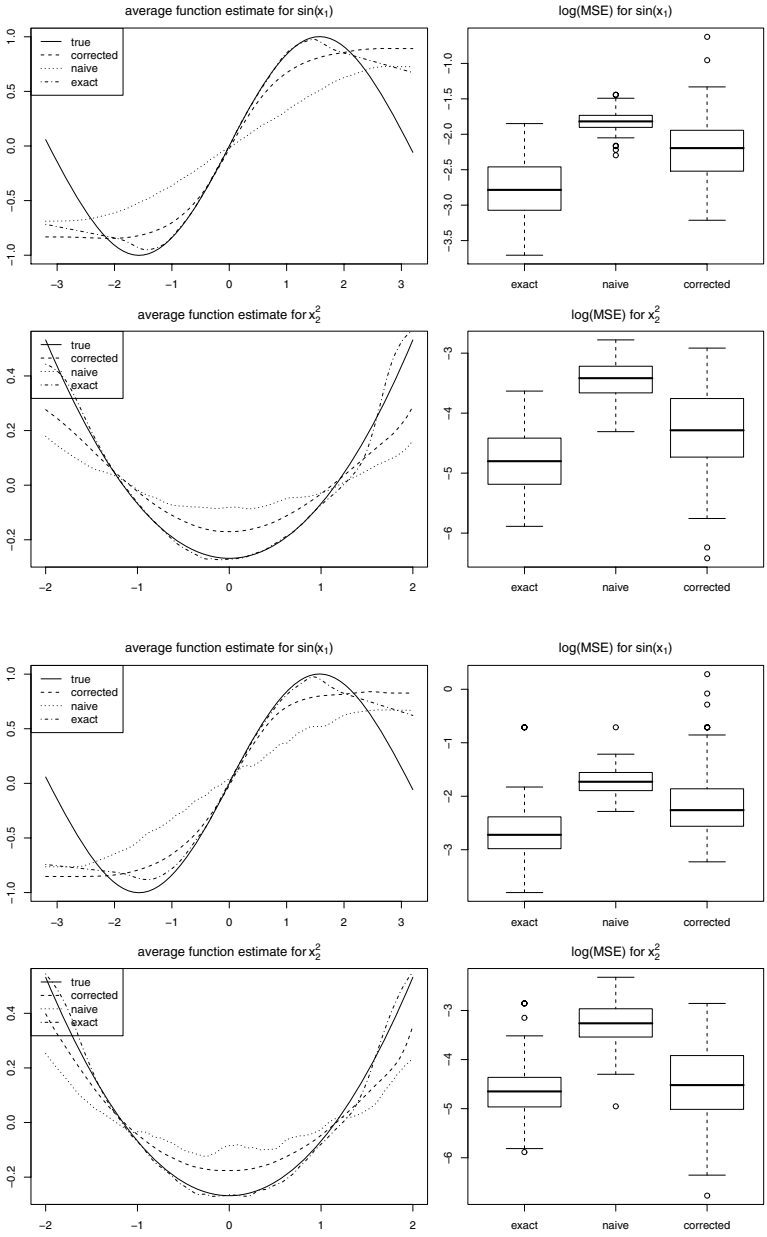


Fig. 4 Average estimates and boxplots for two response types in scenario (c) (binomial with ten replications in the upper two rows, poisson in the lower two rows).

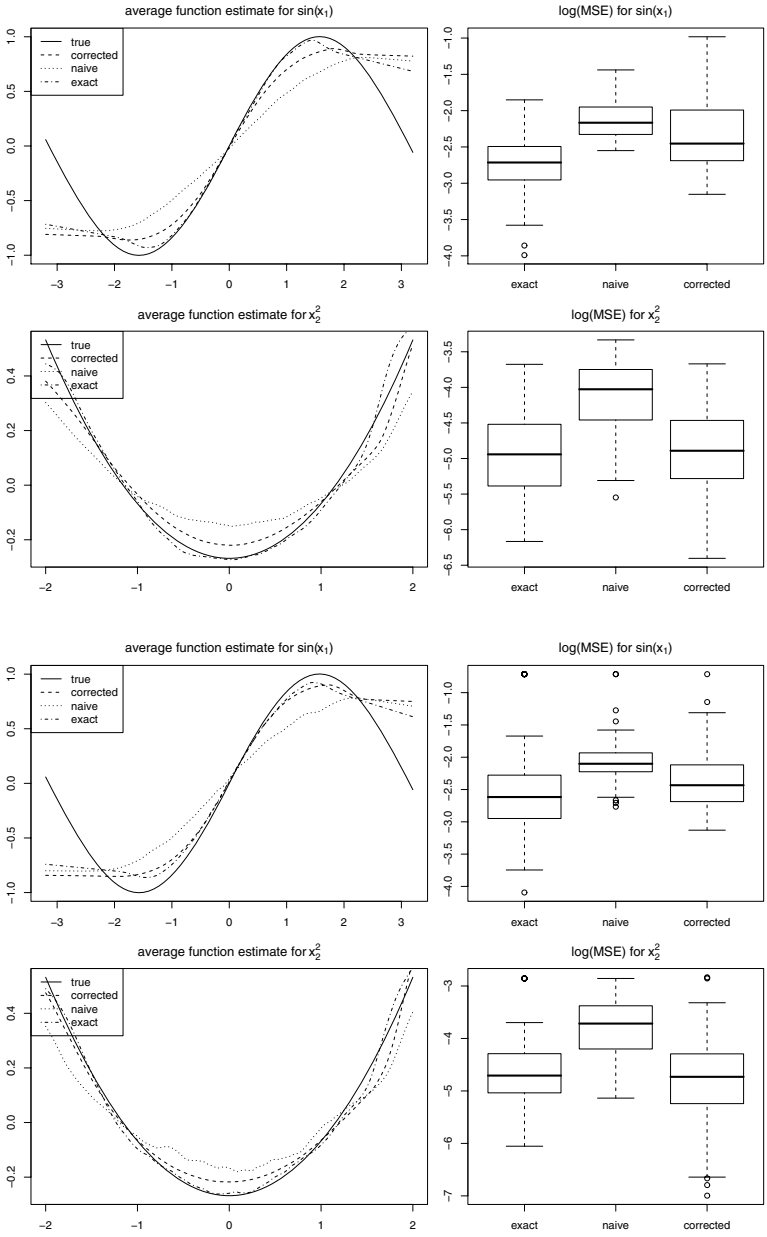


Fig. 5 Average estimates and boxplots for two response types in scenario (c') (binomial with ten replications in the upper two rows, poisson in the lower two rows).

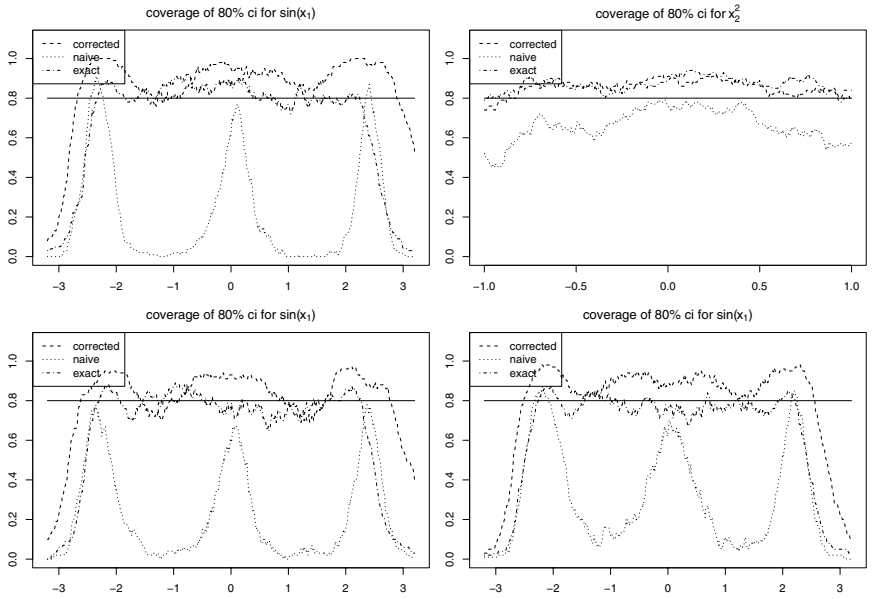


Fig. 6 Average coverage probabilities of 80% credible intervals for different effects in scenario (a) (upper left), scenario (b) (upper right), scenario (c) (lower left) and scenario (c') (lower right).

Finally, Figure 6 visualizes average coverage probabilities for different effects in the four scenarios. Again we find the impact of flattened estimation when using the naive approach: The empirical coverages are far too low not only for the effects of covariates measured with error but also for the square function in scenario (b). In contrast, the coverages of the corrected estimates are on average close to the nominal value all over the relevant covariate domain. Only at the boundaries, where data become sparse, the empirical coverage decreases. Note also, that using the true covariate values actually leads to somewhat too conservative credible intervals – an artefact that is found frequently in the context of Bayesian credible intervals.

In summary, our simulations allow the following conclusions to be drawn:

- MCMC-based imputation of the true covariate values allows to correct for the adverse effects of covariates measured with error. The correction effect is particularly expressed for the nonparametric effects of the covariates with measurement error while the amount of correction varies for effects of covariates observed exactly.
- In our simulation, measurement error had the expected impact on naive nonparametric regression results, i.e. nonparametric effects are underestimated and far too smooth.
- Ignoring measurement error also has dramatic impact on the coverage properties of the credible intervals.

We confirmed our findings in a second simulation study with smaller measurement error variances with practically the same results (not shown).

5 Incident Heart Failure in the ARIC Study

Our proposed methodology was motivated by the analysis of time to event data from the Atherosclerosis Risk in Communities (ARIC) study. ARIC is a large multipurpose epidemiological study conducted in four US communities (Forsyth County, NC; suburban Minneapolis, MN; Washington County, MD; and Jackson, MS). From 1987 through 1989, 15,792 male and female volunteers aged 45 through 64 were recruited from these communities for a baseline and three subsequent visits. The baseline visit (visit 1) included at-home interviews, laboratory measurements, and clinic examinations. The study participants returned for additional visits in 1990-92 (visit 2), 1993-95 (visit 3), and 1996-98 (visit 4). Details of the enrollment process and the study procedures are fully described by The ARIC INVESTIGATORS (1989).

Time to event data is observed continuously for multiple end points, but we focus here on the event *detection of heart failure* (HF), the inability of the heart to pump blood with normal efficiency. After exclusion of 752 participants with prevalent heart failure, 14,857 ARIC study participants were followed for incident heart failure hospitalization or death from 1987 to 2002. During a mean follow-up of 13.2 years, 1,193 participants developed HF (Kottgen et al. 2007).

The relationship between various risk factors, such as race, age or sex, and progression time to heart failure may be confounded by a series of baseline covariates. Two such important confounders are the baseline systolic blood pressure (SBP) and the baseline kidney function as measured by the glomerular filtration rate (GFR). Both SBP and GFR are measured with moderate error and their corresponding dose/response functions are expected to be non-linear. Taking into account these features of the data is necessary for satisfactory inference and can be handled using the methodology and software introduced in this paper. A reasonable approach to statistical modeling of the present data is to consider a survival model for time to heart failure with the following log-hazard function

$$\log\{\lambda_0(t)\} + f_1\{\log(\text{SBP} - 50)\} + f_2\{\log(\text{GFR})\} + \gamma_1 \text{sex} + \gamma_2 \text{AA} + \gamma_3 \text{age}, \quad (2)$$

where $\lambda_0(t)$ is the baseline hazard, $f_1(\cdot)$ and $f_2(\cdot)$ are unspecified smooth functions modeled as penalized splines, sex is a $0/1$ variable with 1 corresponding to males, AA is a $0/1$ variable with 1 corresponding to African Americans, and age being the baseline age. For $f_1(\cdot)$ and $f_2(\cdot)$ we used degree 1 penalized B-splines with 30 equidistant knots. We also employed quantile based knots but found that they produce very wiggly estimates both with and without measurement error correction in this example. This is probably due to the concentration of observations in a smaller part of the domain, that is more prevalent in the large data set of the application than in the comparable small simulation data sets.

Table 2 Corrected and naive posterior mean estimates, and 80% credible intervals for the parametric effects

	corrected			naive		
	$\hat{\gamma}$	80% ci		$\hat{\gamma}$	80% ci	
intercept	-8.577	-9.264	-7.938	-8.861	-9.402	-8.314
male	0.419	0.341	0.495	0.421	0.340	0.506
african american	0.355	0.254	0.451	0.350	0.262	0.440
age at first visit	0.083	0.075	0.091	0.081	0.074	0.089

In model (2), SBP represents the true long term average SBP and GFR represents the true filtration rate of the kidney at the time it was measured. Both variables are measured with error and replication studies are used to estimate the variance of the error process. To obtain the measurement error variance of $\log(\text{SBP} - 50)$ we use a replication study from the Framingham Heart Study described in Carroll et al. (2006), pages 112-114. In short, the Framingham study consists of a series of exams taken two years apart. The estimated measurement error using exams 2 and 3 is $\hat{\tau}^2_{\text{SBP}} = 0.01259$, which in the ARIC study corresponds to a reliability of 81%. Thus, in our model $\log(\text{SBP} - 50)$ is the true long term average $\log(\text{SBP} - 50)$ over a 2 year period.

There are important technical differences between measuring blood pressure with a sphygmomanometer and measuring the filtration rate of the kidney. Indeed, GFR can only be obtained through a long and awkward procedure that is impractical for routine analyses, as required by medical practice and large epidemiological studies. Instead, the estimated GFR (eGFR) is used in practice and is obtained from a prediction equation based on creatinine, gender and age (Hsu et al. (2005), Kottgen et al. (2007), Cheng & Crainiceanu (2009)). More precisely, the eGFR is predicted from the following equation:

$$\text{eGFR} = 186.3 * (\text{Serum Creatinine})^{-1.154} * (\text{Age})^{-0.203} * (0.742)^{(1-sex)} * (1.21)^{(\text{AA})}.$$

Thus, the eGFR measurement contains at least two non-ignorable sources of error: 1) the biological variability unaccounted for by the prediction equation; and 2) the laboratory variability associated with urine serum creatinine. To assess the variability of eGFR, a replication study was conducted in the Third National Health and Nutrition Examination Survey (NHANES III). Duplicate eGFR measurements were obtained for each of 513 participants aged 45 to 64 with eGFR ≥ 60 from two visits at a median of 17 days apart (Coresh et al. 2002). We assumed a classical measurement error model for log(eGFR) and calculated the measurement error variance as $\hat{\tau}^2_u = \frac{1}{2}\sum_{i=1}^{513}(w_{i1} - w_{i2})^2$, where w_{im} is the observed log(eGFR) for subject i at visit m. The estimated measurement error variance was $\hat{\tau}^2_u = 0.009743$ corresponding to a reliability of 0.80 in the ARIC data set and will be treated as a constant in our subsequent analyses.

Figure 7 and Table 2 summarize the results of both a naive and a measurement error corrected analysis. While the estimated baseline hazard rate remains practically

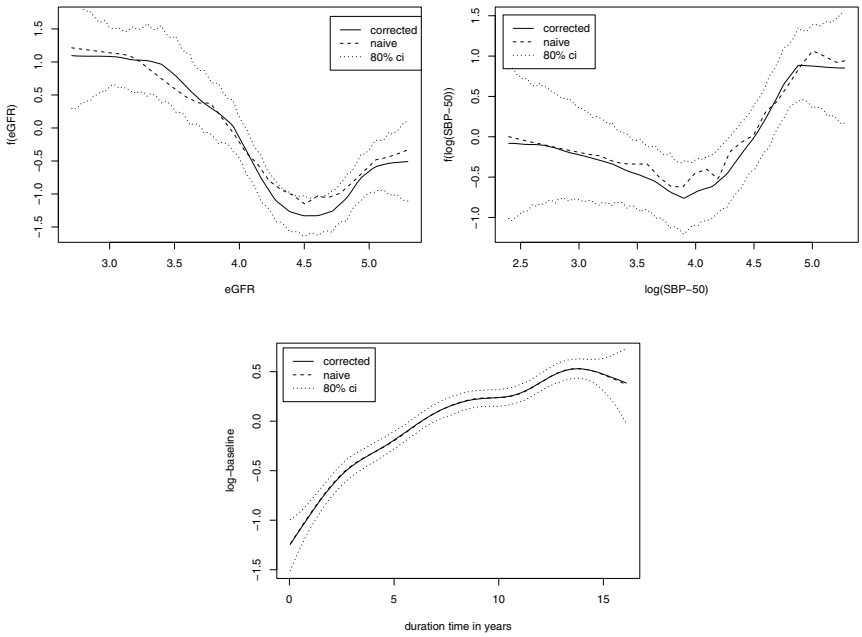


Fig. 7 Corrected and naive posterior mean estimates for the nonparametric effects of eGFR and log(SBP-50), and the log-baseline hazard rate with 80% pointwise credible intervals.

unchanged when correcting for measurement error, there are obvious changes in the results for SBP and eGRF. In particular, the local minima at 4.5 (eGFR) and 4.0 (SBP) are underestimated due to oversmoothing in the naive analysis. This effect is expressed more clearly for eGFR where the reliability is smaller and therefore the (relative) measurement error is larger.

Since the data set of the application is much larger than the data sets employed in the simulation, it is also worthwhile to consider the impact of rounding on the computing times again. With two valid decimal places, the corrected analysis (28,000 MCMC iterations on a dual core processor PC with 3Ghz CPU) including the imputation for two covariates took about 99 minutes, which is very competitive taking the complexity of the model and size of the data set into account. When increasing the number of valid decimal places, computing times increase to 215 minutes for 4 decimal places with visually indistinguishable results.

6 Summary

We have introduced a flexible Bayesian imputation scheme for correcting for measurement error in a large class of semiparametric regression models including models for the expectation in exponential family regression and models for the hazard rate in the case of survival data. The model specification permits quite flexible structures involving several nonparametric effects and several covariates measured with error. A variety of situations has been studied in a simulation study, indicating that the proposed algorithm works well even in complicated settings. The approach has been implemented in a user-friendly and efficient software package, allowing for easy access to the new methods. Moreover, the software supports a number of extended modeling possibilities not considered in this paper. To be more specific, varying-coefficient terms, interaction surfaces, spatial effects, or time-varying effects in survival can be augmented to the model specification if needed. This large flexibility of the model class is available due to the modular structure of MCMC simulations that makes all modeling components introduced previously to Bayesian semiparametric regression readily available as components in the measurement error correction approach.

A frequent drawback of approaches based on MCMC simulations are long computation times and difficulties in mixing and convergence. We circumvent both by considering a specialized implementation that relies on numerically fast sparse matrix computations in combination with efficient storage and rounding schemes. In addition, MCMC makes model combinations accessible that would require quite involved methodological treatment and computations in a frequentist approach.

Acknowledgements The authors thank Thomas Augustin and Ludwig Fahrmeir for valuable discussions at various stages of preparing this paper. The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, and N01-HC-55022. The authors thank the staff and participants of the ARIC study for their important contributions. The support for Ciprian Crainiceanu was provided by contracts N01-HC-55020 and R01-DK-076770-01.

References

The ARIC INVESTIGATORS (1989). The Atherosclerosis Risk in Communities (ARIC) study: design and objectives, *American Journal of Epidemiology* **129**: 687—702.

Berry, S. M., Carroll, R. J. & Ruppert, D. (2002). Bayesian Smoothing and Regression Splines for Measurement Error Problems, *Journal of the American Statistical Association* **97**: 160–169.

Biller, C. (2000). Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models, *Journal of Computational and Graphical Statistics* **9**: 122–140.

Brezger, A., Kneib, T. & Lang, S. (2005). BayesX: Analysing Bayesian structured additive regression models, *Journal of Statistical Software* **14**: (11).

Brezger, A. & Lang, S. (2006). Generalized additive regression based on Bayesian P-splines, *Computational Statistics and Data Analysis* **50**: 967–991,

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models - a modern perspective (2nd edition)*, Chapman & Hall / CRC, New York.

Cheng, Y.-J. & Crainiceanu, C. M. (2009). Cox models with smooth functional effects of covariates measured with error, *Journal of the American Statistical Association*, to appear.

Coresh, J., Astor, B., McQuillan, G., Kusek, J., Greene, T., Van Lente, F. & Levey, A. (2002). Calibration and random variation of the serum creatinine assay as critical elements of using equations to estimate glomerular filtration rate, *American Journal of Kidney Diseases* **39**: 920—929.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998). Automatic Bayesian curve fitting, *Journal of the Royal Statistical Society Series B* **60**: 333–350.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder), *Statistical Science* **11**: 89–121.

Fahrmeir, L. & Kneib, T. (2009). Propriety of Posteriors in Structured Additive Regression Models: Theory and Empirical Evidence, *Journal of Statistical Planning and Inference* **139**: 843–859.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer, New York.

Fahrmeir, L. & Lang, S. (2001). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors, *Applied Statistics*, **50**: 201–220.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective, *Statistica Sinica* **14**: 731–761.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear models, *Statistics and Computing* **7**: 57–68.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology*, Chapman & Hall / CRC, Boca Raton.

Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models, *Journal of the American Statistical Association* **101**: 1065–1075.

Hobert, J. P. & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models, *Journal of the American Statistical Association* **91**: 1461–1473.

Hsu, C. C., Kao, W. H., Coresh, J., Pankow, J. S., Marsh-Manzi, J., Boerwinkle, E. & Bray, M. S. (2005). Apoliprotein E and progression of chronic kidney disease, *Journal of the American Medical Association* **293**: 2892–2899.

Jullion, A. & Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models, *Computational Statistics & Data Analysis* **51**: 2542–2558

Kneib, T. & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression, *Scandinavian Journal of Statistics* **34**: 207–228.

Kottgen, A., Russell, S. D., Loehr, L. R., Crainiceanu, C. M., Rosamond, W. D., Chang, P. P., Chambless, L. E. & Coresh, J. (2007). Reduced kidney function as a risk factor for incident heart failure: The Atherosclerosis Risk in Communities (ARIC) study, *Journal of the American Society of Nephrology* **18**: 1307–1315.

Richardson, S. & Gilks, W. R. (1993). Conditional independence models for epidemiolgoical studies with covariate measurement error, *Statistics in Medicine* **12**: 1703–1722.

Rue, H. (2001). Fast sampling of Gaussian Markov Random Fields with Applications, *Journal of the Royal Statistical Society B* **63**: 325–338.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003), *Semiparametric Regression*, University Press, Cambridge.

Stephens, D. A. & Dellaportas, P. (1992). Bayesian analysis of generalised linear models with covariate measurement error, *in* Bernardo, J. M., Berger, J. O., Dawid, A. P. & Smith, A. F. M. (eds), *Bayesian Statistics 4*, Oxford University Press.

Wang, C.-Y. & Pepe, M. S. (2001). Expected estimating equations to accomodate covariate measurement error, *Journal of the Royal Statistical Society B* **62**: 509–524.

Wood, S. N. (2006). *Generalized Additive Models*, Chapman & Hall / CRC, Boca Raton.

Determinants of the Socioeconomic and Spatial Pattern of Undernutrition by Sex in India: A Geoadditive Semi-parametric Regression Approach

Christiane Belitz, Judith Hübner, Stephan Klasen and Stefan Lang

Abstract In this paper, we use geoadditive semiparametric regression models to study the determinants of chronic undernutrition of boys and girls in India in 1998/99. A particular focus of our paper is to explain the strong regional pattern in undernutrition and sex differences in determinants of undernutrition. We find that determinants associated with competition for household resources and cultural factors are more important for the nutrition of girls than boys, while boys' nutrition reacts more sensitively to nutrition and medical care access. With our models we are able to explain a large portion of the spatial pattern of undernutrition of boys and girls, but significant spatial patterns remain. We are also able to fully explain the spatial pattern of sex differences in undernutrition with our empirical model.

1 Introduction

Extremely high prevalence of childhood undernutrition as well as very large gender bias in various indicators (including infant and child mortality, education, and employment) are two of the most severe development problems in India. Using the

Christiane Belitz
Institut für Statistik, Ludwigstr. 33, Ludwig-Maximilians-Universität München, Germany

Judith Hübner
Institut für Mathematik, Boltzmannstr. 3, Technische Universität München, Germany

Stephan Klasen
Department of Economics, University of Göttingen, Platz der Göttinger Sieben 3, D-37073 Göttingen, Germany, URL: http://www.uni-goettingen.de/en/64786.html, e-mail: sklasen@uni-goettingen.de

Stefan Lang
Department of Statistics, University of Innsbruck, Universitätsstr. 15, A-6020 Innsbruck, Austria, URL: http://www.uibk.ac.at/statistics/personal/lang/, e-mail: stefan.lang@uibk.ac.at

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_9,

most commonly used indicator of childhood undernutrition, insufficient weight for age (measured as the share of children being more than 2 standard deviations below an international reference standard for weight for age), some 47 percent of children below 5 were underweight in 1998/99, among the five worst performers in the world (IIPS 2000).[1] In addition, as shown by Klasen & Wink (2002), Klasen & Wink (2003) and Sen (2003), India also belongs to a small group of countries with extremely large gender bias in mortality. In fact, Klasen & Wink (2002) and Klasen & Wink (2003) estimate that some 39 million or 7.9 percent of the female population was "missing" in the 2001 census and have been victims of gender inequality in survival.

Undernutrition, overall infant and child mortality, and sex differentials in mortality have a strong regional pattern which has been noted repeatedly in the literature, see Murthi, Guio & Dreze (1995), Agarwal (1994), Dreze & Sen (1995), Dreze & Sen (2001), Klasen & Wink (2002), Klasen & Wink (2003). All three indicators show a much worse performance in Northern States, while Southern India and East and Northeastern India performs much better on all three, with the Southern state of Kerala being the well-known star performer in these (as well as other dimensions of development). This is documented in Table 1 which shows overall infant and under five mortality rates and by sex, sex ratios (males per females), undernutrition rates overall and by sex for India's largest states where this information is available from the 2001 census or the 1998/99 Family Health Survey. While there are significant variations within each region (in the case of mortality rates also due to sample size issues) the regional patterns are remarkably similar, with Northern India standing out as the worst performer in all dimensions and Southern India doing best in most of them. Within Southern India, it is Kerala that clearly stands out as the area with the lowest overall infant and child mortality rates and undernutrition as well as the lowest gender gaps in these indicators.

At the same time, the regional patterns differ greatly by indicator. While Kerala is doing best in all of them, there are some states that do better in overall mortality rates than in gender gaps, and for others, the reverse is the case. It is also notable that the sex differentials in undernutrition tend to be smaller than the differentials in mortality, while the regional differences in overall undernutrition are sizable.[2]

These large geographic differentials at the district level can also be seen in the four maps of Figure 1. They show the spatial pattern of undernutrition for girls and boys separately, without controlling for covariates. The first two show the spatially smoothed effects (see below for a discussion of how these effects are calculated), while the latter two show significant positive (white) and negative (black) spatial effects, with the left graph showing significant overall effects (i.e. for both sexes combined) while the right map shows significant sex differences by district. A very

[1] For a discussion of potential problems and biases when comparing undernutrition rates across the developing world, see Klasen (2008). While it is argued there that measurement problems contribute to the fact that South Asian children are reported to be the worst nourished of the world and that undernutrition might be as severe in other regions including in Africa, there is no doubt that undernutrition is an extremely serious problem in India.

[2] For further discussion on these relatively small differentials in undernutrition, see Svedberg (2002) and IIPS (2000).

Table 1 The table shows overall infant and under five mortality rates and its respective female-male ratio, sex ratios (males per females), undernutrition rates overall and its female-male ratio for India's largest states.

State	Sex ratio (M/F) 2001	Infant Mort. all	IMR F/M	U5M all	U5M F/M	Stunt.	Stunt. F/M	Under-weight Rates	Under-weight F/M
All-India	1.072	66.2	0.98	96.0	1.19	44.9	1.064	46.7	1.078
North	1.113	77.61	1.05	113.07	1.32	51.3	1.067	48.1	1.074
Delhi	1.218	48.1	1.01	58.1	0.78	36.7	1.085	34.6	0.877
Haryana	1.161	56.0	0.89	73.6	1.35	50.0	1.116	34.6	1.195
Himachal Pradesh	1.033	34.8	0.80	49.4	0.67	41.5	0.775	43.7	0.919
Jammu & Kashmir	1.121	64.9	0.95	79.2	1.27	38.9	0.942	34.6	0.903
Punjab	1.142	53.9	1.28	75.5	1.80	39.1	1.034	28.7	1.122
Rajasthan	1.086	78.5	1.11	114.7	1.31	52.0	1.069	50.7	1.057
Madhya Pradesh	1.088	85.7	1.03	139.1	1.25	51.1	1.073	55.2	1.091
Uttar Pradesh	1.114	85.4	1.05	122.6	1.35	55.4	1.079	51.4	1.085
East	1.070	63.97	0.84	93.23	1.01	47.1	1.125	52.3	1.082
Bihar	1.088	71.5	0.89	108.1	1.10	53.6	1.028	54.3	1.059
Orissa	1.029	79.4	0.91	113.8	0.86	44.0	0.993	54.5	0.993
West Bengal	1.071	49.1	0.75	68.4	1.00	41.7	1.286	49.1	1.146
Northeast	1.068	61.86	0.90	89.13	0.88	46.0	0.930	34.9	1.021
Arunachal Pradesh	1.120	64.0	0.93	99.5	0.88	26.6	0.938	24.2	0.840
Assam	1.070	66.6	0.85	95.8	0.77	50.4	0.986	36.1	1.066
Manipur	1.022	33.7	0.70	61.0	1.43	31.2	0.902	27.6	1.160
Meghalaya	1.029	89.1	0.79	115.6	1.00	45.1	0.844	37.9	1.082
Mizoram	1.070	36.4	1.36	55.6	2.04	34.5	0.791	27.7	0.861
Nagaland	1.111	44.1	1.19	88.4	0.84	32.4	0.667	23.6	0.660
Sikkim	1.143	42.5	0.97	57.7	0.98	32.2	1.035	20.7	1.161
Tripura	1.055	42.7	1.17	45.0	1.05	40.5	0.731	42.6	0.827
West	1.085	48.46	0.85	65.72	1.09	41.0	1.059	47.9	1.092
Goa	1.041	34.7	0.69	41.8	0.76	18.1	0.451	28.6	0.609
Gujarat	1.087	60.0	0.78	78.8	0.96	43.6	1.079	45.1	1.249
Maharashtra	1.085	42.6	0.88	59.2	1.15	40.0	1.057	49.6	1.016
South	1.012	49.00	1.01	68.25	1.14	33.1	1.032	37.3	1.092
Andhra Pradesh	1.022	64.0	1.09	93.7	1.38	38.6	1.064	37.7	1.145
Karnataka	1.036	49.9	0.93	66.6	0.97	36.5	1.089	43.8	1.083
Kerala	0.945	14.9	0.52	16.2	0.46	22.0	0.978	26.9	1.053
Tamil Nadu	1.013	47.3	1.23	65.1	1.34	29.3	0.973	36.6	1.053

clear pattern of undernutrition is visible for both girls and boys. In North-Central India (particularly Uttar Pradesh (UP), Madhya Pradesh (MP), Rajasthan, and Orissa), both sexes suffer from significant undernutrition, while in the very North, the East, and the South West, they are doing significantly better. This spatial pattern seems to be more pronounced for girls than boys. As a result, the significance map of the sex differences in undernutrition by district shows girls doing significantly worse than boys in UP, MP, and West Bengal, while they are doing significantly better in the relatively small Northeastern states (e.g. Assam, Nagaland, Tripura).

Understanding the spatial pattern of undernutrition is important in its own right to design appropriate policies to combat undernutrition. In addition, given the close presumed relationship between undernutrition and child mortality [3], understanding the determinants of undernutrition for each sex should help in an understanding of the sex-specific differences in mortality. Given the strong regional pattern in overall mortality, undernutrition, and sex differences in mortality, it would be critical to analyze the regional pattern of undernutrition by sex to understand the contribution of sex differences in undernutrition to these regional patterns.

The purpose of this paper is therefore to analyze the determinants of undernutrition by sex at the district level in India using the 1998/99 India Family Health Survey. In our analysis, we use a comprehensive set of covariates that have been suggested in the literature based on either theoretical considerations or empirical evidence. The particular methodological innovations to the analysis of this problem are that we use particularly flexible semi-parametric regression models to model non-linear effects and that we explicitly include (smooth) spatial effects in our models. Model selection and choice of covariates is enhanced by recent methodology for simultaneous selection of relevant determinants of undernutrition and estimation of regression coefficients.

We find that factors that are related to competition for resources at the household level have a larger impact on undernutrition for girls than for boys. In contrast, boys undernutrition levels appear more influenced by care and health-seeking behavior, including pre-natal care and breast-feeding practises. Our models are able to explain a significant share of the spatial pattern of undernutrition for both sexes. But a significant spatial pattern in overall undernutrition remains which we are unable to explain with the variables we have at our disposal.

At the same time, our models are able to fully explain the spatial patterns of sex differences in undernutrition, suggesting that our socioeconomic determinants are able to fully capture the observed spatial pattern of sex differences in undernutrition in India (shown in Figure 1).

2 The Data

This analysis is based on micro data from the second National Family Health Survey (NFHS-2) from India which was conducted in the years 1998 and 1999. The sample is representative of the population and covers more than 90 000 ever-married women in the age group 15 – 49 from 26 Indian states (IIPS (2000), p. xiii). In addition, the survey collected detailed health and anthropometric information on 32 393 children born in the 3 years preceding the survey (IIPS (2000), p. xix). The NFHS-2 provides retrospective information on fertility, mortality, family planing, domestic violence, education, standard of living and important aspects of nutrition, health and health care. Complete data of 24 989 weighed and measured children under 4 years are

[3] See, for example, Pelletier (1994) and Osmani (1990) for a discussion.

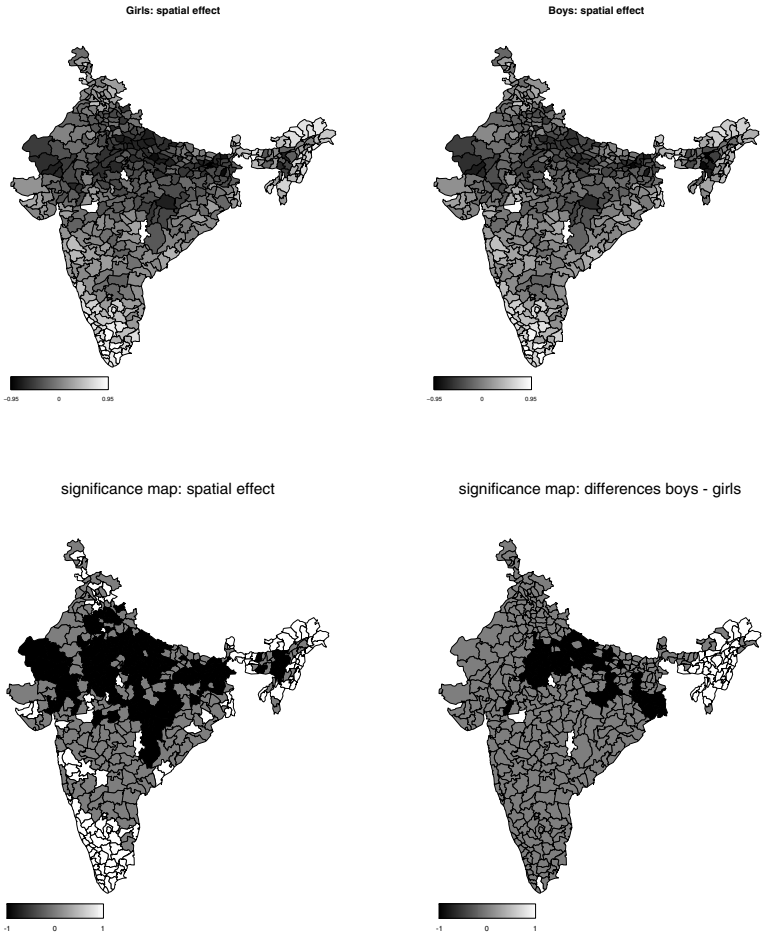


Fig. 1 Top panel: Spatially smoothed Z-score for boys and girls without controlling for covariates. Bottom left panel: Significance map for the smoothed spatial effect for both sexes. Regions with average Z-score significantly above (below) the overall average Z-score are colored in white (black). Bottom right panel: Significance map of the difference Z-score between boys and girls. White (black) denotes regions with significant higher (lower) average Z-score for girls compared to boys.

available. There are 13 090 observations of male and 11 899 observations of female children. Although the exact spatial location of the households is not available in the publicly available data, the International Institute for Population Sciences made the district locations of the households available to us which we can use in our analysis below.

3 Measurement and Determinants of Undernutrition

Nutritional status is the result of the complex interaction between the food a child gets to eat, its overall state of health, and the environment in which the child lives. Undernutrition refers to any imbalance in satisfying nutrition requirements and is therefore the result of a combination of inadequate intake of protein, energy and micronutrients and frequent infections or diseases. In line with the literature on the subject, undernutrition among children is taken here to manifest itself in growth failure: undernourished children are shorter and lighter than they should be for their age, see de Onis, M., de Frongillo & Blossner, M. (2000), p. 8 – 12, UNICEF (1998), p. 10 – 14 and www.who.int/nut/nutrition2.htm.

3.1 Measurement

Undernutrition among children is usually measured by determining the anthropometric status of the child. Researchers distinguish between three types of undernutrition (WHO (2002), p. 3 - 4):

- Wasting: insufficient weight for height indicating acute undernutrition.
- Stunting: insufficient height for age indicating chronic undernutrition.
- Underweight: insufficient weight for age which can be a result of both stunting and wasting.

In this paper we focus on the influences on stunting, because the NFHS-2 does not contain a lot of information about the recent past. Therefore it is not possible to analyze acute undernutrition with any precision.

To get a measure of undernutrition in a population, young children are weighed and measured and the results are compared to those of a "reference population" known to have grown well. The reference standard typically used for the calculation is the NCHS/CDC Growth Standard (National Center for Health Statistics/ Center for Disease Control) that has been recommended for international use by the World Health Organisation (WHO), see WHO (2002), p. 4 - 6.

The international reference growth curves were formulated in the 1970s by combining growth data from two distinct data sets. For children under 24 months data from a study of white, largely bottle-fed middle-class children from the longitudinal Fels study (Ohio Fels Research Institute) from 1929 – 1974 were used, while for older children the standard is based on data of three cross-sectional USA representative surveys conducted between 1960 and 1975, see WHO (2002), p. 4 - 6. There are some questions regarding the appropriateness of this standard for international comparisons of undernutrition (Klasen 2008), but for comparisons within a country at one point in time these problems are likely to be small and the Nutrition Foundation of India has concluded that this standard recommended by the WHO is applicable to Indian children, see Mishra, Lahiri & Luther (1999), p. 7. There are also some other

more technical problems with the standard which led WHO to recently produce a new growth standard for children.[4]

Undernutrition (Stunting, Wasting and Underweight) for a child i is typically determined using a Z-score which is defined as

$$Z_i = \frac{AI_i - MAI}{\sigma}, \tag{1}$$

where AI refers to the individual's anthropometric indicator (e.g. height at a certain age), MAI refers to the median of the individuals' anthropometric indicator of the reference population, and σ refers to the standard deviation of the individuals' anthropometric indicator of the reference population (WHO (2002), p. 6 - 7).

The percentage of children whose Z-scores are below -2 standard deviations from the median of the reference category are considered as undernourished (stunted, wasted and underweight), while those with Z-scores below -3 are considered severely undernourished (WHO (2002), p. 6 - 7 and Kandala, Fahrmeir, Klasen & Priebe (2008), p. 4). In this analysis the Z-score is used as a continuous variable to use the maximum amount of information available in the data set.

In accordance with the conceptual framework developed by the United Nations Children's Fund one can distinguish between immediate, underlying and basic determinants. This framework incorporates both biological and socioeconomic causes, and encompasses causes at both micro and macro levels (UNICEF (1998), p. 23 - 34 and Smith & Haddad (1999), p. 3 - 5).

3.2 Determinants of Undernutrition

The immediate determinants of child's nutritional status manifest themselves at the level of the individual human being. The two most significant immediate causes of malnutrition are inadequate dietary intake and illness. The immediate determinants, in turn, are influenced by three underlying determinants manifesting themselves at the household level. These are inadequate access to food, unhealthy environment (including insufficient health services) and inadequate care for children and women. The basic determinants include the potential resources available to a country or community, which are limited by the natural environment, access to technology, and

[4] The NCHS/CDC Growth Standard is beset with other problems: the splicing of two different data sets causes considerable discontinuity at the age of 24 months (see below). A further problem is that the bottle-fed children appear to have grown and put on weight more rapidly in the first six months than exclusively breast-fed children which erroneously suggests that exclusive breast-feeding for six months might contribute to undernutrition (Klasen & Moradi (2000), p. 3). The recently published new international growth standard for children is based on the growth experience of six well-to-do samples from across the world. Using this growth standard has a slight influence on overall rates of stunting, wasting, and underweight rates, but is unlikely to affect determinants of undernutrition. For a discussion of this new growth standard, which included a sample from India, see WHO (2006) and Klasen (2008).

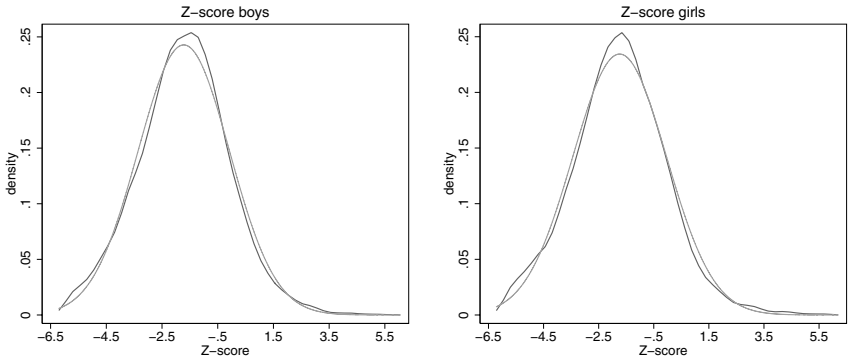


Fig. 2 Kernel density estimates for the distribution of the Z-score.

the quality of human resources (UNICEF (1998), p. 23 - 34 and Smith & Haddad (1999), p. 3 - 5).

In order to capture this complex chain of causation, researchers have either focused on a particular level of causality, have estimated structural equations that address the interactions, have used graphical chain models to assess the causal pathways, or have used multi-level modeling techniques (Kandala, Fahrmeir, Klasen & Priebe 2008, Harttgen & Misselhorn 2006, Caputo, Foraita, Klasen & Pigeot, I. 2003). With the data available, it is not always possible to clearly separate immediate from underlying or underlying from basic determinants. In this paper we estimate a reduced form equation that mainly models factors that are underlying determinants of undernutrition.

4 Variables Included in the Regression Model

In line with the discussion above, we study the determinants of the height for age Z-score as a continuous response to use the maximum amount of information available in the data set. The kernel density estimates of the distribution of the Z-scores (for boys and girls), together with a normal density, give clear evidence that a Gaussian model is a reasonable choice for inference, see Figure 2.

We include covariates in three ways into the model. First as categorical covariates where we expect linear effects and where a categorical treatment is suggested by the form of the data available, second as continuous covariates where we allow for a flexible functional form, and third as spatial covariates as we discussed above.

Categorical Covariates

Empirical distributions of categorical covariates, together with codings used in the analysis, are given in Table 2. In line with the conceptual framework of the underlying determinants, several of the variables are indicators of the health environment and health access (in particular, pre-natal care indicators, vaccination indicators, toilet access), the care environment (e.g. preceding birth interval, first milk), and some are indicators of competition for nutritional resources (e.g. twins, household size, preceding birth interval, first born, planned child). We also control for rural/urban as well as religion of the households which might also affect the resource, health environment, and care situation and practises of the households.[5]

Spatial Covariates

We use the district as the geographic unit of analysis. In the data set there are observations of 438 different districts from 26 states. The districts are given numerically and can be matched to an Indian map showing district boundaries.

Continuous Covariates

After checking for possible non-linear influences, all continuous covariates are modeled nonparametrically in this approach to capture non-linearities and their differences by sex. Table 4 gives an overview of continuous covariates included into the model.

Linear Index: Household's Economic Status

Most of the covariates are given directly in the DHS data sets. Since in the DHS data sets neither household income nor consumption expenditures are reported, we had to create a new variable that captures the economic resource base of the household. Hence, to get a proxy for long-run household wealth we constructed, following common practise in the literature, a linear index from a set of asset indicators using principal components analysis (PCA) to derive the weights a_i. The factor loadings of the PCA determine the weights a_i. The linear index W_j capturing a household's economic status is calculated as

$$W_j = \sum_{i=1}^{n=24} \frac{a_i(w_{ji} - w_i)}{s_i} = \frac{a_1(w_{j1} - w_1)}{s_1} + \ldots + \frac{a_{24}(w_{j24} - w_{24})}{s_{24}},$$

[5] Including other categorical covariates into the model, such as "access to good water quality", "child had fever recently", "child had cough recently", "child had diarrhea recently", "child's sibling has died" and "mother has access to massmedia regularly", turned out to be insignificant or to have too high correlations with other covariates. Thus, they were omitted in our analysis.

Table 2 Overview of categorical covariates included in the model. (*) It is recommended by the World Health Organisation (WHO) that mothers breastfeed their children within one hour after birth, because the hormone oxytocin is released resulting in uterine contractions what facilitates the expulsion of the placenta. This is important for the regeneration of the mother. Children also benefit from the first milk because it provides natural immunity. (**) The covariate *vacC* is age-dependent, because older children are more likely to have received more vaccinations than younger children. According to international recommendations of the Indian government children should be completely vaccinated within their first year of life. A child is regarded as completely vaccinated if it had received the vaccinations against tuberculosis (BCG), diphtheria, pertussis and tetanus (DPT), polio and measles, see Table 3.

Categorical covariates	Frequency in %	Effect-coding	Content
birthinC	18.91	-1 $\leq$ 24 months	Preceding birth
	81.09	1 $>$ 24 months	interval?
born1stC	70.60	-1 no	First born
	29.40	1 yes	child?
firstmC*	88.03	1 no	Child got
	11.97	-1 yes	first milk?
hhsizeH	33.32	hhsmallH: 1 $\leq$ 5 HH-members	Size of household
	50.38	Ref.-kat.:-1 6 - 10 HH-members	in which the
	16.30	hhlargeH: 1 $\geq$ 10HH-members	child lives?
ironfolM	7.05	-1 no	Mother received iron folic
	62.95	1 yes	tablets during pregnancy?
plannedC	22.71	1 no	Child was
	77.29	-1 yes	planned?
precareM	28.67	-1 no	Mother received medical
	71.33	1 yes	care during pregnancy?
religM	74.68	Ref.-kat.:-1 Hinduism	
	13.75	relislaM: 1 Islam	Religion
	6.90	relchriM: 1 Christianity	mother
	2.18	relsikhM: 1 Sikh	belongs to?
	2.49	relotheM: 1 Other religion	
ruralH	27.10	-1 no	Rural place
	72.90	1 yes	of residence?
tetanusM	20.90	-1 no	Mother received tetanus
	79.10	1 yes	injection during pregnancy?
toiletH	60.73	-1 no	Household has toilet
	39.27	1 yes	facility of any kind (pit, latrine, flush)
twinC	99.05	-1 single birth	Child was born
	0.95	1 multiple birth	under multiple birth?
vacC**	43.08	-1 no	Child is vaccinated
	56.92	1 yes	according to its age?

Table 3 Coding for the covariate *vacC*.

	received vaccinations		
Age of child	no vaccinations	some of the vaccinations	all vaccinations
0 to 1 month	category 1	category 1	category 1
2 to 11 months	category 0	category 1	category 1
12 to 35 months	category 0	category 0	category 1

with a_i weight (scoring factor) for the i-th asset, determined by the PCA, w_{ji} is the household's value, in which the j-th child lives, for the i-th asset, w_i is the mean of the i-th asset calculated over all households and s_i is the standard deviation of the i-th asset calculated over all households. This method for constructing an index

Table 4 Overview of continuous covariates included in the model.

Continuous covariates	Mean	Unit of measurement	Content
agebirM	24.11	years	Mother's age at child's birth?
ageC	17.27	months	Child's age?
bfmC	13.86	months	Months child was breastfed?
bmiM	19.80	kg/m^2	Mother's body mass index?
ecstatH	–0.01	index	Household's economic status
educM	4.07	years of education	Mother's educational attainment?
heightM	151.55	cm	Mother's height?
womstatM	–0.005	index	Mother's women's status?

indicating a households economic status was suggested by Filmer & Pritchett (1998), recommended by UNICEF (www.childinfo.org/MICS2/finques/gj00106a.htm), and is now routinely used when analyzing DHS surveys. The asset indicators included in our index are given in Table 5.[6]

According to the findings of Filmer and Pritchett the constructed index for India can be assessed as reliable in three dimensions. It is internally coherent and produces clean separations across the poor, middle and rich households for each asset individually. It is robust to the assets included. That means that the asset index produces very similar classifications when different subsets of variables are used in its construction. Furthermore it produces reasonable comparisons with poverty and output across states (Filmer & Pritchett (1998), p. 8 – 11).

However, one criticism of this index is a problem with urban/rural comparisons. It might be that rural households seem to be less wealthy because of non-availability of infrastructure (electricity, piped water, ...). We tried to mitigate this effect by including the ownership of agricultural goods (land, machines, live stock) into the index.

Table 5 shows the weights for the different factors included in the asset index. The direction of influence of each factor on the index is as expected and the different factors have remarkably similar weights, i.e. a similar influence on the resulting asset index.

Linear Index: Mother's Women's Status

There is a sizable literature that demonstrates that women's status has a significant impact on health outcomes for children.[7] Also, Smith et al. (2003) found that women's status has a particularly important impact on undernutrition in a multi-country micro data analysis using the Demographic and Health Surveys. Since no measure about the women's status is directly given in the DHS data set we constructed in analogy to the

[6] The asset indicators included in our index are not exactly the same as suggested by Filmer & Pritchett (1998)

[7] See, for example, World Bank (2001) for a survey.

Table 5 Asset variables for household's economic status index.

	Weight	Mean	Std. dev.
	a_i	w_i	s_i
Household ownership of certain durables			
Clock	0.34700	0.70221	0.45730
Fan	0.40979	0.46699	0.49892
Sewing machine	0.35091	0.23345	0.42304
Refrigerator	0.35212	0.10802	0.31041
Radio	0.30388	0.40013	0.48993
Television	0.42673	0.36033	0.48011
Bicycle	0.19174	0.45350	0.49784
Motor bicycle	0.34675	0.12071	0.32580
Car	0.18189	0.01802	0.13301
Characteristics of household's dwelling			
Piped drinking water	0.36868	0.39062	0.48790
Drinking water from spring or well	–0.34723	0.54401	0.49807
Drinking water from open source (surface)	–0.03251	0.05754	0.23288
Drinking water from other source	0.00724	0.00783	0.08814
Flush toilet	0.39012	0.23723	0.42539
Pit toilet	0.07606	0.15560	0.36248
No toilet	–0.39625	0.60717	0.48839
Electricity	0.33101	0.64215	0.47938
Number of rooms > 4	0.12388	0.16467	0.37089
Kitchen as a separate room	0.20913	0.52879	0.49918
Quality of dwelling materials: low, middle, high	–0.33276	1.02598	0.79082
Main cooking fuel is wood/dung/coal	–0.38474	0.75293	0.43132
Ownership of land or agricultural goods			
Ownership of land < 0.4; $0.4 \leq$ Owns. ≤ 2; > 2	0.62205	0.68298	0.77290
Ownership of live stock	0.57735	0.53690	0.49865
Ownership of agricultural tools/ machines	0.52889	0.10266	0.30353

index about a household's economic status a linear index about the women's status. Women's status is defined to be women's power relative to men's. Women with a low status tend to have weaker control over resources in their households, tighter constraints on their time, more restricted access to information and health services, and poorer mental health, self-esteem and self-confidence. The variables listed in Table 6 are thought to be closely related with women's own nutritional status, the quality of care they receive, their own children's birth weights, and the quality of care provided to their children (Smith et al. 2003).

To build an index about women's status was suggested by Smith et al. (2003). The index we constructed is slightly different to the one proposed to fit the needs of our analysis here.[8]

[8] Our index contains three parts: relative power of women within the household, decision making and domestic violence. Smith et al. built two indices, one about relative power of women within the household and another about relative power of women outside of the household (societal gender equality). As we investigate the influence of underlying determinants (and not basic determinants), we didn't include an index about societal gender equality. We used the PCA to derive the weights, whereas Smith et al. used factor analysis.

Table 6 Asset variables for women's status index.

Indicators of women's status	Weight a_i	Mean b_i	Std. dev. s_i
Woman's position			
Woman works for cash income	–0.01479	0.17449	0.37954
Woman's age at first marriage	0.72535	17.64164	3.35581
Woman's and her partner's age difference (in %)	0.63161	–17.40122	11.14080
Difference in the woman's and her partner's years of education	0.27336	–2.48114	4.16671
Decision making and necessity of partner's permission			
Decision making regarding ...			
... medical care	–0.39697	0.50014	0.50001
... purchasing jewelry	–0.38190	0.48529	0.49979
Partner's permission needed for ...			
... visit of market	–0.59405	0.74324	0.43685
... visit of friends and relatives	–0.58624	0.79681	0.40238
Domestic violence			
Woman's opinion about domestic violence	–0.30728	0.57742	0.49398
Woman has been beaten by her partner	–0.69067	0.16762	0.37354
Frequency of beating in last 12 months	–0.65465	0.02708	0.16233

The index about the women's status B_j is calculated as

$$B_j = \sum_{i=1}^{n=11} \frac{a_i(b_{ji} - b_i)}{s_i} = \frac{a_1(b_{j1} - b_1)}{s_1} + \ldots + \frac{a_{11}(b_{j11} - b_{11})}{s_{11}},$$

where a_i is the weight (scoring factor) for the i-th variable, b_{ji} is the value of the i-the variable of the mother of the j-th child, b_i is the mean of the variable i calculated over all mothers and s_i is the standard deviation of the variable i calculated over all mothers.

Table 6 shows the components for the women's status index. Also here, the effects are all as expected. Only the influence of women working for cash income is surprisingly small and negative suggesting that this factor is not highly correlated with the overall women's status index.

5 Statistical Methodology - Semiparametric Regression Analysis

The data are given in the form (Z_i, x_i, v_i, s_i), $i = 1, \ldots, n$, where Z is the continuous variable of primary interest, x is a vector of continuous covariates with possibly nonlinear effects on Z, v is a vector of mostly categorical covariates with linear effects on Z, and s is a spatial index that indicates the geographical region the observation pertains to. In our application Z corresponds to the Z-score as a measure of chronic undernutrition, x includes the continuous variables listed in Table 4, v includes the categorical variables listed in Table 2, and s is the district in India where the mother and her family lives.

Traditional linear models assume that the effects of the covariates in x are linear. Nonlinear effects of covariates can be handled via variable transformation or polynomial regression, but the approach is rather cumbersome and time consuming. Moreover, complicated nonlinear functional forms cannot be detected within the traditional linear modeling framework.

In this paper we apply modern semiparametric regression techniques that can handle the following nonstandard requirements:

- Nonlinear covariate effects can be estimated in a nonparametric and therefore automated way. Even complex nonlinear functions can be detected with the approach.
- Complex interactions between covariates are easily incorporated into the models. In our application we will estimate a two dimensional nonparametric effect of the age of the child and the duration of breastfeeding. This type of modeling is necessary because the effect of breastfeeding depends on the age of the child. For instance, a duration of breastfeeding of 2 months should have a different effect for a child that is two months old compared to a child that is 2 years old.
- The approach is also able to deal with spatial heterogeneity either by incorporating an additional (smooth) spatial effect or by spatially smoothing the residuals after estimation.
- Model choice and selection of relevant effects is enhanced by simultaneous selection of covariates and estimation of regression parameters. The methodology used is able to

  – decide whether a particular covariate enters the model,
  – decide whether a continuous covariate enters the model linearly or nonlinearly,
  – decide whether a spatial effect enters the model,
  – select complex interaction effects between the sex of the child and other covariates
  – select the degree of smoothness of nonlinear covariate, spatial or interaction effects.

  Selection of relevant terms is particularly important because there is a lack of economic theory suggesting which of the determinants of undernutrition interact with sex.

Usually, many of the estimated covariate effects have a particular simple functional form. Even linear effects are frequently estimated. Sometimes it is argued that the classical parametric approach might do the job equally well. Note, however, that it is not sufficient to be able to reproduce the final model. The statistical model must be rich enough to be able to detect everything that *might* happen in theory. As an example take the effect of the age of the child. The literature suggests a particularly complex nonlinear relationship. For the first 4-6 months it is assumed that on average the nutritional status is comparable to the reference population. Thereafter the nutritional status worsens considerably and stabilizes at a lower level after 24 months. Within a classical linear model the assumed effect cannot be modeled.

In this paper we proceed largely in three steps:

1. In a first step we estimate (for both sexes) the model

$$Z_i = f(s_i) + \varepsilon_i,$$

where $f(s_i)$ is a smooth spatial effect of the district in India where the mother and her child lives. The estimated function $\hat{f}(s)$ is then an estimate of the mean Z-score in district s. The effect will be smooth because the estimation technique guarantees that estimates of neighboring districts are more alike than estimates of districts that are far away.
2. In a second step we specify the semiparametric model

$$\begin{aligned} Z = {} & f_1(ageC, bfmC) + g_1(ageC, bfmC) \cdot sex + \\ & f_2(agebirM) + g_2(agebirM) \cdot sex + f_3(bmiM) + g_3(bmiM) \cdot sex + \\ & f_4(heightM) + g_4(heightM) \cdot sex + f_5(womstatM) + g_5(womstatM) \cdot sex + \\ & f_6(ecstatM) + g_6(ecstatM) \cdot sex + f_7(eduC) + f_7(eduC) \cdot sex + \\ & \gamma_0 + \gamma_1\, sex + \ldots + \varepsilon, \end{aligned}$$

where $f_1 - f_7$ are smooth (possibly) nonlinear functions of the covariates $ageC - eduC$, $g_1 - g_7$ are nonlinear interaction effects with sex, and $\gamma_0 + \gamma_1\, sex + \ldots$ are effects of sex and other categorical covariates including the respective interactions with sex. Functions f_1 and g_1 are two-dimensional smooth functions of *ageC* and *bfmC*. The effect of *ageC* and *bfmC* is specified in an non-additive manner because an interaction effect between the two determinants is likely a priori. All functions are smooth in the sense that they are continuous and differentiable. A specific functional form (e.g. linear or quadratic) is not assumed. The interpretation of f_j and g_j depends on the coding of the binary variable *sex*. Since *sex* is in effect-coding, the functions f_j can be interpreted as average effects of the respective covariates, and g_j respectively $-g_j$ is the deviation from f_j for $sex = 1$ (females) and $sex = -1$ (males), respectively. This is an example of models with structured additive predictor (Fahrmeir, Kneib & Lang 2004) because the effects of covariates are additive (as in the traditional linear model) but nonlinear.
The estimation approach simultaneously selects relevant terms as well as the degree of nonlinearity and estimates the parameters. Model selection is done by minimizing a version of the Akaike Information Criterion (AIC). We use the modified AIC criterion of Hurvich, Simonoff & Tsai (1998) which corrects for bias in regression models. The selected model is

$$\begin{aligned} Z = {} & f_1(ageC, bfmC) + g_1(ageC, bfmC) \cdot sex + f_2(agebirM) + \\ & f_3(bmiM) + f_5(womstatM) + g_5(womstatM) \cdot sex + \\ & f_6(ecstatM) + \gamma_0 + \gamma_1\, sex + \gamma_2\, heightM + \gamma_3\, eduC + \ldots + \varepsilon, \end{aligned} \tag{2}$$

Table 7 Linear effects of *educM* and *heightM*.

Covariate	coeff.	std. dev.	95% CI
educM	0.0261	0.0027	0.0207–0.0312
heightM	0.0475	0.0017	0.0444–0.0507

which is much more parsimonious than the specified start model. Relevant interactions with sex are found for the continuous variables *ageC*, *bfmC*, *womstatM* and the categorical variables *twinC*, *tetanusM*, *toiletH* and *religM*. The effects of *heightM* and *eduC* are linear rather than nonlinear as specified a priori, and they show expected effects (see Table 7). As will be seen in the next section, the results, particulary the interaction effects, show two general findings: the competition for resources worsens the nutritional status of girls compared to boys and boys are particularly affected by health behavior. In order to see whether the interactions of some of the covariates with sex aid our interpretation we present in section 6 the results of a slightly modified version of model (2). The presented model additionally includes interactions of sex with the categorical covariates *birthinC*, *firstmC*, *plannedC* and *hhsizeH*.

3. In a last step we spatially smooth the residuals $\hat{\varepsilon}_i$ of the preceding semiparametric model, i.e. we estimate the model

$$\hat{\varepsilon}_i = f(s_i) + u_i.$$

By comparing the spatial effect of the residuals with the spatial effect of the first step we are able to assess how much of the spatial variation of undernutrition can be explained by the covariates in the regression of the second step.

Details about the semiparametric regression models used in this paper are given in Fahrmeir, Kneib & Lang (2004), Lang & Brezger (2004), Brezger & Lang (2006) and particularly Belitz & Lang (2008). Good introductions to semiparametric regression models are the monographs by Fahrmeir & Tutz (2001), Ruppert, Wand & Carroll (2003), Wood (2006) and in German by Fahrmeir, Kneib & Lang (2009). We used the software package BayesX for estimation, see Belitz, Brezger, Kneib & Lang (2009). The homepage of BayesX contains also a number of tutorials, see

`http://www.stat.uni-muenchen.de/~bayesx/`

6 Results

Since we have three types of variables, the results are presented in three different forms below. We first discuss the linear effects which are shown in Figure 3. The visual presentation of linear effects proved to be superior to the usual regression output because the interpretation of the interaction effects is greatly facilitated. As

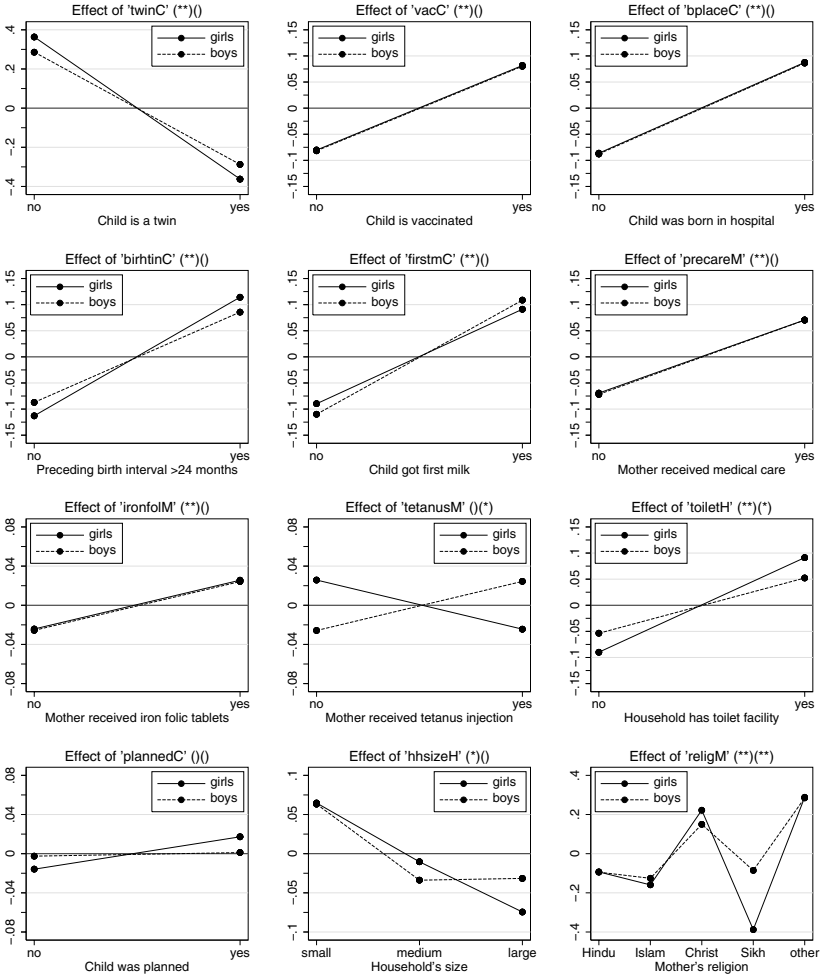


Fig. 3 Effects of categorical covariates. Symbols included in brackets indicate the significance of effects. The two brackets correspond to the respective main effects and the interaction effects. Empty brackets indicate that the respective effect is not included in the model. One, respectively two stars indicate significance at a level of 20%, respectively 5%.

the dependent variable is the Z-score, a negative coefficient means that the covariate effect lowers the nutritional status of the child, while a positive one increases it.

Many of the effects are as expected and in the same direction for boys and girls. In particular, being a twin, having a short preceding birth interval, living in a large household, not being breastfed immediately after birth, and having poor access to prenatal care is all associated with poorer nutrition, as is being born to a shorter mother (indicating a genetic transmission as well as possibly pointing to inter-generational

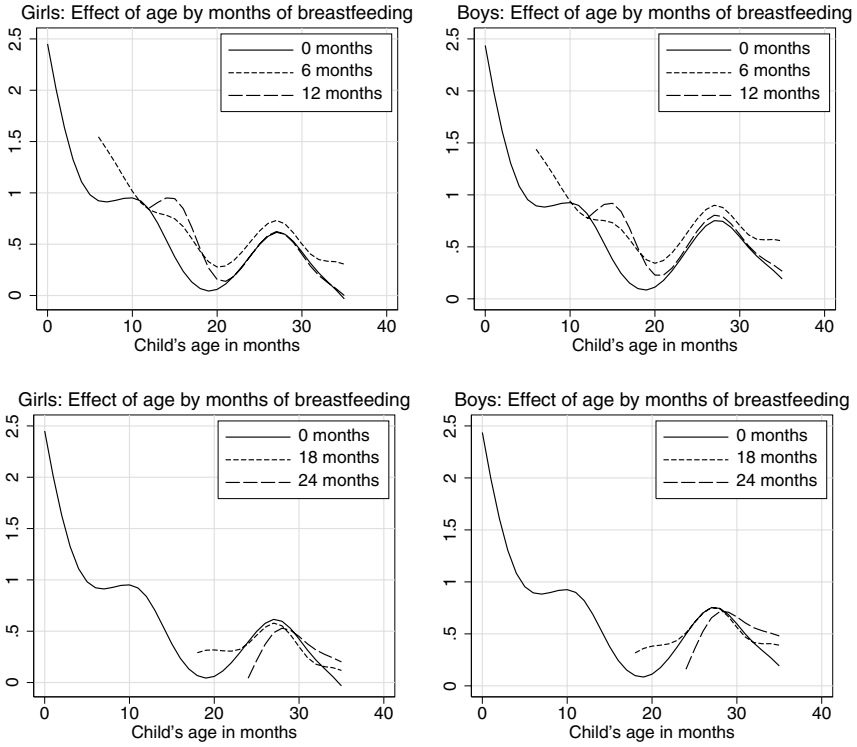


Fig. 4 Nonlinear effects of the age of the child for different durations of breastfeeding.

transmission of economic status). There is a rather strong effect of parental religion with children of Christian, Hindu, and other religions being better nourished, while Muslims and Sikh are worse nourished, suggesting significant cultural differences in care practises.

While the linear effects are rather similar for boys and girls, there are some notable and systematic differences. In particular, it appears that the nutritional status of girls reacts more sensitively to competition for resources within the household. The effect of being a twin, living in a large household, not being planned, and having a short preceding birth interval are more negative for girls than boys. Also, the cultural environment seems to matter more for girls than boys with stronger positive effects for Christian and other religions and stronger negative effects for Islam and Sikh. In contrast, the coefficients for boys indicate a greater vulnerability to inadequate care and nutrition practises. The timing of first milk is more important, as is the tetanus injection of the mother. This is further confirmed when examining breast-feeding patterns (see below).

The non-linear effects are shown in Figures 4 to 6. Figure 4 shows the combined effect of age and breast-feeding on nutritional status. The age effect shows that the

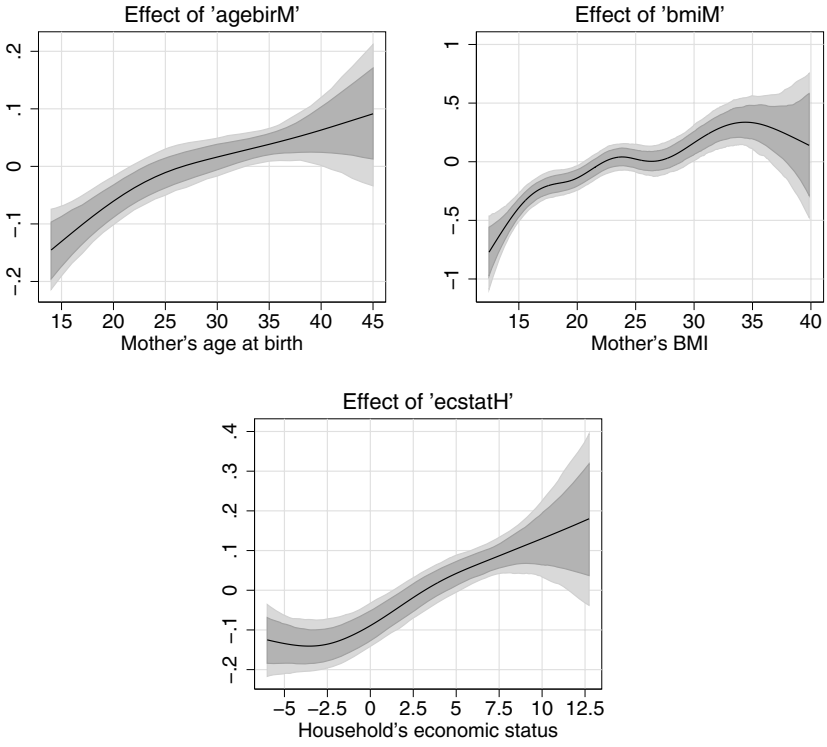


Fig. 5 Nonlinear effects of *bmiM*, *agebirthM* and *ecstatH*

nutritional status of children in India rapidly deteriorates between age 0 and about 20 months after which it oscillates. This is in line with findings from other studies (e.g. Kandala, Lang, Klasen & Fahrmeir 2001, Klasen & Moradi 2000) and indicates that children are not born chronically malnourished but develop this as a result of disease and inadequate nutritional intake. The sudden improvement of the nutritional status around 24 months is an artifact of the reference standard as at this age, children switch from being compared to the better nourished reference children from the white, bottle-fed Fels study, to the worse nourished reference children derived from a cross-section of the US population. It thus represents no real improvement and was one of the reasons for the development of a new reference standard.

The different curves for children with different breast-feeding durations are also instructive and support the greater sensitivity of boys to breast-feeding. Boys that are breastfed for 6 or twelve months have a better nutritional status throughout, while the effect for girls is weaker. Long breast-feeding durations (18 or 24 months) carry no benefits, however, and are probably an indicator of poor availability of alternative nutrition.

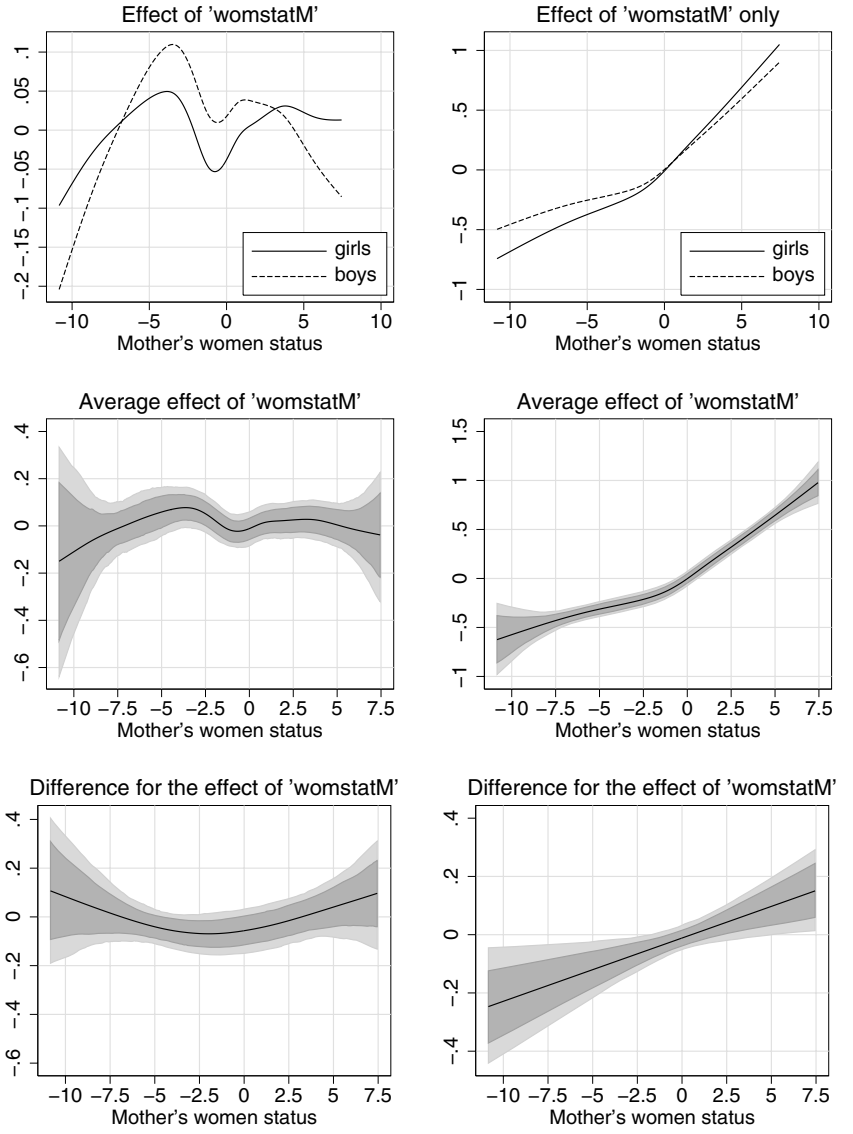


Fig. 6 Nonlinear effects of women's status.

As shown in Figure 5, there are strong effects of mother's age at birth, her BMI, as well as the household's economic status on the nutrition of her child. These effects did not differ significantly for the two sexes.

The effects of the women's status (Figure 6) variable is surprising. For girls, it shows a U-shape, for boys a more or less linear decline. These results should,

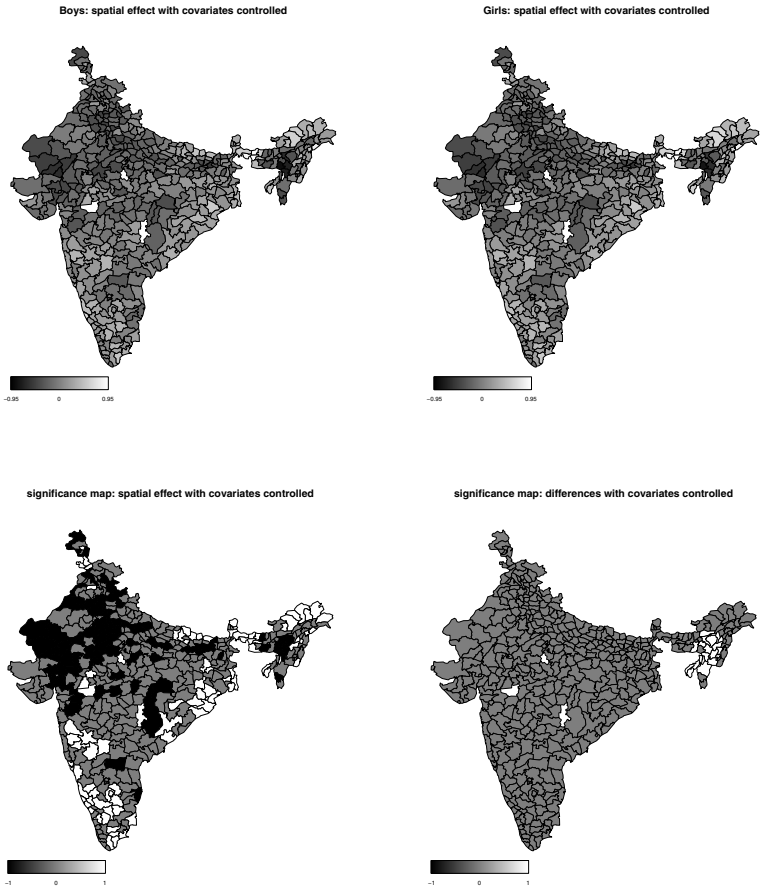


Fig. 7 Top panel: Spatially smoothed residuals. Bottom left panel: Significance map for spatial effect. Regions with average residuals significantly above (below) zero are colored in white (black). Bottom right panel: Significance map of the difference residuals between boys and girls. White (black) denotes regions with significant higher (lower) average residuals for girls compared to boys.

however, be treated with caution. Women's status is highly correlated with other covariates used in the regression and in fact, if one just considers the univariate impact of women's status on the stunting Z-score, the effect is strongly positive for both girls and boys (with a stronger effect for girls, see the right-hand panel of Figure 6). Thus women's relative status has a positive impact, but this is mediated via the other effects. The conditional direct effect (i.e. after controlling for the other covariates) is only positive for high relative women's status for girls, and negative

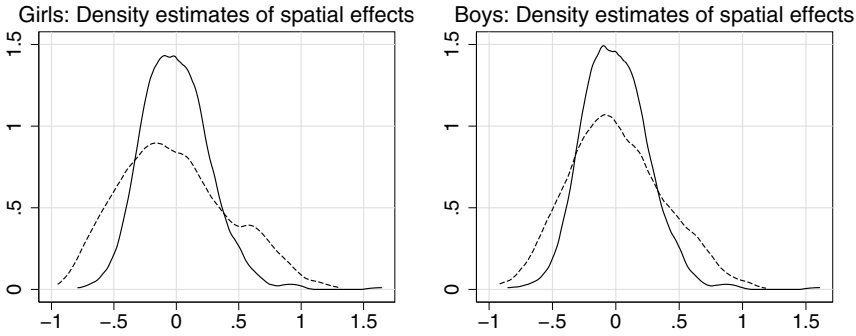


Fig. 8 Density estimates of the spatially smooth effects. The solid lines correspond to the spatial model for the residuals of the regression, i.e. covariate effects are controlled. The dashed lines correspond to the spatial model for the Z-score without controlling for covariates.

for boys which seems plausible if one can assume that high status mother's exhibit, *ceteris paribus*, a preference for favoring their daughters.

We then examine the structure of spatial residuals after controlling for covariates to see whether we have been able to explain the spatial pattern of undernutrition. So one can compare the maps of undernutrition before controlling for covariates (Figure 1) with the one after controlling for covariates (Figure 7). Before discussing the details of this comparison, it appears that the before and after maps look quite similar and thus might suggest that our empirical model has not been able to capture the spatial differential very well. While this is partially the case and will be discussed presently, Figure 8 shows that we have been able to significantly reduce the spatial residuals through our empirical model. Compared to the distribution of spatial effects before using covariates (dotted line), the solid line shows a much tighter distribution of the district-level spatial residuals, suggesting that we have been able to significantly explain the spatial distribution of undernutrition and thus reduce the spatial residuals. When examining the spatial pattern of the residuals, it becomes apparent that the spatial pattern of mother's education, women's status, mother's BMI, and household economic status significantly contributed to explaining the spatial disparity in undernutrition.

Nevertheless, we have to admit that a significant spatial pattern of undernutrition remains and the resulting spatial pattern looks, at first glance, similar to the spatial pattern observed before the use of covariates. But apart from the overall reduction of these spatial effects, there are also some notable shifts in the residual spatial pattern. In particular, the areas of unexplained poor nutritional status have now shifted from the Central-North to the North-West, i.e. from Uttar Pradesh, Bihar, Jarkhand, and Madhia Pradesh, to Rajasthan, Haryana, and Uttaranchal. Conversely, new areas of 'better than expected' female nutrition appear in the East, in West Bengal and parts of Orissa, while undernutrition in Assam, Manipur, Mizoram, and Triupura are no longer better than expected.

What are we to make of these significant residual spatial effects? They are unlikely to be due to the usual arguments advanced by scholars of regional differences in India, such as different female roles and differential female autonomy, or different public action in the fields of education, health and nutrition, or different religions (e.g. Agarwal 1994, Dreze & Sen 2001, Dyson & Moore 1983 , Klasen & Wink 2003) as we have tried to control for these effects, to the extent possible, explicitly through our covariates. One possible explanation could be that our covariates are insufficiently capturing these differential, for example by neglecting the quality of education and health services, or the success of other public interventions in improving nutrition and health of children.[9]

A second possible explanation is that certain aspects of public commitment and public activism are not sufficiently captured by our variables. In particular, it is notable that public activism for health, education, land reform, and inequality, have been particularly strong in the Indian state of Kerala but also in West Bengal, the two states with the strongest remaining positive effects. Conversely, the areas of significantly poorer than expected performance are concentrated in areas which recently witnessed the rise of Hindu nationalism, the ascendancy of the Hindu nationalist BJP to political prominence, and related incidences of communal violence between Muslims and Hindus.[10]. A third possible explanation is that some cultural institutions that affect the treatment of children are not closely correlated with religious affiliation or our measures of female autonomy and might therefore account for the remaining regional pattern.[11] Lastly, there could be climatic factors that help to explain these different patterns of undernutrition. With all four explanations, it might be the case that they have a larger impact on the treatment of female than on male children and can thus explain the stronger residual spatial pattern for girls. We do not have the data available to investigate these hypotheses which we hope will stimulate further analysis of these remaining spatial patterns of undernutrition.

Regarding the spatial pattern of the sex differences in undernutrition, our model seems to perform very well. As shown in the lower right map on Figure 3, there are hardly any significant sex differences remaining, except for a few districts in Tripura, Mizoram, and Manipur. Thus it appears that we are fully able to account for the spatial pattern of sex difference in undernutrition which are largely driven by the gender-specific effects regarding competition and care that we discussed above.

7 Conclusion

In this paper, we used geoadditive semiparametric regression models to study the determinants of chronic undernutrition of boys and girls in India in 1998/99. A particular focus of our paper was to explain the strong regional pattern in undernutrition

[9] See Dreze & Sen (2001) for a discussion of these issues.

[10] For more details, see Dreze & Sen (2001), Sen (1998) and Sen (2005)

[11] See, for example, Basu (1992) and Basu & Jeffery (1996) for a discussion.

and sex differences in determinants and regional pattern of undernutrition. We find that determinants associated with competition for household resources and cultural factors are more important for the nutrition of girls than boys, while boys' nutrition reacts more sensitively to nutrition and medical care access. With our models we are able to explain a large portion of the spatial pattern of undernutrition, but significant spatial effects remain, with the South-West and East having significantly lower, and the North-West significantly higher undernutrition rates. The remaining spatial patterns, that are slightly different for boys than for girls, are intriguing and call for more detailed analysis which were not possible with our data set.

Acknowledgements We would like to thank Monica Das Gupta, Davic Sahn, Paul Schultz, Amartya Sen, and Lisa Smitz as well as participants at workshops in Munich, Tübingen, IFPRI, and Göttingen for helpful comments and discussion. Funding from the German Research Foundation is gratefully acknowledged. We also thank the IIPS for giving us access to the district location of households.

References

Agarwal, B. (1994). *A field of one's own*, Cambridge University Press.

Basu, A.M. (1992). *Culture, the Status of Women, and Demographic Behavior*, Oxford University Press.

Basu, A.M. & Jeffery, R. (1996). *Girl's Schooling, Women's Autonomy and Fertility Change in South Asia*, Sage.

Belitz, C., Brezger, A., Kneib, T. & Lang, S. (2009). BayesX Manuals.

Belitz, C. & Lang, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics and Data Analysis* **53**: 61–81.

Brezger, A. & Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis* **50**: 967–991.

Caputo, A., Foraita, R., Klasen, S. & Pigeot, I. (2003). Undernutrition in Benin : An Analysis based on Graphical Models, *Social Science and Medicine* **56**: 1677–1691.

Dreze, J. & Sen, A. (1995). *India. Economic Development and Social Opportunity*, Oxford University Press.

Dreze, J. & Sen, A. (2001): *India. Development and Participation*, Oxford University Press.

Dyson, T. & Moore, M. (1983). On Kinship Structure, Female Autonomy, and Demographic Behavior in India, *Population and Development Review* **9**: 35–57.

Fahrmeir, L., Kneib, Th. & Lang, S. (2004). Penalized additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**: 715-745.

Fahrmeir, L., Kneib, T. & Lang, S. (2009). *Regression. Modelle, Methoden und Anwendungen*, Springer–Verlag.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*, Springer–Verlag, New York.

Filmer, D. & Pritchett, L. (1998). Estimating Wealth Effects without Expenditure Data – Tears: An Application to Educational Enrollments in States of India, *World Bank Policy Research Working Paper No. 1994*, Development Economics Research Group.

International Institute for Population Sciences (1998). *National Family Health Survey (NFHS-2), 1998-99*, IIPS. Mumbai, India.

International Institute for Population Sciences (2000). *National Family Health Survey (NFHS-2), 1998-99*, IIPS. Mumbai, India.

Harttgen, K. & Misselhorn, M. (2006). A Mulitlevel Approach to Explain Child Mortality and Undernutrition in South Asia and Sub-Saharan Africa, *Discussion paper, University of Göttingen*.

Hurvich, C.M., Simonoff, J.S. & Tsai, C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society B* **60**: 271–293.

Kandala, N.B., Fahrmeir, L., Klasen, S. & Priebe, J. (2008). Geo-additive models of childhood undernutrition in three Sub-Saharan African countries. *Population, Space and Place*, **14**.

Kandala, N.B., Lang, S., Klasen, S. & Fahrmeir, L. (2001). Semiparametric Analysis of the Socio-Demographic and Spatial Determinants of Undernutrition in Two African Countries. *Research in Official Statistics* **1**: 81-100.

Kandala, N.B., Lang, S. & Klasen, S. (2001). Semiparametric Analysis of Childhood Undernutrition in Developing Countries. *In* George, E. (ed), *Monographs of Official Statistics; Bayesian methods with applications to science, policy and official statistics*.

Klasen, S. (2008). Poverty, undernutrition, and child mortality: Some inter-regional puzzles and their implications for research and policy, *Journal of Economic Inequality* **6**: 89–115.

Klasen, S. & Moradi, A. (2000). The Nutritional Status of Elites in India, Kenya and Zambia: An appropriate guide for developing reference standards for undernutrition?, *Sonderforschungsbereich 386: Discussion Paper No. 217.*

Klasen, S. & Wink, C. (2002). A Turning Point in Gender Bias in Mortality? An Update on the Number of Missing Women, *Population and Development Review* **28**: 285–312.

Klasen, S. & Wink, C. (2003). Missing Women: Revisiting the Debate, *Feminist Economics* **9**: 263–299.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**: 183-212.

Mishra, V.K., Lahiri, S. & Luther, N.Y. (1999). Child Nutrition in India. *National Family Health Survey Subject Reports Number 14*, International Institute for Population Sciences.

Murthi, M., Guio, A.C., & Dreze, J. (1995). Mortality, Fertility, and Gender Bias in India: A District-Level Analysis. *Population and Development Review*, **21**: 745–782.

de Onis, M., de Frongillo, E.A. & Blossner, M. (2000). Is malnutrition declining? An analysis of changes in levels of child malnutrition since 1980. *Bulletin ot the World Health Organisation* **78**.

Osmani, S.R. (1990). Nutrition and the Economics of Food: Implications of Some Recent Controversies. *in* Dreze, J. & Sen, A. (eds), *The Political Economy of Hunger*, Springer–Verlag, 10–13.

Pelletier, D. (1994). The relationship between child anthropometry and mortality in developing countries, *Journal of Nutrition* **124**: 2047S–2081S.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge University Press.

Sen, A. (1998). *Development as Freedom*, Knopf, New York.

Sen, A. (2003). Missing Women-revisited, *British Medical Journal*, **327**: 1297.

Sen, A. (2005). *The argumentative Indian*, Allyn Lane, London.

Smith, L.C., Ramakrishnan, U., Haddad, L., Martorell, R. & Ndiaye, A. (2003). The Importance of Women's Status for Child Nutrition in Developing Countries, *IFPRI Research Report No. 131*, International Food Policy Research Institute. Washington DC, USA.

Smith, L.C. & Haddad, L. (1999). Explaining Child Malnutrition in Developing Countries: A Cross-Country Analysis, *FCND Discussion Paper No. 60*, Food Consumption and Nutrition Division, International Food Policy Research Institute, USA.

Svedberg, P. (2002). Hunger in India- Facts and Challenges, *Little Magazine*.

UNICEF (1998). *The State of the World's Children. Focus on Nutrition*, Oxford University Press.

WHO (2002). *Global Database on Child Growth and Malnutrition*, WHO, Department of Nutrition for Health and Development.

WHO (2006). WHO Child Growth Standards based on length/height, weight, and age, *Acta Paediatrica Supplement* **450**: 76–85.

World Bank (2001). *Engendering Development*, The World Bank, Washington DC.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall.

Boosting for Estimating Spatially Structured Additive Models

Nikolay Robinzonov and Torsten Hothorn

Abstract Spatially structured additive models offer the flexibility to estimate regression relationships for spatially and temporally correlated data. Here, we focus on the estimation of conditional deer browsing probabilities in the National Park "Bayerischer Wald". The models are fitted using a componentwise boosting algorithm. Smooth and non-smooth base learners for the spatial component of the models are compared. A benchmark comparison indicates that browsing intensities may be best described by non-smooth base learners allowing for abrupt changes in the regression relationship.

1 Introduction

Biological diversity and forest health are major contributors to the ecological and economical prosperity of a country. This is what makes the conversion of mono-species into mixed-species forests an important concern of forest management and policy in Central Europe (Knoke et al. 2008). Recent research shows not only positive ecological effects of mixed-species forests (e.g. Fritz 2006) but also positive economic consequences (Knoke & Seifert 2008). Like any other living environment, the development of forests is strongly conditioned on a balanced and consistent regeneration. Whether natural or artificial, the regeneration is challenged at a very early stage by browsing damage caused by various game species. In middle Europe, especially, roe and red deer are the most common species accountable for browsing on young trees. This activity is certainly natural by definition. However, the eradication of large predators, the conversion of the landscape and the fostering of trophy animals have given rise to increased number of deer and subsequently to

Nikolay Robinzonov and Torsten Hothorn
Institut für Statistik, Ludwigstraße 33, Ludwig-Maximilians-Universität München, Germany,
e-mail: nikolay.robinzonov@stat.uni-muenchen.de,
Torsten.Hothorn@stat.uni-muenchen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_10,

Fig. 1 The National Park "Bayerischer Wald". The southern grey colored region is the district of "Rachel-Lusen" where our studies take part.

intensified browsing pressure in the past centuries. The consequences of excessive browsing activity often lead to forest growth retardation and homogenization (Eiberle & Nigg 1987, Eiberle 1989, Ammer 1996, Motta 2003, Weisberg et al. 2005).

Forest regeneration is monitored on a regular basis by the Bavarian Forest Administration (Forstliches Gutachten 2006). This Bavarian-wide survey is conducted every three years and takes place in all 745 game management districts (Hegegemeinschaften) in Bavaria. Preventive measures are proposed following the survey's results. In case of an estimated browsing quota above the specified thresholds, the local authorities consider a protection of the most vulnerable areas. An often used practice is to recommend for intensified deer harvesting in the corresponding areas. Whether the impact of game on the forest regeneration is correctly measured remains a matter of debate (e.g. Prien 1997, Rüegg 1999, Moog 2008). Developing precise measures which reflect the true condition of the forest's regeneration is thus crucial and non-trivial.

Our focus is on surveys conducted to estimate the local conditional probability of a young tree to be affected by deer browsing, as recommended for monitoring of the influence of game on forest regeneration (Rüegg 1999). For the beech species (*Fagus sylvatica*) to be found in a certain area, this quantity reflects the exposure to deer browsing and is the basis for subsequent management decisions. Here, we are concerned with the estimation of such conditional browsing probabilities.

We evaluate and compare boosting algorithms for fitting structured additive models (Fahrmeir et al. 2004) to deer browsing intensities. This article aims to make in brief space a comparison of smooth and non-smooth model components for capturing the spatio-temporal variation in such data. Our investigations are based on two surveys conducted in 1991 and in 2002 in the district of "Rachel-Lusen", the southern part of the National Park "Bayerischer Wald" depicted in Figure 1.

2 Methods

The main purpose of a deer browsing survey is to estimate the probability of deer browsing on young trees. More specifically, the conditional probability of a young tree of a certain species at a given location to suffer from deer browsing is the quantity of interest. The tree height is an important exploratory variable for deer browsing and thus needs to be included in the model. In addition, unobserved heterogeneity in the browsing damage will be considered by allowing for *spatial* and *spatio-temporal* components to enter the model. Commonly, other covariates describing the forest ecosystem are not measured and are thus not included in our investigations. For the sake of simplicity, we restrict our attention to beeches.

The general idea of our modelling strategy is as follows. The logit-transformed conditional probability of browsing damage is linked to the tree height and spatial and spatio-temporal effects by the regression function f such that

$$\begin{aligned} \operatorname{logit}(\mathbb{P}(Y = 1 | \mathsf{height}, \mathsf{space}, \mathsf{time})) &= f(\mathsf{height}, \mathsf{space}, \mathsf{time}) \\ &= f_{height}(\mathsf{height}) + f_{spatial}(\mathsf{space}) \\ &\quad + f_{spatemp}(\mathsf{space}, \mathsf{time}) \end{aligned} \tag{1}$$

where the predictor space stays for a two-dimensional covariate of northing and easting, height is a one-dimensional continuous variable representing trees' height and time is an ordered factor with levels 1991 and 2002.

Therefore, we differentiate between three types of variability: such caused by the trees' height and captured by f_{height}, solely spatial variability explained by the two-dimensional smooth function $f_{spatial}$ and time-dependent heterogeneity modelled by the multi-dimensional smooth function $f_{spatemp}$. Thus far we have sketched model (1) for a general view of our estimation strategy. In the subsequent chapters we consider the component pieces of three possible approaches meant to accomplish this strategy.

2.1 Spatio-Temporal Structured Additive Models

The next two methods originate from the family of ensemble construction algorithms. The original idea of ensemble methods is to use reweighted original data to obtain

a linear combination of some model fittings methods (Bühlmann 2004). We interchangeably refer to these fitting methods as *base procedures* or *weak learners*. The literature on ensemble methods is diverse and wide-ranging, but the two prominent classes that draw our attention are *bagging* (Breiman 1996) (or its successor *random forests*, Breiman 2001) and *boosting* (e.g. Bühlmann & Hothorn 2007, Hastie et al. 2009). Although bagging does not directly strengthen our case, it is worth understanding the difference between them. Bagging is a *parallel* ensemble method which averages the coefficients of the whole ensemble, while boosting methods (also called *incremental learning*) are sequential ensemble algorithms which update the coefficients iteratively. All these ideas have merits and demerits but in contrast to bagging, boosting retains the especially useful initial structure of the base procedures, hence allowing for better interpretation.

Both of our ensemble methods are boosting techniques which solely differ in the choice of their spatial and spatio-temporal base procedures. The first boosting method is a structured additive regression (GAMBoost) model for describing the probability of browsing damage:

$$\begin{aligned}\text{logit}(\mathbb{P}(Y=1|\mathsf{height},\mathsf{space},\mathsf{time})) &= f_{str}(\mathsf{height},\mathsf{space},\mathsf{time}) \\ &= f_{bheight}(\mathsf{height}) + f_{bspatial}(\mathsf{space}) \\ &\quad + f_{bspatemp}(\mathsf{space},\mathsf{time})\end{aligned} \tag{2}$$

where $f_{bheight}$ is a an additively structured, P-Spline function of height, $f_{bspatial}$ is an additively structured, bivariate P-Spline tensor function (Kneib et al. 2009) of easting and northing or shortly space and $f_{bspatemp}$ is essentially the same as $f_{bspatial}$ but applied only for the year 2002 (see (7) below). The objective is to obtain an estimate $\hat{f}_{str}$ of the function f_{str}. In theory, this approximation is commonly based on the expectation of some prespecified loss function $L(y,\pi(f_{str}))$, in practice we aim at minimizing its empirical version

$$\hat{f}_{str} = \arg\min_{f_{str}} \frac{1}{n}\sum_{i=1}^{n} L(y_i,\pi_i(f_{str})) \tag{3}$$

where $\pi_i(f_{str}) = \text{logit}^{-1}(f_{str}(\mathsf{height}_i,\mathsf{space}_i,\mathsf{time}_i))$ denotes the inverse of the logit function. A discussion of the specification of several loss functions can be found in Hastie et al. (2009, chap. 10), Bühlmann & Hothorn (2007), Friedman (2001) and in Lutz et al. (2008). We aim at minimizing the negative log-likelihood

$$L(y_i,\pi_i(f_{str})) = -(y_i \log(\pi_i(f_{str})) + (1-y_i)\log(1-\pi_i(f_{str}))). \tag{4}$$

As mentioned above, each function in (2) has an additive structure which means in particular that the model can be decomposed in

$$\hat{f}_{bheight}(\text{height}) = \nu \sum_{m=0}^{M} \hat{h}^{[m]}_{height}(\text{height}) \tag{5}$$

$$\hat{f}_{bspatial}(\text{space}) = \nu \sum_{m=0}^{M} \hat{h}^{[m]}_{spatial}(\text{space}) \tag{6}$$

$$\begin{aligned} \hat{f}_{bspatemp}(\text{space},\text{time}) &= \nu \sum_{m=0}^{M} \hat{h}^{[m]}_{spatemp}(\text{space},\text{time}) \\ &= \begin{cases} \nu \sum\limits_{m=0}^{M} \hat{h}^{[m]}_{spatial}(\text{space}), & \text{time} = 2002, \\ 0, & \text{time} = 1991, \end{cases} \end{aligned} \tag{7}$$

where the weak learner $\hat{h}^{[m]}_{height}$ is a smooth penalized B-Spline function (P-Spline, Eilers & Marx 1996), $\hat{h}^{[m]}_{spatial}$ is a smooth bivariate P-Spline based surface and $\nu \in (0,1)$ is a *shrinkage* parameter (Friedman 2001) or *step size*. Thus, our choice of weak learners are basis expansions themselves which means that

$$\hat{h}^{[m]}_{height}(\text{height}) = \sum_{k=1}^{K} \hat{\gamma}^{[m]}_{height,k} b_k(\text{height}) \tag{8}$$

$$\hat{h}^{[m]}_{spatial}(\text{space}) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \hat{\gamma}^{[m]}_{space_{k_1,k_2}} b_{k_1,k_2}(\text{space}) \tag{9}$$

where the b_k's represent K completely known univariate basis functions, b_{k_1,k_2} tensor product functions with

$$b_{k_1,k_2}(\text{space}) = b_{k_1,k_2}(\text{easting},\text{northing}) = b_{k_1}(\text{easting}) b_{k_2}(\text{northing})$$

and $\hat{\gamma}^{[m]}_{height}$ and $\hat{\gamma}^{[m]}_{space}$ are regression coefficients which scale these basis functions (see Kneib et al. 2009, Wood 2006). $\hat{\gamma}^{[0]}_{height}$ and $\hat{\gamma}^{[0]}_{space}$ are arbitrarily chosen start vectors of parameters. Note that the time-dependent effect in (7) is interpreted as the spatial difference between the years 1991 and 2002. It should be further noted that K, K_1 and K_2 are known in advance (specified by the user) and M is the major tuning parameter for boosting which we discuss below.

All parameters, i.e. all $\hat{\gamma}^{[m]}$s, of this additive expansion will be determined iteratively by successively improving (updating) them and accumulating the whole estimation in $\hat{f}_{str}$. Hence, the step size ν can be thought of as an improvement penalty which prevents the model from taking the full contribution of the updates.

The minimization problem (3) is solved iteratively by *componentwise* boosting which chooses at each step the "best" base procedure amongst (5)-(7), i.e. the one that contributes to the fit most. One option to attain this is via the so called *steepest-descent* optimization which relies of the negative gradient

Componentwise boosting

1. Initialize $\hat{f}_{str} = \text{offset}$, set $m = 0$.
2. $m = m + 1$.
3. Compute the negative gradient: $g_i = -\left[\frac{\partial}{\partial f_{str}} L(y_i, \pi_i(f_{str}))\right]_{f_{str}=\hat{f}_{str}}, i = 1, \ldots, n$.
4. Fit all base procedures to the negative gradient and select the best one according to $\hat{s}_m = \underset{s \in \{intercept, height, spatial, spatemp\}}{\arg\min} = \sum_{i=1}^{n} (g_i - \hat{h}_s^{[m]})^2$.
5. Update $\hat{f}^{str} := \hat{f}^{str} + \nu \hat{h}_{\hat{s}_m}^{[m]}$.
6. Iterate 2-5 until $m = M$.

$$g_i = -\left[\frac{\partial}{\partial f_{str}} L(y_i, \pi_i(f_{str}))\right]_{f_{str}=\hat{f}_{str}}, i = 1, \ldots, n \tag{10}$$

being computed at each step and subsequently fitted against each base procedure separately, i.e. the negative gradient is used as a *pseudo-response* in each step m. The negative gradient (10) indicates the direction of the locally greatest decrease in the loss. The most "valuable" covariate has the highest correlation with the negative gradient and is therefore chosen for fitting. In this way we incorporate a simultaneous variable selection procedure.

Schmid & Hothorn (2008) carried out an extensive analysis of the main hyper-parameters' effects on boosting, such as the maximum step number M, the step size ν, the smoothing parameters for the P-Splines and the number of knots. Their results confirmed the common knowledge that there is a minimum number of necessary knots needed to capture the curvature of the function and that the algorithm is not sensitive to this choice (20-50 knots should be sufficient). They also found that $\nu = 0.1$ is a reasonable choice for the step size, whose altering interplays with the computational time only, i.e. smaller ν increases the computational burden but does not deteriorate the fitting quality. The same holds for the P-Spline smoothing parameters which essentially penalize the flexibility of the base procedure through its degrees of freedom. Choosing larger values leads to fewer degrees of freedom, which translates into larger bias but smaller variance. This follows the prescriptions of the recommended strategy for boosting (Bühlmann & Yu 2003, Schmid & Hothorn 2008). Again, reasonable altering of this parameter reflects solely in the computational time.

Aside from obtaining the stopping condition M (which will be discussed later), we are ready to summarize componentwise boosting in the following algorithm:

Researchers in many fields have found the cross-validatory assessment of tuning parameters attractive in the era of plentiful computing power. By splitting the original (training) set into k roughly equally sized parts, one can use $k-1$ parts to train the model and the last kth part to test it. This is known as a k-fold cross-validation. A known issue of cross-validation is the systematic partition of the training set rising up the risk of error patterns. That is, the training set is not a random sample from the available data, but chosen to disproportionally represent the classes, especially to

over-represent rare classes. Therefore, we alleviate this to some degree by using the *bootstrap* algorithm (Efron 1979). We perform a random sampling with replacement of the original data set, i.e. the n sample points are assumed to be multinomially distributed with equal probabilities $1/n$. After the sampling we have a new training set of size n with some sample points chosen once, some more than once and some of them being completely omitted (usually $\sim 37\%$). Those omitted sample points are regarded as our test set in order to quantify performance. We choose some large value for M, say 2000, and perform 25 bootstrapped samples with each $m = 1, \ldots, M$. The optimal m is reported according to the average out-of-sample risk, also referred to as *out-of-bag*, minimization of the loss function.

2.2 Tree Based Learners

There are regions in the National Park "Bayerischer Wald" which are not affected by deer browsing and others with disproportionally higher risk of browsing. This is due to the irregular distribution of regeneration areas in the National Park and to other environmental factors. Therefore, we might wish to reconsider the smooth relationship between the response and the predictors made so far. By putting the smooth assumption of the underlying function f_{str} under careful scrutiny we aim to improve the performance of regression setting (2). Having covariates at different scales we find *regression trees* (Breiman et al. 1983) to be an attractive way to express knowledge and aid forest decision-making. A "natural" candidate for a decision tree based learner is the spatio-temporal component due to the different scales of space and time. The spatial component space is another good option for a tree based modelling due to the probable coarse relationship between the space and the browsing probability which we suspect. We let the smooth P-Spline based learner h_{height} remain unchanged. Therefore, we have a very similar general structure to (2)

$$\begin{aligned} \text{logit}(\mathbb{P}(Y = 1|\text{height, space, time})) &= f_{bb}(\text{height, space, time}) \\ &= f_{bheight}(\text{height}) + f_{bbspatial}(\text{space}) \\ &\quad + f_{bbspatempt}(\text{space, time}) \end{aligned} \tag{11}$$

with $f_{bheight}$ being exactly the same as in (5) and modified learners

$$\hat{f}_{bbspatial}(\text{space}) = \nu \sum_{i=1}^{M} \hat{h}^{[m]}_{spatialtree}(\text{space}) \tag{12}$$

$$\hat{f}_{bbspatemp}(\text{space, time}) = \nu \sum_{i=1}^{M} \hat{h}^{[m]}_{spatemptree}(\text{space, time}). \tag{13}$$

The model (11) is referred to as a Black-Box model. We should avoid overinterpreting the result of tree based learners too much. However, the height component remains perfectly interpretable. We choose the unbiased recursive partitioning framework of

Hothorn et al. (2006) to grow binary trees. The spatial component has the additive form

$$\hat{h}^{[m]}_{spatialtree}(\text{space}) = \sum_{j=1}^{J} \hat{\gamma}^{[m]}_{spatialtree,j}\, I(\text{space} \in R^{[m]}_j) \tag{14}$$

and the spatio-temporal component is represented by

$$\hat{h}^{[m]}_{spatemptree}(\text{space}, \text{time}) = \sum_{j=1}^{J^*} \hat{\gamma}^{[m]}_{spatemptree,j}\, I((\text{space}, \text{space}) \in R^{*[m]}_j). \tag{15}$$

Here I denotes the indicator function, $R^{[m]}_j, j = 1, \ldots, J$ are disjoint regions which collectively cover the space of all joint values of the predictor variables in space (recall that $\text{space} = \{\text{easting}, \text{northing}\}$). The superscript $[m]$ in $R^{[m]}_j$ means that this region is defined by the terminal nodes of the tree at the mth boosting iteration. $R^{*[m]}_j$ are the respective regions for space and time. Thus, we compute a sequence of simple binary trees with a maximal depth of, say, five. The task at each step is to find a new tree to describe the prediction residuals (pseudo-response) of the preceding tree succinctly. The next tree will then be fitted to the new residuals and will further partition the residual variance for the data, given the preceding sequence of trees.

2.3 Generalized Additive Model

The last method under test is a Generalized Additive Model (GAM) (Hastie & Tibshirani 1990). Once we are familiar with the underlying structure of the GAMBoost model the GAM model can be seen as a simplified special case of (2) with $\nu = 1, M = 1$ and with no componentwise selection carried out. This means that we have the following structure

$$\begin{aligned} \text{logit}(\mathbb{P}(Y = 1 | \text{height, space, time})) &= f(\text{height, space, time}) \\ &= f_{height}(\text{height}) + f_{spatial}(\text{space}) \\ &\quad + f_{spatemp}(\text{space}, \text{time}) \end{aligned} \tag{16}$$

where

$$\hat{f}_{height}(\text{height}) = \sum_{k=1}^{K} \hat{\gamma}_{height,k}\, b_k(\text{height}) \tag{17}$$

$$\hat{f}_{spatial}(\text{space}) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \hat{\gamma}_{space_{k_1,k_2}}\, b_{k_1,k_2}(\text{space}) \tag{18}$$

$$\hat{f}_{spatemp}(\text{space},\text{time}) = \begin{cases} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \hat{\gamma}_{space_{k_1,k_2}} b_{k_1,k_2}(\text{space}), & \text{time} = 2002, \\ 0, & \text{time} = 1991. \end{cases} \tag{19}$$

The interpretation of the basis function b_k, b_{k_1,k_2} and their scaling parameters remains the same as in (8) and (9). A similar model to (16) has been proposed by Augustin et al. (2009, p. 11). A major difference between their model and the specification above is the time component being a continuous predictor smoothly modelled through a cubic regression spline basis functions. This is what they call a *3-d tensor product smoother for space and time*. It is also worth mentioning that their model restrain from the pure spatial component $f_{spatial}$ and relies solely on the multi-dimensional function $f_{spatemp}$ to capture the spatial variability.

3 Results

In this section we exemplify the different models on the map of the National Park. Initially we depict the fitted surfaces and denote the cases of browsing damage throughout these surfaces. In addition we perform a model comparison in order to quantify the prediction performance of the models.

3.1 Model Illustrations

In a first step we visualize the browsing probability estimates obtained by the GAMBoost model (2), the Black-Box model (11) and the common GAM as in (16). Figure 2 illustrates the estimation produced by the GAMBoost model for an average beech tree 60 cm in height. The white areas indicate regions with higher risk of browsing damage. The darker the regions, the smaller the estimated probability of browsing. Further we depict black circles proportional to the absolute number of damaged trees in the corresponding sample points of the National Park.

The GAMBoost model proposes the smoothest fit amongst all models. The model detects the risky regions in 2002 rather well, encompassing the black circles with smooth light regions and fitting the northern high-level areas to low risk probabilities. However, the relatively even empirical distribution of damaged cases in 1991 leaves the impression of too smooth surface, i.e. possible underfitting. The GAMBoost model is also an example of why fine tuning of the hyper parameters should be undertaken with greater care in the presence of tensor P-Spline base learners. The claims we made about the "informal" impact of the step size, the number of knots and the smoothing parameters do hold in this case. However, the maximum number of boosting steps markedly increases if bivariate base learners are considered. One could falsely define too small M for an upper bound of the boosting steps. Therefore, boosting would continuously improve its prediction power within the proposed values

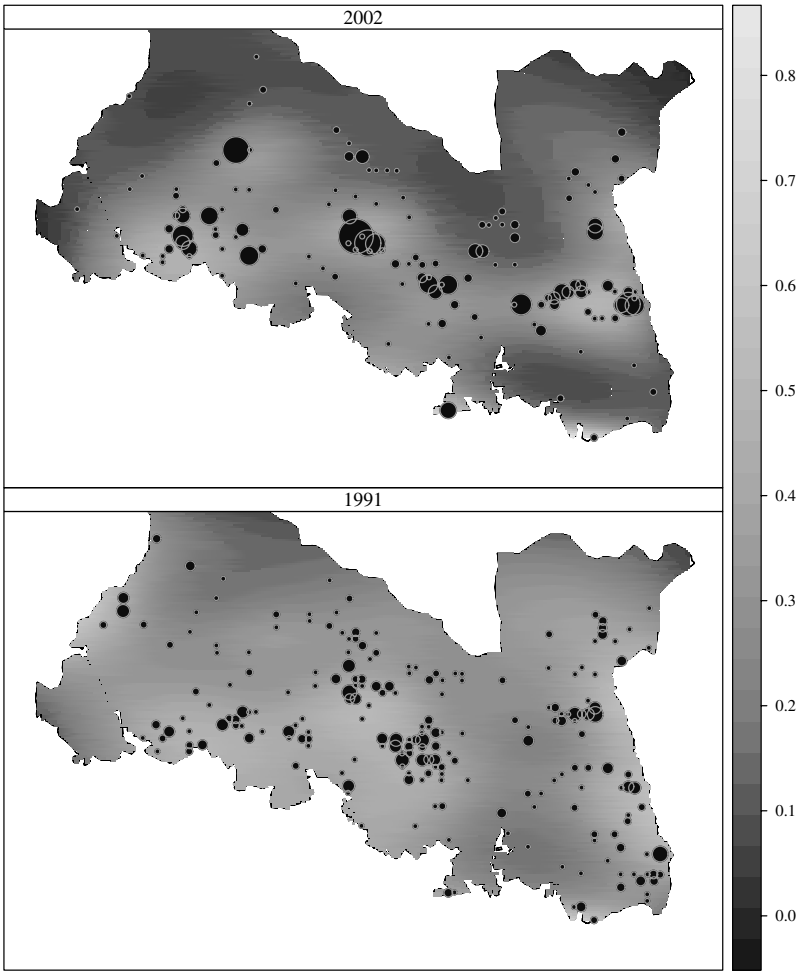


Fig. 2 Spatial component space fitted by the GAMBoost model for an average beech tree at the height of 60 cm in the years 1991 (bottom) and 2002 (top). The diameter of the black circles is proportional to the absolute number of browsed trees.

of M and will always find the optimal M at the border, i.e. at the last step. This is due to the insufficient degrees of freedom leading to a very modest amount of progress towards optimality, i.e. the optimal step number is basically never reached. Therefore, the "standard" amount for degrees of freedom df $\in (4, 6)$ for the univariate P-Spline learners, seems to be a very challenging choice for tensor P-Spline learners in terms of a reasonable computation time. We use df $= 12$ in order to speed up computations and

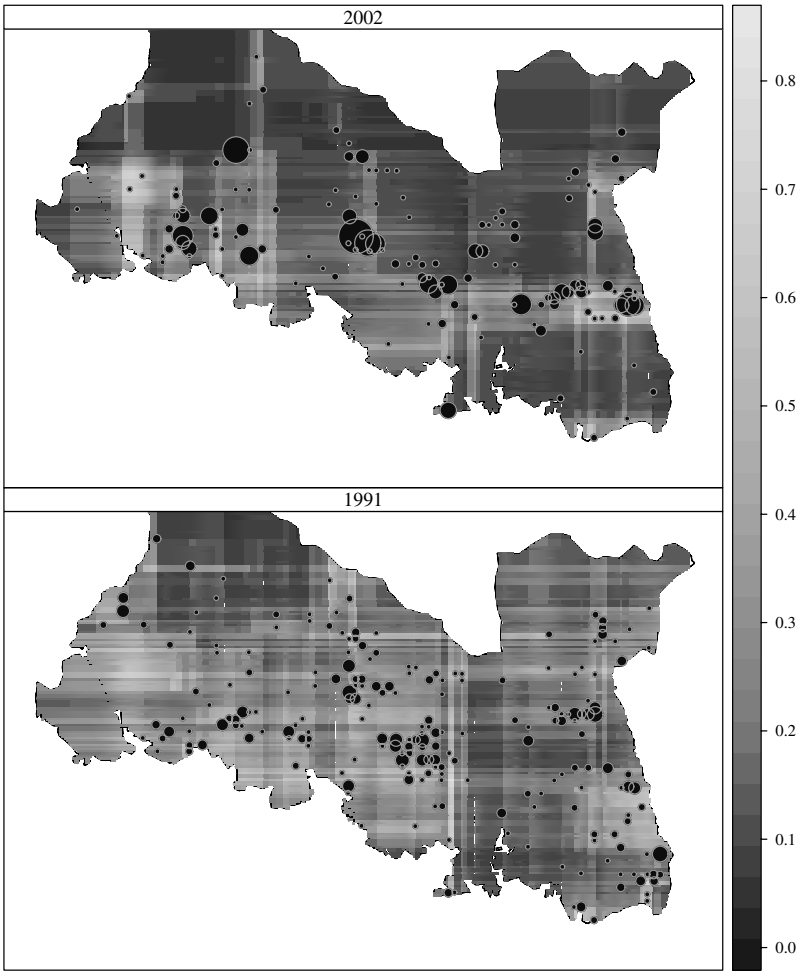


Fig. 3 Spatial Component space fitted by the Black-Box model for an average beech tree 60 cm in height in the years 1991 (bottom) and 2002 (top). The diameter of the black circles is proportional to the absolute number of browsed trees.

to ensure that $M = 2000$ is sufficient to find an optimal number of boosting iterations. Alternatively one could dampen the learning rate less severely by increasing the step size ν or altering the number of the spline knots.

Figure 3 represents the estimation produced by the Black-Box model for an average beech tree 60 cm in height. The color codes are the same as in the example above. The inherent coarse structure in the fit might look less attractive than Figure 2

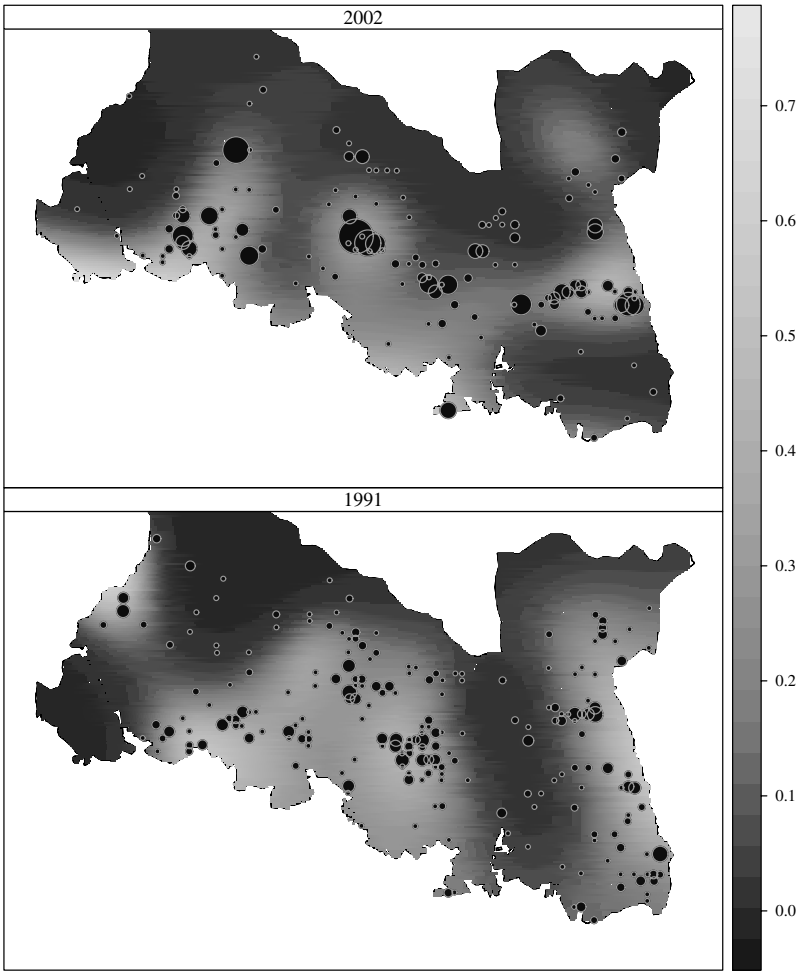


Fig. 4 Spatial Component space fitted by the Generalized Additive Model for an average 60 cm tree in the years 1991 (bottom) and 2002 (top). The diameter of the black circles is proportional to the absolute number of browsed trees.

but in the next section we will perform a formal bootstrap based model inference and will compare the prediction power of all models in fair conditions. Although not as straightforward as in Figure 2, the general pattern for the risky regions in the central and south-western parts of the National Park in 2002 remains visible.

The final example is depicted in Figure 4 representing the GAM model. It proposes a similar structure to Figure 2 with nicely shaped smooth peaks in the risky areas.

In the next section we carry out a model comparison of the prediction quality of the different models.

3.2 Model Comparison

Eight models were fitted to the beech browsing data. Our three candidate models GAMBoost (2) , Black-Box (11) and GAM (16) and their simplified versions including several restrictions are summarized in Table 1. The single column which requires additional clarification is the second column termed "Label". The Label concisely represents the restrictions which we apply to the models. For instance, the label "A" refers to the simplest and fully constrained model with a single intercept as a covariate. "B" denotes a model which considers the height variable only hence ignoring the spatial and the spatio-temporal effects. "C" means a model with the height predictor being constrained to zero and "D" denotes the most complex model which considers all predictors.

We quantify the prediction power of each model using the out-of-bootstrap empirical distribution of the negative log-likelihood. For the boosting algorithms this is done for the *optimal* number of boosting steps. It may appear tedious to bootstrap the step number of boosting within a bootstrapping assessment but the distributional properties of boosting are usually hard to be tracked analytically. Therefore, we perform bootstrapping twice: in the first place seeking for an optimal step number M and secondly for a formal performance assessment.

The results of the performance assessment are shown in Figure 5. Each boxplot represents 25 out-of-bootstrap values of the negative log-likelihood function based on the different models from Table 1. The first four light gray colored boxes represent the common GAM models. The highly constrained models "A" and "B" are not boosted and are primarily used to strengthen the credibility of the other models. The distinct risk collapse in all "C" models compared to "A" and "B" suggest the significant importance of the spatio-temporal effects on the browsing probability. It is further apparent that the height does contribute to the fit at least in the smooth specifications, i.e. "C" has a clearly higher risk than the largest model specification "D" for GAM and GAMBoost, whereas this is not the case in the Black-Box specification.

Further we evidenced that boosting the smooth relationship between the response and the covariates is superior to the common GAM. This can be seen from the juxtaposition of the third and fourth light gray boxplots from left and the two right most boxplots in Figure 5. A drawback of GAMBoost, however, could arise out of the preliminary stage of fine tuning which ensures reasonable computation costs.

It is instantly apparent that the Black-Box model performs the best compared to the other strategies. Thereupon, we empirically showed that expecting the underlying structure to be smooth, does indeed, degrade performance in this case.

Table 1 An overview over all models under test.

Class	Label	Model Specification	Details
GAM	A	$f = 1$	An intercept model averaging the logits in the whole area of investigation.
	B	$f = f_{height}$	A model with restrictions $f_{spatial} = f_{spatemp} \equiv 0$ which allows only for the height effect.
	C	$f = f_{spatial} + f_{spatemp}$	A model with a restriction $f_{height} \equiv 0$ thiw only allowing for spatial and spatio-temporal effects.
	D	$f = f_{spatial} + f_{spatemp} + f_{height}$	The full model defined in (16)
Black-Box	C	$f_{bb} = f_{bbspatial} + f_{bbspatemp}$	A model with restrictions $f_{bheight} \equiv 0$ thus only allowing for spatial and spatio-temporal effects.
	D	$f_{bb} = f_{bbspatial} + f_{bbspatemp} + f_{bheight}$	The full model defined in (11).
GAMBoost	C	$f_{str} = f_{bspatial} + f_{bspatemp}$	A model with restrictions $f_{bheight} \equiv 0$ thus only allowing for spatial and spatio-temporal effects.
	D	$f_{str} = f_{bspatial} + f_{bspatemp} + f_{bheight}$	The full model defined in (2).

4 Discussion

The focus of our study was on the comparison of three modelling techniques for estimating the real forest situation with respect to the beeches in the district of "Rachel-Lusen" in the National Park "Bayerischer Wald". We specified a structured additive model which accounts for the trees' variation in height, as well as for spatial and spatio-temporal effects. The aim was to estimate a smooth surface representing the browsing probabilities on young beech tree within the borders of "Rachel-Lusen" district. We provided a boosted version of the GAM model, i.e. the GAMBoost model, which succeeded to outperform the classical GAM model in terms of stronger minimization of the out-of-sample risk.

We found that the spatial component does contribute to the fit considerably. The same holds for the trees' height which should be considered when estimating the browsing probability in regeneration areas.

The assumption of smooth relationship between the response and the covariates did not proof to be the most credible one amongst our model choices. A simple recursive partitioning of the space predictor via tree based learners, i.e. the Black-Box model, proved to obtain by far the smallest out-of-sample risk than its smooth competitors. In addition, the Black-Box model showed more efficient computational time and required more indulgent parameter tuning effort compared to the GAMBoost model. Apparently the spatial effect of the Black-Box model was very strong and the time predictor seemed to have no effect on the response in this specification.

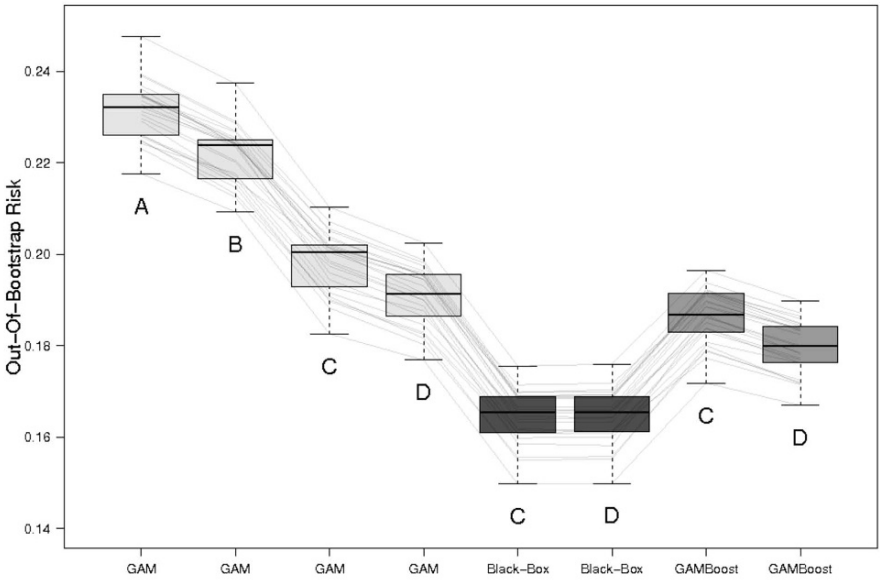


Fig. 5 Out-Of-Bootstrap assessment of the different models defined in Table 1. Each boxplot contains 25 values of negative log-likelihood function.

References

Ammer, C. (1996). Impact of ungulates on structure and dynamics of natural regeneration of mixed mountain forests in the Bavarian Alps, *Forest Ecology and Management* **88**: 43–53.

Augustin, N., Musio, M., von Wilpert E, K., Kublin, Wood, S. & Schumacher, M. (2009). Modelling spatio-temporal forest health monitoring data.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

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

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1983). Classification and regression trees, *Wadsworth, Belmont, California* .

Bühlmann, P. (2004). Bagging, boosting and ensemble methods, *in* J. Gentle, W. Härdle & Y. Mori (eds), *Handbook of Computational Statistics: Concepts and Methods*, Springer.

Bühlmann, P. & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting, *Statistical Science* **22**(4): 477–505.

Bühlmann, P. & Yu, B. (2003). Boosting with the l_2 loss: Regression and classification, *Journal of the American Statistical Association* **98**: 324–339.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**: 1–26.

Eiberle, K. (1989). Über den Einfluss des Wildverbisses auf die Mortalität von jungen Waldbäumen in der oberen Montanstufe, *Schweizer Zeitschrift für Forstwesen* **140**: 1031–1042.

Eiberle, K. & Nigg, H. (1987). Grundlagen zur Beurteilung des Wildverbisses im Gebirgswald, *Schweizer Zeitschrift für Forstwesen* **138**: 474–785.

Eilers, P. & Marx, B. (1996). Flexible smoothing with B-splines and penalties, *Statistical Science* **11**: 89–102.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective, *Statistica Sinica* **14**: 731–761.

Forstliches Gutachten (2006). *Forstliche Gutachten zur Situation der Waldverjungung 2006*, Bayerische Staatsministerium für Landwirtschaft und Forsten.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine, *The Annals of Statistics* **29**: 1189–1232.

Fritz, P. (2006). *Ökologischer Waldumbau in Deutschland - Fragen, Antworten, Perspektiven*, Oekom, München.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall/CRC.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics* **15**: 651–674.

Kneib, T., Hothorn, T. & Tutz, G. (2009). Variable selection and model choice in geoadditive regression models, *Biometrics* **65**: 626–634.

Knoke, T., Ammer, C., Stimm, B. & Mosandl, R. (2008). Admixing broad-leaved to coniferous tree species–A review on yield, ecological stability and economics, *European Journal of Forest Research* **127**: 89–101.

Knoke, T. & Seifert, T. (2008). Integrating selected ecological effects of mixed european beech–Norway spruce stands in bioeconomic modelling, *Ecological Modelling* **210**: 487–498.

Lutz, R., Kalisch, M. & Bühlmann, P. (2008). Robustified l_2 boosting, *Computational Statistics & Data Analysis* **52**: 3331–3341.

Moog, M. (2008). *Bewertung von Wildschäden im Wald*, Neumann-Neudamm, Melsungen.

Motta, R. (2003). Ungulate impact on rowan (*Sorbus aucuparia* l) and norway spruce (*Picea abies* (l) karst) height structure in mountain forests in the italian alps, *Forest Ecology and Management* **181**: 139–150.

Prien, S. (1997). *Wildschäden im Wald*, Paul Parey, Berlin.

Rüegg, D. (1999). Zur Erhebung des Einflusses von Wildtieren auf die Waldverüngung, *Schweizer Zeitschrift für Forstwesen* **150**: 327–331.

Schmid, M. & Hothorn, T. (2008). Boosting additive models using component-wise P-splines, *Computational Statistics & Data Analysis* **53**: 298–311.

Weisberg, P., Bonavia, F. & Bugmann, H. (2005). Modeling the interacting effects of browsing and shading on mountain forest regeneration (*Picea abies*), *Ecological Modelling* **185**: 213–230.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC.

Generalized Linear Mixed Models Based on Boosting

Gerhard Tutz and Andreas Groll

Abstract A likelihood-based boosting approach for fitting generalized linear mixed models is presented. In contrast to common procedures it can be used in high-dimensional settings where a large number of potentially influential explanatory variables is available. Constructed as a componentwise boosting method it is able to perform variable selection with the complexity of the resulting estimator being determined by information criteria. The method is investigated in simulation studies and illustrated by using a real data set.

Key words: Generalized linear mixed model; Boosting; Linear models; Variable selection

1 Introduction

Generalized linear mixed models (GLMMs) as an extension of generalized linear models that incorporate random effects have been an area of intensive research. Various methods have been proposed ranging from numerical integration techniques (for example Booth & Hobert 1999) over "joint maximization methods" (Breslow & Clayton 1993, Schall 1991), in which parameters and random effects are estimated simultaneously, to fully Bayesian approaches (Fahrmeir & Lang 2001). Overviews on current methods are found in McCulloch & Searle (2001) and Fahrmeir & Tutz (2001). Due to the already heavy computational problems in GLMMs modelling usually is restricted to few predictor variables. When many predictors are given, the selection of predictors is often based on test statistics with the usual problems of forward-backward algorithms with stability of estimates.

Gerhard Tutz and Andreas Groll
Institut für Statistik, Ludwigstraße 33, Ludwig-Maximilians-Universität München, Germany,
e-mail: gerhard.tutz@stat.uni-muenchen.de,andreas.groll@stat.uni-muenchen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_11,

In the present article boosting techniques for the selection of predictors are proposed. Boosting was developed within the machine learning community as a method to improve classification. A first breakthrough was the AdaBoost algorithm proposed by Freund & Schapire (1996). Breiman (1998) considered the AdaBoost algorithm as a gradient descent optimization technique and Friedman (2001) extended boosting methods to include regression problems. Bühlmann & Yu (2003) succeeded in proving an exponential dependence between the bias and the variance of the boosted model, which explains the resistance against overfitting. They showed how to fit smoothing splines by boosting base learners and introduced the idea of componentwise boosting, which may be exploited to select predictors. For a detailed overview of componentwise boosting, see Bühlmann & Yu (2003) and Bühlmann & Hothorn (2008).

The paper is structured as follows. In Section 2 we introduce the generalized linear mixed model. In Section 3 we present the boosting algorithm with its computational details and give further information about starting values, stopping criteria and selection. Then the performance of the boosting algorithm is investigated in two simulation studies, one for the random intercept Poisson model and one for the random intercept Bernoulli model. An application to the Multicenter AIDS Cohort Study (MACS, see Kaslow et al. 1987, Zeger & Diggle 1994) is considered in Section 4, which is based on the CD4 data and deals with gay or bisexual men infected with HIV.

2 Generalized Linear Mixed Models - GLMM

Let y_{it} denote observation t in cluster i, $i = 1,\dots,n$, $t = 1,\dots,T_i$, collected in $\mathbf{y}_i^T = (y_{i1},\dots,y_{iT_i})$. Let $\mathbf{x}_{it}^T = (x_{it1},\dots,x_{itp})$ be the covariate vector associated with fixed effects and $\mathbf{z}_{it}^T = (z_{it1},\dots,z_{its})$ the covariate vector associated with random effects. It is assumed that the observations y_{it} are conditionally independent with means $\mu_{it} = E(\mathbf{y}_{it}|\mathbf{b}_i,\mathbf{x}_{it},\mathbf{z}_{it})$ and variances $var(y_{it}|\mathbf{b}_i) = \phi\upsilon(\mu_{it})$, where $\upsilon(.)$ is a known variance function and ϕ is a scale parameter. The generalized linear mixed model that we consider in the following has the form

$$g(\mu_{it}) = \beta_0 + \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{z}_{it}^T\mathbf{b}_i = \beta_0 + \eta_{it}^{\text{par}} + \eta_{it}^{\text{rand}}, \tag{1}$$

where g is a monotonic and continuously differentiable link function, β_0 is the intercept, $\eta_{it}^{\text{par}} = \mathbf{x}_{it}^T\boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_1,\dots,\beta_p)$ and $\eta_{it}^{\text{rand}} = \mathbf{z}_{it}^T\mathbf{b}_i$ contains the cluster-specific random effects $\mathbf{b}_i \sim N(\mathbf{0},\mathbf{Q})$, with covariance matrix $\mathbf{Q}$.

An alternative form that we also use in the following is

$$\mu_{it} = h(\eta_{it}), \quad \eta_{it} = \beta_0 + \eta_{it}^{\text{par}} + \eta_{it}^{\text{rand}},$$

where $h = g^{-1}$ is the inverse link function.

A closed representation of model (1) is obtained by using matrix notation. Let $\mathbf{X}_i^T = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT_i})$ denote the design matrix of the i-th covariate and $\tilde{\boldsymbol{\beta}}^T = (\beta_0, \boldsymbol{\beta}^T)$ the linear parameter vector including intercept. Let $\tilde{\mathbf{X}}_i = [\mathbf{1}, \mathbf{X}_i]$ be the corresponding design matrix, where $\mathbf{1}^T = (1, \ldots, 1)$ is a vector of ones having suitable length. By collecting observations within one cluster the model has the form

$$g(\boldsymbol{\mu}_i) = \tilde{\mathbf{X}}_i \tilde{\boldsymbol{\beta}} + \mathbf{Z}_i \mathbf{b}_i,$$

where $\mathbf{Z}_i^T = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iT_i})$. For all observations one obtains

$$g(\boldsymbol{\mu}) = \tilde{\mathbf{X}} \tilde{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{b},$$

with $\tilde{\mathbf{X}}^T = [\tilde{\mathbf{X}}_1^T, \ldots, \tilde{\mathbf{X}}_n^T]$ and block-diagonal matrix $\mathbf{Z} = Blockdiag(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$. For the random effect $\mathbf{b}$ one has a normal distribution with covariance matrix $\mathbf{Q}_\mathbf{b} = Blockdiag(\mathbf{Q}, \ldots, \mathbf{Q})$.

Focusing on generalized linear mixed models we assume that the conditional density of y_{it}, given explanatory variables and the random effect $\mathbf{b}_i$, is of exponential family type

$$f(y_{it}|\mathbf{X}_i, \mathbf{b}_i) = \exp\left\{ \frac{(y_{it}\theta_{it} - \kappa(\theta_{it}))}{\phi} + c(y_{it}, \phi) \right\}, \tag{2}$$

where $\theta_{it} = \theta(\mu_{it})$ denotes the natural parameter, $\kappa(\theta_{it})$ is a specific function corresponding to the type of exponential family, $c(.)$ the log normalization constant and ϕ the dispersion parameter (compare Fahrmeir & Tutz 2001).

One popular method to maximize generalized linear mixed models is penalized quasi-likelihood (PQL), which has been suggested by Breslow & Clayton (1993), Lin & Breslow (1996) and Breslow & Lin (1995). Typically the covariance matrix $\mathbf{Q}(\boldsymbol{\rho})$ of the random effects $\mathbf{b}_i$ dependents on an unknown parameter vector $\boldsymbol{\rho}$. In penalization-based concepts the joint likelihood-function is specified by the parameter vector of the covariance structure $\boldsymbol{\rho}$ together with the dispersion parameter ϕ, which are collected in $\boldsymbol{\gamma}^T = (\phi, \boldsymbol{\rho}^T)$ and parameter vector $\boldsymbol{\delta}^T = (\beta_0, \boldsymbol{\beta}^T, \mathbf{b}^T)$, $\mathbf{b}^T = (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)$. The corresponding log-likelihood is

$$l(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log\left(\int f(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\gamma}) p(\mathbf{b}_i, \boldsymbol{\gamma}) d\mathbf{b}_i \right), \tag{3}$$

where $p(\mathbf{b}_i, \boldsymbol{\gamma})$ denotes the density of the random effects. Approximation of (3) along the lines of Breslow & Clayton (1993) yields the penalized likelihood

$$l^P(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log(f(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\gamma})) - \frac{1}{2}\mathbf{b}^T \mathbf{Q}(\boldsymbol{\rho})^{-1}\mathbf{b}, \tag{4}$$

where the penalty term $\mathbf{b}^T \mathbf{Q}(\boldsymbol{\rho})^{-1}\mathbf{b}$ is due to the approximation based on the Laplace method.

PQL usually works within the profile likelihood concept. It is distinguished between the estimation of $\boldsymbol{\delta}$ given the plugged-in estimate $\hat{\boldsymbol{\gamma}}$ resulting in the profile-

likelihood $l^P(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}})$ and the estimation of $\boldsymbol{\gamma}$. The PQL method is implemented in the macro `GLIMMIX` and proc `GLMMIX` in SAS (Wolfinger 1994), in the `glmmPQL` and `gamm` functions of the R-packages `MASS` (Venables & Ripley 2002) and `mgcv` (Wood 2006). Further notes were given by Wolfinger & O'Connell (1993), Littell et al. (1996) and Vonesh (1996).

3 Boosted Generalized Linear Mixed Models - bGLMM

Boosting originates in the machine learning community where it has been proposed as a technique to improve classification procedures by combining estimates with reweighted observations. Since it has been shown in Breiman (1999) and Friedman (2001) that reweighting corresponds to minimizing iteratively a loss function, boosting has been extended to regression problems in a L2-estimation framework by Bühlmann & Yu (2003). The boosting algorithm presented in this paper is based on the likelihood function and works by iterative fitting of residuals using "weak learners" and implies selection of components.

3.1 The Boosting Algorithm

The following algorithm uses componentwise boosting. Componentwise boosting means that only one component of the predictor, in our case one linear term, is fitted at a time. More precisely, a model containing the intercept and only one linear term $x_r\beta_r$ is fitted in one iteration step. We will use the notation $\mathbf{x}_{i.r}^T = (x_{i1r}, \ldots, x_{iT_ir})$ for the covariate vector of the r-th linear effect in cluster i and define $\mathbf{x}_{.r}^T = (\mathbf{x}_{1.r}^T, \ldots, \mathbf{x}_{n.r}^T), r = 1, \ldots, p$. Hence the corresponding r-th design matrix containing intercept and only r-th covariate vector is given by

$$\mathbf{X}_{i.r} = [\mathbf{1}, \mathbf{x}_{i.r}] \quad \text{and} \quad \mathbf{X}_{.r} = [\mathbf{1}, \mathbf{x}_{.r}],$$

for cluster i and the whole sample, respectively. For cluster i the predictor that contains only the r-th covariate has the form $\boldsymbol{\eta}_{ir} = \mathbf{X}_{i.r}\tilde{\boldsymbol{\beta}}_r + \mathbf{Z}_i\mathbf{b}_i$, with $\tilde{\boldsymbol{\beta}}_r^T = (\beta_0, \beta_r)$, and for the whole sample one obtains

$$\boldsymbol{\eta}_r = \mathbf{X}_{.r}\tilde{\boldsymbol{\beta}}_r + \mathbf{Z}\mathbf{b}.$$

In the following boosting algorithm the vectors $\tilde{\boldsymbol{\beta}}_r^T = (\beta_0, \beta_r)$ and $\boldsymbol{\delta}_r^T = (\beta_0, \beta_r, \mathbf{b}^T)$ contain only the r-th fixed effect.

Algorithm `bGLMM`

1. *Initialization*
   Compute starting values $\hat{\boldsymbol{\mu}}^{(0)}, \mathbf{Q}^{(0)}, \mathbf{b}^{(0)}$ (see Section 3.2.3) and set $\boldsymbol{\eta}^{(0)} = \mathbf{X}\hat{\hat{\boldsymbol{\beta}}}^{(0)} + \mathbf{Z}\mathbf{b}^{(0)}$.
2. *Iteration*
   For $l = 1, 2, \ldots, l_{max}$

   a. *Refitting of residuals*
      i. Computation of parameters
      For $r \in \{1, \ldots, p\}$ derive the penalized score function $\mathbf{s}_r^P(\boldsymbol{\delta}) = \partial l^P / \partial \boldsymbol{\delta}_r$ and the penalized pseudo Fisher matrix $\mathbf{F}_r^P(\boldsymbol{\delta})$ (see Section 3.2.1). Based on the general form of one step in Fisher scoring given by

$$\hat{\boldsymbol{\delta}}^{(l)} = \hat{\boldsymbol{\delta}}^{(l-1)} + (\mathbf{F}^P(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\mathbf{s}^P(\hat{\boldsymbol{\delta}}^{(l-1)}).$$

      an update of the r-th component is computed. Because the fit is within an iterative procedure it is sufficient to use just one single step. In order to obtain an additive correction of the already fitted terms (the offset), we use one step in Fisher scoring with starting value $\boldsymbol{\delta} = \mathbf{0}$. Therefore Fisher scoring for the r-th component takes the simpler form

$$\hat{\boldsymbol{\delta}}_r^{(l)} = (\mathbf{F}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\mathbf{s}_r(\hat{\boldsymbol{\delta}}^{(l-1)}) \tag{5}$$

      with variance-covariance components being replaced by their current estimates $\hat{\mathbf{Q}}^{(l-1)}$.

      ii. Selection step
      Select from $r \in \{1, \ldots, p\}$ the component j that leads to the smallest $AIC_r^{(l)}$ or $BIC_r^{(l)}$ as given in Section 3.2.3 and select the corresponding $(\hat{\boldsymbol{\delta}}_j^{(l)})^T = \left(\hat{\beta}_0^*, \hat{\beta}_j^*, (\hat{\mathbf{b}}^*)^T\right)$.

      iii. Update
      Set

$$\hat{\beta}_0^{(l)} = \hat{\beta}_0^{(l-1)} + \hat{\beta}_0^*, \qquad \hat{\mathbf{b}}^{(l)} = \hat{\mathbf{b}}^{(l-1)} + \hat{\mathbf{b}}^*$$

      and for $r = 1, \ldots, p$ set

$$\hat{\beta}_r^{(l)} = \begin{cases} \hat{\beta}_r^{(l-1)} & \text{if } r \neq j \\ \hat{\beta}_r^{(l-1)} + \hat{\beta}_r^* & \text{if } r = j, \end{cases}$$

$$(\hat{\boldsymbol{\delta}}^{(l)})^T = \left(\hat{\beta}_0^{(l)}, \hat{\beta}_1^{(l)}, \ldots, \hat{\beta}_p^{(l)}, (\hat{\mathbf{b}}^{(l)})^T\right).$$

      With $\mathbf{A} := [\mathbf{X}, \mathbf{Z}]$ update

$$\hat{\boldsymbol{\eta}}^{(l)} = \mathbf{A}\hat{\boldsymbol{\delta}}^{(l)}$$

b. *Computation of variance-covariance components*
Estimates of $\hat{\mathbf{Q}}^{(l)}$ are obtained as approximate REML-type estimates or alternative methods (see Section 3.2.2)

3.2 Computational Details of bGLMM

In the following we give a more detailed description of the single steps of the `bGLMM` algorithm. First we describe the derivation of the score function and the Fisher matrix. Then two estimation techniques for the variance-covariance components are given. Finally, we give details of the computation of starting values and the selection procedure.

3.2.1 Score Function and Fisher Matrix

In this section we specify more precisely the single components which are derived in step 2 (a) of the `bGLMM` algorithm. For $r \in \{1,\ldots,p\}$ the penalized score function $\mathbf{s}_r^P(\boldsymbol{\delta}) = \partial l^P / \partial \boldsymbol{\delta}_r$, obtained by differentiating the log-likelihood from equation (4), has vector components

$$\begin{aligned}
\mathbf{s}^P_{\tilde{\boldsymbol{\beta}}r} &= \sum_{i=1}^{n} \mathbf{X}_{ir}^T \mathbf{D}_i(\boldsymbol{\delta}) \boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\delta})), \\
\mathbf{s}^P_{ir} &= \mathbf{Z}_i^T \mathbf{D}_i(\boldsymbol{\delta}) \boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\delta})) - \mathbf{Q}^{-1}\mathbf{b}_i, \quad i = 1,\ldots,n,
\end{aligned}$$

with $\mathbf{D}_i(\boldsymbol{\delta}) = \partial h(\boldsymbol{\eta}_i)/\partial \boldsymbol{\eta}, \boldsymbol{\Sigma}_i(\boldsymbol{\delta}) = cov(\mathbf{y}_i|\tilde{\boldsymbol{\beta}}, \mathbf{b}_i)$, and $\boldsymbol{\mu}_i(\boldsymbol{\delta}) = h(\boldsymbol{\eta}_i)$. The vector $\mathbf{s}^P_{\tilde{\boldsymbol{\beta}}r}$ has dimension $p+1$, while the vectors $\mathbf{s}^P_{ir}$ are of dimension s. Note that $\mathbf{s}_r^P(\boldsymbol{\delta})$ could be seen as a penalized score function because of the term $\mathbf{Q}^{-1}\mathbf{b}_i$.
The penalized pseudo-Fisher matrix $\mathbf{F}_r^P(\boldsymbol{\delta})$, $r \in \{1,\ldots,p\}$, which is partitioned into

$$\mathbf{F}_r^P(\boldsymbol{\delta}) = \begin{bmatrix}
\mathbf{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}r} & \mathbf{F}_{\tilde{\boldsymbol{\beta}}1r} & \mathbf{F}_{\tilde{\boldsymbol{\beta}}2r} & \cdots & \mathbf{F}_{\tilde{\boldsymbol{\beta}}nr} \\
\mathbf{F}_{1\tilde{\boldsymbol{\beta}}r} & \mathbf{F}_{11r} & & & 0 \\
\mathbf{F}_{2\tilde{\boldsymbol{\beta}}r} & & \mathbf{F}_{22r} & & \\
\vdots & & & \ddots & \\
\mathbf{F}_{n\tilde{\boldsymbol{\beta}}r} & 0 & & & \mathbf{F}_{nnr}
\end{bmatrix},$$

has single components

$$\mathbf{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}r} = -E\left(\frac{\partial^2 l^P(\boldsymbol{\delta})}{\partial\tilde{\boldsymbol{\beta}}_r\partial\tilde{\boldsymbol{\beta}}_r^T}\right) = \sum_{i=1}^{n}\mathbf{X}_{ir}^T\mathbf{D}_i(\boldsymbol{\delta})\boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1}\mathbf{D}_i(\boldsymbol{\delta})\mathbf{X}_{ir},$$

$$\mathbf{F}_{\tilde{\boldsymbol{\beta}}ir} = \mathbf{F}_{i\tilde{\boldsymbol{\beta}}r}^T = -E\left(\frac{\partial^2 l^P(\boldsymbol{\delta})}{\partial\tilde{\boldsymbol{\beta}}_r\partial\mathbf{b}_i^T}\right) = \mathbf{X}_{ir}^T\mathbf{D}_i(\boldsymbol{\delta})\boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1}\mathbf{D}_i(\boldsymbol{\delta})\mathbf{Z}_i,$$

$$\mathbf{F}_{iir} = -E\left(\frac{\partial^2 l^P(\boldsymbol{\delta})}{\partial\mathbf{b}_i\partial\mathbf{b}_i^T}\right) = \mathbf{Z}_i^T\mathbf{D}_i(\boldsymbol{\delta})\boldsymbol{\Sigma}_i(\boldsymbol{\delta})^{-1}\mathbf{D}_i(\boldsymbol{\delta})\mathbf{Z}_i + \mathbf{Q}^{-1}.$$

3.2.2 Variance-covariance Components

For the estimation of variances (Breslow & Clayton 1993) maximize the profile likelihood that is associated with the normal theory model. By replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ one maximizes

$$\begin{aligned} l(\mathbf{Q}_\mathbf{b}) = &-\frac{1}{2}\log(|\mathbf{V}(\hat{\boldsymbol{\delta}})|) - \frac{1}{2}\log(|\mathbf{X}^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}})\mathbf{X}|) \\ &-\frac{1}{2}(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}) - \mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\hat{\boldsymbol{\delta}})(\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{\delta}}) - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{6}$$

with respect to $\mathbf{Q}_\mathbf{b}$, with the pseudo-observations $\tilde{\boldsymbol{\eta}}(\boldsymbol{\delta}) = \mathbf{A}\boldsymbol{\delta} + \mathbf{D}^{-1}(\boldsymbol{\delta})(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\delta}))$ and with matrices $\mathbf{V}(\boldsymbol{\delta}) = \mathbf{W}^{-1}(\boldsymbol{\delta}) + \mathbf{Z}\mathbf{Q}_\mathbf{b}\mathbf{Z}^T$, $\mathbf{Q}_\mathbf{b} = Blockdiag(\mathbf{Q},\ldots,\mathbf{Q})$ and $\mathbf{W}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta})\mathbf{D}(\boldsymbol{\delta})^T$. Having calculated $\hat{\boldsymbol{\delta}}^{(l)}$ in the l-th boosting iteration, we obtain the estimator $\hat{\mathbf{Q}}_\mathbf{b}^{(l)}$, which is an approximate REML-type estimate for $\mathbf{Q}_\mathbf{b}$.

An alternative estimate, which can be derived as an approximate EM algorithm, uses the posterior mode estimates and posterior curvatures. One derives $(\mathbf{F}^P(\hat{\boldsymbol{\delta}}^{(l)}))^{-1}$, the inverse of the penalized pseudo Fisher matrix of the full model using the posterior mode estimates $\hat{\boldsymbol{\delta}}^{(l)}$ to obtain the posterior curvatures $\hat{\mathbf{V}}_{ii}^{(l)}$. Now compute $\hat{\mathbf{Q}}^{(l)}$ by

$$\hat{\mathbf{Q}}^{(l)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{V}}_{ii}^{(l)} + \hat{\mathbf{b}}_i^{(l)}(\hat{\mathbf{b}}_i^{(l)})^T). \tag{7}$$

In general, the $\mathbf{V}_{ii}$ are derived via the formula

$$\mathbf{V}_{ii} = \mathbf{F}_{ii}^{-1} + \mathbf{F}_{ii}^{-1}\mathbf{F}_{i\tilde{\boldsymbol{\beta}}}(\mathbf{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} - \sum_{i=1}^{n}\mathbf{F}_{\tilde{\boldsymbol{\beta}}i}\mathbf{F}_{ii}^{-1}\mathbf{F}_{i\tilde{\boldsymbol{\beta}}})^{-1}\mathbf{F}_{\tilde{\boldsymbol{\beta}}i}\mathbf{F}_{ii}^{-1},$$

where $\mathbf{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}}, \mathbf{F}_{i\tilde{\boldsymbol{\beta}}}, \mathbf{F}_{ii}$ are the elements of the penalized pseudo Fisher matrix $\mathbf{F}^P(\boldsymbol{\delta})$ of the full model, for details see for example Fahrmeir & Tutz (2001).

3.2.3 Starting Values, Stopping Criteria and Selection in bGLMM

We compute the starting values $\hat{\boldsymbol{\mu}}^{(0)}, \mathbf{Q}^{(0)}$ from step 1. of the `bGLMM` algorithm by fitting the simple global intercept model with random effects given by

$$g(\mu_{it}) = \beta_0 + \mathbf{z}_{it}^T \mathbf{b}_i. \tag{8}$$

This can be done very easily, e.g. by using the R-function `glmmPQL` (Wood 2006) from the `MASS` library (Venables & Ripley 2002).

To find the appropriate complexity of our model we use the effective degrees of freedom, which corresponds to the trace of the hat matrix (Hastie & Tibshirani 1990). In the following we derive the hat matrix corresponding to the l-th boosting step for the r-th component (compare Tutz & Binder 2006, Leitenstorfer 2008). Let $\mathbf{A}_r := [\mathbf{X}_r, \mathbf{Z}]$ and $\mathbf{K} = Blockdiag(0, 0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$ be a block diagonal penalty matrix with a diagonal of two zeros corresponding to intercept and r-th fixed effect and n times the matrix $\mathbf{Q}^{-1}$. Then the Fisher matrix $\mathbf{F}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)})$ and the score vector $\mathbf{s}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)})$ are given in closed form as

$$\mathbf{F}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)}) = \mathbf{A}_r \mathbf{W}_l \mathbf{A}_r + \mathbf{K}$$

and

$$\mathbf{s}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)}) = \mathbf{A}_r^T \mathbf{W}_l \mathbf{D}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}) - \mathbf{K} \hat{\boldsymbol{\delta}}_r^{(l-1)}$$

where $\mathbf{W}_l = \mathbf{W}(\hat{\boldsymbol{\delta}}^{(l-1)}), \mathbf{D}_l = \mathbf{D}(\hat{\boldsymbol{\delta}}^{(l-1)}), \boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}(\hat{\boldsymbol{\delta}}^{(l-1)})$ and $\hat{\boldsymbol{\mu}}^{(l-1)} = h(\hat{\boldsymbol{\eta}}^{(l-1)}) = h(\mathbf{A}\hat{\boldsymbol{\delta}}^{(l-1)})$. For $r = 1, \ldots, m$ the refit in the l-th iteration step by Fisher scoring (5) is given by

$$\begin{aligned} \hat{\boldsymbol{\delta}}_r^{(l)} &= (\mathbf{F}_r^P(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1} \mathbf{s}_r(\hat{\boldsymbol{\delta}}^{(l-1)}) \\ &= (\mathbf{A}_r \mathbf{W}_l \mathbf{A}_r + \mathbf{K})^{-1} \mathbf{A}_r^T \mathbf{W}_l \mathbf{D}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}). \end{aligned}$$

We define the predictor corresponding to the r-th refit in the l-th iteration step as

$$\begin{aligned} \hat{\boldsymbol{\eta}}_r^{(l)} &:= \hat{\boldsymbol{\eta}}^{(l-1)} + \mathbf{A}_r \hat{\boldsymbol{\delta}}_r^{(l)}, \\ \hat{\boldsymbol{\eta}}_r^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)} &= \mathbf{A}_r \hat{\boldsymbol{\delta}}_r^{(l)} \\ &= \mathbf{A}_r (\mathbf{A}_r \mathbf{W}_l \mathbf{A}_r + \mathbf{K})^{-1} \mathbf{A}_r^T \mathbf{W}_l \mathbf{D}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}). \end{aligned}$$

Taylor approximation of first order $h(\hat{\boldsymbol{\eta}}) = h(\boldsymbol{\eta}) + \frac{\partial h(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^T}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ yields

$$\begin{aligned} \hat{\boldsymbol{\mu}}_r^{(l)} &\approx \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{D}_l (\hat{\boldsymbol{\eta}}_r^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)}), \\ \hat{\boldsymbol{\eta}}_r^{(l)} - \hat{\boldsymbol{\eta}}^{(l-1)} &\approx \mathbf{D}_l^{-1} (\hat{\boldsymbol{\mu}}_r^{(l)} - \hat{\boldsymbol{\mu}}^{(l-1)}), \end{aligned}$$

and therefore

$$\mathbf{D}_l^{-1}(\hat{\boldsymbol{\mu}}_r^{(l)} - \hat{\boldsymbol{\mu}}^{(l-1)}) \approx \mathbf{A}_r\,(\mathbf{A}_r\mathbf{W}_l\mathbf{A}_r + \mathbf{K})^{-1}\,\mathbf{A}_r^T\mathbf{W}_l\mathbf{D}_l^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}).$$

Multiplication with $\mathbf{W}_l^{1/2}$ and using $\mathbf{W}^{1/2}\mathbf{D}^{-1} = \boldsymbol{\Sigma}^{-1/2}$ yields

$$\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}}_r^{(l)} - \hat{\boldsymbol{\mu}}^{(l-1)}) \approx \tilde{\mathbf{H}}_r^{(l)}\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}),$$

where $\tilde{\mathbf{H}}_r^{(l)} := \mathbf{W}_l^{1/2}\mathbf{A}_r\,(\mathbf{A}_r\mathbf{W}_l\mathbf{A}_r + \mathbf{K})^{-1}\,\mathbf{A}_r^T\mathbf{W}_l^{1/2}$ denotes the usual generalized ridge regression hat-matrix. Defining $\mathbf{M}_r^{(l)} := \boldsymbol{\Sigma}_l^{1/2}\tilde{\mathbf{H}}_r^{(l)}\boldsymbol{\Sigma}_l^{-1/2}$ yields the approximation

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_r^{(l)} &\approx \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{M}_r^{(l)}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}) \\
&= \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{M}_r^{(l)}[(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-2)}) - (\hat{\boldsymbol{\mu}}^{(l-1)} - \hat{\boldsymbol{\mu}}^{(l-2)})] \\
&\approx \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{M}_r^{(l)}[(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-2)}) - \mathbf{M}_r^{(l-1)}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-2)})].
\end{aligned}$$

The hat matrix corresponding to the global intercept model from equation (8) is

$$\mathbf{M}^{(0)} = \mathbf{A}_1(\mathbf{A}_1^T\mathbf{W}_1\mathbf{A}_1 + \mathbf{K}_1)\mathbf{A}_1^T\mathbf{W}_1,$$

with matrices $\mathbf{A}_1 := [\mathbf{1}, \mathbf{Z}]$ and $\mathbf{K}_1 := \mathit{Blockdiag}(0, \mathbf{Q}^{-1}, \ldots, \mathbf{Q}^{-1})$. As the approximation $\hat{\boldsymbol{\mu}}^{(0)} \approx \mathbf{M}^{(0)}\mathbf{y}$ holds, one obtains

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_r^{(1)} &\approx \hat{\boldsymbol{\mu}}^{(0)} + \mathbf{M}_r^{(1)}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)}) \\
&\approx \mathbf{M}^{(0)}\mathbf{y} + \mathbf{M}_r^{(1)}(\mathbf{I} - \mathbf{M}^{(0)})\mathbf{y}.
\end{aligned}$$

In the following, to indicate that the hat matrices of the former steps have been fixed, let $j_k \in \{1, \ldots, p\}$ denote the index of the component selected in boosting step k. Then we can abbreviate $\mathbf{M}_{j_k} := \mathbf{M}_{j_k}^{(k)}$ for the matrix corresponding to the component that has been selected in the k-th iteration. Further, in a recursive manner, we get

$$\hat{\boldsymbol{\mu}}_r^{(l)} \approx \mathbf{H}_r^{(l)}\mathbf{y},$$

where

$$\begin{aligned}
\mathbf{H}_r^{(l)} &= \mathbf{I} - (\mathbf{I} - \mathbf{M}_r^{(l)})(\mathbf{I} - \mathbf{M}_{j_{l-1}})(\mathbf{I} - \mathbf{M}_{j_{l-2}}) \cdot \ldots \cdot (\mathbf{I} - \mathbf{M}^{(0)}) \\
&= \mathbf{M}_r^{(l)} \prod_{i=0}^{l-1}(\mathbf{I} - \mathbf{M}_{j_i}) + \sum_{k=0}^{l-1} \mathbf{M}_{j_k} \prod_{i=0}^{k-1}(\mathbf{I} - \mathbf{M}_{j_i}) \\
&= \sum_{k=0}^{l} \mathbf{M}_{j_k} \prod_{i=0}^{k-1}(\mathbf{I} - \mathbf{M}_{j_i}),
\end{aligned}$$

is the hat matrix corresponding to the l-th boosting step considering the r-th component, whereas $\mathbf{M}_{j_l} := \mathbf{M}_r^{(l)}$ is not fixed yet.

In general, given hat matrix $\mathbf{H}$, the complexity of the model may be determined by the information criteria. We will use

$$AIC = -2\,l(\boldsymbol{\mu}) + 2\,\text{trace}\,(\mathbf{H}), \tag{9}$$

$$BIC = -2\,l(\boldsymbol{\mu}) + 2\,\text{trace}\,(\mathbf{H})\,log(n), \tag{10}$$

where

$$l(\boldsymbol{\mu}) = \sum_{i=1}^{n} l_i(\boldsymbol{\mu}_i) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\boldsymbol{\mu}_i) \tag{11}$$

denotes the general log-likelihood and $l_i(\boldsymbol{\mu}_i)$ the log-likelihood contribution of $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i)$. In general, the log-likelihood (4) can also be written with $\boldsymbol{\mu}$ instead of $\boldsymbol{\delta}$ in the argument, considering the definition of the natural parameter $\theta = \theta(\boldsymbol{\mu})$ in (2) and using $\boldsymbol{\mu} = h(\boldsymbol{\eta}) = h(\boldsymbol{\eta}(\boldsymbol{\delta}))$. In (9) and (10) the nonpenalized log-likelihood is used.

For exponential family distributions $\log f(\mathbf{y}_i|\boldsymbol{\mu}_i)$ has a well-known form. For example in the case of binary responses, one obtains

$$\log f(\mathbf{y}_i|\boldsymbol{\mu}_i) = \sum_{t=1}^{T_i} y_{it} \log \mu_{it} + (1 - y_{it}) \log (1 - \mu_{it}),$$

whereas in the case of Poisson responses, one has

$$\log f(\mathbf{y}_i|\boldsymbol{\mu}_i) = \sum_{t=1}^{T_i} y_{it} \log \mu_{it} - \mu_{it}.$$

Based on (11), the information criteria (9) and (10) used in the l-th boosting step, considering the r-th component, have the form

$$AIC_r^{(l)} = -2\,l(\hat{\boldsymbol{\mu}}_r^{(l)}) + 2\,\text{trace}\,(\mathbf{H}_r^{(l)}),$$

$$BIC_r^{(l)} = -2\,l(\hat{\boldsymbol{\mu}}_r^{(l)}) + 2\,\text{trace}\,(\mathbf{H}_r^{(l)})\,log(n),$$

with

$$l(\hat{\boldsymbol{\mu}}_r^{(l)}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_{ir}^{(l)}). \tag{12}$$

In the l-th step one selects from $r \in \{1, \ldots, p\}$ the component j_l that minimizes $AIC_r^{(l)}$ or $BIC_r^{(l)}$ and obtains $AIC^{(l)} := AIC_{j_l}^{(l)}$. We choose a number l_{max} of maximal boosting steps, e.g. $l_{max} = 1000$, and stop the algorithm at iteration l_{max}. Then we select from $\mathscr{L} := \{1, 2, \ldots, l_{max}\}$ the component l_{opt}, where $AIC^{(l)}$ or $BIC^{(l)}$ is smallest, that is

$$l_{opt} = \arg\min_{l \in \mathscr{L}} AIC^{(l)},$$
$$l_{opt} = \arg\min_{l \in \mathscr{L}} BIC^{(l)}.$$

Finally, we obtain the parameter estimates $\hat{\boldsymbol{\delta}}^{(l_{opt})}, \hat{\mathbf{Q}}^{(l_{opt})}$ and the corresponding fit $\hat{\boldsymbol{\mu}}^{(l_{opt})}$.

It should be noted that similar to Tutz & Reithinger (2006) our selection step reflects the complexity of the refitted model, which is in contrast to established componentwise boosting procedures. For example Bühlmann & Yu (2003), select the component that maximally improves the fit and then evaluate if the fit including model complexity deteriorates. The procedure proposed here selects the component such that the new lack-of-fit, including the augmented complexity, is minimized.

3.3 Simulation Study

In the following simulation studies the performance of the `bGLMM` algorithm is compared to alternative approaches.

Poisson Link The underlying model is the random intercept Poisson model

$$\eta_{it} = \sum_{j=1}^{p} x_{itj}\beta_j + b_i, \quad i = 1,\ldots,40, \quad t = 1,\ldots,10,$$
$$E[y_{it}] = \exp(\eta_{it}) := \lambda_{it}, \qquad y_{it} \sim \text{Pois}(\lambda_{it}),$$

with linear effects given by $\beta_1 = -4, \beta_2 = -6, \beta_3 = 10$ and $\beta_j = 0, j = 4,\ldots,50$. We choose the different settings $p = 3,5,10,20,50$. For $j = 1,\ldots,50$ the vectors $\mathbf{x}_{it}^T = (x_{it1},\ldots,x_{it50})$ follow a uniform distribution within the interval $[-0.3, 0.3]$. The number of observations is determined by $n = 40, T_i := T = 10, i = 1,\ldots,n$. The random effect and the noise variable have been specified by $b_i \sim N(0,\sigma_b^2)$ with $\sigma_b^2 = 0.6$.

The performance of estimators is evaluated separately for the structural components and the variance. We compare the results of our `bGLMM` algorithm with the results obtained by the R-function `glmmPQL` recommended in Wood (2006). The `glmmPQL` routine is supplied with the `MASS` library (Venables & Ripley 2002). It operates by iteratively calling the R-function `lme` from the `nlme` library and returns the fitted `lme` model object for the working model at convergence. For more details about the `lme` function, see Pinheiro & Bates (2000).

By averaging across 50 training data sets we consider mean squared errors for $\boldsymbol{\beta}$ and σ_b given by

$$\text{mse}_{\boldsymbol{\beta}} := ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2, \qquad \text{mse}_{\sigma_b} := ||\sigma_b - \hat{\sigma}_b||^2.$$

Table 1 Generalized linear mixed model (`glmmPQL`) and boosting (`bGLMM`) on Poisson data

	`glmmPQL`		`bGLMM` (EM)				`bGLMM` (REML)			
p	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	falsepos	falseneg	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	falsepos	falseneg
3	0.088	0.004	0.104	0.006	0	0	0.100	0.004	0	0
5	0.124	0.004	0.108	0.006	0.10	0	0.101	0.004	0.02	0
10	0.218	0.004	0.110	0.006	0.34	0	0.101	0.004	0.04	0
20	0.537	0.004	0.118	0.006	0.66	0	0.108	0.004	0.10	0
50	2.013	0.005	0.143	0.008	1.68	0	0.124	0.007	0.30	0

To avoid that single outliers distort the analysis, we present the medians of both quantities in Table 1. The corresponding boxplots are shown in Figure 1. Additionally, we present boxplots of the σ_b-difference

$$\triangle_{\sigma_b} := \sigma_b - \hat{\sigma}_b$$

in Figure 2, to investigate the bias of estimates the true value $\sigma_b = \sqrt{0.6}$.

Additional information on the performance of the algorithm was collected in *falseneg*, the mean over all 50 simulations of the number of variables $\beta_j, j = 1,2,3$, that were not selected and in *falsepos*, the mean over all 50 simulations of the number of variables $\beta_j, j = 4,\ldots,50$, that were selected. Notice at this point, that the `glmmPQL` function is not able to perform variable selection and therefore always estimates all p parameters β_j.

The results for varying number p of covariates $x_{it1},\ldots,x_{itp}$ are summarized in Table 1. For the computation of the random effects variance-covariance components $\mathbf{Q}$ we used the two estimation techniques given in Section 3.2.2. The results using the EM-type estimates $\hat{\mathbf{Q}}$ from (7) are found in the `bGLMM` (EM) column of Table 1, results for the REML-type estimates $\hat{\mathbf{Q}}$, obtained by maximization of the profile likelihood in (6), are given in the third column. The corresponding results can be found in the `bGLMM` (REML) column of Table 1. It is seen that boosting estimates distinctly outperform the simple PQL algorithm when redundant variables are present. REML type estimates turned out to be more stable than the EM-type estimates.

To illustrate how the `bGLMM` algorithm works we show in Figure 3 the paths of the three coefficients β_1, β_2 and β_3 for all simulations. It is seen that the algorithm always starts with updating coefficient β_3, which has the most influence on $\boldsymbol{\eta}$, as it has the biggest absolute value. Next, the coefficient β_2 is updated, while β_1 is the last coefficient which is refitted.

Bernoulli Link

The underlying model is the random intercept Bernoulli model

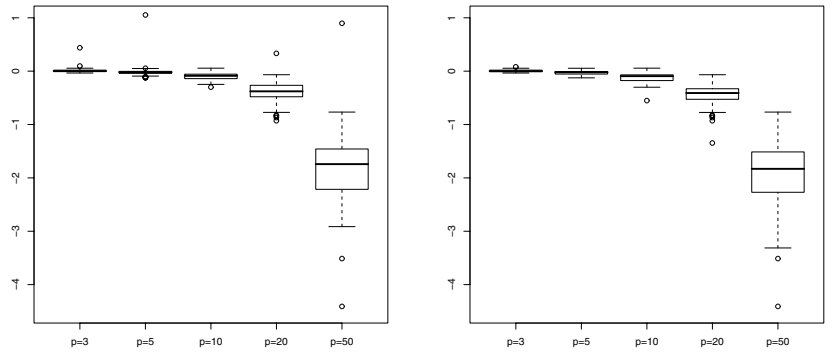


Fig. 1 Boxplots of $(\text{mse}_{\boldsymbol{\beta}}^{\text{bGLMM}} - \text{mse}_{\boldsymbol{\beta}}^{\text{glmmPQL}})$ for the EM model (left, without few extreme outliers) and the REML model (right)

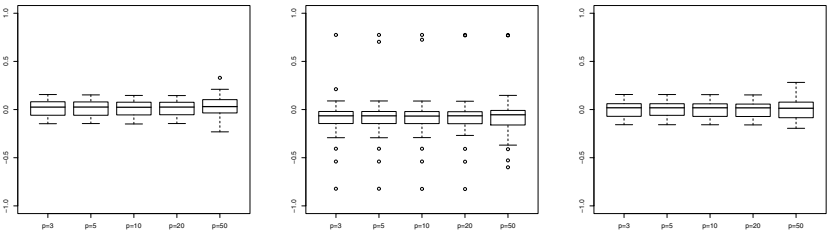


Fig. 2 Boxplots of $\triangle_{\sigma_b}$ for the `glmmPQL` model (left), for the `bGLMM` EM model (middle) and for the `bGLMM` REML model (right)

$$\eta_{it} = \sum_{j=1}^{p} x_{it\,j}\beta_j + b_i, \quad i = 1,\ldots,40, \quad t = 1,\ldots,10$$

$$E[y_{it}] = \frac{\exp(\eta_{it})}{1+\exp(\eta_{it})} := \pi_{it} \qquad y_{it} \sim \text{B}(1,\pi_{it})$$

with linear effects given by $\beta_1 = -5, \beta_2 = -10, \beta_3 = 15$ and $\beta_j = 0, j = 4,\ldots,50$. Again we choose the different settings $p = 3,5,10,20,50$. For $j = 1,\ldots,50$ the vectors $\mathbf{x}_{it}^T = (x_{it1},\ldots,x_{it50})$ have been drawn independently with components following a uniform distribution within the interval $[-0.1,0.1]$. The number of observations remains $n = 40, T_i := T = 10, \forall i = 1,\ldots,n$. The random effect and the noise variable have been specified by $b_i \sim N(0,\sigma_b^2)$ with $\sigma_b^2 = 0.6$.

Again, we evaluate the performance of estimators separately for structural components and variance and compare the results of our `bGLMM` algorithm with the results achieved via the `glmmPQL` function (Wood 2006). Therefore we use the same goodness-of-fit criteria as in the Poisson case.

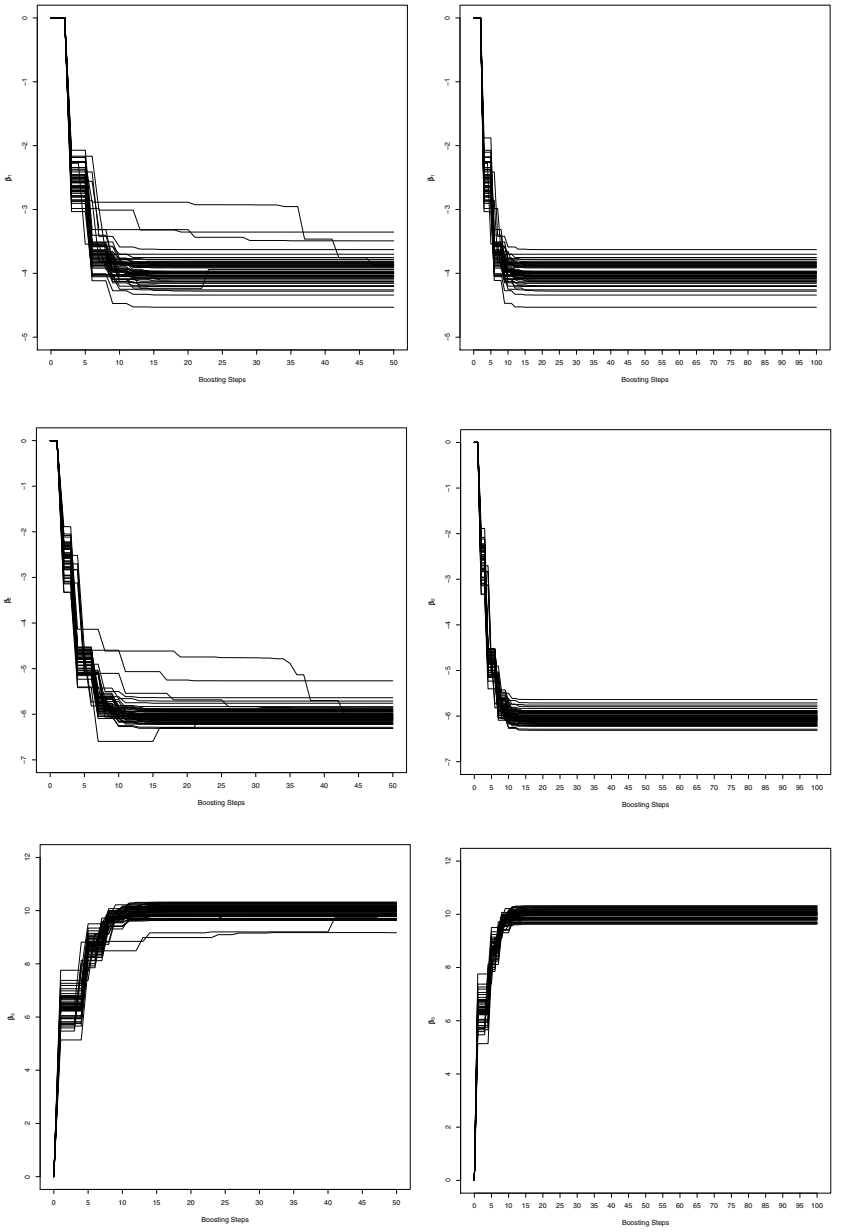


Fig. 3 Coefficient paths of β_1, β_2 and β_3 calculated by `bGLMM` algorithm for the generalized linear mixed Poisson EM (left) and REML (right) model in the $p = 20$ case

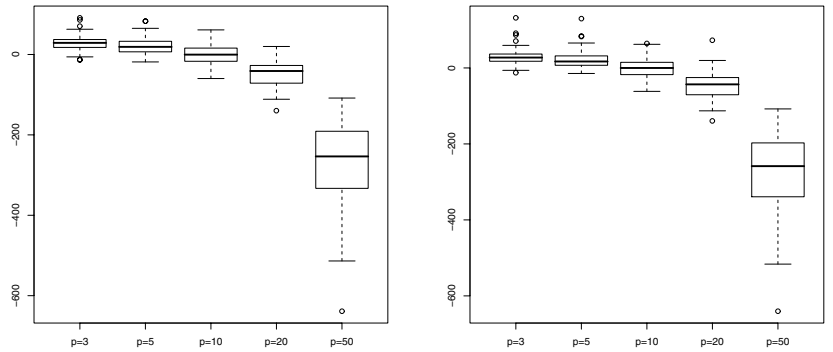


Fig. 4 Boxplots of $(\text{mse}_{\boldsymbol{\beta}}^{\text{bGLMM}} - \text{mse}_{\boldsymbol{\beta}}^{\text{glmmPQL}})$ for the EM model (left) and the REML model (right)

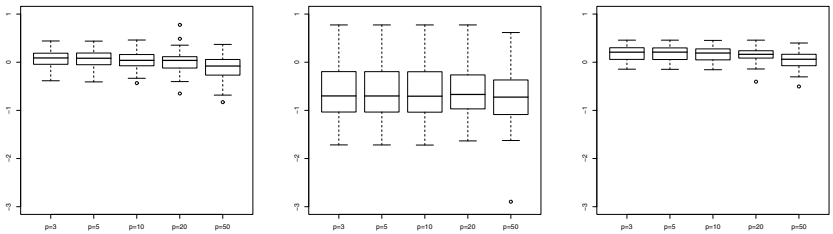


Fig. 5 Boxplots of $\triangle_{\sigma_b}$ for the `glmmPQL` model (left), for the `bGLMM` EM model (middle) and for the `bGLMM` REML model (right)

Table 2 Generalized linear mixed model (`glmmPQL`) and boosting (`bGLMM`) on Bernoulli data

	`glmmPQL`		`bGLMM` (EM)				`bGLMM` (REML)			
p	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	falsepos	falseneg	$\text{mse}_{\boldsymbol{\beta}}$	mse_{σ_b}	falsepos	falseneg
3	9.70	0.016	36.66	0.552	0	0.84	36.92	0.043	0	0.86
5	19.80	0.014	36.66	0.553	0.02	0.82	36.93	0.044	0.02	0.86
10	44.92	0.012	39.93	0.554	0.10	0.82	37.92	0.036	0.10	0.86
20	90.82	0.015	34.29	0.553	0.12	0.66	35.08	0.029	0.14	0.72
50	294.01	0.030	48.39	0.525	0.46	0.66	45.08	0.016	0.46	0.60

The results for varying number p of covariates $x_{it1}, \ldots, x_{itp}$ and for the two different estimation methods for the random effects variance-covariance components $\mathbf{Q}$ are summarized in Table 2. Table 2 as well as Figures 4 to 5 show that in the Bernoulli case boosting is less convincing than in the Poisson case, in particular in terms of mse_{σ_b} . But the general trend, that, in case of many covariates, the $\boldsymbol{\beta}$-fit that is achieved using the `bGLMM` algorithm outperforms the fit obtained by the `glmmPQL`

Table 3 Estimates for the AIDS Cohort Study MACS with `glmmPQL` function (standard deviations in brackets) and `bGLMM` algorithm

	`glmmPQL`	`bGLMM` (EM)	`bGLMM` (REML)
Intercept	6.5547 (0.018)	6.5362	6.5362
Time	-0.2210 (0.011)	-0.2191	-0.2191
Time^2	-0.0156 (0.010)	-0.0197	-0.0197
Drugs	0.0126 (0.010)	0	0
Partners	0.0385 (0.010)	0.0568	0.0568
Packs of Cigarettes	0.057 (0.013)	0	0
Mental illness score (cesd)	-0.0304 (0.010)	-0.0388	-0.0388
Age	0.0020 (0.018)	0	0
σ_b^2	0.3025	0.3549	0.3539
Φ	66.8224	76.0228	76.0228

function, can still be observed. When variable selection is needed boosting estimates of $\boldsymbol{\beta}$ are distinctly better than estimates obtained by the `glmmPQL` function.

4 Application to CD4 Data

The data were collected within the Multicenter AIDS Cohort Study (MACS). In the study about 5000 infected gay or bisexual men from Baltimore, Pittsburgh, Chicago and Los Angeles have been observed since 1984 (see Kaslow et al. 1987, Zeger & Diggle 1994). The human immune deficiency virus (HIV) causes AIDS by attacking an immune cell called the CD4+ cell which coordinates the body's immunoresponse to infectious viruses and hence reduces a person's resistance against infection. According to Diggle et al. (2002) an uninfected individual has around 110 cells per milliliter of blood and since the number of CD4+ cells decreases with time from infection, one can use an infected person's CD4+ cell number to check disease progression. Within the MACS, $n = 369$ seroconverters with a total of $\sum_{i=1}^{n} T_i = 2376$ measurements were included with the number of CD4+ cells being the interesting response variable. Covariates include years since seroconversion ranging from 3 years before to 6 years after seroconversion, packs of cigarettes a day, recreational drug use (yes/no), number of sexual partners, age and a mental illness score (cesd). For observation t of individual i, the model that is considered in the following has the form

$$\begin{aligned} g(\mu_{it}) &= \beta_0 + \eta_{it}^{\text{par}} + \eta_{it}^{\text{rand}} \\ &= \beta_0 + \text{time}_{it}\beta_1 + \text{time}_{it}^2\beta_2 + \text{drugs}_{it}\beta_3 + \text{partners}_{it}\beta_4 \\ &\quad + \text{cigarettes}_{it}\beta_5 + \text{cesd}_{it}\beta_6 + \text{age}_{it}\beta_7 + b_i, \end{aligned}$$

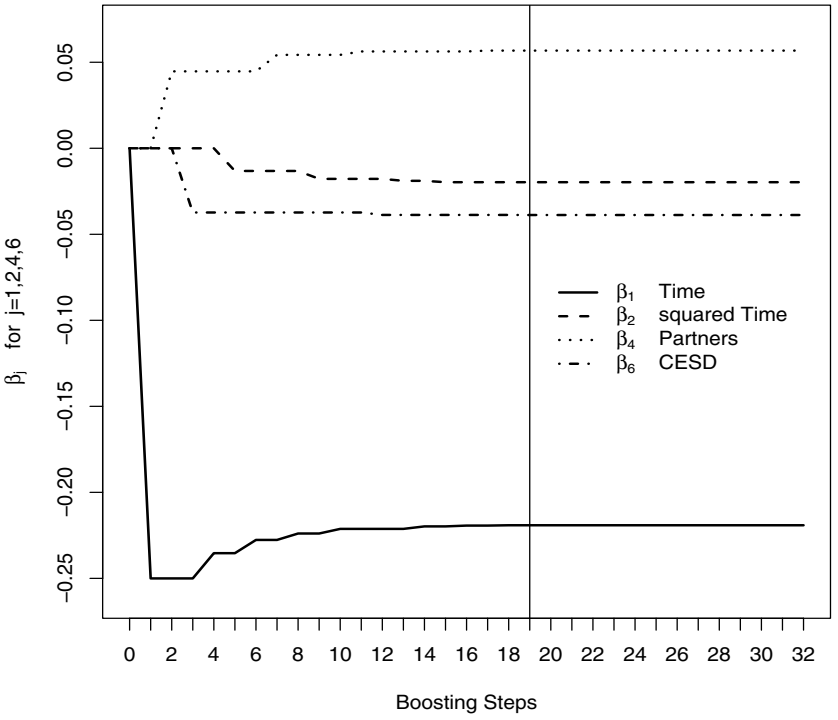


Fig. 6 Coefficient paths of $\beta_i \neq 0$ calculated by the generalized linear mixed Poisson REML model

with $b_i \sim N(0, \sigma_b^2)$. Our main objective is the typical time course of CD4+ decay and the variability across subjects. As the time effect may be nonlinear, we additionally consider the covariate "squared time". We fit an overdispersed Poisson model with natural link. The overdispersion parameter Φ is estimated by use of Pearson residuals

$$\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{(v(\hat{\mu}_{it}))^{\frac{1}{2}}}$$

by

$$\hat{\Phi} = \frac{1}{N - \text{trace}(\mathbf{H})} \sum_{i=1}^{n} \sum_{1}^{T_i} \hat{r}_{it}^2, \qquad N = \sum_{i=1}^{n} T_i.$$

For the estimation procedure we have standardized all covariates. The results for the `bGLMM` algorithm and for the `glmmPQL` function are given in Table 3. It is seen that the two boosting algorithms yield nearly the same estimates. The incorporated

selection procedure suggests that drug use, pack of cigarettes a day and age are not needed in the predictor.

The maximal number of boosting steps has been chosen as $l_{max} = 100$ and the algorithm selected $l_{opt} = 19$ as optimal number of boosting steps. Coefficients build ups-for coefficients are found in Figure 6, with the vertical line indicating the optimal stopping point l_{opt}. It is seen that coefficient estimates are very stable after about 10 boosting steps.

5 Concluding Remarks

Algorithms are derived that allow to estimate generalized mixed models with high-dimensional predictor structure. The incorporated selection procedure reduces the predictor space when redundant variables are present. Although penalized quasi-likelihood estimators work also in cases up to 50 predictors, performance deteriorates when many spurious variables are present. In these cases boosting approaches show better performance even in the binary response case. For low-dimensional settings boosting for binary responses still needs to be improved.

The approach proposed here can be extended to incorporate nonparametric effects. Let $\mathbf{u}_{it}^T = (u_{it1}, \ldots, u_{itm})^T$ be the covariate vector consisting of m different covariates associated with these nonparametric effects. The generalized semiparametric mixed model has the form

$$\begin{aligned} g(\mu_{it}) &= x_{it}^T \boldsymbol{\beta} + \sum_{j=1}^{m} \alpha_{(j)}(u_{itj}) + w_{it}^T \mathbf{b}_i \\ &= \eta_{it}^{\text{par}} + \eta_{it}^{\text{add}} + \eta_{it}^{\text{rand}}, \end{aligned}$$

where g is a monotonic differentiable link function, $\eta_{it}^{\text{par}} = x_{it}^T \boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$, now including the intercept, $\eta_{it}^{\text{add}} = \sum_{j=1}^{m} \alpha_{(j)}(u_{itj})$ is an additive term with unspecified influence functions $\alpha_{(1)}, \ldots, \alpha_{(m)}$ and finally $\eta_{it}^{\text{rand}} = w_i^T \mathbf{b}_i$ contains the cluster-specific random effects $\mathbf{b}_i \sim N(0, \mathbf{Q})$, where $\mathbf{Q}$ is a known or unknown covariance matrix. By expanding nonparametric effects in basis functions and using a weak learner that refers to the updating of all coefficients corresponding to one nonparametric effect the model may be estimated with an incorporated selection procedure.

References

Booth, J. G. & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *J. R. Statist. Soc* **B 61**: 265–285.

Breiman, L. (1998). Arcing classifiers, *Annals of Statistics* **26**: 801–849.

Breiman, L. (1999). Prediction games and arcing algorithms, *Neural Computation* **11**: 1493–1517.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed model, *Journal of the American Statistical Association* **88**: 9–25.

Breslow, N. E. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**: 81–91.

Bühlmann, P. & Hothorn, T. (2008). Boosting algorithms: regularization, prediction and model fitting, *Statistical Science* . accepted.

Bühlmann, P. & Yu, B. (2003). Boosting with the L2 loss: Regression and classification, *Journal of the American Statistical Association* **98**: 324–339.

Diggle, P. J., Heagerly, P., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*, Oxford University Press, Oxford.

Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors, *Applied Statistics* **50**: 201–220.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer-Verlag, New York.

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm, *Proceedings of the Thirteenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp. 148–156.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics* **29**: 337–407.

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

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. & Rinaldo, C. R. (1987). The multicenter aids cohort study: rationale, organization and selected characteristic of the participants, *American Journal of Epidemiology* **126**: 310–318.

Leitenstorfer, F. (2008). *Boosting in Nonparametric Regression: Constrained and Unconstrained Modeling Approaches*, Verlag Dr. Hut, München.

Lin, X. & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion, *Journal of the American Statistical Association* **91**: 1007–1016.

Littell, R., Milliken, G., Stroup, W. & Wolfinger, R. (1996). *SAS System for Mixed Models*, SAS Institue Inc., Cary, NC.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear and Mixed Models*, Wiley, New York.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*, Springer, New York.

Schall, R. (1991). Estimation in generalised linear models with random effects, *Biometrika* **78**: 719–727.

Tutz, G. & Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting, *Biometrics* **62**: 961–971.

Tutz, G. & Reithinger, F. (2006). A boosting approach to flexible semiparametric mixed models, *Statistics in medicine* **26**: 2872–2900.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, fourth edn, Springer, New York.

Vonesh, E. F. (1996). A note on the use of laplace's approximatio for nonlinear mixed-effects models, *Biometrika* **83**: 447–452.

Wolfinger, R. W. (1994). Laplace's approximation for nonlinear mixed models, *Biometrika* **80**: 791–795.

Wolfinger, R. W. & O'Connell, M. (1993). Generalized linear mixed models; a pseudolikelihood approach, *Journal Statist. Comput. Simulation* **48**: 233–243.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman & Hall, London.

Zeger, S. L. & Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters, *Biometrics* **50**: 689–699.

Measurement and Predictors of a Negative Attitude towards Statistics among LMU Students

Carolin Strobl, Christian Dittrich, Christian Seiler, Sandra Hackensperger and Friedrich Leisch

Abstract The measurement of the attitude towards statistics and the relationship between the attitude towards statistics and several socio-demographic and educational factors was investigated in a survey on over 600 students of the Ludgwig-Maximilians-Universität (LMU). The attitude towards statistics was measured by means of the Affect and Cognitive Competence scales of the Survey of Attitudes Towards Statistics (SATS, Schau et al. 1995), that proved to be well suited for identifying students with high levels of negative attitude against statistics, even though potential effects of the translation into German were noticeable for the positively worded items. Predictors found relevant for a negative attitude towards statistics were gender, mathematics taken as an intensive course in high school, prior (perceived) mathematics achievement, prior mathematics experience as well as two of the newly included items on students' strategy applied in mathematics courses in high school: Students who named practicing as their strategy were less likely, while students who named memorizing as their strategy were more likely to show a negative attitude towards statistics.

1 Introduction

The issue of mathematics and statistics anxiety among college students has been a subject of psychological and educational research for decades, headed by early publications such as Dreger & Aitken (1957) on "number anxiety" in general. Cruise et al. (1985) define statistics anxiety as "the feeling of anxiety encountered when taking a statistics course or doing statistical analysis". This kind of anxiety can pose a major problem to both students and instructors in many applied sciences, where

Carolin Strobl, Friedrich Leisch, Christian Dittrich, Christian Seiler and Sandra Hackensperger
Institut für Statistik, Ludwigstraße 33, Ludwig-Maximilians-Universität München, Germany,
e-mail: carolin.strobl,friedrich.leisch@stat.uni-muenchen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_12,

statistics and methodology classes are both mandatory and necessary to provide essential academic skills.

The results of the experimental study of Ashcraft & Kirk (2001) could even show that students with a high level of mathematics anxiety, who usually perform worse not only in mathematics exams but also in working memory tests involving numbers, could do as well as subjects from a control group when permitted to use pencil and paper for computations. This finding indicates that mathematics anxiety directly affects cognitive processes such as the working memory, which in return results in poor test performance. A similar effect is likely to hold for statistics anxiety and may affect the students performance in statistics and methodology classes in university.

While Mills (2004) reports that in her sample of over 200 business students positive attitudes about statistics were more frequent than negative attitudes, many authors have estimated that between 70% (Zeidner 1991) and up to 80% (Onwuegbuzie et al. 2000) of college students enrolled in various major subjects experience more or less severe forms of statistics anxiety.

Different potential predictors of statistics anxiety have been investigated in empirical studies on statistics anxiety among college students (see, e.g., Zeidner 1991, Wilson 1997, Galagedera et al. 2000, Fullerton & Umphrey 2001, Onwuegbuzie 2001, Baloglu 2003, Carmona 2004, Mills 2004). Some potential predictors, like prior math achievement, show a persistent association with statistics anxiety, while the influence of age and gender, for instance, is still subject to discussion.

Instruments available for assessing statistics anxiety include, amongst many others (cf., e.g., the overviews in Schau et al. 1995, Fullerton & Umphrey 2002), the Statistical Anxiety Rating Scale (STARS, Cruise et al. 1985), the Attitude Towards Statistics scale (ATS, Wise 1985) and the Survey of Attitudes Towards Statistics (SATS, Schau et al. 1995). Fullerton & Umphrey (2002) state that "all [considered] instruments showed a high correlation between positive attitudes towards statistics and high course grades."

The SATS, that was used in the study presented here, was designed to meet several key characteristics that were not covered by the then existing scales (Schau et al. 1995). In order to reflect the most important dimensions of attitudes toward statistics a panel of instructors and students identified the dimensions by consensus. In the evaluation of the concurrent validity of the SATS with respect to the ATS scale of Wise (1985) Pearson correlation coefficients indicated a high positive correlation for the Affect and Cognitive Competence scales and medium positive correlations for the Value and Difficulty scales with the ATS Course scale (Schau et al. 1995). The correlations with the ATS Field scale were medium positive for the SATS Affect and Cognitive Competence scale and high for the Value scale, while the correlation for Difficulty was not significant. Schau et al. (1995) conclude that "[t]hese results suggest substantial correspondence between the ATS Course scale and two dimensions of the SATS, Affect and Cognitive Competence". Therefore, these two scales of the SATS, together with some additional items, were used in the questionnaire for this study.

Note that some of the additional items of the original SATS do not apply to the German school and university system: For example, the number of years of

high school mathematics taken is the same for all regular high school attendees in Germany. On the other hand in the German high school system the students have a choice for or against mathematics as an intensive course for the last two years of high school. Therefore we included mathematics as an intensive course, rather than the original SATS item on the number of years of high school mathematics taken, in our study.

A new type of item that was also included in the questionnaire is concerned with the strategy applied in previous math courses, because informal conversations with students in introductory statistics courses gave us the impression that i) there are structural differences in the students' strategies when preparing for a test, that have been adopted in school and ii) students with high levels of statistics anxiety might use suboptimal strategies.

Throughout this paper we will first investigate the behavior of the SATS scales Affect and Cognitive Competence in our sample (with a focus on potential effects of their translation into German) and then use a simple indicator, that is computed from these two scales, to identify relevant predictor variables of a negative attitude towards statistics or even statistics anxiety in LMU students.

Students with different major subjects (psychology, sociology, business studies, economics and few others) were included in the study. However, the sample contained no students of Ludwig Fahrmeir, because their attitude towards statistics would have been positively biased by attending his lecture.

In addition to aspects of previous math performance, experience and training, that have been found to be relevant in previous studies (e.g., Zeidner 1991, Carmona 2004), we included in this study not only the current age of the students but also their time and activity since high school graduation. These variables are in the same spirit as the time since last exhibited to a mathematics course, which has been shown to be related to statistics anxiety by Wilson (1997). The information on time and activity since high school graduation is particularly important because some of our students, especially in psychology and sociology, do not directly go from high school to university and might be differentially affected by statistics anxiety.

2 Method

The study was conducted at the LMU in Munich, Germany, with participants of three introductory statistics courses as subjects. The student sample as well as the design and distribution of the questionnaire are described in detail in the following.

2.1 Participants

689 first year students from empirical social and business sciences were recruited as volunteers from three mandatory introductory statistics courses. Of these students

294 were majors in business studies, 103 in economics, 113 in psychology, 84 in sociology, 60 in business and economics education and 35 in others or did not report. The average age was 21.56 years with a minimum of 17 (the average German student enters university at age 18 or 19) and a maximum of 51 years. The time since high school graduation ranged from 36 years ago to within the last year. Of the sample of 689 students 402 were female, 273 male, the rest did not report. 84.5% of the participants named German as their first language, 87.1% had achieved their high school diploma in Germany. 10.1% had achieved their high school diploma by means of second-chance education. 54% of the participants went directly from high school to university. Another 17% served in military or civil service (which is mandatory for most male students in Germany) before entering university, 14% were working, 19% attended industrial training, 16% spent the time otherwise (multiple answers were possible). 36% of the students had chosen mathematics as an intensive course in high school.

Of the 689 initial observations 24 were deleted for the item analysis of the SATS Affect and Cognitive Competence scales because they had missing values in these scales. This leaves 665 observations for the item analysis in Section 3.1 and the computation of the indicator for a negative attitude towards statistics in Section 3.2. Of these observations 63 had missing values in one or more of the socio-demographic and educational items and 3 stated implausible values for their time of high school graduation. These observations were deleted for the rest of the analysis, leaving 599 observations for the analysis of potential predictors of a negative attitude towards statistics in Section 3.3.

2.2 Procedure and Instrument

All three introductory statistics courses were visited in the first week of the students' first term to ensure that their attitude towards statistics was not affected by previous course experience. The participants were free to volunteer or hand in blank questionnaires anonymously. By completing the form they were informed that they gave permission to use their responses for research purposes. Results of this and further analyses will be reported to the students.

The questionnaire consisted of the two SATS scales Affect and Cognitive Competence presented as 7-point Likert-type scales. In addition the questionnaire contained 3 items on prior experience covering mathematics experience and attitude in school ("I liked mathematics in school.", "In school I was scared of mathematics." and "In school I was good in mathematics.") and 3 items on strategy applied in previous math courses ("My strategy in mathematics was to try to understand the underlying concepts.", "... was to practice on as many problems as possible." and "... was to memorize as much as possible."). These items were also presented as 7-point Likert-type scales.

Socio-demographic and educational items covered major subject, age, gender, first language, high school attended in Germany, time since high school graduation,

high school diploma by means of second-chance education, activity since high school graduation, favorite subject in school, mathematics taken as an intensive course in high school and mathematics grade in final exams.

2.3 Software

All following analyses were conducted by means of the `R` system for statistical computing version 2.5.1 (R Development Core Team 2008) and the functions `princomp` for principle components analysis, `hclust` for cluster analysis, `heatmap` for illustrating cluster analysis results in a heat map, `glm` for logistic regression (all available in the standard `stats` package) as well as the function `cforest` for random forests (available in the add-on package `party`, Hothorn et al. 2006).

3 Results

Because of the fact that the two SATS scales Affect and Cognitive Competence were translated into a different language and were used separately from the rest of the instrument, that had been evaluated as a whole by Schau et al. (1995), an exploratory item analysis was conducted. (Positively worded item responses were reversed so that a high value indicates a negative attitude towards statistics or even statistics anxiety in the following.)

This section describes the results of the item analysis as well as the construction of the negative attitude indicator, that was then used to identify relevant predictor variables for a negative attitude towards statistics.

3.1 Item Analysis for SATS Scales

Because of the fact that the two SATS scales Affect and Cognitive Competence were translated into a different language and were used separately from the rest of the instrument, that had been evaluated as a whole by Schau et al. (1995), an exploratory item analysis was conducted. (Positively worded item responses were reversed so that a high value indicates a negative attitude towards statistics or even statistics anxiety in the following.)

The items showed satisfactory correlations, but revealed an interesting pattern in an exploratory principal components analysis: When we look, e.g., at the factor loadings of the two principal components for all items in Figure 1, we find that some items deviate from the pattern. Those items that do not group with the items from the same scale (Affect vs. Cognitive Competence as indicated by circular and triangular symbols) are all positively worded items (indicated by filled symbols).

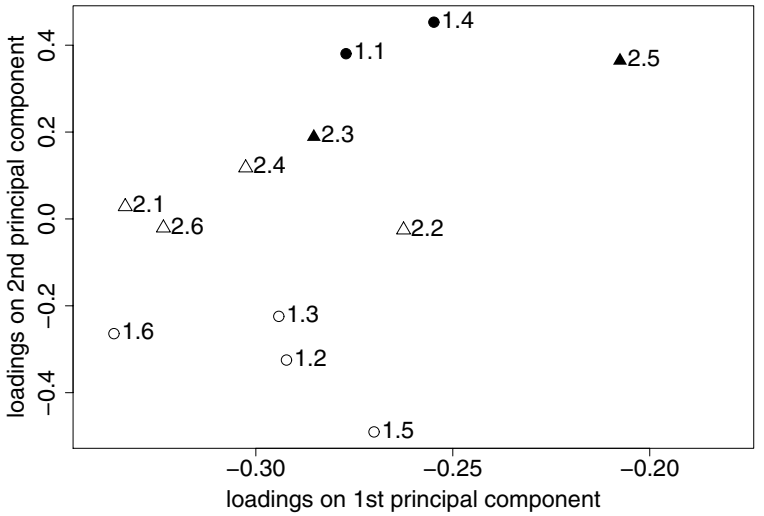


Fig. 1 Loadings on the first two principal components for each of the 12 STAS items. Affect items are indicated by circular and Cognitive Competence items by triangular symbols. Positively worded items are indicated by filled, negatively worded items by unfilled symbols.

The positively worded items that are separated from the rest are the items 1.1 ("I will like statistics.", translated as "Ich werde Statistik mögen.") and 1.4 ("I will enjoy taking statistics courses.", translated as "Ich werde Spaß am Statistik Unterricht haben.") of the Affect scale and the item 2.5 ("I can learn statistics.", translated as "Ich kann Statistik lernen.") of the Cognitive Competence scale, while the positively worded item 2.3 ("I will understand statistics equations.", translated as "Ich werde die statistischen Formeln verstehen.") of the Cognitive Competence scale is situated closer to the negatively worded items of this scale in the principal component loadings plot.

Further investigation of the reversely worded items, not displayed here, revealed that their empirical distributions tended to be less skewed than those of the other items, indicating that less subjects reported extreme values in these items. Correspondingly the discriminatory power, measured by the t-statistic of the comparison between the upper and lower quartile, was found to be lower for the reversely worded items. Note also that the positively worded items tend to group late with the other items in the cluster analysis presented later.

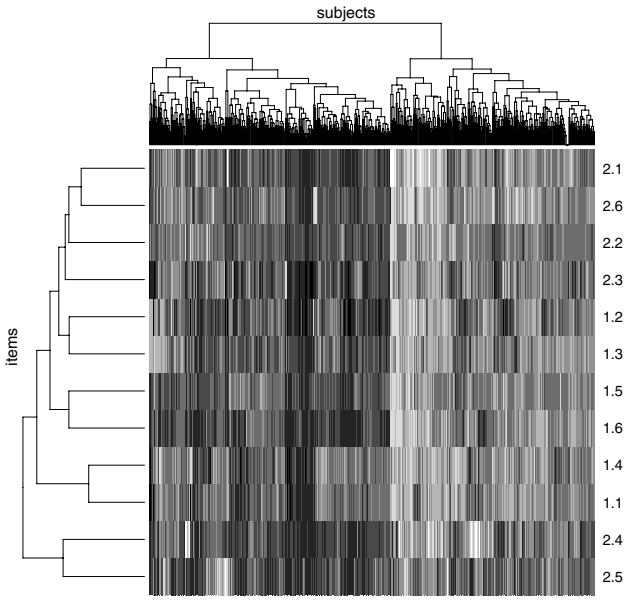


Fig. 2 Heatmap of the dendrogram of the 12 STAS items against the dendrogram of the subjects. Light segments indicate a highly negative attitude towards statistics.

3.2 Negative Attitude Indicator

Because the assumption of a Likert-scale for the entire instrument might not be justified for the two sub-scales of the SATS used here, our aim was to identify groups of students with a similar answering pattern, that could be used as categories of a working response for identifying predictors associated with certain attitudes towards statistics.

As depicted in the heatmap in Figure 2, when the subjects are hierarchically clustered with respect to the distances in their item response patterns we can visually identify two groups of students: The left cluster with a tendency towards highly positive attitudes towards statistics (as indicated by the majority of dark segments) and the right cluster with a tendency towards highly negative attitudes or anxiety towards statistics (as indicated by the majority of light segments).

This binary division will be used in the following as an easy and intuitive dichotomization of the working response. For ease of description, the right cluster containing $45,86\%$ of the sample will be labeled "anxious" in the following, while the left cluster will be labeled "not anxious". Note that this division is not based on an arbitrarily chosen threshold or an expected a priori percentage of anxious students, but is data driven and reflects actual differences in the response patterns of the subjects in the underlying sample. In return, this means that the partition is not meant to be transferred to other samples as a diagnostic for statistics anxiety.

3.3 Predictors of a Negative Attitude towards Statistics

Our aim is to identify predictor variables associated with a negative attitude towards statistics. Usually, logistic regression would be employed for this task when a binary indicator, as the one created from the SATS items in the previous section, is used as a response variable. However, here we are particularly interested in, and have to assume, complex interactions between the predictors – and even after dichotomizing the items, a logistic regression model is not feasible (or instable when forward selection is used) when we want to include interactions of order higher than two. Therefore we use random forests to preselect a set of variables that are associated with a negative attitude towards statistics. An advantage of random forests in this context is that their variable importance measures reflect not only the main effect of a variable but also any effects it has in interactions of potentially high order with other variables (Lunetta et al. 2004).

Another advantage of using random forest variable importance measures for variable selection is that the results are not affected by the kind of instability caused by order effects that affects stepwise variable selection approaches for, e.g., logistic regression. Therefore Rossi et al. (2005) use random forest variable importance measures to support the stepwise variable selection approaches of logistic regression to identify relevant predictors that determine once-only contact in community mental health service. On the other hand, the random forest variable importance of a predictor is not interpretable with respect to the form or direction of the association like a coefficient estimate in a parametric model. Therefore, we will later return to a logistic regression model to regain some interpretability, even though this model has to be limited to two-fold interactions for feasibility and may not sufficiently represent higher-order interactions.

Note that, from a theoretical point of view, the results of Leeb & Pötscher (2006) indicate that statistical inference on model parameters after variable selection on the same data set is not reliable. This problem is not limited to variable selection with random forests, but we recommend that the coefficients of the logistic regression model presented below be interpreted as merely descriptive indicators of the direction of effect of the predictor variables on the negative attitude indicator.

A random forest was created with the model parameters set such as to guarantee unbiased variable selection (cf. Strobl et al. 2007) and $\sqrt{p}$, where p is the number of potential predictor variables, randomly preselected variables (parameter `mtry`). The number of trees in the random forest was set to 1000 to guarantee highly stable results. The resulting variable importance values are depicted in Figure 3.

By visual inspection we decided to include all predictors that exceeded a nominal variable importance value of 0.001 in the further analysis. This conservative threshold seems well suited to separate between predictors whose importance varies only randomly around zero from those that systematically exceed zero: The fact that below this threshold positive and negative values occur with comparable amplitude indicates that this degree of variation is only du to random sampling. Alternative approaches for variable selection with random forests (such as Diaz-Uriarte & Alvarez de Andrés 2006, Breiman & Cutler 2008, Rodenburg et al. 2008) may appear

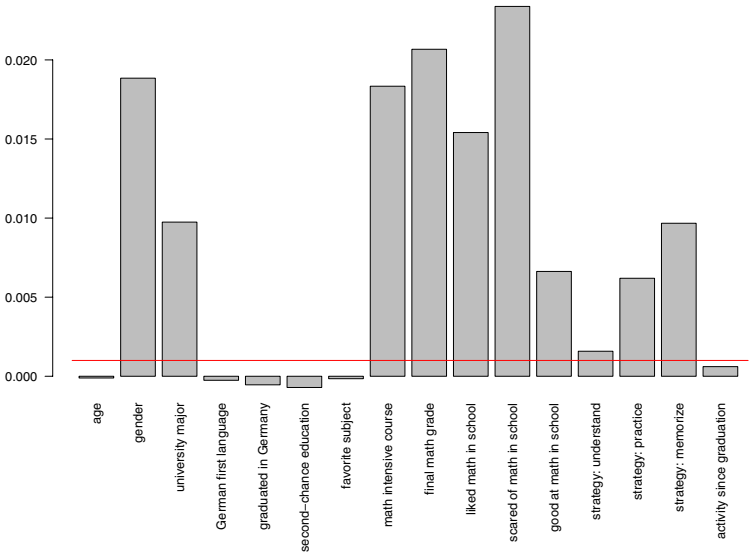


Fig. 3 Variable importance measures computet by means of a random forest for potential predictors of a negative attitude towards statistics. Small variations around zero can be due to random sampling.

more statistically advanced, but have been shown to be affected by undesired artifacts (Strobl & Zeileis 2008, Strobl et al. 2008)

The importance of one strategy item only marginally exceeds the conservative threshold. However, we decided to employ a conservative variable selection strategy and include all strategy items in the further analysis.

Variables whose importance exceed the threshold are: Gender, university major, math intensive course in high school, mathematics grade in final exams in high school, the prior experience items “I liked mathematics in school.”, “I was scared of mathematics in school.” and “I was good at mathematics in school.” as well as the strategy items “My strategy in math was to try to understand the underlying concepts.”, “... was to practice on as many problems as possible.” and “... was to memorize as much as possible.”. The strategy item “My strategy in math was to try to understand the underlying concepts.” only marginally exceeds the threshold and will later turn out to be excluded from the logistic regression model, while the other variables by far exceed it.

This list of variables is now further explored in a logistic regression model that is determined by means of combined forward and backward stepwise selection based on the AIC criterion. Due to the limited amount of data it is not possible to allow for interactions of high order, as automatically reflected by a recursive partitioning method like random forests. However, in order to grasp as many interactions as possible, all two-fold interactions were included in the stepwise selection process. A

Table 1 Summary of the logistic regression model determined by means of combined forward and backward stepwise selection. (Note that the variable gender is an indicator of female gender, that a nominally "high" grade in the German high school system indicates poor performance, and that the item "I liked mathematics in school." was reversed so that approval now indicates a negative attitude.)

	Estimate	Std. Error
(intercept)	−1.18	0.39
gender (female)	0.86	0.20
math intensive course	−0.89	0.22
final math grade (poor)	0.15	0.15
liked math in school (reversed)	0.06	0.28
scared of math in school	3.26	1.13
strategy: practice	−0.88	0.50
strategy: memorize	0.68	0.31
gender (female) × scared of math in school	−1.11	0.71
math intensive course × liked math in school (reversed)	1.30	0.54
final math grade (poor) × scared of math in school	−0.47	0.33
final math grade (poor) × strategy: practice	0.54	0.20

summary of the resulting logistic regression model is given in Table 1. As compared to the model derived by means of combined forward and backward stepwise selection presented here, the model derived from mere forward stepwise selection included the same main effects but excluded one interaction. The model derived from mere backward stepwise selection was much less sparse and included additional main effects of major subject in university and the strategy item "My strategy in math was to try to understand the underlying concepts." as well as additional interactions with the prior experience in school items. Of the two sparse models the one resulting from combined forward and backward stepwise selection had a slightly higher prediction accuracy, cf. Table 2. Therefore this model is presented in detail here.

Note that variables that had high random forest variable importance values, but show no or only little relevance in the logistic regression model, may work in higher order interactions that cannot be reflected in the logistic regression model because many combinations of predictor variable levels are too sparse to estimate effects with reasonable estimation error. One such example may be the final mathematics grade, that shows a particularly high variable importance in the random forest but only a moderate effect in the logistic regression model with main effects and two-fold interactions. Also the major subject in university, that showed a decent variable importance in the random forest, was not included in the logistic regression model that resulted from the combined forward and backward selection. However, the variable was included in the more extensive backward selection model. The same holds for the strategy item "My strategy in math was to try to understand the underlying concepts.", that showed only a small variable importance in the random forest but was also included in the extensive backward selection model. These differences in the logistic regression models determined by stepwise model selection with different starting models can be considered as an indicator of instability in the stepwise selection process due to order effects as pointed out by Rossi et al. (2005).

Table 2 Prediction accuracies on learning data for all considered models. The out-of-bag prediction accuracy of the random forest model gives a realistic estimation of the prediction accuracy on a test data set.

Model	Accuracy
random forest	71.1%
random forest (out-of-bag)	63.3%
logistic regression (forward)	68.6%
logistic regression (backward)	70.4%
logistic regression (combined)	69.3%

The prediction accuracies for the different models considered here are compared in Table 2. The random forest model shows the highest prediction accuracy on the learning data set. Of the logistic regression models the least sparse model resulting from backward selection has the highest prediction accuracy on the learning data as expected. However, in order to give an idea of the prediction accuracy that could be achieved on a new test data set from the same data generating process, random forests offer the possibility to compute the so called "out-of-bag" prediction accuracy: For each tree in the random forest the prediction accuracy is assessed only for those observations that were not included in the bootstrap sample on which the tree was built. Together these observations form a "built-in" test sample for the random forest.

4 Discussion and Conclusion

Overall the results of the item analysis show that the items of the SATS Affect and Cognitive Competence scales, even when translated into German, have satisfactory correlations and are well suited for partitioning the sample with respect to attitude towards statistics. Interestingly the pattern found here, that (at least when translated into German) the positively worded items of the SATS produce different response patterns than the negatively worded items, had not been reported before. The negatively worded items seem to have the tendency to be slightly more selective in the English version of the instrument as well (Schau 2007). However, the effect seems to be more pronounced in the German translation.

To diminish the possibility of translation errors a fluent German and native English speaker was asked to review the translations. Only for the item 1.4 ("I will enjoy taking statistics courses.", translated as "Ich werde Spaß am Statistik Unterricht haben.") the German translation may have a more positive meaning than the English version. However, the item 2.5 ("I can learn statistics.", translated as "Ich kann Statistik lernen."), the translation of which is literal, showed a much stronger deviation from the negatively worded items in the exploratory item analysis. Therefore we believe that the difference between positively and negatively worded items is not due to translation errors but rather due to a different perception or interpretation by the subjects: Our hypothesis is that the positively worded items are perceived much

more positive by German students in the German version than by American students in the English version, so that the German students find the phrases overstated and cannot identify with them – and therefore respond only moderately. Besides affirming popular stereotypes and scientifically documented patterns (Schroll-Machl 2002) concerning a general lack of enthusiasm in the German style of communication, this finding indicates that a direct translation of attitude scales into a different language does not necessarily reproduce the characteristics of the original instrument.

The results of the logistic regression model indicate in the main effects that female students, students who had bad grades in their final mathematics exams, students who were scared of mathematics in school and students whose strategy was to memorize as much as possible are more likely to be anxious, while those that took mathematics as an intensive course in high school and whose strategy was to practice as much as possible are less likely to be anxious of statistics. In the interactions we find that the protective effect of mathematics as an intensive course was outweighed in the interaction with a negative attitude in response to the item "I liked mathematics in school." so that overall students who took mathematics as an intensive course in high school but did not like it are more likely to be anxious of statistics in university. Also the effect of the final mathematics grade was aggravated in interaction with the strategy to practice as much as possible, so that students who (at least claim to have) practiced a lot but still received poor mathematics grades in high school are more likely to be anxious of statistics in university.

On the other hand the interactions between gender and approval to the item "I was scared of mathematics in school." as well as between the final mathematics grade and approval to the item "I was scared of mathematics in school." merely soften the effect so that students that share both characteristics are not as likely to be anxious of statistics in university as would be indicated by the additive effects of the corresponding items.

Even though we included respective items in our survey and the sample size (with 28% of our sample reporting that they did not go directly from high school or civil/military service to university) would have been sufficient to detect an effect present in the sample, we could not replicate the effect of the time since the last math course reported in Wilson (1997) and our expectation that the time and activity since high school graduation affected the attitude towards statistics in university was not supported. On the other hand the effects of gender and poor prior achievement found here replicate the results of Wilson (1997), Fullerton & Umphrey (2001), Mills (2004) as well as Zeidner (1991) and Carmona (2004), and indicate that female students and students with negative prior mathematics experience are more likely to develop a negative attitude towards statistics or even statistics anxiety in university.

The effects of the newly included items on strategy applied in previous math courses for predicting a negative attitude towards statistics support our previous impression that the students' strategy is correlated with their negative attitude towards statistics. Practicing has a protective effect – as long as this strategy is successful and does not go along with poor grades, as indicated by the respective interaction effect. This protective effect of practicing could either be that practicing helps students feel more secure about their abilities and the subject matter, or that anxious students rather

choose other suboptimal strategies, such as memorizing, instead of practicing for a test. The effect of memorizing, on the other hand, could either be that it increases anxiety because memorized knowledge is not as reliable as knowledge achieved by understanding or practicing, or that memorizing is used as a strategy only by those students who are so anxious that they see no other chance to pass an exam.

For statistics instructors in university it may be scary but insightful that a non-negligible percentage of their students come from high-school with suboptimal preparation strategies: about 12.35% of the students in our sample (16.88% in social and 9.51% in business sciences) agreed or strongly agreed to using memorizing as their strategy in math courses.

Teaching students such successful strategies for mastering mathematics courses – as early as possible in their school career, before they accustom to suboptimal learning strategies – could thus serve both as an intervention against a negative attitude towards statistics or even statistics anxiety as well as for improving the students' long time statistics achievement.

In a follow-up study the exam grade achieved and the attitude towards statistics at the end of the first introductory course will be investigated for those students who agreed to participate by providing their matriculation number. The aim of this follow-up is i) to assess the influence of a negative attitude towards statistics or statistics anxiety on test performance and compare the results from this sample to previous findings in the literature (Zeidner 1991, Zimmer & Fuller 1996), and ii) to further investigate if the conservative answering tendency in the positively worded SATS items in persistent or moderated in the second presentation.

Acknowledgements Christian Dittrich, Christian Seiler and Sandra Hackensperger were students at the LMU Department of Statistics and analyzed the data of this study during their course "Statistisches Praktikum". They have attended several lectures given by Ludwig Fahrmeir and – as for so many of us – their attitude towards statistics has highly profited from his teachings.

References

Ashcraft, M. A. & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance, *Journal of Experimental Psychology: General* **130**(2): 224.

Baloglu, M. (2003). Individual differences in statistics anxiety among college students, *Personality and Individual Differences* **34**(5): 855–865.

Breiman, L. & Cutler, A. (2008). Random forests – Classification manual. Website accessed in 1/2008; http://www.math.usu.edu/∼adele/forests.

Carmona, J. (2004). Mathematical background and attitudes toward statistics in a sample of undergraduate students, *Paper Presented at the 10th International Congress on Mathematical Education (ICME 2004), Copenhagen, Denmark* .

Cruise, R. J., Cash, R. W. & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety, *Proceedings of the 1985 Statistical Education Section of the American Statistical Association, Las Vegas, NV, USA*, pp. 92–98.

Diaz-Uriarte, R. & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest, *BMC Bioinformatics* **7:3**.

Dreger, R. M. & Aitken, L. R. (1957). The identification of number anxiety in a college population, *Journal of Educational Psychology* **48**: 344–351.

Fullerton, J. A. & Umphrey, D. (2001). An analysis of attitudes toward statistics: Gender differences among advertising majors, *Paper presented at the 84th Annual Meeting of the Association for Education in Journalism and Mass Communication 2001, Washington, DC, USA* .

Fullerton, J. A. & Umphrey, D. (2002). Statistics anxiety an math aversion among advertising students, *Journal of Advertising Education* **6**(2).

Galagedera, D., Woodward, G. & Degamboda, S. (2000). A investigation of how perceptions of mathematics ability can affect elementary statistics performance, *International Journal of Mathematical Education in Science and Technology* **31**(5): 679–689.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework, *Journal of Computational and Graphical Statistics* **15**(3): 651–674.

Leeb, H. & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators?, *The Annals of Statistics* **34**(5): 2554–2591.

Lunetta, K. L., Hayward, L. B., Segal, J. & Eerdewegh, P. V. (2004). Screening large-scale association study data: Exploiting interactions using random forests, *BMC Genetics* **5:32**.

Mills, J. (2004). Students' attitudes toward statistics: Implications for the future, *College Student Journal* **38**(3): 349–361.

Onwuegbuzie, A. J. (2001). Statistics anxiety and the role of self-perceptions, *Journal of Educational Research* pp. 323–330.

Onwuegbuzie, A. J., Slate, J. R., Paterson, F. R. A., Watson, M. H. & Schwartz, R. A. (2000). Factors associated with achievement in educational research courses, *Research in the Schools* **7**: 53–65.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

Rodenburg, W., Heidema, A. G., Boer, J. M., Bovee-Oudenhoven, I. M., Feskens, E. J., Mariman, E. C. & Keijer, J. (2008). A framework to identify physiological responses in microarray based gene expression studies: Selection and interpretation of biologically relevant genes, *Physiological Genomics* **33**(1): 78–90.

Rossi, A., Amaddeo, F., Sandri, M. & Tansella, M. (2005). Determinants of once-only contact in a community-based psychiatric service, *Social Psychiatry and Psychiatric Epidemiology* **40**(1): 50–56.

Schau, C. (2007). personal correspondence.

Schau, C., Stevens, J., Dauphinee, T. L. & Vecchio, A. D. (1995). The development and validation of the survey of attitudes toward statistics, *Educational and Psychological Measurement* **55**(5): 868–875.

Schroll-Machl, S. (2002). *Die Deutschen - Wir Deutsche*, Vandenhoeck & Ruprecht, Göttingen.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests, *BMC Bioinformatics* **9:307**.

Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics* **8:25**.

Strobl, C. & Zeileis, A. (2008). Danger: High power! – Exploring the statistical properties of a test for random forest variable importance, *Proceedings of the 18th International Conference on Computational Statistics, Porto, Portugal*.

Wilson, V. (1997). Factors related to anxiety in the graduate statistics classroom, *Paper presented at the Annual Meeting of the Mid-South Educational Research Association 1997, Memphis, TN, USA*.

Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics, *Educational and Psychological Measurement* **45**(2): 401–405.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels, *British Journal of Educational Psychology* **61**(3): 319–328.

Zimmer, J. C. & Fuller, D. K. (1996). Factors affecting undergraduate performance in statistics: A review in literature, *Paper Presented at the Annual Meeting of the Mid-South Educational Research Association, Tuscaloose, AL, USA* .

Graphical Chain Models and their Application

Iris Pigeot, Stephan Klasen and Ronja Foraita

Abstract Graphical models are a powerful tool to analyze multivariate data sets that allow to reveal direct and indirect relationships and to visualize the association structure in a graph. As with any statistical analysis, however, the obtained results partly reflect the uncertainty being inherent in any type of data and depend on the selected variables to be included in the analysis, the coding of these variables and the selection strategy used to fit the graphical models to the data. This paper suggests that these issues may be even more crucial for graphical models than for simple regression analyses due to the large number of variables considered which means that a fitted graphical model has to be interpreted with caution. Sensitivity analyses might be recommended to assess the stability of the obtained results. This will be illustrated using a data set on undernutrition in Benin.

1 Introduction

The selection of an adequate model is a crucial task when modeling complex association structures. The results of a particular analysis about direct and indirect effects of covariates on response variables and the corresponding substantive conclusions can be strongly affected by the choice of the underlying model. Each of the candidate models has advantages but also limitations that impact the most relevant questions to be answered by the analysis, namely how do the variables involved affect our

Iris Pigeot, Ronja Foraita
Bremen Institute for Prevention Research and Social Medicine (BIPS), University of Bremen, D-28357 Bremen, Germany, URL: www.bips.uni-bremen.de,
e-mail: pigeot@bips.uni-bremen.de

Stephan Klasen
Department of Economics, University of Göttingen, D-37073 Göttingen, Germany, URL: www.uni-goettingen.de/en/64094.html,
e-mail: sklasen@uni-goettingen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_13,

outcomes of interest and are there possibly interactions between them which also influence our response variable. In addition to the choice of the statistical model itself, the selection and the coding of the variables to be included in the analysis is another critical aspect in an empirical investigation.

In this paper we focus on the application of graphical models that are still a rather novel, though powerful statistical tool to analyze multivariate data sets. Graphical models are specifically suited for the analysis of complex association structures and provide a graphical representation of certain independence properties among the variables of interest. We restrict ourselves here to so-called graphical chain models that allow to reveal indirect associations and to identify hidden relationships. We demonstrate the challenges that are related to the interpretation of graphical chain models by especially investigating their robustness with respect to a change of the variables included in the analysis or a different coding.

The challenges that are illustrated by using a highly complex data example are of course not limited to these statistical techniques. The complexity of the model, however, adds to the difficulties which are inherent to any statistical modeling approach in an empirical investigation.

We will illustrate the above mentioned challenges in obtaining valid and meaningful results by considering the example of childhood undernutrition which is one of the most important health problems in developing countries. That is we are interested in modeling the determinants of undernutrition among children which is a complex undertaking. Although it seems as if the determinants of undernutrition are quite clear, namely inadequate dietary intake and incidence, severity, and duration of disease, these factors themselves are related to a large number of intermediate, underlying, and basic causes operating at the household, community, or national level (UNICEF 1998). Among the most important factors are probably the education, wealth, and income situation of the parents, household size, birth order, religion, and sex of the child, the availability of clean water, adequate sanitation, immunization, and primary health care services, and the level of disease prevalence in the surrounding community. Noteworthy, the association structure between these factors is assumed to be fairly complex. In fact, UNICEF has made a useful distinction between immediate, intermediate, and underlying causes of undernutrition. Any empirical strategy that attempts to identify the determinants of undernutrition must recognize the existence of such a dependence chain. This suggests that a simple multivariate regression model is not appropriate to capture the indirect associations and the overall complex association structure.

To determine whether an individual child suffers from undernutrition of the three forms, i.e. insufficient height for age (stunting) indicating chronic undernutrition, insufficient weight for height (wasting) indicating acute undernutrition, and insufficient weight for age (underweight) indicating acute and/or chronic undernutrition, the anthropometric indicator of the child is compared with a reference population by means of a Z-score:

$$Z_i = \frac{AI_i - MAI}{\sigma}$$

where AI_i refers to the individuals anthropometric indicator (weight at a certain height, height at a certain age, or weight at a certain age), MAI refers to the median of the reference population, and σ refers to the standard deviation of the reference population (Gorstein et al. 1994, WHO 1995). The Z-score thus measures the distance, expressed in standard deviations of the reference population, between the individuals anthropometry and the median of the reference population, where both populations are presumed to be normally distributed. While the average Z-score is likely to give an accurate picture of undernutrition at the population level, for an individual child it might be misleading as genetic influences of the parents are likely to affect it and thus bias the findings. There is also some on-going debate whether this might bias findings on undernutrition between different continents as there might be genetic differences particularly in the height potential of children (WHO 1995, Klasen 2003, Klasen 2008). In particular, there is a question whether the very high reported rates of undernutrition in South Asia, compared to Sub-Saharan Africa, are partly related to this question.

For the purpose of this paper we use data from the 1996 Demographic and Health Survey to fit a graphical chain model for undernutrition in Benin, West Africa. Our discussion of the challenges related to the modeling of the association structure will be based on two analyses conducted by the authors (Caputo et al. 2003, Foraita et al. 2008). The research work on the analysis of undernutrition with the help of graphical chain models was started within a subproject of the DFG-funded Collaborative Research Center 386 “Statistical Analysis of Discrete Structures: Modelling and Application in Biometrics and Econometrics" which was successfully coordinated by Ludwig Fahrmeir. In the first analysis we made full use of the data being available for Benin. In contrast, the second analysis had to be restricted to variables that were available for both Benin and Bangladesh for comparative reasons. The comparison of these two analyses sheds some light on the differences in the results explaining undernutrition in Benin. In addition to discussing the major differences between the two resulting models we will also describe the most important overlaps.

The paper is organized as follows. Section 2 gives an introduction to the theory of graphical chain models. In Section 3 we briefly describe the selection strategy we used for fitting such a model to our multivariate data set. We then present the data set in Section 4 where we also provide some descriptive statistics. Section 5 gives a detailed discussion of the results, while Section 6 concludes.

2 Graphical Chain Models

Graphical models are probability models for multivariate observations to analyze and visualize conditional relationships between random variables encoded by a conditional independence graph. In contrast to regression models, graphical modeling is concerned with identifying association structures for all study variables, including those which usually are regarded as explanatory. They are therefore appropriate in situations where complex associations have to be dealt with. Due to the visualization

in graphs, these models make it easier to display complex dependence structures. Furthermore, they can handle simultaneously categorical and continuous variables.

We denote an arbitrary graph by $G = (V, E)$ where $V = \{1, \ldots, K\}$ is a set of vertices representing the components of a multivariate random vector $X_V = (X_1, \ldots, X_K)$ and $E \subseteq V \times V$ is a set of edges. For $i, j \in V$, there is a symmetric association between two vertices i and j and a line in the graph (also called undirected edge) if $(i, j) \wedge (j, i) \in E$ whereas $(i, j) \in E \wedge (j, i) \notin E$ corresponds to an asymmetric association and an arrow in the graph (also known as directed edge), pointing from i to j. Semi-directed cycles are not allowed, i.e. sequences $a = i_0, \ldots, i_r = a$ with $(i_{k-1}, i_k) \in E \wedge (i_k, i_{k-1}) \notin E$ for at least one value of k.

The structure of the conditional relationships among random variables can be explored with the help of Markov properties (Lauritzen & Wermuth 1989, Frydenberg 1990). For instance, the pairwise Markov property claims

$$X_i \perp\!\!\!\perp X_j | X_{V^* \setminus \{i,j\}} \text{ whenever } (i, j), (j, i) \notin E$$

where V^* consists of all variables prior to or at the same level as i and j and the symbol $\perp\!\!\!\perp$ stands for conditional independence between X_i and X_j given $X_{V^* \setminus \{i,j\}}$. This implies that a missing edge can be interpreted as conditional independence. However this is only justified if the underlying multivariate statistical distribution fulfills the Markov properties since they lead to a factorization of the multivariate density and thus to a decomposition into smaller models and equivalently cliques, which are maximal complete subgraphs.

Graphical chain models are suitable to account for prior substantial knowledge of an underlying dependence structure by forming a dependence chain where all variables are partitioned into an ordered sequence of disjoint subsets $V_1 \cup \ldots \cup V_R$. The subsets are called blocks and all edges within V_r are undirected and all edges between V_r and V_s are directed from V_r to V_s for $r < s$. The blocks $V_1, \ldots, V_R$ are ordered due to subject-matter knowledge, so that the rightmost block contains the pure explanatory variables, the leftmost block the pure responses and the blocks between contain variables that are simultaneously responses to variables in previous blocks and potentially explanatory to variables in future blocks. Variables in these in-between blocks are intermediates and, in contrast to usual regression models, allow for modeling possibly indirect influences. Variables in the same block are assumed to be on equal footing, i.e. no sensible response-explanatory relationship can be assumed within this subset. The random vector X_V is divided into subvectors $X_{V_1}, \ldots, X_{V_R}$ such that the joint density $f(x_V)$ factorizes into a product of conditional densities as

$$f(x_V) = f(x_{V_1}) \prod_{r=2}^{R} f(x_{V_r}) | x_{V_1}, \ldots, x_{V_{r-1}}). \tag{1}$$

Each of the factors in (1) corresponds to the distribution of variables at one level conditional on variables at all lower levels. Thus, one may regard a graphical chain model as a sequence of regression models that describe these conditional distributions and the choice of the recursive structure reflects that one is specifically interested in the latter.

If mixed models are investigated, i.e. models including continuous as well as discrete variables, the distribution considered is the Conditional Gaussian distribution (CG-distribution), where the continuous variables are multivariate normal given the discrete. For further reading we refer to Lauritzen (1996), Cox & Wermuth (1996), Edwards (2000) and Green et al. (2003) and the references therein.

3 Model Selection

In our study we have to deal with mixed variables and also with a large number of variables. Thus, we are confronted with the problem to select those variables that are most influential for the response and to find the most appropriate association structure among them. A possible solution to this problem is the data-driven Cox-Wermuth selection strategy (Cox & Wermuth 1993, Cox & Wermuth 1994). This strategy exploits that each conditional density of the factorization is described by a system of multiple univariate regressions. The kind of regression used depends on the measurement scale of the involved univariate response. A problem of this strategy is that fitting multiple univariate regressions neglects the multivariate structure of the data and the validity of the equivalence of the Markovian properties is not ensured for the whole graph. Nevertheless, for large and complex graphs with mixed variables it is still the only feasible computer algorithm which is implemented in the software GraphFitI (Blauth et al. 2000).

The Cox-Wermuth selection strategy consists of roughly two steps: First, a screening for second-order interactions and non-linearities is performed (Cox & Wermuth 1994); second, a system of forward and backward regressions depending on the scale of the response variable is carried out.

In the screening procedure, the search for second-order interactions is based on the calculation of t-statistics derived from trivariate regressions, such as X_a on $X_b, X_c, X_b X_c$ with $X_a \in V_s$ and $X_b, X_c \in V_r, s \geq r$, where each X_a has to be regressed on all possible pairs of variables in the same block and in previous blocks as well as on their pairwise interaction. In case of large sample sizes and if there is no interaction, the t-statistics approximately follow a standard normal distribution. The ordered t-statistics are plotted against their expected values obtained from the standard normal distribution. If the assumption of no interactions is fulfilled, the points spread along the diagonal. Checking for non-linearities is performed similarly. All interactions and non-linearities with a $|t|$-value > 4 are considered in further steps.

To derive the graph a multivariate response model is needed for each V_r given $V_1 \cup \ldots \cup V_{r-1}$. The Cox-Wermuth strategy splits the problem of multivariate regressions into a system of univariate regressions for each variable X_a on the remaining variables in the same block and on all explanatories in the previous blocks. First, a forward selection investigates whether the detected interactions or non-linearities from the screening step have to be added into the set of covariates regarding X_a. This selection is based on statistical tests with $\alpha = 0.1$. The corresponding p-values have, however, to be interpreted in an exploratory sense since no adjustment for multiplicity takes

place. Then, a backward selection strategy for X_a is used on the preliminary set of covariates. In each step, the covariate with the smallest corresponding $|t|$-value is excluded until the remaining covariates all come up with a p-value smaller than 0.05. After that the remaining variables are checked again for interactions and non-linearities. All qualitative interaction terms and mixed interactions terms are included in the model equation. Again, a backward selection as described above is carried out. Finally, all quantitative interaction terms and non-linearities are introduced into the model. The final backward selection leads to the reduced model that should capture the underlying association structure.

4 Data Set

The data is part of the 1996 Demographic and Health Surveys (DHS, Macro 1996). These surveys are conducted regularly by the National Statistical Institutes in collaboration with Macro International, a US-based company that operates on behalf of the US Agency for International Development, in several countries of Africa, Asia, Latin America and the Near East. The DHS is based on a representative sample of women of reproductive age. These women are administered an extensive questionnaire covering a broad range of items regarding household structure, socioeconomic status, health access and behavior, fertility behavior, reproductive health, and HIV/AIDS. The questionnaire also contains items about the children including prenatal and postnatal care, nutrition, health, immunization, and care practices. Some parts or questions of the survey have been disregarded in some countries. In this study we focus on the DHS data set from Benin and involve only children between twelve and 35 months. We focus on these age group since by that age the children surely have been introduced to additional foods and water and therefore have already been through the weaning crisis associated with this transition. For older children the DHS survey does not collect data about their nutrition and health status. If the respondent has more than one child belonging to this age group we only select the younger one.

We compare the data sets from Caputo et al. (2003) and Foraita et al. (2008) which in this paper are abbreviated with *A* and *B* respectively. While Caputo et al. (2003) only focus on Benin, Foraita et al. (2008) compare the different patterns of malnutrition in Benin and Bangladesh. Although both papers are based on the DHS 1996 Benin data set, they vary in the variables that causes different total sample sizes ($N_A = 1076, N_B = 1122$) since both data sets are constrained to complete cases. Additionally, some variables have a different coding scheme or the reference category has changed (see Section 4.1 for more details).

4.1 Summary Measures

In order to capture the determinants of undernutrition and not to miss a relevant influence, a large number of variables are included in the model. In this section, we briefly introduce the variables, their scales and coding. Table 1 gives absolute and relative frequencies of binary and polytomous variables and Table 2 summarizes mean, median and the 25^{th} and 75^{th} percentile of continuous variables. This distinction between the various scales is not only convenient for their presentation, but also needed for choosing the adequate regression models in later analysis.

The response variables *stunting* (*St*) and *wasting* (*Wa*) are both continuous anthropometric indicators that measure malnutrition using the *Z*–score. Stunting reflects chronic malnutrition, whereas wasting stands for acute malnutrition (see Section 1). The different composition of the two analysis data sets has already an impact on the stunting and wasting. To be more specific, in data set *B* the children are slightly more stunted and less wasted than in data set *A*. The investigation of the impact of the child's nourishment focuses on the quality of food, measured by the number of meals containing *protein* (*P*) during one day. In data set *A*, *P* shows the absolute frequency of meals containing milk, meat, egg, fish or poultry whereas in data set *B* only the meals containing milk or meat are counted. The difference in this operationalization heavily affects the distribution of *P*: for data set *A*, 20% of all children had no protein in their meal compared to 65% in data set *B*. Data set *A* contains the further aspect of food quantity (*F*) which counts the number of meals a child has had during the day. Since this variable was not available for the Bangladesh data set, the number of meals was not included in data set *B*. Comparing both data sets to further nutritional variables, no essential difference can be seen with respect to the time when the children are *put to breast* or the duration of exclusively *breast–feeding* that is on average unusually long with about 19 months which has to be interpreted as an indicator of high poverty and lack of alternative food. The security of nutrition for the child is represented by the mother's body mass index *BMI*. A large BMI can be interpreted as sufficient nourishment of the whole family, whereas a low BMI indicates an uncertain nourishment.

Another important influence on the child's physical status is its current health situation. Therefore, the variable *ill* counts children who suffered from diarrhea or cough during the last two weeks before the interview. Due to the short observation time, one may presume an effect on *wasting*.

In both data sets half of the children have to be regarded as ill. In both data sets nearly 78% of the mothers have access to modern health care, measured by *prenatal and birth attendance* score (*BPA*). The variable *vaccination* (*V*) counts the number of vaccinations a child has already had. It may be considered as a substitute of health knowledge, but also of access to health care. Furthermore, the access to clean water and clean sanitation is important. These variables have been operationalized differently for both data sets. In data set *A* only piped water and flush toilet or all kinds of pit latrines has been categorized as high quality (around 24%) compared to data set *B* where piped as well as well water (50%) and only flush toilets but no open latrines (13%) has been regarded as high quality.

Table 1 Absolute and relative frequencies of binary and polytomous variables.

			A ($N = 1076$)		B ($N = 1122$)	
	Variable	Category	Freq	%	Freq	%
P	***protein* intakes yesterday**	0	216	20.1	734	65.4
		1	607	56.4	307	27.4
		2	207	19.2	81	7.2
		3	46	4.3	-	-
ILL	child was *ill* during the last 14 days	no (*A*)	539	50.1	560	49.9
		yes (*B*)	537	49.9	562	50.1
PB	when child *put to breast*	immediately	241	22.4	249	22.2
		within 6 hours	326	30.3	340	30.3
		first day	271	25.2	282	25.1
		2 days or more	238	22.1	251	22.4
BPA	*prenatal and birth attendance*	nothing	28	2.6	29	2.6
		other	133	12.4	136	12.1
		traditional	78	7.3	84	7.5
		modern	837	77.8	873	77.8
W	**source of drinking *water***	low quality	822	76.4	566	50.5
		high quality	254	23.6	556	49.6
T	**type of *toilet* facility**	low quality	876	81.4	972	86.6
		high quality	200	18.6	150	13.4
Rel	*religion*	Islam (*B*)	235	21.8	245	21.8
		Traditional	278	25.84	290	25.8
		Christianity	246	39.6	444	39.6
		no religion (*A*)	137	12.7	143	12.8
Sex	*sex* of child	male	551	51.2	575	51.3
		female	525	48.8	547	48.8
HH	relationship to *household head*	relative	174	16.2	183	16.3
		wife	842	78.3	871	77.6
		head (*B*)	40	3.7	43	3.8
		not related (*A*)	20	1.9	25	2.2
H	*house* quality	low quality	555	51.6	580	51.7
		high quality	521	48.4	542	48.3
Wo	current type of *employment*	paid employee	83	7.1	85	7.6
		self-employed	891	82.8	933	83.2
		unpaid worker (*A*)	44	4.1	45	4.0
		did not work (*B*)	58	5.4	59	5.3

A and *B* mark different reference categories in the respective data sets and bold written variables mark a different coding scheme in both data sets.

Additionally, various socioeconomic factors have been included into both data sets like the current type of *employment* of the mother (*Wo*), having a different reference category in the data sets, three proxies for the economic situation of the household (house quality (*H*), durable goods (*G*) and mother's height (*Ht*)) or the educational

Table 2 Summary measures of the continuous variables.

		A				*B*			
	Variable	Mean	Median	Q_1	Q_3	Mean	Median	Q_1	Q_3
St	*stunting* (*Z*-score·100)	**-147.3**	**-153.0**	**-232.5**	-62.0	**-148.2**	**-154.5**	**-233.0**	-62.0
Wa	*wasting* (*Z*-score·100)	**-93.6**	**-95.0**	**-165.5**	**-20.5**	**-92.7**	**-94.0**	**-164.0**	**-20.0**
F	*food*	**4.2**	**4.0**	**-3.0**	**5.0**	-	-	-	-
BF	duration of *breast-feeding* in months	18.6	18.0	15.0	22.0	18.6	18.0	15.0	22.0
BMI	*body mass index*	**21.3**	20.8	19.2	**22.5**	**21.4**	20.8	19.2	**22.6**
Ht	mother's *height*	158.2	**158.2**	154.0	162.2	158.2	**158.3**	154.0	162.2
V	*vaccination* score	6.2	8.0	5.0	8.0	6.2	8.0	5.0	8.0
Bo	*birth order number*	4.1	4.0	2.0	6.0	4.1	4.0	2.0	6.0
Age	*age* in months	22.5	22.0	16.0	28.0	22.5	22.0	16.0	28.0
HM	no. of *household members*	8.9	8.0	5.0	11.0	8.9	8.0	5.0	11.0
TC	*total children* ever born	**4.2**	**4.0**	**2.0**	**6.0**	-	-	-	-
CD	*deceased children* in %	12.5	0.0	0.0	25.0	12.5	0.0	0.0	25.0
1^{st}	age of mother at *first birth*	19.0	19.0	17.0	21.0	19.0	19.0	17.0	21.0
G	durable *goods* in %	23.8	28.6	14.2	28.6	23.9	28.6	14.3	28.6
EM	*mother's education* in years	0.9	0.0	0.0	0.0	0.9	0.0	0.0	0.0
EP	*partner's eduction* in years	2.3	0.0	0.0	4.0	2.3	0.0	0.0	4.0

Q_1: 25^{th} percentile; Q_3: 75^{th} percentile. Bold figures mark differences in data sets.

background in the family. The duration of education in years is very short for mothers with a 75^{th} percentile of 0 years and it is slightly better for their partners with around 4 years.

4.2 Dependence Chain

In line with UNICEF (1998), Caputo et al. (2003) and Foraita et al. (2008) postulated a dependence chain as follows (see Figure 1), from left to right: in the first block we put our pure response variables *wasting* and *stunting*.

The second block includes all variables that have an immediate influence on these *Z*-scores, where we include variables relating to nutritional intake and illness episodes (*ill*, *protein* and *food* for data set *A*).

The next block includes intermediate variables that reflect care practices, health knowledge, and access to water and sanitary services. It includes child care (*breast-feeding*, time *put to breast*, *vaccination*, *prenatal and birth attendance*, mother's *BMI*) and sanitary facilities (quality of drinking *water*, quality of *toilet* facilities).

The last block on the right includes basic variables affecting the ability of households to take care for their children, including demographic factors (*age*, *sex*, *birth order number* and additionally in data set *A* the variable *total children ever born*, *religion*, number of *household members*, number of *deceased children*, age of mother at *first birth*, relation between mother and *household head*), socioeconomic factors (*house* quality, fraction of durable *goods*, *mother's education*, *partner's education*,

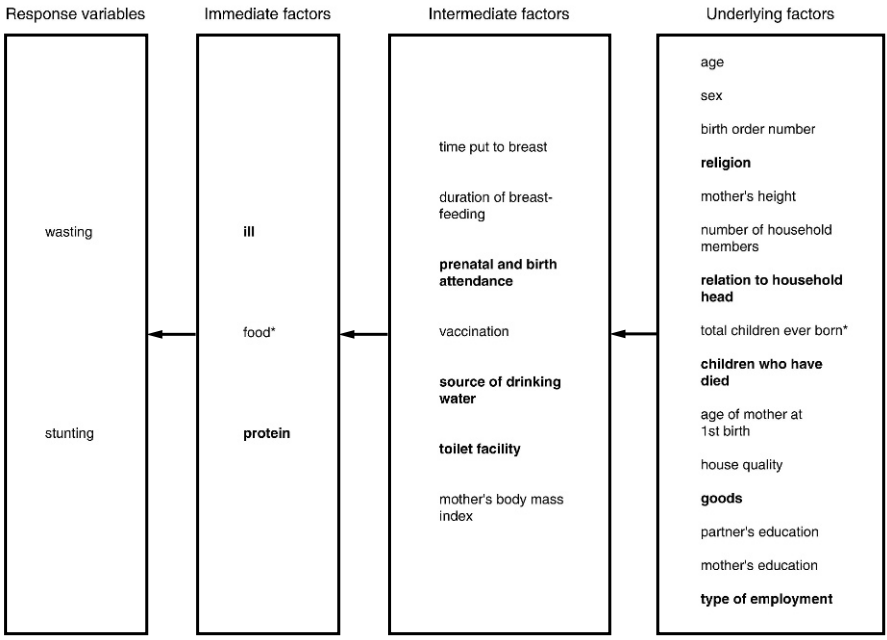


Fig. 1 The postulated chain. Pure responses are the undernutrition variables *wasting* and *stunting*. Immediate factors reflect food quality and the state of health, intermediate factors are variables of health care, health knowledge, food security, and sanitary facilities. Demographic and socioeconomic factors are put in the block of the underlying factors. Bold written variables indicate different coding schemes in both data sets and * indicates variables that are only included in data set *A*.

current type of *employment*) and mother's *height* as combination of socioeconomic and nutritional aspects that we call underlying factors. In the appendix a more detailed description of the variables is given.

5 Results

Figures 2–4 show the fitted graphical chain models for data set *A* and *B* and their common edges. The common edges in both analyses are shown in black in Figures 2 and 3 and are separately presented in Figure 4. The edges that are specific to analysis *A* and *B* are shown in grey in Figures 2 and 3, respectively.

While the figures may at first glance appear rather complicated, closer inspection reveals a number of interesting points.

Although both data sets differ only slightly, the graphs show several notable differences. This is surely on the one hand due to the omission of the variables *total children ever born* and in particular *food* that attracts many influences from the intermediate and underlying factors and on the other hand due to the change of some

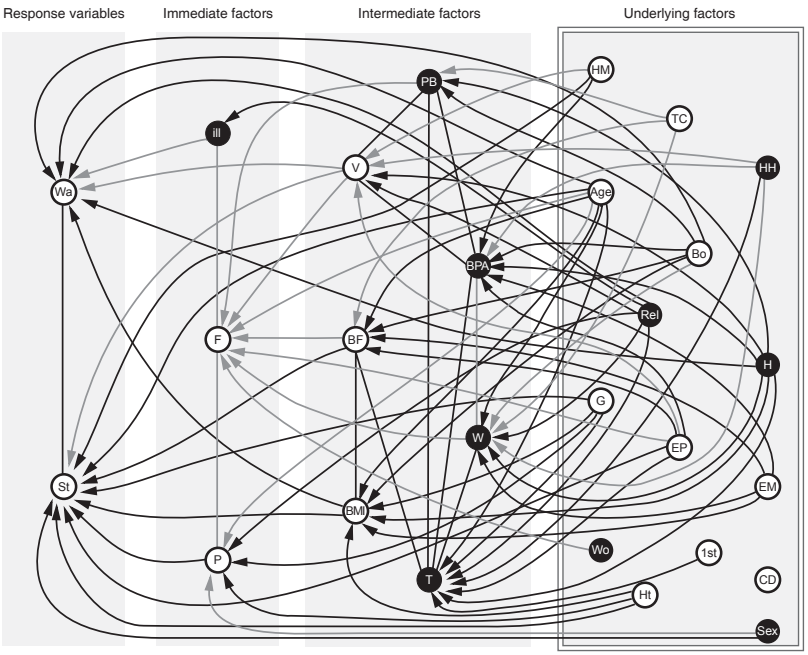


Fig. 2 Undernutrition in Benin – Fitted graphical chain model using data set *A*. Black dots represent discrete and white circles continuous variables. The double-lined box indicates that the associations among the variables within this box are not shown. Edges in black are common to both analyses; edges marked in grey are only specific to the analysis using data set *A*. For abbreviations of variables see Tables 1 and 2 or Figure 4.

categorical variables as well as the inclusion of further children which affects the variability between the data sets. Especially the variable *food* acts as some kind of hub in data set *A* that forwards the influences of many intermediate and underlying factor though its connections to *protein* and *ill*. In data set *B* it seems that *ill* has partly inherited the role since it works as endpoint for many underlying factors, with the difference that these influences are not carried forward to the response variables.

Although we have not substantially changed the data sets, we can see in Figures 2 and 3 that there is a certain amount of uncertainty in the data we have to be aware of. Edges that we detected in both analyses, seem to be more stable. Hence, our interpretation will only be based on those pathways.

First, in both data sets *protein* consumption has a direct influence on *stunting*, even though the variable has been recoded for data set *B*. Second, there are many direct and indirect influences from those variables reflecting the economic condition of the household, e.g. proxied by *partner's eduction*, *house quality* and *goods*. Third, mother's *BMI* and duration of *breast–feeding* are important players as intervening variable between underlying factors and the response variables *stunting* and *wasting*. However, the role of mother's *BMI* in the network is twofold. It can be regarded as

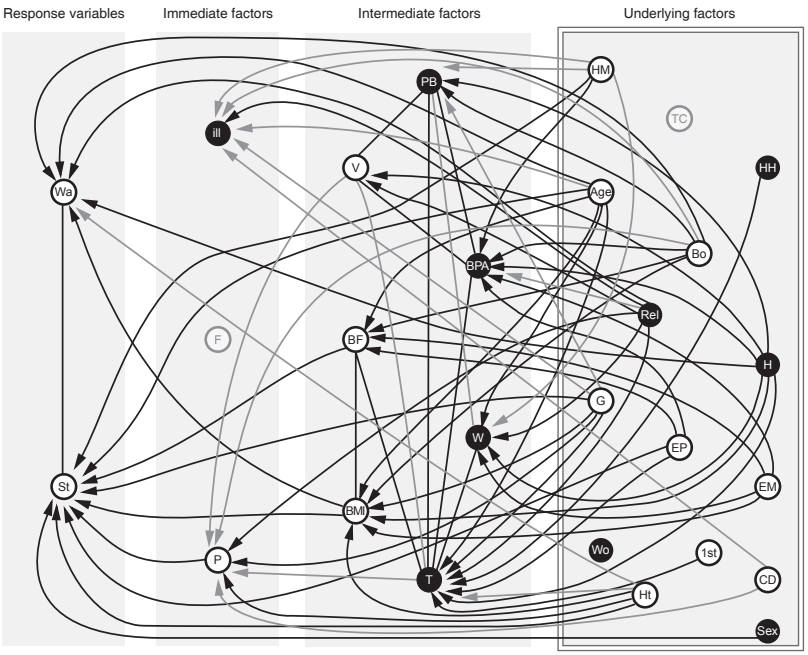


Fig. 3 Undernutrition in Benin – Fitted graphical chain model using data set *B*. The variables food (*F*) and total children ever born (*TC*) are not included in fitting the graph. Edges in black are common to both analyses; edges marked in grey are only specific to the analysis using data set *B*.

a variable that in some extent reflects the economic situation of the household in the sense that children of well-nourished mothers are also well-nourished. But *BMI* may also partly capture the genetic influences of the mother's anthropometry on her children. *Breast–feeding* is directly associated with *stunting* and has an indirect influence on *wasting* through *BMI*. The WHO (WHO 1995) recommends that after six months of exclusively breast–feeding, children need additional food. Hence extended periods of breast–feeding may be an indicator of the inability of the household to provide for such supplemental foods. Table 2 shows that on average the children are breast-fed for more than 18 months. The 25%-percentile equals 15 months. Forth, the eduction of the mother shows a rather indirect influence on the response via *BMI*, duration of *breast-feeding*, source of drinking *water* and *prenatal and birth attendance* which means that especially *stunting* is influenced in numerous ways by *mother's education*. The results suggest that more years of education is associated with a shorter duration of breast-feeding and better nourished mothers. It is more likely that these mothers make use of a modern health service and have access to clean water. Fifth, the *toilet* variable is noticeable since many influences point on it, but *toilet* itself is connected with *stunting* only via the *breast-feeding* link. Sixth, *religion* is an important factor in Benin. It is associated with many aspects in the dependence chain of undernutrition. It is directly linked with *wasting*, the current

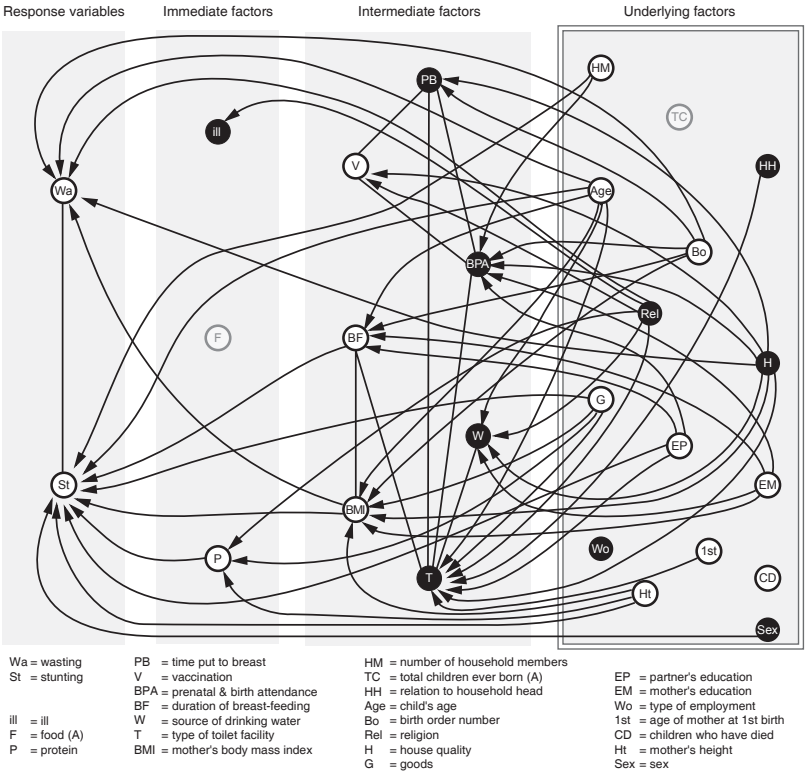


Fig. 4 Common edges in data sets *A* and *B*.

health situation of the child (*ill*), quality of food (*P*), access to health care (*V*) and to the access of clean water (*W*) and sanitation (*T*). Generally, belonging to any kind of religion seems to be favorable for children in Benin.

6 Discussion

As it became evident in the last section, the results that we obtain from a statistical analysis with graphical chain models should be interpreted with caution and critically reflected regarding their substantive message. Although the most important conclusions from the above exploratory data analysis remained stable regardless of the variables included and their coding, some results were rather different in both graphs. This problem is of course more severe in analyses with a huge number of variables involved that all may explain the response directly or indirectly. This is also

not so surprising as the omission of some variables will naturally lead to the effect being captured by closely correlated ones. A related problem is due to the fact that there is always more than one model which is consistent with the data, and typically different model selection strategies will lead to different results.

To get a better understanding of the mechanisms that led to these differences we again fitted a graphical chain model to the original data set *A* where we used the same coding but left out the variables *food* and *total children ever born*. The resulting Figure 5 (see Appendix) differs from Figure 2 mostly in those edges that are due to the above variables; only very few other edges are affected. Figure 3 that is based on data set *B* shows much more differences compared to Figure 5 which means that relatively small differences in coding and number of observations can have a substantial difference on the detected associations. Examples are that *ill* has a direct influence on *wasting*, or that *vaccination* directly influences *stunting* which is each shown in Figure 5, but does not appear in Figure 3. Thus, it is strongly recommended to not only think about the variables to be included but to also carefully think about the coding of variables and to carry out some sensitivity analysis based on various coding schemes.

Whereas a simple linear regression model can be assessed by the coefficient of determination R^2, no comparable measure exists to assess a graphical model. Thus, it is recommended to at least perform some kind of sensitivity analysis. For this purpose, often we suggest to estimate different reasonable models and compare their most important results. As an alternative, the bootstrap offers a valuable opportunity in two ways (Friedman et al. 1999b, Friedman et al. 1999a, Steck & Jaakkola 2004). On the one hand, it allows to generate repeated samples out of the original one which can be used to fit the model of interest repeatedly and to compare the variety of selected models. This gives an idea about the stability of the originally selected one. (For graphical models with only discrete data, the R-package gmvalid (Foraita & Sobotka 2008) provides functions that apply the bootstrap to investigate the uncertainty of graphical models.) The resulting models can, on the other hand, be exploited to derive measures of uncertainty which are especially appropriate to assess the validity of a selected graphical model. The development of such measures and their evaluation by means of real data examples and simulated data sets is currently under research by the authors.

References

Blauth, A., Pigeot, I. & Bry, F. (2000). Interactive analysis of high–dimensional association structures with graphical models, *Metrika* **51**: 53–65.

Caputo, A., Foraita, R., Klasen, S. & Pigeot, I. (2003). Undernutrition in Benin - An analysis based on graphical models, *Social Science & Medicine* **56**: 1677–1697.

Cox, D. R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion), *Statistical Science* **8**: 204–283.

Cox, D. R. & Wermuth, N. (1994). Tests of linearity, multivariate normality and the adequacy of linear scores, *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **43**: 347–355.

Cox, D. R. & Wermuth, N. (1996). *Multivariate Dependencies*, Chapman & Hall, London.

Edwards, D. (2000). *Introduction to Graphical Modelling*, 2 edn, Springer, New York.

Foraita, R., Klasen, S. & Pigeot, I. (2008). Using graphical chain models to analyze differences in structural correlates of undernutrition in Benin and Bangladesh, *Economics and Human Biology* **6**: 398—-419.

Foraita, R. & Sobotka, F. (2008). *gmvalid: Validation of graphical models*. R package version 1.2.

Friedman, N., Goldszmidt, M. & Wyner, A. (1999a). Data analysis with bayesian networks: A bootstrap approach.

Friedman, N., Goldszmidt, M. & Wyner, A. (1999b). On the application of the bootstrap for computing confidence measures on features of induced bayesian networks.

Frydenberg, M. (1990). The chain graph markov property, *Scandinavian Journal of Statistics* **17**: 333–353.

Gorstein, J., Sullivan, K., Yip, R., de Onis, M., Trowbridge, F., Fajans, P. & Clugston, G. (1994). Issues in the assessment of nutritional status using anthropometry., *Bull World Health Organ* **72**: 273–283.

Green, P. J., Hjort, N. L. & Richardson, S. (eds) (2003). *Highly Structured Stochastic Systems*, Oxford University Press, Oxford.

Klasen, S. (2003). Malnourished and surviving in South Asia, better nourished and dying young in Africa: what can explain this puzzle?, *Measurement and Assessment of Food Deprivation and Undernutrition*, FAO, Rome, pp. 283–287.

Klasen, S. (2008). Poverty, undernutrition, and child mortality: Some inter-regional puzzles and their implications for research and policy, *Journal of Economic Inequality* **6**: 89–115.

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

Lauritzen, S. L. & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative, *Annals of Statistics* **17**: 31–57.

Macro (1996). MEASURE DHS datasets Bangladesh, Benin. `www.measuredhs.com`, last accessed: 10. Aug 2009.

Steck, H. & Jaakkola, T. S. (2004). Bias-corrected bootstrap and model uncertainty, *in* S. Thrun, L. Saul & B. Schölkopf (eds), *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA.

UNICEF (1998). *The State of the World's Children: Focus on Nutrition.*, UNICEF, New York.

WHO (1995). Physical status: The use and interpretation of anthropometry, *WHO Technical Report Series 854*, WHO, Geneva.

Appendix

Table 3 Further explanations of some variables. Abbreviations *A* or *B* indicate the respective data set.

Variable		Category	Comments
ill	child was *ill*	no yes	child suffered from diarrhoea or cough during the last 14 days
P	*Protein* intakes yesterday (*A*)	0-3	remembered number of meals containing milk, meat, egg, fish or poultry
P	*Protein* intakes yesterday (*B*)	0-2	remembered number of meals containing milk or meat
W	source of drinking *water* (*A*)	low quality high quality	unprotected well or surface water piped or well water
W	source of drinking *water* (*B*)	low quality high quality	unprotected well or surface water piped water
T	type of *toilet* facility (*A*)	low quality high quality	well water, open latrine, no facility or “other" toilets flush toilet, pit toilet latrine, open latrine
T	type of *toilet* facility (*B*)	low quality high quality	no facility or “other" toilets flush toilet or pit toilet latrine
G	durable *goods* in %	$[0, 1]$	averaged sum score out of if the house has electricity, radio, television, refrigerator, bicycle, motorcycle, car and telephone
H	*house* quality	low quality high quality	main floor material is natural all other materials (i.e. wood, cement ...)

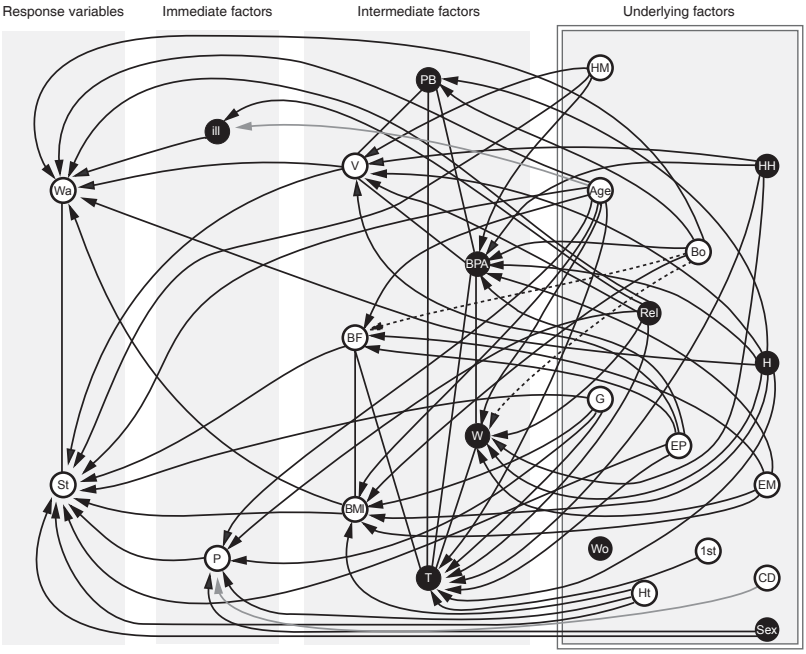


Fig. 5 Fitted graphical model using data set *A* without variables *food* and *total children ever born*. The associations $Age \rightarrow ill$ and $CD \rightarrow P$ are new (in grey), whereas the edges $Bo \rightarrow BF$ and $Bo \rightarrow W$ disappeared (marked as dotted lines).

Indirect Comparison of Interaction Graphs

Ulrich Mansmann, Markus Schmidberger, Ralf Strobl and Vindi Jurinovic

Abstract A strategy for testing differential conditional independence structures (CIS) between two graphs is introduced. The graphs have the same set of nodes and are estimated from data sampled under two different conditions. The test uses the entire pathplot in a Lasso regression as the information on how a node connects with the remaining nodes in the graph.
The interpretation of the paths as random processes allows defining stopping times which make the statistical properties of the test statistic accessible to analytic reasoning. A resampling approach is proposed to calculated p-values simultaneously for a hierarchical testing procedure. The hierarchical testing steps through a given hierarchy of clusters. First, collective effects are measured at the coarsest level possible (the global null hypothesis that no node in the graph shows a differential CIS). If the global null hypothesis can be rejected, finer resolution levels are tested for an effect until the level of individual nodes is reached.
The strategy is applied to association patterns of categories from the ICF in patients under post-acute rehabilitation. The patients are characterized by two different conditions. A comprehensive understanding of differences in the conditional independence structures between the patient groups is pivotal for evidence-based intervention design on the policy, the service and the clinical level related to their treatment.
Due to extensive computation, parallel computing offers an effective approach to implement our explorative tool and to locate nodes in a graph which show differential CIS between two conditions.

Ulrich Mansmann, Markus Schmidberger and Vindi Jurinovic
IBE, LMU Munich, Germany, e-mail: mansmann@ibe.med.uni-muenchen.de

Ralf Strobl
Institute for Health and Rehabilitation Sciences, LMU Munich, Germany

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_14,

1 Introduction

We present a statistical strategy to detect changes in the conditional independence structure (CIS) between elements under different conditions. For example, the elements could be the genes which are annotated to a certain pathway. The conditions may be defined by two different diseases and two datasets containing the corresponding gene expression information measured in tissues from the respective patients. Finally, the CIS between the genes of the pathway may be estimated by an appropriate method (Schäfer & Strimmer 2005, Meinshausen & Bühlmann 2006, Wainwright et al. 2006, Friedman et al. 2007, Banerjee et al. 2008).
The detection of nodes which show differences in the way they connect to other nodes is straightforward by visual inspection of both graphs. But, it is difficult to decide which of the detected nodes show a differential CIS between both conditions caused by systematic differences (true positives) and which are statistical artefacts caused by the algorithm or random fluctuation in the data (false positives). A similar problem exists for the nodes with equal CIS's between both conditions. It is difficult to discriminate between true or false negatives. The goal of this paper is to present a strategy to detect a set of nodes with differential CIS under a controlled error rate.
The proposed strategy to detect the set of nodes is called indirect because an explicit estimation of the CIS between nodes is avoided. A direct test calculates the test statsitic from the estimated graphs. For example it can be based on a resampling (permutation) approach which works as follows:

- Choose a metric to measure differential connectivity between two graphs. This can be done by the Structural Hamming Distance (*SHD*).
- Estimate the two graphs by a specific algorithm from the given data and determine the SHD_{obs} between both graphs.
- Permute the data units between both data sets, estimate both graphs for permutation i and calculate the specific SHD_i $(i = 1, \ldots, R)$.
- Determine a permutation p-value by $\#\{SHD_{obs} < SHD_i\}/R$.

Related ideas can be found in Balasubramanian et al. (2004) or Ruschhaupt (2008). The strategy proposed will use a global test for a set of nodes. Furthermore, a given hierarchy of clusters within the set of nodes is considered. The hierarchy has to be derived from specific domain knowledge. For each cluster C we will test the null hypothesis $H_{0,C}$: The cluster C does not contain any node with a differential CIS to other nodes of the graph.
The hierarchical testing steps through a given hierarchy of clusters. First, collective effects are measured at the coarsest level possible (the global null hypothesis that no node in the graph shows a differential CIS). If the global null hypothesis can be rejected, finer resolution levels are tested for an effect until the level of individual nodes is reached.
Meinshausen (2008) developed an attractive approach for hierarchical testing which will be used to solve our problem.
In computational biology, it might for example be interesting to use the Gene Ontology (Ashburner et al. 2000) when testing for the differential connectivity derived

for genes of a specific pathway or functional group. But, the Gene Ontology does not posses the hierarchical nature of the hierarchies used by Meinshausen, although the approach can be made feasible (with some more cumbersome notation) for Gene Ontology and related hierarchies derived from genomic domain knowledge (Goeman & Mansmann 2008).
Since simple hierarchies do not exist for problems in computational biology, we study for illustrative reasons an example from human functioning where the nodes are respective categories defined by the International Classification of Functioning, Disability and Health (ICF, WHO (2001)).
The paper is organized as follows: Section 2 introduces the methodological aspects of the test statistic with which we compare graphs. It states theorems to describe properties of the test statistics, and defines the sampling approach to perform the hierarchical test procedure. Section 3 presents the example and Section 4 will discuss our approach. The Appendix offers some results to the properties of the test statistics.

2 Methods

Consider the p-dimensional multivariate distributed random variable $X = (X_1, \ldots, X_p)$ which is the outcome of a Markov random field (MRF). A Markov random field is specified by an undirected graph $G = (N, E)$, with node set $N = 1, 2, \ldots, p$ and edge set $E \subset N \times N$. The structure of this graph encodes certain conditional independence assumptions among subsets of the p-dimensional random variable X, where variable X_i is associated with node $i \in N$.
For multivariate Gaussian data, the article Meinshausen & Bühlmann (2006) solved the fundamental problem of estimating the structure of the underlying graph given a set of n samples from the MRF and showed that L_1-regularization can lead to practical algorithms with strong theoretical guarantees. For multivariate binary data, Wainwright et al. (2006) provides comparable results. Both methods use L_1-regularized regression (linear and logistic), in which the neighbourhood of any given node is estimated by performing regression subject to an L_1-constraint. Neighbourhood selection estimates the CIS separately for each node in the graph and is hence equivalent to variable selection for regression models. The proposed neighbourhood selection schemes are consistent for sparse high-dimensional graphs. Consistency depends on the choice of the penalty parameter which can be derived from controlling the probability of falsely joining some distinct connectivity components of the graph.

2.1 Defining the Test Statistic

For the specific node i the corresponding path plot of the regression coefficients for the L_1-regularized regression can be interpreted as a $p-1$ dimensional random process indexed by the penalty parameter λ: $B^{(i)}(\omega, \lambda) = (\beta_\lambda^{i,j})_{j \in N \setminus \{i\}}$ where $\lambda \geq 0$.

The randomness is introduced by the random data sample. For each of the $p-1$ paths of $B^{(i)}$ it is possible to determine the stopping time $\tau^{(i,j)} = \min\{\lambda > 0 : \beta_\lambda^{(i,j)} = 0\}$, $j \in N \setminus \{i\}$. For a fixed set of nodes $N = 1,\ldots,p$ we observe two i.i.d. samples of sizes n and m: $D_X = \{x^{(1)},\ldots,x^{(n)}\}$ and $D_Y = \{y^{(1)},\ldots,y^{(m)}\}$ with $x^{(i)} = (x_1^{(i)},\ldots,x_p^{(i)})$ and $y^{(k)} = (y_1^{(k)},\ldots,y_p^{(k)})$. For node i in node set N the path plots derived from both data sets are compared by counting the number of common non-zero regression coefficients given penalty parameter λ_x for the path plot derived from data D_X and penalty parameter λ_y for the path plot derived from data D_Y : $\Psi_i(\lambda_x,\lambda_y)$. This function is integrated over the range of the penalty:

$$\Psi_i = \iint\limits_{[0,\infty[\times[0,\infty[} \Psi_i(\lambda_x,\lambda_y)d\lambda_x d\lambda_y \tag{1}$$

The random variable Ψ_i can also be calculated from the stopping times introduced above:

$$\Psi_i = \sum_{j\in N\setminus\{i\}} \tau_X^{(i,j)} \cdot \tau_Y^{(i,j)} \tag{2}$$

where stopping times $\tau_X(\tau_Y)$ are derived from the path plot inferred from data $D_X(D_Y)$.
It is also possible to calculate a Ψ_N for the entire graph or a Ψ_C related to the subset $C \subset N$ of nodes:

$$\Psi_N = \sum_{i\in N} \Psi_i \text{ and } \Psi_C = \sum_{i\in C\subset N} \Psi_i \tag{3}$$

We define

$$\Psi_i^{max} := \sum_{j\in N\setminus\{i\}} \max\{\tau_X^{(i,j)}, \tau_Y^{(i,j)}\}^2 \tag{4}$$

The following theorem supports the intuition that the value of Ψ_i is larger when the same *MRF* defines the distribution of the data in both conditions than when the distributions are defined by two different *MRF*s.

Theorem 1. *Let P_1 and P_2 be two equal MRFs over the same set of nodes N, D_X and D_Y two i.i.d. samples from both MRFs of size n resp. $m(n) = n \cdot \frac{1-\pi}{\pi}$. The quantity $\pi = \frac{n}{n+m}$ is the fixed percentage of the sample size of D_X on the total number of observations. Then for an arbitrary small $\varepsilon > 0$ and each node $i \in N$ there is an $n(\varepsilon)$ such that for all $n > n(\varepsilon)$ it holds*

$$P[\Psi_i^{max} \leq \Psi_i + \varepsilon] > 1 - \varepsilon. \tag{5}$$

A sketch of the proof is given in the Appendix.

2.2 A Permutation Test

Theorem 2. *Let P_1 and P_2 be two identical Markov random fields (MRFs) which generate the data under the two conditions of interest. Let Q be the mixture distribution created from P_1 and P_2 with mixture proportion π (for component P_1). For an arbitrary small $\varepsilon > 0$, a value $w > 0$, and each node $i \in N$ there is an $n(\varepsilon)$ such that for all $n > n(\varepsilon)$ it holds*

$$|P[\Psi_i^* > w] - P[\Psi_i^{\#} > w]| < \varepsilon. \tag{6}$$

The value of Ψ_i^* is calculated from the data sets D_X^* [n i.i.d. samples from P_1] and D_Y^* [$m = n\cdot(1-\pi)/\pi$ i.i.d. samples from P_2] and $\Psi_i^{\#}$ is calculated from the data sets D'_X [n i.i.d. samples from Q] and D'_Y [$m = n\cdot(1-\pi)/\pi$ i.i.d. samples from Q].

A sketch of the proof and a possible extension to a wider class of null-hypotheses is given in the Appendix.

The theorem above states for each node in N: Under the null-hypothesis (*equal MRFs*) the distribution of the test statistics can be generated from a permutation sampling of the observed data.

The permutation procedure calculates S samples simultaneously for each node i: $\Psi_i^{\#(S)}$. The permutation p-value for node i is derived as $p_i = |\{r : \Psi_i^{\#(S)} > \Psi_i^{\#}\}|/S$. It is straight forward to extend the theorem to test statistics for the set of nodes ($\Psi_N = \sum_{i\in N} \Psi_i$) or specific subsets of nodes ($\Psi_C = \sum_{i\in C\subset N} \Psi_i$). Correspondingly, it is possible to calculate permutation p-values for arbitrary sets of nodes.

2.3 Hierarchical Testing

Now it is straightforward to combine our approach with the hierarchical testing principle of Meinshausen (2008). The principle allows using the same resampling sample to calculate p-values for each element of the hierarchy.

For the following, we assume that a hierarchy $\aleph$ is given, which is a set of clusters $C \subset \{1,\ldots,p\}$. The cardinality of a cluster C is denoted by $|C|$. The root node $\{1,\ldots,p\}$ contains all nodes of the graph and has cardinality p. The hierarchical structure implies that any two clusters $C, C' \in \aleph$ either have an empty intersection, or that one cluster is a subset of the other.

To take the multiplicity of the testing problem into account, p-values have to be adjusted. Define for cluster C the adjusted p-value as $p_{adj}^C := p_C \cdot \frac{p}{|C|}$ where p is the total number of nodes in the graph and $|C|$ is the number of nodes in the cluster of interest. The adjustment amounts to multiplying the p-value of each cluster C with the inverse of the fraction $|C|/p$ of variables it contains. The adjustment is thus resolution dependent. At coarse resolutions, the penalty for multiplicity is weak, and it increases for finer resolution levels. The p-value of the root node is thus unadjusted, whereas individual variables receive a Bonferroni-type adjustment.

The hierarchical testing procedure rejects now all hypotheses $H_{0,C}$ with $C \in \aleph$ for which (a) the adjusted p-value p^C_{adj} is below or equal to the specified level and (b) the parent node is rejected (this is always considered to be fulfilled for the root node).
Note that condition (b) is not a severe restriction. The null hypothesis $H_{0,C}$ of a cluster C is by definition always true if the null hypothesis $H_{0,pa(C)}$ is true for the parent cluster $pa(C)$. Hence it makes sense to stop testing in subtrees of clusters whose null hypothesis could not be rejected.
Using the definition of a hierarchically adjusted p-value $p^C_{h,adj} = \max_{D \in \aleph, C \subseteq D} p^D_{adj}$, the set of clusters which are rejected in the hierarchy $\aleph$ on the level α is then given by $C_{rejected} = \{C \in \aleph, p^C_{h,adj} < \alpha\}$.
Control of the family-wise error rate can now be achieved. The set of clusters that fulfill the null hypothesis $H_{0,C}$ is denoted by $C_0 = \{C \in \aleph, H_{0,C} \text{ is fulfilled}\}$. Family-wise error rate control entails that the probability of rejecting any cluster in C_0 is smaller than the pre-specified level α.

Theorem 3. *For $C_{rejected}$ and C_0 as defined above, the family-wise error rate is controlled at level α:*

$$P(C_{rejected} \cap C_0 = \emptyset) = 1 - \alpha \tag{7}$$

Proof is given by Meinshausen (2008).

2.4 Computational Issues

Calculations are done in the statistical computing software `R (V 2.9.0)` (R Development Core Team 2009). The working horse of the Lasso Regression is the computationally efficient gradient-ascent algorithm as proposed by Goeman (2009b) and implemented in Goeman (2009a). The permutation test was parallelized.
A total of 1000 samples were created to perform the comparison of the CIS between both conditions. The calculation is very computer intensive and sample calculations are independent from each other. Therefore, the functions and data are distributed to different processors. In the `R` language different technologies and approaches for parallel computing exist (Schmidberger et al. 2009). We use the snow package (`snow`) with the Rmpi package (`Rmpi`) for the communication between the processors. The code is executed at 1000 processors using the super computer HLRB2 at the Leibniz-Rechenzentrum in Munich (Germany). To guarantee a different random number stream in every `R` session, an additional `R` package rlecuyer (`rlecuyer`) is applied.
An `R`-package which offers the needed algorithms is under preparation.

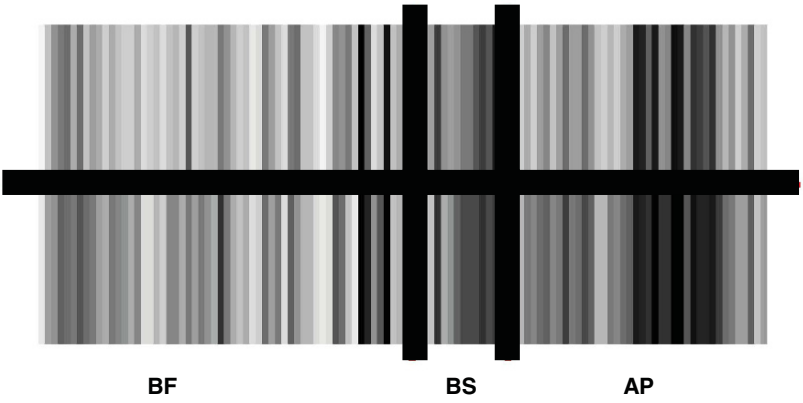


Fig. 1 Relative frequences of impairments and limitations. Patients with condition B are above the horizontal black line, patients with condition A below. BF = Body Functions, BS = Body Structures, AP = Activities & Participation

3 Example

The example studies a multivariate binary situation. It is taken from the field of rehabilitation science.

Functioning and disability are universal human experiences. Over the life span people may experience limitations in functioning in relation to health conditions including an acute disease or injury, a chronic condition, or aging. A standard language for the analysis of functioning is provided by the International Classification of Functioning, Disability and Health (ICF, WHO (2001)).

A secondary analysis of observational cross sectional data of patients from five early post-acute rehabilitation units is performed. The ICF is used to measure functioning and contextual factors. We look at the components Body Functions, Body Structures, and Activities & Participation. The presence of an impairment or limitation was binary coded for each of the 122 categories considered.

616 patients (mean age 63 years, 46% male) were included. 56% had health condition A. Figure 1 shows the profiles of the 122 ICF-categories measured during post-acute rehabilitation between the 343 patients with health condition A and the 273 patients with health condition B.

Besides the comparison of functional profiles it is important to understand stability and distinctiveness of functioning across health conditions. This can be achieved by revealing patterns of associations between distinct aspects of functioning, the ICF categories.

Based on the proposal of Wainwright et al. (2006), CIS graphs can be estimated for each condition and visually compared as done in Figure 2.

Besides a simple visual inspection for differential CIS we apply the combination between our test and the hierarchical testing procedure. The hierarchy is defined by

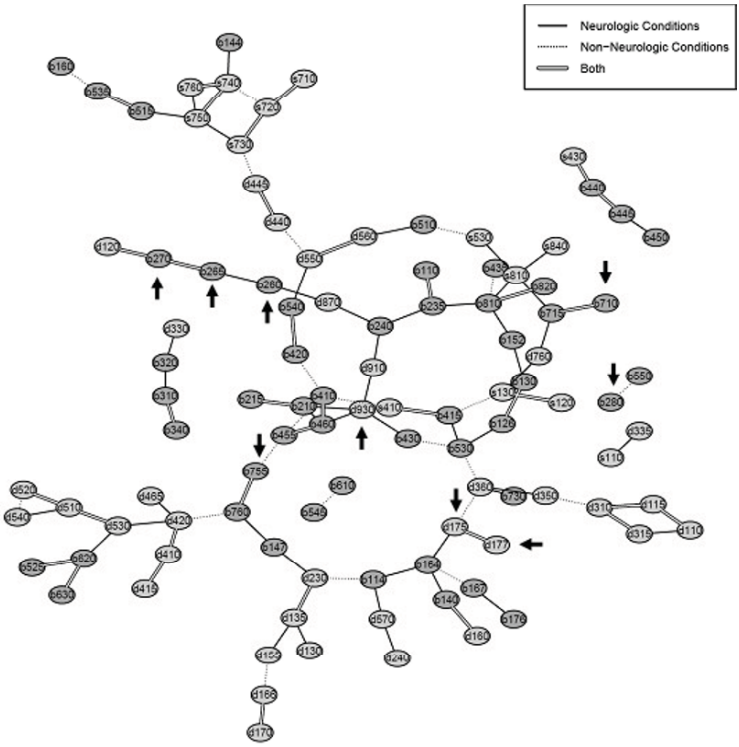


Fig. 2 Estimated CIS graphs and visual presentation of common edges as well as different edges within both graphs. The arrows point to the nodes with significant differential CIS on a 5% significance level ($\alpha = 0.05$). The three ICF components are presented in different colours: orange: Body Functions, green: Body Structures, blue: Activities & Participation

the ICF itself (WHO 2001). The root node contains all 122 ICF categories analyzed. The first level of the hierarchy is defined by the ICF-components. The second level is determined by the ICF-chapters; the first branching level in the classification gives the third level. The fourth level is given by the single categories of interest.
Table 1 gives the nodes with a significant differential CIS according to the hierarchical test procedure. The complete results are presented in the Appendix.

Our strategy was able to detect a set of nodes with differential CIS in the multivariate ICF profiles between patients under two different disease conditions as well as clusters in the hierarchy (result not shown). The identified set holds a family wise error rate of 5%, i.e. the probability of containing at least one false positive node under the null-hypothesis is 5%. The result on differential CIS is combined with two graph estimates and their naive comparison with respect to present or missing edges between both graphs by adding black arrows to the naive graph comparison. Some of the arrows point to nodes ($b260, b280, 285, b755, d175, d930$) which under condition A show a different connectivity than under condition B. Some of the arrows point

Table 1 nodes with significant differential CIS

Code	title	p.value	p.value.adj
b260	Proprioceptive function	0.00018	0.02056
b265	Touch function	1e-05	0.00124
b270	Sensory functions related to temperature and other stimuli	1e-05	0.00124
b280	Sensation of pain	0.00015	0.01722
b710	Mobility of joint functions	4e-05	0.00497
b755	Involuntary movement reaction functions	1e-05	0.0014
d175	Solving problems	1e-05	0.01244
d177	Making decisions	0.00025	0.02855
d930	Religion and spirituality	1e-05	0.0056

to nodes ($b170$, $b270$, $d177$) that show the same connectivity in both graphs. This can be understood by looking more closely to the null-hypothesis which is rejected: differential CIS can also be produced by different regression coefficients given the same connectivity. For example the odds ratio between $b710$ and $b715$ ($b270$ and $d120$, $b270$ and $b265$) is 3.91 (19.27, 29.12) in patients with condition A and 13.12 (10.89, 19.29) in patients with condition B.

4 Discussion

The estimation of complex graphs from observed data is a challenging task and different strategies were developed for the case of sparse graphical structures (Schäfer & Strimmer 2005, Meinshausen & Bühlmann 2006, Wainwright et al. 2006, Friedman et al. 2007, Kalisch & Bühlmann 2007, Banerjee et al. 2008). Especially, the estimation of the conditional independence structure (CIS) in a multivariate observation is of high interest. Different data sets of the same multidimensional random variable from different conditions may be available. They may produce different CIS-graph estimates. A natural question is if the observed data give convincing evidence for a systematic difference between the CIS-graphs behind the distributions of the observed data. The meaning of convincing evidence has to be operationalized in statistical terms. This is achieved by using the family wise error rate.

It is our intention to compare the conditional independence structure (CIS) between two multivariate Gaussian or binary distributions (Ising model). Multivariate Gaussian or binary distributions define Markov random fields (MRF) and imply a graph with the nodes defined by the single components of the multivariate random variable and the edges between the nodes by the CIS. Besides different set of edges, further differences in the CISs between two distributions are created by the strength of the conditional dependencies between specific nodes. We present a test statistic which addresses both CIS-aspects.

The measure to compare two graphs uses the idea that the connectivity of a node with

the remaining nodes in a graph relates to a variable selection problem in a regression setting. In this context, the usefulness of L_1-regularized regression was demonstrated by several authors (Wainwright et al. 2006, Meinshausen & Bühlmann 2006, Friedman et al. 2007). The differential connectivity of node i in both graphs is defined as follows: (1) for each node $j \neq i$ we determine the minimal penalty parameter which shrinks the corresponding regression coefficient to zero $(\tau_X^{(i,j)}, \tau_Y^{(i,j)})$ in both conditions; (2) for each $j \neq i$ we calculate $\tau_X^{(i,j)} \cdot \tau_Y^{(i,j)}$; (3) we determine the test statistic Ψ_i as the sum of the products over all $j \neq i$. The test statistics for the entire graph or a subset C of nodes is $\Psi_N = \sum_{i \in N} \Psi_i$ and $\Psi_C = \sum_{i \in C \subset N} \Psi_i$ respectively. The test statistic is motivated by the intuition that equal *MRF*s in both conditions will produce a large value of Ψ_i, Ψ_N or Ψ_C.

It is shown in the Appendix that the $\tau^{(i,j)}$s can be calculated in principle and that formal statements about their properties can be derived.

The null-hypothesis of equal *MRF*s for both conditions is tested by a permutation approach. It allows formulating global tests on sets of nodes as well as tests for single nodes. This enables searching for a set of nodes with differential CIS in a hierarchical testing procedure. The motivation for hierarchical testing can be summarized as follows:

- Any differential connectivity at all? The CIS of a group of nodes can be tested between both graphs whether all nodes have the same CIS under each condition.
- Differential CIS in sub-clusters? If it is established that a cluster of nodes does indeed contain nodes with differential CIS, it is desirable to attribute it to one or several sub-clusters

If possible, the differential CIS in a cluster of variables is attributed to its sub-clusters. In each sub-cluster, it is again examined whether the collective effect can be attributed to even smaller sub-clusters of nodes. The procedure retains the smallest possible clusters which exhibit a significant differential CIS or helps to detect single nodes with differential CIS.

Our approach avoids estimating graphs explicitly. We did not put the direct and the indirect approach side by side. Therefore, no detailed analysis of their pros and cons is available. The indirect approach does not fix the value of a regularization parameter which has to be done when the estimate of an explicit graph is needed. The direct approach needs explicit graphs since differential CIS may be measured by the Structural Hamming Distance (*SHD*) (Kalisch & Bühlmann 2007) or other suitable methods. The hierarchical test procedure can be applied for the direct as well as the indirect approach.

One potential advantage of the proposed test statistics is its generalizability to more than two graphs. The comparison of several graphs based on the bivariate *SHD* measure is cumbersome and not illuminative. For node i and data from conditions U, V, W, X, Y, and Z (for example) it is possible to modify Ψ_i by $\Psi_i = \sum_{j \in N \setminus \{i\}} \tau_U^{(i,j)} \cdots \tau_Z^{(i,j)}$. This is the subject of further research.

In this paper we study the null-hypothesis of two equal Markov random fields. More general forms of the null-hypothesis ($H_{0,general}$) may be of interest: *The MRFs are*

different but the graphs related to both distributions have a Structural Hamming Distance of 0. A CIS graph G defines a family of probability measures (multivariate Gaussian or multivariate binary Ising Model) p_G by the corresponding CIS. The null-hypothesis $H_{0,general}$ states that the distributions which generate the data under both conditions belong to p_G. Since the mixture of two distributions from p_G is in general not in p_G it follows that the permutation approach used so far will not work anymore. We assume that $H_{0,general}$ can be tested by replacing the permutation by a more complicated resampling procedure. It may be based on a sampling scheme which creates new versions of data D_X and D_Y under the restriction that the graphs behind the distribution of X and Y have the same set of edges, $SHD(G_X, G_Y) = 0$. The proof of this assumption and the development of an efficient algorithm are topics of ongoing research.

The strategy presented in our paper depends on the implicit assumption that the data is created by *MRF*s (multivariate Gaussian or multivariate binary Ising Model). We used this assumption implicitly in our example. Tools for model validation and model assessment in a setting comparable to the data presented are under development (Gneiting 2008). It is a second line of our research to implement efficient validation strategies to assess model assumptions. A mixture of binary Ising models is in general not a binary Ising model. A mixture of binary Ising models with the same conditional independence graph does not need to have the same graph anymore because of confounding. We tried to reduce confounding by using fixed covariates in the Lasso regression. The algorithm provided by Goeman (2008) allows controlling for confounding by the incorporation of fixed covariates.

The strategy presented offers an explorative tool to detect nodes in a graph with the potential of a relevant impact on the regulatory process between interacting units in a complex process. The findings introduce a practical algorithm with theoretical guarantees. We see our result as the first step on the way to a meta-analysis of graphs. A meta-analysis of graphs is only useful if the graphs available for aggregation are *homogeneous*. The definition of the homogeneity of graphs $G_1, \ldots, G_K$ by a pairwise Structural Hamming Distance of 0 is not sufficient to describe homogeneity in a correct way. The assessment of homogeneity of graphs needs procedures like the one presented.

Acknowledgements This work is supported by the LMU*innovativ* project *Analysis and Modelling of Complex Systems in Biology and Medicine (Cluster B, Expression Analyses)*.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology:a tool for the unification of biology., *Nature Genetics* **25**: 25–29.

Balasubramanian, R., LaFramboise, T., Scholtens, D. & Gentleman, R. (2004). A graph-theoretic approach to testing associations between disparate sources of functional genomics data., *Bioin-*

formatics **20**(18): 3353–3362.
Banerjee, O., Ghaoui, L. E. & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data, *Journal of Machine Learning Research* pp. 485–516.
Friedman, J., Hastie, T. & Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* .
Gneiting, T. (2008). Editorial: Probabilistic forecasting, *Journal of the Royal Statistical Society: Series A* **17**: 319–321.
Goeman, J. (2008). *penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model*. `http://www.msbi.nl/goeman`. R package version 0.9-22.
Goeman, J. (2009a). L1 and l2 penalized regression models. `http://www.msbi.nl/goeman`. R package version 0.9-24.
Goeman, J. (2009b). L1 penalized estimation in the cox proportional hazards model., *Biometrical Journal* .
Goeman, J. J. & Mansmann, U. (2008). Multiple testing on the directed acyclic graph of gene ontology., *Bioinformatics* **24**(4): 537–544.
Kalisch, M. & Bühlmann, P. (2007). Estimating high dimensional acyclic graphs with the pc-algorithm, *Journal of Machine Learning Research* **8**: 613–636.
Meinshausen, N. (2008). Hierarchical testing of variable importance, *Biometrika* **95**: 265–276.
Meinshausen, N. & Bühlmann, P. (2006). High dimensional graphs and variable selection with the Lasso, *The Annals of Statistics* **34**: 1436–1462.
R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`.
Ruschhaupt, M. (2008). *Erzeugung von positiv definiten Matrizen mit Nebenbedingungen zur Validierung von Netzwerkalgorithmen für Microarray-Daten*, PhD thesis, LMU München, Fakultät für Mathematik, Informatik und Statistik, München, Germany. `http://edoc.ub.uni-muenchen.de/view/subjects/fak16.html`.
Schäfer, J. & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics., *Statistical Applications in Genetics and Molecular Biology* **4**: Article32.
Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L. & Mansmann, U. (2009). State of the art in parallel computing with R, *Journal of Statistical Software* **31**(1). `http://www.jstatsoft.org/v31/i01/`.
Wainwright, M. J., Ravikumar, P. & Lafferty, J. D. (2006). High dimensional graphical model selection using l1-regularized logistic regression, *Proceedings of Advances in neural information processing systems* **9**: 1465–1472.
WHO (2001). International classification of functioning, disability and health (icf), *Nature Genetics* .

Appendix

Properties of the Stopping Times

We look at the following penalized quadratic form

$$QF(\beta,\lambda) = (\beta-\beta^*)^t \cdot \Sigma \cdot (\beta-\beta^*) - \lambda \cdot \sum_{i=1}^{d} |\beta_i| \tag{8}$$

where β is a d-dimensional parameter vector, β^* a d-dimensional fixed vector, Σ a $d \times d$ matrix, and $\lambda \geq 0$. For given λ, QF is maximized by $\beta^{\#}(\lambda)$

$$\beta_j^{\#}(\lambda) = 2 \cdot \beta_j^* + \lambda \cdot \sum_k (-1)^{j+k} \mathrm{Det}(\Sigma_{jk}) / \mathrm{Det}(\Sigma) \tag{9}$$

for $j = 1, \ldots, d$. Σ_{jk} is a quadratic matrix derived from matrix Σ by removing line j and column k.
The penalized log-likelihood for a linear regression problem $y_i \sim \beta \cdot x_i$ is given by QF where $\Sigma = X^t \cdot X$ and $\beta^* = y \cdot X^t \cdot (X^t \cdot X)^{-1}$. The penalized log-likelihood for a logistic regression problem $\mathrm{logit}(p_i) = \beta \cdot x_i$ is approximated by QF where β^* is the ML estimate and Σ is the corresponding Fisher Information Matrix.
From (1.8) it is possible to calculate the minimal penalty parameter λ which shrinks the coefficient to 0. In terms of the notation introduced in the methods section if follows

$$\tau^{(i,j)} = \min\{\lambda > 0 : \beta_\lambda^{(i,j)} = 0\} = 2 \cdot \beta_j^* \cdot \frac{\mathrm{Det}(\Sigma)}{\sum_k (-1)^{j+k+1} \mathrm{Det}(\Sigma_{jk})} \tag{10}$$

The matrix Σ is derived from observed data and varies around the true value Σ^0. The variability of $\tau^{(i,j)}$ can be controlled by the following property of determinants:

$$\begin{aligned} \mathrm{Det}(\Sigma^0 + \delta \cdot \Lambda) &= \mathrm{Det}\left(\left(\Sigma^0\right)^{-1}\left(I + \delta \cdot \Lambda \cdot \left(\Sigma^0\right)^{-1}\right)\right) \\ &= \mathrm{Det}\left(\left(\Sigma^0\right)^{-1}\right) \cdot \left(I + \delta \cdot \mathrm{trace}\left(\Lambda \cdot \left(\Sigma^0\right)^{-1}\right)\right) \end{aligned}$$

It holds that

$$\begin{aligned} \tau^{(i,j)} &= \min\{\lambda > 0 : \beta_\lambda^{(i,j)} = 0\} \\ &= 2 \cdot \beta_j^0 \cdot \frac{\mathrm{Det}(\Sigma^0)}{\sum_k (-1)^{j+k+1} \mathrm{Det}(\Sigma_{jk}^0)} + \frac{\mathrm{Det}(\Sigma^0)}{\sum_k (-1)^{j+k+1} \mathrm{Det}(\Sigma_{jk}^0)} \cdot \varepsilon_{ij} \end{aligned}$$

where β_j^0 is the true regression coefficient. The random variables ε_{ij} concentrate on a small neighbourhood of 0 (depending of the sample size).

Statistical Properties of Ψ_i

Sketch of proof of Theorem 1.1: Without loss of generality we choose $r = 1$. We denote a variable or parameter which belongs to *MRF r* and the regression of node j on node i by $..{}_r^{(i,j)}$. We also introduce for the *MRF* $P_r (r = 1, 2)$ the notation $\Omega_r^{(i,j)} =$

$\frac{\mathrm{Det}(\Sigma^0_{jk,r})}{\sum_k (-1)^{j+k+1}\mathrm{Det}(\Sigma^0_{jk,r})}$ where $\Sigma^0_{jk,r}$ is Σ^0 which belongs to *MRF* $r(r=1,2)$ without row j and column k.
The inequality claimed in the theorem follows from

$$\Psi_i^{max} = \sum_{j \in n \setminus \{i\}} \max\{\tau_X^{(i,j)}, \tau_Y^{(i,j)}\}^2 \tag{11}$$

using the approximative representation of $\tau^{(i,j)}$ by the true parameter values and a controlled error term $\tilde{\varepsilon}$. It follows

$$\Psi_i^{max} \leq \sum_{j \in n \setminus \{i\}} \max\{\beta_1^{(i,j)} \cdot \Omega_1^{(i,j)}, \beta_2^{(i,j)} \cdot \Omega_2^{(i,j)}\}^2 + \tilde{\varepsilon} \tag{12}$$

under the null-hypothesis $\beta_1^{(i,j)} = \beta_2^{(i,j)} = \beta^{(i,j)}$ and $\Omega_1^{(i,j)} = c \cdot \Omega_2^{(i,j)} = \Omega^{(i,j)}$ where $c = c(n,\pi)$. As a consequence,

$$\Psi_i^{max} \leq \sum_{j \in n \setminus \{i\}} \{\beta^{(i,j)} \cdot \Omega^{(i,j)}\}^2 \cdot c^{-1} + \tilde{\varepsilon} \tag{13}$$

using again the approximative representation of $\tau^{(i,j)}$ it follows

$$\Psi_i^{max} = \sum_{j \in n \setminus \{i\}} \tau_X^{(i,j)} \cdot \tau_Y^{(i,j)} + 2 \cdot \tilde{\varepsilon} \tag{14}$$

where $P[|\tilde{\varepsilon}| < \varepsilon] > 1 - \varepsilon$ for $n > n(\varepsilon)$. This proves the theorem.

Sketch of proof of Theorem 1.2: The proof follows from Theorem 1 by the argument that $P_1 = P_2$ and therefore $P[\Psi_i^* \leq \Psi_i^\#] > 1 - \varepsilon$ as well as $P[\Psi_i^\# \leq \Psi_i^*] > 1 - \varepsilon$ for an arbitrary small $\varepsilon > 0$ and $n > n(\varepsilon)$. This implies $P[|\Psi_i^\# - \Psi_i^*| < \varepsilon] > 1 - \varepsilon$ and

$$\begin{aligned} P[\Psi_i^* > w] &= P[\Psi_i^* - \Psi_i^\# + \Psi_i^\# > w] = P[\Psi_i^\# > w + \Psi_i^\# - \Psi_i^*] \\ &= P[\{\Psi_i^\# > w + \Psi_i^\# - \Psi_i^*\} \cap \{|\Psi_i^\# - \Psi_i^*| < \delta\}] + \\ &\quad P[\{\Psi_i^\# > w + \Psi_i^\# - \Psi_i^*\} \cap \{|\Psi_i^\# - \Psi_i^*| > \delta\}]. \end{aligned}$$

For appropriately small δ and sufficiently large n it holds

$$P[\Psi_i^* > w] \geq P[\Psi_i^\# > w + \delta] \cdot (1 - \delta) \tag{15}$$

$$P[\Psi_i^* > w] \leq P[\Psi_i^\# > w - \delta] + \delta \tag{16}$$

which proves the theorem.

Results of the Hierarchical Test Procedure on Differential Conditional Independence Structure for each Node (ICF Category)

Code	title	p.value	p.value.adj
b110	Consciousness functions	0.26916	1
b114	Orientation functions	0.19031	1
b126	Temperament and personality functions	0.07132	1
b130	Energy and drive functions	0.40246	1
b134	Sleep functions	0.1266	1
b140	Attention functions	0.81069	1
b144	Memory functions	0.13688	1
b147	Psychomotor functions	0.03534	1
b152	Emotional functions	0.57016	1
b156	Perceptual functions	0.67479	1
b160	Thought functions	0.0098	1
b164	Higher-level cognitive functions	0.00572	0.64073
b167	Mental functions of language	0.0526	1
b176	Mental function of sequencing complex movements	0.0067	0.75053
b180	Experience of self and time functions	0.80941	1
b210	Seeing functions	0.21501	1
b215	Functions of structures adjoining the eye	0.85915	1
b230	Hearing functions	0.46113	1
b235	Vestibular functions	0.95262	1
b240	Sensations associated with hearing and vestibular function	0.19814	1
b260	Proprioceptive function	0.00018	0.02056
b265	Touch function	1e-05	0.00124
b270	Sensory functions related to temperature and other stimuli	1e-05	0.00124
b280	Sensation of pain	0.00015	0.01722
b310	Voice functions	<0.00001	0.18667
b340	Alternative vocalization functions	0.09032	1
b410	Heart functions	0.79414	1
b415	Blood vessel functions	0.01307	1
b420	Blood pressure functions	0.00783	0.87703
b430	Haematological system functions	0.15039	1
b435	Immunological system functions	0.0033	0.36913
b440	Respiration functions	0.1212	1
b445	Respiratory muscle functions	0.10322	1
b450	Additional respiratory functions	0.70879	1
b455	Exercise tolerance functions	0.81957	1
b460	Sensations associated with cardiovascular and respiratory functions	0.47558	1
b510	Ingestion functions	0.01285	1
b515	Digestive functions	0.79141	1
b525	Defecation functions	0.34564	1
b530	Weight maintenance functions	0.12755	1

Code	title	p.value	p.value.adj
b535	Sensations associated with the digestive system	0.21833	1
b540	General metabolic functions	0.17652	1
b545	Water, mineral and electrolyte balance functions	0.50622	1
b550	Thermoregulatory functions	0.98723	1
b610	Urinary excretory functionsn	0.20772	1
b620	Urination functions	0.85082	1
b630	Sensations associated with urinary functions	0.13142	1
b710	Mobility of joint functions	4e-05	0.00497
b715	Stability of joint functions	0.00113	0.12604
b730	Muscle power functions	0.16681	1
b735	Muscle tone functions	0.0137	1
b755	Involuntary movement reaction functions	1e-05	0.0014
b760	Control of voluntary movement functions	0.02091	1
b770	Gait pattern functions	0.22097	1
b780	Sensations related to muscles and movement functions	0.92971	1
b810	Protective functions of the skin	0.56761	1
b820	Repair functions of the skin	0.93102	1
s110	Structure of brain	0.0736	1
s120	Spinal cord and related structures	0.05937	1
s130	Structure of meninges	0.13523	1
s410	Structure of cardiovascular system	0.72956	1
s430	Structure of respiratory system	0.12588	1
s530	Structure of stomach	0.01038	1
s710	Structure of head and neck region	0.22969	1
s720	Structure of shoulder region	0.46041	1
s730	Structure of upper extremity	0.38268	1
s740	Structure of pelvic region	0.1352	1
s750	Structure of lower extremity	0.5652	1
s760	Structure of trunk	0.47323	1
s810	Structure of areas of skin	0.4446	1
s840	Structure of hair	0.15527	1
d110	Watching	0.03047	1
d115	Listening	0.0462	1
d120	Other purposeful sensing	0.27109	1
d130	Copying	0.02501	1
d135	Rehearsing	0.08179	1
d155	Acquiring skills	0.40423	1
d160	Focusing attention	0.52833	1
d166	Reading	0.06237	1
d170	Writing	0.82457	1
d175	Solving problems	<0.00001	0.01244
d177	Making decisions	0.00025	0.02855
d230	Carrying out daily routine	0.96039	1
d240	Handling stress and other psychological demands	0.49906	1

Code	title	p.value	p.value.adj
d310	Communicating with - receiving - spoken messages	0.42238	1
d315	Communicating with - receiving - nonverbal messages	0.83199	1
d330	Speaking	0.73189	1
d335	Producing nonverbal messages	0.01302	1
d350	Conversation	0.42381	1
d360	Using communication devices and techniques	<0.00001	1
d410	Changing basic body position	3e-05	1
d415	Maintaining a body position	0.26954	1
d420	Transferring oneself	0.94277	1
d430	Lifting and carrying objects	0.31586	1
d440	Fine hand use	0.85019	1
d445	Hand and arm use	0.22326	1
d450	Walking	0.32237	1
d460	Moving around in different locations	0.00711	0.79633
d465	Moving around using equipment	0.04467	1
d510	Washing oneself	0.44569	1
d520	Caring for body parts	0.17469	1
d530	Toileting	0.29863	1
d540	Dressing	0.98399	1
d550	Eating	0.48303	1
d560	Drinking	0.0371	1
d570	Looking after ones health	0.00608	0.68055
d760	Family relationships	0.03614	1
d870	Economic self-sufficiency	0.04971	1
d910	Community life	0.07671	1
d930	Religion and spirituality	<0.00001	0.0056
d940	Human rights	0.12703	1

Modelling, Estimation and Visualization of Multivariate Dependence for High-frequency Data

Erik Brodin and Claudia Klüppelberg

Abstract Dependence modelling and estimation is a key issue in the assessment of financial risk. It is common knowledge meanwhile that the multivariate normal model with linear correlation as its natural dependence measure is by no means an ideal model. We suggest a large class of models and a dependence function, which allows us to capture the complete extreme dependence structure of a portfolio. We also present a simple nonparametric estimation procedure of this function. To show our new method at work we apply it to a financial data set of high-frequency stock data and estimate the extreme dependence in the data. Among the results in the investigation we show that the extreme dependence is the same for different time scales. This is consistent with the result on high-frequency FX data reported in Hauksson et al. (2001). Hence, the different asset classes seem to share the same time scaling for extreme dependence. This time scaling property of high-frequency data is also explained from a theoretical point of view.

Key words: Risk management; extreme risk assessment; high-frequency data; multivariate extreme value statistics; multivariate models; tail dependence function

1 Multivariate Risk Assessment for Extreme Risk

Estimation of dependence within a portfolio based on high-frequency data faces various problems:

- data are not normal: they are skewed and heavy-tailed

Erik Brodin
Department of Mathematical Sciences, Chalmers University of Technology, SE-412 96 Göteborg, Sweden, URL: www.chalmers.se/math/EN/, e-mail: ebrodin@math.chalmers.se

Claudia Klüppelberg
Center for Mathematical Sciences, Technische Universität München, D-85747 Garching, Germany, URL: www-m4.ma.tum.de, e-mail: cklu@ma.tum.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_15,

- one-dimensional data are uncorrelated but not iid
- multivariate high-frequency data are not synchronised
- data are discrete-valued for a very high frequency
- most likely there is microstructure noise in the data
- there is seasonality in the data
- higher moments may not exist
- the multivariate distribution may not be elliptical
- dependence may not be symmetric

We are interested here in the influence of the multivariate dependence within the portfolio. We first recall that under the condition that the portfolio P/L follows a multivariate normal distribution and, if there is no serial dependence, the portfolio P/L standard deviation σ is calculated by the square root of its variance

$$\sigma^2 = \sum_{i=1}^{n} w_i^2 \sigma_i^2 + \sum_{i \neq j} w_i w_j \sigma_i \sigma_j \rho_{ij}\,, \tag{1}$$

where the portfolio consists of n different instruments with nominal amount w_i invested into asset i. The standard deviation of asset i is given by σ_i and the pairwise correlation coefficients are ρ_{ij} $(i, j = 1, \ldots, n)$.

Definition 1. For two random variables X and Y their *linear correlation* is defined as

$$\rho_L(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}\,,$$

where $\operatorname{cov}(X,Y) = E((X - EX)(Y - EY))$ is the covariance of X and Y, and $\operatorname{var}(X)$ and $\operatorname{var}(Y)$ are the variances of X and Y, respectively.

Correlation measures linear dependence: we have $|\rho_L(X,Y)| = 1$ if and only if $Y = aX + b$ with probability 1 for $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$. Furthermore, correlation is invariant under strictly increasing *linear* transformations; i.e. for $\alpha, \gamma \in \mathbb{R} \setminus \{0\}$ and $\beta, \delta \in \mathbb{R}$

$$\rho_L(\alpha X + \beta, \gamma Y + \delta) = \operatorname{sign}(\alpha\gamma)\, \rho_L(X,Y)\,.$$

Also for high-dimensional models correlation is easy to handle. For random (column) vectors $X, Y \in \mathbb{R}^n$ we denote by $\operatorname{cov}(X,Y) = E((X - EX)(Y - EY)^T)$ the covariance matrix of X and Y. Then for $m \times n$ matrices A, B and vectors $a, b \in \mathbb{R}^m$ we calculate

$$\operatorname{cov}(AX + a, BY + b) = A \operatorname{cov}(X,Y)\, B^T\,,$$

where B^T denotes the transpose of the matrix B. From this it follows for $w \in \mathbb{R}^n$ that

$$\operatorname{var}(w^T X) = w^T \operatorname{cov}(X,X) w\,,$$

which is exactly formula (1) above. The popularity of correlation is also based on the fact that it is very easy to calculate and estimate. It is a natural dependence measure for elliptical distributions such as the multivariate normal or t distributions, provided

second moments exist. Within the context of linear models correlation has also proved as a useful tool for dimension reduction (e.g. by factor analysis), an important issue in risk management; see Klüppelberg & Kuhn (2009) for a new approach to dimension reduction for financial data.

Multivariate portfolios, however, are often not elliptically distributed, and there may be a more complex dependence structure than linear dependence. Indeed, data may be uncorrelated, i.e. with correlation 0, but still may be highly dependent. In the context of risk management, when measuring extreme risk, modelling dependence by correlation may be grossly misleading; see e.g. Embrechts et al. (2002).

We turn to a measure for tail dependence, which relates large values of the components of a portfolio; see e.g. Joe (1997). In the bivariate context, consider random variables X and Y with marginal distribution functions G_X and G_Y and (generalized) inverses $G_X^{\leftarrow}$ and $G_Y^{\leftarrow}$. For any distribution function G its *generalized inverse or quantile function* is defined as

$$G^{\leftarrow}(t) = \inf\{x \in \mathbb{R} \mid G(x) \geq t\}, \quad 0 < t < 1.$$

If G is strictly increasing, then $G^{\leftarrow}$ coincides with the usual inverse of G.

Definition 2. The *upper tail dependence coefficient* of (X,Y) is defined by

$$\rho_U = \lim_{u \uparrow 1} P(Y > G_X^{\leftarrow}(u) \mid X > G_Y^{\leftarrow}(u)), \tag{2}$$

provided the limit exists. If $\rho_U \in (0,1]$, then X and Y are called *asymptotically upper tail dependent*, if $\rho_U = 0$, they are called *asymptotically upper tail independent*.

For some situations, this measure may be an appropriate extreme dependence measure; this is true, in particular, when the bivariate distribution is symmetric; see Example 1. However, ρ_U is not a very informative measure, since the extreme dependence around the line with angle $\pi/4$ does not reveal much about what happens elsewhere; see e.g. the asymmetric model in Example 2. As a remedy we suggest an extension of the upper tail dependence coefficient to a function of the angle, which measures extreme dependence in any direction in the first quadrant of $\mathbb{R}^2$. Its derivation is based on multivariate extreme value theory and we indicate this relationship in Section 2. We shall, however, refrain from a precise derivation and rather refer to Hsing et al. (2004) for details. We also want to emphasize that one-dimensional extreme value theory has been applied successfully to risk management problems; see Embrechts (2000). We remark further that one-dimensional extreme value theory has meanwhile reached a consolidated state; we refer to Embrechts et al. (1997) or Coles (2001) as standard references.

We will illustrate our results by a direct application to a real data set. The complete data set we investigated consists of high-frequency data for three different stocks: Intel, Cisco and General Motors (GM). We have full sample paths of the price data of the stocks between February and October 2002 from the Trades and Quotes TAQ database of the New York Stock Exchange (NYSE); i.e. our data consists of all trading dates [in seconds] and corresponding prices [in cents]. They are depicted in

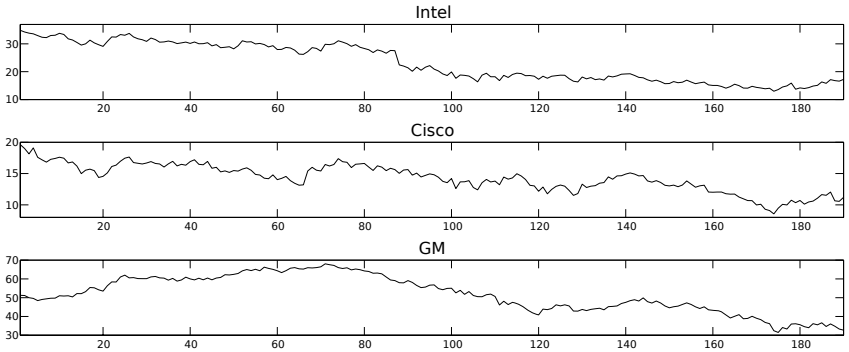


Fig. 1 Example of stock prices: Intel, Cisco and GM: Feb-Oct 2002.

Figure 1. This dataset has to be filtered in different steps due to false data, seasonality and serial dependence; see Section 4 for details.

After these filtering steps the residuals can be assumed iid and the dependence structure between the three stocks can be investigated. The first row of Figure 2 shows scatter plots of different combinations of the filtered stocks. We have estimated the means, variances and the correlation of the data in each scatter plot. The second row shows simulated normal data with the estimated parameters.

For extreme risk assessment one is particularly interested in the left lower corner and we have zoomed into this corner to get a more precise account of the dependence there; see Figure 3. None of the normal models seem to be able to capture the dependence structure in this area.

Our paper is organized as follows. After introducing the tail dependence function in Section 2 we shall present some examples including an asymmetric Pareto model and the bivariate normal model.

In Section 3 we introduce a simple nonparametric estimation procedure of the tail dependence function. We show its performance in various simulation examples and plots.

In Section 4 we investigate our high-frequency data in more detail and estimate their tail dependence function. We also show various plots to visualize our results. Finally, in Section 5 we conclude the paper with a summary of our findings.

2 Measuring Extreme Dependence

Although the upper tail dependence coefficient and its functional extension we are aiming at can be defined for random vectors of any dimension, we restrict ourselves in our presentation to the bivariate case. For a general treatment in any dimension we refer to Hsing et al. (2004).

Suppose $(X_i, Y_i)_{i=1,\dots,n}$ is a sequence of iid vectors and (X, Y) is a generic random vector with the same distribution function $G(x,y) = P(X \le x, Y \le y)$ for $(x,y) \in \mathbb{R}^2$

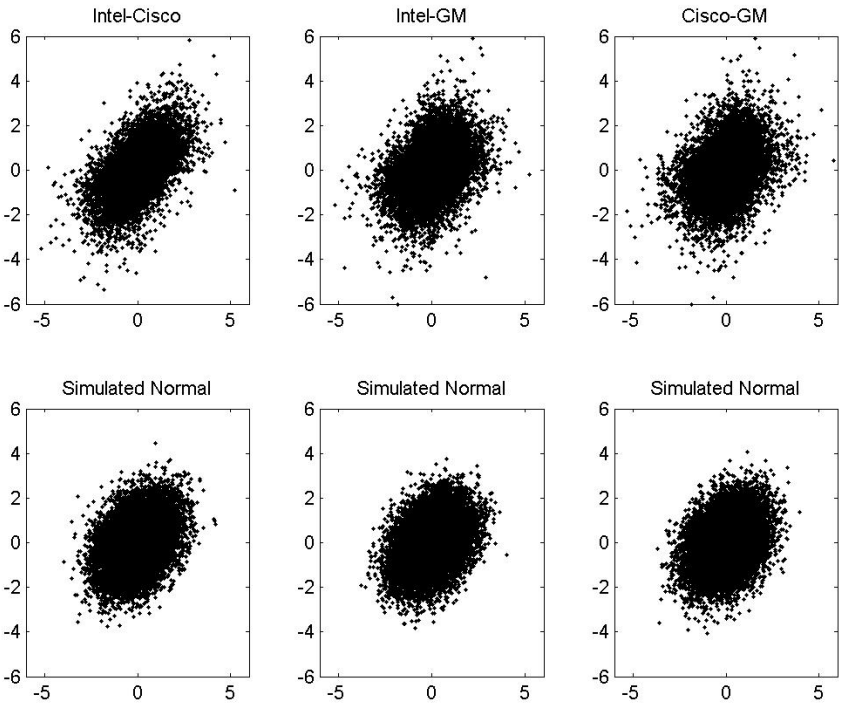


Fig. 2 Bivariate stock data versus simulated normal with same (estimated) means and variances.

with continuous marginals. For $n \in \mathbb{N}$ define the vector of componentwise maxima $M_n = (\max_{i=1,\ldots,n} X_i\,, \max_{i=1,\ldots,n} Y_i)$. As a first goal we want to describe the behavior of M_n for large n.

It is a standard approach in extreme value theory to first transform the marginals to some appropriate common distribution and then model the dependence structure separately. As copulas have become a fairly standard notion for modelling dependence we follow this approach and transform the marginal distributions G_X and G_Y to uniform (0,1). Then we have a bivariate uniform distribution, which is called a *copula* and is given for $0 < u, v < 1$ by

$$C_G(u,v) = P(G_X(X) \le u\,, G_Y(Y) \le v) = P(X \le G_X^{\leftarrow}(u)\,, Y \le G_Y^{\leftarrow}(v))\,.$$

For more details on copulas and dependence structures in general we refer to Joe (1997); for applications of copulas in risk management see Embrechts et al. (2001).The transformation of the marginals to uniforms is illustrated in Figure 4.

Under weak regularity conditions on the bivariate distribution function G we obtain

$$\lim_{n\to\infty} P\left(\max_{i=1,\ldots,n} G_X(X_i) \le 1 + \frac{1}{n}\ln u\,,\ \max_{i=1,\ldots,n} G_Y(Y_i) \le 1 + \frac{1}{n}\ln v\right)$$

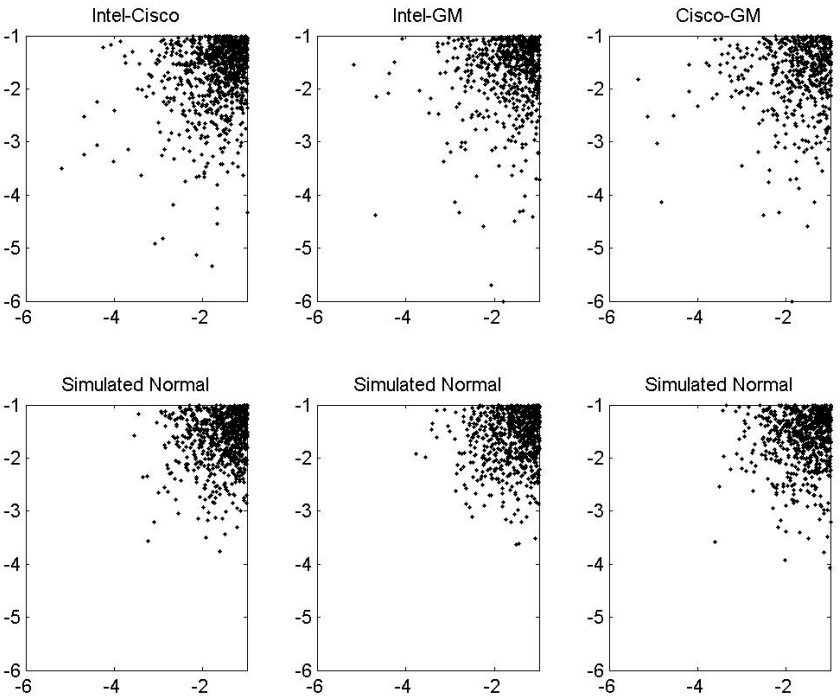


Fig. 3 Bivariate stock data versus simulated normal with same (estimated) means and variances.

$$= \exp(-\Lambda(-\ln u, -\ln v))) = C(u,v), \quad 0 \le u,v \le 1.$$

Such a copula is called *extreme copula* and satisfies for all $t > 0$

$$C^t(u,v) = C(u^t, v^t), \quad 0 < u,v < 1.$$

$C(u,v)$ has various integral representations. The *Pickands' representation* yields an extreme event intensity measure (we write $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$):

$$\begin{aligned} \Lambda(x,y) &= \lim_{n\to\infty} nP\left(G_X(X) > 1 - \frac{x}{n} \text{ or } G_Y(Y) > 1 - \frac{y}{n}\right) \\ &= \int_0^{\pi/2} \left(\frac{x}{1 \vee \cot\theta} \vee \frac{y}{1 \vee \tan\theta}\right) \Phi(d\theta), \quad x,y \ge 0. \end{aligned} \tag{1}$$

Φ is a finite measure on $(0, \pi/2)$ satisfying $\int_0^{\pi/2}(1 \wedge \tan\theta)\,\Phi(d\theta) = \int_0^{\pi/2}(1 \wedge \cot\theta)\,\Phi(d\theta) = 1$. The definition of Λ as a limit of $n \times$ *success probability* is a version of the classical limit theorem of Poisson. For large n the measure Λ can be interpreted as the mean number of data in a strip near the upper and right boundary of the uniform distribution; see Figure 4. We also recall some properties of $\tan\theta = \frac{1}{\cot\theta} = \frac{\sin\theta}{\cos\theta}$: $\tan 0 = 0$, $\tan\theta$ is increasing in $\theta \in (0, \pi/2)$ and $\lim_{\theta\to\pi/2} \tan\theta = \infty$. Then $\cot\theta$ is its reflection on the 45 degree line, correspond-

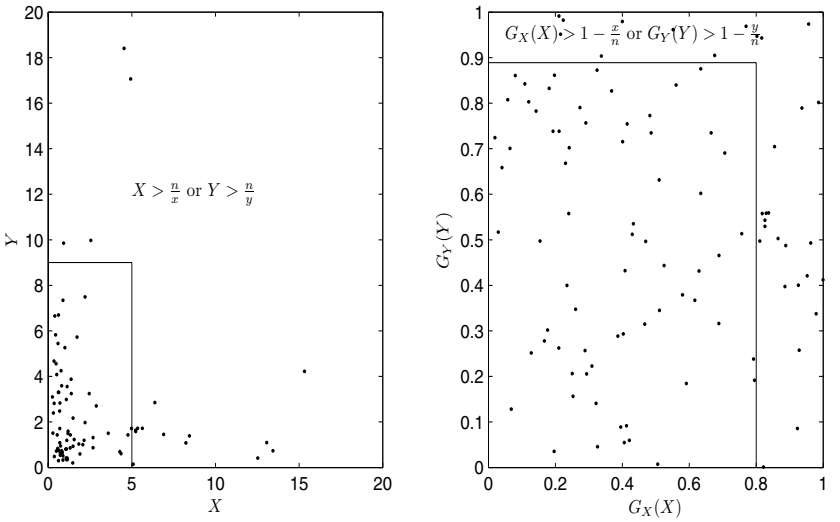


Fig. 4 Left plot: Simulated data for X and Y Fréchet distributed with distribution functions $G_X(x) = G_Y(x) = \exp(-1/x)$ for $x > 0$ with region of large data points indicated. The range of the data is in total $[0,410]$ for X and $[0,115]$ for Y; for reasons of presentation 14 extremely large points had to be left out.
Right plot: Illustration of the intensity measure Λ as defined in equation (1): Λ measures the probability in the strip near the upper and right boundary of the uniform distribution.

ing to $\theta = \pi/4$. Moreover, $\tan(\pi/4) = \cot(\pi/4) = 1$ and $\cot(\frac{\pi}{2} - \theta) = 1/\cot\theta$ for $\theta \in (0, \pi/2)$. Finally, arctan is the inverse function of tan.

The fact that $\Lambda(x,y) = x\Lambda(1, y/x)$ motivates the following definition.

Definition 3. For any random vector (X,Y) such that (1) holds we define the *dependence function* as

$$\psi(\theta) = \Lambda(1, \cot\theta), \quad 0 < \theta < \pi/2.$$

Note that $\psi(\cdot)$ is a function of the angle θ only and measures dependence in any direction of the positive quadrant of a bivariate distribution.

The following result shows that $\psi(\cdot)$ allows us to approximate for large x_1 and y_1 the probability for X or Y to become large. We write $a(x) \sim b(x)$ as $x \to x_0$ for $\lim_{x \to x_0} a(x)/b(x) = 1$. We also denote by $\overline{G}(\cdot) = 1 - G(\cdot)$ the tail of G.

Proposition 1. *Let (X,Y) be a random vector. If $x_1, y_1 \to \infty$ such that $P(X > x_1)/P(Y > y_1) \to \tan\theta$, then the following quotient converges for all $\theta \in (0, \pi/2)$,*

$$\frac{P(X > x_1 \text{ or } Y > y_1)}{P(X > x_1)}.$$

Furthermore, the limit is the dependence function $\psi(\theta)$.

Proof. From (4) we have for large x_1, y_1 and $x = n\overline{G}_X(x_1)$ and $y = n\overline{G}_Y(y_1)$ as $n \to \infty$ (note that x, x_1, y, y_1 depend on n):

$$P(X > x_1 \text{ or } Y > y_1) \sim \frac{1}{n}\Lambda(n\overline{G}_X(x_1), n\overline{G}_Y(y_1))$$
$$= \overline{G}_X(x_1)\Lambda\left(1, \frac{\overline{G}_Y(y_1)}{\overline{G}_X(x_1)}\right) = \overline{G}_X(x_1)\psi\left(\arctan\left(\frac{\overline{G}_X(x_1)}{\overline{G}_Y(y_1)}\right)\right). \tag{2}$$

We set

$$\theta = \arctan\left(\frac{\overline{G}_X(x_1)}{\overline{G}_Y(y_1)}\right)$$

and obtain the result.

The following corollary summarizes some obvious results; the symmetry property of part (d) is new and will prove useful for estimation purposes.

Corollary 1. *(a) For X and Y independent we calculate*

$$\frac{P(X > x_1 \text{ or } Y > y_1)}{P(X > x_1)} \sim \frac{P(X > x_1) + P(Y > y_1)}{P(X > x_1)} \to 1 + \cot\theta =: \psi_0(\theta)$$

for $x_1, y_1 \to \infty$ such that $P(Y > y_1)/P(X > x_1) \to \cot\theta$.
(b) For X and Y completely dependent, i.e. $X = g(Y)$ with probability 1 for some increasing function g, we obtain

$$\frac{P(X > x_1 \text{ or } Y > y_1)}{P(X > x_1)} = \frac{P(X > x_1) \vee P(X > y_1)}{P(X > x_1)} \to 1 \vee \cot\theta =: \psi_1(\theta)$$

for $x_1, y_1 \to \infty$ such that $P(Y > y_1)/P(X > x_1) \to \cot\theta$.
(c) $\psi_1(\theta) \le \psi(\theta) \le \psi_0(\theta)$ for $0 < \theta < \pi/2$.
(d) $\psi_{Y,X}(\theta) = \cot\theta\, \psi_{X,Y}(\pi/2 - \theta)$.

Proof. It only remains to proof part (d). By Example 4.3 in Hsing et al. (2004) we have together with the change of variables $x = t\tan\theta$,

$$1 + \cot\theta - \psi_{X,Y}(\theta) = \lim_{t\to\infty} P(X > G_X^{\leftarrow}(1 - 1/(t\tan\theta)) | Y > G_Y^{\leftarrow}(1 - 1/t))$$
$$= \lim_{t\to\infty} P(Y > G_Y^{\leftarrow}(1 - 1/t) | X > G_X^{\leftarrow}(1 - 1/(t\tan\theta))) \frac{P(X > G_X^{\leftarrow}(1 - 1/(t\tan\theta)))}{P(Y > G_Y^{\leftarrow}(1 - 1/t))}$$
$$= \cot\theta\,(1 + \tan\theta - \psi_{Y,X}(\pi/2 - \theta)) = \cot\theta + 1 - \cot\theta\, \psi_{Y,X}(\pi/2 - \theta).$$

We normalize $\psi(\cdot)$ to the interval $[0, 1]$ as follows.

Definition 4. The normalized function

$$\rho(\theta) = \frac{\psi_0(\theta) - \psi(\theta)}{\psi_0(\theta) - \psi_1(\theta)} = \frac{1 + \cot\theta - \psi(\theta)}{1 \wedge \cot\theta}, \quad 0 < \theta < \pi/2,$$

we call *tail dependence function*.

Note that ρ describes the tail dependence of (X,Y) in any direction of the bivariate distribution on the positive quadrant of $\mathbb{R}^2$.

By this definition we have $\rho(\theta) \in [0,1]$ for all $0 < \theta < \pi/2$, $\rho(\theta) \equiv 0$ in case of independence and $\rho(\theta) \equiv 1$ in case of complete dependence. Consequently, $\rho(\theta)$ being close to 0/1 corresponds to weak/strong extreme dependence.

Remark 1. (i) (Relation between tail dependence function and Pickands' dependence function.) We can write an extreme copula as

$$C(u,v) = \exp\left(\log(uv) A\left(\frac{\log(v)}{\log(uv)}\right)\right), \quad 0 < u,v < 1.$$

The function $A : [0,1] \to [\frac{1}{2}, 1]$ is called *Pickands' dependence function*. $A \equiv 1$ corresponds to independence and $A(t) = t \vee (1-t)$ to total dependence. Using $-\Lambda(-\log(u), -\log(v)) = \log(C(u,v))$ we have the following relation between ρ and A:

$$\rho(\theta) = \frac{(1 + \cot\theta)(1 - A\left(\frac{\cot\theta}{1+\cot\theta}\right))}{1 \wedge \cot\theta}, \quad 0 < \theta < \pi/2.$$

(ii) For elliptical copula models a new semi-parametric approach for extreme dependence modelling was suggested and investigated in Klüppelberg et al. (2007, 2008).

The function $\rho(\cdot)$ is invariant under monotone transformation of the marginal distributions. We show this by calculating it as a function of the copula.

Proposition 2. *Let (X,Y) be a random vector with continuous marginal distribution functions G_X and G_Y. Then $G_X(X) \stackrel{d}{=} U$ and $G_Y(Y) \stackrel{d}{=} V$ for uniform random variables U and V with the same dependence structure as (X,Y). Denote by $C(u,v) = P(U \le u, V \le v)$ the corresponding copula. We also relate the arguments by $G_X(x_1) = u$ and $G_Y(y_1) = v$. Then, provided that the limits exist,*

$$\rho(\theta) = \lim_{\substack{u,v \to 1 \\ (1-u)/(1-v) \to \tan\theta}} \frac{1 - u - v + C(u,v)}{(1-u) \wedge (1-v)}, \quad 0 < \theta < \pi/2.$$

Proof.

$$\psi(\theta) = \lim_{\substack{x_1, y_1 \to \infty \\ \overline{G}_X(x_1)/\overline{G}_Y(y_1) \to \tan\theta}} \frac{1 - P(X \le x_1, Y \le y_1)}{P(X > x_1)} = \lim_{\substack{u,v \to 1 \\ (1-u)/(1-v) \to \tan\theta}} \frac{1 - C(u,v)}{1-u}.$$

Remark 2. Note also that the quantity $\rho(\pi/4)$ is nothing but the *(upper) tail dependence coefficient* ρ_U as defined in (2). Thus, the function ρ extends this notion from

a single direction, the 45 degree line corresponding to $\theta = \pi/4$, to all directions in $(0, \pi/2)$.

This extension is illustrated by the following examples.

Example 1. [Gumbel copula]
Let (X,Y) be a bivariate random vector with dependence structure given by a Gumbel copula for $\delta \in [1,\infty)$:

$$C(u,v) = \exp\left\{-\left[(-\ln u)^\delta + (-\ln v)^\delta\right]^{1/\delta}\right\}, \quad 0 < u,v < 1\,. \tag{3}$$

The dependence arises from δ. To calculate $\psi(\theta)$ we use the relationship of ψ to its copula. We use also the fact that for $u, v \to 1$ we have

$$\frac{-\ln v}{-\ln u} \sim \frac{1-v}{1-u} \to \cot\theta\,.$$

Then by continuity of u^x in x we obtain for $u, v \to 1$ such that $(1-v)/(1-u) \to \cot\theta$

$$1 - C(u,v) = 1 - \exp\left(\ln u\left[1 + \left(\frac{-\ln v}{-\ln u}\right)^\delta\right]^{1/\delta}\right) \sim 1 - u^{(1+(\cot\theta)^\delta)^{1/\delta}}\,.$$

Using the l'Hospital rule and the fact that $u \to 1$, we obtain

$$\frac{1-C(u,v)}{1-u} \to \left(1 + (\cot\theta)^\delta\right)^{1/\delta},$$

and hence

$$\rho(\theta) = \frac{1 + \cot\theta - \left(1 + (\cot\theta)^\delta\right)^{1/\delta}}{1 \wedge \cot\theta}, \quad 0 < \theta < \pi/2\,.$$

We also obtain the well-known upper tail dependence coefficient $\rho_U = \rho(\pi/4) = 2 - 2^{1/\delta}$.

Our next result concerns models, whose extreme dependence vanishes in the limit.

Proposition 3. *Let (X,Y) be a random vector with continuous marginal distribution functions G_X and G_Y. If $\rho(\theta_0) = 0$ for some $\theta_0 \in (0,\pi/2)$ then $\rho(\theta) = 0$ for all $\theta \in (0,\pi/2)$.*

Proof. From Corollary 1(d) we have

$$\rho_{Y,X}(\pi/2 - \theta) = \rho(\theta), \quad 0 < \pi/2 < 1\,. \tag{4}$$

Now note that $P(X > G_X^{\leftarrow}(1 - 1/(t\tan\theta)) | Y > G_Y^{\leftarrow}(1-1/t))$ is decreasing in θ, hence if $\rho(\theta_0) = 0$ then $\rho(\theta) = 0$ for $\theta > \theta_0$. Now, assume that $\rho(\pi/4) = 0$ so that

$\rho_{Y,X}(\pi/4) = 0$ by (4). This results in $\rho(\theta) = 0$ and $\rho_{Y,X}(\theta) = 0$ for $\theta > \pi/4$, i.e $\rho \equiv 0$ by (4) and monotonicity. Hence, we only have to show that $\rho(\theta_0) = 0$ for some $\theta_0 \in (0, \pi/2)$ implies $\rho(\pi/4) = 0$. This is trivial for $\theta_0 < \pi/4$ by monotonicity. For $\theta_0 > \pi/4$, (4) gives $\rho_{Y,X}(\pi/2 - \theta_0) = 0$ for $\pi/2 - \theta_0 < \pi/4$, so that $\rho_{Y,X}(\pi/4) = \rho(\pi/4) = 0$ and this finishes the proof.

We conclude with the multivariate normal distribution. It is well-known (see e.g. Embrechts et al. 2001, 2002) that for correlation $\rho < 1$ the upper tail dependence coefficient is $\rho_U = 0$. Consequently, Proposition 3 gives the following result.

Corollary 2. *For a bivariate normal distribution with correlation $\rho < 1$ we have $\rho \equiv 0$.*

The following example is a typical model to capture risk in the extremes.

Example 2. [Asymmetric Pareto model]
For $p_1, p_2 \in (0,1)$ set $\overline{p}_1 = 1 - p_1$ and $\overline{p}_2 = 1 - p_2$ and consider the model

$$X = p_1 Z_1 \vee \overline{p}_1 Z_2 \quad \text{and} \quad Y = p_2 Z_1 \vee \overline{p}_2 Z_3$$

with Z_1, Z_2, Z_3 iid Pareto(1) distributed; i.e., $P(Z_i > x) = x^{-1}$ for $x \geq 1$. Clearly, the dependence between X and Y arises from the common component Z_1. Hence the dependence is stronger for larger values of p_1, p_2. We calculate the function ρ, and observe first that by independence of the Z_i for $x \to \infty$,

$$\begin{aligned} P(X > x) &= 1 - P(p_1 Z_1 \vee \overline{p}_1 Z_2 \leq x) = 1 - P(p_1 Z_1 \leq x) P(\overline{p}_1 Z_2 \leq x) \\ &= 1 - \left(1 - \frac{p_1}{x}\right)\left(1 - \frac{\overline{p}_1}{x}\right) \sim \frac{1}{x}(p_1 + \overline{p}_1) = \frac{1}{x}. \end{aligned}$$

Consequently, we choose $y = x \tan\theta$, which satisfies the conditions of Proposition 1 and calculate similarly,

$$\begin{aligned} &P(X > x \text{ or } Y > x \tan\theta) = 1 - P(X \leq x, Y \leq x \tan\theta) \\ &= 1 - P\left(Z_1 \leq \frac{x}{p_1} \wedge \frac{x \tan\theta}{p_2}\right) P\left(Z_2 \leq \frac{x}{\overline{p}_1}\right) P\left(Z_3 \leq \frac{x \tan\theta}{\overline{p}_2}\right) \\ &\sim \frac{1}{x}(p_1 \vee p_2 \cot\theta + \overline{p}_1 + \overline{p}_2 \cot\theta), \end{aligned}$$

which implies $\psi(\theta) = 1 + \cot\theta - p_1 \wedge p_2 \cot\theta$ for $0 < \theta < \pi/2$ and

$$\rho(\theta) = \frac{p_1 \wedge p_2 \cot\theta}{1 \wedge \cot\theta}, \quad 0 < \theta < \pi/2.$$

An important class of distributions are those with Pareto-like tails. Proposition 4 ensures that, within this class, multivariate returns on different timescales have the same extremal (spatial) dependence, provided the observations are independent and have no time series structure. Hence, one can take advantage of the fact that a higher frequency results in a larger sample and is easier to estimate. We shall illustrate this in

Section 4.5. This version of the proof of Proposition 4 was kindly communicated to the first author by Patrik Albin. Also, one can find a similar proposition in Hauksson et al. (2001) in the setting of multivariate regular variation.

Proposition 4. *Let (X,Y) be a random vector with marginal tails $\overline{G}_X$ and $\overline{G}_Y$ that are regularly varying at infinity, with indices $\alpha < 0$ and $\beta < 0$, respectively. Denote by X^{*n} the sum of n iid copies of X and define Y^{*n} analogously. If the limit*

$$\lim_{t\to\infty} P\big(X > G_X^{\leftarrow}(1-\lambda/t)\,\big|\,Y > G_Y^{\leftarrow}(1-1/t)\big) = L(\lambda) \quad \textit{exists for } \lambda > 0, \tag{5}$$

then the following hold:
(a) $P\big(X^{*n} > x, Y^{*n} > y\big) \sim nP(X > x, Y > y)$ *as* $x,y \to \infty$*;*
(b) The marginal tails $\overline{G}_{X^{*n}}$ *and* $\overline{G}_{Y^{*n}}$ *of* X^{*n} *and* Y^{*n} *satisfy for all* $n \geq 2$

$$\lim_{t\to\infty} P\big(X^{*n} > G_{X^{*n}}^{\leftarrow}(1-\lambda/t)\,\big|\,Y^{*n} > G_{Y^{*n}}^{\leftarrow}(1-1/t)\big) = L(\lambda) \quad \textit{for } \lambda > 0\,.$$

Proof. (a) The one-dimensional version of this result goes back to Feller and has been extended to the larger class of subexponential random variables (see e.g. Embrechts et al. 1997, Appendix A3); i.e. we have

$$P\big(X^{*n} > t\big) \sim nP(X > t) \quad \text{and} \quad P\big(Y^{*n} > t\big) \sim nP(Y > t) \quad \text{as } t \to \infty. \tag{6}$$

We prove a bivariate version of this result. For $\varepsilon > 0$ sufficiently small, we have

$$\begin{aligned}
P\big(X^{*n} > x, Y^{*n} > y\big) &\leq \sum_{i=1}^{n}\sum_{j=1}^{n} P\big(X_i > (1-(n-1)\varepsilon)x, Y_j > (1-(n-1)\varepsilon)y\big) \\
&\quad + \sum_{1\leq i\neq k\leq n}\sum_{j=1}^{n} P\big(X_i > \varepsilon x, X_k > \varepsilon x, Y_j > (1-(n-1)\varepsilon)y\big) \\
&\quad + \sum_{i=1}^{n}\sum_{1\leq j\neq l\leq n} P\big(X_i > (1-(n-1)\varepsilon)x, Y_j > \varepsilon y, Y_l > \varepsilon y\big) \\
&\quad + \sum_{1\leq i\neq k\leq n}\sum_{1\leq j\neq l\leq n} P\big(X_i > \varepsilon x, X_k > \varepsilon x, Y_j > \varepsilon y, Y_l > \varepsilon y\big) \\
&\leq nP(X > (1-(n-1)\varepsilon)x, Y > (1-(n-1)\varepsilon)y) \\
&\quad + n^2 P(X > (1-(n-1)\varepsilon)x)\,P(Y > (1-(n-1)\varepsilon)y) \\
&\quad + 2n^2 P(X > \varepsilon x, Y > \varepsilon y)\,(P(X > \varepsilon x) + P(Y > \varepsilon y)) \\
&\quad + n^3 P(X > \varepsilon x)\,P(Y > \varepsilon y)\,(P(X > \varepsilon x) + P(Y > \varepsilon y)) \\
&\quad + n^2 P(X > (1-(n-1)\varepsilon)x)\,P(Y > (1-(n-1)\varepsilon)y) \\
&\quad + n^2 P(X > \varepsilon x, Y > \varepsilon y)^2 \\
&\quad + n^3 P(X > \varepsilon x, Y > \varepsilon y)\,P(X > \varepsilon x)\,P(Y > \varepsilon y) \\
&\quad + n^4 P(X > \varepsilon x)^2\,P(Y > \varepsilon y)^2 \\
&\sim nP(X > (1-(n-1)\varepsilon)x, Y > (1-(n-1)\varepsilon)y) \quad \text{as } x,y \to \infty
\end{aligned} \tag{7}$$

by (5) together with the regular variation properties. Now, using again an $\varepsilon > 0$ and properties of disjoints sets together with the Boolean inequality, we estimate

$$\begin{aligned}
&P\big(X^{*n} > x, Y^{*n} > y\big) \\
&\geq \sum_{i=1}^{n} P\big(X_i > (1+(n-1)\varepsilon)x, Y_i > (1+(n-1)\varepsilon)y \\
&\qquad\qquad \bigcap_{j\neq i}\{-\varepsilon x \leq X_j \leq x, -\varepsilon y \leq Y_j \leq y\}\big) \\
&\geq \sum_{i=1}^{n} P\big(X_i > (1+(n-1)\varepsilon)x, Y_i > (1+(n-1)\varepsilon)y\big) \\
&\quad - \sum_{i=1}^{n}\sum_{j\neq i} P\big(X_i > (1+(n-1)\varepsilon)x, Y_i > (1+(n-1)\varepsilon)y, X_j \notin [-\varepsilon x, x]\big) \\
&\quad - \sum_{i=1}^{n}\sum_{j\neq i} P\big(X_i > (1+(n-1)\varepsilon)x, Y_i > (1+(n-1)\varepsilon)y, Y_j \notin [-\varepsilon y, y]\big) \\
&\sim nP(X > (1+(n-1)\varepsilon)x, Y > (1+(n-1)\varepsilon)y) \quad \text{as } x,y \to \infty.
\end{aligned}$$

(b) Proposition 1.5.15 of Bingham et al. (1987) ensures that the generalized inverses satisfy as $t \to \infty$,

$$G_X^{\leftarrow}(1-1/t) \sim G_{X^{*n}}^{\leftarrow}(1-n/t) \quad \text{and} \quad G_Y^{\leftarrow}(1-1/t) \sim G_{Y^{*n}}^{\leftarrow}(1-n/t). \tag{8}$$

In particular, $G_X^{\leftarrow}(1-1/\cdot)$ and $G_{X^{*n}}^{\leftarrow}(1-1/\cdot)$ are regularly varying with index $1/\alpha$, while $G_Y^{\leftarrow}(1-1/\cdot)$ and $G_{Y^{*n}}^{\leftarrow}(1-1/\cdot)$ are regularly varying with index $1/\beta$.

By (5)-(7), we have (with ε not the same as before)

$$\begin{aligned}
&\limsup_{t\to\infty} P\big(X^{*n} > G_{X^{*n}}^{\leftarrow}(1-\lambda/t) \,\big|\, Y^{*n} > G_{Y^{*n}}^{\leftarrow}(1-1/t)\big) \\
&\leq \limsup_{t\to\infty} \frac{nP\big(X > (1-\varepsilon)^{\beta/\alpha} G_X^{\leftarrow}(1-\lambda/(nt)), Y > (1-\varepsilon)G_Y^{\leftarrow}(1-1/(nt))\big)}{nP\big(Y > (1+\varepsilon)G_Y^{\leftarrow}(1-1/(nt))\big)} \\
&\leq \limsup_{t\to\infty} \frac{P\big(X > G_X^{\leftarrow}(1-(1-2\varepsilon)^{\beta}\lambda/(nt)), Y > G_Y^{\leftarrow}(1-(1-2\varepsilon)^{\beta}/(nt))\big)}{((1+\varepsilon)/(1-3\varepsilon))^{\beta} P\big(Y > (1-3\varepsilon)G_Y^{\leftarrow}(1-1/(nt))\big]} \\
&\leq \Big(\frac{1-3\varepsilon}{1+\varepsilon}\Big)^{\beta} \limsup_{t\to\infty} P\Big(X > G_X^{\leftarrow}\Big(1-\frac{(1-2\varepsilon)^{\beta}\lambda}{nt}\Big) \,\Big|\, Y > G_Y^{\leftarrow}\Big(1-\frac{(1-2\varepsilon)^{\beta}}{(nt)}\Big)\Big) \\
&= \Big(\frac{1-3\varepsilon}{1+\varepsilon}\Big)^{\beta} L(\lambda) \\
&\to L(\lambda) \quad \text{as } \varepsilon \downarrow 0.
\end{aligned}$$

Analogously follows from the reverse inequality in (a)

$$\liminf_{t\to\infty} P\big(X^{*n} > G_{X^{*n}}^{\leftarrow}(1-\lambda/t) \,\big|\, Y^{*n} > G_{Y^{*n}}^{\leftarrow}(1-1/t)\big) \geq L(\lambda).$$

Remark 3. In Example 4.3 in Hsing et al. (2004) we have, for X and Y random variables with continuous distributions G_X and G_Y,

$$\lim_{t\to\infty} P(X > G_X^{\leftarrow}(1-1/(t\tan\theta))|Y > G_Y^{\leftarrow}(1-1/t)) = (1\wedge\cot\theta)\rho(\theta)\,, 0<\theta<\pi/2\,.$$

Hence, for X and Y random variables with Pareto-like tails, setting $\lambda = 1/\tan\theta$ and $L(\cot\theta) = (1\wedge\cot\theta)\rho(\theta)$ we conclude $L(\lambda) = (1\wedge\lambda)\rho(\arctan(1/\lambda))$ for $\lambda > 0$.

Corollary 3. *Denote by $\psi(\theta)$ the dependence function of (X,Y). Let X^{*n} and Y^{*n} be the sum of n iid copies of X and Y, respectively, and denote by $\psi^{*n}(\cdot)$ the dependence function of (X^{*n},Y^{*n}) for $n\geq 2$. Then $\psi^{*n}(\theta) = \psi(\theta)$ for all $0<\theta<\pi/2$. The same holds for the tail dependence function $\rho(\theta)$.*

3 Extreme Dependence Estimation

To assess extreme dependence in data we estimate the tail dependence function $\rho(\cdot)$ on the positive quadrant. We use a nonparametric estimator as suggested in Hsing et al. (2004) based on the empirical distribution function, which yields a simple nonparametric estimator of $\psi(\cdot)$ and hence of $\rho(\cdot)$. Recall that the empirical distribution function given by

$$\widehat{G}_X(x) = \widehat{P}_n(X\leq x) = \frac{1}{n}\sum_{j=1}^{n} I(X_i\leq x)\,,\quad x\in\mathbb{R}\,,$$

is the standard estimator for the distribution function G_X of iid data ($I(\mathbf{A})$ denotes the indicator function of the set $\mathbf{A}$). The empirical distribution function can be rewritten in terms of the ranks of the sample variables X_i for $i = 1,\ldots,n$ and we write

$$\widehat{G}_X(X_i) = \widehat{P}_n(X\leq X_i) = \frac{1}{n}\mathrm{rank}(X_i)\,.$$

We still have to explain one important issue of our estimation procedure. Recall from (1), denoting by $\overline{G}_X(\cdot) = 1-G_X(\cdot)$ and $\overline{G}_Y(\cdot) = 1-G_Y(\cdot)$ for continuous G_X and G_Y, that

$$\begin{aligned}\Lambda_n(x,y) &:= nP\left(G_X(X) > 1-\frac{x}{n} \text{ or } G_Y(Y) > 1-\frac{y}{n}\right)\\ &= nP\left(n\overline{G}_X(X)\leq x \text{ or } n\overline{G}_Y(Y)\leq y\right)\\ &= nP(n(\overline{G}_X(X),\overline{G}_Y(Y))\in\mathbf{A})\\ &\to \Lambda(x,y)\quad n\to\infty\,.\end{aligned} \tag{9}$$

By a continuity argument we can replace $n\in\mathbb{N}$ by $t\in(0,\infty)$ and also replace in a first step the probability measure P by its empirical counterpart $\widehat{P}_n$. Then we obtain

$$\widehat{\Lambda}_{t,n}(x,y) = t\widehat{P}_n\left(t\left(\overline{G}_X(X),\overline{G}_Y(Y)\right)\in \mathbf{A}\right) = \frac{t}{n}\sum_{i=1}^{n} I(t\left(\overline{G}_X(X),\overline{G}_Y(Y)\right)\in \mathbf{A}).$$

Now estimate the two distribution tails by their empirical counterparts:

$$\widehat{\overline{G}}_X(X_i) := \frac{1}{n}R_i^X := \frac{1}{n}\text{rank}(-X_i) \quad \text{and} \quad \widehat{\overline{G}}_Y(Y_i) := \frac{1}{n}R_i^Y := \frac{1}{n}\text{rank}(-Y_i).$$

Then setting $\varepsilon = t/n$ we obtain

$$\widehat{\Lambda}_{\varepsilon,n}(\mathbf{A}) = \varepsilon\sum_{i=1}^{n} I(\varepsilon\left(R_i^X, R_i^Y\right)\in \mathbf{A}).$$

This yields in combination with Definition 4 an estimator for the function ρ:

$$\widehat{\rho}_{\varepsilon,n}(\theta) = \frac{1+\cot\theta - \widehat{\Lambda}_{\varepsilon,n}(1,\cot\theta)}{1\wedge\cot\theta}, \quad 0\le\theta\le\frac{\pi}{2}, \tag{10}$$

where $\widehat{\Lambda}_{\varepsilon,n}(1,\cot\theta)$ can be rewritten as

$$\varepsilon\sum_{i=1}^{n} I(R_i^X \le \varepsilon^{-1} \text{ or } R_i^Y \le \varepsilon^{-1}\cot\theta), \quad 0\le\theta\le\frac{\pi}{2}. \tag{11}$$

Choosing ε is not an easy task and when θ approaches $\pi/2$ increasingly fewer points are used in the estimation. In Hsing et al. (2004) this problem was solved by letting ε decrease slightly as θ approaches $\pi/2$. A much better solution is provided by the symmetry proved in Corollary 1(d) in combination with (4): the extreme dependence of (X,Y) for $\theta\in[\pi/4,\pi/2]$ is the same as the extreme dependence of (Y,X) for $\theta\in[0,\pi/4]$. Consequently, we estimate $\rho_{\varepsilon,n}(\theta)$ by estimating $\rho_{X,Y}(\theta)$ by

$$\widehat{\rho}_{\varepsilon,n}(\theta) := \begin{cases} \widehat{\rho}_{\varepsilon,n}^{XY}(\theta), & 0<\theta<\pi/4, \\ \widehat{\rho}_{\varepsilon,n}^{YX}(\pi/2-\theta), & \pi/4\le\theta<\pi/2. \end{cases} \tag{12}$$

In the following remark we summarize some important properties of $\widehat{\rho}_{\varepsilon,n}$.

Remark 4. (i) Estimator (12) has good convergence properties: for appropriately small ε and $n\to\infty$ it converges in probability and almost surely; see Hsing et al. (2004) and references therein.
(ii) To assess asymptotic dependence involves passing to a limit function, which for a finite sample is simply impossible. Consequently, for X and Y independent, even for very small ε it is highly possible that the estimated tail dependence function will be positive. This can be made precise by calculating

$$\begin{aligned}
&\varepsilon \sum_{i=1}^{n} I(R_i^X \le \varepsilon^{-1} \text{ or } R_i^Y \le \varepsilon^{-1}\cot\theta) \\
&= \varepsilon \sum_{i=1}^{n} (I(R_i^X \le \varepsilon^{-1}) + I(R_i^Y \le \varepsilon^{-1}\cot\theta)) - I(R_i^X \le \varepsilon^{-1} \text{ and } R_i^Y \le \varepsilon^{-1}\cot\theta)) \\
&= 1 + \cot\theta - \varepsilon \sum_{i=1}^{n} I(R_i^X \le \varepsilon^{-1} \text{ and } R_i^Y \le \varepsilon^{-1}\cot\theta)\,.
\end{aligned}$$

Now, independent samples for X and Y yield for fixed n, ε and θ

$$\sum_{i=1}^{n} I(R_i^X \le \varepsilon^{-1} \text{ and } R_i^Y \le \varepsilon^{-1}\cot\theta) \sim \text{Bin}\left(\frac{\cot\theta}{\varepsilon^2 n^2}, n\right).$$

Hence,

$$E\left(\varepsilon \sum_{i=1}^{n} I(R_i^X \le \varepsilon^{-1} \text{ or } R_i^Y \le \varepsilon^{-1}\cot\theta)\right) = 1 + \cot\theta - \frac{\cot\theta}{\varepsilon n}, \quad 0 < \theta < \frac{\pi}{2},$$

giving

$$E\left(\widehat{\rho}_{\varepsilon,n}(\theta)\right) = \begin{cases} \dfrac{\cot\theta}{\varepsilon n}, & 0 < \theta < \dfrac{\pi}{4}, \\ \dfrac{\cot(\pi/2-\theta)}{\varepsilon n}, & \dfrac{\pi}{4} \le \theta < \dfrac{\pi}{2}. \end{cases} \tag{13}$$

In much the same fashion we get

$$\text{Var}(\widehat{\rho}_{\varepsilon,n}(\theta)) = \begin{cases} \dfrac{\cot\theta}{n} - \dfrac{\cot^2\theta}{\varepsilon^2 n^3}, & 0 < \theta < \dfrac{\pi}{4}, \\ \dfrac{\cot(\pi/2-\theta)}{n} - \dfrac{\cot^2(\pi/2-\theta)}{\varepsilon^2 n^3}, & \dfrac{\pi}{4} \le \theta < \dfrac{\pi}{2}. \end{cases}$$

(iii) Inspecting equation 11 one can see that choosing an ε is equivalent to the estimation of $\rho(\theta)$ based on the $1/\varepsilon$ largest values of X. Hence, it is natural to see $1/\varepsilon$ as a threshold of the data and we will therefore use this term.
(iv) The estimator $\widehat{\rho}_{\varepsilon,n}$ has the advantage that it is only based on the ranks of the data. Consequently, it can be smoothed in the usual way. For instance, by averaging it over a window of size $2m+1$ for $m \in \mathbb{N}$, we call this smoothed estimator $\widehat{\rho}^{(m)}_{\varepsilon,n}(\cdot)$.

In the second column of Figures 5 and 6 we estimated $\rho(\theta)$ for the Gumbel copula (cf. Example 1) and the asymmetric Pareto model (cf. Example 2). The estimated tail dependence function is indeed (except for $\theta \in \{0, \pi/2\}$, where $E(\widehat{\rho}_{\varepsilon,n}(\cdot))$ has singularities) far away from $E(\widehat{\rho}_{\varepsilon,n}(\cdot))$. For our sample size and the chosen ε it is smaller than 0.075 for the interval depicted. Given that the variance is of the order n^{-1} the estimated extreme dependence in our data is significant.

Example 3. [Gumbel copula: continuation of Example 1]
In Figure 5 we simulated the model (with student-t marginals with 8 degrees of

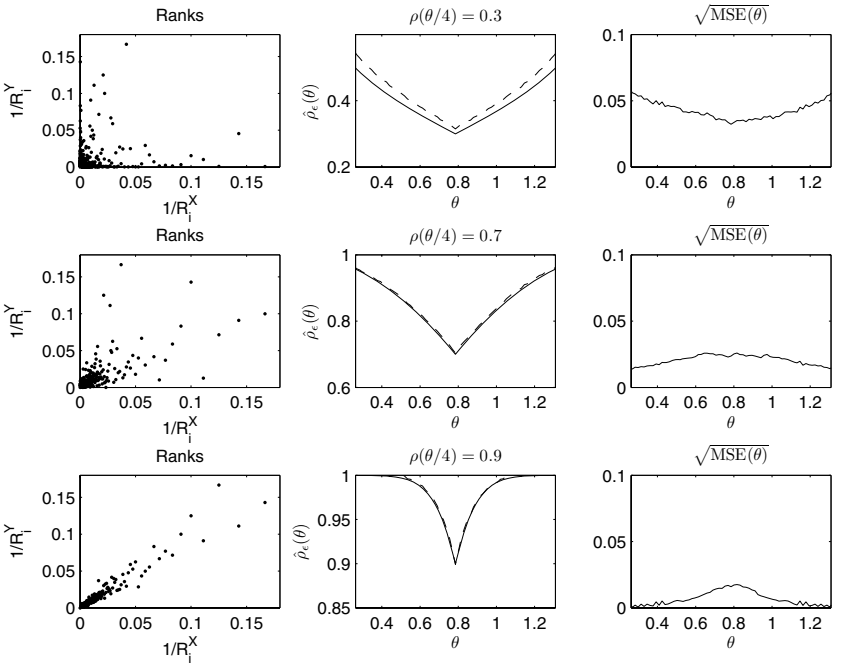


Fig. 5 Simulated Gumbel copula model for $\rho(\theta/4) = 0.3$ (upper row), $\rho(\theta/4) = 0.7$ (middle row), $\rho(\theta/4) = 0.9$ (lower row).
Left column: Plots of ranks $(1/R_i^X, 1/R_i^Y)$, with points close to $(1,1)$ truncated.
Middle column: Plots of $\widehat{\rho}(\theta)$ (dashed) overlaid with true function $\rho(\theta)$ (solid).
Right column: Estimation error in terms of $\sqrt{\mathrm{MSE}(\theta)}$.

freedom) for $n = 10\,000$ iid observations of (X,Y) 100 times. We estimate the tail dependence function $\rho(\cdot)$ for this model with $\varepsilon = 1/200$. We stay away from the boundaries $\theta = 0$ and $\theta = \pi/2$, since in the numerator of (10) we have the difference of two quantities which both tend to ∞ as $\theta \to 0$. The three sets of plots on the three rows correspond to the cases: $\rho(\pi/4) = 0.3$ (upper row), $\rho(\pi/4) = 0.7$ (middle row) and $\rho(\pi/4) = 0.9$ (lower row). On each row the left plots contain ranks $\left(1/R_i^X, 1/R_i^Y\right)$, $1 \le i \le n$, of a simulated sample of size 10 000. Points on the axes correspond to independent extreme points; all points in the open quadrant exhibit some extreme dependence structure. Completely dependent points are to be found on the 45-degree line. The level of dependence is manifested by the data scattered around this diagonal. The true functions $\rho(\theta)$ in (5) (solid) are overlaid with the estimated mean of $\widehat{\rho}_{\varepsilon,n}(\theta)$ (dashed) based on the simulated sample. The right plot depicts the squareroot of the estimated mean squared error. Note that $\rho(\pi/4)$ is the upper tail dependence coefficient, which is an appropriate and simple measure of extreme dependence for this symmetric model.

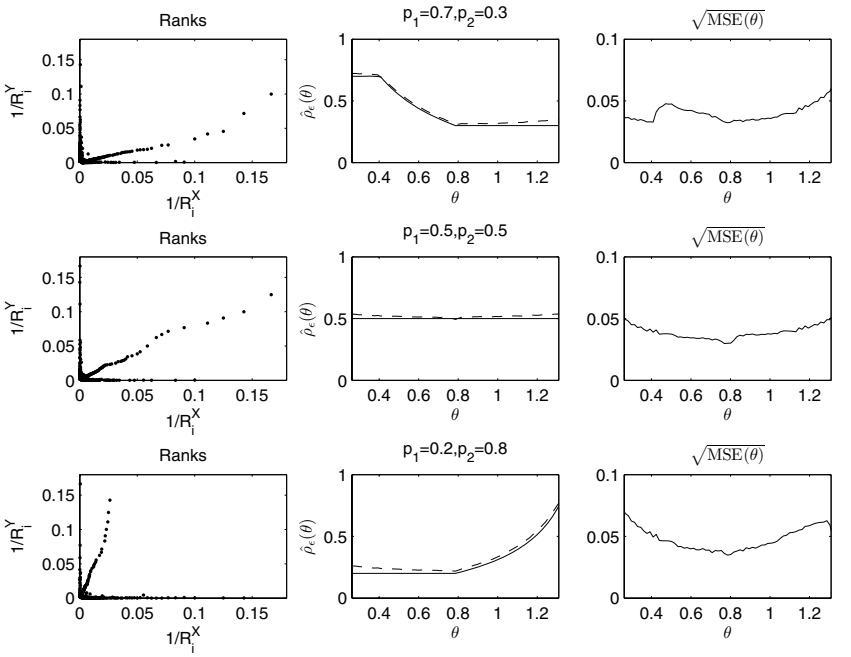


Fig. 6 Simulated asymmetric Pareto model with $\rho(\theta/4) = 0.3$ (upper row), $\rho(\theta/4) = 0.7$ (middle row), $\rho(\theta/4) = 0.9$ (lower row).
Left column: Plots of ranks $(1/R_i^X, 1/R_i^Y)$, with points close to $(1,1)$ truncated.
Middle column: Plots of $\widehat{\rho}(\theta)$ (dashed) overlaid with true function $\rho(\theta)$ (solid).
Right column: Estimation error in terms of $\sqrt{\mathrm{MSE}(\theta)}$.

Example 4. [Asymmetric Pareto model: continuation of Example 2]
In Figure 6 we simulated this model for $n = 10\,000$ iid observations of (X,Y) with $\varepsilon = 1/200$ 100 times. The three sets of plots on the three rows correspond to the cases: $(p_1\,, p_2) = (0.7\,, 0.3)$, $(p_1\,, p_2) = (0.5\,, 0.5)$ and $(p_1\,, p_2) = (0.2\,, 0.8)$. On each row the left plots contain ranks $\left(1/R_i^X, 1/R_i^Y\right)$, $1 \le i \le n$ of a simulated sample of size 10 000. The true functions $\rho(\theta)$ in (5) (solid) are overlaid with the estimated mean of $\widehat{\rho}_{\varepsilon,n}(\theta)$ (dashed) based on the simulated sample. The right plot depicts the squareroot of the estimated mean squared error.

In the first row of plots, ρ is larger for small θ than for large θ; this is reflected by the left plot in which the violation of independence can be seen to be more severe below the diagonal. In the second row of plots, ρ is constant; which is reflected by having a portion of extreme points lined up on the diagonal in the left plot. The third row of plots is the converse situation to the first row, which is reflected by the pattern of extreme points above the diagonal. This is an example of a situation where the tail dependence coefficient does not convey a good picture of extreme dependence, in that $\rho(\pi/4)$ is not sufficient to describe the full dependence structure of this model.

4 High-frequency Financial Data

We have tick-by-tick data of the *Trades and Quotes* database, in terms of trading times [in seconds] and prices [in 1 cent units] of three stocks traded between February and October 2002 on NYSE and Nasdaq. The stocks are General Motors (GM) from NYSE, and Intel and Cisco both from Nasdaq. One major difference between the two stock markets is that on NYSE trading is made on the floor while Nasdaq has electronic trading. We shall analyze the extreme dependence between the three stocks using the tail dependence function ρ. A study with focus on bivariate dependence structures on FX spot data has been performed by Breymann et al. (2003) and Dias & Embrechts (2003). Also, FX spot data was studied within the concept of multivariate regular variation in Hauksson et al. (2001). For cleaning and deseasonalizing our data we mainly follow the methods applied in these papers; see also there for further references. In these papers *parametric bivariate copulas* were fitted to FX spot data in both non-extreme and extreme regions. Our study considers the extremal dependence for stock data, which is estimated *nonparametrically*. One main difference between stock data and FX spot data is that FX spot data is traded 24 hours per day. In contrast, NYSE for instance, has regular opening hours between 9.30 and 16.00 on working days. This introduces additional complexity into our data analysis, and we have to deal with this problem.

When dealing with extremes it is of importance to use as much data as possible, since extremes are consequences of rare events. However, we can not simply use the full samples of all stocks as each single time series is not stationary and, even worse, for high-frequency data the different time series are not synchronized. As a remedy for the non-synchronous data we take subsamples of logreturns on specific timescales. If one chooses a relatively high frequency, one is confronted with the problem that tick prices are discrete, and also microstructure noise effects can enter. We chose 5 minutes logreturns as the lowest frequency, thus avoiding microstructure noise effects.

There are a number of issues which appear when dealing with high-frequency data and we will describe them in turns.

4.1 Cleaning the Data

A full sample path of stock data contains a huge amount of information. At Nasdaq there is almost a trade every second. However, some ticks are false, mostly due to fake quotes and decimal errors.

To be able to continue the analysis one has to clean the data. This is done by filtering the data and removing values that differ too much from their neighboring values in the sense of logarithmic differences. Also, sometimes false values may come in clusters, which one also has to deal with. The selection of thresholds for removing a bad tick was done by visually inspecting the time series before and after the cleaning. When a false tick was observed it was replaced by a value based on

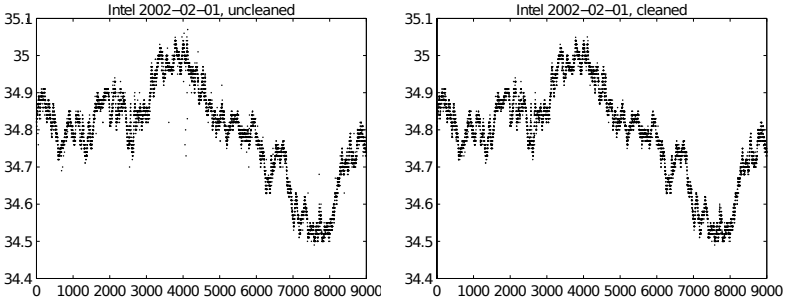


Fig. 7 Intel ticks during 9:35 to 11:45 on February 1, 2002. Left: Raw data. Right: Data cleaned from false ticks as described in Section 4.1.

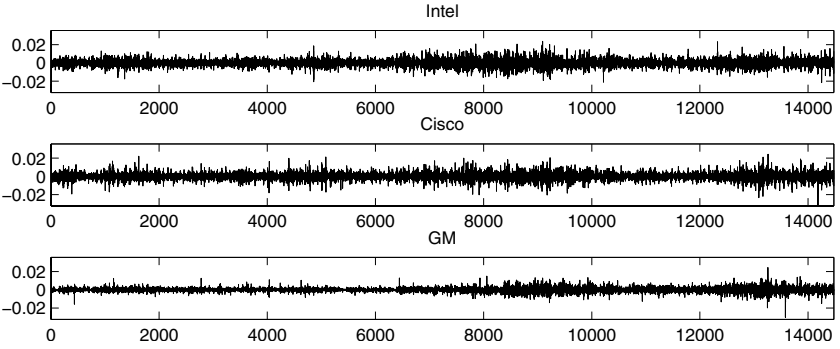


Fig. 8 Synchronized 5 minutes logreturns for Intel, Cisco and GM between 9:35 and 16:00 during February 1 to October 31, 2002

linear interpolation with its neighbors. In this way less than one percent of the data was removed.

The thresholds for logarithmic difference were set to 0.1% for Intel and Cisco and to 0.2% for GM, respectively. The reason for different thresholds is that Intel and Cisco are traded at a much higher frequency. A result of the cleaning procedure can be seen in Figure 7. We repeated our analysis after altering the thresholds slightly. However, this sensitivity analysis did basically not change the results.

When dealing with information from a stock exchange one is faced with the problem that they do not trade for 24 hours resulting in a gap of information, when the stock market is closed over night. However, Nasdaq and NYSE have off-hour trading, but prices behave differently than prices during the regular opening times as the trading rules differ. To obtain synchronised data we only considered the stocks between 9:35 to 16:00 from Monday to Friday using the previous tick method, which results in 77 five minutes logreturns per day. Also, there were a couple of holidays where no data were available. Finally we had 14 476 synchronized observations (5 minutes logreturns) for each stock, which we plot in Figure 8

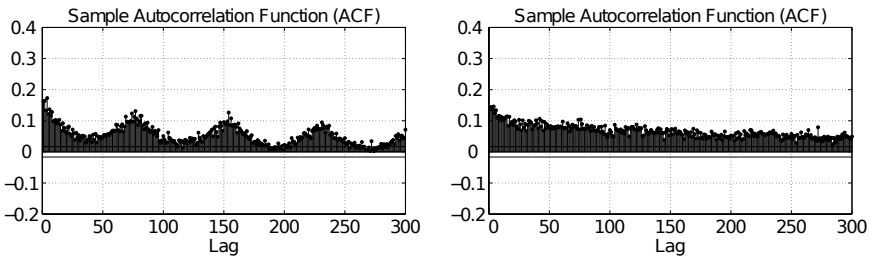


Fig. 9 Autocorrelation function of squared 5 minutes logreturns for Intel.
Left: Original data: Visible is the cycle of 77 lags indicating daily seasonality.
Right: Deseasonalized data as in (15) based on daily seasonality.

4.2 Deseasonalizing the Data

When investigating the 5 minutes logreturns closer one can detect seasonality in the data. In Figure 9, we depict the autocorrelation of the squared logreturns for Intel. Here one can see the daily seasonality. A comparison to the FX data in Breymann et al. (2003) shows that FX data have a much clearer weekly seasonality.

To be able to remove the seasonality, there are two main approaches. The first one is to time-change the logreturns to a business clock instead of the physical clock. The second is to use volatility weighting. We chose the second one as it is not clear how to choose a business clock for multivariate time series.

Volatility weighting divides a period (we first take a week) into several smaller subperiods and then estimates the seasonality effect in each subperiod in terms of volatility. Then each subperiod is deseasonalized separately by devolatization. We chose 5 minutes intervals as subperiods. This means that our observed returns, $\tilde{x}_t$, is a realization of the process

$$\tilde{x}_t = \mu + v_t x_t.$$

where x_t are the deseasonalized returns, μ is a constant drift and v_t is the seasonality coefficient (volatility weights), estimated by

$$\widehat{v}_\tau = \sqrt{\frac{1}{N^\tau}\sum_{i=1}^{N^\tau}(\tilde{x}_{t_i+\tau})^2}. \tag{14}$$

Here N^τ is the number of weeks, during which we have observed our stocks in the given subperiod $\tau \in \{0,5,10,15,\ldots\}$ (in minutes), and t_i denotes the start of week i which always is on Monday at 9:35. Also, τ has to be corrected for nights and weekends. We estimate μ with the sample mean $\widehat{\mu}$ of the logreturns. Hence, the deseasonalized 5 minutes logreturns are

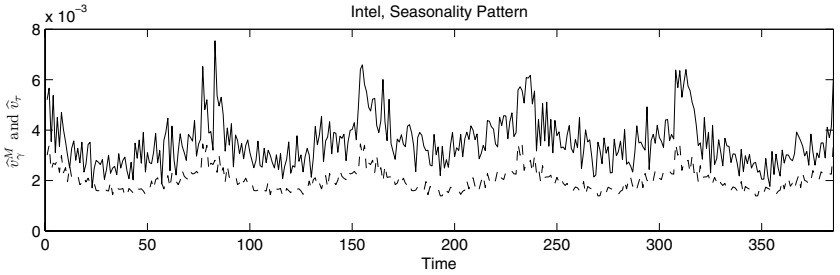


Fig. 10 Volatility weights $\widehat{v}_\tau$ and $\widehat{v}_\gamma^M$ for 5 minutes Intel logreturns: weekly estimated by (14), $\widehat{v}_\tau$, (solid) and daily robust estimated by (16), $\widehat{v}_\gamma^M$, (dashed).

$$x_t = \frac{\tilde{x}_t - \widehat{\mu}}{\widehat{v}_t}. \tag{15}$$

However, as we only have about 40 weeks the estimated volatility weights are quite noisy, see Figure 10. This is due to the fact that single large values can dominate $\widehat{v}_\tau^2$: the mean taken over 40 weeks is not sufficiently smooth.

To overcome this problem we first assume a daily seasonality instead of the weekly. This can be motivated by the fact that the different days do not seem to differ to a higher degree; see Figure 10. However, single large values still dominate the volatility weights, which is unsatisfactory.

Consequently, we use a robust estimator based on the median and absolute values:

$$\widehat{v}_\gamma^M = \text{median}_{i=1,\ldots,N^\gamma} |\tilde{x}_{t_i+\gamma}|. \tag{16}$$

Here N^γ is the number of days, during which we have observed our stocks in the given subperiod $\gamma \in \{0,5,\ldots,385\}$ (in minutes). We can now observe the stylistic pattern of the autocorrelation of squared logreturns in Figure 9 for our deseasonalized time series using the robustly estimated volatility weights.

The depicted volatility weights can be seen in Figure 10. One can clearly see that trading is more intense at the beginning and at the end of a day. We also observe that the robustly estimated volatility weights are much more stable. The deseasonalization removes seasonality in the squared logreturns, which are right skewed, hence the difference in magnitude for the two estimation methods.

When comparing the two different deseasonalization methods the robust one leaves more larger absolute values in the data, which occur in low trading time. The non-robust version decreases them as large values contribute much more to the volatility weights. Hence, the non-robust version of the deseasonalization makes the time series smoother than the robust method does.

Table 1 AIC-based values for (r,m,p,q,ν) and t and normally distributed residuals with corresponding likelihood; the last column presents our model.

Stock	t	normal	our model
Intel	55531 (2,3,5,2,8.2)	55869 (2,3,4,1,-)	55534 (0,1,5,2,8.2)
Cisco	55805 (0,1,3,3,6.5)	56596 (0,1,4,0,-)	55807 (0,1,1,1,6.5)
GM	57343 (2,2,5,1,5.5)	58467 (1,2,3,1,-)	57355 (0,3,1,1,5.5)

4.3 Filtering the Data

Because of the dependence, which we have observed in the autocorrelation for the squared logreturns, we will assume a stochastic volatility model for each stock. We model the mean by an ARMA process and use the standard GARCH(p,q) model for the martingale part. The model selection is based on the AIC criterion, the results are summarized in Table 1.

We model the logreturns for different equidistant frequencies by

$$x_t = \mu_t + \sigma_t z_t$$

with $\mu_t = c + \sum_{i=1}^{r} \phi_i x_{t-i} + \sum_{i=1}^{m} \theta_i \varepsilon_{t-i}$ and $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}^2$, where $\varepsilon_t = \sigma_t z_t$. We model the z_t by a standardnormal or a student-t distribution with ν degrees of freedom. The overall fit of the model was assessed by a residual analysis. We applied the Ljung-Box test for serial correlation, where we tested the residuals and the squared residuals, and the Kolmogorov-Smirnov test for goodness-of-fit of the normal and student-t distribution.

As we only have logreturns for 9:35-16:00 Monday to Friday we will make an error if we fit the time series model to our data without taking the missing values into account. We have used three different approaches to circumvent this problem:

(1) We (wrongly) fit the ARMA-GARCH model directly to the deseasonalized data, ignoring the missing observations during the nights completely.
(2) We estimated the logreturns during the nights by 5 minutes logreturns using the (wrong) square root scaling (the correct but complicated scaling constants have been calculated by Drost & Nijman 1993). Then we deseasonalize and fit the time series.
(3) We fit different MA(1)-GARCH(1,1) models for each day. In this case we used the estimated volatility of the previous day as the initial value.

One comment to the second approach is that the deseasonalized nightly logreturns should have the same distribution as the deseasonalized daily logreturns. We have tested this assumption via QQ-plots with bootstrapped confidence interval. Using ordinary bootstrap we can conclude that the deseaonalized nightly logreturns do not have the same distribution as the deseasonalized daily logreturns. However, as we have dependence in our time series one should use a bootstrap method which takes this into consideration. Using block bootstrap we can not reject the hypothesis that

Table 2 Estimated α by the Hill estimator for the loss region.

Stock	Intel	Cisco	GM
$\widehat{\alpha}$	5.6	4.36	4.3

the deseaonalized nightly logreturns have the same distribution as the deseasonalized daily logreturns.

If we compare the methods (1)-(3) we conclude that the first and second behave very similar with respect to the parameter estimation. For the third method this estimation was difficult. Even if we only use a three parameter model the estimation is not stable. Based on the Ljung-Box test for serial correlation, both residuals and squared residuals, the two first methods out-perform the third. Also, for the final result in Section 4.4 the outcome is similar. We concluded that the error of using a false approach (among these three) is minimal and concentrated on the first method for simplicity.

In Table 1 we have selected the model by AIC criteria, also giving the likelihood of the selected model. The selected optimal order of the ARMA model $m, r \in (0, \ldots, 5)$ and the order of the GARCH model $p, q \in (0, \ldots, 6)$ for normal and also the degree of freedom ν for t-distributed innovations are given in the first two columns.

As we want to keep the number of parameters as low as possible, we performed a sensitivity analysis based on the likelihood of the model. In this way we found the model given in column 3 of Table 1, which we will use in the sequel. Our analysis also confirmed the common knowledge that residuals are heavy-tailed; i.e. the t-distribution outperforms by far the normal distribution.

Concerning the Ljung-Box test, we could not reject independence of the residuals or the squared residuals for all time series. In Table 3 we show the p-values for a selection of lags for squared logreturns. We have also looked at the p-values up to 50 lags. However, all time series failed the Kolmogorov-Smirnov test, actually for all models presented in Table 1.

Diagnostic tools from extreme value theory (see e.g. Embrechts et al. 1997, Section 6.1) show, however, clearly that all three filtered time series are heavy-tailed. Consequently, we model the far out distribution tail of all residuals as regularly varying and estimate the tail index α by the Hill estimator (see e.g. Embrechts et al. (1997), Section 6.4). We summarize the result in Table 2.

Due to the devolatizaton a 10 minutes logreturn is obtained as a linear combination of two 5 minutes logreturns and so logreturns should have the same tail-parameter for different frequencies. However, for higher timescales the tail-parameter increases slightly, even if one compares the filtered 5 minutes logreturns with 45 minutes, but still remains heavy-tailed. This is well known and reported, for instance, in Müller et al. (1998).

We have also investigated the cross-correlation between the stocks. In Table 4 we display the first four lags. The other lags were smaller in absolute magnitude.

From Table 4 one can see that GM tends to follow Intel and Cisco more than vice versa. A formal test for uncorrelation of two time series tests this hypothesis for each specific lag based on asymptotic normality of the cross-correlation function (see e.g.

Table 3 p-values from the Ljung-Box test of filtered squared residuals. We have tested 1, 5, and 10 lags of the 5 minutes logreturns.

Stock	1	5	10
Intel	0.39	0.06	0.07
Cisco	0.74	0.86	0.77
GM	0.92	0.96	0.84

Table 4 Cross-correlation of the first four lags for the filtered 5 minutes logreturns.

Stocks	-4	-3	-2	-1	0	1	2	3	4
Intel-Cisco	-0.01	0.01	0.01	0.05	0.56	0.03	0.02	0.01	-0.00
Intel-GM	0.02	0.02	0.05	0.04	0.35	0.02	0.01	-0.00	-0.02
Cisco-GM	0.03	0.01	0.04	0.04	0.33	0.03	0.01	-0.00	-0.02

Brockwell & Davis 1987, Theorem 11.2.2) The uncorrelation hypothesis is rejected if the corresponding estimate has absolute value larger than 0.017.

For the 15 minutes data, there is some cross-correlation between GM and Intel and GM and Cisco for the first lag, but none significant between Intel and Cisco. For the 30 minutes data, there is no significant cross-correlation at all.

Such tests have to be interpreted with caution for various reasons. First of all there is the usual problem that a test should be performed not only on each lag separately. Furthermore, the amount of high-frequency data is so large that a formal test rejects already for very small cross-correlation: for 5 min the rejection level is 0.017, for 30 min it is 0.042.

4.4 Analyzing the Extreme Dependence

Recall the estimator $\widehat{\rho}_{\varepsilon,n}$ from (12), where $\varepsilon = t/n$ represents the proportion of upper order statistics used for the estimation, which itself has to be estimated; cf. Remark 4(iii). The estimation of ε involves in extreme value statistics a variance-bias tradeoff; i.e. it is tricky and time-consuming, but important. We have used two approaches.

Firstly, by plotting the estimated tail dependence function for different choices of ε visual inspection clearly showed the influence of the variance/bias, when using different thresholds. For high threshold, i.e. small ε, the estimated tail dependence function was rather rough showing the high variation of the estimator. When decreasing the threshold the estimated tail dependence function became very smooth, which we interpreted in analogy to tail index estimation as a bias.

Secondly, we studied plots of $\widehat{\rho}_{\varepsilon}(\theta)$ as a function of ε for fixed θ. This was done for $\theta = \pi/4 \pm 0, \pi/12, \pi/6$. Here we looked for regions where $\rho(\theta)$ was stable. The case $\theta = \pi/4$ can be seen in Figure 11. We want to mention that for other choices of θ the stability plots were not equally convincing.

As a result of our diagnostics we fixed $\varepsilon = 1/650$, which represents about 4.5% of the data. In Figure 12 we can see the resulting estimated tail dependence function.

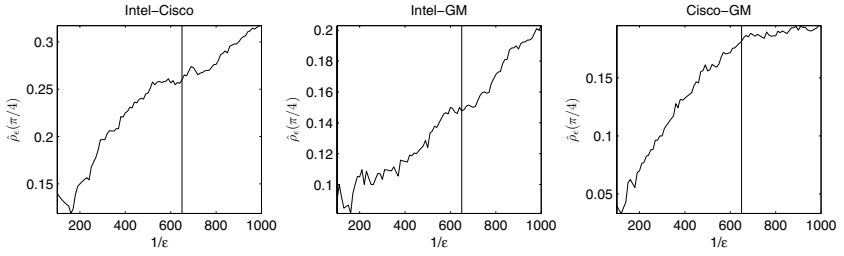


Fig. 11 For the 5 minutes logreturns: $\widehat{\rho}_\varepsilon(\pi/4)$ as a function of ε for $\varepsilon = 1/100, \ldots, 1/1000$.

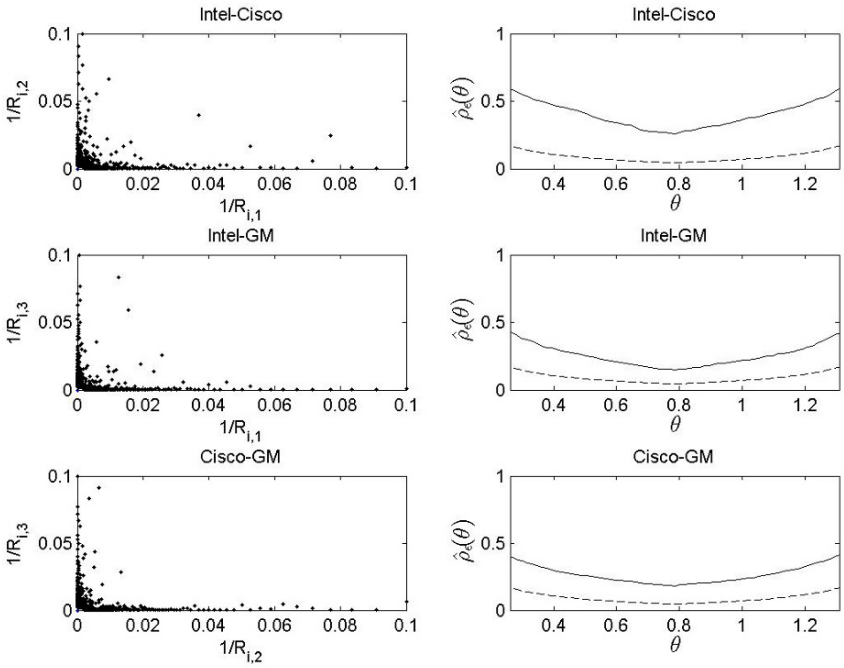


Fig. 12 Left plots: $1/R_{i,j}$, where $R_{i,j} = \text{rank}(-X_{i,j})$. Right plots: Estimators of $\rho(\theta)$ (solid line). For sake of reference we have also plotted the expected dependence $E(\widehat{\rho}_{\varepsilon,n}(\cdot))$ from (13) for independent samples (dashed line).

We conclude that for all bivariate combinations of our data tail dependence can be modelled symmetric and is significantly stronger than for the independent case. Not surprisingly, dependence is highest between Intel and Cisco, presumably due to branch dependence, besides being both traded at Nasdaq. The dependence of GM and Cisco is slightly higher than of GM and Intel. The symmetry in the dependence reflects that we have three major stocks and can also be viewed as underlying market dependence. We also notice that the estimated tail dependence function looks similar as the tail dependence function of a bivariate extreme value distribution with a Gumbel copula.

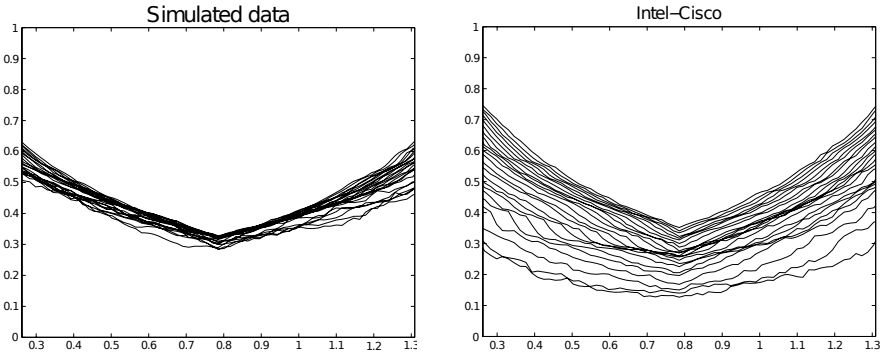


Fig. 13 Estimated tail dependence function $\widehat{\rho}_\varepsilon(\theta)$ for different ε (1/200 to 1/1200). Left: Simulated data from the t-Gumbel model with parameters estimated from the Intel-Cisco filtered 5 minutes logreturns. Right: Filtered 5 minutes logreturns, Intel-Cisco.

It is now tempting to fit a distribution with t marginals (a common model in econometrics, called the t-GARCH) with degree of freedom from Table 1 and a Gumbel copula (cf. Example 1,3) pairwise to our data or even to the three-dimensional sample. We know already from Section 4.3 that the t distribution is not a good model for the marginals. However, our concern is now for the extreme dependence structure, and it turns out that the Gumbel copula, although an extreme value copula, is not a valid model. The dependence structure in our data is far more complex. This can be illustrated by viewing the tail dependence function for different ε (1/200 to 1/1200) compared to data simulated from the above t-Gumbel model with the same sample size, presented in Figure 13. Recall from Example 1 that the Gumbel coupla gives tail dependence function

$$\rho(\theta) = \frac{1+\cot\theta-(1+(\cot\theta)^\delta)^{(1/\delta)}}{1\wedge\cot\theta}, \quad 0<\theta<\pi/2.$$

We estimated δ from the upper tail dependence coefficient $\widehat{\rho}_\varepsilon(\pi/4) = 2-2^{(1/\widehat{\delta})}$, the value of the estimated tail dependence function at $\pi/4$. We obtained $\widehat{\delta} = 1.25$. Now we generate a sample of the same size as the 5 minutes logreturns with a Gumble($\widehat{\delta}$) copula and t-distributed marginals. We compare the estimated tail dependence functions for different ε and present the results in Figure 13.

Notice that the simulated data behave much more stably with respect to changes of ε, while the real data reacts heavily on such changes. Using a parametric model such as the Gumble copula would only be an approximation based on one given threshold.

Table 5 Correlation change for different timescales.

Stock	Intel-Cisco	Intel-GM	Cisco-GM
5	0.56	0.35	0.33
15	0.62	0.41	0.38
30	0.65	0.43	0.39
45	0.66	0.46	0.41

4.5 Different Timescales

As a result of our statistical analysis of the marginal data, the one-dimensional logreturns exhibit Pareto-like tails. If the stocks came from a three-dimensional exponential Lévy process with appropriate dependence structure, then the extreme dependence would be the same for all time scales, i.e. 5 minutes logreturns of the three stocks would have the same dependence structure as daily logreturns. This applies in particular to extreme dependence and is in accordance with Proposition 4. Note that our data do not satisfy the independence condition of Proposition 4. Extreme value estimates, however, often extend properties from independent data to dependent data.

We shall at least perform a statistical test to our data, whether there is a change in the extreme dependence on different timescales by analyzing logreturns of 5, 15, 30 and 45 minutes frequencies.

To this end we performed the same filtering steps as for the 5 minutes logreturns again for the 15, 30 and 45 minutes logreturns obtained from the raw data. Then we fitted a MA(1)-GARCH(1,1) model with student-t distributed residuals to the deseasonalized data of the 15, 30 and 45 minutes logreturns. For the 5 minutes logreturns we keep the model from Section 4.4.

For the 15, 30 and 45 minutes logreturns we applied the Ljung-Box test for serial correlation, where we tested the residuals and the squared residuals, and the Kolmogorov-Smirnov test for goodness-of-fit of the student-t density. Observe that the degrees of freedom is not the same as in Table 1 for the different timescales. All the filtered time series passed the Ljung-Box test and the filtered 30 and 45 minutes logreturns passed the Kolmogorov-Smirnov test.

For the residuals we again estimate the dependence between the different stocks.

A comparisons of the linear dependence for different timescales is presented in Table 5. Here we can see that the correlation increases for higher frequencies; this effect is well-known and also called the Epps effect; see Zhang (2006).

Next we estimate the tail dependence function for the different frequencies. To compensate for the increasing lack of data for low frequencies, the ε is always chosen so that ε times the number of observations is the same for all frequencies. Hence we always consider the same quantile.

As the sample of the 45 minutes logreturns is only about 10 percent in size of the 5 minutes logreturns, they set the standard for the other frequencies. We increased the threshold $1/\varepsilon$ until the estimated tail dependence function behaved stably for the 45 minutes logreturns. We also studied a plot of $\widehat{\rho}_\varepsilon(\theta)$ for various values of θ,

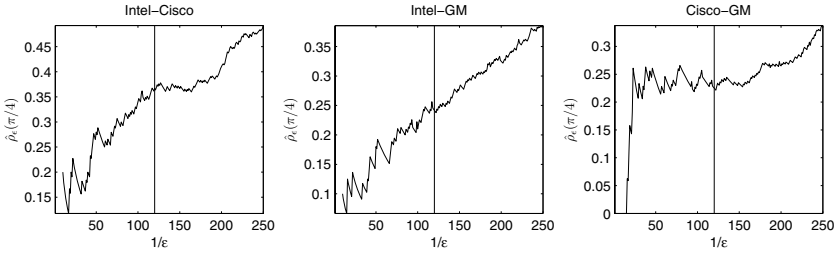


Fig. 14 For the 45 minutes filtered logreturns we depict $\widehat{\rho}_\varepsilon(\pi/4)$ as a function of ε for $\varepsilon = 1/10, \ldots, 1/250$.

when altering ε. For $\theta = \pi/4$ the result can be seen in Figure 14. Finally, we chose $\varepsilon = 1/120$, which represents about eight percent of the data. We want to remark that, in view of Figure 11, we presumably introduced a bias into our estimation.

Also, by using straight forward bootstrap techniques one can present bootstrap confidence intervals. In Figure 16 we depict the tail dependence function for Intel-Cisco on the timescales 5 minutes and 45 minutes.

From Figures 15 and 16 we can conclude that the tail dependence is approximately the same for different timescales. This also holds for different ε but there are some variations if we increase the threshold. If we lower the threshold, then the similarities between the different timescales become more pronounced. We recall that in Table 2 on p. 9 in Breymann et al. (2003) the tail dependence coefficient $\rho(\pi/4)$ is estimated via a parametric model for different timescales for DEM (Deutsche Mark) and JPY (Japanese Yen). Even for the unfiltered data in that paper the estimator for $\rho(\pi/4)$ looks stable. If we increase the timescale to, for instant, two hours the extreme dependence starts to deviate unless we lower the threshold and use as much as 15% of the data. This is consistent with the result on high-frequency FX data reported in Hauksson et al. (2001). Hence, the two different asset classes seem to share the same time scaling for extreme dependence. As pointed out earlier in this section, the time scaling is also explained from a theoretical point of view via Proposition 4.

From the above analysis we conclude that we can estimate extreme dependence for lower frequencies by estimating it for high frequencies, where enough data are available.

Another possibility to achieve a more stable estimation procedure is to use subsampling, based on different samples of the same frequency, obtained by time shifts. We performed the estimation separately for each subsample and, at the end, averaged over all estimated tail dependence functions. The subsamples proved to be very stable in the basic statistics, the estimates for the ARMA and GARCH parameters, and also the properties of the residuals. However, for the estimated tail dependence functions we cannot report significant improvement, in particular, when compared to the tail dependence function estimated from higher frequencies.

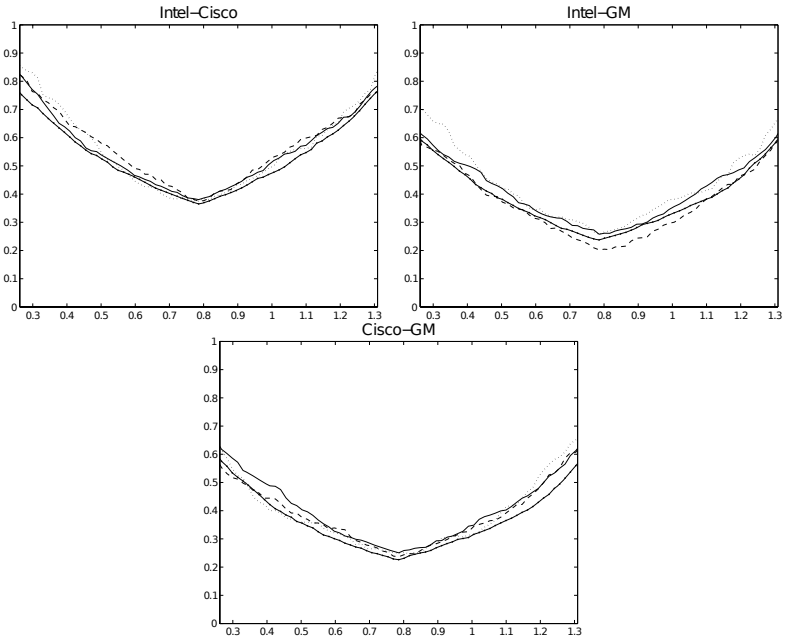


Fig. 15 Estimated tail dependence function $\widehat{\rho}_\varepsilon(\theta)$ of filtered logreturns for different frequencies. Five minutes (straight-dotted), 15 minutes (straight), 30 minutes (dashed) and 45 minutes (dotted).

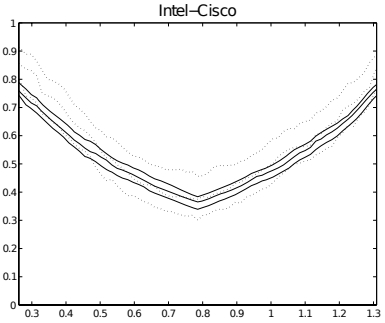


Fig. 16 Estimated tail dependence function $\widehat{\rho}_\varepsilon(\theta)$ with bootstrap confidence intervals (100 resamples) of filtered logreturns for timescale 5 and 45 minutes. Five minutes with corresponding confidence intervals (straight) and 45 minutes with corresponding confidence intervals (dotted).

4.6 Dependence Under Filtering

Recall that we have in principle prices which are multiples of one cent, but there are values our logreturn will never take. Moreover, we have an unnaturally large amount of zeroes and small values. However, concerning extreme dependence we can rest assured that this does not affect the tails.

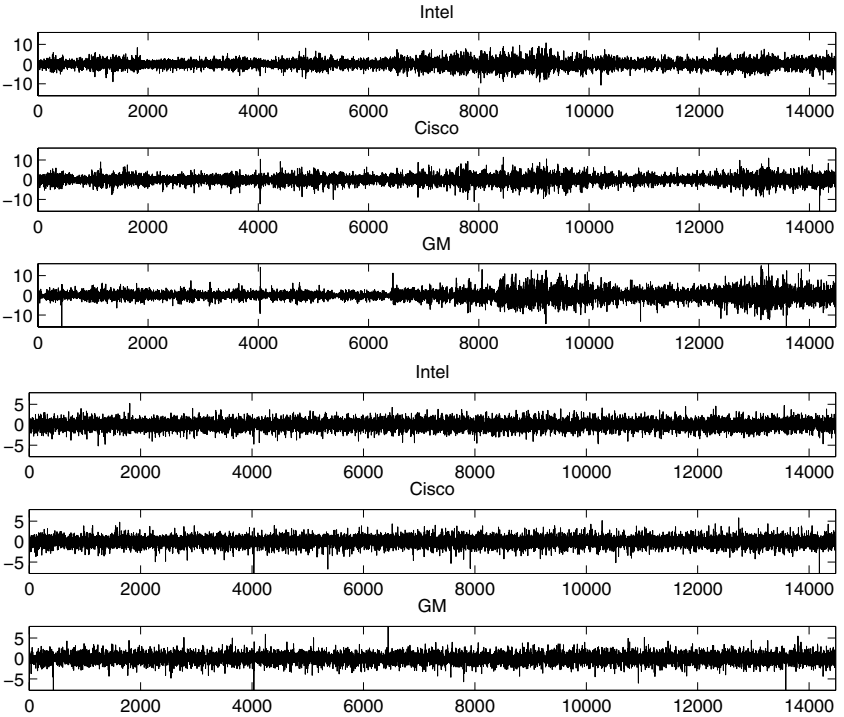


Fig. 17 Upper three plots: Deseasonalized 5 minutes logreturns modelled by daily seasonality and coefficients estimated by the robust method (16). Lower three plots: GARCH filtered logreturns based on our model in Table 1. Compare to the raw data in Figure 8.

Table 6 Estimated correlation for the 5 minutes logreturns for different steps in the data analysis.

Stocks	Original	Deseasonalized	Filtered
Intel-Cisco	0.57	0.56	0.56
Intel-GM	0.36	0.36	0.35
Cisco-GM	0.33	0.33	0.33

Now we shall investigate, how the dependence structure has changed during the filtering steps. In Table 6 we can see the correlation between the 5 minutes logreturns for the different steps of the filtering.

It is satisfactory to see that the different filtering steps have obviously not changed the correlation and hence not changed the linear dependence between the different stocks. This also holds for other timescales.

Now we turn to an account of extreme dependence before and after filtering. When examining the logreturns in Figure 8 one can clearly see the dominating volatile periods. The same holds for Figure 17. Taking the same ε for the raw data and the filtered returns yields for the raw data an over representation of the volatile periods. This implies that one would consider in fact only a much smaller time period for the extreme value analysis. So theoretically, there is no reason, why extreme dependence

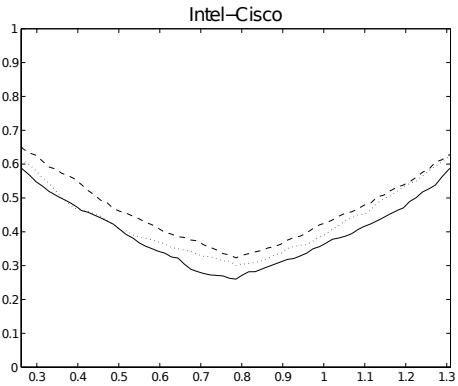


Fig. 18 Estimated tail dependence function $\widehat{\rho}_{\varepsilon}(\theta)$ for 5 minutes Intel logreturns. Dashed: Unfiltered data. Dotted: Deseasonalized data. Solid: GARCH filtered data.

before and after filtering should be similar. In Figure 18 we have plotted the estimated tail dependence function for Intel and Cisco after each filtering step for 5 minutes logreturns. We have used the same ε as in Section 4.4 and the same ε for the different filtering steps. One can see that there seems to be only a small difference in magnitude, not in shape. This also holds for different choices of ε and different timescales. Consequently, for our data the rather complicated filtering procedure seems to be obsolete for a realistic account of the extreme dependence.

5 Conclusion

We have introduced a new estimator for the tail dependence function, which is tailor made to assess the extreme dependence structure in data. As it measures dependence in every direction it is in principle also able to measure extreme dependence for data with asymmetric dependence structure. We show the performance of this function for high-frequency data for varying frequencies.

After giving some theoretical results, which are important in the high-frequency context, we clean the data carefully and perform some basic statistics. We then show the tail dependence function at work for our data and estimate extreme dependence for high-frequency stock data.

We have investigated the extremal dependence between Intel, Cisco and GM for different time scales. We can conclude for the filtered data:

- All three stocks have heavy tails. Within the 5,10,15,45 minutes frequencies we observed that a lower frequency gives lighter tails.
- We can work with the hypothesis that the square root scaled deseasonalized nightly logreturns have the same distribution as the deseasonalized daily logreturns.
- There is (weak) cross-correlation between the stocks for frequencies of up to 30 minutes, it disappears for lower frequencies.

- The extreme dependence is symmetric which means that the stocks influence each other to the same degree. This can be interpreted as market dependence.
- The IT stocks (Cisco and Intel) have stronger dependence indicating branch dependence.
- Extreme dependence is there, but moderate. We have the same extreme dependence for different timescales. This is consistent with the result on high-frequency FX data reported in Hauksson et al. (2001). Hence, the two different asset classes seems to share the same time scaling for extreme dependence. The time scaling is also explained from a theoretical point of view via Proposition 4.
- The filtering steps do not alter the extreme dependence to a high degree.
- Higher correlation does not necessarily lead to stronger extreme dependence.

Our analysis shows again that extreme value theory has to be applied with care. To obtain a realistic picture about the extreme dependence structure in real data it is not enough to describe it by one single number. Another obvious lesson to draw from our analysis is that it is important to use reference results such as simulations from exact models. Moreover, a message, which we can not repeat too often, one should be careful when selecting the threshold.

Acknowledgements E.B. takes pleasure to thank Patrik Albin for generously providing the version the proof of Proposition 4, Holger Rootzén for fruitful discussions, Catalin Starica for suggesting the median filter and the Stochastic Center, Chalmers for a travelling grant. He also thanks the Center for Mathematical Sciences of the Munich University of Technology for a very stimulating and friendly atmosphere during a much needed research stay.

References

Bingham, N.H., Goldie, C.M. & Teugels, J.L. (1987) *Regular Variation.* Cambridge University Press. Cambridge.

Breymann, W. Dias, A. & Embrechts, P. (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance* **3**: 1–14.

Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods,* 2nd edition. Springer, New York.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

Dias, A. & Embrechts, P. (2003). Dynamic copula models for multivariate high-frequency data in finance. *Preprint, ETH Zurich.*

Drost, F. C. & Nijman, T.E. (1993). Temporal aggregation of GARCH processes *Econometrica* **61**: 909–927.

Embrechts, P. (Ed.) (2000). *Extremes and Integrated Risk Management.* UBS Warburg and Risk Books.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.

Embrechts, P., Lindskog, F. & McNeil, A. (2001). Modelling dependence with copulas and applications to risk management. In: Rachev, S. (Ed.) *Handbook of Heavy Tailed Distributions in Finance.* Elsevier, Chapter 8, pp. 329–384.

Embrechts, P., McNeil, A. & Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. In: Dempster, M. and Moffatt, H.K. (Eds.) *Risk Management: Value at Risk and Beyond.* Cambridge University Press, Cambridge.

Hauksson, H., Dacorogna, M., Domenig, T., Müller, U. & Samorodnitsky, G. (2001) Multivariate extremes, aggregation and risk estimation. *Quantitative Finance* **1**(1): 79–95.

Hsing, T., Klüppelberg, C. & Kuhn, G. (2004). Dependence estimation and visualization in multivariate extremes with applications to financial data. *Extremes* **7**: 99–121.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman & Hall, London.

Klüppelberg, C. & Kuhn, G. (2009). Copula structure analysis. *J. Royal Stat. Soc., Series B* **71**(3), 737–753.

Klüppelberg, C., Kuhn, G. & Peng, L. (2008). Semi-parametric models for the multivariate tail dependence function - the asymptotically dependent case. *Scand. J. Stat.* **35**(4): 701–718.

Klüppelberg, C., Kuhn, G. & Peng, L. (2007). Estimating the tail dependence of an elliptical distribution. *Bernoulli* **13**(1): 229–251.

Müller U. A., A. Dacorogna M. M. & Pictet, O. V. (1998). Heavy tails in high-frequency financial data. In: R.J. Adler, R. E. Feldman and M. S. Taqqu (Eds.) *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions*, Birkhäuser, Boston, MA, pp. 55–77.

Zhang, L. (2009). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics.* To appear.

Ordinal- and Continuous-Response Stochastic Volatility Models for Price Changes: An Empirical Comparison

Claudia Czado, Gernot Müller and Thi-Ngoc-Giau Nguyen

Abstract Ordinal stochastic volatility (OSV) models were recently developed and fitted by Müller & Czado (2009) to account for the discreteness of financial price changes, while allowing for stochastic volatility (SV). The model allows for exogenous factors both on the mean and volatility level. A Bayesian approach using Markov Chain Monte Carlo (MCMC) is followed to facilitate estimation in these parameter driven models. In this paper the applicability of the OSV model to financial stocks with different levels of trading activity is investigated and the influence of time between trades, volume, day time and the number of quotes between trades is determined. In a second focus we compare the performance of OSV models and SV models. The analysis shows that the OSV models which account for the discreteness of the price changes perform quite well when applied to such data sets.

1 Introduction

Modeling price changes in financial markets is a challenging task especially when models have to account for salient features such as fat tail distributions and volatility clustering. An additional difficulty is to allow for the discreteness of price changes. These are still present after the US market graduation to decimalization of possible tick sizes. Recently, Müller & Czado (2009) introduced the class of ordinal stochastic volatility (OSV) models, which utilizes the advantages of continuous-response stochastic volatility (SV) models (see Ghysels et al. (1996) and more lately Shephard (2006)) such as fat tails and persistence through autoregressive terms in the volatility process, while adjusting for the discreteness of the price changes.

OSV models are based on a threshold approach, where the hidden continuous process follows a SV model, thus providing a more realistic extension of the ordered

Claudia Czado, Gernot Müller and Thi-Ngoc-Giau Nguyen
Zentrum Mathematik, Technische Universität München, 85747 Garching, Germany,
e-mail: cczado@ma.tum.de, mueller@ma.tum.de, ntngiau2002@yahoo.com

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_16,

probit model suggested by Hausman et al. (1992). In addition we allow for exogenous variables both on the mean and variance level of the hidden process. Parameter estimation in OSV models using maximum likelihood is not feasible, since first the hidden SV process has no closed form of the likelihood and second the threshold approach induces the need to evaluate multidimensional integrals with dimension equal to the length of the financial time series. Therefore Müller & Czado (2009) follow a Bayesian approach. Here Markov Chain Monte Carlo (MCMC) methods allow for sampling from the posterior distributions of model parameters and the hidden process variables.

While Müller & Czado (2009) provided the model specification, developed and implemented the necessary estimation techniques, this paper explores the applicability of the OSV model to financial stocks with different levels of trading activity. In particular, we investigate which exogenous factors such as volume, daytime, time elapsed between trades and the number of quotes between trades have influence on the mean and variance level of the hidden process and thus on the discrete price changes. A second focus of this paper is to compare the performance of the OSV and SV models when these are fitted to such discrete price changes.

Alternative discrete price change models are based on rounding and decomposition ideas. Following the rounding approach Harris (1990) models discrete prices by assuming constant variances of the underlying efficient price, while Hasbrouck (1999a) models efficient prices for bid and ask prices separately using GARCH dynamics for the volatility of the efficient price processes. Hasbrouck (1999a) proposes to use non-Gaussian, non-linear state space estimation of Kitagawa (1987). Other works of Manrique & Shephard (1997), Hasbrouck (1999), Hasbrouck (2003) and Hasbrouck (2004) also use MCMC techniques for estimation.

Decomposition models for discrete price changes assume that the price change is a product of usually three random variables: a price change indicator, the direction of the price change, and the size of the price change. Rydberg & Shephard (2003) and Liesenfeld et al. (2006) follow this approach. Russell & Engle (2005) introduce a joint model of price changes and time elapsed between trades (duration) where price changes follow an autoregressive conditional multinomial (ACM) model and durations the autoregressive conditional duration (ACD) model of Engle & Russell (1998). A common feature of these models is that the time dependence is solely induced by lagged endogenous variables, while our OSV specification allows for parameter driven time dynamics.

The paper is organized as follows: Section 2 introduces the OSV and SV model specifications and summarizes their estimation using MCMC methods. It also considers the problem of model selection among OSV, among SV and between OSV and SV models. The data application to three NYSE stocks with different trading levels from the TAQ data base are given in Section 3. Special emphasis is given to model interpretation and model selection. The paper closes with a summary and outlines further research.

2 Ordinal- and Continuous-Response Stochastic Volatility Models

In this section we recall the OSV and SV model specifications and briefly summarize MCMC techniques which have been developed to estimate these models. Furthermore, we discuss methods of model selection within and between the two model classes.

2.1 OSV and SV Model Specification and Interpretation

As introduced by Müller & Czado (2009) we consider the following stochastic volatility model for an ordinal valued time series $\{Y_{t_i}, i = 1,\dots,I\}$, where t_i, $i = 1,\dots,I$ denote the possibly unequally spaced observation times. In this model the response Y_{t_i} with K possible values is viewed as a censored observation from a hidden continuous variable $Y^*_{t_i}$ which follows a stochastic volatility model, i.e.

$$\begin{aligned} Y_{t_i} = k \;\Leftrightarrow\; & Y^*_{t_i} \in [c_{k-1}, c_k)\,, \\ & Y^*_{t_i} = \mathbf{x}'_{t_i}\beta + \exp(h^*_{t_i}/2)\varepsilon^*_{t_i}, \\ & h^*_{t_i} = \mathbf{z}'_{t_i}\alpha + \phi(h^*_{t_{i-1}} - \mathbf{z}'_{t_{i-1}}\alpha) + \sigma\eta^*_{t_i}, \end{aligned} \tag{1}$$

where $c_0 = -\infty < c_1 < \cdots < c_{K-1} < c_K = +\infty$ are unknown threshold parameters (also called cutpoints). Moreover, $\mathbf{x}_{t_i}$ and $\mathbf{z}_{t_i}$ are p and q dimensional covariate vectors on the hidden mean and log volatility level, respectively. Associated with these covariate vectors are unknown regression parameters β and α, respectively. The parameter ϕ is an unknown autocorrelation parameter and σ^2 an unknown variance parameter on the hidden log volatility scale. The error variables $\varepsilon^*_{t_i}$ and $\eta^*_{t_i}$ are assumed to be i.i.d. standard normal, with independence also between $\{\varepsilon^*_{t_i}, i = 1,\dots,I\}$ and $\{\eta^*_{t_i}, i = 1,\dots,I\}$. For t_0 we assume $\mathbf{z}_0 := (0,\dots,0)'$ and that h^*_0 follows a known distribution. Finally, for identifiability reasons we have to fix a threshold parameter, and hence we set $c_1 = 0$. The model specified by (1) is abbreviated by $OSV(X_1,\dots,X_p;Z_1,\dots,Z_q)$, where $(X_1,\dots,X_p)$ and $(Z_1,\dots,Z_q)$ represent the names of the covariates with corresponding observation vectors $\mathbf{x}_{t_i}$ and $\mathbf{z}_{t_i}$ at time t_i, respectively.

To interpret such a model, denote the mean and variance of the hidden process at t_i by μ_{t_i} and $\sigma^2_{t_i}$, respectively. As μ_{t_i} is increased holding $\sigma^2_{t_i}$ fixed, we see that the probability of a large (small) category is increased (decreased). For fixed μ_{t_i}, we see that if $\sigma^2_{t_i}$ is increased the probability of extreme categories is increased. These two situations are illustrated in Figure 1.

Furthermore, the OSV model allows to quantify the probability $p^k_{t_i} := P(Y_{t_i} = k)$ for observing a specific category k at time t_i. This probability is given by

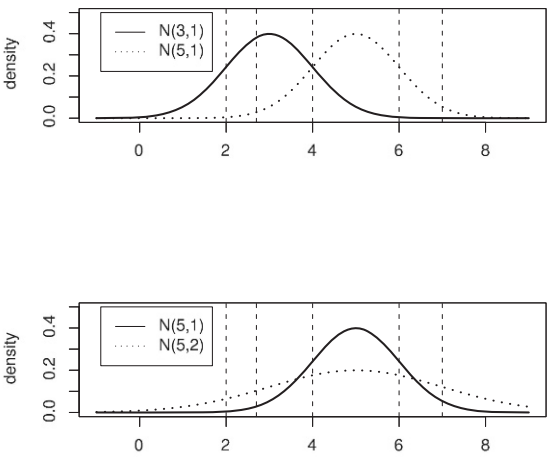


Fig. 1 Category probabilities (visualized as area under the curve between adjacent threshold bounds) as mean and variance of a hidden normally distributed random variable vary

$$p_{t_i}^k = \begin{cases} \Phi\left((c_1 - \mathbf{x}_{t_i}'\beta)/\exp(h_{t_i}^*/2)\right) & \text{for } k = 1, \\ \Phi\left((c_k - \mathbf{x}_{t_i}'\beta)/\exp(h_{t_i}^*/2)\right) - \Phi\left((c_{k-1} - \mathbf{x}_{t_i}'\beta)/\exp(h_{t_i}^*/2)\right) & \\ \qquad \text{for } k = 2, \ldots, K-1, & \\ 1 - \Phi\left((c_{K-1} - \mathbf{x}_{t_i}'\beta)/\exp(h_{t_i}^*/2)\right) & \text{for } k = K, \end{cases}$$

where $\Phi()$ denotes the cumulative distribution function of a standard normal random variable. Therefore the model is able to identify time points where there is a large probability of extreme small or large category labels. Note that no symmetry assumptions about the occurrence of large/small categories are present in the model specification.

We conclude this subsection by presenting the ordinary stochastic volatility model. For a real valued time series $\{Y_{t_i}^c, i = 1, \ldots, I\}$ the ordinary SV model is specified by

$$\begin{aligned} Y_{t_i}^c &= \mathbf{x}_{t_i}'\beta + \exp(h_{t_i}/2)\varepsilon_{t_i} \\ h_{t_i} &= \mathbf{z}_{t_i}'\alpha + \phi(h_{t_{i-1}} - \mathbf{z}_{t_{i-1}}'\alpha) + \sigma\eta_{t_i}, \end{aligned} \tag{2}$$

where $\mathbf{x}_{t_i}, \beta, \mathbf{z}_{t_i}, \alpha, \phi$ and σ^2 are specified as in the OSV model. The error variables ε_{t_i} and η_{t_i} are assumed to be i.i.d. standard normal, with independence also between $\{\varepsilon_{t_i}, i = 1, \ldots, I\}$ and $\{\eta_{t_i}, i = 1, \ldots, I\}$. Analogously to the OSV case, the model specified by (2) is denoted by $SV(X_1, \ldots, X_p; Z_1, \ldots, Z_q)$.

In our application we use $OSV(X_1, \ldots, X_p; Z_1, \ldots, Z_q)$ models for the category labels of the associated price change classes, whereas $SV(X_1, \ldots, X_p; Z_1, \ldots, Z_q)$ models are applied to the observed price changes directly.

2.2 Bayesian Inference for OSV and SV Models

Bayesian inference for the SV models was thoroughly investigated in Chib et al. (2002). They used an estimation procedure based on a state space approximation which we just briefly recall. Obviously, in model (2) one can equivalently write

$$\log\left(Y_{t_i}^c - \mathbf{x}_{t_i}'\beta\right)^2 = h_{t_i} + \log\varepsilon_{t_i}^2.$$

Kim et al. (1998) have shown that the distribution of $\log\varepsilon_{t_i}^2$ can be approximated very well by a seven-component mixture of normals. In particular, one can assume $\log\varepsilon_{t_i}^2 \approx \sum_{k=1}^7 q_k u_{t_i}^{(k)}$ where $u_{t_i}^{(k)}$ is normally distributed with mean m_k and variance v_k^2 independent of t_i. Moreover, the random variables $\{u_{t_i}^{(k)} \mid i = 1,\ldots,I,\ k = 1,\ldots,7\}$ are independent. The quantity q_k denotes the probability that the mixture component k occurs. These probabilities are also independent of t and are given in Table 1 of Chib et al. (2002) together with the corresponding means and variances. Let $s_{t_i} \in \{1,\ldots,7\}$ denote the component of the mixture that occurs at time t_i and let $\pi(s_{t_i})$ denote the prior for s_{t_i}, where $\pi(s_{t_i}=k) = q_k$. Then, by setting $\tilde{Y}_{t_i}^c := \log\left(Y_{t_i}^c - \mathbf{x}_{t_i}'\beta\right)^2$, one arrives at

$$\tilde{Y}_{t_i}^c = h_{t_i} + u_{t_i}^{(s_{t_i})}$$

which, together with the second equation of (2), gives the desired state space representation.

The inference for the OSV models is even more complicated, since a straightforward extension of the algorithm by Chib et al. (2002) shows an unacceptable bad mixing of the chains. Therefore, Müller & Czado (2009) developed a grouped-move multigrid Monte Carlo (GM-MGMC) algorithm which exhibits fast convergence of the produced Markov chains. Since the SV model given by (2) is a submodel of the OSV model, we use the same sampling scheme also for the SV model, of course reduced by the sampling of the cutpoints which do not appear in the SV model, and the variables $Y_{t_i}^*$, $i = 1,\ldots,I$, which are observed in the SV case.

Each iteration of the GM-MGMC sampler consists of three parts. In the first part, the parameter vector β is drawn in a block update from a $(p+1)$-variate normal distribution, the latent variables $Y_{t_i}^*$, $i = 1,\ldots,I$, from truncated univariate normals, and the cutpoints c_k, $k = 2,\ldots,K-1$, from uniform distributions. In the second part, the grouped move step is performed. Here one draws a transformation element γ^2 from a Gamma distribution and updates β, $(Y_{t_1}^*,\ldots,Y_{t_I}^*)$, and $\mathbf{c}$ by multiplication by the element $\gamma = \sqrt{\gamma^2}$. The third part starts with computation of the state space approximation, i.e. by computing $\tilde{Y}_{t_i}^* = \log(Y_{t_i}^* - \mathbf{x}_{t_i}'\beta)^2$ for $i = 1,\ldots,I$. Then s_{t_i}, $i = 1,\ldots,I$, are updated in single updates, and (α,ϕ,σ) by a Metropolis-Hastings step. Finally, the log volatilities $h_{t_1}^*,\ldots,h_{t_I}^*$ are drawn in one block using the simulation smoother of De Jong & Shephard (1995). For more details on the updates we refer to Müller & Czado (2009).

For the Bayesian approach one also has to specify the prior distributions for $\mathbf{c}$, β, h_0^*, α, ϕ, and σ. Assuming prior independence the joint prior density can be written as

$$\pi(\mathbf{c},\beta,h_0^*,\alpha,\phi,\sigma) = \pi(\mathbf{c})\pi(\beta)\pi(h_0^*)\pi(\alpha_1)\cdots\pi(\alpha_q)\pi(\phi)\pi(\sigma).$$

For β a multivariate normal prior distribution is chosen, for h_0^* the Dirac measure at 0, and for the remaining parameters uniform priors. In particular,

$$\begin{aligned} \pi(\mathbf{c}) &= \mathbf{I}_{\{0<c_2<\cdots<c_{K-1}<C\}}, & \pi(\beta) &= N_{p+1}(\beta\,|\,\mathbf{b}_0,B_0),\\ \pi(h_0^*) &= \mathbf{I}_{\{h_0^*=0\}}, & \pi(\alpha_j) &= \mathbf{I}_{(-C_\alpha,C_\alpha)}(\alpha_j), \quad j=1,\ldots,q,\\ \pi(\phi) &= \mathbf{I}_{(-1,1)}(\phi), & \pi(\sigma) &= \mathbf{I}_{(0,C_\sigma)}(\sigma), \end{aligned}$$

where $C>0$, $C_\alpha>0$, and $C_\sigma>0$ are (known) hyperparameters, as well as the mean vector $\mathbf{b}_0$ and the covariance matrix B_0.

2.3 Model Selection

We now look at some criteria for model selection among OSV models, among SV models, and between OSV and SV models.

Model Selection Between OSV Models

We consider a model specification to be reasonable when credible intervals do not contain zero for all parameters. However model selection among such reasonable models is difficult since the likelihood cannot be evaluated simply for OSV models, thus the often used deviance information criteria (DIC) of Spiegelhalter et al. (2002) or score measures discussed in Gneiting & Raftery (2007) cannot be computed directly. Therefore we consider the following simple model selection criteria.

To choose among OSV models we first derive estimates of the ordinal categories for each t_i based on the MCMC iteration values. Note that the hidden volatility for each t_i is updated in each MCMC iteration, but we use only the average value of the log volatility estimates at t_i over all MCMC iterations. These averages are denoted by $\hat{h}_{t_i}^*$ and are used to derive fitted values for the hidden process. Let $\beta^r,\alpha^r,\sigma^r,\phi^r$ and c_k^r, $k=2,\ldots,K-1$ denote the rth MCMC iterate of β,α,σ,ϕ and c_k, $k=2,\ldots,K-1$, respectively for $r=1,\ldots,R$. The estimated log volatilities $\hat{h}_{t_i}^*$ allow to derive fitted hidden process variables $y_{t_i}^{*r}$ defined by

$$y_{t_i}^{*r} := \mathbf{x}_{t_i}'\beta^r + \exp(\hat{h}_{t_i}^*/2)\varepsilon_{t_i}^{*r},$$

where $\varepsilon_{t_i}^{*r}$ are i.i.d. standard normal observations. Finally find category k such that $y_{t_i}^{*r}\in[c_{k-1}^r,c_k^r)$ and set

$$y_{t_i}^r := k.$$

The ordinal category at time t_i is now fitted by the empirical median of $\{y^r_{t_i}, r = 1,\ldots,R\}$, which we denote as $\hat{y}_{t_i}$.

To construct interval estimates for the ordinal categories we define

$$y^{*r}_{t_i,1-\alpha} := \mathbf{x}'_{t_i}\beta^r + \exp(\hat{h}^*_{t_i}/2)z_{1-\alpha},$$
$$y^{*r}_{t_i,\alpha} := \mathbf{x}'_{t_i}\beta^r - \exp(\hat{h}^*_{t_i}/2)z_{\alpha},$$

where z_δ denotes the δ quantile of a standard normally distributed random variable. Then we find categories $k_{1-\alpha}$ such that $y^{*r}_{t_i,1-\alpha} \in [c^r_{k-1}, c^r_k)$ and k_α such that $y^{*r}_{t_i,\alpha} \in [c^r_{k-1}, c^r_k)$, respectively, and set

$$y^r_{t_i,1-\alpha} := k_{1-\alpha} \quad \text{and} \quad y^r_{t_i,\alpha} := k_\alpha.$$

The interval estimate for a category at a time t_i is now defined as the interval $[\hat{y}_{t_i,\alpha}, \hat{y}_{t_i,1-\alpha}]$ where $\hat{y}_{t_i,\alpha}$ and $\hat{y}_{t_i,1-\alpha}$ denote the empirical medians of $\{y^r_{t_i,\alpha}, r = 1,\ldots,R\}$ and $\{y^r_{t_i,1-\alpha}, r = 1,\ldots,R\}$, respectively.

Alternatively we could consider a $100(1-\alpha)\%$ credible interval, which is given by $[\hat{y}^B_{t_i,\alpha}, \hat{y}^B_{t_i,1-\alpha}]$, where $\hat{y}^B_{t_i,\alpha}$ ($\hat{y}^B_{t_i,1-\alpha}$) denotes the empirical α ($1-\alpha$) quantile of $\{y^r_{t_i}, r = 1,\ldots,R\}$. Since the fitted category $y^r_{t_i}$ of the rth MCMC iterate takes on only a few values, the empirical α and $(1-\alpha)$ quantiles are not well defined. Therefore we will not follow this approach.

To choose among several OSV specifications we now count the times the observed category coincides with the fitted category as well as how many times the interval estimate covers the observed category. We choose the model with the highest correctly fitted and covered categories as the best model. Note that the observed coverage percentage is not identical with $100(1-\alpha)$ for the α value used in the construction of the interval estimates, since category values for different time points are dependent.

Model Selection Between SV Models

For the SV models we follow a similar approach as for the OSV models. First let $\hat{h}^c_{t_i}$ denote the average value of the log volatility estimates at time t_i over all MCMC iterations. Again let β^r, α^r, σ^r, and ϕ^r denote the rth MCMC iterate of β, α, σ, and ϕ for $r = 1,\ldots,R$ for the SV model, respectively. Define

$$y^{c,r}_{t_i} := \mathbf{x}'_{t_i}\beta^r + \exp(\hat{h}^c_{t_i}/2)\varepsilon^r_{t_i}$$
$$y^{c,r}_{t_i,1-\alpha} := \mathbf{x}'_{t_i}\beta^r + \exp(\hat{h}^c_{t_i}/2)z_{1-\alpha}$$
$$y^{c,r}_{t_i,\alpha} := \mathbf{x}'_{t_i}\beta^r - \exp(\hat{h}^c_{t_i}/2)z_{\alpha},$$

where $\varepsilon^r_{t_i}$ are i.i.d. standard normal. Now determine the median of $\{y^{c,r}_{t_i}, r = 1,\ldots,R\}$, $\{y^{c,r}_{t_i,1-\alpha}, r = 1,\ldots,R\}$ and $\{y^{c,r}_{t_i,\alpha}, r = 1,\ldots,R\}$, and denote them by $\hat{y}^c_{t_i}, \hat{y}^c_{t_i,1-\alpha}$ and $\hat{y}^c_{t_i,\alpha}$, respectively. Since $\hat{y}^c_{t_i}$ is real-valued, it is not informative to count the times the observed value is equal the fitted value $\hat{y}^c_{t_i}$ for all t_i. Hence, we only count the number of times the observed value is covered by the interval $[\hat{y}^c_{t_i,\alpha}, \hat{y}^c_{t_i,1-\alpha}]$ for all t_i.

Model Selection Between OSV and SV Models

The coverage percentage by the interval estimate for the OSV and SV, respectively, is used as a measure how good the model explains the observed values. A larger percentage gives a better fit.

3 Application

In this section we investigate the applicability of the OSV model to financial stocks with different levels of trading activity, and determine the influence of time between trades, volume, day time and the number of quotes between trades. Moreover, we compare the performance of OSV models and SV models using suitable model selection criteria.

3.1 Data

To investigate the gain of the OSV model over a corresponding SV model for the price changes we selected three stocks traded at the NYSE, reflecting stocks which are traded at a low, medium and high level. We chose the Fremont General Corporation (FMT), the Agilent Technologies (Agilent) and the International Business Machine Cooperation (IBM) from the TAQ data base for a low, medium and high level of trading, respectively. The data was collected between November 1-30, 2000 excluding November 23, 24 (thanksgiving).

Table 1 contains trading characteristics for the three stocks during the investigated time period. The absolute values of extremal price changes increase as trading activity increases (cf. rows 'price diff. between t_{i-1} and t_i'), indicating a higher volatility for more frequently traded stocks. As expected, the median and maximum time between trades decreases as the level of trading increases. For the number of quotes between trades we see a different behavior; while the medium number of quotes remains constant, the maximal number of quotes is the same for FMT and IBM, while it is lower for Agilent. Finally, Agilent has the highest maximum volume per trade among these three stocks.

To illustrate the discreteness of the observed price changes we recorded the number of occurrences of tick changes of size $\leq -3/16, -2/16, -1/16, 0, 1/16, 2/16, \geq 3/16$ together with their percentages in Table 2. For each of the tick change size we associate a category label (necessary for the OSV formulation) also given in Table 2. We see that the observed price changes are quite symmetric around 0 during the investigated time period and that a zero price change is observed most often.

The considered OSV and SV models allow for covariates on the mean and volatility level. To get an idea of possible day time effects we report the corresponding observed median values of price, price change, time between trades, number of quotes

Table 1 Observed characteristics of the FMT, Agilent and IBM stocks between Nov. 1 - 30, 2000

		minimum	median	maximum
FMT	price (dollar)	2 7/16	4 5/16	5 5/16
	price diff. between t_{i-1} and t_i (dollar)	−4/16	0	2/16
	time diff. between t_{i-1} and t_i (seconds)	0	192	4001
	number of quotes between t_{i-1} and t_i	0	1	24
	volume per trade	100	1000	122400
Agilent	price (dollar)	38 1/16	46 3/16	53 15/16
	price diff. between t_{i-1} and t_i (dollar)	−11/16	0	8/16
	time diff. between t_{i-1} and t_i (seconds)	0	11	276
	number of quotes between t_{i-1} und t_i	0	1	14
	volume per trade	100	500	247000
IBM	price (dollar)	91 10/16	99 7/16	104 5/16
	price diff. between t_{i-1} and t_i (dollar)	−13/16	0	14/16
	time diff. between t_{i-1} and t_i (seconds)	0	7	150
	number of quotes between t_{i-1} and t_i	0	1	24
	volume per trade	100	1000	225000

Table 2 Observed price changes together with category label, frequency and relative frequency in percent for the FMT, Agilent and IBM stocks from Nov. 1-30, 2000

	price difference	$\leq -3/16$	$-2/16$	$-1/16$	0	$1/16$	$2/16$	$\geq 3/16$
FMT	category	1	2	3	4	5	6	7
	frequency	3	25	229	755	227	28	0
	rel. freq. (%)	0.2	2.0	18.1	59.6	17.9	2.2	0.0
Agilent	category	1	2	3	4	5	6	7
	frequency	196	939	4662	16599	4747	863	216
	rel. freq. (%)	0.7	3.3	16.5	58.8	16.8	3.1	0.8
IBM	category	1	2	3	4	5	6	7
	frequency	585	3090	10251	22286	11161	2546	613
	rel. freq. (%)	1.2	6.1	20.3	44.1	22.1	5.0	1.2

and volume in Table 3. All stocks show larger (smaller) time intervals between trades during midday (opening and closing times), however the median price change is constant over the day time indicating no effect on the mean level of the hidden process. With regard to the volatility we also recorded the minimal and maximal price changes during trading hours in Table 4. Here we see less changes for different trading hours for the FMT and Agilent stocks compared to the IBM stock. This may indicate a day time effect on the volatility level for IBM stocks, which is detected by a corresponding OSV model specification.

Comparing Table 3 with Table 4 we might identify covariates on the volatility level. For example the median volume value exhibits a similar pattern as the pattern of volatility changes for the FMT and IBM stocks, indicating that volume has some explanatory power for the volatility of the price changes. For Agilent stocks the patterns of volume and volatility of the price changes do not match as well. For the other covariates the identification is less pronounced, so we consider them all as potentially useful covariates and let the statistical models identify them.

Table 3 Observed median number of price, price change, time between trades, number of quotes between trades and volume for different trading hours of the FMT, Agilent and IBM stock between Nov. 1 -30, 2000

	day time	9:30-10	10-11	11-12	12-1	1-2	2-3	3-4
	price (dollar)	4 8/16	4 3/16	4 4/16	4 4/16	4 8/16	4 10/16	4 3/16
	price diff. (dollar)	0	0	0	0	0	0	0
FMT	time diff. (sec.)	89	176	210	256	182.5	208	165
	no. of quotes	1	1	1	2	1	1	1
	volume	1000	1000	1000	1000	1000	1000	1000
	price (dollar)	46 8/16	46 9/16	46 2/16	46 4/16	45 14/16	46 2/16	46 4/16
	price diff. (dollar)	0	0	0	0	0	0	0
Agilent	time diff. (sec.)	7	10	11	12	12	11	10
	no. of quotes	1	1	1	1	1	1	1
	volume	600	600	500	500	500	500	500
	price (dollar)	99 3/16	99 9/16	99 8/16	99 11/16	99 11/16	99 7/16	99 4/16
	price diff. (dollar)	0	0	0	0	0	0	0
IBM	time diff. (sec.)	6	6	7	9	9	7	6
	no. of quotes	1	1	1	1	1	1	1
	volume	1300	1000	800	600	700	800	1000

Table 4 Minimal and maximal price changes for different trading hours of the FMT, Agilent and IBM stock between Nov. 1-30, 2000

	day time	9:30-10	10-11	11-12	12-1	1-2	2-3	3-4
FMT	min. price change	−2/16	−2/16	−2/16	−4/16	−3/16	−2/16	−2/16
	max. price change	2/16	2/16	2/16	2/16	2/16	2/16	2/16
Agilent	min. price change	−6/16	−5/16	−4/16	−11/16	−5/16	−4/16	−8/16
	max. price change	8/16	6/16	5/16	5/16	4/16	7/16	7/16
IBM	min. price change	−13/16	−8/16	−4/16	−9/16	−4/16	−5/16	−8/16
	max. price change	14/16	10/16	5/16	10/16	4/16	6/16	9/16

3.2 OSV Models

As response we choose the category corresponding to the price change at trading time t_i, denoted by y_{t_i}. To model a possibly present dependence between the current price change category and the previous one, we use the lagged price change as a covariate on the mean level and denote it by LAG1 (no other covariates turned out to be significant for the mean level in our analysis). In addition, we allow for an intercept parameter on the mean level. For possible covariates on the volatility level we use volume (V), daytime (D), time elapsed between trades (T) and the number of quotes between trades (Q). For numerical stability we use centered and standardized versions of these variables. For reasons of identifiability, no intercept is included in the term $\mathbf{z}'_{t_i}\alpha$.

For all three stocks we ran a variety of models involving V, D, T and Q as well as quadratic functions of these. In the following we only present models where all covariates are significant, i.e. their individual 80% credible intervals do not contain zero. For all models we ran 20000 MCMC iterations of the GM-MGMC algorithm.

Appropriate burnin values were determined using trace plots. Furthermore, the estimated autocorrelations among the MCMC iterations suggested to take a subsample of every 20th iteration.

Fremont General Cooperation

The left panel of Table 5 presents, for three different OSV model specifications, the estimated posterior medians and means of each parameter together with a 80% credible interval for the subsampled MCMC iterations after burnin. Figure 2 shows estimated posterior densities for all parameters of the $OSV(1, LAG1; V, T)$ model. We see a symmetric behavior of the posteriors for the cutpoint parameters and regression parameters and slightly skewed distributions for σ and ϕ. The posterior density estimates for the remaining two OSV specifications show a similar behavior and are therefore omitted.

Interpreting the results for the OSV specifications, we see from the negative sign of LAG1, that an higher (lower) previous price change category decreases the probability of an higher (lower) current price change category, a fact which can be observed directly from the data, where often a positive price change is followed by a negative one and vice versa. A higher volume, a larger time interval between trades and a larger number of quotes increase the log volatility, thus the probability of observing an extreme positive or negative price change is increased.

It remains to choose among the three OSV specifications. Since the models $OSV(1, LAG1; V, T)$, $OSV(1, LAG1; T, Q)$ are nested within $OSV(1, LAG1; V, T, Q)$, the significance of the parameter estimates established by the credible intervals may lead to a slight preference for the $OSV(1, LAG1; V, T, Q)$ model specification. This is also confirmed, when we calculate the fitted price change categories (see Section 2.3) and compare them to the observed price categories. Moreover, we determine fitted interval bounds for the price change category and check how many times they are covering the observed price change category. The percentage of correctly fitted categories is 59.67%, 59.43% and 59.75% for the models $OSV(1, LAG1; V, T)$, $OSV(1, LAG1; T, Q)$ and $OSV(1, LAG1; V, T, Q)$, respectively. The corresponding values for the percentage of correctly covered categories are 96.45%, 96.37% and 96.53%, respectively. This may lead again to a slight preference for the large model.

Agilent Technologies

For the Agilent stock we found only a single OSV specification with significant parameter estimates, whose summary statistics are given in the left panel of Table 6. It is a different specification as for FMT stocks. The effect of the previous price change category for the Agilent stocks is similar to that one for the FMT stocks, and the autocorrelation of the hidden log volatilities is quite the same. A notable difference is the effect of the number of quotes between trades on the price change

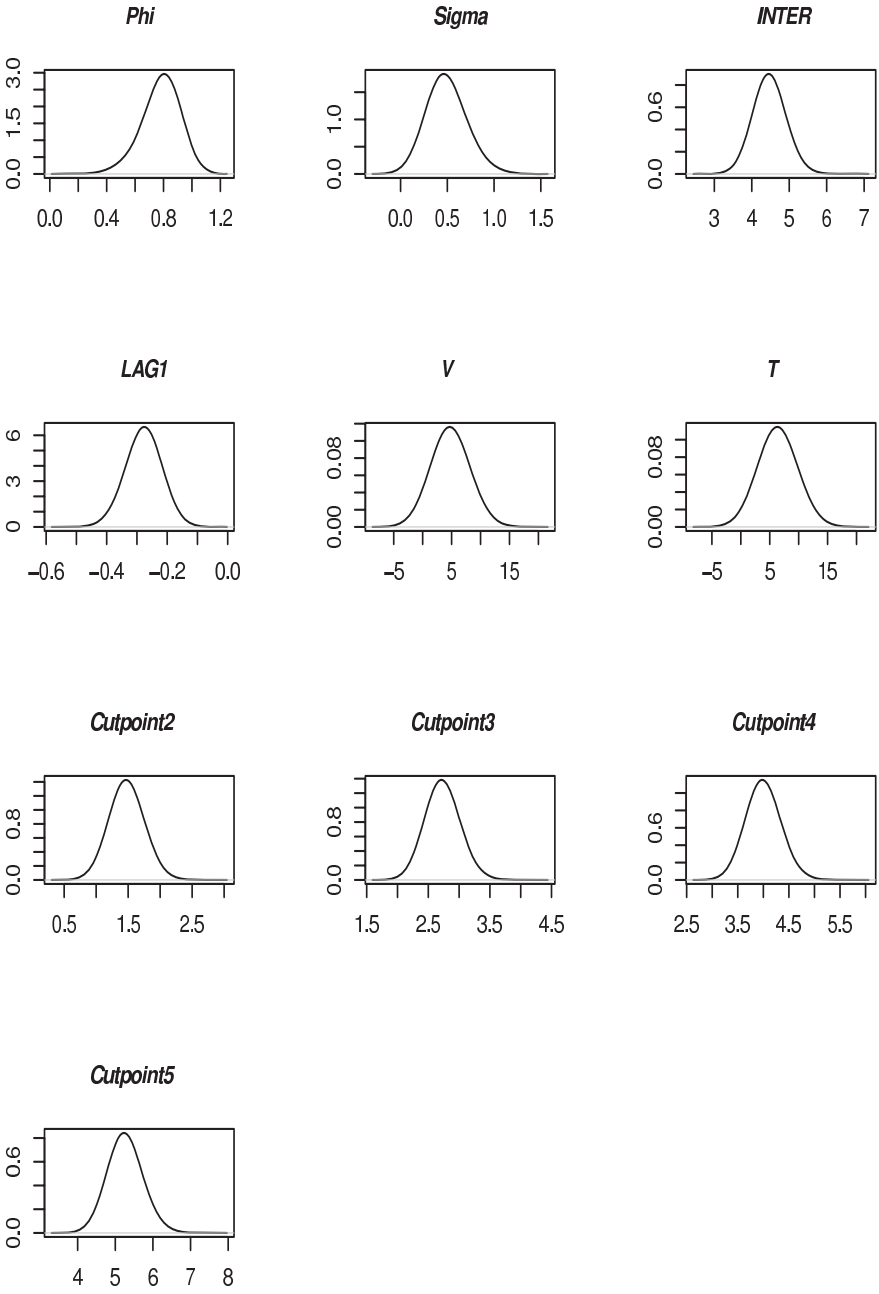


Fig. 2 Estimated posterior density for $OSV(1, LAG1; V, T)$ parameters for FMT stocks

Table 5 Estimated posterior means, medians and quantiles of three OSV (left panel) and three SV (right panel) model specifications with significant parameters fitted for FMT stocks based on the subsampled MCMC iterations after burnin

parameter	10%	90%	median	mean	10%	90%	median	mean
	$OSV(1, LAG1; V, T)$				$SV(1; V, T)$			
ϕ	0.64	0.89	0.80	0.78	0.75	0.79	0.77	0.77
σ	0.32	0.70	0.47	0.49	9.17	11.86	9.90	10.21
c_2	1.25	1.72	1.47	1.48				
c_3	2.50	2.98	2.72	2.73				
c_4	3.73	4.29	3.99	4.00				
c_5	4.88	5.68	5.24	5.26				
1	4.11	4.86	4.46	4.47	$1.1 \cdot 10^{-6}$	$7.1 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$
LAG1	−0.33	−0.23	−0.28	−0.28				
V	1.86	7.83	4.66	4.79	14.14	24.96	19.48	19.51
T	3.45	9.46	6.32	6.39	25.09	36.75	31.09	31.04
	$OSV(1, LAG1; V, Q)$				$SV(1; V, Q)$			
ϕ	0.44	0.85	0.74	0.69	0.75	0.79	0.77	0.77
σ	0.39	0.93	0.58	0.63	9.18	11.82	9.94	10.19
c_2	1.47	2.00	1.72	1.73				
c_3	2.78	3.42	3.07	3.08				
c_4	4.01	4.73	4.34	4.36				
c_5	5.21	6.20	5.65	5.68				
1	4.45	5.38	4.87	4.89	$1.1 \cdot 10^{-6}$	$7.1 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$
LAG1	−0.34	−0.24	−0.30	−0.29				
V	1.49	7.57	4.76	4.64	13.67	25.19	19.28	19.44
Q	1.72	6.84	4.24	4.27	15.44	26.29	21.36	21.10
	$OSV(1, LAG1; V, T, Q)$				$SV(1; V, T, Q)$			
ϕ	0.72	0.91	0.83	0.82	0.76	0.79	0.77	0.77
σ	0.26	0.58	0.40	0.42	9.12	11.71	9.88	10.11
c_2	1.01	1.58	1.24	1.27				
c_3	2.14	2.84	2.42	2.46				
c_4	3.38	4.13	3.71	3.74				
c_5	4.52	5.45	4.94	4.96				
1	3.77	4.65	4.18	4.19	$1.1 \cdot 10^{-6}$	$7.1 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$
LAG1	−0.33	−0.22	−0.27	−0.27				
V	1.87	7.54	4.58	4.59	13.82	24.53	19.71	19.35
T	1.55	6.42	4.02	4.00	15.06	26.38	20.73	20.74
Q	3.63	8.93	6.26	6.30	25.13	36.67	31.00	30.98

categories. Here the parameter estimate has a negative sign, thus the probability of extreme price change categories is decreased when the number of quotes is increased.

Table 6 Estimated posterior means, medians and quantiles of the $OSV(1, LAG1; T, Q)$ (left panel) and $SV(1; T)$ fitted for Agilent stocks based on subsampled MCMC iterations after burnin

parameter	10%	90%	median	mean	10%	90%	median	mean
	$OSV(1, LAG1; T, Q)$				$SV(1; T)$			
ϕ	0.80	0.84	0.82	0.82	0.74	0.80	0.77	0.77
σ	0.46	0.52	0.49	0.49	9.12	10.37	9.67	9.72
c_2	1.28	1.33	1.30	1.30				
c_3	2.24	2.31	2.28	2.28				
c_4	3.42	3.52	3.47	3.47				
c_5	4.39	4.52	4.46	4.46				
c_6	5.36	5.56	5.47	5.47				
1	3.68	3.81	3.74	3.74	$1.1 \cdot 10^{-6}$	$7.1 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$	$4.1 \cdot 10^{-6}$
LAG1	−0.23	−0.21	−0.22	−0.22				
T	22.12	27.43	24.78	24.80	62.57	181.79	121.86	121.89
Q	−10.11	−5.25	−7.64	−7.64				

International Business Machines Cooperation

For the IBM stocks we have two OSV model specifications where all parameter estimates are significant (see the left panel of Table 7). The effect of the number of quotes is similar to that one of the Agilent stock. The full specification also includes a significantly negative daytime parameter, indicating a lower probability of extreme price change categories for later in the day than in the morning. This corresponds to the fact, that often the highest volatility during a day can be observed directly after opening of the exchange. The percentage of correctly fitted response categories is 41.54% for the $OSV(1, LAG1; V, Q)$ model compared to 41.48% for the $OSV(1, LAG1; V, T, Q, D)$ model. The percentage of correctly covered response categories is 92.94% for the for the $OSV(1, LAG1; V, Q)$ model compared to 92.93% for the $OSV(1, LAG1; V, T, Q, D)$ model. Hence, we prefer the simpler one of the two OSV model specifications.

In summary, we see that different OSV models are specified for the different stocks. Whereas there is a negative parameter estimate for the number of quotes between two subsequent trades of the Agilent and the IBM stock, the opposite is true for the less frequently traded FMT stock. Therefore the probability of extreme price changes seems to decrease for more frequently traded stocks when the number of quotes between trades increases, whereas this probability increases for less frequently traded stocks. In addition, the trading frequency influences the magnitude of autocorrelation present in the log volatilities. The highest autocorrelation was observed for the IBM stock. Daytime effects on the hidden volatility are not significant in our three preferred models. The effect of the time elapsed between trades on the log volatility is always positive. This indicates that larger time differences between two subsequent trades usually lead to a higher volatility. The positive regression coefficient for volume induces a larger volatility for larger volumes, which results in higher probabilities for the occurrence of extreme price change categories.

Table 7 Estimated posterior means, medians and quantiles of two OSV (left panel) and one SV (right panel) model specifications with significant parameters fitted for IBM stocks based on recorded MCMC iterations

parameter	10%	90%	median	mean	10%	90%	median	mean
	$OSV(1, LAG1; V, Q)$							
ϕ	0.93	0.94	0.94	0.94				
σ	0.20	0.23	0.21	0.21				
c_2	0.93	0.95	0.94	0.94				
c_3	1.65	1.69	1.67	1.67				
c_4	2.52	2.57	2.54	2.54				
c_5	3.33	3.40	3.37	3.37				
c_6	4.10	4.21	4.16	4.16				
1	3.04	3.12	3.08	3.08				
LAG1	−0.25	−0.24	−0.24	−0.24				
V	5.54	10.25	7.98	7.98				
Q	−9.17	−5.07	−7.12	−7.12				
	$OSV(1, LAG1; V, T, Q, D)$				$SV(1; V, T)$			
ϕ	0.92	0.94	0.93	0.93	0.67	0.69	0.68	0.68
σ	0.21	0.24	0.22	0.23	0.07	0.14	0.09	0.10
c_2	0.90	0.93	0.92	0.91				
c_3	1.66	1.70	1.68	1.68				
c_4	2.51	2.55	2.53	2.53				
c_5	3.34	3.41	3.38	3.38				
c_6	4.14	4.26	4.20	4.20				
1	3.04	3.12	3.08	3.08	$1.9 \cdot 10^{-5}$	$7.5 \cdot 10^{-4}$	$3.7 \cdot 10^{-4}$	$3.7 \cdot 10^{-4}$
LAG1	−0.25	−0.24	−0.24	−0.24				
V	5.22	9.76	7.43	7.46	0.54	9.60	4.91	4.99
T	33.36	38.85	36.08	36.02	23.15	37.59	29.28	29.99
Q	−8.37	−3.90	−6.06	−6.07				
D	−35.08	−22.16	−28.12	−28.26				

3.3 SV Models

For the SV setup we use the observed price changes as response and ignore their discrete nature. For each of the three stocks we investigated different SV specifications. A first difference to the OSV specifications are that none of the covariates LAG1, V, T, Q, and D for the mean level are significant. Therefore all SV models include only an intercept parameter in the mean level, which is significant but very close to zero. For the log volatilities we find significant covariates, which we present in the following. Again we ran 20000 MCMC iterations and determined appropriate burnin values and subsampling rates.

Fremont General Corporation

Three significant SV specifications were found for the FMT stocks and the results are summarized in the right panel of Table 5. The highest coverage percentage is achieved using the $SV(1;V,T,Q)$, which we select as best model among the SV models for the FMT stocks.

Agilent Technologies

For the Agilent stocks only a single SV specification produces significant parameter estimates and the results are presented in the right panel of Table 6. From this we see that only the time elapsed between trades has a significant effect on the price changes. A larger time interval between trades produces a larger volatility, i.e. extreme price changes become more likely.

International Business Machines Cooperation

For the frequently traded IBM stocks only the $SV(1;V,T)$ model produces significant posterior parameter estimates. The results presented in right panel of Table 7 show that both volume and time elapsed between trades increase the volatility, thus making more extreme price changes more likely.

3.4 Comparison Between OSV and SV Models

We now compare all selected OSV and SV models by using the coverage percentages. These are reported in Table 8. We see that there is a clear preference for the OSV specifications for Agilent and IBM stocks, while for the FMT stock a slight preference for the SV specification is visible. A graphical illustration of this is given in Figure 3 where the interval estimates are plotted for the last 100 observations together with the observed values.

As a final comparison we estimate posterior densities of the volatilities for each price change category using the competing OSV and SV specifications for all three stocks. The corresponding plots are shown in Figure 4. The OSV specifications nicely identify different volatility patterns. In particular, extreme price categories correspond to larger volatilities. The competing SV specification for the IBM stocks shows a similar pattern. However, the SV specifications for the FMT and the Agilent stocks lead to quite different density estimates.

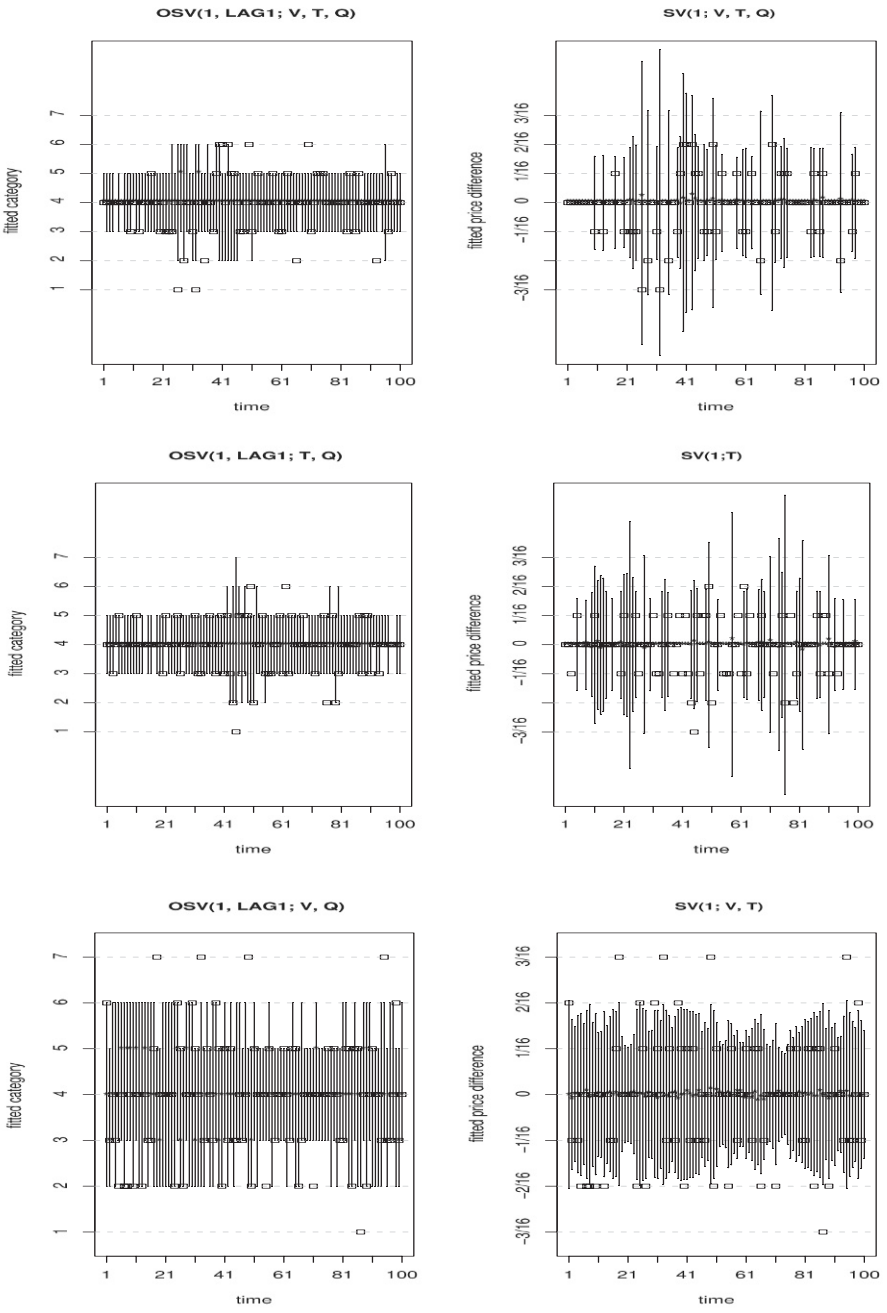


Fig. 3 Fitted categories and fitted price differences of OSV and SV model of the last 100 observations together with interval estimates for FMT (top row), Agilent (middle row) and IBM (bottom row) stocks, respectively

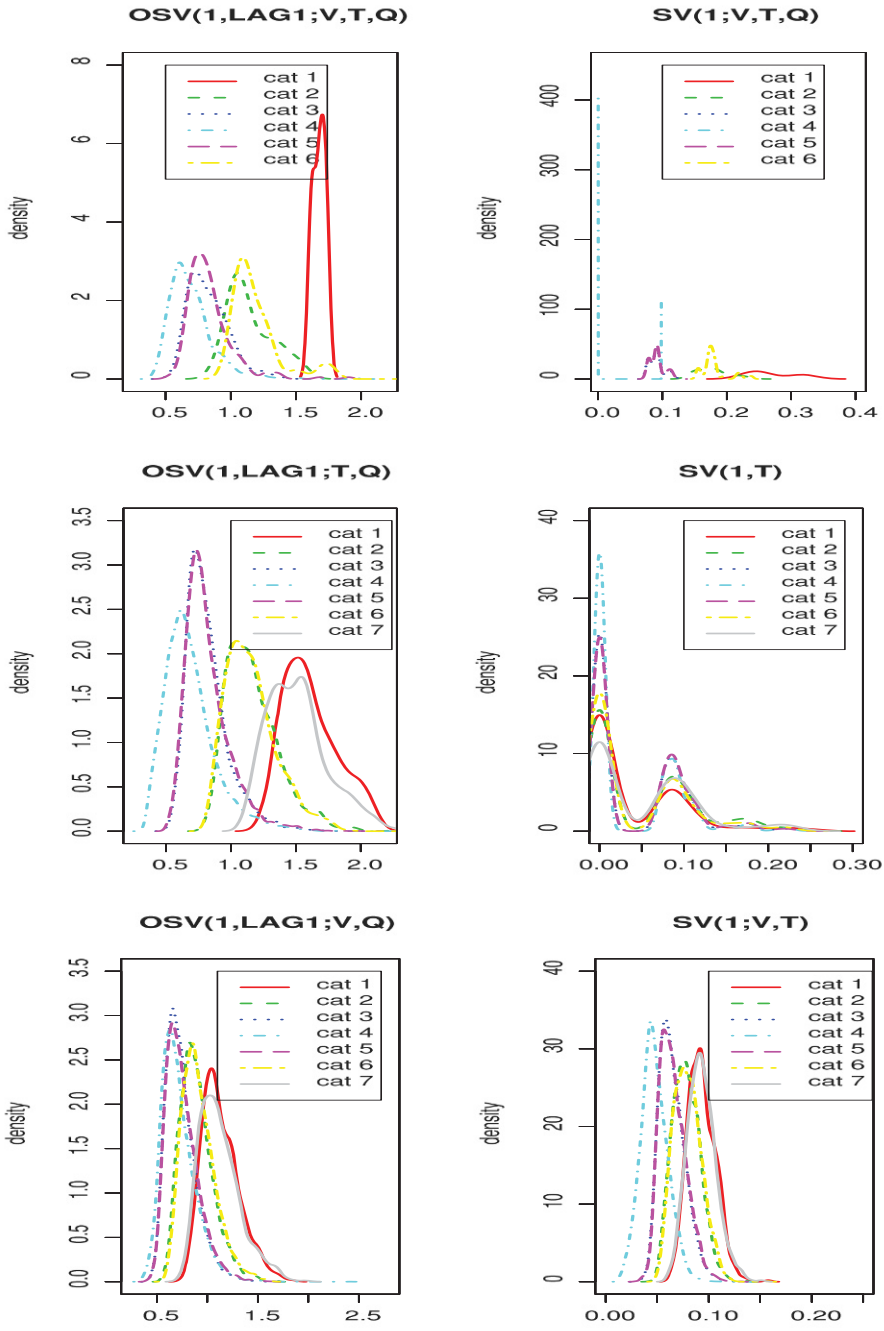


Fig. 4 Estimated posterior densities of the (hidden) volatilities for each category of OSV and SV model for FMT (top row), Agilent (middle row) and IBM (bottom row) stocks, respectively

Table 8 Percentage of correctly covered observations of different OSV and SV specifications for FMT, Agilent and IBM stocks

	OSV specifications		SV specifications	
FMT	$OSV(1, LAG1; V, T, Q)$	1223/1267 = 96.53%	$SV(1, LAG1; V, T, Q)$	1267/1267 = 100.00%
Agilent	$OSV(1, LAG1; T, Q)$	26738/28222 = 94.74%	$SV(1; T)$	20980/28222 = 74.34%
IBM	$OSV(1, LAG1; V, Q)$	46965/50532 = 92.94%	$SV(1; V, T)$	42811/50532 = 84.72%

4 Summary and Discussion

In this paper we presented the results of a Bayesian analysis of two model class specifications for financial price changes. Estimation is facilitated using MCMC methods. The OSV specification explicitly accounts for the discrete values of the price changes, while the SV specification ignores it. The OSV model captures the influence of the previous price change, whereas for the SV models this influence is not significant. In addition we see that volume, time between trades and the number of quotes between trades are important factors determining the volatility. Useful model specifications depend on the trading activity of the stock. In particular, a higher number of quotes between trades increases the volatility for less frequently traded stocks, whereas the opposite pattern is observed for stocks which are more frequently traded. As expected a larger duration between trades increases the volatility. A quadratic day time effect was not significant indicating that there was no strong volatility smile present in the data.

When comparing the OSV and SV models we see that the OSV models perform better (at least for the more frequently traded Agilent and IBM stocks) than the SV models with regard to the coverage proportion of interval estimates. However, more precise model comparison criteria for comparing non nested models with numerical intractable likelihoods in a Bayesian setup are needed and subject to current research. Finally, the OSV and SV model specifications lead to different density estimates for the volatility within the price change classes. However, the density estimates coming from the OSV specifications are quite convincing, since here extreme categories always come along with higher values of the volatility estimates.

Overall we conclude that the OSV models which account for the discreteness of the price changes perform quite well, when applied to data sets as considered in this analysis. Although it is computationally more involved to fit the OSV model to the data, the OSV model is tailored to the structure of ordinal-response data and hence most suitable for price changes.

Acknowledgements Claudia Czado is supported by the *Deutsche Forschungsgemeinschaft*.

References

Chib, S., Nardari, F. & Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models, *Journal of Econometrics* **108**: 281–316.

De Jong, P. & Shephard, N. (1995). The simulation smoother for time series models, *Biometrika* **82**: 339–350.

Engle, R. F. & Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data, *Econometrica* **66**: 1127–1162.

Ghysels, E., Harvey, A. C. & Renault, E. (1996). Stochastic volatility, *in* C. R. Rao & G. S. Maddala (eds), *Statistical Methods in Finance*, North-Holland, Amsterdam, pp. 119–191.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**: 359–378.

Harris, L. (1990). Estimation of stock variances and serial covariances from discrete observations, *Journal of Financial and Quantitative Analysis* **25**: 291–306.

Hasbrouck, J. (1999). Security bid/ask dynamics with discreteness and clustering, *Journal of Financial Markets* **2**: 1–28.

Hasbrouck, J. (1999a). The dynamics of discrete bid and ask quotes, *Journal of Finance* **54**: 2109–2142.

Hasbrouck, J. (2003). Markov chain Monte Carlo methods for Bayesian estimation of microstructure models (computational appendix to: Hasbrouck, J. (2004). Liquidity in the futures pits: inferring market dynamics from incomplete data, *Journal of Financial and Quantitative Analysis* **39**: 305–326), available at http://pages.stern.nyu.edu/∼jhasbrou.

Hasbrouck, J. (2004). Liquidity in the futures pits: inferring market dynamics from incomplete data, *Journal of Financial and Quantitative Analysis* **39**: 305–326.

Hausman, J. A., Lo, A. W. & MacKinlay, A. C. (1992). An ordered probit analysis of transaction stock prices, *Journal of Financial Economics* **31**: 319–379.

Kim, S., Shephard, N. & Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models, *Review of Economic Studies* **65**: 361–393.

Kitagawa, G. (1987). Non-Gaussian state space modeling of nonstationary time series, *Journal of the American Statistical Association* **82**: 1032–1041.

Liesenfeld, R., Nolte, I. & Pohlmeier, W. (2006). Modelling financial transaction price movements: a dynamic integer count data model, *Empirical Economics* **30**: 795–825.

Manrique, A. & Shephard, N. (1997). Likelihood analysis of a discrete bid/ask price model for a common stock quoted on the NYSE, *Nuffield College Economics papers*, W-15.

Müller, G. & Czado, C. (2009). Stochastic volatility models for ordinal valued time series with application to finance, *Statistical Modelling* **9**: 69–95.

Russell, J. R. & Engle, R. F. (2005). A discrete-state continuous-time model of financial transactions prices and times: the autoregressive conditional multinomial-autoregressive conditional duration model, *Journal of Business and Economic Statistics* **23**: 166–180.

Rydberg, T. H. & Shephard, N. (2003). Dynamics of trade-by-trade price movements: decomposition and models, *Journal of Financial Econometrics* **1**: 2–25.

Shephard, N. (2006). Stochastic volatility models, *in* S. N. Durlauf & L. E. Blume (eds), *New Palgrave Dictionary of Economics*, 2nd ed., Palgrave Macmillan. doi: 10.1007/s001090000086.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society B* **64**: 583–639.

Copula Choice with Factor Credit Portfolio Models

Alfred Hamerle and Kilian Plank

Abstract Over the last couple of years we could observe a strong growth of copula based credit portfolio models. So far the major interest has revolved the ability of certain copula families to map specific phenomena such as default clustering or the evolution of prices (e.g., credit derivatives prices). Still few questions have been posed regarding copula selection. This is surprising as the problem of estimating the dependence structure is even unresolved with simple traditional models. For statistical tests of credit portfolio models in general the literature found density-based tests like that of Berkowitz (2001) the most reasonable option. In this text, we examine its power characteristics concerning factor portfolio models in more detail. Our results suggest that both the copula family as well as the level of dependence is generally very difficult to identify.

1 Introduction

In recent years a large number of credit portfolio models have been developed using copulas implicitly or explicitly to model dependence. The major share of publications revolves the question which copula family represents a better explanation of specific empirical phenomena (e.g. Frey & McNeil 2003, Aas 2004). For example, as for credit derivatives pricing a dominant issue was and still is to find a copula which better reproduces market risk premia for default clustering (e.g. Moosbrucker 2006). However, in most of these empirical applications a copula family is not selected by statistical methods but rather "ad hoc", i.e., based on qualitative criteria. Indeed, it

Alfred Hamerle
Lehrstuhl für Statistik, Wirtschaftswissenschaftliche Fakultät, Universität Regensburg, Germany, e-mail: alfred.hamerle@wiwi.uni-regensburg.de

Kilian Plank
Lehrstuhl für Statistik, Wirtschaftswissenschaftliche Fakultät, Universität Regensburg, Germany, e-mail: kilian.plank@wiwi.uni-regensburg.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_17,

appears that selection of a copula is not an easy task, especially due to the fact that the dependence structure is commonly latent. Often many copula families appear to be equally suitable. In a credit portfolio context the problem is even worse if only default rates are available instead of individual default data. Furthermore, credit data sets intrinsically suffer from the rare event issue or simply of variance. Default events as well as losses are more seldom and recorded at lower frequency as, for example, asset values. Thus, the problem of credit portfolio model selection and validation is pending and appears to become even more relevant given the enormous modeling freedom enabled by the copula concept as well as the large set of already available models.

Hamerle & Roesch (2005) and Moosbrucker (2006) inquire the consequences of misspecification of the dependence structure on the estimation and forecasting results but they do not treat the selection of the dependence structure itself. This is no surprise as the whole topic is still in its infancy. Many of the theoretical articles on copula selection and goodness of fit (GoF) tests appeared only recently (e.g. Chen et al. 2004, Fermanian 2005, Genest et al. 2006, Dobric & Schmidt 2007). These are tests on the copula directly. However, our interest in copula selection is indirect as the copula forms merely one part of the overall portfolio model.

For this specific problem only two tests have been suggested so far. First, there is the class of quantile tests inheriting from the market risk literature (Kupiec 1995). These tests compare observed and theoretical proportions of quantile exceedances. Although more information may be extracted via subsampling, as suggested by Lopez & Saidenberg (2000), such binary tests usually require a large number of observations.

The second test is a density test due to Berkowitz (2001). Originally suggested for market risk models, it was soon applied to credit portfolio model validation (Frerichs & Loeffler 2003). This test is clearly superior to quantile tests since it relies on the whole distribution instead of specific distribution quantiles. In any case, none of the extant articles investigates the role of the dependence structure for model selection.

In this article, we compare the power of the Berkowitz test to identify different copulas[1], i.e., we analyze rejection frequencies for a set of copulas using one of them for the data generating process. Within this context the similarity of factor models based on different copulas is examined. In order to apply the test to count data we make use of a specific probability transformation suggested by Hamerle & Plank (2009).

The article is organized as follows. In the next section we shortly summarize five portfolio models admitting a random factor representation of the copula. Subsequently, the Berkowitz test is described. Within this context we shortly outline the modified probability transformation suggested by Hamerle & Plank (2009). Finally, in Section 4, the results of our power tests are presented and discussed.

[1] Note that by "copula" we denote copula family plus parameterization.

2 Factor Models

In this article, we focus on credit portfolio models establishing the dependence structure via factor conditional default indicators. Factor conditional models are very popular in credit risk management since the additional structure admits straightforward interpretation and tractability. In the following subsections we discuss three model types. We start with the most popular one, the Gaussian single risk factor model.

2.1 Gaussian Single Risk Factor Model

Credit portfolio models typically comprise two components: (1) a set of obligors having certain probability of default[2] and (2) a dependence structure establishing default dependencies among obligors. In the Gaussian single risk factor model (also called "Gaussian copula model") a portfolio of n obligors is considered. Default of obligor $i \in I, I = \{1,2,\ldots,n\}$, depends on its "normalized asset return" V_i. V_i is a latent variable comprising two terms

$$V_i = \sqrt{\rho}Y + \sqrt{1-\rho}\varepsilon_i \tag{1}$$

a common (systematic) factor Y and an idiosyncratic factor ε_i. Both factors are standard normal distributed, i.e., $Y, \varepsilon_i \sim \mathcal{N}(0,1)$ and independent. As a result, the correlation between the asset returns of obligors i and j ($i \neq j$) is given by ρ.

Now default of obligor i is modeled as the event that asset return V_i falls short of a threshold c_i. Let D_i denote a default indicator, then

$$D_i = 1 \Leftrightarrow V_i \leq c_i \tag{2}$$

Based on this, the unconditional probability of default (PD) of borrower i is $\lambda_i = \mathbb{P}(D_i = 1) = \Phi(c_i)$ where Φ denotes the standard Gaussian cumulative distribution function. Conversely, given the unconditional PD λ_i (which is usually implied from a public rating) the threshold is given by $c_i = \Phi^{-1}(\lambda_i)$.

2.2 t-Copula Factor Model

The Gaussian copula from the last subsection has no tail dependence. From a modeling point of view this implies that joint tail events have relatively little probability. An alternative is the t-copula which has tail dependence. The above factor model can be easily altered in order to have a t-copula. This is simply accomplished by multiplication of V_i by ν/W, i.e.,

[2] A related component is loss given default but we neglect this aspect here.

$$V_i = \sqrt{\rho}\frac{\nu}{W}Y + \sqrt{1-\rho}\frac{\nu}{W}\varepsilon_i \tag{3}$$

where W is an independent Chi-Squared distributed random variable with ν degrees of freedom.

2.3 Archimedean Copula Factor Models

A general method for the construction of Archimedean copulas (called "frailty construction") is due to Marshall & Olkin (1988). The procedure may be described as follows. Let $\mathbb{P}(D_i = 1) = \lambda_i, i = 1, \ldots, n$, be the unconditional PD of i and Y a positive latent random factor. Then, let the conditional PD be defined as follows

$$\lambda_i(y) = \mathbb{P}(D_i = 1 \mid Y = y) = \lambda_i^y \tag{4}$$

As defaults are assumed to be independent conditional on Y, the joint conditional default probability is simply the product of the marginal conditional PDs

$$\mathbb{P}(D_1 = 1, \cdots, D_n = 1 \mid Y = y) = \prod_{i=1}^{n} \lambda_i^y \tag{5}$$

The unconditional joint probability arises by integration over the factor

$$\mathbb{P}(D_1 = 1, \cdots, D_n = 1) = \int_{-\infty}^{+\infty} \left(\prod_{i=1}^{n} \lambda_i^y \right) dF_Y \tag{6}$$

This in turn may be expressed as Laplace transform (LT) φ_Y^{-1} of the factor[3]

$$\begin{aligned} \mathbb{P}(D_1 = 1, \cdots, D_n = 1) &= \int_{-\infty}^{+\infty} \left(\prod_{i=1}^{n} \lambda_i^y \right) dF_Y \\ &= \int_{-\infty}^{+\infty} exp\left(ln\left(\prod_{i=1}^{n} \lambda_i^y \right) \right) dF_Y \\ &= \int_{-\infty}^{+\infty} exp\left(y \sum_{i=1}^{n} ln\left(\lambda_i\right) \right) dF_Y \\ &= \varphi_Y^{-1}\left(-\sum_{i=1}^{n} ln\left(\lambda_i\right) \right) \end{aligned} \tag{7}$$

Now, we can apply the same LT representation to each margin. To that end, we write the unconditional PD of obligor i as expectation of the conditional PD

[3] We denote the LT with superscript -1 just in order to be in line with the usual notation in the literature.

Table 1 Some Factor Distributions with Proper Generators and Inverses.

Copula	LT	Inverse	Factor Dist.
Clayton	$\varphi(u) = \left(u^{-\delta} - 1\right)$	$\varphi^{-1}(s) = (1+s)^{-1/\delta}$	$Y \sim \Gamma(\frac{1}{\delta})$
Gumbel	$\varphi(u) = \left(-ln(u)\right)^{\delta}$	$\varphi^{-1}(s) = exp\left(-s^{1/\delta}\right)$	$Y \sim \alpha$ stable with $\alpha = 1/\delta$
Frank	$\varphi(u) = -ln\left(\frac{exp(-\delta u)-1}{exp(-\delta)-1}\right)$	$\varphi^{-1}(s) = -\frac{1}{\delta} ln\left(1 - e^{-s}\left(1 - e^{-\delta}\right)\right)$	$Y \sim$ log series on $\mathbb{N}$ with $\alpha = 1 - e^{-\delta}$

$$\mathbb{P}(D_i = 1) = \int_{-\infty}^{+\infty} \lambda_i^y dF_Y \tag{8}$$

Again, this may be expressed as Laplace transform

$$\begin{aligned} \mathbb{P}(D_i = 1) &= \int_{-\infty}^{+\infty} exp\left(ln\left(\lambda_i^y\right)\right) dF_Y \\ &= \varphi_Y^{-1}\left(-ln\left(\lambda_i\right)\right) \end{aligned} \tag{9}$$

The latter expression lends itself for substitution in (7) as a proper inverse of a LT always exists

$$\begin{aligned} \lambda_i = \varphi_Y^{-1}\left(-ln\left(\lambda_i\right)\right) \Leftrightarrow \\ \varphi_Y\left(\lambda_i\right) = -ln\left(\lambda_i\right) \end{aligned} \tag{10}$$

After substitution in (7) we obtain

$$\mathbb{P}(D_1 = 1, \cdots, D_n = 1) = \varphi_Y^{-1}\left(\sum_{i=1}^{n} \varphi_Y\left(\lambda_i\right)\right) \tag{11}$$

and the copula is

$$C(u_1, \ldots, u_n) = \varphi_Y^{-1}\left(\sum_{i=1}^{n} \varphi_Y\left(u_i\right)\right) \tag{12}$$

Note that $\lambda_i(y) \to 0$ for larger values of y as in the previous factor models.

For the Clayton, Gumbel, and Frank copula family the factor distribution and Laplace transform (LT) are given in Table 1. For example, for the Clayton copula $Y \sim \Gamma(1/\delta)$ and the resulting conditional default probability is $\lambda_i(y) = e^{-y\left(\lambda_i^{-\delta}-1\right)}$.

3 The Berkowitz Test

In this section we shortly explain the density test of Berkowitz (2001). Later on we have a closer look at probability transforms of discrete distributions.

3.1 The Test Explained

Berkowitz (2001) originally suggested a test to evaluate density forecasts. The basic principle is to generate pseudo observations by first transforming observed losses $(l_t)_{t=1}^T$ via their probability integral (PIT) $F_L(l_t)$[4]. These pseudo observations are transformed in a second step to standard Gaussian random variables (r.v.), so that under H_0

$$z_t = \Phi^{-1}\left(F_L(l_t)\right) \sim N(0,1) \tag{13}$$

Let $z = (z_1, \ldots, z_T)$ denote the vector of twice transformed default counts. The two moments of the Gaussian distribution can be tested by means of a likelihood ratio (LR) test the statistic of which is given by

$$\mathrm{LR} = 2\left[\ell\left(\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2\right) - \ell\left(\mu = 0, \sigma^2 = 1\right)\right] \sim \chi^2(2) \tag{14}$$

i.e., chi-squared distributed with two degrees of freedom. The log-likelihood function is given by

$$\begin{aligned} \ell(\mu, \sigma^2) &= \ln\left(\prod_{t=1}^{T} \phi\left(\frac{z_t - \mu}{\sigma}\right)\right) \\ &= \ln\left(\prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(z_t-\mu)^2}{\sigma^2}}\right) \end{aligned} \tag{15}$$

The ML estimates of μ and σ^2 are given by

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{T}\sum_{t=1}^{T} z_t \text{ and } \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{T}\sum_{t=1}^{T} (z_t - \hat{\mu}_{ML})^2 \tag{16}$$

Additional tests for normality are possible which allow for the third and fourth moments (see Berkowitz (2001)). Frerichs & Loeffler (2003) tested the approach of Doornik & Hansen (1994) but found no clear improvement as compared with LR.

3.2 Discrete PIT

In this section we shortly dwell on a specific problem which may arise when using the Berkowitz test with discrete data. In that case probability transforms lead to discrete mass concentrations which are obviously not uniform or Gaussian, respectively. The reason for this phenomenon are ties. For instance, let $\tilde{U}$ denote the r.v. describing the pseudo observations $\tilde{u}_t = F_L(l_t)$. Then, there is usually a pronounced mass peak

[4] $F_L(l_t)$ denotes the cdf of L.

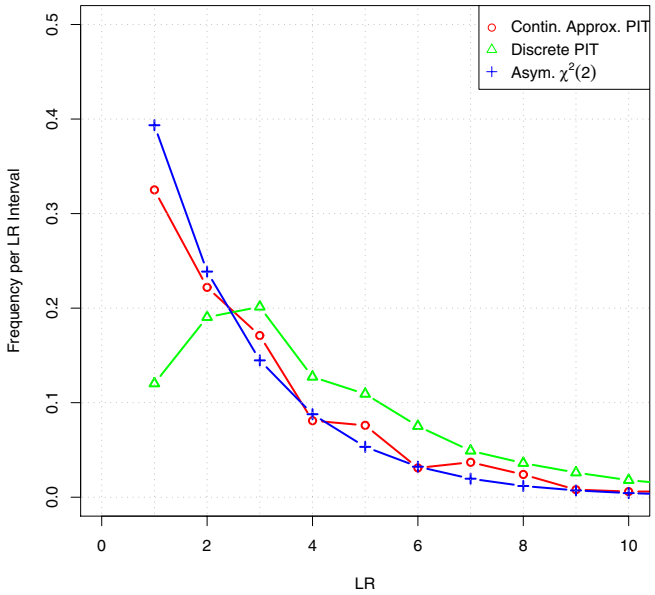


Fig. 1 Simulated LR with continuous PIT modification, with discrete PIT and asymptotic $\chi^2(2)$. Gaussian 1-factor model with $N = 1000$, $T = 5$, $\tau = 0.3$, and $\lambda = 0.02$.

at $F_L(0)$, i.e., $\mathbb{P}\big(\tilde{U} = F_L(0)\big) > 0$ but $\mathbb{P}\big(\tilde{U} < F_L(0)\big) = 0$. Ties are particularly likely when the simulated portfolio is small or when default correlation is high. Without appropriate modifications a PIT does not result in standard uniform samples and $\chi^2(2)$ is a poor approximation for the distribution of the LR statistic. Although this issue has not been mentioned or inquired in the previous literature, it may be substantial, especially when there are singularities of high mass, e.g., large shares of zero defaults.

Formally, let $\tilde{u}_t = F_L(l_t)$ and $\tilde{u}_t = \tilde{u}^{(k)}$, i.e., $\tilde{u}_t$ is identical to the kth element in the ordered sequence of possible realisations $\tilde{u}^{(1)}, \ldots, \tilde{u}^{(K)}$ of $\tilde{U} = F_L(L)$. Then, Hamerle & Plank (2009) suggest to replace the pseudo observations $F_L(l_t)$ by random variables drawn from

$$U\left(\tilde{u}^{(k-1)}, \tilde{u}^{(k)}\right) \tag{17}$$

where $\tilde{u}^{(k-1)} = 0$ for $k = 1$.

As an example, consider Figure 1 which compares theoretical $\chi^2(2)$, the simulated LR distribution based on discrete PIT and the simulated LR distribution based on the modified PIT (17). It is obvious that while the continuous version of the PIT is close to the limiting distribution a simple discrete PIT is usually not. Commonly, as T or n decreases, the discrete PIT deviates more significantly from the limiting distribution.

Table 2 Copula families and their index.

k	Copula Family
1	Gaussian
2	t
3	Clayton
4	Frank
5	Gumbel

4 Simulation Study and Analyses

The major question of this article is whether the Berkowitz test is able to identify the true model with reasonable power for realistic credit default data constellations. If the Berkowitz procedure fails, i.e., if the power is unsatisfactory, the ensuing question is necessarily whether this implies a forecasting problem.

We expect that identification becomes easier when default dependence increases as well as when sample size increases. This is a common result in the literature (e.g. Nikoloulopoulos & Karlis 2008). Furthermore, previous research suggests that copula constellations leading to similar loss distributions do exist. For example, Hamerle & Roesch (2005) found that a Gaussian copula with correlation ρ_1 implies a similar loss distribution as a t copula with appropriate $\rho_2 < \rho_1$. We want to extend these results to different copula factor models in this section.

4.1 Default Count Distributions

To get an impression of the distributions we are working with, Figure 2 depicts the evolution of default count distributions of different copula families for varying levels of Kendall's Tau τ. The latter is our measure of dependence throughout this article. Generally, $-1 \leq \tau \leq +1$ and $\tau = +1$ implies perfect dependence in terms of ranks. The default count distributions in Figure 2 are based on $N = 1000$ obligors, a sample (i.e., time-series) length of $T = 5$, and a homogeneous unconditional default probability of $\lambda = 0.02$.

As expected, we observe that all but one distributions "converge" as $\tau \to 0$ since the underlying copula converges to the product copula. The limiting default distribution is binomial. An exception is the t-copula which does not converge since uncorrelated t-distributed r.v. are not independent (Lemma 3.5 McNeil et al. 2005). Furthermore, as τ increases, we obtain typical extreme shapes with large mass about zero and thick tail. An extraordinary shape can be observed with Frank's family with a second peak shifting to the right and flattening as τ increases.

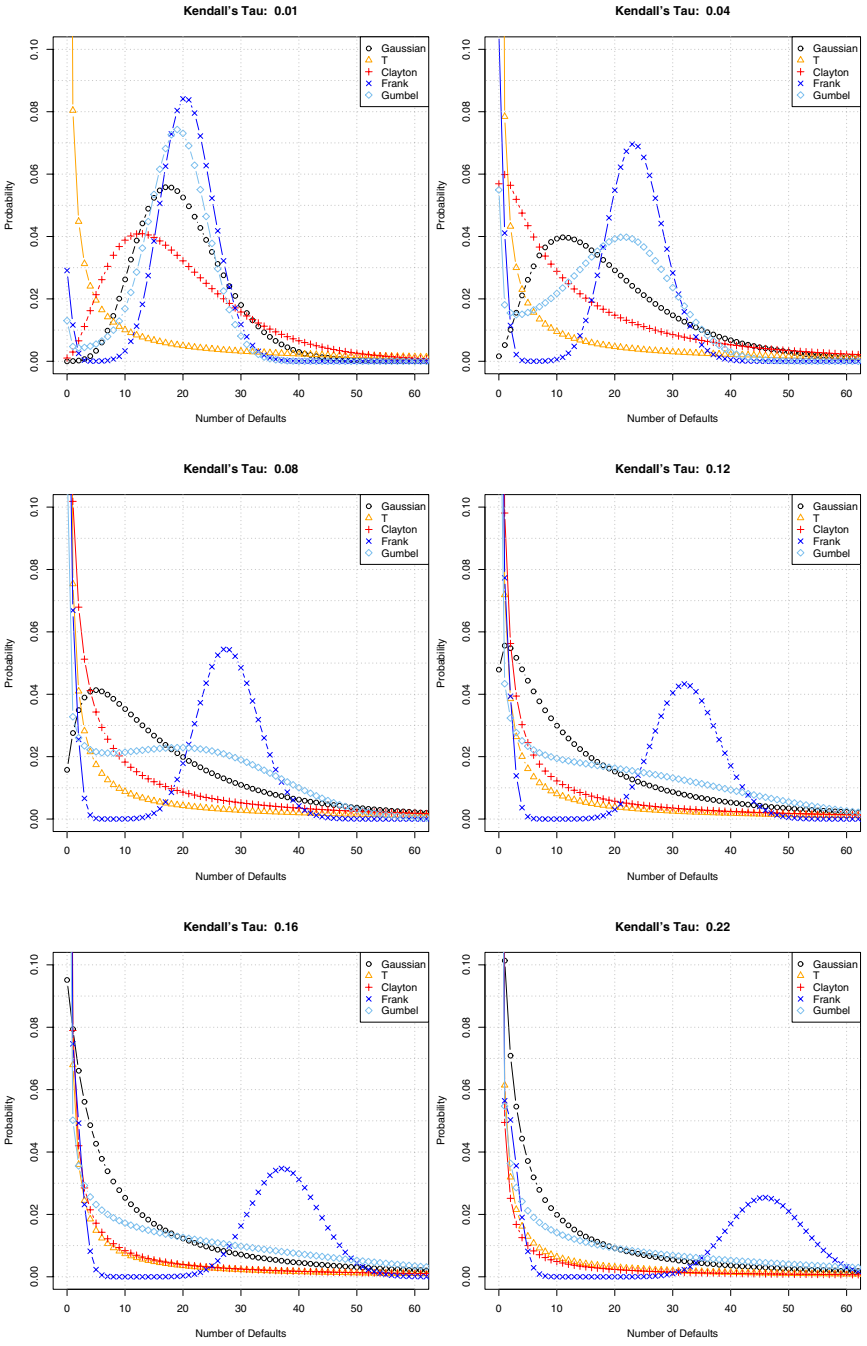


Fig. 2 Default count distributions of a set of copula-based factor models for different levels of Kendall's Tau.

4.2 Power Tests

In this section, we investigate the rejection frequencies of each copula for joint hypotheses on family and level of dependence. Our procedure is as follows.

Let the copula families be indexed by $k \in \{1,2,3,4,5\}$ (Table 2).

First, we calculate for each copula $C_\tau^{(k)}$ the unconditional default distribution as well as the $1-\alpha$ quantile of the corresponding LR statistic. We choose three levels of association, $\tau \in \{0.02, 0.1, 0.22\}$. As there are five copula families these are $3 \times 5 = 15$ simulation points altogether. Note that a copula is identified as a pair (τ, k). Within the context of hypothesis tests we denote by $\tilde{\tau}$ and $\tilde{k}$ true levels and by τ_0 and k_0 hypothesized levels. Now, given initial pairs (τ_0, k_0) and $(\tilde{\tau}, \tilde{k})$ the simulation proceeds as follows

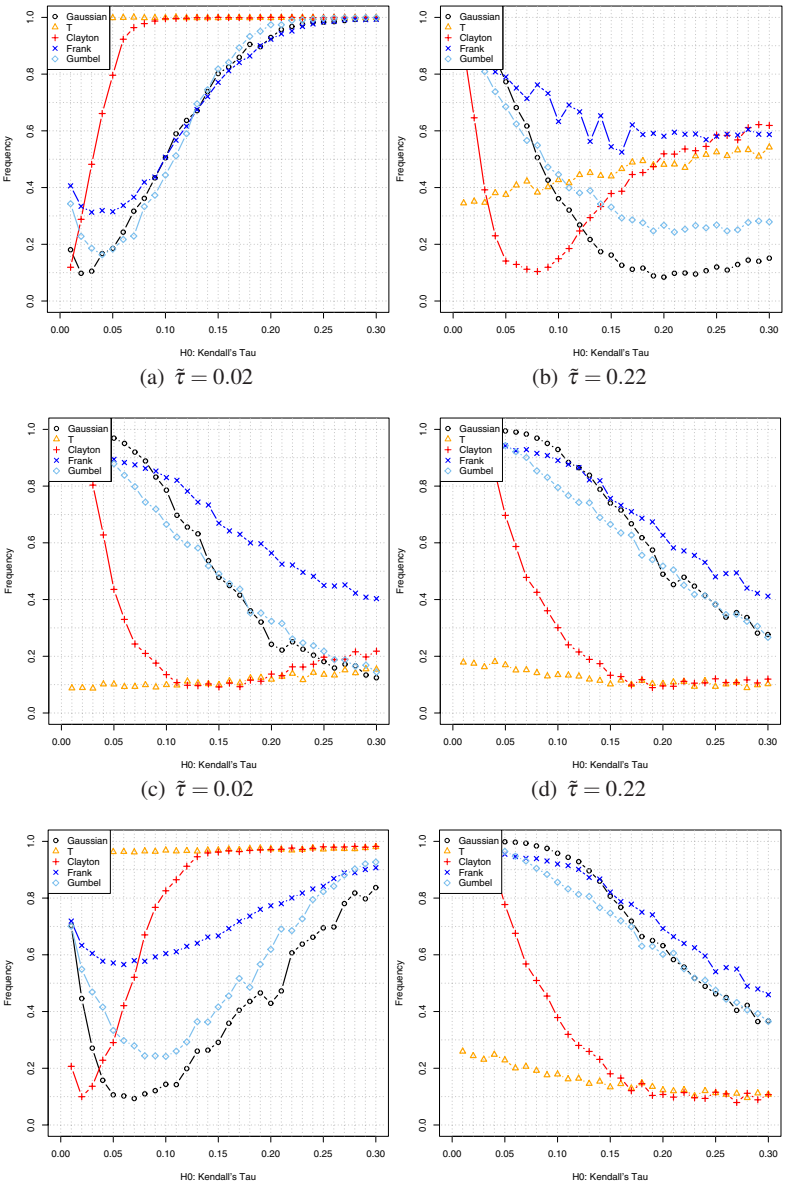
1. Draw a default count sample of length T from the true model $(\tilde{\tau}, \tilde{k})$.
2. Perform a Berkowitz test for $H_0 : (\tau, k) = (\tau_0, k_0)$ on the sample generated in step (1).
3. Go to step (1) m times and calculate the relative frequency of rejection.
4. Choose another hypothesis (τ_0, k_0) and go to step (1) until all hypotheses have been tested.
5. Choose another true pair $(\tilde{\tau}, \tilde{k})$ and go to step (1) until all true pairs have been checked.

We expect the true copula family $\tilde{k}$ to lead to the lowest curve about the true level $\tilde{\tau}$ and to reach the test size there. Formally, this is the point where the hypothesized level of Kendall's Tau, τ_0, is equal to the true level, i.e., $\tau_0 = \tilde{\tau}$ and $k_0 = \tilde{k}$.

Figures 3 and 4 show rejection frequencies for $T = 5$ and $T = 25$, respectively. Each diagram relates to one true level of $\tilde{\tau}$. Furthermore, each row relates to a different true copula family. Note that we do not show rejection frequencies for a given null hypothesis and different true values but for a given true value and different hypotheses.

Let us start with some general observations. First, as specified, the curve of the true family attains its minimum at the true level of Tau. Second, all curves become less peaked as τ increases. This is to be expected because (see Figure 2) our focal unconditional distributions resemble each other more and more as τ increases[5]. Third, all curves become more peaked as T increases. This is also an expected result as longer samples convey more distributional information. As a result, the power of the Berkowitz test grows. Fourth, the t-copula model shows consistently flat power while others, like the Clayton model, is commonly much more peaked. Finally, in many cases, i.e., for many copulas, there are alternative copulas the power of which is very low or even equal to the size of the test. This means that there are often alternative models which cannot be distinguished by the Berkowitz test. Note that this finding holds for both $T = 5$ and $T = 25$.

[5] Note that above we stated that the power generally increases as τ increases. This statement, however, only relates to multivariate tests of the copula directly. In our case, however, we test a sum of dependent random variables.

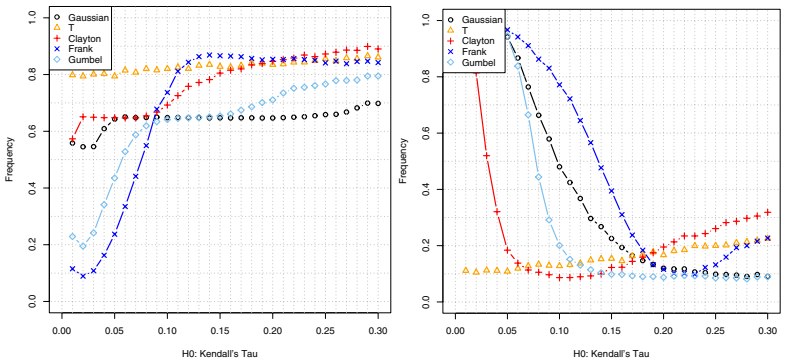


(a) $\tilde{\tau} = 0.02$ (b) $\tilde{\tau} = 0.22$

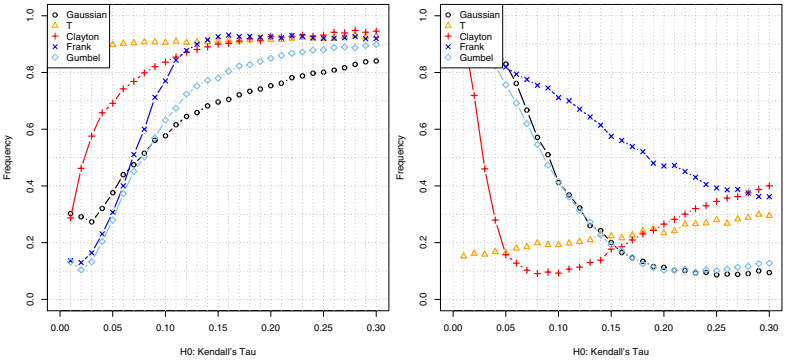
(c) $\tilde{\tau} = 0.02$ (d) $\tilde{\tau} = 0.22$

(e) $\tilde{\tau} = 0.02$ (f) $\tilde{\tau} = 0.22$

Fig. 3 LR rejection frequencies for $T = 5$. Abscissa: τ_0, ordinate: rejection frequency. True family: (a)-(b): **Gaussian**, (c)-(d): **T**, (e)-(f): **Clayton**, (g)-(h): **Frank**, (i)-(j): **Gumbel**.



(g) $\tilde{\tau} = 0.02$ (h) $\tilde{\tau} = 0.22$



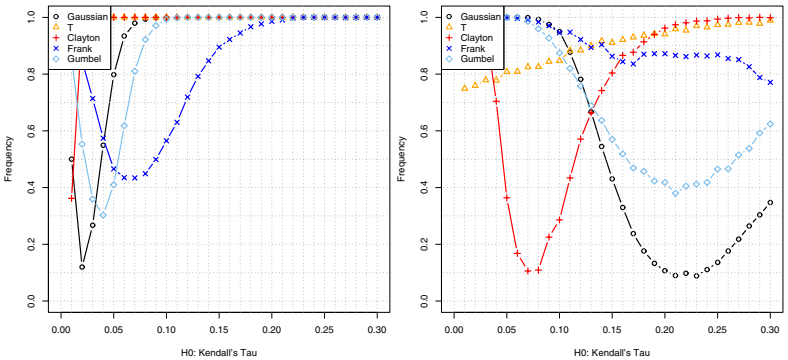
(i) $\tilde{\tau} = 0.02$ (j) $\tilde{\tau} = 0.22$

Fig. 3 LR rejection frequencies for $T = 5$. Abscissa: τ_0, ordinate: rejection frequency. True family: (a)-(b): **Gaussian**, (c)-(d): **T**, (e)-(f): **Clayton**, (g)-(h): **Frank**, (i)-(j): **Gumbel**.

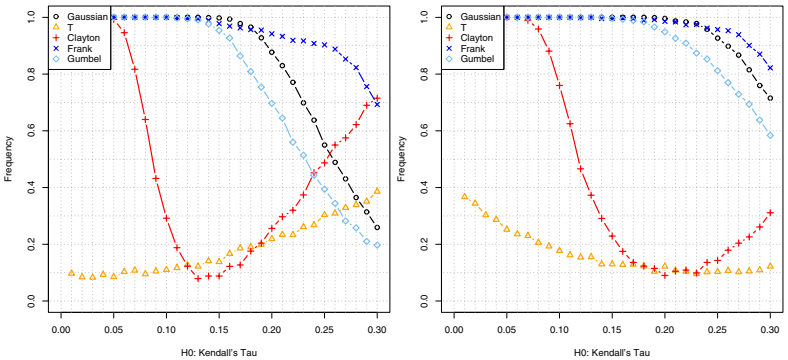
We study some of these aspects in detail now for the case that a Gaussian model is true. Consider Figure 3(a)-(c). As $\tau \to 0$ the rejection curves become more peaked. Conversely, as τ rises it becomes increasingly difficult to reject hypotheses in the neighbourhood of $\tilde{\tau}$. For very low levels of τ (e.g. $\tilde{\tau} = 0.02$) Gaussian, Gumbel, as well as Clayton copula are hardly distinguishable.

The t-copula models does not match this pattern as it does not tend to the product copula as $\tau \to 0$. We take a very low level of degrees of freedom ($\nu = 3$) implying a consistently high share of extreme events. The t-copula model rejection frequency decreases as association increases. For low levels of association the t-model may be rejected with very high probability while for higher levels of association (i.e., high $\tilde{\tau}$) it is difficult to reject a t-model with low τ_0.

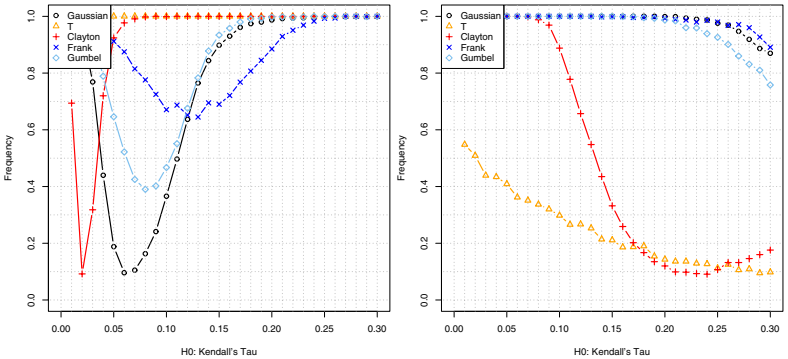
By contrast the Clayton copula based model attains low levels of power in all cases of $\tilde{\tau}$. Curiously enough, the power to reject the Clayton drops down to α for certain $\tau_0 < \tilde{\tau}$, an observation holding for any level of $\tilde{\tau}$. For example, consider Figure 3c. In this case, a model with Clayton copula and $\tau \approx 0.08$ can only be rejected in



(a) $\tilde{\tau} = 0.02$ (b) $\tilde{\tau} = 0.22$

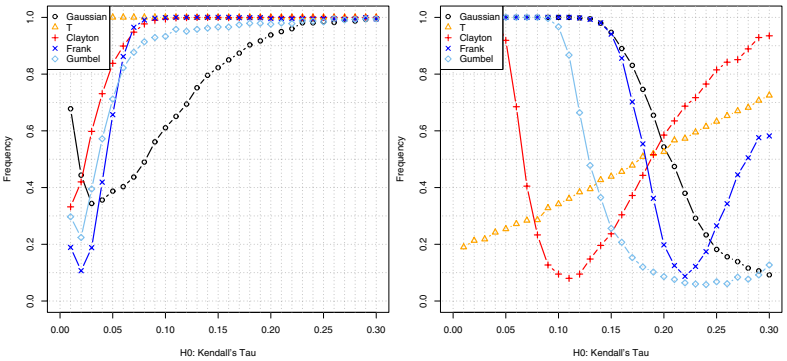


(c) $\tilde{\tau} = 0.02$ (d) $\tilde{\tau} = 0.22$

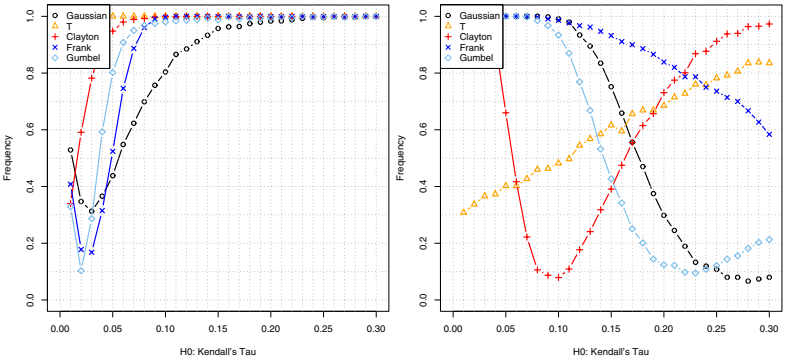


(e) $\tilde{\tau} = 0.02$ (f) $\tilde{\tau} = 0.22$

Fig. 4 LR rejection frequencies for $T = 25$. Abscissa: τ_0, ordinate: rejection frequency. True family: (a)-(b): **Gaussian**, (c)-(d): **T**, (e)-(f): **Clayton**, (g)-(h): **Frank**, (i)-(j): **Gumbel**.



(g) $\tilde{\tau} = 0.02$ (h) $\tilde{\tau} = 0.22$



(i) $\tilde{\tau} = 0.02$ (j) $\tilde{\tau} = 0.22$

Fig. 4 LR rejection frequencies for $T = 25$. Abscissa: τ_0, ordinate: rejection frequency. True family: (a)-(b): **Gaussian**, (c)-(d): **T**, (e)-(f): **Clayton**, (g)-(h): **Frank**, (i)-(j): **Gumbel**.

$\alpha \times 100$ % cases. In other words, the Berkowitz test does not allow to distinguish between H_0 : Gaussian copula, $\tau_0 = 0.22$ and H_0 : Clayton copula, $\tau_0 = 0.08$. This observation carries over to $T = 25$.

An explanation for that is easily found inspecting the default count distributions. Compare for example in Figure 2 the graphs for $\tau = 0.22$ and $\tau = 0.08$ of the Gaussian and Clayton model, respectively. The graphs of the Clayton copula based loss distribution for $\tau = 0.08$ is very similar to that of the Gaussian copula based loss distribution for $\tau = 0.22$. The same holds true for other constellations. As a result, we may state that misspecification in these cases appears to be a minor problem, at least in terms of unconditional prediction.

Most of the above results for the Gaussian copula as the true dependence model carry over to other families. The major observations from the graphs are as follows. First, there is consistently little power to identify the correct t copula. Irrespective of what level of τ is true, various alternative hypotheses have power as high as α. At low levels of $\tilde{\tau}$ Gaussian and Gumbel copulas with high τ_0 cannot be distinguished

statistically. The Clayton family provides indistinguishable alternatives for all $\tilde{\tau}$. Second, Clayton DGPs may be explained by Gaussian models for low $\tilde{\tau}$ and by t models for higher $\tilde{\tau}$. Third, Frank's family is particularly difficult to identify in terms of family membership. Several families attain the test size for higher levels of $\tilde{\tau}$. On the other hand, Frank's family has comparatively good power to find the right copula within the family. Similar results hold for the Gumbel family.

5 Conclusion

In this article we extended preliminary research on copula selection based on the Berkowitz test. We found that unequivocal identification power only exists in special cases. Default data commonly admit different fitting copulas. Clearly, when similarities of default count distributions are strong the Berkowitz test is unable to detect any difference, too. To that end, we showed the evolution of default count distributions and explained our test results with the degree of mutual agreement. Gaussian, Clayton, and Gumbel proved very flexible. These families have low rejection frequencies at some biased level of τ almost irrespective of what copula is true. This confirms and extends results of Hamerle & Roesch (2005) for the relation of Gaussian and t copula.

References

Aas, K. (2004). Modelling the dependence structure of financial assets: a survey of four copulas, *Technical report*, Norwegian Computing Center.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management, *Journal of Business and Economic Statistics* **19**(4): 465–474.

Chen, X., Fan, Y. & Patton, A. (2004). Simple tests for models of dependence between multiple financial time series with applications to u.s. equity returns and exchange rates, *Technical report*, Financial Markets Group, International Asset Management.

Dobric, J. & Schmidt, F. (2007). A goodness of fit test for copulas based on rosenblatt's transformation, *Computational Statistics and Data Analysis* **51**(9): 4633–4642.

Doornik, J. & Hansen, H. (1994). An omnibus test for univariate and multivariate normality, *Technical report*, University of Oxford, University of Copenhagen.

Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas, *Journal of Multivariate Analysis* **95**: 119–152.

Frerichs, H. & Loeffler, G. (2003). Evaluating credit risk models using loss density forecasts, *Journal of Risk* **5**(4): 1–23.

Frey, R. & McNeil, A. (2003). Dependent defaults in models of portfolio credit risk, *Journal of Risk* **6**(1): 59–92.

Genest, C., Quessy, J. & Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation, *Scandinavian Journal of Statistics* **33**: 337–366.

Hamerle, A. & Plank, K. (2009). A note on the Berkowitz test with discrete distributions, *Journal of Risk Model Validation* **3**(2): 3–10.

Hamerle, A. & Roesch, D. (2005). Misspecified copulas in credit risk models: How good is gaussian?, *Journal of Risk* **8**(1): 41–58.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models, *Journal of Derivatives* **3**: 73–84.

Lopez, J. & Saidenberg, M. (2000). Evaluating credit risk models, *Journal of Banking and Finance* **24**: 151–165.

Marshall, A. & Olkin, I. (1988). Families of multivariate distributions, *Journal of the American Statistical Association* **83**: 834–841.

McNeil, A., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management. Concepts, Techniques, Tools*, Princeton University Press.

Moosbrucker, T. (2006). Copulas from infinitely divisible distributions - applications to credit value at risk, *Technical report*, University of Cologne.

Nikoloulopoulos, A. & Karlis, D. (2008). Copula model evaluation based on parametric bootstrap, *Computational Statistics and Data Analysis* p. in press.

Penalized Estimation for Integer Autoregressive Models

Konstantinos Fokianos

Abstract The integer autoregressive model of order p can be employed for the analysis of discrete–valued time series data. It can be shown, under some conditions, that its correlation structure is identical to that of the usual autoregressive process. The model is usually fitted by the method of least squares. However, consider an alternative estimation scheme, which is based on minimizing the least squares criterion subject to some constraints on the parameters of interest. The ridge type of constraints are used in this article and it is shown that under some reasonable conditions on the penalty parameter, the resulting estimates have less mean square error than that of the ordinary least squares. A real data set and some limited simulations support further the results.

1 Introduction

Ludwig Fahrmeir, whom this volume honors, has made seminal contributions to the statistical analysis of integer valued time series by promoting the idea of generalized linear models for inference. In particular, I would like to mention the articles Fahrmeir & Kaufmann (1985) and Fahrmeir & Kaufmann (1987) and the text Fahrmeir & Tutz (2001) which deal respectively with the following;

- the development of maximum likelihood estimation for the regression parameters of a generalized linear model with independent data for both canonical and non-canonical link functions,
- the extension of these results to categorical time series,
- the presentation of the above in a coherent piece of work.

Konstantinos Fokianos
Department of Mathematics & Statistics, University of Cyprus, PO BOX 20537, Nicosia, Cyprus,
URL: http://www.ucy.ac.cy/goto/mathstatistics/en-US/HOME.aspx,
e-mail: fokianos@ucy.ac.cy

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_18,

The results of these references have influenced considerably my research on time series, see e.g. Kedem & Fokianos (2002). On a more personal level, I wish to express my gratitude to Ludwig Fahrmeir for inviting me to Munich on a number of occasions and for giving me the opportunity to discuss with him several issues of mutual interest and to gain important insight.

Integer valued time series occur in diverse applications and therefore statistical methodology should be developed to take into account the discrete nature of the data. In this work, attention is focused on the so called integer autoregressive models of order p–denoted by INAR(p). These processes provide a class of models whose second order structure is identical to that of the standard AR(p) models and estimation can be carried out by standard least squares techniques.

The question of interest in this manuscript is whether the least squares estimators can become more efficient and under what conditions. It is shown that increase in efficiency can be achieved by introducing the so–called penalized least squares criterion (7) for estimation. In particular, it is shown that there are two cases that need to be considered. The first is when the true parameter vector that generates the process assumes "large" values componentwise; then minimization of (7) does not offer any improvement over the standard least squares estimators. On the other hand, when the true parameter vector values are assumed to be "small" , then it is possible to gain in efficiency. Here, the term efficiency, refers to mean square error improvement, since it is well known that penalized estimators are usually biased. The same phenomenon occurs in linear models theory, namely the method of ridge regression. It is well known that the mean square error of ridge estimators is less than the mean square error of the ordinary least squares estimators for some values of the ridge parameter. It is conjectured that the results carry over to the dependent data case under some reasonable assumptions. Some research advocating the use of shrinkage estimators in time series can be found in the recent article by Taniguchi & Hirukawa (2005).

When using penalized criteria for inference, there is an extra complexity introduced, that is the choice of the penalty parameter–see (7). It is a common practice to use cross–validation methods but their performance is questionable, especially in the time series context. Therefore, it is proposed to estimate the regularization parameter by using the AIC. Real data show-see Section 4–that a unique minimizer exists but the method requires more research.

The paper starts with Section 2 where INAR(p) processes are briefly reviewed and the least squares approach to the problem of estimation is discussed. The asymptotic distribution of least squares estimators is also stated. Section 3 introduces the penalized least squares estimator and discuss their asymptotic properties, see Theorems 2 and 3, which constitute the main results. Section 4 complements the presentation by some simulated and real data examples. The article concludes with some comments and an appendix.

2 Integer Autoregressive Processes and Inference

This section reviews briefly some probabilistic properties of the integer autoregressive processes and discuss estimation of unknown parameter by conditional least squares inference.

2.1 Integer Autoregressive Processes

Integer autoregressive processes have been introduced by Al-Osh & Alzaid (1987) and Alzaid & Al-Osh (1990) as a convenient way to transfer the usual autoregressive structure to discrete valued time series. The main concept towards this unification is given by the notion of thinning which is defined by the following:

Definition 1. Suppose that X is a non–negative integer random variable and let $\alpha \in [0,1]$. Then, the thinning operator, denoted by $\circ$, is defined as

$$\alpha \circ X = \begin{cases} \sum_{i=1}^{X} Y_i, & \text{if } X > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\{Y_i\}$ is a sequence of independent and identically distributed Bernoulli random variables–independent of X–with success probability α. The sequence $\{Y_i\}$ is termed as a counting series.

Definition 1 allows for specification of the integer autoregressive process of order p. More specifically, suppose that for $i = 1,2,\ldots,p$, $\alpha_i \in [0,1)$ and let $\{\varepsilon_t\}$ be a sequence of independent and identically distributed nonnegative integer valued random variables with $\mathrm{E}[\varepsilon_t] = \mu$ and $\mathrm{Var}[\varepsilon_t] = \sigma^2$. The following process

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \varepsilon_t, \tag{1}$$

is called integer autoregressive process of order p and is denoted by INAR(p). It should be noted that the Bernoulli variables used for defining the random variable $\alpha_1 \circ X_{t-1}$ are independent of those involved in the definition of $\alpha_2 \circ X_{t-2}$, and so on. This assumption guarantees that the INAR(p) process has the classical AR(p) correlation structure, see Du & Li (1991). A unique stationary and ergodic solution of (1) exists if

$$\sum_{j=1}^{p} \alpha_j < 1. \tag{2}$$

Various other authors have studied the above model, including Al-Osh & Alzaid (1987), Alzaid & Al-Osh (1990), McKenzie (1985), McKenzie (1986) and McKenzie (1988). Some very recent work extending the model in different directions can be found in the papers by Ferland et al. (2006), Neal & Subba Rao (2007), Zheng et al. (2006) and Zhu & Joe (2006).

2.2 Conditional Least Squares Inference

In what follows consider the INAR(p) model defined by (1). The $(p+1)$–parameter vector $\beta = (\mu, \alpha_1, \ldots, \alpha_p)'$ belongs to the $[0,\infty] \times [0,1)^p$ and it is usually estimated by conditional least squares method. Suppose that $\mathscr{F}_t$ is the σ–field generated by the past information, say $X_1, X_2, \ldots, X_t$. The conditional least squares estimator of β is calculated ny minimizing the following sum of squares:

$$S(\beta) = \sum_{t=p+1}^{N} (X_t - E(X_t|\mathscr{F}_{t-1}))^2 = \sum_{t=p+1}^{N} (X_t - \mu - \sum_{i=1}^{p} \alpha_i X_{t-i})^2. \tag{3}$$

Denote by $\hat{\beta}$ the value that minimizes the above expression and notice that standard arguments show that (see Brockwell & Davis (1991), for example)

$$\hat{\beta} = Q^{-1} r \tag{4}$$

where the $(p+1) \times (p+1)$ matrix Q is equal to

$$Q = \begin{bmatrix} N-p & \sum_{t=p+1}^{N} X_{t-1} & \cdots & \sum_{t=p+1}^{N} X_{t-p} \\ \sum_{t=p+1}^{N} X_{t-1} & \sum_{t=p+1}^{N} X_{t-1}^2 & \cdots & \sum_{t=p+1}^{N} X_{t-p} X_{t-1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{t=p+1}^{N} X_{t-p} & \sum_{t=p+1}^{N} X_{t-1} X_{t-p} & \cdots & \sum_{t=p+1}^{N} X_{t-p}^2 \end{bmatrix},$$

and the $(p+1)$–dimensional vector r is defined by

$$r = \left(\sum_{t=p+1}^{N} X_t, \sum_{t=p+1}^{N} X_t X_{t-1}, \ldots, \sum_{t=p+1}^{N} X_t X_{t-p} \right)'.$$

Then the following theorem holds true for the estimator $\hat{\beta}$:

Theorem 1. (Du & Li 1991) Suppose that $\hat{\beta}$ is the conditional least squares estimator defined by means of minimizing (3) for the INAR(p) model (1). In addition, assume that the error process has $\mathrm{E}[\varepsilon_t] = \mu$, $\mathrm{Var}[\varepsilon_t] = \sigma^2$ and $\mathrm{E}[\varepsilon_t^3] < \infty$. Suppose that condition (2) is satisfied and let μ_x to denote the mean of the stationary distribution of the INAR(p) model (1). Then

$$\sqrt{N}\left(\hat{\beta} - \beta\right) \to N_{p+1}(0, V^{-1} W V^{-1}),$$

where the $(p+1) \times (p+1)$ matrix $V = [v_{ij}]$ is defined by

$$v_{ij} = \begin{cases} 1, & i = 1, j = 1, \\ \mu_x, & i = 1, j > 1 \text{ or } i > 1, j = 1, \\ \mathrm{E}[X_{p+1-i} X_{p+1-j}] + \mu_x^2, & i, j \geq 2. \end{cases}$$

Furthermore, the $(p+1) \times (p+1)$ matrix $W = [w_{ij}]$ is given by

$$w_{ij} = \begin{cases} \mathrm{E}[(X_{p+1} - \sum_{k=1}^{p} \alpha_k X_{k-p} - \mu)^2], & i = 1, j = 1, \\ \mathrm{E}[X_{p+1-i}(X_{p+1} - \sum_{k=1}^{p} \alpha_k X_{k-p} - \mu)^2] & j = 1, i > 1, \\ \mathrm{E}[X_{p+1-j}(X_{p+1} - \sum_{k=1}^{p} \alpha_k X_{k-p} - \mu)^2] & i = 1, j > 1, \\ \mathrm{E}[X_{p+1-i}X_{p+1-j}(X_{p+1} - \sum_{k=1}^{p} \alpha_k X_{k-p} - \mu)^2], & i, j \geq 2, \end{cases}$$

where expectation is taken with respect to the stationary distribution.

In addition, it can be shown that the estimator $\hat{\beta}$ is strongly consistent. Theorem 1 is proved by standard arguments from martingale theory, see Klimko & Nelson (1978) and Hall & Heyde (1980), for more. A consistent estimator of the matrix V is given by

$$\hat{V} = \frac{1}{N} Q. \tag{5}$$

Indeed, $\lim_{N\to\infty} \hat{V} = V$, in probability, because of the ergodicity of the process. Similarly, the matrix W is estimated by means of

$$\hat{W} = \frac{1}{N} \sum_{t=p+1}^{N} \left((X_t - \sum_{i=1}^{p} \alpha_i X_{t-i} - \mu)^2 \begin{bmatrix} 1 & X_{t-1} & \dots & X_{t-p} \\ X_{t-1} & X_{t-1}^2 & \dots & X_{t-p}X_{t-1} \\ \dots & \dots & \dots & \dots \\ X_{t-p} & X_{t-1}X_{t-p} & \dots & X_{t-p}^2 \end{bmatrix} \right). \tag{6}$$

Therefore, a consistent estimator of the asymptotic covariance matrix of $\hat{\beta}$ is given by $\hat{V}^{-1}\hat{W}\hat{V}^{-1}$–see theorem 1.

3 Penalized Conditional Least Squares Inference

We suggest estimation of the unknown parameter vector β of the INAR(p) by penalizing the conditional least square criterion with a quadratic penalty. As it is the case with the ridge regression, see Hoerl & Kennard (1970a), Hoerl & Kennard (1970b), it is anticipated that the mean square error of the estimates is minimized by some value of the ridge parameter. Therefore, the choice of the ridge (or regularization) parameter is important and its selection is taken up in Section 4.2 where a proposal is made by using the so–called AIC criterion; Akaike (1974). In the following, the first issue is to show how ridge inference proceeds and then apply the resulting estimators to the problem of prediction.

Ridge coefficients are defined by minimization of the following penalized sum of squares

$$\begin{aligned} S_p(\beta) &= S(\beta) + \lambda_N \sum_{j=1}^{p} \alpha_j^2 \\ &= \sum_{t=p+1}^{N} (X_t - \sum_{j=1}^{p} \alpha_j X_{t-j} - \mu)^2 + \lambda_N \sum_{j=1}^{p} \alpha_j^2 \end{aligned} \tag{7}$$

where $\lambda_N \geq 0$, is the so called regularization parameter. When $\lambda_N = 0$, the ordinary CLS estimator is obtained while if $\lambda_N \to \infty$ then all the coefficients shrink towards zero. An alternative way of obtaining the above penalized sum of squares is to postulate the constraint $\sum_{j=1}^{p} \alpha_j^2 \leq t$. Obviously the parameter t is inversely related with λ_N but both constraints are equivalent.

The penalized CLS estimator of $\hat{\beta}$ will be denoted by $\hat{\beta}^{\lambda}$ and it is easily obtained by

$$\hat{\beta}^{\lambda} = (Q + \lambda_N D_{p+1})^{-1} r. \tag{8}$$

The matrix Q and the vector r have been defined immediately after (4) and the $(p+1) \times (p+1)$ matrix D_{p+1} is given by

$$D_{p+1} = \begin{bmatrix} 0 & 0 \\ 0 & I_p \end{bmatrix},$$

where I_p is the diagonal matrix of order p. It is recognized that the penalized CLS estimator is of the same form as the ordinary ridge regression estimator. It is expected therefore that for a suitably chosen value of the regularization parameter, the mean square error of $\hat{\beta}^{\lambda}$ will be less or equal than that of $\hat{\beta}$. In what follows, it is shown that when the true parameter values are small, then a more efficient estimator–in the sense of mean square error–is obtained by means of minimizing (7) provided that the regularization parameter λ_N is of order N.

We study the asymptotic properties of $\hat{\beta}^{\lambda}$ in the following theorem whose proof is postponed in the appendix.

Theorem 2. Assume the same conditions as in Theorem 1. Assume further that λ_N is such that $\lambda_N/\sqrt{N} \to \lambda_0 \geq 0$. Then

$$\sqrt{N}(\hat{\beta}^{\lambda} - \beta) \to N_{p+1}(-\lambda_0 V^{-1} b, V^{-1} W V^{-1})$$

in distribution, as $N \to \infty$. The matrices V and W have been defined in Theorem 1 and the $(p+1)$–dimensional vector b is given by $b = (0, \alpha_1, \ldots, \alpha_p)'$.

The above theorem shows that when $N \to \infty$, then the penalized CLS (8) are asymptotically normal but biased while their asymptotic covariance matrix is given by the same formula that corresponds to the ordinary CLS estimators–see Theorem 1. Hence, there seems to be no particular improvement when using the penalized CLS estimator unless $\lambda_N = o(\sqrt{N})$, and this is in agreement with the asymptotic results for least squares regression with independent data obtained by Knight & Fu (2000, Th. 2). Theorem 2 implies that when the true parameter values are large and $\lambda_0 > 0$, then the bias of the restricted estimators might be of considerable magnitude.

Suppose now that the data are generated by the INAR(p) process (1) where the vector of unknown parameters satisfies

$$\beta_N = \beta + \frac{c}{\sqrt{N}},$$

for some vector of the form $c = (0, c_1, \ldots, c_p)'$, such that condition (2) is satisfied. Then the second part of the following theorem shows that for small α_j's there is a gain when using the ridge regression. The proof of the theorem is along the lines of Theorem 2 and therefore it is omitted.

Theorem 3. Assume the same conditions as in Theorem 1. Assume further that $\beta_N = \beta + c/\sqrt{N}$ where c is of the form $c = (0, c_1, \ldots, c_p)'$ such that condition (2) holds true. Let $\hat{\beta}^{\lambda}$ be the penalized CLS (8). Then

1. If $\lambda_N/\sqrt{N} \to \lambda_0 \geq 0$, then

$$\sqrt{N}(\hat{\beta}^{\lambda} - \beta) \to N_{p+1}(-\lambda_0 V^{-1} b, V^{-1} W V^{-1}),$$

   in distribution, as $N \to \infty$.
2. If $\alpha_i = 0$ for $i = 1, 2, \ldots, p$ so that $\beta = (\mu, 0, \ldots, 0)'$ and $\lambda_N/N \to \lambda_0 \geq 0$, then

$$\sqrt{N}(\hat{\beta}^{\lambda} - c/\sqrt{N}) \to N_{p+1}(-\lambda_0 \tilde{V}^{-1} c, \tilde{V}^{-1} W \tilde{V}^{-1}),$$

   in distribution, as $N \to \infty$, where $\tilde{V} = V + \lambda_0 D_{p+1}$.

The above notation is the same as that of Theorem 2.

The second part of the above theorem shows that for large sample sizes, the asymptotic distribution of $\hat{\beta}^{\lambda}$ is a multivariate normal provided that the choice of λ_N is of order N and the true parameter value is relatively small. In particular, certain choices of λ_N yield to consistent estimators which are asymptotically normally distributed. However, other choices of λ_N yield to biased estimators. We anticipate though that the bias will be small and regularization will provide estimates with smaller mean square error.

An estimator of the asymptotic covariance matrix is given by

$$\left(\hat{V} + \frac{\lambda_N}{N} D\right)^{-1} \hat{W} \left(\hat{V} + \frac{\lambda_N}{N} D\right)^{-1} \tag{9}$$

where all the matrices are evaluated at $\hat{\beta}^{\lambda}$. The matrices $\hat{V}$, $\hat{W}$ have been defined by (5) and (6), respectively. For large N, and if $\lambda_N = o(\sqrt{N})$, formula (9) reduces to that used for the asymptotic variance estimator of the conditional LS estimator $\hat{\beta}$-see Theorem 1.

4 Examples

A limited simulation study and a real data example are presented to complement the theoretical findings.

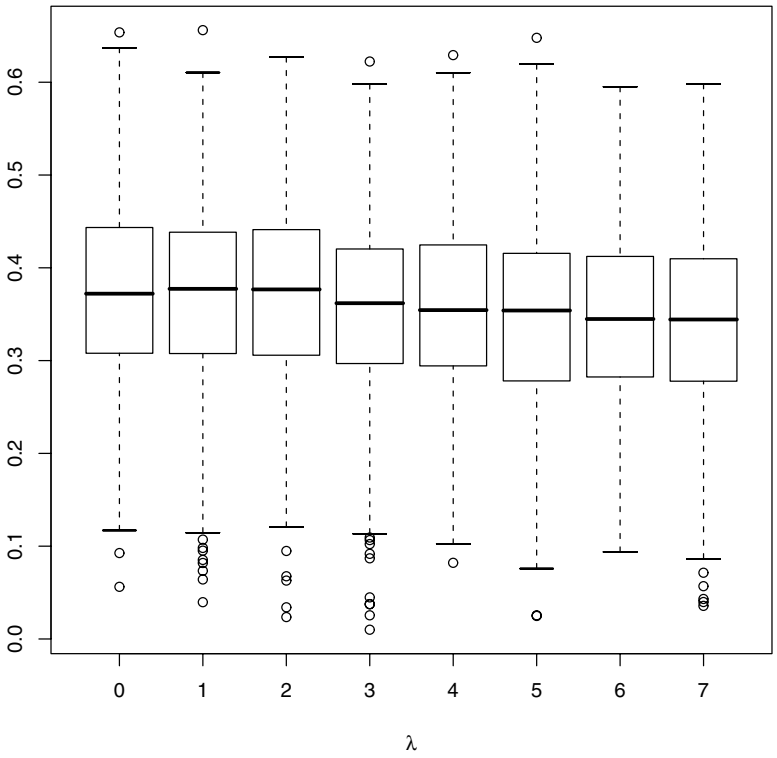


Fig. 1 Boxplots of the distribution of $\hat{\alpha}_1^{\lambda}$ for various values of the penalty parameter.

4.1 Simulations

To study the empirical performance of the penalized LS estimators for the INAR(p) model, a limited simulation study is presented. First, data are generated by the INAR(1) process

$$X_t = \alpha_1 \circ X_{t-1} + \varepsilon_t,$$

for $t = 1, 2, \ldots, N$, where the error sequence is assumed to be i.i.d. Poisson with mean μ. The computation were carried out by the statistical language `R` and all simulation output is based on 1000 runs.

The asymptotic normality of the restricted estimators is demonstrated for various values of the penalty parameter–see Figure 1–where the boxplots of the distribution of $\hat{\alpha}_1^{\lambda}$ are shown for $\lambda = 0, 1, 2, \ldots, 7$. The sample size is $N = 100$, $\alpha_1 = 0.4$ and $\mu = 1$. The asserted asymptotic normality is in agreement with the simulation findings.

Table 1 Penalized estimators for 100 observations from the INAR(2) model with true parameters $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$ and for different values of the Poisson mean μ. The regularization parameter varies from 0 to 10 by 0.5 and the number of simulations is 1000.

λ	$\mu = 0.50$				$\mu = 1.00$			
	$\hat{\mu}^\lambda$	$\hat{\alpha}_1^\lambda$	$\hat{\alpha}_2^\lambda$	$e_1(\lambda)$	$\hat{\mu}^\lambda$	$\hat{\alpha}_1^\lambda$	$\hat{\alpha}_2^\lambda$	$e_1(\lambda)$
0.0	0.531	0.0852	0.171	1.000	1.07	0.0821	0.170	1.000
0.5	0.531	0.0831	0.172	0.947	1.06	0.0816	0.171	0.906
1.0	0.535	0.0835	0.159	0.958	1.05	0.0927	0.167	0.893
1.5	0.537	0.0802	0.166	0.947	1.05	0.0865	0.169	0.886
2.0	0.535	0.0808	0.163	0.944	1.08	0.0805	0.165	0.962
2.5	0.545	0.0779	0.159	0.998	1.08	0.0766	0.164	0.955
3.0	0.539	0.0789	0.158	0.918	1.07	0.0792	0.166	0.904
3.5	0.549	0.0788	0.151	0.946	1.08	0.0756	0.163	0.893
4.0	0.553	0.0785	0.152	0.915	1.08	0.0798	0.164	0.977
4.5	0.549	0.0738	0.150	0.926	1.07	0.0820	0.161	0.916
5.0	0.555	0.0731	0.147	0.948	1.07	0.0835	0.158	0.825
5.5	0.549	0.0715	0.145	0.905	1.08	0.0793	0.158	0.884
6.0	0.548	0.0723	0.147	0.893	1.08	0.0773	0.163	0.848
6.5	0.548	0.0757	0.144	0.891	1.09	0.0813	0.158	0.906
7.0	0.556	0.0721	0.143	0.900	1.10	0.0730	0.159	0.932
7.5	0.554	0.0757	0.144	0.896	1.08	0.0826	0.154	0.837
8.0	0.563	0.0740	0.135	0.960	1.09	0.0798	0.155	0.830
8.5	0.560	0.0699	0.140	0.897	1.10	0.0751	0.151	0.902
9.0	0.561	0.0738	0.130	0.915	1.10	0.0758	0.148	0.922
9.5	0.566	0.0688	0.138	0.926	1.10	0.0763	0.150	0.936
10.0	0.562	0.0697	0.133	0.901	1.10	0.0768	0.150	0.894

Notice that $\lambda = 0$ corresponds to the ordinary CLS estimators while for large values of λ, the resulting estimator is more biased compared to the CLS estimator.

Furthermore, consider the INAR(2) model

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \varepsilon_t$$

where ε_t are assumed to be Poisson random variables with mean μ, as before. Table 1 shows the results of 1000 simulations when there are 100 observations available from the process at hand. Note that the resulting penalized estimators are biased as it was claimed before. However, their relative efficiency to the ordinary least squares estimators is superior in both cases considered. The quantity $e_1(\lambda)$–that is the efficiency–has been defined by

$$e_1(\lambda) = \frac{\text{MSE}(\hat{\beta}^\lambda)}{\text{MSE}(\hat{\beta})},$$

and it is the ratio of the mean square error of the constrained estimator to the mean square error of the unconstrained estimator. Table 2 shows the same results but for $N = 500$. Here, most of the values of $e_1(\lambda)$ fluctuate around unity showing that there is no any improvement by penalization.

Table 2 Penalized estimators for 500 observations from the INAR(2) model with true parameters $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$ and for different values of the Poisson mean μ. The regularization parameter varies from 0 to 10 by 0.5 and the number of simulations is 1000.

λ	$\mu = 0.50$				$\mu = 1.00$			
	$\hat{\mu}^\lambda$	$\hat{\alpha}_1^\lambda$	$\hat{\alpha}_2^\lambda$	$e_1(\lambda)$	$\hat{\mu}^\lambda$	$\hat{\alpha}_1^\lambda$	$\hat{\alpha}_2^\lambda$	$e_1(\lambda)$
0.0	0.508	0.0945	0.193	1.000	1.01	0.0982	0.193	1.000
0.5	0.507	0.0970	0.194	0.995	1.01	0.0969	0.195	1.018
1.0	0.507	0.0963	0.192	1.111	1.01	0.0981	0.192	0.971
1.5	0.508	0.0968	0.193	1.034	1.01	0.0979	0.193	1.032
2.0	0.508	0.0971	0.191	1.022	1.01	0.0957	0.193	1.055
2.5	0.509	0.0949	0.191	1.044	1.01	0.0964	0.193	0.993
3.0	0.508	0.0962	0.192	1.084	1.01	0.0946	0.194	0.973
3.5	0.509	0.0978	0.188	1.067	1.02	0.0954	0.192	1.008
4.0	0.509	0.0941	0.189	1.018	1.02	0.0977	0.190	1.013
4.5	0.513	0.0921	0.189	1.019	1.02	0.0943	0.191	1.074
5.0	0.511	0.0955	0.191	1.072	1.02	0.0964	0.192	1.044
5.5	0.510	0.0972	0.187	1.036	1.02	0.0977	0.192	1.068
6.0	0.511	0.0955	0.186	1.003	1.02	0.0962	0.191	0.950
6.5	0.517	0.0921	0.185	1.084	1.02	0.0965	0.191	0.984
7.0	0.514	0.0948	0.187	1.009	1.03	0.0934	0.188	1.008
7.5	0.513	0.0936	0.187	1.004	1.02	0.0962	0.191	1.061
8.0	0.514	0.0924	0.186	0.978	1.02	0.0945	0.188	0.964
8.5	0.517	0.0917	0.183	1.002	1.02	0.0967	0.187	0.990
9.0	0.512	0.0936	0.187	0.955	1.02	0.0975	0.190	0.970
9.5	0.516	0.0944	0.182	1.044	1.02	0.0975	0.189	1.007
10.0	0.519	0.0910	0.183	1.054	1.02	0.0919	0.188	1.041

4.2 Data Example

The Westgren gold particle data is used to demonstrate the penalized least squares estimation method. The data consists of consecutive count measurements of gold particles in a well defined colloidal solution of equally spaced points in time. These data have been analyzed by various authors, including Guttorp (1991), Grunwald et al. (2000) and more recently by Jung & Tremayne (2006). In particular, the first 370 observations are used throughout the subsequent analysis, along the lines of Jung & Tremayne (2006).

To analyze the data, consider the INAR(p) model (1) for $p = 1,2,3,4$. For comparison purposes, which are described below, the first four observations are removed and all models were fitted based on the 366 observations. Figure 2 shows the values of AIC for each INAR model fitted to the data, defined as

$$\mathrm{AIC}(p,\lambda) = 366 \log\left(\frac{\sum_{t=5}^{370}(X_t - \hat{\mu}^\lambda - \sum_{i=1}^{p}\hat{\alpha}_i^\lambda X_{t-i})^2}{366}\right) + 2\mathrm{df}_\lambda \qquad (10)$$

where the quantity df_λ is called the effective degrees of freedom as in the ordinary ridge regression. In other words, recall the definition of Q from (4) and set

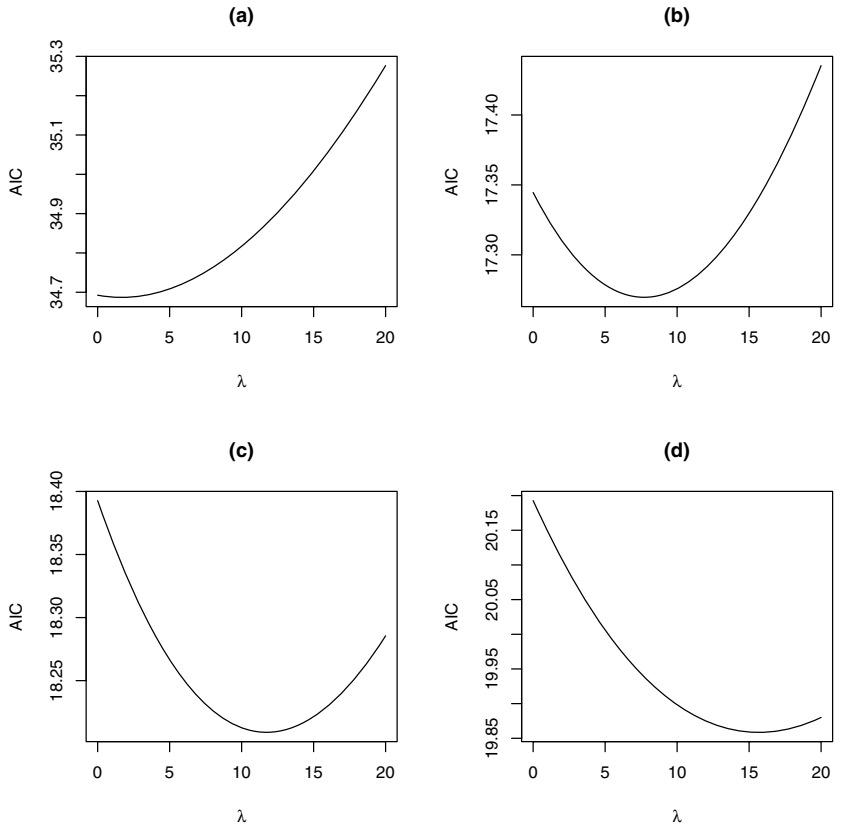


Fig. 2 Selection of λ by AIC for the gold particle data. (a) INAR(1), (b) INAR(2), (c) INAR(3), (d) INAR(4).

$$X = \begin{bmatrix} 1 & X_p & X_{p-1} & \dots & X_1 \\ 1 & X_{p+1} & X_p & \dots & X_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n-1} & X_{n-2} & \dots & X_{n-p} \end{bmatrix}.$$

it is clear that $Q = X'X$ and therefore the effective degrees of freedom are defined by

$$\mathrm{df}_\lambda = \mathrm{tr}\left(X(Q + \lambda D_{p+1})^{-1}X'\right),$$

where $\mathrm{tr}(.)$ denotes the trace of a matrix. Although estimation proceeds from the least squares, it can be argued that the AIC is the expected Kullback–Leibler distance of the maximum Gaussian likelihood model relative to the true distribution of the process, see Brockwell & Davis (1991, p. 306).

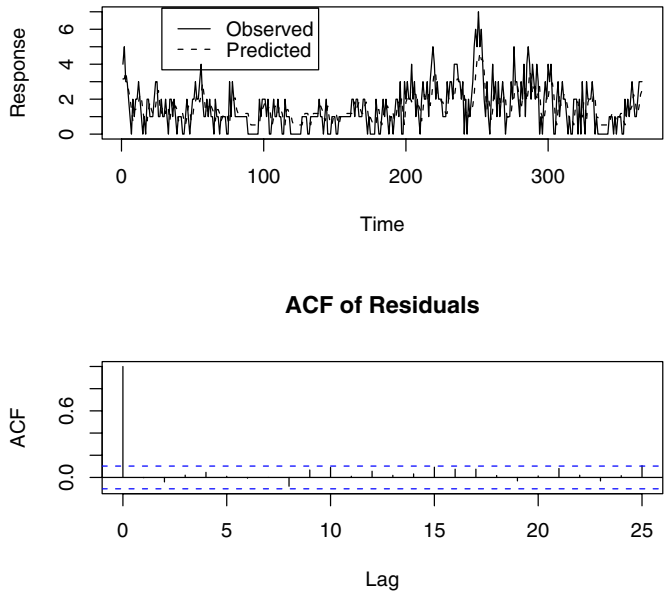


Fig. 3 Diagnostics for the INAR(2) model fitted to the gold particle data by minimizing (7) using $\lambda = 7.65$.

Using the above definition, and turning back to Figure 2, we note that the plot suggests the existence of a value of λ such that (10) attains a minimum. Notice that the values of the penalty parameter λ varies between 0 and 20 for a fine grid of values. When comparing the AIC from all different models, the INAR(2) yields its minimum value–in fact for all λ. Therefore the point that was made by previous authors that the INAR(2) model fits these data well is iterated further–see Jung & Tremayne (2006).

Hence this model is used for data fitting at the value of λ that minimizes (10). It turns out that $\lambda_{\text{opt}} = 7.65$ and the corresponding estimators are given by $\hat{\alpha}_1^{7.65} = 0.43082$, $\hat{\alpha}_2^{7.65} = 0.22685$ and $\hat{\mu}^{7.65} = 0.52510$. Figure 3 shows some further diagnostics for the model at hand. The upper panel shows a plot of the observed versus the predicted data while the lower plot shows the autocorrelation function of residuals. Both graphs indicate the adequacy of the INAR(2) model.

5 Discussion

This article introduces the ridge regression idea to the INAR processes. It was shown by theory and some supporting simulations that improvement over ordinary CLS is possible given a good choice of the regularization parameter. The choice of the regularization parameter is based on the minimization of the AIC and it was shown that for the Westgren gold particle data the method appears to work nicely. However, further investigation is needed to understand the results obtained from such procedure.

Integer autoregressive models have been generalized in different directions by several authors. For instance, Latour (1998) studies generalized integer valued autoregressive models of order p. This class of models is based on generalization of the thinning operator but their second order properties are similar to those of INAR(p) models. Hence the results reported here should be applicable in this class of models as well.

Another interesting class of models is that of conditional linear AR(1) models (see Grunwald et al. (2000)) specified by the following

$$m(X_t) = \alpha_1 X_{t-1} + \mu,$$

where $m(X_t) = \mathrm{E}[X_t \mid X_{t-1}]$, with X_t a time–homogeneous first-order Markov process. This class of model includes several AR(1) models proposed in the literature for non–Gaussian data. Inference is carried out either by maximum likelihood or by least squares. Therefore, the proposed ridge methods should apply to those models as well.

In a different direction, the recent contribution of Zhu & Joe (2006) extends the INAR(p) to include covariates. Estimation of regression coefficients is based on maximum likelihood and therefore the ridge constraints can be easily incorporated so that (7) is of the form of maximizing a penalized log likelihood function.

As a final remark, alternative penalties can be used so that model selection can be combined with estimation. For instance, consider penalty function of the following form

$$J(\beta) = \sum_{j=1}^{p} |\beta_j|^q,$$

where $q > 0$. The choices of $q = 1, 2$ yield to the Lasso (Tibshirani 1996) and ridge estimators respectively. In general these estimators were introduced by Frank & Friedman (1993) and were termed as Bridge estimators. When $q \leq 1$, the penalty function has the neat property to set some of the regression coefficient equal to 0, that is it can be used for model selection and estimation.

References

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **AC–19**: 716–723.

Al-Osh, M. A. & Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process, *Journal of Time Series Analysis* **8**: 261–275.

Alzaid, A. A. & Al-Osh, M. (1990). An integer-valued *p*th-order autoregressive structure (INAR(*p*)) process, *Journal of Applied Probability* **27**: 314–324.

Brockwell, P. J. & Davis, R. A. (1991). *Time Series: Data Analysis and Theory*, 2nd edn, Springer, New York.

Du, J. G. & Li, Y. (1991). The integer-valued autoregressive INAR(p) model, *Journal of Time Series Analysis* **12**: 129–142.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic Normality of the maximum likelihood estimates in generalized linear models, *Annals of Statistics* **13**: 342–368.

Fahrmeir, L. & Kaufmann, H. (1987). Regression Models for Nonstationary Categorical Time Series, *Journal of Time Series Analysis* **8**: 147–160.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer, New York.

Ferland, R., Latour, A. & Oraichi, D. (2006). Integer–valued GARCH processes, *Journal of Time Series Analysis Analysis* **27**: 923–942.

Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics* **35**: 109–148. (with discussion).

Grunwald, G. K., Hyndman, R. J., Tedesco, L. & Tweedie, R. L. (2000). Non-Gaussian conditional linear AR(1) models, *Australian & New Zealand Journal of Statistics* **42**: 479–495.

Guttorp, P. (1991). *Statistical Inference for Branching Processes*, Wiley, New York.

Hall, P. & Heyde, C. C. (1980). *Martingale Limit Theory and its Applications*, Academic Press, New York.

Hoerl, A. E. & Kennard, R. W. (1970a). Ridge regression:Applications to non–orthogonal problems, *Technometrics* **12**: 69–82.

Hoerl, A. E. & Kennard, R. W. (1970b). Ridge regression:Biased estimation for non–orthogonal problems, *Technometrics* **12**: 55–67.

Jung, R. C. & Tremayne, A. R. (2006). Coherent forecasting in integer time series models, *International Journal of Forecasting* **22**: 223–238.

Kedem, B. & Fokianos, K. (2002). *Regression Models for Time Series Analysis*, Wiley, Hoboken, NJ.

Klimko, L. A. & Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes, *The Annals of Statistics* **6**: 629–642.

Knight, K. & Fu, W. (2000). Asymptotics for lasso–type estimators, *Annals of Statistics* **28**: 1356–1378.

Latour, A. (1998). Existence and stochastic structure of a non-negative integer-valued autoregressive process, *Journal of Time Series Analysis* **19**: 439–455.

McKenzie, E. (1985). Some simple models for discrete variate time series, *Water Resources Bulletin* **21**: 645–650.

McKenzie, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions, *Advances in Applied Probability* **18**: 679–705.

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts, *Advances in Applied Probability* **20**: 822–835.

Neal, P. & Subba Rao, T. (2007). MCMC for integer–valued ARMA processes, *Journal of Time Series Analysis* **28**: 92–100.

Taniguchi, M. & Hirukawa, J. (2005). The Stein-James estimator for short- and long-memory Gaussian processes, *Biometrika* **92**: 737–746.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**: 267–288.

Zheng, H., Basawa, I. V. & Datta, S. (2006). Inference for the *p*th–order random coefficient integer–valued process, *Journal of Time Series Analysis* **27**: 411–440.

Zhu, R. & Joe, H. (2006). Modelling count data time series with Markov processes based on binomial thinning, *Journal of Time Series Analysis* **27**: 725–738.

Appendix

Suppose that $M_N^0 = -2^{-1}(\partial S_p(\beta)/\partial\mu) = \sum_{t=1}^{N}(X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \mu)$ and put $M_0^0 = 0$. Then

$$\begin{aligned}\mathrm{E}(M_N^0|\mathscr{F}_{N-1}) &= \mathrm{E}\left(M_{N-1}^0 + X_N - \sum_{i=1}^{p}\alpha_i X_{N-i} - \mu|\mathscr{F}_{N-1}\right)\\ &= M_{N-1}^0 + \mathrm{E}\left(X_N - \sum_{i=1}^{p}\alpha_i X_{N-i} - \mu|\mathscr{F}_{N-1}\right) = M_{N-1}^0,\end{aligned}$$

from the properties of the INAR(p) processes. Thus, the sequence $\{M_N^0, \mathscr{F}_N, N \geq 0\}$ forms a martingale which is square integrable. Furthermore, if condition (2) is fulfilled, then the sequence X_t is stationary and ergodic. Hence, from the ergodic theorem,

$$\frac{1}{N}\sum_{t=p+1}^{N}\left(X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \mu\right)^2 \to \mathrm{E}\left(X_{p+1} - \sum_{i=1}^{p}\alpha_i X_{p-i} - \mu\right)^2 \equiv \sigma_1^2,$$

almost surely, as $N \to \infty$. Therefore, by (Hall & Heyde 1980, Cor. 3.2) we obtain that

$$\frac{1}{\sqrt{N}}M_N^0 \to N(0, \sigma_1^2),$$

in distribution, as $N \to \infty$. Along the same lines, it can be shown that if $M_N^j = -2^{-1}(\partial S_p(\beta)/\partial\alpha_j) = \sum_{t=p+1}^{N} X_{t-j}(X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \mu) - \lambda_N\alpha_j$, for $j = 1, 2, \ldots, p$, then $\tilde{M}_N^j = M_N^j + \lambda_N\alpha_j$ is a martingale that satisfies

$$\frac{1}{N}\sum_{t=p+1}^{N} X_{t-j}^2\left(X_t - \sum_{i=1}^{p}\alpha_i X_{t-i} - \mu\right)^2 \to \mathrm{E}\left(X_{p+1-j}^2(X_{p+1} - \sum_{i=1}^{p}\alpha_i X_{p-i} - \mu)^2\right) \equiv \sigma_j^2,$$

almost surely, and

$$\frac{1}{\sqrt{N}}\tilde{M}_N^j \longrightarrow N(0, \sigma_j^2)$$

for all $j = 1, \ldots, p$. Using the assumption that $\lambda_N/\sqrt{N} \to \lambda_0 \geq 0$ then

$$\frac{1}{\sqrt{N}} M_N^j \longrightarrow N(-\lambda_0 \alpha_j, \sigma_j^2).$$

By the Cramer-Wold device and the properties of the INAR(p) process, it can be shown that

$$\frac{1}{\sqrt{N}} \begin{pmatrix} M_N^0 \\ M_N^1 \\ \vdots \\ M_N^p \end{pmatrix} \rightarrow N_{p+1}\left(-\lambda_0 b, W\right),$$

in distribution, as $N \rightarrow \infty$.

Recall the penalized conditional least squares estimators, given by (8). It can be shown that

$$\sqrt{N}(\hat{\beta}^{\lambda} - \beta) = \left(\frac{1}{N} Q + \frac{\lambda_N}{N} D_{p+1}\right)^{-1} \frac{1}{\sqrt{N}} \begin{pmatrix} M_N^0 \\ M_N^1 \\ \vdots \\ M_N^p \end{pmatrix} \rightarrow N_{p+1}(-\lambda_0 V^{-1} b, V^{-1} W V^{-1}),$$

in distribution, as $N \rightarrow \infty$. The theorem is proved.

Bayesian Inference for a Periodic Stochastic Volatility Model of Intraday Electricity Prices

Michael Stanley Smith

Abstract The Gaussian stochastic volatility model is extended to allow for periodic autoregressions (PAR) in both the level and log-volatility process. Each PAR is represented as a first order vector autoregression for a longitudinal vector of length equal to the period. The periodic stochastic volatility model is therefore expressed as a multivariate stochastic volatility model. Bayesian posterior inference is computed using a Markov chain Monte Carlo scheme for the multivariate representation. A circular prior that exploits the periodicity is suggested for the log-variance of the log-volatilities. The approach is applied to estimate a periodic stochastic volatility model for half-hourly electricity prices with period $m = 48$. Demand and day types are included in both the mean and log-volatility equations as exogenous effects. A nonlinear relationship between demand and mean prices is uncovered which is consistent with economic theory, and the predictive density of prices evaluated over a horizon of one week. Overall, the approach is shown to have strong potential for the modelling of periodic heteroscedastic data.

Key words: Periodic Autoregression; Bayesian MCMC; Electricity Price Forecasting; Multivariate Stochastic Volatility; Vector Autoregression; Longitudinal Model; Heteroscedasticity

1 Introduction

The univariate stochastic volatility model has attracted a great deal of attention by researchers over the past fifteen years. Shephard (2005) gives an overview of the development of the model, along with a collection of selected readings. Bayesian inference, computed via Markov chain Monte Carlo (MCMC) methods, has proven

Michael Stanley Smith
Melbourne Business School, University of Melbourne, 200 Leicester Street, Carlton, Victoria 3053, Australia. e-mail: mike.smith@mbs.edu

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_19,

popular for this class of models; for example, see Jacquier et al. (1994), Chib et al. (2002; 2006) among others. There have also been a number of extensions to the multivariate case, with recent surveys given by Asai et al. (2006) and Chib et al. (2009). At the same time, periodic autoregressions (PAR), popularised by Pagano (1978), have increasingly been used to model data that exhibit periodicity in their dependence structures; see Franses & Paap (2004) for a recent overview. In this paper, the univariate stochastic volatility model is extended to the case where both the level and log-volatility process follow Gaussian PARs. The resulting model is labelled here the Gaussian periodic stochastic volatility (PSV) model. This is different than the model of the same name suggested by Tsiakas (2006), which instead accounts for periodicity by introducing exogenous effects only.

Pagano (1978) observed that a PAR can be expressed as a vector autoregression (VAR) for a longitudinal vector of contiguous observations of length equal to the period, and vice versa. The VAR is assumed to be first order, which is not as restrictive as one might expect because PAR models with lag length no more than the period can always be represented in this fashion (Franses & Paap 2004; pp. 31–35). When the period is long, the number of parameters can be large, so that sparse lag structures are often considered. If the PAR only has non-zero lags for the k immediately preceding time points, plus one at the period, then the disturbance to the VAR representation has a band k precision (inverse covariance) matrix. In the longitudinal literature this corresponds to assuming the disturbance vector is Markov of order k (Smith & Kohn 2002). Using the VAR representation, the PSV can be expressed as a multivariate stochastic volatility model, for which Bayesian inference can be computed using a MCMC algorithm. It is important to note here that the PSV cannot be expressed as a factor model, as is currently popular for the multivariate modelling of stock returns; see, for example, Pitt & Shephard (1999) and Chib et al. (2006).

Exogenous variables are considered for both the mean of the series, and also for the mean of the log-volatility process. To enforce the band structure of the precision matrices a transformation to an unconstrained parameterisation as originally suggested by Panagiotelis & Smith (2008) is used. A circular prior is suggested for the log-variance of the log-volatility process that shrinks together values adjacent in time. Such a prior exploits the unique structure of the periodic model.

The PSV is used to model half-hourly electricity prices from a contemporary wholesale electricity market. In such markets electricity is traded at an intraday frequency in reference to a spot price. The dynamics of this spot price are totally unlike that of other financial assets. There is signal in at least the first and second moments, and a strong diurnal pattern exists in all features; see Karakatsani & Bunn (2008). There is a growing literature on the time series modeling of electricity prices at an intraday level; see, for example, Barlow (2002), Conejo et al. (2005), Haldrup & Nielsen (2006), Weron & Misiorek (2008) and Panagiotelis & Smith (2008), among others. Recently, Guthrie & Videbeck (2007) show that a PAR is effective in capturing signal in the first moment of intraday prices in the New Zealand market; see also Broszkiewicz-Suwaj et al. (2004) for a PAR fit to NordPool data. Our analysis extends this approach to the second moment. Day type and electricity demand are also included as exogenous effects in both level and volatility equations. By using flexible

functional forms, the relationship between demand and mean prices is estimated and matches that anticipated by economic theory.

2 Periodic Autoregressions

A zero mean series $\{x(1),\ldots,x(n)\}$ follows a periodic autoregressive (PAR) process of period m if

$$x(s) = \sum_{j\in L^{(i)}} \rho_j^{(i)} x(s-j) + \sigma^{(i)} z(s), \quad \text{for } s = 1,\ldots,n, \tag{1}$$

where $z(s)$ is a white noise process and $L^{(i)}$ is a set of positive integers. The autoregressive coefficients $\rho^{(i)}$, variances $(\sigma^{(i)})^2$ and even lag structure $L^{(i)}$ may vary over the period, which is denoted by index $i = s \bmod m$. PAR models have been used to account for seasonally varying dependence in quarterly ($m = 4$) and monthly ($m = 12$) data; for example, see Osborn & Smith (1989). However, they can also be used to capture diurnally varying dependence in data collected at an intraday resolution. In this study the data are observed at a half-hourly resolution and $m = 48$ throughout, so that i corresponds to the half hour, and t the day, of time point s.

Let $x_{it} = x(s)$, where $i = s \bmod m$, $t = [s/m] + 1$ and $[a]$ is the integer part of a. Then, Pagano (1978) observed that a PAR can be expressed as a vector autoregression (VAR) for the longitudinal vector $x_t = (x_{1t}, x_{2t}, \ldots, x_{mt})'$ and vice versa. Following Franses & Paap (2004; Chapter 3) if $L = \max_i(L^{(i)}) \leq m$ then the PAR in equation (1) can be represented by the first order VAR

$$\Phi^A x_t = \Phi^B x_{t-1} + z_t, \text{ for } t = 1,\ldots,T.$$

Here, $n = Tm$, z_t is a vector of independent white noise processes, and $\Phi^A = \{\phi_{i,j}^A\}$ and $\Phi_{i,j}^B = \{\phi_{i,j}^B\}$ are sparse matrices with elements

$$\phi_{i,j}^A = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } j > i \\ -\rho_{i-j}^{(i)} & \text{if } j < i \end{cases}, \quad \phi_{i,j}^B = \begin{cases} \rho_{i+m-j}^{(i)} & \text{if } i+m-j \leq L \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where $\rho_j^{(i)} = 0$ if $j \notin L^{(i)}$. Multiplying on the left by $(\Phi^A)^{-1}$ gives:

$$x_t = \Phi x_{t-1} + e_t, \text{ for } t = 1,\ldots,T, \tag{3}$$

where $\Phi = (\Phi^A)^{-1}\Phi^B$ and $\text{Var}(e_t)^{-1} = \Sigma^{-1} = (\Phi^A)'\text{diag}(\sigma^{(1)},\ldots,\sigma^{(m)})^{-2}\Phi^A$. Only stationary series are considered in this paper, which occurs for both the longitudinal VAR process and underlying PAR when all the eigenvalues of Φ lie inside the unit circle (Franses & Paap 2004, pp. 34–39).

PAR models are highly parameterised when m is large, so that sparse lag structures are often assumed to make the model more parsimonious. If $L^{(i)} = \{1,2,\ldots,k,m\}$ for $k < m$, then the precision matrix Σ^{-1} is band k and Φ is a sparse matrix. For example, if the data are intraday, then this lag structure relates an observation to that k intraday periods immediately prior, plus the observation at the same time the previous day. Assuming that Σ^{-1} is band k also corresponds to assuming the elements of the longitudinal vector e_t are Markov of order k; for example, see Pourahmadi (1999) and Smith & Kohn (2002).

In this paper Gaussian PAR models are employed, so that $z(s) \sim N(0,1)$ in equation (1) and e_t are an independent $N(0,\Sigma)$ series in equation (3). The precision matrix Σ^{-1} is assumed band k. In the empirical work in Section 5.2 a low bandwidth of $k = 2$ is assumed, and to make the representation more parsimonious, Φ is also assumed to be a diagonal matrix. This is an approximation to the full structure of Φ that proves useful when estimating high dimensional models, such as that examined here where $m = 48$. However, the approach outlined in this paper applies to Φ with any pattern, including when patterned as $(\Phi^A)^{-1}\Phi^B$.

Last, it is noted that Panagiotelis & Smith (2008) employ such a PAR with $k = 2$ for the mean-corrected logarithm of electricity prices, but where the disturbances e_t are distributed multivariate skew t. They also approximate Φ, but with a diagonal matrix with three additional non-zero elements in the upper right hand corner. The idea of this approximation is to relate each observation x_{it} to those in the preceding $k = 2$ half hours, and also that at the same time the previous day, but in a more parsimonious manner than the patterned matrix $(\Phi^A)^{-1}\Phi^B$.

3 Periodic Stochastic Volatility Model

3.1 The Model

Using the first order VAR representation of a PAR in Section 2, the Gaussian stochastic volatility model can be extended to incorporate periodicity as follows. Let $y_t = (y_{1t},\ldots,y_{mt})'$ be a longitudinal vector of $m = 48$ half-hourly electricity spot prices observed on day t. Exogenous linear effects are introduced into the mean with a $(mp \times 1)$ vector of coefficients $\beta = (\beta_1',\ldots,\beta_m')'$, where β_i is a $(p \times 1)$ vector of coefficients for half hour i. The model for the observations is:

$$y_t = X_t\beta + e_t\,, \;\; \text{where } e_t = \Phi e_{t-1} + H_t^{1/2}u_t\,, \tag{4}$$

for $t = 1,\ldots,T$. Here, X_t is the $(m \times mp)$ block matrix of regressors observed at time t corresponding to β and $e_t = (e_{1t},e_{2t},...,e_{mt})'$. The mean-corrected longitudinal vector e_t follows a first order Gaussian VAR with autoregressive coefficient matrix $\Phi = \{\phi_{i,j}\}$. The disturbance has variance $H_t = \text{diag}(\exp(h_{1t}),...,\exp(h_{mt}))$ and correlation matrix C, so that the u_t are distributed independently $N(0,C)$. The inverse

correlation matrix C^{-1} is assumed band k_1, so that the elements of the longitudinal vector u_t are Markov of order k_1.

The log-volatilities also follow a PAR expressed as a first order VAR with linear exogenous mean effects, so that if $h_t = (h_{1t}, \ldots, h_{mt})'$, then

$$h_t = Z_t \alpha + \xi_t, \text{ where } \xi_t = \Psi \xi_{t-1} + \eta_t, \tag{5}$$

for $t = 1, \ldots, T$. The errors η_t are independently distributed $N(0, \Sigma)$, α is a $(mq \times 1)$ vector of coefficients for q exogenous effects and Z_t is the corresponding $(m \times mq)$ matrix of independent variables. The mean-corrected latent log-volatilities $\xi_t = (\xi_{1t}, \xi_{2t}, ..., \xi_{mt})'$ follow the VAR with autoregressive coefficient matrix $\Psi = \{\psi_{i,j}\}$. The precision matrix Σ^{-1} is assumed to be band k_2, so that the elements of the longitudinal vector ξ_t are Markov of order k_2. In the state space literature, equations (4) and (5) are referred to as the observation and transition equations, respectively.

3.2 Matrix Parameterisations

To ensure that C is constrained to the space of correlation matrices with a band k_1 inverse, a re-parameterisation suggested by Panagiotelis & Smith (2008) and Smith & Cottet (2006) is employed. Define a band k positive definite matrix Ω_C such that

$$C^{-1} = [\text{diag}(\Omega_C^{-1})]^{1/2} \Omega_C [\text{diag}(\Omega_C^{-1})]^{1/2}.$$

The parameterisation employed is the upper triangular Cholesky factor $R_C = \{r_{C,ij}\}$, such that $\Omega_C = R_C' R_C$. Note that C is a correlation matrix with $m(m-1)/2$ free elements, so that to identify the parameterisation the elements $r_{C,ii} = 1$ for all i. This parameterisation is particularly useful for two reasons. First, the upper bandwidth of R_C is the same as the bandwidth of C^{-1}, so that simply setting $r_{C,ij} = 0$ for all i, j such that $j - i > k_1$, ensures C^{-1} is band k_1. Second, the non-fixed elements of R_C are unconstrained and easier to generate in the sampling scheme in Section 4 than the elements of Ω_C when m is large.

For $\Sigma = \{\sigma_{ij}\}$ the variances $D = \text{diag}(\sigma_{11}, \ldots, \sigma_{mm})$ are isolated, so that $\Sigma = D^{1/2} B D^{1/2}$. This enables informative priors to be placed on the variances of the elements of η_t as discussed in Section 3.4. The correlation matrix B is parameterised in the same manner as C. That is, using the Cholesky factor $R_B = \{r_{B,ij}\}$ of a positive definite matrix $\Omega_B = R_B' R_B$, such that

$$B^{-1} = [\text{diag}(\Omega_B^{-1})]^{1/2} \Omega_B [\text{diag}(\Omega_B^{-1})]^{1/2}.$$

Again, for identification $r_{B,ii} = 1$ for all i, and setting $r_{B,ij} = 0$ for all i, j such that $j - i > k_2$, ensures that Σ^{-1} is band k_2. Again, the non-fixed elements of R_B are unconstrained and relatively easy to generate in the sampling scheme. While there is no reason why the bandwidth of C^{-1} and B^{-1} cannot differ, $k_1 = k_2 = 2$ in our empirical work.

3.3 The Augmented Likelihood

As noted by previous authors the likelihood is unavailable in closed form for such stochastic volatility models. Instead, focus is usually on the likelihood augmented with the latent volatilities, which is employed here. To simplify the analysis estimation is undertaken conditional on the pre-period observation vector y_0. In addition, a stationary distribution for the process $\{\xi_t\}$ is assumed, so that the marginal distribution $h_1|\alpha,\Psi,\Sigma \sim N(Z_1\alpha,\Gamma)$, where Γ is a closed form function of Σ and Ψ (Hamilton 1994; p.265). Let $y = \{y_1,...,y_T\}, h = \{h_1,...,h_T\}$ and Π be the set of model parameters, then the likelihood augmented with the latent volatilities h is

$$\begin{aligned} p(y,h|\Pi,y_0) &= p(y_1,h_1|\Pi,y_0)\prod_{t=2}^{T} p(y_t,h_t|\Pi,y_{t-1},h_{t-1}) \\ &= p(h_1|\Pi)\prod_{t=1}^{T} p(y_t|\Pi,y_{t-1},h_t)\prod_{t=2}^{T} p(h_t|\Pi,h_{t-1}). \end{aligned} \tag{6}$$

The Jacobian of the transformation between u_t and y_t is $|H_t^{-1/2}|$, so that

$$\begin{aligned} p(y,h|\Pi,y_0) \propto\ & |\Gamma|^{-1/2}\exp\left\{-\frac{1}{2}(h_1 - Z_1\alpha)'\Gamma^{-1}(h_1 - Z_1\alpha)\right\} \\ & |C|^{-T/2}\prod_{t=1}^{T}|H_t^{-1/2}|\exp\left\{-\frac{1}{2}u_t'C^{-1}u_t\right\} \\ & (|D||B|)^{-(T-1)/2}\prod_{t=2}^{T}\exp\left\{-\frac{1}{2}\eta_t'\Sigma^{-1}\eta_t\right\}. \end{aligned} \tag{7}$$

3.4 Priors

In previous Bayesian estimation of the univariate stochastic volatility model informative priors for the conditional variance of the log-volatilities have often been employed (Kim et al. 1998; Chib et al. 2002). In the periodic case here informative priors which shrink together adjacent elements of diag(Σ) (and the first and last elements) may improve inference. This motivates a Gaussian circular prior similar to the pairwise shrinkage priors employed in Bayesian regression smoothing (Lang & Brezger 2004). Specifically, if $\delta_i = \log(\sigma_{ii})$ then the circular prior has $\delta_i|\delta_{i-1} \sim N(\delta_{i-1},\tau_\delta^2)$ for $i = 2,\ldots,m$, and $\delta_1 \sim N(\delta_m,\tau_\delta^2)$. This results in the informative prior for the vector $\delta = (\delta_1,\ldots,\delta_m)'$

$$p(\delta|\tau_\delta^2) \propto (\tau_\delta^2)^{-m/2}\exp\left\{-\frac{1}{2\tau_\delta^2}\delta' W\delta\right\},$$

with highly sparse precision matrix $W = \{w_{i,j}\}$ which has non-zero elements $w_{i,i} = 2$ for $i = 1,\ldots,m$; $w_{i,i+1} = w_{i+1,i} = -1$ for $i = 1,\ldots,m-1$; and $w_{1,m} = w_{m,1} = -1$. The parameter τ_δ^2 is interpretable as a shrinkage parameter and a conjugate $IG(1.01, 0.01)$ hyperprior is assumed, which is shown not to dominate it's marginal posterior in the empirical work.

Let ϕ and ψ be the non-zero elements of Φ and Ψ, respectively. The priors $p(\phi)$ and $p(\psi)$ are uniform on the region where $\{e_t\}$ and $\{\xi_t\}$ are stationary. That is, $p(\phi) \propto I(\Phi \in \mathscr{S})$ and $p(\psi) \propto I(\Psi \in \mathscr{S})$, where $\mathscr{S}$ is the space of $(m \times m)$ matrices with eigenvalues inside the unit circle, and the indicator function $I(A) = 1$ if A is true, and $I(A) = 0$ otherwise. When Φ and Ψ are diagonal, this simplifies to the priors $p(\phi) \propto \prod_{j=1}^m I(-1 < \phi_{j,j} < 1)$ and $p(\psi) \propto \prod_{j=1}^m I(-1 < \psi_{j,j} < 1)$. The coefficients α and β of the exogenous effects, and the non-fixed elements of the Cholesky factors R_B and R_C, all have flat priors, although it is straightforward to incorporate informative priors if required.

4 Bayesian Posterior Inference

Because it is difficult to compute the marginal posterior distribution of the parameters analytically, MCMC is used to evaluate the posterior distribution. Such an approach has proven effective in obtaining inference in univariate stochastic volatility models (Jacquier et al. 1994; Kim et al. 1998; Chib et al. 2002) as well as in multivariate extensions (Pitt & Shephard 1999; Chib et al. 2006; Chan et al. 2006; Smith & Pitts 2006).

4.1 Sampling Scheme

The following sampling scheme is employed to obtain Monte Carlo iterates from the augmented posterior density $p(\Pi, h, \tau_\delta^2 | y, y_0)$, where $\Pi = \{\beta, \phi, R_C, \alpha, \psi, R_B, \delta\}$, which are used to construct posterior inference. Below, the notation $\{\Pi \backslash A\}$ denotes the parameter set Π without element A, $h_{b,t}$ denotes a contiguous sub-vector of h_t and $h_{\backslash b,t}$ denotes h without $h_{b,t}$.

Sampling Scheme
Generate sequentially from each of the following conditional posterior distributions:

(1) $p(\phi_{i,j} | \{\Pi \backslash \phi_{i,j}\}, h, y, y_0)$ for all non-zero elements of Φ
(2) $p(r_{C,ij} | \{\Pi \backslash r_{C,ij}\}, h, y, y_0)$ for i, j such that $0 < j - i < k_1$
(3) $p(h_{b,t} | h_{\backslash b,t}, \Pi, y, y_0)$, for all blocks b and $t = 1, ..., T$
(4) $p(\psi_{i,j} | \{\Pi \backslash \psi_{i,j}\}, h, y, y_0)$ for all non-zero elements of Ψ
(5) $p(r_{B,ij} | \{\Pi \backslash r_{B,ij}\}, h, y, y_0)$ for i, j such that $0 < j - i < k_2$
(6a) $p(\delta_i | \{\Pi \backslash \delta_i\}, \tau_\delta^2, h, y, y_0)$, for $i = 1, ..., m$
(6b) $p(\tau_\delta^2 | \delta)$

(7) $p(\beta|\{\Pi\backslash\beta\},h,y,y_0)$
(8) $p(\alpha|\{\Pi\backslash\alpha\},h,y,y_0)$

The main features of the sampling scheme are briefly outlined below, while a more detailed description is given in the Appendix.

In Step (2) C is generated through its parameterisation R_C, where the non-fixed upper triangular elements $r_{C,ij}$ are generated one at a time using a Metropolis-Hastings step. The proposal is a normal approximation to the conditional distribution centred around its mode, obtained using quasi-Newton-Raphson with numerical derivatives. This approach to generating a correlation matrix is simpler than generating directly from the elements of Ω_C. This is because the upper and lower bounds for each of the elements of Ω_C need recomputing at each sweep to ensure Ω_C is positive definite (Barnard et. al. 2000; Chan et al. 2006) which slows the sampling scheme when m is large. In Step (5) the non-fixed upper triangular elements of R_B are generated one element at a time using a normal approximation as a Metropolis-Hasting proposal in a similar manner as for the non-fixed elements of R_C.

Following Shephard & Pitt (1997) a Metropolis-Hastings step is used to generate the log-volatility vector $h_{b,t}$ in Step (3). The proposal is a Gaussian approximation to the conditional posterior centred around its mode obtained by Newton-Raphson. The gradient and Hessian of the log-density are calculated analytically and are provided in the Appendix. In the empirical work in Section 5.2 h_t is partitioned into 8 blocks of 6 elements each.

In Step (1) the distribution of each non-zero $\phi_{i,j}$ is constrained Gaussian. In Step (4) a Metropolis-Hastings step is employed with constrained Gaussian proposal equal to the conditional density omitting the term $p(h_1|\Pi)$. A Metropolis-Hastings step is used in Step (6a) based on a Gaussian approximation to the conditional posterior. The smoothing parameter of the circular prior in Step (6b) is generated directly from its inverse gamma posterior distribution.

In Step (7) the conditional density of β is recognizable as a normal conditional distribution. In applications where m and/or p are large (such as the application in this paper where $mp = 432$) it would not normally be computationally feasible to generate β as a single vector. However, because C^{-1} is band k_1, the precision matrix of the conditional distribution of β is block diagonal and vectors with large values of m can be generated. In Step (8) the vector α is generated using a Metropolis-Hastings step. The proposal is based on the conditional posterior omitting the term $p(h_1|\Pi)$, which is recognizable as a normal density.

4.2 Posterior Inference and Forecasts

After the Markov chain has converged, Monte Carlo iterates are obtained from the augmented posterior density $p(h,\Pi,\tau_\delta^2|y)$. These can be used to construct the full spectrum of posterior inference, including estimates for the marginal posterior means of the parameters, log-volatilities and shrinkage parameters, which are used as point estimates. In addition, $100(1-\alpha)\%$ posterior probability intervals for each parameter

or volatility can be computed by ranking the Monte Carlo iterates and counting off $100\alpha/2\%$ of the upper and lower values.

Let $y^f = \{y_{T+1}, \ldots, y_{T+T_2}\}$ and $h^f = \{h_{T+1}, \ldots, h_{T+T_2}\}$ be T_2 future values of the longitudinal process and associated log-volatilities. Then appending the following steps to the sampling scheme in Section 4.1 produces iterates from the predictive distribution $p(h^f, y^f | y, y_0)$.

Forecasting Scheme
For $t = 1, ..., T_2$ generate from the conditional distributions:

(F1) $p(h_{T+t} | h_{T+t-1}, \Pi)$
(F2) $p(y_{T+t} | h_{T+t}, y_{T+t-1}, \Pi)$

Here, $h_t | h_{t-1}, \Pi$ is normal with mean $Z_t\alpha + \Psi\xi_{t-1}$ and variance Σ and $y_t | h_t, y_{t-1}, \Pi$ is also normal with mean $X_t\beta + \Phi e_{t-1}$ and variance $H_t^{1/2} C H_t^{1/2}$. The Monte Carlo estimates of the predictive means $E(h^f | y, y_0)$ and $E(y^f | y, y_0)$ can be used as point forecasts. Monte Carlo estimates of the predictive probability intervals can be computed in the same manner as the parameter posterior probability intervals and used as forecast intervals.

5 Intraday Electricity Prices

5.1 The Australian Electricity Market and Spot Price

During the past twenty years governments in many developed countries have introduced wholesale electricity markets. One of the earliest of these is the Australian National Electricity Market (NEM), which has been in operation since 1998, although some earlier state-based markets have been in operation since 1994. Due to inter-connection constraints and transmission loss, electricity is poorly transportable over long distances and the price for electricity varies by location. In particular, prices paid by customers in the state of New South Wales (NSW) are obtained by referring transmission to a bulk supply point on the western edge of the state capital Sydney, which is the spot price examined here.

Wholesale markets around the world operate by varying rules, and the NEM works as follows. Generating utilities submit supply curves (or stacks) to the NEM management company (NEMMCO) prior to generation. A generator will offer to supply more electricity the higher the price. Based on short-term demand forecasts[1] NEMMCO matches forecast demand for each future five minute period against the average of the supply curves submitted by all generators for that 5 minute period (this is the industry supply curve). The generators are then scheduled to dispatch in bid-price order until forecast demand is fully matched. The spot price for the five minute period is set equal to the most expensive generation capacity that is

[1] See Cottet & Smith (2003) and Soares & Medeiros (2008) for a discussion of methods used to obtain short-term demand forecasts.

ultimately dispatched to meet demand. Half-hourly prices are then created as the average of the six five minute periods, and this is the data examined here. The rules of the NEM, along with the fact that electricity is a flow commodity that cannot be stored economically, induces a strong systematic component to intraday electricity prices. As documented by Escribano et al. (2002), Knittel & Roberts (2005) and Koopman et al. (2007), these features are common to all but a few markets.

The spot price occurs at the equilibrium of supply and demand. Demand can be observed exactly because, in the absence of blackouts or load-shedding, it is equal to system load which is read off the grid. Spot prices tend to vary positively with demand patterns, although the relationship is nonlinear because as demand increases beyond a certain point generation capacity with rapidly increasing marginal cost requires dispatching. This issue is discussed in greater detail in the empirical analysis in Section 5.2.

Obtaining supply-side data for analysis is more difficult, although the impact of supply on the spot price is likely to be two-fold. The first is potential intra and inter-day autocorrelation in both price and it's second moment due to the serial dependence in factors that affect the industry supply curve. The second is the existence of price spikes. These can be due to either under-forecasting of short-term demand by NEMMCO, or to a sudden supply disruption. In both cases a spike will occur because expensive generating capacity with a very short ramping time needs to be brought online at short notice to maintain system stability. In this case, the prices will often mean-revert quickly because within one hour cheaper generating capacity can usually be ramped to a level to meet the shortfall in supply. In effect, these spikes are caused by the shape of the aggregate industry supply curve. The time-to-ramp constraints result in a kink in this curve at a certain capacity level after which the curve has a steep upward slope. Sudden unanticipated changes in demand or supply can shift the point of intersection to the right of this kink, resulting in a price spike. In Section 5.2 it is demonstrated how estimation of average supply curves, including the location of the kink, at different times of the day is possible using the price data.

Figure 1 plots the year 2000 spot price (in dollars per kilowatt hour; \$/KWh) and corresponding half-hourly total NSW system demand (in gigawatt hours; GWh) in four panels that correspond to the southern hemisphere seasons. The data exhibit strong daily periodicity and there is likely to be persistence in the first and second moments, both between and within days. Large spikes in prices also occur and rapidly revert to the mean price level. Strong seasonal patterns appear to exist with the average winter price being higher than the average summer price, although summer proves to be the most volatile season. Such features have been observed in the NEM since its inception.

These empirical features in the data are the result of the market design discussed above, along with the strong systematic component to demand. They motivate the adoption of a periodic time series model to account for the strong diurnal variation of all aspects of the data. Demand for electricity can be included as an exogenous effect. Features of prices that are likely to be present are a nonlinear relationship

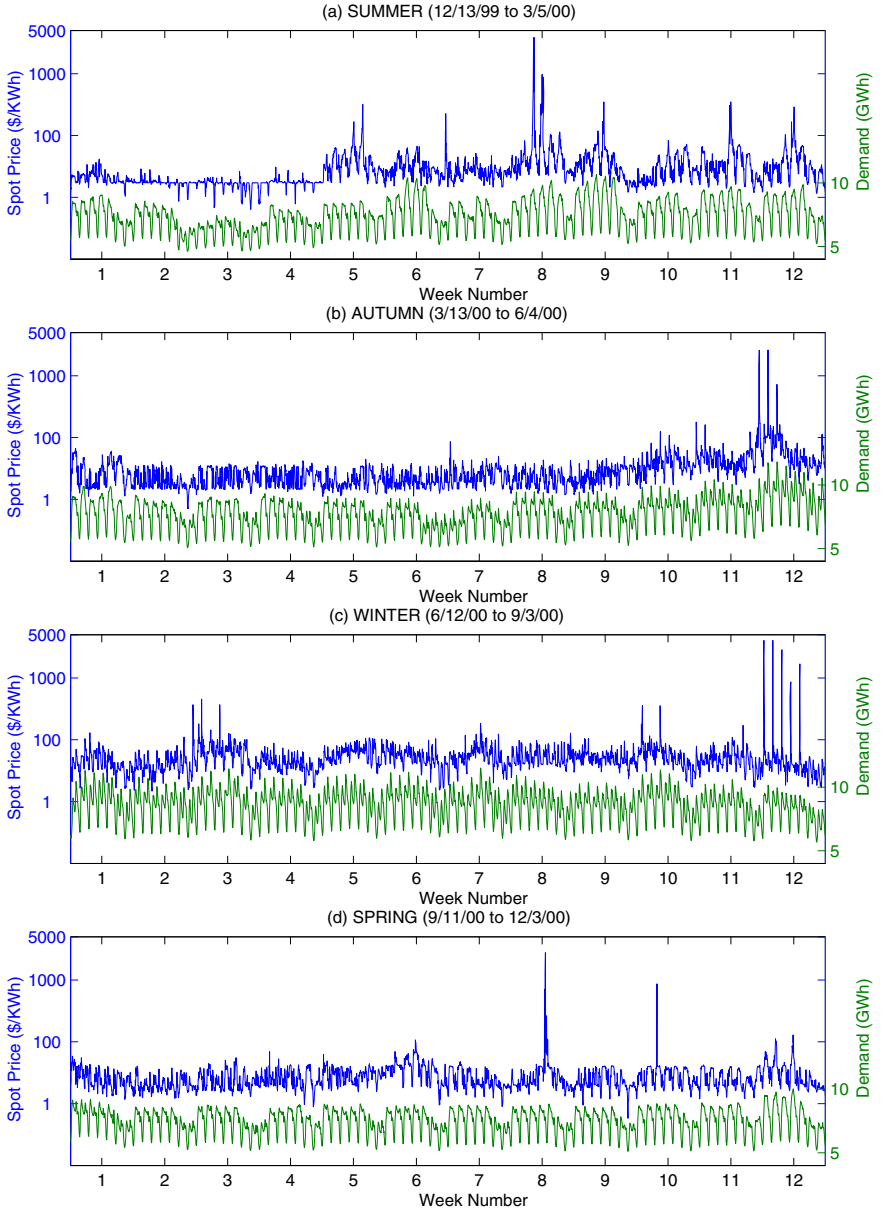


Fig. 1 The upper series is the NSW electricity spot price (in dollars per KiloWatt hour on the left-hand logarithmic axis), while the lower series is demand (in GigaWatt hours on the right-hand axis). Each panel corresponds to one of the four southern hemisphere seasons in 2000.

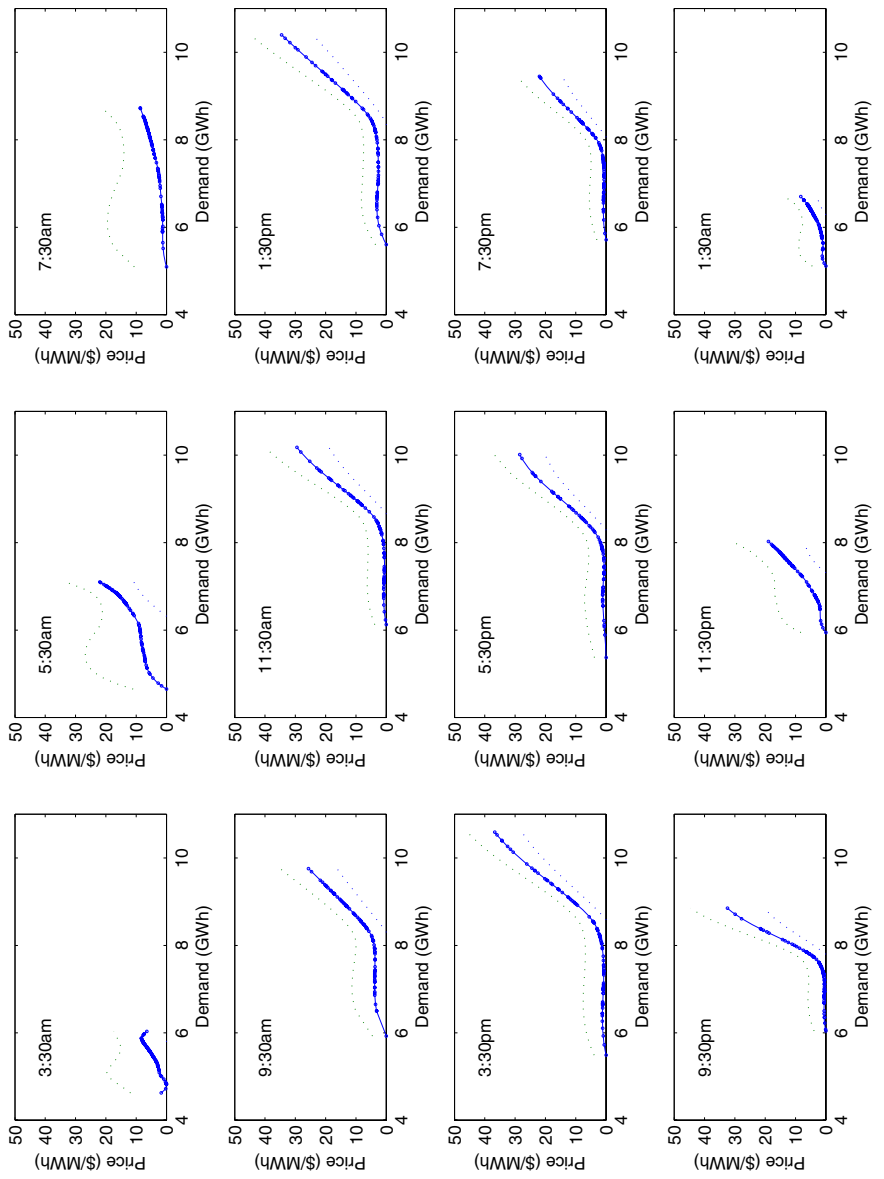


Fig. 2 Estimates of the exogenous effect of demand on average price at twelve equally-spaced half hours. The estimated functions are plotted in bold over their observed domain, normalised so that they start at zero. The dashed lines represent 90% posterior probability intervals for these effects. Demand is measured in GigaWatt hours (GWh) and price in dollars per MegaWatt hour ($/MWh).

with demand and periodic autocorrelation in the first and second moments. The autocorrelation may be between observations on an intraday or inter-day basis, or both. For a recent discussion of the implication of market structure on the dynamics of electricity prices see Karakatsani & Bunn (2008).

5.2 Empirical Analysis of NSW Spot Price

The PSV model is used to analyse the first $T = 84$ days of half-hourly summer electricity spot prices depicted in Figure 1(a). This is a particularly challenging set of data to fit because in the first 28 days low demand often resulted in fixed price baseline generation capacity setting the marginal price, which is not the case for the latter period. The first element in the longitudinal vector corresponds to 03:30 (approximately the overnight demand and price low) and the last to 03:00, so that the period is $m = 48$. The bands of Σ^{-1} and C^{-1} are $k_1 = k_2 = 2$ and the autoregressive coefficient matrices Φ and Ψ are assumed to be diagonal.

The following exogenous variables are employed in the observation and transition equations. A constant was employed in both equations, along with four dummy variables to capture any day type effects on Monday, Friday, Saturday and Sunday/Public Holidays. The exogenous effect of demand is included in the observation equation using a flexible functional form to capture any nonlinearity. In particular, if demand DEM is normalized between 0 and 1 for each half hour, then the radial basis terms $|DEM - K_j|^2 \log(|DEM - K_j|)$, for $K_j = 0.3, 0.6$ and 0.9, are included, along with a linear term in DEM. For the transition equation only a linear demand effect is included.

The circular prior outlined in Section 3.4 is used for the log-variances δ, smoothing the elements of δ across the day and stabilizing the posterior distribution of the latent volatilities h_t and the other parameters in the transition equation. Note that similar circular priors may also be placed on other parameters, although flat priors are adopted for this analysis. The sampling scheme was run for a burnin of 15,000 iterates, after which the Markov chain is assumed to converge to the augmented likelihood. The subsequent 25,000 iterates form the Monte Carlo sample from which inference is computed.

The estimated impact of demand for electricity on the first moment of the spot price using the flexible functional form is plotted for 12 equally-spaced half hours in Figure 2. The relationships are largely monotonic and kinked. The location where the curvature changes can be understood as the maximum level of demand which will be matched by the supply of efficient baseline generation at each given half hour. Estimation of the location of the change of curvature is important for risk management because increases in demand beyond this level correspond to steep increases in the expected spot price.

This nonlinear relationship is consistent with economic theory. The instantaneous demand curve is largely insensitive to price changes because the increased cost of electricity is born by the retailing utility in the short-run, not the end-user. However,

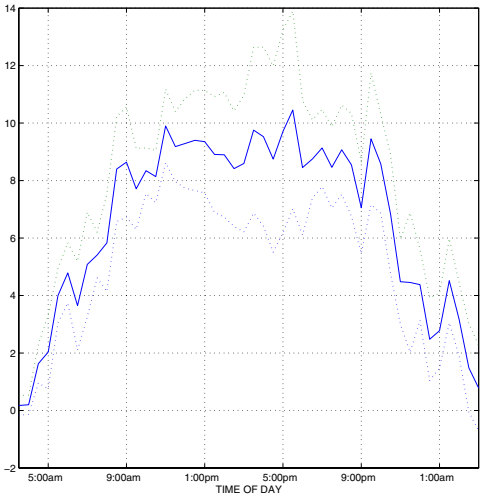


Fig. 3 Estimates of the linear coefficients of the effect of demand on price volatility against time of day. The 90% posterior probability intervals for the linear coefficients are also plotted as light dotted lines.

demand varies day-to-day at any given half hour period based on season, day type, weather and other factors. Therefore, as demand varies over the three months of data the equilibrium price effectively traces out the kinked supply curve for each half hour, resulting in the relationship between demand and price depicted in Figure 2.

Figure 3 plots the estimated linear coefficients of demand in the transition equation against the time-of-day, along with 90% posterior probability intervals. This shows that the log-volatility of price is also highly affected by changes in demand, but in a manner that differs over the day. Price volatility is much more sensitive to changes in demand during the high-load period 09:00 to 23:00.

The estimates of the posterior means $E[C|y]$ and $E[B|y]$ are presented in Figure 4. The absolute values of the elements are plotted for ease of exposition, although almost every element was positive. (Note that because C^{-1} and B^{-1} are banded, this does not mean that C and B are sparse). There is strong residual intraday autocorrelation in both the level and log-volatility of prices. However, in the second moment the correlation is close to block diagonal, with three distinctly correlated periods of the day: 09:00 to 12:00 (morning work-hours), 13:00 to 18:00 (afternoon work-hours) and 18:00 to 23:30 (evening). A similar block structure for intraday dependence was found by Panagiotelis & Smith (2008) using a related model and more recent NSW data. Guthrie & Videbeck (2007) also found a similar block pattern using New Zealand data. Disappointingly, the estimates of ϕ show little evidence of residual inter-day autocorrelation in the level of prices, but there is significant, albeit minor, inter-day autocorrelation ψ in the mean-corrected log-volatilities. However, when the model is refit fixing $C = B = I$ (thereby removing the intraday serial dependence) the inter-day autocorrelation becomes substantial in both the observation and transition equations. Similarly, when the model is refit with no exogenous demand effect, both intraday and inter-day autocorrelation become substantial in both equations. The day type dummy variables only have a minor affect on the level of prices, which is not

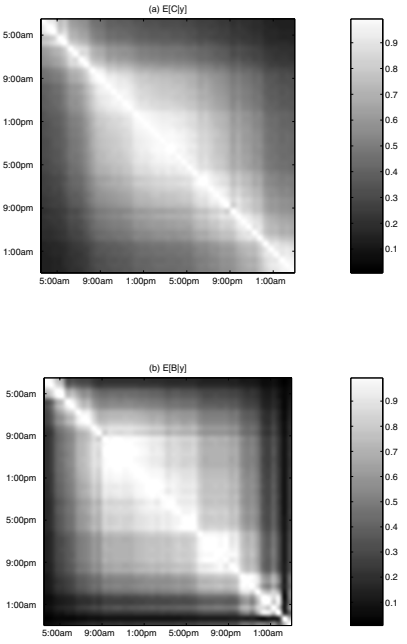


Fig. 4 Image plots of the Monte Carlo estimates of $E[C|y]$ in panel (a) and $E[B|y]$ in panel (b). For expositional purposes the absolute values of the elements are plotted, although there were almost no negative elements.

surprising given that demand is accounted for directly. However, while not presented here, there is a substantial day type effect in the second moment. For a full exposition of the empirical results, see Smith & Cottet (2006).

Figure 5(a) plots the Monte Carlo estimate of $E[\sigma_{ii}|y] = E[\exp(\delta_i)|y]$ against the time-of-day. To demonstrate the impact of the circular pairwise shrinkage prior the posterior means are also plotted when estimated assuming a flat prior on δ, so that $p(D) \propto \prod_{i=1}^{m} 1/\sigma_{ii}$. Figure 5(b) shows the point estimates of the log-volatility process $\{h_{it}\}$ for $i = 25$ (that is, at 15:30) for the two different priors on δ. The circular prior stabilises the variance of the mean-corrected latent volatility process. Figure 5(c) plots the posterior distribution of τ_δ^2. It demonstrates that the $IG(1.01, 0.01)$ hyper-prior for τ_δ^2 does not dominate the likelihood in determining the optimal level of shrinkage.

Figure 6(a) plots on the logarithmic scale the fitted values, which are defined as $\hat{y}_{it} = E[y_{it}|\Pi = \hat{\Pi}]$, where $\hat{\Pi}$ is the Monte Carlo estimate of $E[\Pi|y]$. The model proves flexible enough to enable prices to be recovered well during the earlier low demand period (days 1-28), where fixed price baseline generation often sets the marginal price, and also at the latter period (days 29-84) where during much of the day more expensive generation capacity is being dispatched to meet demand, so that the marginal price is significantly higher. A comparison with the observed price data in Figure 1(a) show the strong intraday pattern to the signal is captured effectively by the periodic model. Figure 6(b) plots the estimated posterior mean of the log-volatilities in prices $E[h_{it}|y]$. The model appears to capture the complex variance process well, including the change in the variance between the earlier and

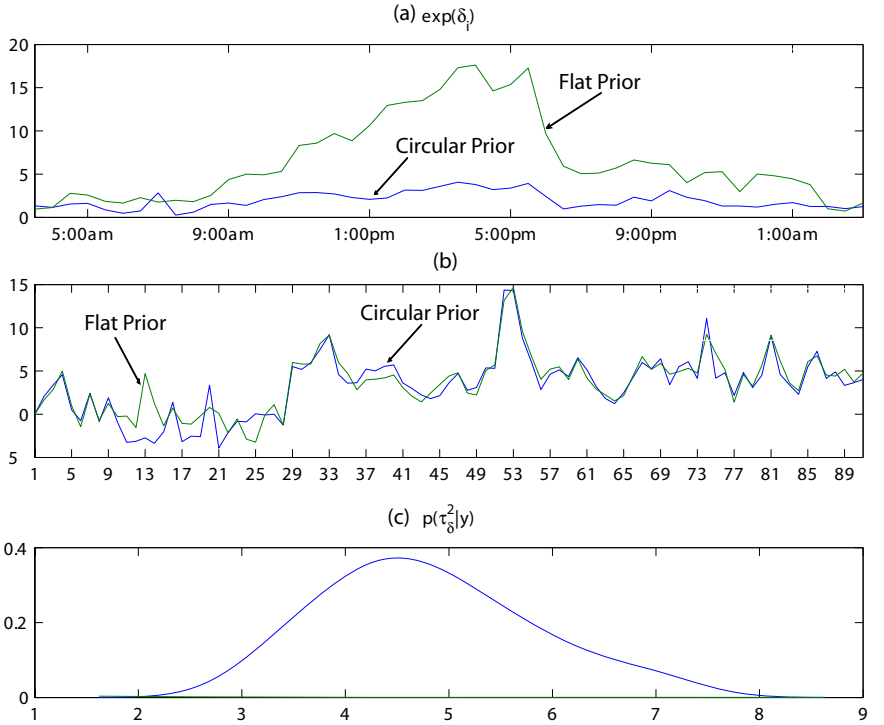


Fig. 5 Panel (a) plots Monte Carlo estimates of $E[\sigma_{it}|y] = E[\exp(\delta_i)|y]$ against intraday period $i = 1, \ldots, 48$ for the flat and circular priors on δ. Panel (b) plots Monte Carlo estimates of $E[h_{it}|y]$ for $i = 25$ (that is, at 15:30) against day $t = 1, \ldots, T$ for the two priors. Panel (c) plots the estimated posterior distribution $p(\tau_\delta^2|y)$ when the IG(1.01,0.01) hyperprior is employed for τ_δ^2.

latter periods. The data contain substantial price spikes during days 43, 52, 53, 54, 60, 74 and 81. These are captured by the model as substantial simultaneous increases in both the first and second moments of prices.

Figure 7(a) plots the estimated predictive mean of prices $E[y^f|y, y_0]$ over the subsequent seven days, along with the actual price. Panel (b) contains an estimate of the standard deviation of prices given by $\hat{\sigma}^f = \exp(\hat{h}^f/2)$, where $\hat{h}^f$ is the Monte Carlo estimate of $E[h^f|y, y_0]$. The forecasts confirm that there is a significant signal in both moments, which when captured allows for a degree of forecastability of the spot price.

6 Discussion

The PSV model can be used to model data that exhibit both serial correlation in the second moment and periodicity. Potential fields of application include the envi-

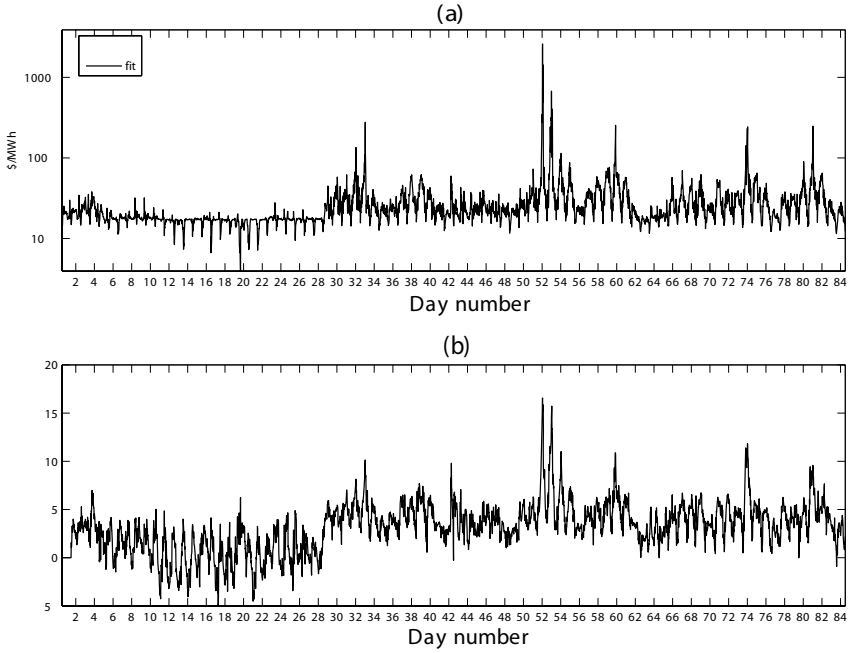


Fig. 6 PSV fit to the summer 99/00 NSW price data. Panel (a) plots the Monte Carlo fitted values $\hat{y}_{it} = E[y_{it}|\Pi = \hat{\Pi}]$ for prices on the log scale. Panel (b) plots the Monte Carlo estimated posterior means of the log-volatilities $E[h_{it}|y]$.

ronmental sciences and economics, where in the latter much macroeconomic data exhibits seasonal heteroscedasticity. The modelling and forecasting of electricity is an important application where time series models that combine a strong periodic structure with serial dependence in the first and second moments are required; see, for example, the discussion in Koopman et al. (2007). When m is large, such as for the half-hourly data examined here, approximating Φ as a diagonal, or sparse triangular matrix as in Panagiotelis & Smith (2008), substantially reduces the computations required in Steps (1) and (4) of the scheme. The band structures for the precision matrices arise from sparse lag structures in the underlying PARs. The reparameterisation of the correlation matrices in Section 3.2 allows for these band structures to be imposed simply. Estimation using MCMC allows for the computation of the full predictive distribution of prices over a horizon, which in the electricity application proves important.

Acknowledgements This work was partially supported by the Australian Research Council Discovery Project DP0985505 'Bayesian Inference for Flexible Parametric Multivariate Econometric Modelling'. The empirical results are drawn from an earlier unpublished manuscript by Smith & Cottet (2006) with the permission of Remy Cottet. The author would like to thank Mohsen Pourahmadi for drawing his attention to the PAR model.

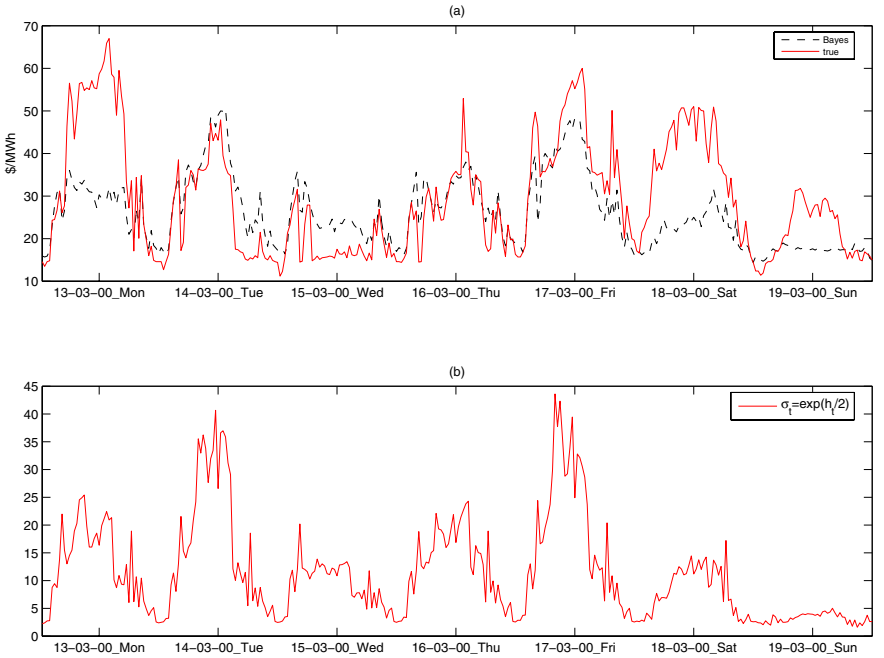


Fig. 7 Panel (a) plots the Monte Carlo estimated predictive mean $E[y^f|y, y_0]$ (dashed line) along with observed price (solid line) over a forecast horizon of seven days. Panel (b) plots the corresponding standard deviations $\hat{\sigma}^f = \exp(\hat{h}^f/2)$, where $\hat{h}^f = E[h^f|y, y_0]$ is the Monte Carlo estimated predictive mean of the future log-volatility vector.

References

Asai, M., McAleer, M. & Yu, J., (2006). Multivariate Stochastic Volatility, *Econometric Reviews*, **25**(2-3): 145–175.

Barlow, M., (2002). A diffusion model for electricity prices, *Mathematical Finance*, **12**: 287–298.

Barnard, J., McCulloch, R. & Meng, X. (2000). Modelling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage, *Statistica Sinica*, **10**: 1281–1311.

Broszkiewicz-Suwaj, A., Makagon, A., Weron, R. & Wylomanska, A. (2004). On detecting and modeling periodic correlation in financial data, *Physica A*, **336**: 196–205.

Chan, D., Kohn, R. & Kirby, C. (2006). Multivariate Stochastic Volatility Models with Correlated Errors, *Econometric Reviews*, **25**(2): 245–274.

Chib, S., Nardari, F. & Shephard, N. (2002). Markov chain Monte Carlo Methods for Stochastic Volatility Models, *Journal of Econometrics*, **108**: 281–316.

Chib, S., Nardari, F. & Shephard, N., (2006). Analysis of high-dimensional multivariate stochastic volatility models, *Journal of Econometrics*, **134**(2): 341–371.

Chib, S., Omori, Y. & Asai, M. (2009). Multivariate Stochastic Volatility, in Andersen, T., Davis, R., Kreiss, J.-P. & Mikosch, T. (eds). *Handbook of Financial time Series*, Springer: Berlin.

Conejo, A., Contreras, J., Espinola, R. & Plazas, M. (2005). Forecasting electricity prices for a day-ahead pool-based electric energy market, *International Journal of Forecasting*, **21**: 435–462.

Cottet, R. & Smith. M., (2003). Bayesian modeling and forecasting of intraday electricity load , *Journal of the American Statistical Association*, **98**: 839–849.

Dhrymes, P. (2000). *Mathematics for Econometrics*, Springer: New York.

Escribano, A., Pena, J. & Villaplana, P. (2002). Modeling electricity prices: International evidence, Working Paper, 02-27, Universidad Carlos III de Madrid.

Franses, P. H. & Paap, R. (2004). *Periodic Time Series Models: Advanced Texts in Econometrics*, OUP.

Guthrie, G. & Videbeck, S. (2007). Electricity spot price dynamics: Beyond financial models, *Energy Policy*, **35**: 5614–5621.

Haldrup, N., & Nielsen, M. (2006). A regime switching long memory model for electricity prices. *Journal of Econometrics*, **135**: 349–376.

Hamilton, J. D. (1994) *Time Series Analysis*, Princeton University Press: New Jersey.

Jacquier, E., Polson, N. & Rossi, P. (1994). Bayesian analysis of stochastic volatility models (with discussion), *Journal of Business and Economic Statistics*, **12**: 371–417.

Karakatsani, N. & Bunn, D. (2008). Intra-day and regime-switching dynamics in electricity price formation. *Energy Economics*, **30**: 1776–1797.

Kim, S., Shephard, N. & Chib, S. (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models, *Review of Economic Studies*, **68**, 361–339.

Knittel, C. & Roberts, M., (2005), 'An empirical examination of restructured electricity prices'. *Energy Economics*, 27, 791-817.

Koopman, S., Ooms, M. & Carnero, M. (2007). Periodic Seasonal Reg-ARFIMA-GARCH Models for Daily Electricity Spot Prices, *Journal of the American Statistical Association*, **102**(477): 16–27.

Lang, S., & Brezger, A. (2004) Bayesian P-Splines, *Journal of Computational and Graphical Statistics*, **13**: 183–212.

Osborn, D. & Smith, J. (1989). The Performance of Periodic Autoregressive Models in Forecasting Seasonal U.K. Consumption, *Journal of Business and Economic Statistics*, **7**(1): 117–127.

Pagano, M. (1978). On Periodic and Multiple Autoregressions, *The Annals of Statistics*, **6**(6): 1310–1317.

Panagiotelis, A., & Smith, M. (2008). Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions, *International Journal of Forecasting*, **24**: 710–727.

Pitt, M. & Shephard, N. (1999). Time-varying covariances: A factor stochastic volatility approach, in J. Bernardo, J. Berger, A. Dawid & A. Smith (eds.), *Bayesian Statistics 6*, OUP: 547–570.

Pourahmadi, M., (1999). Joint mean-covariance models with applications to longitudinal data: Uncontrained parameterisation, *Biometrika*, **86**: 677–669.

Shephard, N. (ed.) (2005). *Stochastic Volatility: Selected Readings: Advanced Texts in Econometrics*, OUP.

Shephard, N. & Pitt, M. (1997). Likelihood analysis of non-gaussian measurement time series, *Biometrika*, **84**: 653–668.

Smith, M. & Cottet, R. (2006). Estimation of a Longitudinal Multivariate Stochastic Volatility Model for the Analysis of Intra-Day Electricity Prices. Unpublished Working Paper ECMT2006-1: Discipline of Econometrics and Business Statistics, University of Sydney. *(Available from first author upon request.)*

Smith, M. & Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data, *Journal of the American Statistical Association*, **97**: 1141–1153.

Smith, M. & Pitts, A., (2006). Foreign Exchange Intervention by the Bank of Japan: Bayesian Analysis using a Bivariate Stochastic Volatility Model, *Econometric Reviews*, **23**(2-3): 425–451.

Soares, L. J. & Medeiros, M. C. (2008). Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data, *International Journal of Forecasting*, **24**: 630–644.

Tsiakas, I. (2006). Periodic Stochastic Volatility and Fat Tails, *Journal of Financial Econometrics*, **4**(1): 90–135.

Weron, R. & Misiorek, A. (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International Journal of Forecasting*, **24**: 744–762.

Appendix

This appendix outlines the evaluation of the posteriors in the sampling scheme in Section 4.1. A detailed exposition can also be found in Smith & Cottet (2006) when Φ and Ψ are strictly diagonal.

(1) Generating from $p(\phi_{i,j}|\{\Pi\backslash\phi_{i,j}\},y,h,y_0)$
The conditional posterior distribution

$$p(\phi_{i,j}|\{\Pi\backslash\phi_{i,j}\},y,h,y_0) \propto p(y,h|\Pi,y_0)p(\phi)$$
$$\propto \exp\left(-\frac{1}{2}\left[\sum_{t=1}^{T}(e_t^* - H_t^{-\frac{1}{2}}\Phi e_{t-1})'C^{-1}(e_t^* - H_t^{-\frac{1}{2}}\Phi e_{t-1})\right]\right) I(a_{ij} < \phi_{i,j} < b_{ij}),$$

where $e_t^* = H_t^{-1/2}e_t$, for $t = 1,...,T$, and $p(\phi_{i,j}|\{\Phi\backslash\phi_{i,j}\}) \propto I(a_{ij} < \phi_{i,j} < b_{ij})$, with a_{ij} and b_{ij} functions of $\{\Phi\backslash\phi_{i,j}\}$. Expanding the term in the exponent gives a quadratic function for $\phi_{i,j}$, so that the distribution is recognisable as constrained Gaussian, which can be sampled via rejection sampling. When Φ is diagonal, $\phi_{i,i}$ are simply constrained to the interval $(-1,1)$. Because $C^{-1} = \{c^{ij}\}$ is a band k_1 matrix, $c^{ij} = 0$ if $|i-j| > k_1$ which substantially speeds computation of the mean and variances.

(2) Generating from $p(r_{C,ij}|\{\Pi\backslash r_{C,ij}\},y,h,y_0)$
First note that $r_{C,ii} = 1$, for $i = 1,\ldots,m$, and that the lower triangular elements of R_C are zero. Then, the conditional distribution of the non-fixed elements of R_C is

$$p(r_{C,ij}|\{\Pi\backslash r_{C,ij}\},h,y,y_0) \propto |C^{-1}|^{T/2}\exp\left\{\frac{-1}{2}\mathrm{tr}C^{-1}M\right\}, \tag{8}$$

where $M = \sum_{t=1}^{T} H_t^{-1/2}(e_t - \Phi e_{t-1})(e_t - \Phi e_{t-1})'H_t^{-1/2}$ and C^{-1} is a function of $r_{C,ij}$ because

$$C^{-1} = \mathrm{diag}(R_C^{-1}R_C^{-1\prime})^{1/2}R_C'R_C\mathrm{diag}(R_C^{-1}R_C^{-1\prime})^{1/2}.$$

To generate $r_{C,ij}$ a Metropolis-Hastings step is used. The proposal density $q(r_{C,ij})$ is Gaussian centred at the mode of the density in equation (8) with variance equal to the inverse of the negative Hessian, which is obtained through numerical methods. (To implement this step the routine 'e04lyf' from the NAG fortran library was used.) The candidate $r_{C,ij}^{\mathrm{new}}$ is accepted over the old value $r_{C,ij}^{\mathrm{old}}$ with probability

$$\min\left\{1, \frac{p(r_{C,ij}^{\mathrm{new}}|\{\Pi\backslash r_{ij}\},h,y,y_0)q(r_{C,ij}^{\mathrm{old}})}{p(r_{C,ij}^{\mathrm{old}}|\{\Pi\backslash r_{ij}\},h,y,y_0)q(r_{C,ij}^{\mathrm{new}})}\right\}.$$

The value of C^{-1} is then computed directly from the resulting iterate of R_C.

(3) Generating from $p(h_{b,t}|h_{\backslash t},\Pi,y,y_0)$
For ease of exposition it is outlined how to generate the full vector h_t, but note that generation of a subvector $h_{b,t}$ is straightforward as it only involves a subset of the

computations outlined. The conditional density of the log-volatility h_t is calculated separately for three different values of t.

(i) For $1 < t < T$,

$$p(h_t|h_{\setminus t},\Pi,y,y_0) = p(y_t|h_t,y_{t-1},\Pi)p(h_t|h_{t-1},\Pi)p(h_{t+1}|h_t,\Pi)$$
$$\propto \exp\left\{-\frac{1}{2}\left[1'h_t+\eta_t'\Sigma^{-1}\eta_t+u_t'C^{-1}u_t+\eta_{t+1}'\Sigma^{-1}\eta_{t+1}\right]\right\},$$

where 1 is a vector of ones.

(ii) For $t = 1$,

$$p(h_1|h_{\setminus 1},\Pi,y,y_0) = p(y_1|h_1,y_0,\Pi)p(h_1|\Pi)p(h_2|h_1,\Pi)$$
$$\propto \exp\left\{-\frac{1}{2}\left[1'h_1+(h_1-Z_1\alpha)'\Gamma^{-1}(h_1-Z_1\alpha)+u_1'C^{-1}u_1+\eta_2'\Sigma^{-1}\eta_2\right]\right\}.$$

(iii) For $t = T$,

$$p(h_T|h_{\setminus T},\Pi,y,y_0) = p(y_T|h_T,y_{T-1},\Pi)p(h_T|h_{T-1},\Pi)$$
$$\propto \exp\left\{-\frac{1}{2}\left[1'h_T+\eta_T'\Sigma^{-1}\eta_T+u_T'C^{-1}u_T\right]\right\}.$$

Let $l(h_t) = \log\left\{p(h_t|h_{\setminus t},\Pi,h,y,y_0)\right\}$, then a normal approximation is taken to the density with mean equal to the mode of $l(h_t)$ and covariance matrix $\left[-\frac{\partial^2 l(h_t)}{\partial h_t \partial h_t'}\right]^{-1}$. The candidate h_t^{new} is accepted over the old value h_t^{old} with probability

$$\min\left\{1, \frac{p(h_t^{\text{new}}|h_{\setminus t},\Pi,y,y_0)q(h_t^{\text{old}})}{p(h_t^{\text{old}}|h_{\setminus t},\Pi,y,y_0)q(h_t^{\text{new}})}\right\}.$$

The mode of l is found using Newton-Raphson with the analytical derivatives, which are evaluated below. These are substantially faster to compute and more accurate than numerical derivatives. Their use significantly improves the efficiency of the sampler. Similar derivations can be found in Chan et al. (2006) and Smith & Pitts (2006, Appendix B) for different multivariate stochastic volatility models. First note that:

$$\frac{\partial u_t}{\partial h_t'} = -\frac{1}{2}\text{diag}\,(u_t)\,,\ \frac{\partial \eta_t}{\partial h_t'} = I \text{ and } \frac{\partial \eta_{t+1}}{\partial h_t'} = -\Psi\,.$$

For $1 < t < T$,

$$\frac{\partial l(h_t)}{\partial h_t} = -\frac{1}{2}\left[1+2\frac{\partial \eta_t'}{\partial h_t}\Sigma^{-1}\eta_t+2\frac{\partial u_t'}{\partial h_t}C^{-1}u_t+2\frac{\partial \eta_{t+1}'}{\partial h_t}\Sigma^{-1}\eta_{t+1}+2\frac{\partial u_{t+1}'}{\partial h_t}C^{-1}u_{t+1}\right]$$
$$= -\frac{1}{2}[1-u_t\odot C^{-1}u_t+2\Sigma^{-1}(h_t-Z_t\alpha-\Psi\xi_{t-1})-2\Psi\Sigma^{-1}(\xi_{t+1}-\Psi(h_t-Z_t\alpha))]\,.$$

For $t = 1$,

$$\frac{\partial l(h_1)}{\partial h_1} = -\frac{1}{2}\left[1 - u_1 \odot C^{-1}u_1 + 2\Gamma^{-1}(h_1 - Z_1\alpha) - 2\Psi\Sigma^{-1}(\xi_2 - \Psi(h_1 - Z_1\alpha))\right] .$$

For $t = T$,

$$\frac{\partial l(h_T)}{\partial h_T} = -\frac{1}{2}\left[1 - u_T \odot C^{-1}u_T + 2\Sigma^{-1}(h_T - Z_T\alpha - \Psi\xi_{T-1})\right] .$$

Here, $\odot$ is the element-by-element matrix product. The following result from Dhrymes (2000, p.153) is used to compute the second derivatives. If $y = x'Ax$, then

$$\frac{\partial^2 y}{\partial z \partial z'} = 2\frac{\partial x}{\partial z'}A\frac{\partial x}{\partial z} + (2x'A \otimes I)\frac{\partial^2 x}{\partial z \partial z'}$$

and that

$$\frac{\partial^2 \eta_t}{\partial h_t \partial h_t'} = 0 \, , \, \frac{\partial^2 \eta_{t+1}}{\partial h_t \partial h_t'} = 0 \text{ and } (u_t'C^{-1} \otimes I)\frac{\partial^2 u_t}{\partial h_t \partial h_t'} = \frac{1}{4}\text{diag}\left(u_t \odot C^{-1}u_t\right) .$$

For $1 < t < T$,

$$\begin{aligned}\frac{\partial^2 l(h_t)}{\partial h_t \partial h_t'} &= -\left[\frac{\partial \eta_t'}{\partial h_t}\Sigma^{-1}\frac{\partial \eta_t}{\partial h_t'} + \frac{\partial u_t'}{\partial h_t}C^{-1}\frac{\partial u_t}{\partial h_t'} + \frac{1}{2}\text{diag}\left(u_t'C^{-1} \odot u_t\right)\right] \\ &= -\left[\frac{1}{4}\text{diag}\left(u_t\right)C^{-1}\text{diag}\left(u_t\right) + \frac{1}{4}\text{diag}\left(u_t'C^{-1} \odot u_t\right) + \Sigma^{-1} + \Psi\Sigma^{-1}\Psi'\right] .\end{aligned}$$

For $t = 1$,

$$\frac{\partial^2 l(h_1)}{\partial h_1 \partial h_1'} = -\left[\frac{1}{4}\text{diag}\left(u_1\right)C^{-1}\text{diag}\left(u_1\right) + \frac{1}{4}\text{diag}\left(u_1'C^{-1} \odot u_1\right) + \Gamma^{-1} + \Psi\Sigma^{-1}\Psi'\right] .$$

For $t = T$,

$$\frac{\partial^2 l(h_T)}{\partial h_T \partial h_T'} = -\left[\frac{1}{4}\text{diag}\left(u_T\right)C^{-1}\text{diag}\left(u_T\right) + \frac{1}{4}\text{diag}\left(u_T'C^{-1} \odot u_T\right) + \Sigma^{-1}\right] .$$

(4) Generating from $p(\psi_{i,j}|\{\Pi \backslash \psi_{i,j}\}, h, y, y_0)$
The conditional posterior distribution of each non-zero autoregressive parameter in the transition equation is

$$p(\psi_{i,j}|\{\Pi \backslash \psi_{i,j}\}, h, y, y_0) \propto p(h_1|\Pi)\prod_{t=2}^{T} p(h_t|h_{t-1}, \Pi)p(\psi)$$

which is not a recognizable due to the term $p(h_1|\Pi)$. Therefore a Metropolis-Hastings step is employed with candidate generated from the approximation:

$$q(\psi_{i,j}) \propto \prod_{t=2}^{T} p(h_t|h_{t-1},\Pi)p(\psi)$$
$$\propto \exp\left(-\frac{1}{2}\left[\sum_{t=2}^{T}(\xi_t-\Psi\xi_{t-1})'\Sigma^{-1}(\xi_t-\Psi\xi_{t-1})\right]\right) I(a_{ij}<\psi_{i,j}<b_{ij}).$$

Expanding the term in the exponent gives a quadratic function for $\psi_{i,j}$, so that the distribution is Gaussian, constrained to (a_{ij},b_{ij}), where a_{ij} and b_{ij} are functions of $\{\Psi\backslash\psi_{i,j}\}$. When Ψ is diagonal $a_{ii}=-1$ and $b_{ii}=1$, so that generation is straightforward. Otherwise, the unconstrained Gaussian can be used for q, but results in a higher rejection rate. Exploiting the band structure of Σ^{-1} substantially speeds the computation of the mean and variance. A new iterate $\psi_{i,j}^{\text{new}}$ is then accepted over the old $\psi_{i,j}^{\text{old}}$ with probability

$$\min\left\{1,\frac{p(h_1|\{\Pi\backslash\psi_{i,j}\},\psi_{i,j}^{\text{new}})}{p(h_1|\{\Pi\backslash\psi_{i,j}\},\psi_{i,j}^{\text{old}})}\right\}.$$

(5) Generating from $p(r_{B,ij}|\{\Pi\backslash r_{B,ij}\},h,y,y_0)$
Note that $r_{B,ii}=1$ for $i=1,\ldots,m$, and that for $0<j-i<k_2$

$$p(r_{B,ij}|\{\Pi\backslash r_{B,ij}\},h,y,y_0) \propto p(h_1|\Pi)\prod_{t=2}^{T}p(h_t|h_{t-1},\Pi)$$
$$\propto |\Gamma|^{-1/2}|B^{-1}|^{(T-1)/2}\exp\left\{-\frac{1}{2}\left(\xi_1'\Gamma^{-1}\xi_1+\sum_{t=2}^{T}(D^{-1/2}\eta_t)'B^{-1}(D^{-1/2}\eta_t)\right)\right\},$$

where $B^{-1}=\text{diag}(R_B^{-1}R_B^{-1\prime})^{1/2}R_B'R_B\text{diag}(R_B^{-1}R_B^{-1\prime})^{1/2}$ and Γ are both highly nonlinear functions of $r_{B,ij}$. Metropolis-Hastings is used to generate from this univariate density with a normal approximation around its mode as a proposal, evaluated using numerical derivatives.

(6a) Generating from $p(\delta_i|\{\Pi\backslash\delta_i\},\tau_\delta^2,h,y,y_0)$
The density

$$p(\delta_i|\{\Pi\backslash\delta_i\},\tau_\delta^2,h,y,y_0) \propto p(h_1|\Pi)\prod_{t=2}^{T}p(h_t|h_{t-1},\Pi)\prod_{k=2}^{m}p(\delta_k|\delta_{k-1},\tau_\delta^2)p(\delta_1|\delta_m,\tau_\delta^2)$$
$$\propto |\Gamma|^{-1/2}\exp\left\{-\frac{1}{2}\left[\eta_1'\Gamma^{-1}\eta_1+\sum_{t=2}^{T}\left(e^{-\delta_i}\eta_{it}^2b^{ii}+2e^{-\delta_i/2}\sum_{j\in\mathscr{C}}\eta_{it}\eta_{jt}b^{ij}e^{-\delta_j/2}\right)\right]\right\}$$
$$\times\exp\left\{-\frac{\delta_i(T-1)}{2}-\frac{S_i}{2\tau_\delta^2}\right\},$$

where $\eta_t=(\eta_{1t},\ldots,\eta_{mt})'$, $\mathscr{C}=\{j|i-k_2\le j\le i+k_2 \text{ and } j\ne i\}$, the matrix $B^{-1}=\{b^{ij}\}$ and

$$S_i = \begin{cases} (\delta_i - \delta_{i-1})^2 + (\delta_{i+1} - \delta_i)^2 & \text{if } 1 < i < m-1 \\ (\delta_1 - \delta_m)^2 + (\delta_2 - \delta_1)^2 & \text{if } i = 1 \\ (\delta_m - \delta_{m-1})^2 + (\delta_1 - \delta_m)^2 & \text{if } i = m\,. \end{cases}$$

This is approximated with a normal density, centred around the mode of the posterior. The mode is obtained using quasi-Newton Raphson with numerical first and second derivatives. An iterate is then generated from the proposal, and then accepted or rejected in a Metropolis-Hastings step.

(6b): Generating from $p(\tau_\delta^2|\delta)$

Assuming an informative IG(a,b) prior, the posterior is $\tau_\delta^2 \sim \text{IG}(\frac{m}{2}+a, \frac{\delta' W \delta}{2}+b)$.

(7) Generating from $p(\beta|\{\Pi\backslash\beta\},h,y,y_0)$

By transforming the data $y_t^* = H_t^{-1/2}(y_t - \Phi y_{t-1})$ and $X_t^* = H_t^{-1/2}(X_t - \Phi X_{t-1})$ for $t = 1,...,T$, the model is then a seemingly unrelated regression model with error covariance matrix C. The conditional density is then a multivariate normal denisty, from which it is easy to generate. Note that because C^{-1} is band k_1, then the posterior precision matrix for β is block diagonal.

(8) Generating from $p(\alpha|\{\Pi\backslash\alpha\},h,y,y_0)$

To generate α a Metropolis-Hastings step is employed. The proposal is based on the augmented likelihood without omitting the term $p(h_1|\Pi)$, which is recognisable as a multivariate normal density in α. Note that because Σ^{-1} is banded, the posterior precision matrix is block diagonal.

Online Change-Point Detection in Categorical Time Series

Michael Höhle

Abstract This contribution considers the monitoring of change-points in categorical time series. In its simplest form these can be binomial or beta-binomial time series modeled by logistic regression or generalized additive models for location, scale and shape. The aim of the monitoring is to online detect a structural change in the intercept of the expectation model based on a cumulative sum approach known from statistical process control. This is then extended to change-point detection in multicategorical regression models such as multinomial or cumulative logit models. Furthermore, a Markov chain based method is given for the approximate computation of the run-length distribution of the proposed CUSUM detectors. The proposed methods are illustrated using three categorical time series representing meat inspection at a Danish abattoir, monitoring the age of varicella cases at a pediatrist and an analysis of German Bundesliga teams by a Bradley-Terry model.

1 Introduction

In the year 2000 and as part of my Ph.D. project, I had the pleasurable experience of getting hold of a copy of Fahrmeir & Tutz (1994b) in my attempt of modeling a multivariate binomial time series of disease treatments in a pig farm. After some enquiries, I ended up implementing the extended Kalman filter approach described in Fahrmeir & Wagenpfeil (1997) and in Section 8.3 of Fahrmeir & Tutz (1994b). With the present contribution I take the opportunity to return to this problem from another point of view while at the same time honoring the work of Ludwig Fahrmeir.

Michael Höhle
Department of Statistics, Ludwig-Maximilians-Universität München, 80539 München, Germany, and Munich Center of Health Sciences, Munich, Germany,
URL: www.stat.uni-muenchen.de/~hoehle,
e-mail: michael.hoehle@stat.uni-muenchen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_20,

Specifically, the focus in this chapter is on monitoring time series with categorical regression models by statistical process control (SPC) methods.

A general introduction to SPC can be found in Montgomery (2005). Hawkins & Olwell (1998) give an in-depth analysis of the CUSUM chart, which is one commonly used SPC method. Detection based on regression charts with normal response can be found in the statistics and engineering literature (Brown et al. 1975, Kim & Siegmund 1989, Basseville & Nikiforov 1998, Lai 1995, Lai & Shan 1999). Generalized linear models based detectors are described in the literature for especially count data time series (Rossi et al. 1999, Skinner et al. 2003, Rogerson & Yamada 2004, Höhle & Paul 2008). For categorical time series, however, less development has been seen – with monitoring of a binomial proportion being the exception (Chen 1978, Reynolds & Stoumbos 2000, Steiner et al. 2000). Retrospective monitoring of multinomial sequences is discussed in Wolfe & Chen (1990). Prospective monitoring of multivariate discrete response variable imposes a great challenge.

The present work contains a novel adaptation of the likelihood ratio based cumulative sum (CUSUM) for the categorical regression context. Accompanying this CUSUM is a newly formulated approximate Markov chain approach for calculating its run-length distribution. Three examples are presented as illustration of the proposed categorical CUSUM: Meat inspection data from a Danish abattoir monitored by a beta-binomial regression model, disease surveillance by a multinomial logit model for the age distribution of varicella cases at a sentinel pediatrist, and finally – in honour of Fahrmeir & Tutz (1994a) – an analysis of paired comparison data for six teams playing in the best German national soccer league (1. Bundesliga). Fahrmeir & Tutz (1994a) analyzed the 1966/67–1986/87 seasons of this example using state-space methodology for categorical time series. My contribution continues their analysis up to the 2008/09 season with a special focus on change-point detection.

The structure of this chapter is as follows. Section 2 provides an introduction to modeling categorical time series while Section 3 contains the novel proposals for performing online change-point detection in such models. Application of the proposed methodology is given in Section 4. Section 5 closes the chapter with a discussion.

2 Modeling Categorical Time Series

Modeling categorical data using appropriate regression models is covered in Agresti (2002) or Fahrmeir & Tutz (1994b). The interest of this chapter lies in using such regression approaches for the modeling of time series with categorical response. Kedem & Fokianos (2002) and also Fahrmeir & Tutz (1994b) provide an introduction to this topic. A *categorical time series* is a time series where the response variable at each time point t takes on *one* of $k \geq 2$ possible categories. Let $\boldsymbol{X}_t = (X_{t1}, \ldots, X_{tk})'$ be a length k vector with $X_{tj}, j = 1, \ldots, k$, being one if the j'th category is observed

at time t and zero otherwise. Consequently, $\sum_{j=1}^{k} X_{tj} = 1$. Assuming that a total n_t of such variables are observed at time t, define $\boldsymbol{Y}_t = \sum_{l=1}^{n_t} \boldsymbol{X}_{t,l}$ as the response of interest. Furthermore, assume that the distribution of $\boldsymbol{Y}_t$ can adequately be described by a multinomial distribution with time series structure, i.e.

$$\boldsymbol{Y}_t \sim \mathrm{M}_k(n_t, \boldsymbol{\pi}_t), \tag{1}$$

for $t = 1,2,\ldots$, $\boldsymbol{\pi}_t = (\pi_{t1},\ldots,\pi_{tk})'$ and $\sum_{j=1}^{k} \pi_{tj} = 1$ for all t. Here $\pi_{tj} = P(Y_t = j|\mathscr{F}_{t-1})$ is the probability for class j at time t and $\mathscr{F}_{t-1}$ denotes the history of the time series up to time $t-1$, i.e. just before but not including time t. When considering a single component $j \in \{1,\ldots,k\}$ of a multinomial distributed $\boldsymbol{Y}_t$, the resulting distribution of Y_{tj} is $\mathrm{Bin}(n_t, \pi_{tj})$. As a consequence, one strategy to describe a multinomial time series is to consider it as a set of independent binomial time series for each component. However, this ignores any correlations between the variables and does not provide a model with total probability 1.

2.1 Binomial and Beta-Binomial Data

The simplest form of categorical data is the case $k = 2$, which describes individuals experiencing an event or items as being faulty. In this case, the resulting distribution of Y_{t1} in (1) is $\mathrm{Bin}(n_t, \pi_{t1})$ while $Y_{t2} = n_t - Y_{t1}$. When modeling binomial data, interest is often in having an additional overdispersion not provided by the multinomial distribution. A parametric tool for such time series is the use of the beta-binomial distribution, i.e. $Y_t \sim \mathrm{BetaBin}(n_t, \pi_t, \sigma_t)$, where $t = 1,2,\ldots$, $0 < \pi_t < 1$ and $\sigma_t > 0$, and having probability mass function (PMF)

$$f(y_t|n_t,\pi_t,\sigma_t) = \left(\frac{\Gamma(n_t+1)\Gamma(y_t+1)}{\Gamma(n_t-y_t+1)}\right) \cdot \left(\frac{\Gamma(y_t+\frac{\pi_t}{\sigma_t}) \cdot \Gamma(\frac{1}{\sigma_t}) \cdot \Gamma(n_t+\frac{1-\pi_t}{\sigma_t}-y_t)}{\Gamma(n_t+\frac{1}{\sigma_t}) \cdot \Gamma(\frac{\pi_t}{\sigma_t}) \cdot \Gamma(\frac{1-\pi_t}{\sigma_t})}\right)$$

mean $E(Y_t) = n_t \cdot \pi_t$ and variance

$$\mathrm{Var}(Y_t) = n_t\pi_t(1-\pi_t)\left(1+(n_t-1)\frac{\sigma_t}{\sigma_t+1}\right).$$

In other words, σ_t is the dispersion parameter and for $\sigma_t \to 0$ the beta-binomial converges to the binomial distribution. Beta-binomial models can be formulated and fitted in the context of generalized additive models for location, scale and shape (GAMLSS, Rigby & Stasinopoulos (2005)). Here, the time varying proportion π_t is modeled by a linear predictor η_t on the logit-scale similar to binomial logit-modeling, i.e.

$$\mathrm{logit}(\pi_t) = \log\left(\frac{\pi_t}{1-\pi_t}\right) = \eta_t = \boldsymbol{z}_t'\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{z}_t$ is a $p \times 1$ vector of covariates and $\boldsymbol{\beta}$ is a $p \times 1$ vector of covariate effects. Additionally, in a GAMLSS the dispersion can be modeled by a separate linear predictor $\log(\sigma_t) = \boldsymbol{w}_t'\boldsymbol{\gamma}$, but for notational and computational simplicity the dispersion is assumed to be time constant and not depending on covariates, i.e. $\sigma_t = \sigma$ for all t.

2.2 Nominal Data

In case the k groups of the response variable lack a natural ordering, i.e. in case of a nominal time series, one uses a multinomial logistic model with one of the categories, say category k, as reference:

$$\log\left(\frac{\pi_{tj}}{\pi_{tk}}\right) = \boldsymbol{z}_t'\boldsymbol{\beta}_j, \quad j = 1,\ldots,k-1.$$

As a result, the category specific probabilities can be computed as

$$\pi_{tj} = \frac{\exp(\boldsymbol{z}_t'\boldsymbol{\beta}_j)}{1+\exp(\boldsymbol{z}_t'\boldsymbol{\beta}_j)}, \quad j = 1,\ldots,k-1, \text{ and}$$
$$\pi_{tk} = \frac{1}{1+\sum_{j=1}^{k-1}\exp(\boldsymbol{z}_t'\boldsymbol{\beta}_j)}.$$

Let $\boldsymbol{y}_{1:N} = (\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N)$ denote the observed time series up to time N given as a $(m \times N)$ matrix, where each $\boldsymbol{y}_t = (y_{t1},\ldots,y_{tk})', t = 1,\ldots,N$ contains information on how the n_t observations fell into the k categories, i.e. $\sum_{j=1}^{k} y_{tj} = n_t$. The likelihood of the above model is given by

$$L(\boldsymbol{\beta};\boldsymbol{y}_{1:N}) = \prod_{t=1}^{N}\prod_{j=1}^{k}\pi_{tj}^{y_{tj}}(\boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_{k-1}')'$. Statistical inference for the model parameters $\boldsymbol{\beta}$ based on this likelihood is described in detail in Fahrmeir & Tutz (1994b, Section 3.4) or Fokianos & Kedem (2003). Asymptotics for such categorical time series is studied in Kaufman (1987) and Fahrmeir & Kaufmann (1987).

2.3 Ordinal Data

If the k categories of the response variable can be considered as ordered, it is beneficial to exploit this additional information in order to obtain more parsimonious models. Denoting the categories of the ordered response variable by the ordered set $\{1,\ldots,k\}$, a *cumulative model* described in, e.g., Fahrmeir & Tutz (1994b, p. 76) for the response at time t looks as follows

$$P(Y_t \leq j) = F(\theta_j + \boldsymbol{z}_t'\boldsymbol{\beta}), \quad j = 1, \ldots, k,$$

with $-\infty = \theta_0 < \theta_1 < \ldots < \theta_k = \infty$ being the set of *threshold parameters*. When using the logistic distribution function $F(x) = \exp(x)/(1+\exp(x))$, the resulting model is called the *proportional odds model*, but in use are also other link functions such as the extreme-minimal-value distribution function. Consequently, the specific category probabilities can be derived as

$$\pi_{tj} = F(\theta_j + \boldsymbol{z}_t'\boldsymbol{\beta}) - F(\theta_{j-1} + \boldsymbol{z}_t'\boldsymbol{\beta}), \quad j = 1, \ldots, k.$$

2.4 Paired Comparisons

One application of the proportional odds model is the analysis of paired-comparison data used to determine preference or strength of items. Such data are typical in sports like chess or tennis, where world rankings of m players are based on pairwise comparisons having categorical outcomes (e.g. win, loose). Other areas of application are consumer preference, sensory studies and studies of animal behavior (Courcoux & Semenou 1997, Bi 2006, Whiting et al. 2006). The basic Bradley-Terry model (Bradley & Terry 1952) is a logistic regression model quantifying the probability of a positive outcome (i.e. winning) for the first mentioned player in a match of two players. Each player $i \in \{1, \ldots, m\}$ has ability or strength $\alpha_i \in \mathbb{R}$, and the probability that a match between the i'th and j'th player results in a win for player i is given by

$$\text{logit}\{P(Y_{ij} = 1)\} = \alpha_i - \alpha_j.$$

In the above, Y_{ij} is a binary random variable with states 1 (i wins) and 2 (i looses). As a consequence, $P(Y_{ij} = 1) = 1/2$ if $\alpha_i = \alpha_j$ and $P(Y_{ij} = 1) > 1/2$ if $\alpha_i > \alpha_j$. To ensure identifiably, one has to impose a constraint such as $\alpha_m = 0$ or $\sum_{i=1}^m \alpha_i = 0$ on the $\alpha's$. Extensions of the Bradley-Terry model consist of letting strength be given by additional covariates such as home court advantages, age or injuries (Agresti 2002). Another common extension is to handle additional tied outcomes or even more complicated ordinal response structure (Tutz 1986).

If the time interval over which the paired-comparisons are performed is long, one might expect the abilities of players to change over time (Fahrmeir & Tutz 1994b, Glickman 1999, Knorr-Held 2000). Following Fahrmeir & Tutz (1994a), a general time-dependent ordinal paired-comparison model including covariates can be formulated as

$$P(Y_{tij} = r) = F(\theta_{tr} + \alpha_{ti} - \alpha_{tj} + \boldsymbol{z}_{tij}'\boldsymbol{\beta}_t) - F(\theta_{t,r-1} + \alpha_{ti} - \alpha_{tj} + \boldsymbol{z}_{tij}'\boldsymbol{\beta}_t), \qquad (3)$$

with $r = 1, \ldots, k$ being the category, $t = 1, 2, \ldots$ denoting time and $i, j = 1, \ldots, m$ being the players compared. For example in the application of Section 4.3, Y_{tij} will denote paired comparisons of six teams within each season of the best German national

soccer league (1. Bundesliga). In what follows, I will assume time constant covariate effects $\boldsymbol{\beta}_t = \boldsymbol{\beta}$ for all t and similar time constant thresholds $\boldsymbol{\theta}_t = (\theta_{t0}, \ldots, \theta_{tk})' = \boldsymbol{\theta} = (\theta_0, \ldots, \theta_k)'$ for all t.

After having presented the basic modeling techniques, the focus is now on the online detection of changepoints in such models.

3 Prospective CUSUM Changepoint Detection

The cumulative sum (CUSUM) detector is a method known from statistical process control for online detecting structural changes in time series. An overview of the method can be found in Hawkins & Olwell (1998). Use of the method for count, binomial or multicategorical time series using regression models is still a developing field. Höhle & Paul (2008) treats one such approach for count data and Grigg & Farewell (2004) provide an overview. For multicategorical time series Topalidou & Psarakis (2009) contains a survey of existing monitoring approaches. My interest is in monitoring a time varying vector of proportions $\boldsymbol{\pi}_t$ in a binomial, beta-binomial or multinomial setting having time-varying n_t. Regression models for categorical time series provide a versatile modeling framework for such data allowing for time trends with seasonality and possible covariate effects. Sections 3.1–3.3 contain my proposal for combining CUSUM detection with categorical time series analysis.

Let $f(\boldsymbol{y}_t; \boldsymbol{\theta})$ denote the PMF of the response variable at time t. While new observations arrive, the aim is to detect as quickly as possible if the parameters of f have changed from the in-control value of $\boldsymbol{\theta}_0$ to the out-of-control value $\boldsymbol{\theta}_1$. Following Frisén (2003), define the *likelihood ratio based CUSUM statistic* as

$$C_s = \max_{1 \leq t \leq s} \left[\sum_{i=t}^{s} \log \left\{ \frac{f(\boldsymbol{y}_i; \boldsymbol{\theta}_1)}{f(\boldsymbol{y}_i; \boldsymbol{\theta}_0)} \right\} \right], \quad s = 1, 2, \ldots. \tag{4}$$

Given a fixed threshold $h > 0$, a change-point is detected at the first time s where $C_s > h$, and hence the resulting stopping time S is defined as

$$S = \min\{s \geq 1 : C_s > h\}. \tag{5}$$

At this time point, enough evidence is found to reject $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ in favor of $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$. Let now $LLR_t = \log f(y_t; \boldsymbol{\theta}_1) - \log f(y_t; \boldsymbol{\theta}_0)$ be shorthand for the loglikelihood ratio at time t in (4). If $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are known, (4) can be written in recursive form

$$C_0 = 0 \quad \text{and} \quad C_s = \max(0, C_{s-1} + LLR_t), \quad \text{for } s \geq 1. \tag{6}$$

One sees that for time points with $LLR_t > 0$, i.e. evidence against in-control, the LLR_t contributions are added up. On the other hand, no credit in the direction of the in-control is given because C_s cannot get below zero.

In practical applications, the in-control and out-of-control parameters are, however, hardly ever known beforehand. A typical procedure in this case is to use his-

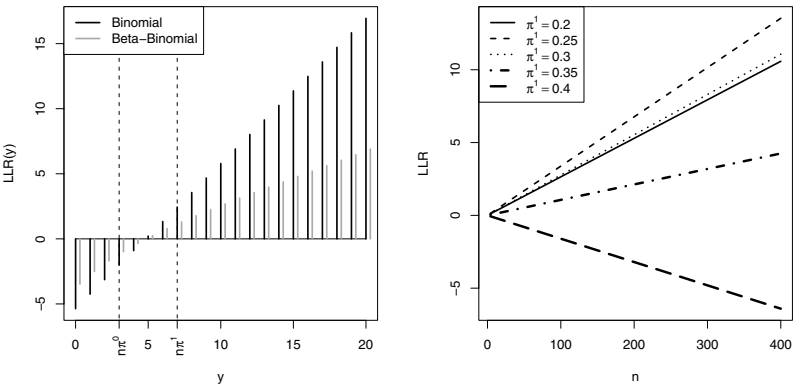


Fig. 1 (a) Loglikelihood ratio (LLR) as a function of y for a binomial distribution with $n = 20$ and $\pi^0 = 0.15$ and $\pi^1 = 0.35$. Also shown are the same LLRs for the corresponding beta-binomial distribution with $\sigma = 0.05$. (b) Binomial LLR as a function of n when $y = 0.25 \cdot n$, $\pi^0 = 0.15$ and when comparing against four different π^1.

torical *phase 1 data* for the estimation of $\boldsymbol{\theta}_0$ with the assumption that these data originate from the in-control state. This estimate is then used as plug-in value in the above CUSUM. Furthermore, the out-of-control parameter $\boldsymbol{\theta}_1$ is specified as a known function of $\boldsymbol{\theta}_0$, e.g. as a known multiplicative increase in the odds. Using categorical regression to model the PMF f as a function of time provides a novel use of statistical process control for monitoring categorical time series. Sections 3.1–3.3 discuss monitoring in case of beta-binomial, multinomial and ordered response. Section 3.4 contains a corresponding method to compute the important run-length distribution of the different CUSUM proposals.

3.1 Binomial and Beta-Binomial CUSUM

Extending the work of Steiner et al. (2000) to a time varying proportion, the aim is to detect a change from odds $\pi_t^0/(1-\pi_t^0)$ to odds $R \cdot \pi_t^0/(1-\pi_t^0)$ for $R > 0$, i.e. let

$$\operatorname{logit}(\pi_t^1) = \operatorname{logit}(\pi_t^0) + \log R. \tag{7}$$

In other words, let $\operatorname{logit}(\pi_t^1) = \operatorname{logit}(\pi_t^0) + \log R$ correspond to such a change in the intercept of the linear predictor in (2). The change-point detection is thus equivalent to a detection from the in-control proportion π_t^0 to the out-of-control proportion π_t^1 in (6) using the beta-binomial PMF as f.

Figure 1(a) illustrates the LLR as a function of the number of positive responses in a binomial distribution for one specific time point (note that t is dropped from the

notation in this example). Starting from $y = 5$ one has $LLR > 0$, i.e. observations with $y \geq 5$ contribute evidence against the null-hypothesis and in favor of the alternative hypothesis. Note also, that the beta-binomial distribution has smaller LLR contributions because the variance of the distribution is larger than for the binomial distribution. Similarly, Figure 1(b) shows that the larger n the larger is the LLR contribution of the observation $y = 0.25 \cdot n$. In other words, the greater n is the more evidence against $H_0 : \pi = 0.15$ there is from an empirical proportion of 0.25. This is of interest in a binomial CUSUM with time varying n_t: the relevance (as measured by its contribution to C_t) of a large proportion of faulty items thus depends on the number of items sampled. However, for $\pi^1 = 0.4$ the value $y = 0.25 \cdot n$ does not provide evidence against H_0 in Figure 1(b). This means that for large out-of-control proportions the observation $y = 0.25 \cdot n$ results in negative LLRs and hence speaks in favor of H_0.

At time t and given the past value of the CUSUM statistic C_{t-1}, the minimum number of cases necessary to reach the threshold h at time t is

$$a_t = \min_{y \in \{0,\ldots,n_t\}} \left\{ \text{LLR}(y; n_t, \pi_t^0, \pi_t^1, \sigma) > h - C_{t-1} \right\}. \tag{8}$$

Note that the set of y fulfilling the above inequality can be empty, in this case a_t does not exists. If a_t exists, the solution of (8) can be derived explicitly for the binomial case as

$$a_t = \max \left\{ 0, \left\lceil \frac{h - C_{t-1} - n_t \cdot (\log(1-\pi_t^1) - \log(1-\pi_t^0))}{\log(\pi_t^1) - \log(\pi_t^0) - \log(1-\pi_t^1) + \log(1-\pi_t^0)} \right\rceil \right\}.$$

In the beta-binomial case the solution has to be found numerically, e.g. by trying possible $y \in \{0, \ldots, n_t\}$ until the first value fulfills the inequality.

3.2 Multinomial CUSUM

This section looks at generalization of the previous binomial CUSUM to the multinomial distribution $M_k(n_t, \boldsymbol{\pi}_t)$ for $k > 2$, and where $\boldsymbol{\pi}_t$ is modeled by multinomial logistic regression. Let $\boldsymbol{\pi}_t^0$ be the in-control probability vector and $\boldsymbol{\pi}_t^1$ the out-of-control probability vector resulting from the models with parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. A simple approach would be to monitor each of the k components separately using the methodology from Section 3.1. However, this would ignore correlations between the measurements with reduced detection power as consequence. Instead, I consider detection as the task of investigating change-points in the linear predictors of the multicategorical logit model. The proposed approach extends the work of Steiner et al. (1999), who monitored surgical performance of a $k = 4$ outcome using two paired binomial CUSUMs with time-constant means.

Based on a multicategorical logit model, let the in-control probabilities for the non-reference categories be

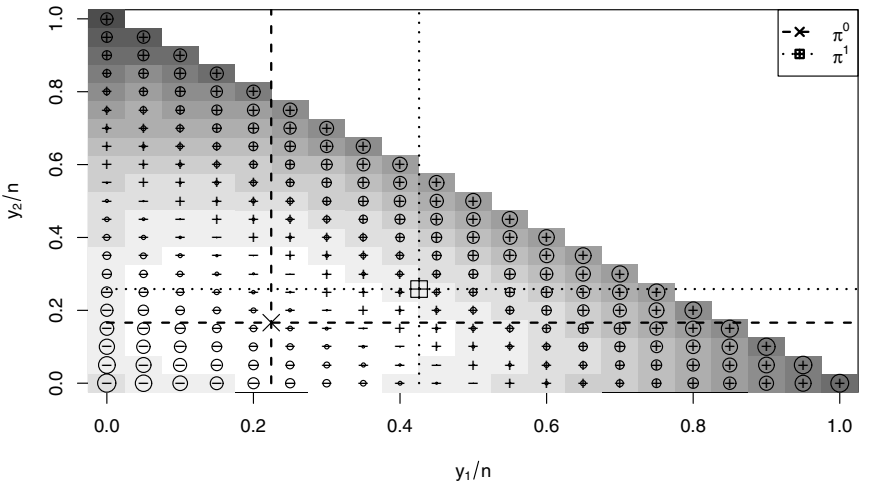


Fig. 2 Illustration of the LLR for a $M_3(20, \boldsymbol{\pi})$ multinomial CUSUM with $\boldsymbol{\pi}^0 = (0.22, 0.17, 0.61)'$ and $\boldsymbol{\pi}^1 = (0.43, 0.26, 0.32)'$. Shown are the first two components y_1 and y_2 of each possible state $\boldsymbol{y}$. Circle sizes indicate magnitude and $\pm$ the sign of the LLR. Also shown are the in-control and out-of-control probabilities. Shading indicates the probability of $\boldsymbol{y}$ in a model with $\boldsymbol{\pi} = \boldsymbol{\pi}^0$ – the whiter the cell the higher is the probability of the corresponding state.

$$\log\left(\frac{\pi^0_{tj}}{\pi^0_{tk}}\right) = \boldsymbol{z}_t'\boldsymbol{\beta}_j, \quad j = 1, \ldots, k-1.$$

As for the binomial CUSUM, the out-of-control probabilities are given by specific changes in the intercept of this model, i.e.

$$\log\left(\frac{\pi^1_{tj}}{\pi^1_{tk}}\right) = \log\left(\frac{\pi^0_{tj}}{\pi^0_{tk}}\right) + \log(R_j), \quad j = 1, \ldots, k-1.$$

Figure 2 illustrates the approach for a $\boldsymbol{Y} \sim M_3(20, \boldsymbol{\pi}^0)$ distribution with $\boldsymbol{\pi}^0 = (0.22, 0.17, 0.61)'$ and $\log(\boldsymbol{R}) = (1.30, 1.10)'$. One observes that many states with high LLR are concurrently very unlikely and that for larger n or k, the approximating multivariate Gaussian distribution can be used to determine states with high enough probability to investigate its LLR.

If the number of possible categories k of the multinomial is very high, log-linear models provide an alternative as done by Qiu (2008). However, in his work time-constant problems are dealt with and the prime goal is to detect a shift in the median of any component without a specific formulation of the alternative. However, a suitable extension of the proposed monitoring in this chapter might be to monitor against an entire set of possible out-of-control models with the different $\boldsymbol{R}$'s specifying different directions.

3.3 Ordinal and Bradley-Terry CUSUM

The multinomial CUSUM proposal from the previous section can be used as a change-point detection approach for ordinal time series: Based on the proportional odds model to generate the in-control and out-of-control proportions. In particular, this approach is considered for the time varying Bradley-Terry model (3) from Section 2.4. Let $\boldsymbol{Y}_t = (Y_{tij}; i = 1,\ldots,m, j = 1,\ldots,m, i \neq j)$ consist of all $K = m \times (m-1)$ paired comparisons occurring at time t, i.e. $\boldsymbol{Y}_t \in \{1,\ldots,k\}^K$. Given the parameters of a time varying Bradley-Terry model, the probability of a state $\boldsymbol{Y}_t = \boldsymbol{y}_t$ can thus be computed as

$$f(\boldsymbol{y}_t; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{m} \prod_{j=1, i\neq j}^{m} f(y_{tij}; \boldsymbol{\alpha}_t, \boldsymbol{\beta}, \boldsymbol{\theta}),$$

where $f(\cdot)$ denotes the PMF given in (3). The interest is now on detecting a structural change in the ability of one or several teams, i.e. $\boldsymbol{\alpha}_t^1 = \boldsymbol{\alpha}_t^0 + \boldsymbol{R}$, where $\boldsymbol{R}$ is a vector of length m with for example one component being different from zero. The LLR in a corresponding CUSUM detector can then be computed as

$$LLR_t = \sum_{i=1}^{m} \sum_{j=1, i\neq j}^{m} \log \frac{f(\boldsymbol{y}_{tij}; \boldsymbol{\alpha}_t^1, \boldsymbol{\beta}, \boldsymbol{\theta})}{f(\boldsymbol{y}_{tij}; \boldsymbol{\alpha}_t^0, \boldsymbol{\beta}, \boldsymbol{\theta})}. \tag{9}$$

3.4 Run-length of Time Varying Categorical CUSUM

The distribution of the stopping time S in (5) for the CUSUMs proposed in sections 3.1–3.3 when data are sampled from either $\boldsymbol{\pi}_t^0$ or $\boldsymbol{\pi}_t^1$ is an important quantity to know when choosing the appropriate threshold h. Specifically, the expected run length $E(S)$ (aka. the average run length (ARL)), the median run length or the probability $P(S \leq s)$ for a specific $s \geq 1$ are often used summaries of the distribution and can be computed once the PMF of S is known. Let $\boldsymbol{\theta}$ be the set of parameters in the multicategorical regression model and let $\boldsymbol{\pi}$ be the resulting proportions under which the distribution of S is to be computed. For example, the above $\boldsymbol{\theta}$ is equal to $\boldsymbol{\theta}_0$ if the in-control ARL is of interest.

Brook & Evans (1972) formulated an approximate approach based on Markov chains to determine the PMF of the stopping time S of a time-constant CUSUM detector. They describe the dynamics of the CUSUM statistic C_t by a Markov chain with a discretized state space of size $M+2$:

$$\begin{aligned}
&\text{State } 0: && C_t = 0\\
&\text{State } i: && C_t \in \left((i-1)\cdot\tfrac{h}{M}, i\cdot\tfrac{h}{M}\right], \; i = 1,2,\ldots,M\\
&\text{State } M+1: && C_t > h
\end{aligned}$$

Note that state $M+1$ is absorbing, i.e. reaching this state results in H_0 being rejected, and therefore no further actions are taken. The discretization of the continuum of values of the CUSUM statistic into a discrete set of states represents an approximation. The size of M controls the quality of the approximation. Adopting this approach to the present time-varying context, let $\boldsymbol{P}_t$ be the $(M+2)\times(M+2)$ transition matrix of $C_t|C_{t-1}$, i.e.

$$p_{tij} = P(C_t \in \text{State } j | C_{t-1} \in \text{State } i), \quad i,j = 0,1,\ldots,M+1$$

Let $a<b$ and $c<d$ represent the lower and upper limits of class j and i, respectively. To operationalize the Markov chain approach one needs to compute

$$p_{t,i,j} = P(a<C_t<b|c<C_{n-1}<d) = \int_c^d \{F_t(b-s)-F_t(a-s)\}\,d\mu(s), \quad (10)$$

where $\mu(x)$ is the unknown distribution function of C_{t-1} conditional on $c<C_{t-1}<d$ and $F_t(\cdot)$ is the distribution function of the likelihood ratio LLR_t at time t when $\boldsymbol{y}_t$ is distributed according to a multinomial distribution with parameters derived from a categorical regression model with parameters $\boldsymbol{\theta}$. Investigations in Hawkins (1992) for the homogeneous case suggest using the uniform distribution for measure $\mu(x)$. Furthermore, he suggests using Simpson's quadrature rule with midpoint $m=(c+d)/2$ to approximate the integral in (10) instead of the Riemann integral used in Brook & Evans (1972). Altogether, Hawkins (1992) adapted to the present time varying case yields

$$\begin{aligned} P(a<C_t<b|c<C_{n-1}<d) \approx{}& \frac{1}{6}\{F_t(b-c)+4F_t(b-m)+F_t(b-d)\} \\ &-\frac{1}{6}\{F_t(a-c)+4F_t(a-f)+F_t(a-d)\}. \end{aligned}$$

Specifically, $F_t(\cdot)$ can be computed for the categorical CUSUM by computing the likelihood ratio of all valid configurations $\boldsymbol{y}_t \in \{0,1,\ldots,n_t\}^k, \sum_{j=1}^k y_{tj} = n_t$, together with the probability $P(\boldsymbol{Y}_t = \boldsymbol{y}_t)$ of its occurrence under $\boldsymbol{\theta}$. However, if n_t or k is large, this enumeration strategy can quickly become infeasible and one would try to identify relevant states with $P(\boldsymbol{y})>\varepsilon$ and approximate $F_t(\cdot)$ by only considering these states in the computations. One strategy to perform this identification could be to compare with the approximating normal distribution.

Borrowing ideas from Bissell (1984), the cumulative probability of an alarm at any step up to time s, $s\geq 1$, is

$$P(S\leq s) = \left[\prod_{t=1}^{s}\boldsymbol{P}_t\right]_{0,M+1},$$

i.e. the required probability is equivalent to the probability of going from state zero at time one to the absorbing state at time s as determined by the s-step transition matrix of the Markov chain. The PMF of S can thus be determined by $P(S=s)=$

$P(S \leq s) - P(S \leq s-1)$, where for $s = 1$ one defines $P(S = 0) = 0$. Hence, $E(S)$ can be computed by the usual expression $\sum_{s=1}^{\infty} s \cdot P(S = s)$. In practice, one would usually compute $P(S \leq s)$ only up to some sufficiently large $s = s_{\max}$ such that $P(S \leq s) \geq 1 - \varepsilon$ for a small ε. This results in a slightly downward bias in the derived ARL. If the Markov chain is homogeneous, then the ARL can alternatively be computed as the first element of $(\boldsymbol{I} - \boldsymbol{R})^{-1} \cdot \mathbf{1}$, where $\boldsymbol{R}$ is obtained from $\boldsymbol{P}$ by deleting the last row and column, $\boldsymbol{I}$ is the identity matrix and $\mathbf{1}$ a vector of ones.

In practice, covariates or n_t are usually not available for future time points. As the predicted in-control and out-of-control probabilities are conditional on these values, it is more practicable to compute $P(S \leq s)$ for phase 2 data where the covariates already have been observed instead of trying to impute them for future time points.

4 Applications

The following three examples illustrate the use of the proposed CUSUM monitoring for categorical time series by applications from veterinary quality control, human epidemiology and – as continuation of Fahrmeir & Tutz (1994b) – sports statistics.

4.1 Meat Inspection

At Danish abattoirs, auditing is performed for each processed pig in order to provide guarantees of meat quality and hygiene and as part of the official control on products of animal origin intended for human consumption (regulated by the European Council Regulation No 854/2004). Figure 3 shows the time series of the weekly proportion of positive audit reports for a specific pig abattoir in Denmark. Reports for a total of 171 weeks are available with monitoring starting in week 1 of 2006.

Using the data of the first two years as phase 1 data, a beta-binomial model with intercept and two sinusoidal components for $\text{logit}(\pi_t)$ is estimated using the R function `gamlss` (Rigby & Stasinopoulos 2005). These estimated values are then used as plug-in values in the model to predict π_t^0 for phase 2. The out-of-control π_t^1 is then defined by specifying $R = 2$ in (7), i.e. a doubling in the odds of a positive audit report is to be detected as quickly as possible. Figure 4 shows the results from this monitoring. After the first change-point is detected the CUSUM statistic is set to zero and monitoring is restarted.

Figure 5 displays the run length distribution when using $h = 4$ by comparing the Markov chain approximation using $M = 5$ with the results of a simulation based on 10000 runs. Note that the Markov chain method provides results much faster than the simulation approach.

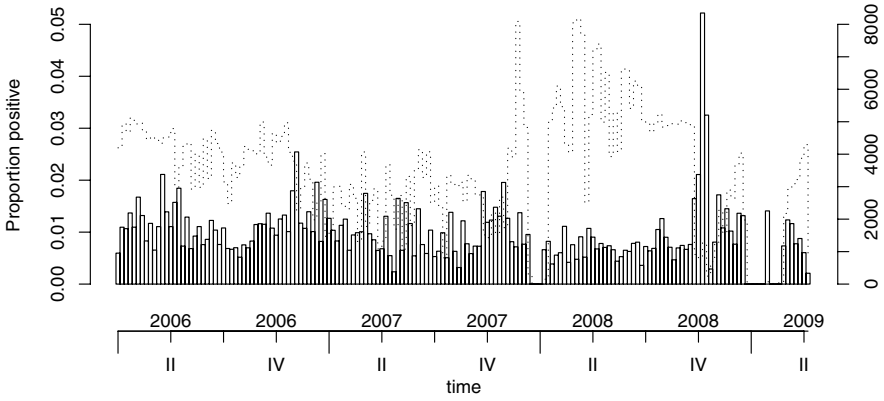


Fig. 3 Weekly proportion y_t/n_t of pigs with positive audit reports indicated by bars (scale on the left axis). The dotted line shows the weekly total number of pigs n_t (scale as on right axis). Roman letters denote quarters of the year.

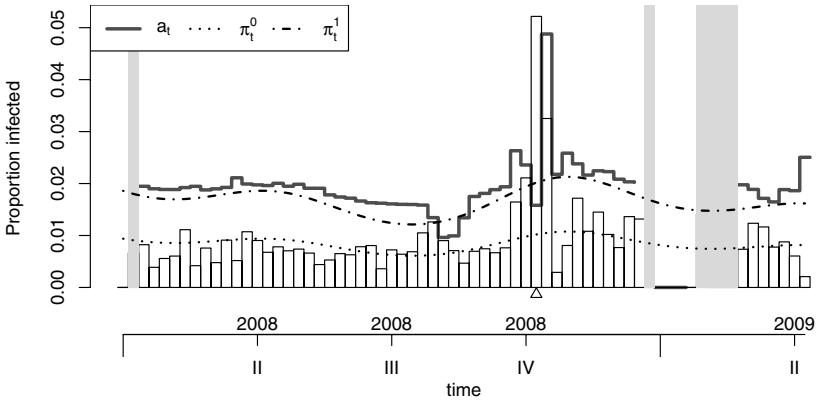


Fig. 4 Results of beta-binomial CUSUM monitoring for phase 2. Shaded bars indicate weeks where $n_t < 200$. The triangle indicates the alarm in week 41 of 2008.

4.2 Agegroups of Varicella Cases

A varicella sentinel was established in April 2005 by the *Arbeitsgemeinschaft Masern und Varizellen* (Robert Koch Institute 2006) to monitor a possible decline in the number of monthly varicella after the introduction of a vaccination recommendation. One particular point of interest is the monitoring of possible shifts in the age distribution of the cases. This is done by dividing the age of cases into one of five groups: <1, 1-2, 3-4, 5-9, and >9 years. A shift in the age distribution is now defined to be a structural change in the proportions $\boldsymbol{\pi}$ controlling which of the five age groups a

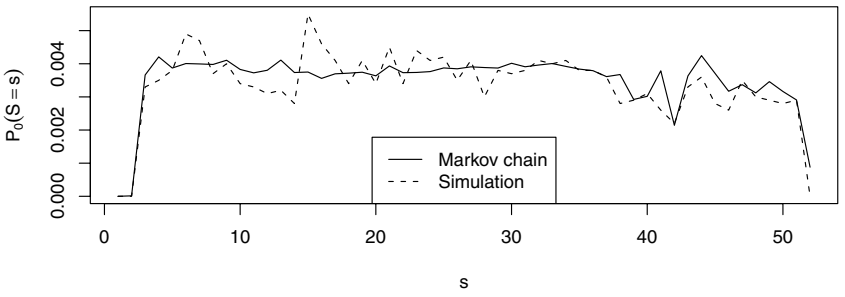


Fig. 5 Comparison of the in-control run-length PMF $P_0(S = s)$ between the Markov chain method and a simulation based on 10000 samples.

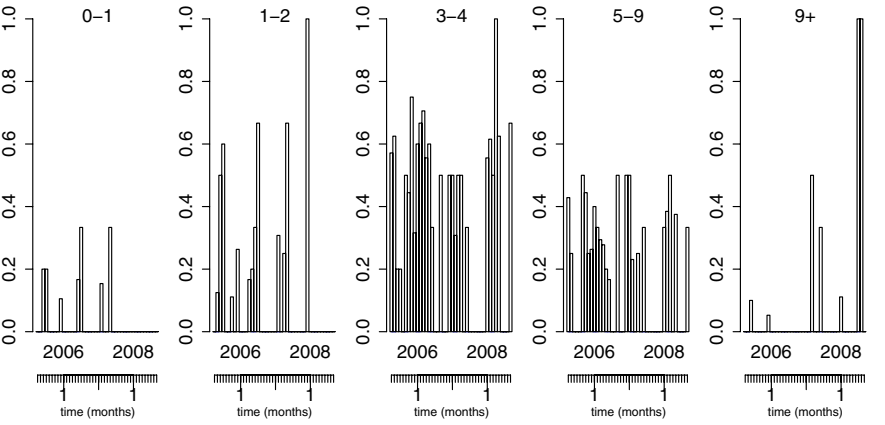


Fig. 6 Multinomial time series of monthly cases at a pediatrist participating in the varicella sentinel surveillance. The values of n_t range from 0 to 19.

case falls into. As proof of concept of the proposed methodology, the time series of a single pediatrist participating in the sentinel is considered. Figure 6 shows the time series of monthly proportions across the five age groups – note that summer vacations result in a seasonal pattern. Using the first 24 months as phase 1 data, a multinomial logistic model using intercept, linear time trend and two seasonal components is fitted by the R function `multinom` (Venables & Ripley 2002). Figure 7 shows the fitted model and the resulting in-control proportions for the five age groups for the subsequent 18 months.

Applying the proposed categorical CUSUM based on the multinomial PMF with the age group 1-2 acting as reference category, one obtains Figure 8. From an epidemiological point of view it is in the 1-2 age group where a decline of cases is expected because primarily this group is vaccinated. Detecting an increase in the remaining

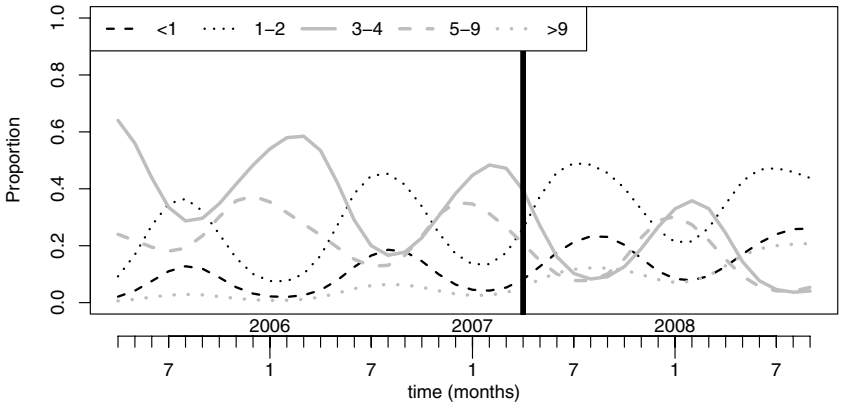


Fig. 7 Age group proportions obtained from fitting a multicategorical logit model to the observed phase 1 data (to the left of the vertical bar). Also shown are the resulting predictions for $\boldsymbol{\pi}^0$ during phase 2 (starting from the vertical bar).

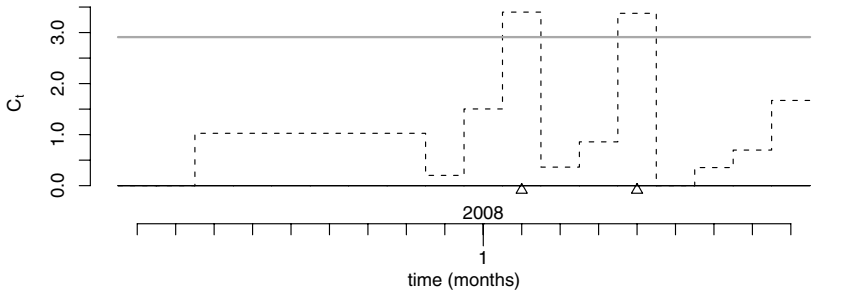


Fig. 8 CUSUM statistic C_t for the pediatrist data together with the threshold $h = 2.911$. Triangles indicate detected change-points.

four groups is one way to identify such a shift. As a consequence, $\log(\boldsymbol{R}) = (1,1,1,1)'$ is used. Figure 8 shows the resulting C_t together with the two detected change-points.

The threshold $h = 2.911$ is selected such that $P_0(S \leq 18) = 0.058$ as computed by the Markov chain approach with $M = 25$. By simulation of the run-length using 10000 runs, one obtains $P_0(S \leq 18) = 0.060$. To get an understanding of the consequences of currently ignored estimation error for the phase 1 parameters, a parametric bootstrap investigation is performed. Let $\widehat{\boldsymbol{\theta}}_0$ represent the estimated phase 1 parameters. In the b'th bootstrap sample, simulate new data phase 1 data $\boldsymbol{y}_{t,b}, t = 1,\ldots,24$ by sampling from a multinomial model with probabilities derived from $\hat{\boldsymbol{\theta}}_0$. Then use this $\boldsymbol{y}_{t,b}$ to estimate the phase 1 parameters $\widehat{\boldsymbol{\theta}}_{0,b}$ and derive $\boldsymbol{\pi}_b^0$ and $\boldsymbol{\pi}_b^1$ from $\widehat{\boldsymbol{\theta}}_{0,b}$ for phase 2. Now use the Markov chain procedure to compute $P_{0,b}(S \leq 18)$. A 95% percentile bootstrap

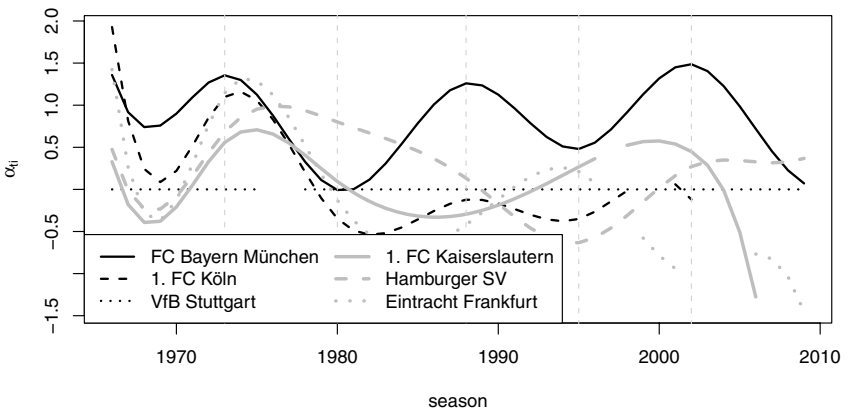


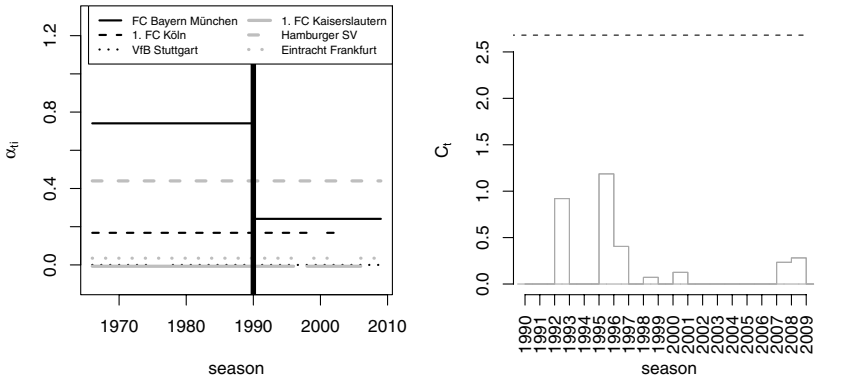
Fig. 9 Abilities α_{ti} of each team as fitted by a proportional odds model described in the text. Seasons without comparisons are indicated by not plotting the ability for this season. The thin shaded lines indicate the know locations.

interval for $P_0(S \leq 18)$ based on 100 bootstrap replications is $(0.013, 0.070)$, which emphasizes the effect of estimation error on the run length properties.

4.3 Strength of Bundesliga Teams

The time series analysed in this section contains the paired comparison data for a subset of six teams playing in the best German national soccer league (1. Bundesliga) as described in Fahrmeir & Tutz (1994a). For each of the 44 seasons from 1966/67–2008/09, all teams play against each other twice – once with the first team having home-court advantage and once with the second team having this advantage. Conceptually, it would have been feasible to perform the comparison based on all teams having played in the primary division since 1965/66, but I conduct the analysis in spirit of Fahrmeir & Tutz (1994a) by using only six teams. Each match has one of three possible outcomes: home team wins, tie and away team wins. In what follows, the ability of each team is assumed constant within the season but varies from season to season, i.e. α_{it} denotes the ability of team i in season $t = 1, \ldots, 44$.

Figure 9 shows the resulting abilities of each team as determined by a Bradley-Terry model fitted using the `vglm` function from package `VGAM` (Yee & Wild 1996, Yee 2008). The team VfB Stuttgart is selected as reference category with $\alpha_{3t} = 0$ for all t. For each team a time trend is modeled by a cubic B-spline with five equidistant interior knots and an intercept, i.e. $\alpha_{it} = f_i(t) = \beta_{i0} + \sum_{k=1}^{8} \beta_{ik} B_k(t)$. This model was found to be the model with equidistant knots minimizing Akaike's information criterion. Seasons where a team did not play in the first division are indicated in Figure 9 by missing abilities for that particular season.



(a) (b)

Fig. 10 (a) In-control abilities fitted from phase 1 data (before the vertical bar). Also shown are the out-of-control abilities for phase 2 (starting at the vertical bar) obtained by prediction from the phase 1 fitted model. (b) CUSUM statistics when monitoring using the in-control and out-of-control abilities from (a) for phase 2. The upper line shows the threshold h.

The estimated abilities only reflect strengths based on the six selected teams. Hence, they do not necessarily reflect the overall strength of the team that season, which explains for example the somewhat weak ability of 1. FC Kaiserslautern in the 1990/91 season where they won the cup. Fahrmeir & Tutz (1994a), with their state space approach also noted the drop for FC Bayern München around 1976-1980, which was due to Franz Beckenbauer leaving the club.

From a sports manager perspective, it could be of interest to online monitor the ability of a team for the purpose of performing strategic interventions. Applying the methodology from Section 2.4, consider the case of monitoring the ability of FC Bayern München starting from year 1990. A Bradley-Terry model with an intercept only is fitted to the data before 1990 and a change of $\boldsymbol{R} = (-0.5, 0, 0, 0, 0)'$ is to be detected for the abilities of all teams except the reference team. Figure 10 illustrates both the abilities obtained from fitting the phase 1 data and the resulting predicted out-of-control abilities for phase 2. The aim is to detect when the strength of FC Bayern München drops by 0.5 units compared to the average strength of 0.741 during the 1965/1966 to 1988/1989 seasons. This means that the probability of winning against VfB Stuttgart at home court drops from 0.742 to 0.636 ($\widehat{\theta_1} = 0.315$).

Using $h = 2.681$, Figure 10(b) shows the resulting C_t statistic of such CUSUM monitoring. No change-points are detected, but one notices the seasons with weaker performance as compared with Figure 9. Run-length computations are not immediately possible in this case as the determination of the distribution function of the LLR requires enumeration over $k^{m(m-1)} = 3^{30} = 2.06 \cdot 10^{14}$ states. As seen from (9), the LLR is a sum over the 30 possible paired-comparisons, i.e. it is the convolution of 30 independent but not identically distributed three-state variables. However,

with the specific value of $\boldsymbol{R}$, where the ability of only one team changes between in-control and out-of-control, only the $2(m-1) = 10$ matches involving FC Bayern München will have a non-zero contribution to the LLR. Hence, it is only necessary to investigate $3^{10} = 59049$ states.

Since the proposed in-control and out-of-control models are time-constant, the in-control ARL can be computed by inversion of the approximate CUSUM transition matrix based on $M = 25$. Using the specified $h = 2.681$ yields an ARL of 100.05. In other words, using $h = 2.681$ means that a structural change from $\boldsymbol{\alpha}^0$ to $\boldsymbol{\alpha}^1$ is detected by pure chance on average every 100.05'th season when the data generating mechanism is $\boldsymbol{\alpha}^0$.

5 Discussion

A likelihood ratio CUSUM method for the online changepoint detection in categorical time series was presented based on categorical regression models, such as the multinomial logit model and the proportional odds model. Altogether, the presented categorical CUSUM together with the proposed run-length computation provides a comprehensive and flexible tool for monitoring categorical data streams of very different nature.

The utilized time series modeling assumed that observations were independent given the time trend and other covariates of the model. This assumption could be relaxed using for example pair-likelihood approaches (Varin & Vidoni 2006) or autoregressive models (Fahrmeir & Kaufmann 1987). It would also be of interest to embed the change-point detection within the non-Gaussian state-space modeling for ordinal time series of Fahrmeir & Tutz (1994a).

The Markov chain approximation for deriving the run length distribution of the proposed CUSUM constitutes a versatile tool for the design of categorical CUSUMs. It also constitutes a much faster alternative to this problem than simulation approaches. Embedding the approach in a numerical search procedure could be useful when performing the reverse ARL computation: Given $\boldsymbol{\pi}^0$, ARL_0, ARL_1 and a direction $\boldsymbol{R}^*, ||\boldsymbol{R}^*|| = 1$, find the corresponding magnitude $c > 0$ such that the desired run-length results are obtained for $\boldsymbol{R} = c \cdot \boldsymbol{R}^*$. Currently, the distribution function of the likelihood ratio is calculated by investigating all possible states – an approach which for large k or n_t can become intractable. Section 4.3 showed that reductions for the number of states to investigate are possible in specific applications. Still, clever approximate strategies are subject to further research – for example by identifying a subset of most probable configurations. Finally, use of the Markov chain approximation is not limited to categorical time series – also the run length of time varying count data CUSUMs can be analyzed. For example, Höhle & Mazick (2009) consider CUSUM detectors for negative binomial time series models with fixed overdispersion parameter which could be analyzed by the proposed Markov chain approach.

Other approaches exist to perform retrospective and prospective monitoring based on regression models. For example the work in Zeileis & Hornik (2007) provides

a general framework for retrospective change-point detection based on fluctuation tests, which also finds prospective use. The method is, for example used in Strobl et al. (2009) to retrospectively assess parameter instability in Bradley-Terry models in a psychometric context. Instead of monitoring against a specific change, another alternative is to try to detect a general change based on model residuals. For this approach, the deviance statistic is an immediate likelihood ratio based alternative suitable for monitoring within the proposed categorical CUSUM framework.

An implementation of the methods is available as functions `categorical-CUSUM` and `LRCUSUM.runlength` in the R package `surveillance` (Höhle 2007) available from CRAN.

Acknowledgements I thank Niels Peter Baadsgaard, Danish Pig Production, Denmark, for providing the abattoir monitoring data; Anette Siedler, Robert Koch Institute, Germany, and the Arbeitsgemeinschaft Masern und Varizellen, Germany, for introducing me to the varicella vaccination data. I also thank Nora Fenske, Department of Statistics, Munich, Germany, for her proofreading of the manuscript.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edn, Wiley.

Basseville, M. & Nikiforov, I. (1998). *Detection of Abrupt Changes: Theory and Application*. Online version of the 1994 book published by Prentice-Hall, Inc. Available from http://www.irisa.fr/sisthem/kniga/.

Bi, J. (2006). *Sensory Discrimination Tests and Measurements: Statistical Principles, Procedures and Tables*, Wiley.

Bissell, A. F. (1984). The performance of control charts and cusums under linear trend, *Applied Statistics* **33**(2): 145–151.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. method of paired comparisons, *Biometrika* **39**(3/4): 324–345.

Brook, D. & Evans, D. A. (1972). An approach to the probability distribution of cusum run length, *Biometrika* **59**(3): 539–549.

Brown, R., Durbin, J. & Evans, J. (1975). Techniques for testing the constancy of regression relationships over time, *Journal of the Royal Statistical Society, Series B* **37**(2): 149–192.

Chen, R. (1978). A surveillance system for congenital malformations, *Journal of the American Statistical Association* **73**: 323–327.

Courcoux, P. & Semenou, M. (1997). Preference data analysis using a paired comparison model, *Food Quality and Preference* **8**(5–6): 353–358.

Fahrmeir, L. & Kaufmann, H. (1987). Regression models for nonstationary categorical time series, *Journal of time series Analysis* **8**: 147–160.

Fahrmeir, L. & Tutz, G. (1994a). Dynamic stochastic models for time-dependt ordered paired comparison systems, *Journal of the American Statistical Association* **89**(428): 1438–1449.

Fahrmeir, L. & Tutz, G. (1994b). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 1st edn, Springer.

Fahrmeir, L. & Wagenpfeil, S. (1997). Penalized likelihood estimation and iterative kalman smoothing for non-gaussian dynamic regression models, *Computational Statistics & Data Analysis* **24**(3): 295–320.

Fokianos, K. & Kedem, B. (2003). Regression theory for categorical time series, *Statistical Science* **18**(3): 357–376.

Frisén, M. (2003). Statistical surveillance: Optimality and methods, *International Statistical Review* **71**(2): 403–434.

Glickman, M. E. (1999). Estimation in large dynamic paired comparison experiments, *Journal of the Royal Statistical Society, Series C* **48**(3): 377–394.

Grigg, O. & Farewell, V. (2004). An overview of risk-adjusted charts, *Journal of the Royal Statistical Society, Series A* **167**(3): 523–539.

Hawkins, D. M. (1992). Evaluation of average run lengths of cumulative sum charts for an arbitrary data distribution, *Communications in Statistics. Simulation and Computation* **21**(4): 1001–1020.

Hawkins, D. M. & Olwell, D. H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*, Statistics for Engineering and Physical Science, Springer.

Höhle, M. (2007). surveillance: An R package for the monitoring of infectious diseases, *Computational Statistics* **22**(4): 571–582.

Höhle, M. & Mazick, A. (2009). Aberration detection in R illustrated by Danish mortality monitoring, *in* T. Kass-Hout & X. Zhang (eds), *Biosurveillance: A Health Protection Priority*, CRC Press. To appear.

Höhle, M. & Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series, *Computational Statistics & Data Analysis* **52**(9): 4357–4368.

Kaufman, H. (1987). Regression models for nonstationary categorical time series: Asmyptotic estimation theory, *Annals of Statistics* **15**: 79–98.

Kedem, B. & Fokianos, K. (2002). *Regression Models for Time Series Analysis*, Wiley.

Kim, H.-J. & Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression, *Biometrika* **76**(3): 409–423.

Knorr-Held, L. (2000). Dynamic rating of sports teams, *The Statistician* **49**(2): 261–276.

Lai, T. (1995). Sequential changepoint detection in quality control and dynamical systems, *Journal of the Royal Statistical Society, Series B* **57**: 613–658.

Lai, T. & Shan, J. (1999). Efficient recursive algorithms for detection of abrupt changes in signals and control systems, *IEEE Transactions on Automatic Control* **44**: 952–966.

Montgomery, D. C. D. (2005). *Introduction to Statistical Quality Control*, 5th edn, John Wiley.

Qiu, P. (2008). Distribution-free multivariate process control based on log-linear modeling, *IEE Transations* **40**(7): 664–677.

Reynolds, M. R. & Stoumbos, Z. G. (2000). A general approach to modeling CUSUM charts for a proportion, *IIE* **32**: 515–535.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion), *Applied Statistics* **54**: 1–38.

Robert Koch Institute (2006). Epidemiologisches Bulletin 33, Available from http://www.rki.de.

Rogerson, P. & Yamada, I. (2004). Approaches to syndromic surveillance when data consist of small regional counts, *Morbidity and Mortality Weekly Report* **53**: 79–85.

Rossi, G., Lampugnani, L. & Marchi, M. (1999). An approximate CUSUM procedure for surveillance of health events, *Statistics in Medicine* **18**: 2111–2122.

Skinner, K., Montgomery, D. & Runger, G. (2003). Process monitoring for multiple count data using generalized linear model-based control charts, *International Journal of Production Research* **41**(6): 1167–180.

Steiner, S. H., Cook, R. J. & Farewell, V. T. (1999). Monitoring paired binary surgical outcomes using cumulative sum charts, *Statistics in Medicine* **18**: 69–86.

Steiner, S. H., Cook, R. J., Farewell, V. T. & Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts, *Biostatistics* **1**(4): 441–452.

Strobl, C., Wickelmaier, F. & Zeileis, A. (2009). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning, *Technical Report 54*, Department of Statistics, University of Munich. Available as `http://epub.ub.uni-muenchen.de/10588/`.

Topalidou, E. & Psarakis, S. (2009). Review of multinomial and multiattribute quality control charts, *Quality and Reliability Engineering International* . In press.

Tutz, G. (1986). Bradley-Terry-Luce models with an ordered response, *Journal of Mathematical Psychology* **30**: 306–316.

Varin, C. & Vidoni, P. (2006). Pairwise likelihood inference for ordinal categorical time series, *Computational Statistics & Data Analysis* **51**: 2365–2373.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th edn, Springer.

Whiting, M. J., Stuart-Fox, D. M., O'Connor, D., Firth, D., Bennett, N. & Blomberg, S. P. (2006). Ultraviolet signals ultra-aggression in a lizard, *Animal Behaviour* **72**(353–363).

Wolfe, D. A. & Chen, Y.-S. (1990). The changepoint problem in a multinomial sequence, *Communications in Statistics – Simulation and Computation* **19**(2): 603–618.

Yee, T. W. (2008). The VGAM package, *R News* **8**(2): 28–39.

Yee, T. W. & Wild, C. J. (1996). Vector generalized additive models, *Journal of the Royal Statistical Society, Series B, Methodological* **58**: 481–493.

Zeileis, A. & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability, *Statistica Neerlandica* **61**(4): 488–508.

Multiple Linear Panel Regression with Multiplicative Random Noise

Hans Schneeweiß and Gerd Ronning

Abstract The paper explores the effect of multiplicative measurement errors on the estimation of a multiple linear panel data model. The conventional fixed effects estimator of the slope parameter vector, which ignores measurement errors, is biased. By correcting for the bias one can construct a consistent and asymptotically normal estimator. In addition, we find a consistent estimate of the asymptotic covariance matrix of this estimator. Measurement errors are sometimes deliberately added to the data in order to minimize their disclosure risk, and then it is often multiplicative errors that are used instead of the more conventional additive errors. Multiplicative errors can be analyzed in a similar way as additive errors, but with some important and consequential differences.

Key words: Panel regression; multiplicative measurement errors; bias correction; asymptotic variance; disclosure control

1 Introduction

The paper explores the effect of measurement errors on the estimation of a multiple linear panel data regression model. The conventional fixed effects least squares estimator, which ignores measurement errors, is biased. By correcting for the bias we can construct consistent and asymptotically normal estimators, where asymptotically here means that the number of sample units tends to infinity.

Hans Schneeweiß
Department of Statistics, University of Munich, Akademiestr. 1, D-80799 München. e-mail: schneew@stat.uni-muenchen.de

Gerd Ronning
Department of Economics, University of Tübingen, Mohlstrasse 36, D-72074 Tübingen. e-mail: gerd.ronning@uni-tuebingen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_21,

Measurement errors can be additive or multiplicative. Additive measurement errors in panel data models have been extensively studied in the literature, Griliches & Hausman (1986), Hsiao & Taylor (1991), Wansbeek & Kooning (1991), Biørn (1996), Wansbeek (2001), Biørn & Krishnakumar (2008). But multiplicative measurement errors, though not uncommon in other models (see, e.g., Hwang 1986, Lin 1989, Carroll et al. 2006), have not found much attention in the context of panel data models.

The present paper was motivated by the various worldwide endeavors to find methods for masking data so that their disclosure risk becomes negligible, see, e.g., Domingo-Ferrer & Saygin (2008). Data and in particular panel data that are released to the public should be not only nominally but also factually anonymous. Making them anonymous in this sense can be done by (slightly) distorting them. The distortion, of course, should be such that the disturbing effects on any subsequent scientific analysis of the data should be minimal or should be amenable to correction. One way of perturbing data is to mix them with random noise, see Kim (1986) as an early reference. This can be done by adding random measurement errors to the data (see, e.g., Brand 2002) or by multiplying them with measurement errors. The latter procedure is often preferred, as it takes automatically into account that large values of a sensitive variable are more prone to disclosure and hence need to be better protected, see also Ronning (2009). In contrast to an additive error, a multiplicative error will distort large values more than small values.

Another aspect of statistical disclosure control techniques is that the procedure used is typically made known to the scientific public. In our case, this means that the measurement error variances and covariances are known to the statistician working with the data.

Although linear panel regressions can also be estimated without this knowledge, we here assume that the error variances and covariances are known. This assumption not only leads to simpler estimators but also to more efficient ones. Indeed, we use this knowledge as prior information to construct consistent estimators of the slope parameters of the model. In addition, we find estimates for the asymptotic variances and covariances of these estimators.

We only deal with one type of estimator, the familiar "within" LS estimator. It uses the "within" variances and covariances for each sample unit over time instead of the overall (total) variances and covariances. In doing so, the unobserved heterogeneity which is present in the panel data is eliminated. There are other estimators that can do the same, especially instrumental variable estimators which use lagged values of the variables as instruments. But in order for them to function properly the variables must be autocorrelated. No such assumptions are needed for the "within" LS estimator.

Although this paper is mainly concerned with multiplicative measurement errors, we also deal briefly with the additive case.

In addition to an i.i.d. component, the measurement errors that we study contain a component which is random over the sample units but constant in time, a case which has been suggested especially for masking panel data. They thus have a common factor structure, see Biørn (1996) and Höhne (2008).

The principles involved for constructing consistent estimators have been developed in the context of a simple linear model with only one slope parameter in Ronning & Schneeweiss (2009). Here we extend the model to the case of a multiple regression.

In Section 2 the linear panel model with measurement errors is presented. Section 3 introduces the within LS (naive) estimator of the slope parameter and derives its bias. In Section 4 a corrected estimator is constructed. Other parameters of interest are briefly dealt with in Section 5. In Section 6 the asymptotic covariance matrix for the corrected slope estimator is presented. Section 7 has a simulation study, where the asymptotic properties of the naive and corrected estimators are studied for a simple linear panel model under small to medium size samples. Section 8 concludes.

2 The Model

For each sampling unit $i = 1, \ldots, N$, assume we have a time series $y_i := (y_{i1}, ..., y_{iT})^\top$ of a response variable y and a matrix

$$X_i = \begin{pmatrix} x_{1i1} & \ldots, & x_{pi1} \\ \vdots & & \vdots \\ x_{1iT} & \ldots, & x_{piT} \end{pmatrix}$$

consisting of time series of p covariates $x_{(k)}$, $k = 1, \ldots, p$. Assume further that there is a linear relationship relating y to X:

$$y_i = (\beta_0 + \alpha_i)\iota_T + X_i\beta + \varepsilon_i, \quad i = 1, \ldots, N, \tag{1}$$

where $\iota_T = (1, ..., 1)^\top$ is a vector of T ones, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})^\top$ is a time series of disturbances, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is the unknown vector of regression parameters, that we wish to estimate, the scalar β_0 is the intercept, and the individual effects α_i are scalars representing the unobserved heterogeneity.

Now assume that the variables y_i and X_i have been perturbed by adding or multiplying random "'measurement"' errors to these data. The randomly perturbed data are denoted by y_i^a and X_i^a and are related to the unperturbed data according to

$$\begin{aligned} X_i^a &= X_i + U_i, \\ y_i^a &= y_i + v_i \end{aligned} \tag{2}$$

in the additive case and

$$\begin{aligned} X_i^a &= X_i \odot (\iota_T \iota_p^\top + U_i), \\ y_i^a &= y_i \odot (\iota_T + v_i) \end{aligned} \tag{3}$$

in the multiplicative case, where U_i is a (T×p) matrix and v_i a (T×1) vector of measurement errors and $\odot$ is the Hadamard product. (We shall only treat the multiplicative case in any detail and shall refer to the additive case only in passing.)

The errors are assumed to have an error component structure such that

$$U_i = \iota_T d_i^\top + U_i^*,$$
$$v_i = \iota_T e_i + v_i^*,$$

where $d_i^\top = (d_{1i}, \ldots, d_{pi})$, and e_i is a scalar.

All variables are taken to be random with the usual independence properties. In particular, the N lists of variables

$$(X_i, \alpha_i, \varepsilon_i, d_i, e_i, U_i^*, v_i^*), \quad i = 1, \ldots, N,$$

are i.i.d. In addition, the T rows of the matrix $(U_i^*, v_i^*, \varepsilon_i)$ are i.i.d. for each i. Furthermore for each i, the set of variable (U_i, v_i) is independent of (X_i, y_i), ε_i is independent of X_i, and (d_i, e_i) is independent of (or at least pairwise uncorrelated with) (U_i^*, v_i^*). All the error terms have expectation zero: $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}d_i = 0$, $\mathbb{E}e_i = 0$, $\mathbb{E}U_i^* = 0$, $\mathbb{E}v_i^* = 0$, and have variances and covariances denoted by $\sigma_\varepsilon^2 := \mathbb{E}\varepsilon_i^2$, $\Sigma_{dd} := \mathbb{E}d_i d_i^\top$, $\sigma_{ee} := \mathbb{E}e_i^2$, $\sigma_{de} := \mathbb{E}d_i e_i$, $\Sigma_{uu}^* := \mathbb{E}u_{it}^* u_{it}^{*\top}$, $\sigma_{uv}^* := \mathbb{E}u_{it}^* v_{it}^*$, $\sigma_{vv}^* := \mathbb{E}v_{it}^{*2}$, where $u_{it}^{*\top}$ is the t-th row of U_i^*. The terms Σ_{uu}, σ_{uv}, and σ_{vv} are similarly defined as Σ_{uu}^*, σ_{uv}^*, and σ_{vv}^*. Moments up to the fourth order are assumed to exist. We also assume $\mathbb{E}\alpha_i = 0$. Two more assumptions will be introduced in Section 3.

3 The Naive Estimator and its Bias

In order to get rid of the individual effects α_i, which are treated as fixed effects, the data matrix (X, y) is premultiplied by the projection matrix

$$P = I_T - \frac{1}{T}\iota_T \iota_T^\top,$$

which has the property $P\iota_T = 0$ and which transforms the data to deviations from their time series means. With these transformed data the within LS estimator of β can be constructed.

If the original unperturbed data y_i and X_i were known, the estimator of β would be given by the solution to the estimating equation

$$S_{xx}\hat{\beta} = s_{xy},$$

where

$$S_{xx} := \frac{1}{N}\sum_{i=1}^{N} \frac{1}{T-1} X_i^\top P X_i,$$

$$s_{xy} := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T-1}X_i^\top P y_i$$

comprise the within covariances of the $x_{(k)}$ and y. Note the division by the degrees of freedom $T-1$ instead of T.

Since the original data are not known to the statistician, this estimator is not feasible. Instead a corresponding estimator $\hat{\beta}^a$ using the perturbed data can be constructed given by the solution to

$$S_{xx}^a \hat{\beta}^a = s_{xy}^a,$$

where

$$S_{xx}^a := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T-1}X_i^{a\top} P X_i^a,$$
$$s_{xy}^a := \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T-1}X_i^{a\top} P y_i^a.$$

We assume that S_{xx}^a is almost surely non-singular, so that $\hat{\beta}^a$ is uniquely defined.

As this estimator does not take into account the fact that the data are perturbed, it is called the naive estimator. It is (asymptotically) biased.

To find the bias we need to compute the probability limits of the second moments used in the construction of $\hat{\beta}^a$. First note that

$$\operatorname{plim} S_{xx} = \mathbb{E}[\frac{1}{T-1}X^\top P X] =: \Sigma_{xx},$$
$$\operatorname{plim} s_{xy} = \mathbb{E}[\frac{1}{T-1}X^\top P y] =: \sigma_{xy}.$$

Note that whenever we compute an expectation we omit the index i since the expectation is independent of i. Note also that Σ_{xx} is not the covariance matrix of the vector $(x_{1it},\ldots,x_{pit})$, which would depend on t in general, but it is the expectation of the empirical within covariance matrix of the data $(x_{1it},\ldots,x_{pit})$ for any i. A similar remark applies to σ_{xy}.

We assume that Σ_{xx} is non-singular.

Now

$$\begin{aligned}\operatorname{plim} S_{xx}^a &= \mathbb{E}[\frac{1}{T-1}X^{a\top} P X^a],\\ &= \mathbb{E}[\frac{1}{T-1}X^\top P X] + \mathbb{E}[\frac{1}{T-1}(X\odot U)^\top P (X\odot U)];\end{aligned}$$

$$\begin{aligned}\operatorname{plim} s_{xy}^a &= \mathbb{E}[\frac{1}{T-1}X^{a\top} P y^a],\\ &= \mathbb{E}[\frac{1}{T-1}X^\top P y] + \mathbb{E}[\frac{1}{T-1}(X\odot U)^\top P (y\odot v)].\end{aligned}$$

Suppressing the index i, partition X into columns: $X = (x_1, \ldots, x_p)$ and similarly $U = (u_1, \ldots, u_p)$, $U^* = (u_1^*, \ldots, u_p^*)$, and $d^\top = (d_1, \ldots, d_p)$, and denote the (l,k)-elements of Σ_{dd} and Σ_{uu}^* by σ_{lk} and σ_{lk}^*, respectively. Then the (k,l)-element of $\mathbb{E}[\frac{1}{T-1}(X \odot U)^\top P(X \odot U)]$ is given by

$$\begin{aligned}
e_{kl} &= \mathbb{E}[\frac{1}{T-1}(x_k \odot u_k)^\top P(x_l \odot u_l)] \\
&= \mathbb{E}\frac{1}{T-1}\mathrm{tr}[P(x_l \odot u_l)(x_k \odot u_k)^\top] \\
&= \mathbb{E}\frac{1}{T-1}\mathrm{tr}[P\{(x_l x_k^\top) \odot (u_l u_k^\top)\}] \\
&= \mathrm{tr}[P\{\mathbb{E}(\frac{1}{T-1}x_l x_k^\top) \odot \mathbb{E}(d_l \iota_T + u_l^*)(d_k \iota_T + u_k^*)^\top\}] \\
&= \mathrm{tr}[P\{\mathbb{E}(\frac{1}{T-1}x_l x_k^\top) \odot (\sigma_{lk}\iota_T \iota_T^\top + \sigma_{lk}^* I_T)\}] \\
&= \sigma_{lk}\mathrm{tr}[P\mathbb{E}(\frac{1}{T-1}x_l x_k^\top)] + \sigma_{lk}^*\mathbb{E}[\mathrm{tr}\{P\,\mathrm{diag}(\frac{1}{T-1}x_l x_k^\top)\}] \\
&= \sigma_{lk}\mathbb{E}[\frac{1}{T-1}x_k^\top P x_l] + \sigma_{lk}^*\frac{T-1}{T}\mathbb{E}(\frac{1}{T-1}x_k^\top x_l),
\end{aligned}$$

where $\mathrm{diag}A$ is the matrix A with all its non-diagonal elements set to zero and where in the last equation we used the easy to prove facts that $\mathrm{tr}PD = \frac{T-1}{T}\mathrm{tr}D$ for any $T \times T$ diagonal matrix D and $\mathrm{tr}(\mathrm{diag}(ab^\top)) = b^\top a$ for any two vectors a and b of equal dimension. It follows that

$$\mathbb{E}[\frac{1}{T-1}(X \odot U)^\top P(X \odot U)] = \Sigma_{dd} \odot \Sigma_{xx} + \Sigma_{uu}^* \odot M_{xx},$$

and similarly,

$$\mathbb{E}[\frac{1}{T-1}(X \odot U)^\top P(y \odot v)] = \sigma_{de} \odot \sigma_{xy} + \sigma_{uv}^* \odot m_{xy},$$

where

$$\begin{aligned}
M_{xx} &:= \mathbb{E}(\frac{1}{T}X^\top X), \\
m_{xy} &:= \mathbb{E}(\frac{1}{T}X^\top y).
\end{aligned}$$

Thus

$$\mathbb{E}[\frac{1}{T-1}X^{a\top} P X^a] = (\iota_p \iota_p^\top + \Sigma_{dd}) \odot \Sigma_{xx} + \Sigma_{uu}^* \odot M_{xx}, \tag{4}$$

$$\mathbb{E}[\frac{1}{T-1}X^{a\top} P y^a] = (\iota_p + \sigma_{de}) \odot \sigma_{xy} + \sigma_{uv}^* \odot m_{xy}. \tag{5}$$

Therefore the probability limit of the naive estimator of β is

$$\operatorname{plim}\hat{\beta}^a = [(\iota_p\iota_p^\top + \Sigma_{dd}) \odot \Sigma_{xx} + \Sigma^*_{uu} \odot M_{xx}]^{-1}[(\iota_p + \sigma_{de}) \odot \sigma_{xy} + \sigma^*_{uv} \odot m_{xy}]. \quad (6)$$

To further evaluate this probability limit and thereby implicitly the bias of $\hat{\beta}^a$ we expand σ_{xy} and m_{xy} using (1):

$$\sigma_{xy} = \Sigma_{xx}\beta, \quad (7)$$
$$m_{xy} = \beta_0 \mathbb{E}\bar{x} + \mathbb{C}ov(\bar{x}, \alpha) + M_{xx}\beta, \quad (8)$$

where $\bar{x} := \frac{1}{T}X^\top\iota$ is the vector of the means of the p time series $x_1, \ldots, x_p$ for any sample unit. We see that the bias depends on the one hand on Σ_{xx} and M_{xx} (and thereby on the joint law governing the p time series $x_1, \ldots, x_p$) and on the other hand on $\mathbb{C}ov(\bar{x}, \alpha)$ (i.e., on the dependency of the individual effects and the regressors). It also depends on β_0.

If the errors have no error component structure (i.e., if $\Sigma_{dd} = 0$ and $\sigma_{de} = 0$, so that $\Sigma^*_{uu} = \Sigma_{uu}$ and $\sigma^*_{uv} = \sigma_{uv}$), then (6) simplifies to

$$\operatorname{plim}\hat{\beta}^a = (\Sigma_{xx} + \Sigma_{uu} \odot M_{xx})^{-1}[\Sigma_{xx}\beta + \sigma_{uv} \odot m_{xy}].$$

If in addition $\sigma_{uv} = 0$, then the bias of $\hat{\beta}^a$ becomes

$$\operatorname{plim}\hat{\beta}^a - \beta = -(\Sigma_{xx} + \Sigma_{uu} \odot M_{xx})^{-1}(\Sigma_{uu} \odot M_{xx})\beta,$$

and the term m_{xy} has no effect on the bias. The minus sign reflects the well-known attenuation effect of measurement errors.

Remark 1: In the case of additive measurement errors (2), we have the much simpler relation

$$\operatorname{plim}\hat{\beta}^a = (\Sigma_{xx} + \Sigma^*_{uu})^{-1}(\Sigma_{xx}\beta + \sigma^*_{uv}).$$

We see that in the additive case the error components d and e have no effect on the bias, and it is only the variances and covariances of the i.i.d. components u^* and v^* that affect the bias.

4 Corrected Estimator

We intend to find a corrected estimating equation for estimating β consistently. However, it turns out that in order to estimate β consistently one has to estimate two nuisance parameters in addition to β, namely, M_{xx} and m_{xy}. Let $\theta = (\beta^\top, \operatorname{vech}^\top(M_{xx}), m_{xy}^\top)^\top$ be the $\frac{1}{2}p(p+5)$-dimensional vector of the parameters to be estimated, then an unbiased vector valued estimating function $\psi := \psi(\theta) := \psi(\theta; X^a, y^a)$ with $\mathbb{E}_\theta\psi(\theta) = 0$ is given by the following three subvectors:

$$\begin{aligned}
\psi_1 &= \left[\left\{\frac{1}{T-1}X^{a\top}PX^a - \Sigma^*_{uu}\odot M_{xx}\right\}\oslash(\iota_p\iota_p^\top+\Sigma_{dd})\right]\beta \\
&\quad - \left\{\frac{1}{T-1}X^{a\top}Py^a - \sigma^*_{uv}\odot m_{xy}\right\}\oslash(\iota_p+\sigma_{de}), \\
\psi_2 &= \operatorname{vech}\left\{\frac{1}{T}X^{a\top}X^a - (\iota_p\iota_p^\top+\Sigma_{dd}+\Sigma^*_{uu})\odot M_{xx}\right\}, \\
\psi_3 &= \frac{1}{T}X^{a\top}y^a - (\iota_p+\sigma_{de}+\sigma^*_{uv})\odot m_{xy},
\end{aligned}$$

such that $\psi = (\psi_1^\top, \psi_2^\top, \psi_3^\top)^\top$, where $\oslash$ denotes Hadamard division and "vech" is the operator that transforms a symmetric matrix into a vector by stacking those parts of the columns of the matrix that lie on and beneath the diagonal one beneath the other, see Lütkepohl (1996). One can easily see that $\mathbb{E}\psi = 0$. Indeed, by (4), (5), and (7),

$$\begin{aligned}
\mathbb{E}\psi_1 &= \{(\iota_p\iota_p^\top+\Sigma_{dd})\odot\Sigma_{xx}\oslash(\iota_p\iota_p^\top+\Sigma_{dd})\}\beta \\
&\quad -(\iota_p+\sigma_{de})\odot\sigma_{xy}\oslash(\iota_p+\sigma_{de}) \\
&= \Sigma_{xx}\beta - \Sigma_{xx}\beta = 0.
\end{aligned}$$

Similarly, $\mathbb{E}\psi_2 = 0$ and $\mathbb{E}\psi_3 = 0$ because by arguments similar to those that led to (4) and (5) one can show that

$$\begin{aligned}
\mathbb{E}\frac{1}{T}X^{a\top}X^a &= (\iota_p\iota_p^\top+\Sigma_{uu})\odot M_{xx}, \\
\mathbb{E}\frac{1}{T}X^{a\top}y^a &= (\iota_p+\sigma_{uv})\odot m_{xy}
\end{aligned}$$

and $\Sigma_{uu} = \Sigma_{dd} + \Sigma^*_{uu}$, $\sigma_{uv} = \sigma_{de} + \sigma^*_{uv}$.

Now a consistent estimator of θ is given by the solution to $\sum_{i=1}^N \psi(\hat\theta; X_i, y_i) = 0$. Thus the corrected estimator for β is given by the solution to

$$S^c_{xx}\hat\beta^c = s^c_{xy} \tag{9}$$

with

$$\begin{aligned}
S^c_{xx} &:= (S^a_{xx} - \Sigma^*_{uu}\odot\hat M_{xx})\oslash(\iota_p\iota_p^\top+\Sigma_{dd}), \\
s^c_{xy} &:= (s^a_{xy} - \sigma^*_{uv}\odot\hat m_{xy})\oslash(\iota_p+\sigma_{de}),
\end{aligned}$$

and

$$\begin{aligned}
\hat M_{xx} &:= \frac{1}{NT}\sum_{i=1}^N X_i^{a\top}X_i^a\oslash(\iota_p\iota_p^\top+\Sigma_{dd}+\Sigma^*_{uu}), \\
\hat m_{xy} &:= \frac{1}{NT}\sum_{i=1}^N X_i^{a\top}y_i^a\oslash(\iota_p+\sigma_{de}+\sigma^*_{uv}).
\end{aligned}$$

The corrected estimator $\hat{\beta}^c$ simplifies if the error components d and e are not present. In this case,

$$\begin{aligned} S_{xx}^c &= S_{xx}^a - \Sigma_{uu} \odot \hat{M}_{xx}, \\ s_{xy}^c &= s_{xy}^a - \sigma_{uv} \odot \hat{m}_{xy}, \\ \hat{M}_{xx} &= \frac{1}{NT} \sum_{i=1}^{N} X_i^{a\top} X_i^a \oslash (\iota_p \iota_p^\top + \Sigma_{uu}), \\ \hat{m}_{xy} &= \frac{1}{NT} \sum_{i=1}^{N} X_i^{a\top} y_i^a \oslash (\iota_p + \sigma_{uv}). \end{aligned}$$

Remark 2: In the case of additive measurement errors, the corrected estimator is given by the solution to

$$(S_{xx}^a - \Sigma_{uu}^*)\hat{\beta}^c = s_{xy}^a - \sigma_{uv}^*,$$

and no nuisance parameters need to be estimated.

5 Residual Variance and Intercept

Apart from the regression parameter β, one may also want to estimate the residual variance σ_ε^2. The usual estimator with the original data is

$$\hat{\sigma}_\varepsilon^2 = s_{yy} - s_{xy}^\top \hat{\beta}$$

with $s_{yy} := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T-1} y_i^\top P y_i$. A corrected estimator with the perturbed data is given by

$$\hat{\sigma}_\varepsilon^{c2} = s_{yy}^c - s_{xy}^{c\top} \hat{\beta}^c,$$

where

$$\begin{aligned} s_{yy}^c &:= (s_{yy}^a - \sigma_{vv}^* \hat{m}_{yy})/(1 + \sigma_{ee}), \\ \hat{m}_{yy} &:= \frac{1}{NT} \sum_{i=1}^{N} y_i^{a\top} y_i^a / (1 + \sigma_{ee} + \sigma_{vv}^*). \end{aligned}$$

Another parameter of interest is the intercept term β_0. It can be estimated with the original data by

$$\hat{\beta}_0 = \bar{\bar{y}} - \bar{\bar{x}}^\top \hat{\beta},$$

where $\bar{\bar{y}} := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it}$ and $\bar{\bar{x}} := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}$ with x_{it} being the rows of X_i. Let $\overline{\overline{y^a}}$ and $\overline{\overline{x^a}}$ be defined in a similar way. As $\mathbb{E}\overline{\overline{y^a}} = \mathbb{E}\bar{\bar{y}}$ and $\mathbb{E}\overline{\overline{x^a}} = \mathbb{E}\bar{\bar{x}}$, the corrected estimator is simply given by

$$\hat{\beta}_0^c = \overline{\overline{y^a}} - \overline{\overline{x^a}}^\top \hat{\beta}^c.$$

6 Asymptotic Covariance Matrix

In general, if an estimator of a parameter vector θ is given by an unbiased estimating function, then the estimator $\hat{\theta}$ is, under some regularity assumptions, consistent and asymptotically normally distributed with a covariance matrix that is given by the sandwich formula (see, e.g., Heyde 1997)

$$\Sigma_{\hat{\theta}} = \frac{1}{N}\left(\mathbb{E}\frac{\partial \psi}{\partial \theta^\top}\right)^{-1} \mathbb{E}\psi\psi^\top \left(\mathbb{E}\frac{\partial \psi^\top}{\partial \theta}\right)^{-1},$$

which is consistently estimated by

$$\hat{\Sigma}_{\hat{\theta}} = \left(\sum_{i=1}^N \frac{\partial \psi_i}{\partial \theta^\top}\right)^{-1} \sum_{i=1}^N \psi_i\psi_i^\top \left(\sum_{i=1}^N \frac{\partial \psi_i^\top}{\partial \theta}\right)^{-1},$$

where ψ_i is short for $\psi(\hat{\theta}; X_i, y_i)$. As in our case ψ consists of three subvectors, the three parts of the sandwich formula as well as the covariance matrix itself each partition into 3×3 submatrices. The submatrix in the upper left corner of $\hat{\Sigma}_{\hat{\theta}}$ is an estimate of the asymptotic covariance matrix $\Sigma_{\hat{\beta}^c}$ of $\hat{\beta}^c$. The nine submatrices ψ_{hj} for $\frac{\partial \psi}{\partial \theta^\top}$ (again suppressing the index i) are given by

$$\psi_{11} := \frac{\partial \psi_1}{\partial \beta^\top} = \left(\frac{1}{T-1} X^{a\top} P X^a - \Sigma^*_{uu} \odot M_{xx}\right) \oslash (\iota_p \iota_p^\top + \Sigma_{dd}),$$

$$\psi_{12} := \frac{\partial \psi_1}{\partial \operatorname{vech}^\top M_{xx}} = -(\beta^\top \otimes I_p) H \operatorname{diag} \operatorname{vech}\{\Sigma^*_{uu} \oslash (\iota_p \iota_p^\top + \Sigma_{dd})\},$$

$$\psi_{13} := \frac{\partial \psi_1}{\partial m_{xy}^\top} = \operatorname{diag}(\sigma^*_{uv} \oslash (\iota_p + \sigma_{de})),$$

$$\psi_{21} := \frac{\partial \psi_2}{\partial \beta^\top} = 0,$$

$$\psi_{22} := \frac{\partial \psi_2}{\partial \operatorname{vech}^\top M_{xx}} = -\operatorname{diag} \operatorname{vech}(\iota_p \iota_p^\top + \Sigma_{dd} + \Sigma^*_{uu}),$$

$$\psi_{23} := \frac{\partial \psi_2}{\partial m_{xy}^\top} = 0,$$
$$\psi_{31} := \frac{\partial \psi_3}{\partial \beta^\top} = 0,$$
$$\psi_{32} := \frac{\partial \psi_3}{\partial \operatorname{vech}^\top M_{xx}} = 0,$$
$$\psi_{33} := \frac{\partial \psi_3}{\partial m_{xy}^\top} = -\operatorname{diag}(\iota_p + \sigma_{de} + \sigma_{uv}^*),$$

where H is a "duplication matrix" that transforms vech$A$ into vec$A$, i.e., vec$A$ = H vech$A$ for any symmetric matrix A, see Lütkepohl (1996, p. 98).

Clearly, if $\sigma_{uv}^* = 0$, we need not estimate m_{xy} and the third subvector ψ_3 of ψ may be dropped.

7 Simulation

In our simulation study we focus on the simple linear model ($p = 1$) with only one slope parameter β to be estimated. We analyze the performance of both the naive estimator $\hat{\beta}^a$ and the corrected estimator $\hat{\beta}^c$ when both x and y are observed with multiplicative measurement errors. We distinguish between the i.i.d. case, where $e_i = d_i = 0$ and the common factor case with non-vanishing d_i and e_i.

For the regressor variable x we assume the stationary AR(1) process

$$x_{it} \;=\; \phi \;+\; \rho\, x_{i,t-1} \;+\; \omega_{it}$$

with $|\rho| < 1$ and ω_{it} normal white noise, independent of the x_{it}, with expectation 0 and variance σ_ω^2. As the x_{it} are stationary, $\mathbb{E}x_{it} := \mu_x$ and $\mathbb{V}x_{it} := \sigma_x^2$ are constant. Given ρ , μ_x, and σ_x^2, the parameters ϕ and σ_ω^2 used to generate the x_{it} are given by

$$\phi = (1-\rho)\mu_x,$$
$$\sigma_\omega^2 = \sigma_x^2(1-\rho^2).$$

In order to study the effect of correlation of the individual effect α with the regressor x, we assume that the α_i are generated by

$$\alpha_i \;=\; (\bar{x}_i - \mathbb{E}\bar{x}_i)\,\lambda \;+\; w_i,$$

where $\bar{x}_i := \frac{1}{T}\sum_t x_{it}$ and w_i is normal white noise with expectation 0 and variance σ_w^2 distributed independently of x. This specification of correlated individual effects has been proposed by Biørn (1996, p. 260). Due to the stationarity of x_{it}, $\mathbb{E}\bar{x}_i = \mu_x$.

In our simulations we fix the correlation between α_i and $\bar{x}_i$, $\rho_{\alpha\bar{x}}$, and the variance of α_i, σ_α^2, and derive the two remaining parameters as follows:

$$\lambda = \frac{\rho_{\alpha\bar{x}}\sigma_\alpha}{\sigma_{\bar{x}}},$$
$$\sigma_w^2 = \sigma_\alpha^2(1-\rho_{\alpha\bar{x}}^2),$$

where $\sigma_{\bar{x}}$ is given by (see Hamilton 1994)

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{T^2}\left\{T + 2(T-1)\rho + 2(T-2)\rho^2 + \cdots + 2\rho^{T-1}\right\}. \tag{10}$$

When studying the common factor structure, we use the specification of Höhne (2008): We set $d_i = e_i$ and use the special structure

$$e_i = \delta D \quad \text{with} \quad D = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5 \end{cases}.$$

This specification implies $\sigma_{dd} = \sigma_{de}$.

The corrected estimator (9) then simplifies to

$$\hat{\beta}^c = \frac{s_{xy}^a - \sigma_{uv}^*\hat{m}_{xy}}{s_{xx}^a - \sigma_{uu}^*\hat{m}_{xx}}$$

with $\hat{m}_{xy} = \frac{1}{NT}\sum_i\sum_t x_{it}^a y_{it}^a/(1+\sigma_{dd}+\sigma_{uv}^*)$ and $\hat{m}_{xx} = \frac{1}{NT}\sum_i\sum_t x_{it}^{a2}/(1+\sigma_{dd}+\sigma_{uu}^*)$. In the i.i.d. case, $\sigma_{dd}=0$, $\sigma_{uv}^* = \sigma_{uv}$, and $\sigma_{uu}^* = \sigma_{uu}$.

The probability limit of the naive estimator according to (6), (7), and (8) becomes

$$\operatorname{plim}\hat{\beta}^a = \frac{(1+\sigma_{dd})\sigma_{xx}\beta + \sigma_{uv}^*(\rho_{\alpha\bar{x}}\sigma_\alpha\sigma_{\bar{x}} + \beta_0\mu_x + m_{xx}\beta)}{(1+\sigma_{dd})\sigma_{xx} + \sigma_{uu}^* m_{xx}}, \tag{11}$$

where (the bar denoting averages over t for fixed i)

$$m_{xx} = \mathbb{E}\overline{x^2} = \mathbb{E}x^2 = \sigma_x^2 + \mu_x^2,$$
$$\sigma_{xx} = \tfrac{T}{T-1}\mathbb{E}(\overline{x^2} - \bar{x}^2) = \tfrac{T}{T-1}(\sigma_x^2 - \sigma_{\bar{x}}^2),$$

and $\sigma_{\bar{x}}^2$ is given by (10).

The following parameters were fixed throughout the whole simulation study:

$$\beta = 1,\ \mu_x = 2,\ \sigma_x^2 = 1.5^2,\ \sigma_\varepsilon^2 = 0.5^2,\ \sigma_\alpha^2 = 1,$$

and for the measurement errors we used

$$\sigma_{uu} = 0.2^2,\ \sigma_{vv} = 0.2^2$$

for the i.i.d. case and

$$\delta = 0.14,\ \sigma_{uu}^* = 0.14^2,\ \sigma_{vv}^* = 0.14^2$$

Table 1 Parameter values in the simulation study

parameter	values used
N	100 ; 1000
T	3 ; 10
ρ	-0.5 ; 0 ; +0.5
$\rho_{uv}(\rho^*_{uv})$	-0.9 ; 0 ; +0.9
$\rho_{\alpha\overline{x}}$	0 ; 0.975
β_0	0 ; 5

for the common factor case. Note that the total variance of measurement error in the latter case is given by $\sigma_{uu} = \delta^2 + \sigma^*_{uu} = 0.14^2 + 0.14^2 = 0.0392$, which is (almost) equal to the variance in the i.i.d. case.

For both the i.i.d. case and the common factor case, we studied the effects of varying the sample size N, the number of waves T, the autoregressive parameter ρ , the correlation between u and v, which we denote by ρ_{uv} (ρ^*_{uv} for the common factor model), the correlation between α and $\overline{x}$, and the intercept term β_0. Table 1 has the details. To save space, not all parameter combinations are shown. In all scenarios we use 2000 replications.

The four tables shown in the appendix contain the simulation results for the i.i.d. case (tables 2 and 3) and for the common factor case (tables 4 and 5). In each case, the second table reports results concerning correlation of individual effects with the regressor. In all four tables, we use the following notation:

$\hat{\beta}^a$ and $s_{\hat{\beta}^a}$ give mean and standard deviation of the 2000 replications concerning the naive estimator, $\hat{\beta}^c$ and $s_{\hat{\beta}^c}$ the corresponding results for the corrected estimator. $\hat{\sigma}_{\hat{\beta}^c}$ denotes the mean of the estimate of the theoretical (asymptotic) standard deviation discussed in Section 6, and $s_{\hat{\sigma}_{\hat{\beta}^c}}$ reports the standard error of this estimate. Finally, $q^{\gamma}_{\hat{\beta}^c}$ is the γ-quantile of the corrected estimator for three different levels of γ.

For large samples ($N = 1000$), our simulations support our theoretical findings: The corrected slope estimator $\hat{\beta}^c$ shows practically no bias, and the average estimate of the theoretical (asymptotic) standard deviation $\hat{\sigma}_{\hat{\beta}^c}$ of the estimator $\hat{\beta}^c$ corresponds very accurately to the empirical standard deviation $s_{\hat{\beta}^c}$ of the estimates $\hat{\beta}^c$ in the simulation runs. The asymptotic results seem to apply almost as well to samples of small to medium size ($N = 100$): the corrected slope estimator shows hardly any bias, and the theoretical standard deviation still corresponds rather closely to the empirical standard deviation. Of course, for smaller N, these standard deviations are (about three times) larger.

The simulations also highlight the considerable amount of bias in the uncorrected (naive) estimator of the slope parameter. The empirical findings on the bias are in accordance with the theoretical result (11). For uncorrelated individual effects and $\beta_0 = 0$, the bias tends to increase for increasing ρ and for decreasing ρ_{uv}. The bias is considerably smaller for errors with a common factor structure, which is plausible considering the fact that, by using inner variances and covariances for constructing the estimator, the common factor is largely eliminated – it is completely eliminated

in the additive case – so that the, much smaller, remaining error components u^* and v^* are now relevant for the bias. The presence of a correlation between individual effect and regressor has only a small effect on the bias. For $\rho_{\alpha\bar{x}} = 0.975$, the bias is somewhat smaller than in the case of no correlation if $\sigma_{uv} \neq 0$. The effect of the intercept β_0 on the bias varies with the values of the other parameters.

The standard deviation of the corrected estimator can also be seen to depend on the various model parameters. It decreases for increasing T, decreasing ρ, and increasing ρ_{uv}. It is a good deal smaller in the common factor case. The dependence on $\rho_{\alpha\bar{x}}$ is negligible. The standard deviation of the corrected estimator is, of course, larger than for the naive estimator, but not very much. The increase in variance is outweighed by the elimination of bias. Finally, it may be noted that the estimate of the standard deviation is very precise in view of its own standard deviation $s_{\hat{\sigma}_{\hat{\beta}^c}}$, in particular for large N.

8 Conclusion

Measurement errors in a linear regression result in biased estimates of the slope parameters when Least Squares (LS) is applied without regard to the measurement errors. This is true both for cross sectional models using ordinary LS as well as for panel data models using within LS (the latter in order to get rid of the unobservable individual effects).

We focus our investigation on multiplicative errors with a common factor structure. They can be treated in a similar way as the more conventional additive errors, but with some characteristic differences. In the bias formula as well as in the expression for the bias corrected estimator, nuisance parameters appear, which have to be estimated, too. Their presence results in a substantially more complicated computation of the asymptotic covariance matrix of the slope estimators than in the additive case. The covariance matrix is computed with the help of the sandwich formula, which, however, has to take the nuisance parameters into account.

An extensive simulation study was carried out. It fully corroborates our theoretical findings on the asymptotics of our estimators and shows that the asymptotic results seem to apply almost as well to samples of small to medium size ($N = 100$). The simulations also make evident the dependence of the asymptotic variance on the various model parameters, e.g., on the autocorrelation of the regressor variable or on the correlation between regressor and individual effect.

Finally, they show how close the asymptotic variance of the corrected estimator may come to that of the uncorrected estimator, at least for large N. Thus the correction is fully justified both on the ground that it eliminates the bias and that it implies only a small increase in variance.

Acknowledgements Financial support by Bundesministerium für Bildung und Forschung (project "Wirtschaftsstatistische Paneldaten und faktische Anonymisierung") is gratefully acknowledged.

References

Biørn, E. (1996), Panel data with measurement errors, in: Mátyás, L. and P. Sevestre (eds.), *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, 2nd. ed., Kluwer, Dordrecht, 236–279.

Biørn, E. & Krishnakumar, J. (2008). Measurement errors and simultaneity. Chapter 10 in: Mátyás, L. and P. Sevestre (eds.), *The Econometrics of Panel Data*, 3rd. ed., Springer, Heidelberg, pp. 323–367.

Brand, R. (2002), Microdata protection through noise addition. In: Domingo-Ferrer. J. (ed.), *Inference Control in Statistical Data Bases. Lecture Notes in Computer Science 2316*, Berlin, Springer-Verlag, pp. 97–116.

Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2006), *Measurement Error in Nonlinear Models*, Chapman and Hall, London.

Domingo-Ferrer, J. & Y. Saygin (2008) (eds), *Privacy in Statistical Databases*. Unesco Chair in Data Privacy International Conference, PSD 2008. Istanbul, Turkey, September 2008, Proceedings. Springer, Berlin.

Griliches, Z. & Hausman, J.A. (1986), Errors in variables in panel data, *Journal of Econometrics* **31**: 93–118.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Heyde C.C. (1997), *Quasi-Likelihood and Its Application*, Springer, Berlin, Heidelberg, New York.

Höhne, J. (2008). Anonymisierungsverfahren für Paneldaten.*Wirtschafts- und Sozialstatistisches Archiv* **2**: 259–275.

Hsiao, C. & Taylor, G. (1991), Some remarks on measurement errors and the identification of panel data models, *Statistica Neerlandica* **45**: 187–194.

Hwang, J.T. (1986), Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy. *Journal of the American Statistical Association* **81**: 680–688.

Kim, J.J. (1986), A Method For Limiting Disclosure in Microdata Based on Random Noise and Transformation, *Proceedings of the Survey Research Methods Section* , American Statistical Association, 370–374.

Lin, A. (1989), Estimation of multiplicative measurement error models and some simulation results. *Economics Letters* **31**: 13–20.

Lütkepohl, H. (1996). *Handbook of Matrices*. Wiley, New York.

Ronning, G. (2009). Measuring Research Intensity From Anonymized Data: Does Multiplicative Noise With Factor Structure Save Results Regarding Quotients? *Jahrbücher für Nationalökonomie und Statistik* **228**, 645–653.

Ronning, G. & Schneeweiss, H. (2009), Panel regression with random noise. CES/ifo Working Paper 2608 (April 2009).
`http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1379622`

Wansbeek, T.J. (2001), GMM estimation in panel data models with measurement errors, *Journal of Econometrics* **104**: 259–268.

Wansbeek, T.J. & Koning, R.H. (1991), Measurement error and panel data, *Statistica Neerlandica* **45**: 85–92.

Table 2 Simulation results for the iid case – uncorrelated individual effects

ρ	ρ_{uv}	β_0	N	$\hat{\beta}^a$	$s_{\hat{\beta}a}$	$\hat{\beta}^c$	$s_{\hat{\beta}c}$	$\hat{\sigma}_{\hat{\beta}c}$	$s_{\hat{\sigma}_{\hat{\beta}c}}$	$q_{\hat{\beta}c}^{0.05}$	$q_{\hat{\beta}c}^{0.50}$	$q_{\hat{\beta}c}^{0.95}$
				T = 3								
-0.5	-0.9	0	100	0.84283	0.05066	0.99924	0.05891	0.05782	0.00872	0.90154	0.99830	1.09681
		0	1000	0.84489	0.01648	0.99997	0.01889	0.01871	0.00100	0.96914	0.99956	1.03066
		5	1000	0.72757	0.02526	1.00045	0.02959	0.02964	0.00136	0.95210	0.99982	1.05111
-0.5	0.0	0	100	0.91859	0.04105	1.00083	0.04540	0.04439	0.00637	0.92737	1.00093	1.07725
		0	1000	0.91875	0.01339	1.00050	0.01475	0.01439	0.00070	0.97608	1.00041	1.02393
		5	1000	0.91855	0.02147	1.00025	0.02355	0.02364	0.00103	0.96253	1.00021	1.04003
-0.5	0.9	0	100	0.99148	0.02554	0.99966	0.02693	0.02673	0.00354	0.95457	0.99940	1.04433
		0	1000	0.99182	0.00800	0.99999	0.00843	0.00858	0.00036	0.98640	0.99997	1.01394
		5	1000	1.10937	0.01599	0.99980	0.01705	0.01662	0.00070	0.97102	1.00001	1.02697
0.0	-0.9	0	100	0.81154	0.05494	1.00371	0.06589	0.06434	0.00956	0.89923	1.00156	1.11200
		0	1000	0.80980	0.01765	1.00030	0.02094	0.02083	0.00112	0.96597	0.99979	1.03525
		5	1000	0.66629	0.02828	1.00111	0.03405	0.03364	0.00160	0.94400	1.00104	1.05641
0.0	0.0	0	100	0.89991	0.04435	1.00107	0.05037	0.04927	0.00701	0.92314	0.99979	1.08438
		0	1000	0.90033	0.01392	1.00044	0.01589	0.01592	0.00076	0.97440	1.00034	1.02637
		5	1000	0.90026	0.02460	1.00040	0.02737	0.02655	0.00116	0.95424	1.00135	1.04465
0.0	0.9	0	100	0.98919	0.02832	0.99923	0.03047	0.03000	0.00371	0.95045	0.99839	1.04919
		0	1000	0.99029	0.00883	1.00030	0.00946	0.00962	0.00039	0.98470	1.00054	1.01586
		5	1000	1.13406	0.01700	0.99973	0.01871	0.01889	0.00077	0.96930	0.99905	1.03116
0.5	-0.9	0	100	0.69584	0.06767	1.00409	0.09125	0.08734	0.01473	0.86124	1.00040	1.16018
		0	1000	0.69625	0.02224	1.00038	0.02893	0.02830	0.00166	0.95370	0.99998	1.04780
		5	1000	0.46477	0.03513	1.00020	0.04680	0.04697	0.00245	0.92510	0.99915	1.07473
0.5	0.0	0	100	0.84122	0.05552	1.00288	0.06816	0.06568	0.00985	0.89316	1.00238	1.11608
		0	1000	0.84001	0.01798	1.00043	0.02185	0.02133	0.00115	0.96501	1.00066	1.03558
		5	1000	0.83971	0.02959	1.00002	0.03566	0.03607	0.00168	0.94274	1.00004	1.05946
0.5	0.9	0	100	0.98248	0.03604	0.99825	0.04094	0.04056	0.00539	0.93236	0.99821	1.06745
		0	1000	0.98331	0.01117	0.99930	0.01280	0.01293	0.00056	0.97738	0.99917	1.02029
		5	1000	1.21447	0.02231	0.99971	0.02622	0.02597	0.00114	0.95614	0.99991	1.04272
				T = 10								
-0.5	-0.9	0	100	0.82120	0.02626	1.00055	0.03112	0.03101	0.00334	0.94925	1.00073	1.05115
		0	1000	0.82125	0.00822	1.00010	0.00979	0.00991	0.00037	0.98374	1.00042	1.01594
		5	1000	0.68563	0.01299	1.00034	0.01567	0.01559	0.00053	0.97401	1.00063	1.02590
-0.5	0.0	0	100	0.90569	0.02138	1.00000	0.02363	0.02341	0.00240	0.96126	0.99982	1.03805
		0	1000	0.90610	0.00665	1.00034	0.00745	0.00753	0.00025	0.98817	1.00035	1.01254
		5	1000	0.90569	0.01107	0.99987	0.01223	0.01228	0.00038	0.98007	0.99950	1.02069
-0.5	0.9	0	100	0.99015	0.01343	0.99960	0.01398	0.01391	0.00131	0.97677	0.99939	1.02260
		0	1000	0.99062	0.00421	1.00001	0.00440	0.00443	0.00013	0.99263	1.00004	1.00721
		5	1000	1.12602	0.00809	0.99985	0.00857	0.00853	0.00026	0.98584	1.00003	1.01350
0.0	-0.9	0	100	0.81098	0.02709	1.00155	0.03199	0.03193	0.00344	0.94922	1.00060	1.05344
		0	1000	0.80996	0.00852	0.99984	0.01021	0.01022	0.00035	0.98369	0.99964	1.01726
		5	1000	0.66583	0.01331	0.99978	0.01616	0.01618	0.00053	0.97342	1.00000	1.02614
0.0	0.0	0	100	0.90057	0.02235	1.00076	0.02511	0.02423	0.00235	0.95949	1.00064	1.04204
		0	1000	0.89992	0.00694	0.99994	0.00781	0.00776	0.00025	0.98693	0.99970	1.01265
		5	1000	0.90017	0.01144	1.00017	0.01275	0.01272	0.00039	0.97952	1.00016	1.02169
0.0	0.9	0	100	0.98946	0.01424	0.99945	0.01490	0.01441	0.00127	0.97498	0.99964	1.02414
		0	1000	0.99005	0.00425	1.00010	0.00443	0.00459	0.00013	0.99278	1.00028	1.00704
		5	1000	1.13415	0.00829	1.00005	0.00892	0.00886	0.00026	0.98581	1.00012	1.01533
0.5	-0.9	0	100	0.77289	0.03006	1.00004	0.03464	0.03513	0.00406	0.94285	0.99969	1.05504
		0	1000	0.77404	0.00960	1.00014	0.01135	0.01127	0.00042	0.98176	1.00005	1.01912
		5	1000	0.60232	0.01555	0.99984	0.01833	0.01812	0.00061	0.96883	0.99987	1.02910
0.5	0.0	0	100	0.88009	0.02384	0.99906	0.02697	0.02662	0.00286	0.95383	0.99827	1.04515
		0	1000	0.88097	0.00761	1.00004	0.00866	0.00852	0.00029	0.98526	1.00003	1.01384
		5	1000	0.88071	0.01264	0.99972	0.01434	0.01412	0.00045	0.97547	0.99953	1.02381
0.5	0.9	0	100	0.98794	0.01498	0.99993	0.01579	0.01593	0.00152	0.97297	0.99996	1.02554
		0	1000	0.98815	0.00473	1.00006	0.00499	0.00508	0.00015	0.99181	1.00017	1.00841
		5	1000	1.15940	0.00912	0.99991	0.00985	0.00991	0.00030	0.98419	0.99993	1.01571

Table 3 Simulation results for the iid case – correlated individual effects ($\rho_{\overline{x}\alpha} = 0.975$)

ρ	ρ_{uv}	β_0	**N**	$\hat{\beta}^a$	$s_{\hat{\beta}^a}$	$\hat{\beta}^c$	$s_{\hat{\beta}^c}$	$\hat{\sigma}_{\hat{\beta}^c}$	$s_{\hat{\sigma}_{\hat{\beta}^c}}$	$q^{0.05}_{\hat{\beta}^c}$	$q^{0.50}_{\hat{\beta}^c}$	$q^{0.95}_{\hat{\beta}^c}$
				T = 3								
-0.5	-0.9	0	100	0.84053	0.05337	1.00471	0.06200	0.05958	0.00911	0.90212	1.00423	1.10365
		0	1000	0.83811	0.01714	1.00015	0.01986	0.01922	0.00104	0.96743	1.00030	1.03357
		5	1000	0.72052	0.02613	1.00045	0.03036	0.03008	0.00143	0.95091	1.00010	1.05134
-0.5	0.0	0	100	0.91995	0.04237	1.00232	0.04676	0.04524	0.00656	0.92776	1.00098	1.08232
		0	1000	0.91829	0.01344	1.00008	0.01479	0.01468	0.00073	0.97606	1.00003	1.02419
		5	1000	0.91802	0.02195	0.99972	0.02404	0.02387	0.00106	0.95999	0.99958	1.03934
-0.5	0.9	0	100	0.99780	0.02599	0.99892	0.02736	0.02709	0.00348	0.95484	0.99884	1.04426
		0	1000	0.99918	0.00829	1.00034	0.00870	0.00870	0.00039	0.98580	1.00019	1.01496
		5	1000	1.11660	0.01549	1.00000	0.01652	0.01684	0.00074	0.97245	0.99991	1.02721
0.0	-0.9	0	100	0.79779	0.05781	1.00260	0.06932	0.06759	0.01064	0.89045	1.00292	1.11316
		0	1000	0.79799	0.01856	1.00037	0.02190	0.02182	0.00123	0.96435	1.00053	1.03682
		5	1000	0.65390	0.02861	1.00026	0.03435	0.03437	0.00167	0.94415	1.00013	1.05946
0.0	0.0	0	100	0.89952	0.04636	1.00019	0.05172	0.05129	0.00788	0.91810	0.99872	1.08645
		0	1000	0.89971	0.01465	0.99969	0.01651	0.01649	0.00084	0.97274	0.99946	1.02676
		5	1000	0.89958	0.02406	0.99957	0.02686	0.02703	0.00124	0.95436	1.00025	1.04308
0.0	0.9	0	100	1.00240	0.02829	1.00026	0.03037	0.03036	0.00390	0.95087	0.99971	1.05111
		0	1000	1.00225	0.00914	1.00011	0.00976	0.00973	0.00041	0.98358	1.00012	1.01679
		5	1000	1.14696	0.01769	1.00071	0.01942	0.01917	0.00080	0.96737	1.00097	1.03275
0.5	-0.9	0	100	0.67100	0.07454	1.00618	0.09625	0.09506	0.01792	0.84582	1.00710	1.16736
		0	1000	0.66977	0.02416	1.00076	0.03056	0.03058	0.00195	0.95008	1.00010	1.05047
		5	1000	0.44067	0.03790	1.00225	0.04982	0.04878	0.00277	0.92001	1.00133	1.08674
0.5	0.0	0	100	0.84164	0.05639	1.00528	0.06957	0.07030	0.01140	0.89500	1.00395	1.12159
		0	1000	0.84049	0.01872	1.00075	0.02284	0.02255	0.00124	0.96320	1.00078	1.03795
		5	1000	0.83951	0.03170	0.99949	0.03795	0.03714	0.00179	0.93862	0.99917	1.06553
0.5	0.9	0	100	1.01084	0.03683	1.00040	0.04250	0.04091	0.00562	0.93068	1.00103	1.07040
		0	1000	1.01019	0.01145	0.99976	0.01313	0.01318	0.00060	0.97824	0.99992	1.02079
		5	1000	1.24025	0.02276	0.99896	0.02715	0.02677	0.00124	0.95333	1.00009	1.04154
				T = 10								
-0.5	-0.9	0	100	0.81771	0.02711	1.00100	0.03188	0.03136	0.00343	0.94769	1.00063	1.05634
		0	1000	0.81742	0.00876	1.00005	0.01018	0.01006	0.00036	0.98373	0.99978	1.01790
		5	1000	0.68154	0.01334	0.99985	0.01567	0.01572	0.00053	0.97358	0.99972	1.02513
-0.5	0.0	0	100	0.90702	0.02155	1.00169	0.02396	0.02386	0.00242	0.96318	1.00161	1.04065
		0	1000	0.90557	0.00685	0.99965	0.00763	0.00762	0.00027	0.98749	0.99941	1.01255
		5	1000	0.90597	0.01148	1.00016	0.01265	0.01234	0.00038	0.97973	1.00036	1.02091
-0.5	0.9	0	100	0.99438	0.01333	1.00001	0.01381	0.01398	0.00133	0.97809	0.99992	1.02319
		0	1000	0.99441	0.00425	0.99998	0.00445	0.00445	0.00013	0.99287	0.99989	1.00734
		5	1000	1.12997	0.00795	0.99996	0.00848	0.00857	0.00026	0.98580	1.00004	1.01395
0.0	-0.9	0	100	0.80262	0.02880	0.99975	0.03337	0.03278	0.00381	0.94516	0.99990	1.05352
		0	1000	0.80328	0.00894	0.99991	0.01040	0.01049	0.00038	0.98309	0.99964	1.01725
		5	1000	0.65928	0.01374	1.00014	0.01631	0.01639	0.00055	0.97357	1.00013	1.02631
0.0	0.0	0	100	0.90022	0.02277	1.00034	0.02588	0.02470	0.00257	0.95914	1.00007	1.04374
		0	1000	0.89992	0.00692	0.99994	0.00784	0.00793	0.00027	0.98698	1.00008	1.01310
		5	1000	0.90027	0.01175	1.00032	0.01308	0.01283	0.00040	0.97830	1.00014	1.02199
0.0	0.9	0	100	0.99655	0.01400	0.99979	0.01466	0.01448	0.00130	0.97635	0.99939	1.02568
		0	1000	0.99642	0.00430	0.99978	0.00450	0.00462	0.00013	0.99239	0.99973	1.00728
		5	1000	1.14052	0.00849	0.99981	0.00904	0.00893	0.00026	0.98499	0.99984	1.01505
0.5	-0.9	0	100	0.76155	0.03299	1.00154	0.03735	0.03662	0.00431	0.93898	1.00155	1.06335
		0	1000	0.76113	0.01053	1.00028	0.01206	0.01178	0.00048	0.98089	1.00048	1.02083
		5	1000	0.58974	0.01635	1.00073	0.01891	0.01853	0.00067	0.96963	1.00048	1.03202
0.5	0.0	0	100	0.88084	0.02420	1.00014	0.02738	0.02759	0.00305	0.95500	0.99943	1.04608
		0	1000	0.88069	0.00780	0.99983	0.00884	0.00882	0.00033	0.98505	0.99990	1.01420
		5	1000	0.88086	0.01268	0.99990	0.01440	0.01435	0.00049	0.97669	0.99978	1.02371
0.5	0.9	0	100	1.00036	0.01479	0.99948	0.01576	0.01609	0.00155	0.97314	0.99947	1.02521
		0	1000	1.00078	0.00493	0.99988	0.00526	0.00514	0.00016	0.99116	0.99989	1.00859
		5	1000	1.17242	0.00943	1.00003	0.01001	0.01007	0.00031	0.98347	0.99992	1.01669

Table 4 Simulation results for the common factor case - uncorrelated individual effects

ρ	ρ^*_{uv}	β_0	N	$\hat{\beta}^a$	$s_{\hat{\beta}^a}$	$\hat{\beta}^c$	$s_{\hat{\beta}^c}$	$\hat{\sigma}_{\hat{\beta}^c}$	$s_{\hat{\sigma}_{\hat{\beta}^c}}$	$q^{0.05}_{\hat{\beta}^c}$	$q^{0.50}_{\hat{\beta}^c}$	$q^{0.95}_{\hat{\beta}^c}$
				T = 3								
-0.5	-0.9	0	100	0.92168	0.04152	1.00010	0.04460	0.04313	0.00642	0.92761	0.99981	1.07544
		0	1000	0.92227	0.01282	1.00024	0.01377	0.01400	0.00069	0.97732	1.00019	1.02295
		5	1000	0.86280	0.01953	0.99959	0.02092	0.02079	0.00095	0.96635	0.99974	1.03359
-0.5	0.0	0	100	0.95940	0.03429	1.00066	0.03576	0.03480	0.00478	0.94435	1.00066	1.06085
		0	1000	0.95913	0.01051	1.00011	0.01099	0.01128	0.00053	0.98239	1.00004	1.01839
		5	1000	0.95953	0.01560	1.00060	0.01636	0.01698	0.00072	0.97373	1.00085	1.02721
-0.5	0.9	0	100	0.99696	0.02475	1.00112	0.02543	0.02474	0.00317	0.96071	1.00055	1.04261
		0	1000	0.99606	0.00766	1.00013	0.00787	0.00796	0.00033	0.98694	1.00001	1.01269
		5	1000	1.05508	0.01205	1.00017	0.01241	0.01273	0.00054	0.97998	1.00013	1.02074
0.0	-0.9	0	100	0.90301	0.04485	1.00063	0.04874	0.04809	0.00727	0.92040	1.00134	1.08118
		0	1000	0.90374	0.01370	1.00017	0.01498	0.01543	0.00075	0.97615	1.00032	1.02483
		5	1000	0.83062	0.02130	1.00035	0.02323	0.02338	0.00101	0.96135	1.00011	1.03899
0.0	0.0	0	100	0.95125	0.03752	1.00247	0.03966	0.03868	0.00525	0.93777	1.00201	1.06933
		0	1000	0.94936	0.01151	1.00017	0.01221	0.01247	0.00054	0.97991	1.00024	1.02007
		5	1000	0.94927	0.01831	0.99999	0.01928	0.01896	0.00083	0.96895	0.99937	1.03203
0.0	0.9	0	100	0.99619	0.02630	1.00137	0.02722	0.02786	0.00357	0.95706	1.00117	1.04480
		0	1000	0.99530	0.00845	1.00038	0.00876	0.00891	0.00037	0.98620	1.00042	1.01513
		5	1000	1.06809	0.01302	0.99992	0.01356	0.01434	0.00058	0.97742	0.99971	1.02176
0.5	-0.9	0	100	0.84030	0.05733	1.00153	0.06412	0.06369	0.00944	0.89518	1.00084	1.10874
		0	1000	0.84089	0.01796	1.00054	0.02013	0.02046	0.00108	0.96704	1.00111	1.03269
		5	1000	0.71981	0.02709	1.00018	0.03092	0.03180	0.00155	0.94952	0.99992	1.05185
0.5	0.0	0	100	0.91643	0.04770	1.00120	0.05248	0.05084	0.00704	0.91432	1.00064	1.08891
		0	1000	0.91629	0.01411	1.00023	0.01556	0.01635	0.00079	0.97557	1.00018	1.02679
		5	1000	0.91598	0.02248	1.00001	0.02459	0.02520	0.00113	0.95859	1.00079	1.04055
0.5	0.9	0	100	0.99077	0.03442	0.99919	0.03700	0.03694	0.00493	0.93823	0.99973	1.06088
		0	1000	0.99157	0.01081	0.99994	0.01158	0.01187	0.00049	0.98069	1.00002	1.01906
		5	1000	1.11278	0.01757	1.00029	0.01924	0.01927	0.00081	0.96939	1.00010	1.03267
				T = 10								
-0.5	-0.9	0	100	0.91033	0.02136	1.00091	0.02312	0.02292	0.00243	0.96371	1.00044	1.03878
		0	1000	0.90978	0.00689	1.00021	0.00748	0.00736	0.00025	0.98809	1.00011	1.01232
		5	1000	0.84082	0.01013	0.99973	0.01094	0.01084	0.00035	0.98225	0.99953	1.01774
-0.5	0.0	0	100	0.95266	0.01726	1.00041	0.01821	0.01842	0.00187	0.97034	1.00104	1.03013
		0	1000	0.95234	0.00559	0.99992	0.00586	0.00589	0.00019	0.99030	0.99973	1.00974
		5	1000	0.95245	0.00829	1.00003	0.00871	0.00879	0.00028	0.98627	0.99991	1.01472
-0.5	0.9	0	100	0.99553	0.01250	1.00030	0.01279	0.01291	0.00126	0.97957	1.00032	1.02092
		0	1000	0.99529	0.00390	1.00003	0.00398	0.00413	0.00012	0.99345	1.00000	1.00653
		5	1000	1.06383	0.00631	1.00007	0.00649	0.00657	0.00020	0.98923	1.00008	1.01081
0.0	-0.9	0	100	0.90314	0.02193	0.99979	0.02380	0.02368	0.00246	0.96016	0.99989	1.03824
		0	1000	0.90339	0.00709	0.99977	0.00764	0.00760	0.00025	0.98729	0.99955	1.01254
		5	1000	0.83103	0.01028	1.00039	0.01132	0.01128	0.00036	0.98180	1.00044	1.01876
0.0	0.0	0	100	0.94956	0.01804	1.00038	0.01908	0.01909	0.00188	0.96886	1.00051	1.03246
		0	1000	0.94935	0.00577	1.00005	0.00611	0.00607	0.00019	0.98984	0.99996	1.01049
		5	1000	0.94918	0.00831	0.99989	0.00875	0.00908	0.00027	0.98582	0.99994	1.01412
0.0	0.9	0	100	0.99476	0.01279	0.99987	0.01320	0.01341	0.00126	0.97766	0.99969	1.02234
		0	1000	0.99496	0.00404	1.00002	0.00414	0.00427	0.00012	0.99310	1.00014	1.00687
		5	1000	1.06785	0.00654	0.99992	0.00680	0.00681	0.00020	0.98851	1.00009	1.01087
0.5	-0.9	0	100	0.88376	0.02337	1.00022	0.02520	0.02647	0.00287	0.95938	1.00069	1.04097
		0	1000	0.88413	0.00761	1.00005	0.00819	0.00846	0.00030	0.98662	0.99975	1.01396
		5	1000	0.79615	0.01146	0.99992	0.01233	0.01273	0.00043	0.97937	0.99999	1.01964
0.5	0.0	0	100	0.93864	0.01981	0.99972	0.02080	0.02096	0.00218	0.96598	0.99918	1.03464
		0	1000	0.93892	0.00625	0.99990	0.00666	0.00669	0.00022	0.98905	0.99997	1.01089
		5	1000	0.93888	0.00930	0.99986	0.00994	0.01005	0.00031	0.98393	0.99990	1.01625
0.5	0.9	0	100	0.99406	0.01409	1.00017	0.01468	0.01486	0.00147	0.97606	0.99993	1.02431
		0	1000	0.99393	0.00438	0.99999	0.00452	0.00474	0.00014	0.99271	1.00014	1.00745
		5	1000	1.08206	0.00709	1.00030	0.00732	0.00761	0.00024	0.98800	1.00019	1.01252

Table 5 Simulation results for the common factor case – correlated individual effects ($\rho_{\overline{\tau}\alpha} = 0.975$)

ρ	ρ^*_{uv}	β_0	**N**	$\hat{\beta}^a$	$s_{\hat{\beta}^a}$	$\hat{\beta}^c$	$s_{\hat{\beta}^c}$	$\hat{\sigma}_{\hat{\beta}^c}$	$s_{\hat{\sigma}_{\hat{\beta}^c}}$	$q^{0.05}_{\hat{\beta}^c}$	$q^{0.50}_{\hat{\beta}^c}$	$q^{0.95}_{\hat{\beta}^c}$
				T = 3								
-0.5	-0.9	0	100	0.91782	0.04274	1.00002	0.04607	0.04415	0.00662	0.92459	0.99971	1.07690
		0	1000	0.91896	0.01327	1.00041	0.01421	0.01426	0.00074	0.97694	1.00022	1.02420
		5	1000	0.86000	0.01952	1.00057	0.02102	0.02103	0.00096	0.96700	1.00026	1.03720
-0.5	0.0	0	100	0.95886	0.03408	1.00023	0.03581	0.03543	0.00493	0.93990	0.99959	1.05895
		0	1000	0.95916	0.01047	1.00016	0.01100	0.01145	0.00055	0.98205	0.99998	1.01846
		5	1000	0.95927	0.01607	1.00027	0.01682	0.01710	0.00075	0.97343	1.00034	1.02748
-0.5	0.9	0	100	0.99855	0.02441	0.99909	0.02508	0.02494	0.00324	0.95802	0.99823	1.04062
		0	1000	0.99947	0.00763	1.00003	0.00785	0.00802	0.00034	0.98723	0.99994	1.01283
		5	1000	1.05841	0.01208	0.99987	0.01244	0.01282	0.00055	0.97906	1.00014	1.01984
0.0	-0.9	0	100	0.89798	0.04530	1.00163	0.04921	0.04965	0.00730	0.92211	1.00204	1.08206
		0	1000	0.89771	0.01391	1.00048	0.01514	0.01605	0.00079	0.97592	1.00017	1.02563
		5	1000	0.82463	0.02116	1.00054	0.02324	0.02384	0.00111	0.96289	1.00009	1.03834
0.0	0.0	0	100	0.95034	0.03779	1.00142	0.04013	0.03985	0.00559	0.93552	1.00022	1.06809
		0	1000	0.94972	0.01195	1.00049	0.01266	0.01280	0.00061	0.97952	1.00040	1.02096
		5	1000	0.94974	0.01800	1.00056	0.01903	0.01922	0.00084	0.96878	1.00066	1.03216
0.0	0.9	0	100	1.00093	0.02660	0.99975	0.02761	0.02786	0.00352	0.95343	0.99948	1.04466
		0	1000	1.00121	0.00842	1.00010	0.00874	0.00896	0.00037	0.98566	1.00017	1.01404
		5	1000	1.07449	0.01362	1.00035	0.01413	0.01445	0.00059	0.97739	1.00037	1.02362
0.5	-0.9	0	100	0.82617	0.06098	1.00224	0.06854	0.06787	0.01118	0.88785	1.00266	1.11473
		0	1000	0.82648	0.01934	0.99981	0.02147	0.02184	0.00131	0.96506	0.99961	1.03423
		5	1000	0.70613	0.02812	1.00042	0.03185	0.03284	0.00180	0.94729	1.00008	1.05256
0.5	0.0	0	100	0.91766	0.04810	1.00294	0.05279	0.05272	0.00754	0.91648	1.00248	1.09066
		0	1000	0.91644	0.01493	1.00057	0.01640	0.01700	0.00087	0.97272	1.00060	1.02776
		5	1000	0.91625	0.02358	1.00029	0.02578	0.02570	0.00120	0.95758	0.99948	1.04234
0.5	0.9	0	100	1.00738	0.03376	1.00192	0.03639	0.03718	0.00478	0.94246	1.00035	1.06331
		0	1000	1.00589	0.01109	1.00046	0.01193	0.01193	0.00050	0.98080	1.00049	1.02032
		5	1000	1.12662	0.01777	1.00008	0.01936	0.01967	0.00085	0.96838	1.00046	1.03190
				T = 10								
-0.5	-0.9	0	100	0.90746	0.02165	1.00006	0.02321	0.02333	0.00253	0.96166	1.00008	1.03801
		0	1000	0.90774	0.00694	1.00009	0.00749	0.00747	0.00027	0.98814	1.00009	1.01199
		5	1000	0.83936	0.01051	1.00020	0.01138	0.01094	0.00037	0.98136	1.00024	1.01794
-0.5	0.0	0	100	0.95225	0.01717	0.99991	0.01808	0.01857	0.00197	0.96915	1.00003	1.02980
		0	1000	0.95226	0.00564	0.99983	0.00595	0.00595	0.00020	0.99002	0.99981	1.00944
		5	1000	0.95224	0.00849	0.99982	0.00889	0.00882	0.00029	0.98532	0.99976	1.01414
-0.5	0.9	0	100	0.99679	0.01262	0.99955	0.01297	0.01294	0.00125	0.97842	0.99915	1.02165
		0	1000	0.99728	0.00393	1.00009	0.00404	0.00414	0.00013	0.99342	1.00001	1.00678
		5	1000	1.06572	0.00634	1.00002	0.00650	0.00660	0.00020	0.98923	1.00003	1.01074
0.0	-0.9	0	100	0.89985	0.02270	0.99969	0.02437	0.02439	0.00260	0.96076	0.99965	1.03900
		0	1000	0.90014	0.00741	0.99985	0.00792	0.00781	0.00027	0.98686	0.99987	1.01319
		5	1000	0.82689	0.01054	0.99964	0.01144	0.01146	0.00038	0.98127	0.99943	1.01889
0.0	0.0	0	100	0.94907	0.01826	0.99986	0.01933	0.01923	0.00189	0.96943	0.99925	1.03314
		0	1000	0.94930	0.00580	1.00001	0.00617	0.00616	0.00020	0.98972	1.00004	1.01010
		5	1000	0.94928	0.00868	0.99997	0.00919	0.00914	0.00028	0.98492	0.99990	1.01492
0.0	0.9	0	100	0.99849	0.01258	1.00025	0.01298	0.01343	0.00127	0.97862	1.00021	1.02179
		0	1000	0.99829	0.00398	0.99997	0.00411	0.00429	0.00013	0.99295	0.99999	1.00680
		5	1000	1.07129	0.00644	0.99995	0.00663	0.00688	0.00020	0.98888	1.00001	1.01068
0.5	-0.9	0	100	0.87789	0.02487	1.00071	0.02645	0.02772	0.00312	0.95739	1.00000	1.04397
		0	1000	0.87742	0.00795	0.99991	0.00849	0.00886	0.00034	0.98612	0.99977	1.01456
		5	1000	0.78963	0.01178	1.00000	0.01259	0.01307	0.00047	0.97969	0.99996	1.02107
0.5	0.0	0	100	0.93853	0.01983	0.99962	0.02099	0.02138	0.00228	0.96545	0.99949	1.03381
		0	1000	0.93903	0.00637	1.00004	0.00676	0.00684	0.00024	0.98908	0.99985	1.01119
		5	1000	0.93888	0.00945	0.99989	0.01007	0.01016	0.00034	0.98330	0.99957	1.01666
0.5	0.9	0	100	1.00034	0.01417	0.99980	0.01476	0.01491	0.00144	0.97494	0.99980	1.02364
		0	1000	1.00054	0.00445	1.00008	0.00461	0.00475	0.00014	0.99230	1.00008	1.00766
		5	1000	1.08837	0.00752	1.00007	0.00762	0.00775	0.00024	0.98745	1.00003	1.01226

A Note on Using Multiple Singular Value Decompositions to Cluster Complex Intracellular Calcium Ion Signals

Josue G. Martinez, Jianhua Z. Huang and Raymond J. Carroll

Abstract Recently (Martinez et al. 2010), we compared calcium ion signaling (Ca^{2+}) between two exposures, where the data present as movies, or, more prosaically, time series of images. They described novel uses of singular value decompositions (SVD) and weighted versions of them (WSVD) to extract the signals from such movies, in a way that is semi-automatic and tuned closely to the actual data and their many complexities. These complexities include the following. First, the images themselves are of no interest: all interest focuses on the behavior of individual cells across time, and thus the cells need to be segmented in an automated manner. Second, the cells themselves have 100+ pixels, so that they form 100+ curves measured over time, so that data compression is required to extract the features of these curves. Third, some of the pixels in some of the cells are subject to image saturation due to bit depth limits, and this saturation needs to be accounted for if one is to normalize the images in a reasonably unbiased manner. Finally, the Ca^{2+} signals have oscillations or waves that vary with time and these signals need to be extracted. Thus, they showed how to use multiple weighted and standard singular value decompositions to detect, extract and clarify the Ca^{2+} signals. In this paper, we show how this signal extraction lends itself to a cluster analysis of the cell behavior, which shows distinctly different patterns of behavior.

1 Introduction

This paper is about understanding different patterns of behavior of calcium ion signaling (Ca^{2+}) in myometrial cells after exposure to 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD). The importance of Ca^{2+} signaling in cell function, e.g, metabolism,

Josue G. Martinez, Jianhua Z. Huang and Raymond J. Carroll
Department of Statistics, Texas A&M University, College Station TX 77843-3143, USA,
e-mail: jgmartinez@stat.tamu.edu, jianhua@stat.tamu.edu,
carroll@stat.tamu.edu

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_22,

contraction, cell death, communication, cell proliferation, has been studied in numerous types of cells; see Putney (1998). TCDD itself is a toxicant by-product of incomplete combustion of fossil fuels, woods and wastes and is known to adversely effect reproduction, development and the immune system.

The data present themselves as movies of 512 images, or time series of images after oxytocin exposure. To best appreciate the complexity of the data, readers should first look at two of the movies, available at

http://statbio.stat.tamu.edu/dataimages.php,

one without and one with TCDD exposure. The first movie is the case with TCDD exposure, "dir2_T.zip", while the second movie is without TCDD exposure, "dir2_C.zip". When unzipped, the movies are in .avi format, and are 30-40MB in size. One can view these, for example, using windows media player. The data consist of 512 images. Myometrial cells can be seen in these images, which start out in their native state and are then exposed to an oxytocin stimulus, at which point Ca^{2+} expression becomes pronounced. The cells themselves are fixed to a substrate and do not move over time. Figure 1 gives a sequence of images in the first 2 minutes of the experiment. The experiment leading to these images is described in detail in Section 2. However, the movies and Figure 1 show that the data are complex, and analysis of them is not simple.

Recently, we (Martinez et al. 2010) described methodology for extracting usable data from these movies over time. We used the singular value decomposition (SVD) in three different ways.

1. First, we used it to detect the Ca^{2+} signal by using the initial first EigenPixel vector. This approach summarizes cell location information across all 512 images instead of using only one image as is typically done for these data.
2. Second, another SVD was used to extract the Ca^{2+} signal from the pixel-wise matrix derived after segmenting the cell region in raw images. These First EigenSignal and First EigenPixel vectors serve as the templates used to "clean up" the signal.
3. Third, we used those candidate EigenSignal and EigenPixel vectors to clarify the Ca^{2+} signal by applying a new weighted SVD, the WSVD, to impute values where saturation occurs in the signal.

In this paper, we aim to show that the EigenSignal vectors can be used effectively to cluster cells within a given treatment, and thus gain insight into the different patterns of variability of cells given an oxytocin stimulus. In particular, we find that cells are characterized by the size of their initial Ca^{2+} response to the oxytocin exposure, and by how quickly the Ca^{2+} response decreases following that exposure. To do this, in Section 2, we describe the experiment. Section 3 briefly describes how one forms the EigenSignal vectors. Section 4 describes the clustering of the cells. Section 5 gives concluding remarks.

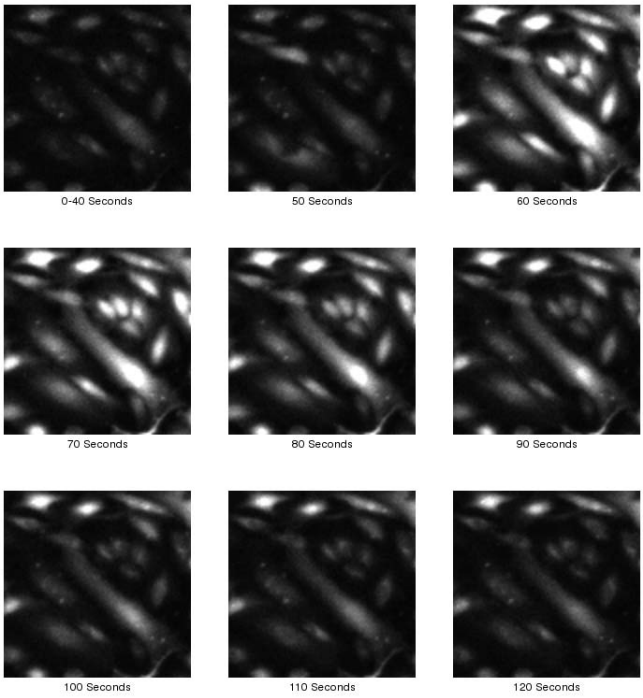


Fig. 1 Oxytocin-induced calcium response in myometrial cells during the first 2 minutes of the experiment. Cells were cultured in a low level of estrogen/progesterone and were treated with 10 nM TCDD for 24 hr.

2 Experiment

The essential statistical details of this experiment are that there are myometrial cells fixed to different substrates, one group of which is exposed to TCDD and the other group is not. Shortly after image capturing commences, the cells are exposed to oxytocin, thus stimulating the Ca^{2+} signal. Our main goal is to understand how cells within a given treatment respond to the stimulus: Martinez et al. (2010) focused on comparing the TCDD exposure to the control. What follows are some of the details of the experiment.

2.1 Treatments

Myometrial cells, which comprise the contractile middle layer of the uterine wall, were cultured in three levels of an estrogen/progesterone hormone combination: basal, low and high. The “basal" level is the one in which the cells where cultured, the

"low" level of hormone is slightly higher than that found in women before pregnancy and the "high" level is the level of a pregnant woman at full term. Our clustering work focuses on the high level of hormone and uses data from two different treatments, control or TCDD.

The treated cells received a 100 nM solution of TCDD 24 hours before the experiment. Cells were cultured on coverglass chambered slides. All cells were then washed and loaded with 3 μM Fluo-4 for 1 hour at 37°C: fluorescent probe Fluo-4 is one of many dyes used to detect changes in Ca^{2+} within cells. Fluo-4 is typically excited by visible light of about 488 nm, and emits about 100 fold greater fluorescence at about 520 nm upon binding free Ca2+. Following loading, cells were washed and placed on the stage of the confocal microscope. Cells were then scanned five times to establish the basal level of Ca^{2+} prior to addition of 20 nM oxytocin, the hormone used in this study to stimulate Ca^{2+} signal in these cells. Scanning continues at 10 second intervals for approximately 85 minutes, leading to 512 images (100×100 pixels) containing 20–50 cells per treatment.

2.2 Imaging

The data captured in these experiments are digital images of Ca^{2+} fluorescence of individual cells. The bit depth of images used in this study is of 8 bits, which translates to 2^8 or 256 possible grayscale values in the image. Unfortunately, it often happens that the maximum concentrations detected in these images are limited by the bit depth. This may sometimes result in saturation and lead to underestimation of changes in Ca^{2+} signals, especially when multiple treatments are performed and accurate evaluation of these differences is required.

Figure 1 shows a response to the oxytocin stimulus, in cells treated with TCDD and cultured in a low estrogen/progesterone hormone level. The reaction due to the oxytocin challenge appears maximal at 60 seconds and then the cells return to their steady state. Notice that not all cells go back to their steady state at the same rate. In fact, there is residual fluorescence in some cells at the top of each of the images in Figure 1, long after the initial peak of fluorescence at 60 seconds.

3 Methods

This section describes the methodology of Martinez et al. (2010) used to obtain the EigenSignal and EigenPixel vectors. Effectively, the algorithm works as follows.

- Do a rough segmentation of each cell.
- Apply the SVD to extract initial EigenSignal and EigenPixel vectors.
- Use these initial vectors to perform more refined segmentation.
- Use a weighted SVD to account for image saturation.

3.1 EigenPixels and EigenSignals

We first describe how to obtain “eigen pixel" and “eigen Ca^{2+} signal" vectors, using the SVD. To accomplish this, we will present the singular value decomposition in the context of our data, assuming that a rough segmentation of the cells has been performed. We represent each cell as a matrix of Ca^{2+} intensity, in grayscale values, that has a number of pixels which comprise the cell, for all 85 minutes of the experiment. Each matrix has n rows and m columns, where n is the number of pixels that represent the cell and m is the number of time points in the experiment. All cells were observed the same number of times so $m = 512$. Let X_k represent the $n \times m$ calcium signal matrix for the kth cell. The singular value decomposition of X_k is

$$X_k = U_k S_k V_k^T. \tag{1}$$

Here V_k is a $m \times n$ matrix whose column vectors, $\mathbf{v}_{kj} \in \mathbb{R}^m$, form an orthonormal basis for the Ca^{2+} signal, and are called EigenSignal vectors. In (1), U_k is a $n \times n$ matrix whose column vectors, $\mathbf{u}_{kj} \in \mathbb{R}^n$, form an orthonormal basis for the pixels of the cell, called EigenPixel vectors. In addition, S_k is a $n \times n$ square matrix of singular values arranged from largest to smallest $s_{k1} \geq s_{k2} \geq ... \geq s_{kn}$.

We can generate a rank-L matrix that approximates X_k by using the first L u_{kj} and v_{kj} vectors, i.e.

$$X_k^L = \sum_{j=1}^{L} u_{kj} s_{kj} v_{kj}^T. \tag{2}$$

In equation (2), X_k^L is the best rank-L matrix that approximates X_k, in the sense that it minimizes the sum of squares difference between X_k^L and X_k among all rank-L matrices, see Trefethen & Bau (1997). Low rank approximations are useful because less data are needed to represent the original matrix; these techniques are often used in image compression. We will use the smallest number of EigenPixel and EigenSignal vectors that summarize both pixel and Ca^{2+} signal information.

3.2 Ca^{2+} Rough Segmentation

As may be apparent from the sequence of images shown in Figure 1, it is difficult to distinguish cell boundaries before oxytocin is delivered. For this reason, in order to determine the location of the cells, as well as their boundaries, it is common to use the brightest image to isolate the cells. We then draw large rectangular regions each containing a cell. If X_k represents the matrix of pixels $\times$ time, we obtain a summary of the pixel information by taking the SVD of X_k and obtaining the first EigenPixel. In the data, the first singular value explains the majority of the variance in these data, hence the first EigenPixel summarizes all the pixel information to one vector. We take this vector and plot it spatially on the corresponding pixel location. What we get is a 2-dimensional image where the pixel intensity reflects the importance of the

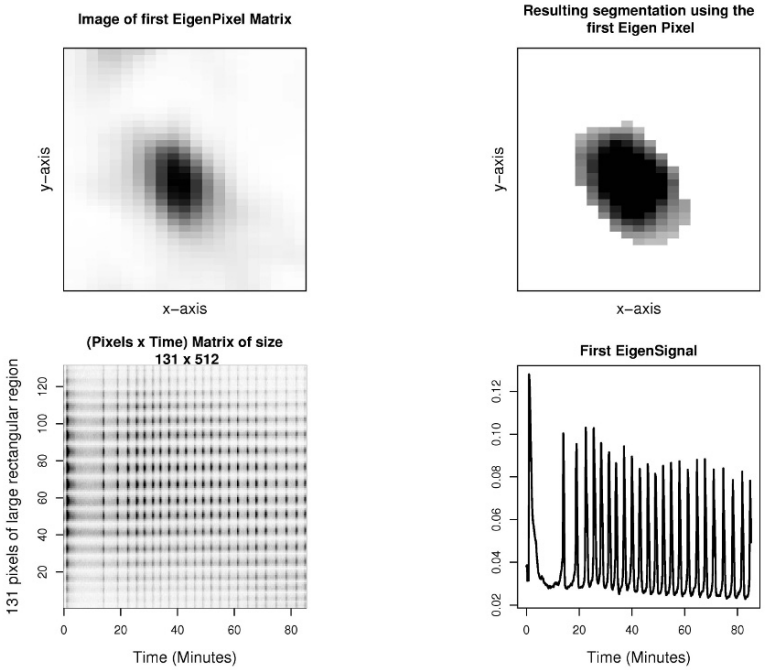


Fig. 2 Top Row: Image of the First EigenPixel vector obtained from the SVD of the rough 777 × 512 pixel-wise matrix and the resulting segmentation of cell 2 after using the First EigenPixel to perform the segmentation. Bottom Row: The corresponding 131 × 512 pixel-wise matrix for this new segmentation and the corresponding first EigenSignal over the 85 minute experiment.

pixel in representing the Ca^{2+} signal of this cell (Top left panel of Figure 2). This image is a better candidate for use in identification of the Ca^{2+} signal than the "peak" image because it summarizes the importance of each pixel across the 512 images in the experiment.

3.3 Ca^{2+} *Final Segmentation*

Once we obtain this first EigenPixel image from X_k we use the EBImage package from Bioconductor to segment and index the cell, R Development Core Team (2008). We first blurred the image to smooth out any noisy pixels. We then used thresholding to pick out the region of high pixel values which usually contains the cell, and finally used a watershedding algorithm to close the cell boundaries and separate other cell chucks that are close together. The result is the final segmentation of the cell shown in the top right panel of Figure 2. In effect, we have chosen the region with highest EigenPixel intensity which in turn should give us the spatial location of the pixels that contain most of the Ca^{2+} signal information. We then collect each of the 131 pixels in

this final segmentation from each of the 512 images and get a matrix representation of the cell, see the bottom left panel of Figure 2. We used this segmentation process to generate contours of each cell, and used these contours to pick out the cell position from every image at every one of the 512 time points. This process yielded 20 to 50 cells from each treatment. In our data, we found that the first EigenSignal and EigenPixel vectors that correspond to this first singular value, summarize almost all the Ca^{2+} signal and pixel information in each of these matrices.

The oscillatory behavior observed in Figures 2 and 3 is present because calcium ions (Ca^{2+}) are responsible for many important physiological functions. In smooth muscle cells that surround hollow organs of the body, transient increases in intracellular Ca^{2+} can be stimulated by a number of hormones to activate smooth muscle contraction. Because sustained elevation of Ca^{2+} is toxic to cells, Ca^{2+} signals in many cell types frequently occur as repetitive increases in Ca^{2+}, referred to as Ca^{2+} oscillations. The periodic Ca^{2+} spikes which increase with increasing hormone concentration are thought to constitute a frequency encoded signal with a high signal-to-noise ratio which limits prolonged exposure of cells to high intracellular Ca^{2+}, see Sneyd et al. (1995). Interestingly, the frequency of Ca^{2+} oscillations in smooth muscle cells is relatively low (e.g., 2 to 10 mHz), see Burghardt et al. (1999), whereas in liver cells which use Ca^{2+} oscillations to stimulate ATP production in mitochondria and the breakdown of glycogen to glucose, the frequency of Ca^{2+} oscillations is much greater (e.g., range from 5 to 100 MHz), see Barhoumi et al. (2002, 2006). The spatial and temporal organization and the control of these intracellular Ca^{2+} signals is of considerable interest to cellular biologists.

3.4 Cell Saturation and the Weighted SVD

The grayscale values of some of the pixels that represent cell 2, shown in Figure 2, reach a ceiling of 255, see Figure 3. The bottom panel of Figure 3 shows the intensity of 20 pixels over time and it is clear that some reach a maximum intensity at values that are larger than 255 and some at much lower values. Martinez et al. (2010) implement a novel weighted SVD, WSVD, using the low rank matrix approximation of Gabriel & Zamir (1979) where they introduce the use of indicators in the weights, as in Beckers & Rixen (2003), to treat the saturated pixels as missing data, with a clever choice of weights that allows for accurate recovery of the original signal.

Briefly, the method works as follows. Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be the first EigenPixel and EigenSignal associated with the second SVD used to extract the putative Ca^{2+} signal, which includes saturated pixels, so that $\boldsymbol{u}$ and $\boldsymbol{v}$ comprise most of the pixel and signal information. The matrix of interest is X_k'. Let the dimensions of the X_k' be $n \times m$; because most of the variation is explained by the first component in the SVD the rank one approximation can be obtained by minimizing the error

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(x'_{kij} - u_i v_j)^2 \tag{3}$$

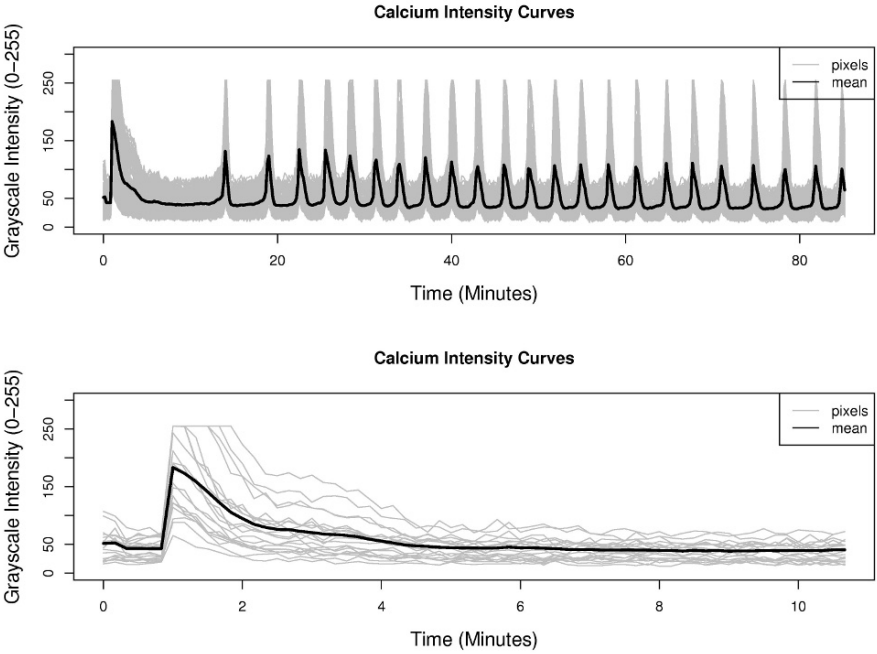


Fig. 3 Top: The Calcium intensity curves over the 85 minute experiment of the 131 pixels in the 131×512 pixel-wise matrix $X_k^{'}$. Bottom: 20 randomly selected pixels from $X_k^{'}$.

with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$. We also wish to weight each term in the double summation so that it removes the influence of saturated pixels and takes into account the appropriate variation. We let the weights be $w_{ij} = \mathrm{I}_{ij}/(u_i v_j)^2$, where $\mathrm{I}_{ij} = 0$ when a pixel is saturared, i.e. $x'_{kij} = 255$, and $\mathrm{I}_{ij} = 1$, otherwise. The minimization problem becomes

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}(x'_{kij} - u_i v_j)^2. \tag{4}$$

We solve the minimization by alternating between u_i and v_j. Fixing j, we can expand the expression in (4), let $A_j(\boldsymbol{u}) = \sum_i \mathrm{I}_{ij}(x'_{kij}/u_i)^2$ and $B_j(\boldsymbol{u}) = \sum_i \mathrm{I}_{ij}(x'_{kij}/u_i)$ and we get that $v'_j = A_j(\boldsymbol{u})/B_j(\boldsymbol{u})$ solves that portion of the minimization. Similarly if we fix i, $u'_i = A_i(\boldsymbol{v})/B_i(\boldsymbol{v})$, where $A_i(\boldsymbol{v}) = \sum_j \mathrm{I}_{ij}(x'_{kij}/v_j)^2$ and $B_i(\boldsymbol{v}) = \sum_j \mathrm{I}_{ij}(x'_{kij}/v_j)$. The new proposed EigenPixel and EigenSignal vectors are $\boldsymbol{u}^{new} = \boldsymbol{u}'/\|\boldsymbol{u}'\|$ and $\boldsymbol{v}^{new} = \boldsymbol{v}'/\|\boldsymbol{v}'\|$ respectively. This gives us a recurrence relation that can used to obtain a clearer version of the EigenPixel and EigenSignal, where the EigenSignal will represent the clarified Ca^{2+} signal of interest. The missing values are imputed by the corresponding $u_i v_j$ after the convergence of the algorithm. We check to make sure that any imputed value for initially saturated pixels does not fall below its saturated value.

4 Clustering

We now aim to understand how the cells within a given treatment respond to the oxytocin exposure. Here we investigate the high hormone level data, both with and without TCDD exposure. To do clustering, we take the final first EigenSignal vectors described in Section 3.4, but restricted to the first 2 minutes after exposure, the so-called “peak" time where the response is the greatest. In other analyses, we found that the greatest difference between the control and TCDD-treated cells occurred in this peak period, at least for peak height and peak area under the curve. Our interest is in understanding if the two groups have structures within themselves, and if the structures are comparable. For clustering, we used the “pam( )" function in the “cluster" package in R to cluster the first EigenSignal vectors. This function implements a more robust version of k-means clustering, and we used an L1 metric to measure distances between EigenSignal vectors.

We first normalized the data so that the initial values were 1.0, and hence what is observed is a type of fold change. Figures 4 and 5 give the results of clustering the first EigenSignal vectors with 3 clusters in both the control and the TCDD-treated cells: the former gives a plot, while the latter gives a heat map of the cells within a cluster over time. Roughly, in both control and TCDD-treated, looking at either the line plots or the heat maps, there are three modes of action, which can be characterized as (a) much reduced peak heights (Cluster 3 in the control and 2 in the TCDD-treated); (b) less rapid decline over time for those with higher peak heights, with for example cluster 1 in the controls and cluster 3 in the TCDD-treated cells showing less decline over time. Looking at the sizes of the clusters, it is immediately apparent that the control cells have overall a greater response than the TCDD-treated cells, a finding shown to be statistically significant in Martinez et al. (2010).

5 Conclusion

Martinez et al. (2010) developed a methodology for extracting low-dimensional representations from images taken over time of cells. The methodology includes cell segmentation, normalization, and imputation to avoid image saturation.

In this note, we have shown that such methods can also be used to cluster cells, and to learn more about the different patterns of response among a group of cells given the same treatment. In the data analysis, the basic responses to oxytocin exposure were qualitatively the same with three modes of action, but the TCDD-exposed cells and the control cells differed in the frequency of the modes of action.

Acknowledgements Martinez was supported by a postdoctoral training grant from the National Cancer institute (CA90301). Carroll and Huang’s research was supported by a grant from the National Cancer Institute (CA57030). Huang was also supported by a grant from the National Science Foundation (DMS-0606580).

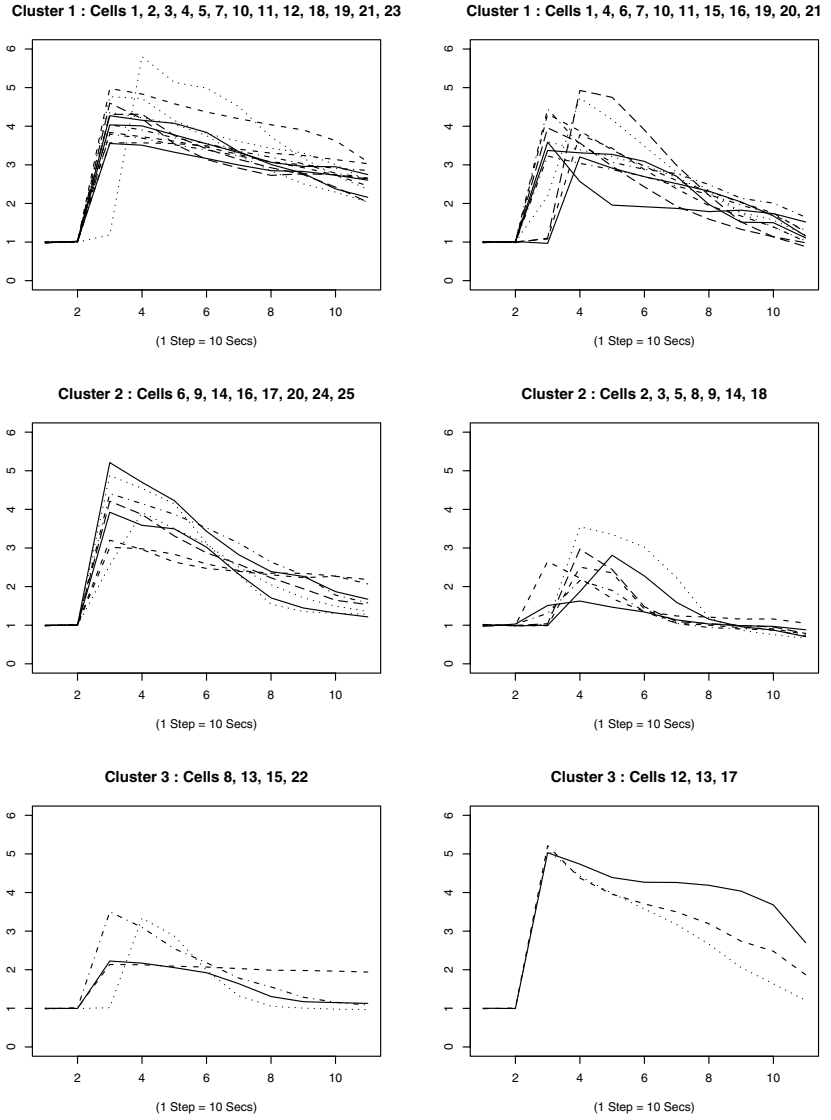


Fig. 4 Three clusters of the first EigenSignal vectors. The clusters on the right are those from cells treated with TCDD, while those on the left are controls. All were from the culture grown in the Low Hormone level.

References

Barhoumi, R., Awooda, I., Mouneimne, Y., Safe, S. & Burghardt, R. C. (2002). Effects of benzo-a-pyrene on oxytocin-induced Ca2+ oscillations in myometrial cells. *Toxicol Lett* **165** 133–141.

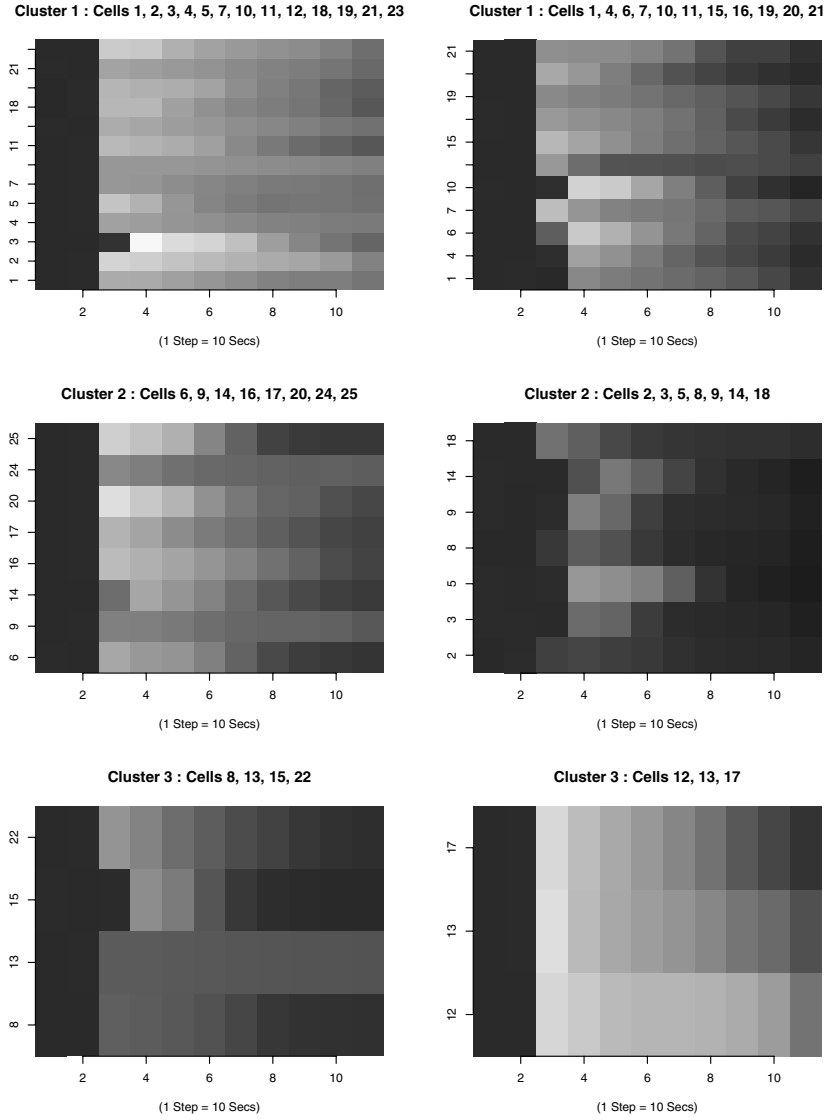


Fig. 5 Matrix plot of the three clusters of the first EigenSignal vectors. The clusters on the right are those from cells treated with TCDD, while those on the left are controls. All were from the culture grown in the Low Hormone level.

Barhoumi, R., Mouneimne, Y., Awooda, I., Safe, S., Donnelly, K. C. & Burghardt, R. C. (2006). Characterization of Calcium Oscillations in Normal & Benzo[a]pyrene-Treated Clone 9 Cells. *Toxicological Sciences* **68** 444–450.

Beckers, J. & Rixen, M. (2003). Eof calculations and data filling from incomplete oceanographic datasets. *J. Atmos. Oceanic Technol.* **20** 1839–1856.

Burghardt, R. C., Barhoumi, R., Sanborn, B. M. & Andersen, J. (1999). Oxytocin-induced Ca2+ responses in human myometrial cells. *Biol Reprod* **60** 777–782.

Gabriel, K. R. & Zamir, S. (1979). Lower Rank Approximation of Matrices by Least Squares with any Choice of Weights. *Technometrics* **21** 489–498.

Martinez, J. G., Huang, J. Z., Burghardt, R. C., Barhoumi, R. & Carroll, R. J. (2010). Use of multiple singular value decompositions to analyze complex intracellular calcium ion signals. *Annals of Applied Statistics*, to appear.

Putney, J. W. (1998). Calcium signaling: up, down, up, down...what's the point? *Science* **279** 191–192.

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`.

Sneyd, J., Keizer, J. & Sanderson, M. J. (1995). Mechanisms of calcium oscillations and waves: A quantitative analysis. *FASEB Journal* **9** 1463–1472.

Trefethen, L. N. & Bau III, D. (1997. *Numerical Linear Algebra* SIAM.

On the self-regularization property of the EM algorithm for Poisson inverse problems

Axel Munk and Mihaela Pricop

Abstract One of the most interesting properties of the EM algorithm for image reconstruction from Poisson data is that, if initialized with a uniform image, the first iterations improve the quality of the reconstruction up to a point and it deteriorates later dramatically. This 'self- regularization' behavior is explained in this article for a very simple noise model. We further study the influence of the scaling of the kernel of the operator involved on the total error of the EM algorithm. This is done in a semi- continuous setting and we compute lower bounds for the L^1 risk. Numerical simulations and an example from fluorescence microscopy illustrate these results.

1 Introduction

1.1 The EM algorithm for Poisson inverse problems

In this article we consider the problem of estimating an intensity λ of a spatial Poisson point process with the help of the EM-ML algorithm based on n independent realizations of another spatial point process, whose intensity μ is related to λ with the help of an integral operator

$$A\lambda = \mu \tag{1}$$

where $A : L^1(\Omega) \to L^1(\Sigma)$ is a linear integral operator with positive kernel $a : \Sigma \times \Omega \to \mathbf{R}$ and $\Omega \subset \mathbf{R}^d$, $\Sigma \subset \mathbf{R}^l$ are closed, bounded subsets of the corresponding Euclidean spaces $\mathbf{R}^d$, respectively $\mathbf{R}^l$.
We have at our disposal the data $\{Y_i\}_{i=1,\ldots,n}$ which are modelled as Poisson dis-

Axel Munk and Mihaela Pricop
Institut für Mathematische Stochastik, Georg August Universität Göttingen, 37077 Göttingen, Germany e-mail: munk@math.uni-goettingen.de, pricop@math.uni-goettingen.de

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_23,

tributed random variables with expectation $\mu(t_{ni}) = \int_\Omega a(t_{ni}, s)\, d\lambda(s)$, where t_{ni} represent the discretization points in the data space corresponding to the sample size n. Estimating the intensity function λ in this setting is a classical theme and has been investigated by various communities. Areas of application include positron emission tomography (Vardi, Shepp & Kaufman 1985), molecular microscopy or various problems in astrophysics (Bertero & Boccacci 1998), (Meinshausen, Rice & Schücker 2006), (Bertero, Boccacci, Desiderà & Vicidomini 2009) among others. Nowadays, there is a vast amount of reconstruction methods and algorithms available, many of them rely on penalisation techniques, also in combination with fast algorithms (see (Natterer 2001) and the references given there). A minimax approach for this statistical inverse problem was studied in (Johnstone & Silverman 1990), (Koo & Chung 1998) or (Korostelëv & Tsybakov 1993). Wavelet- based methods were discussed in (Cavalier & Koo 2002) and (Antoniadis & Bigot 2006).
One of the most prominent algorithms in this context has been suggested by Vardi, Shepp and Kaufman (1985) and results from the expectation maximization (EM) algorithm, introduced by Dempster, Laird and Rubin (1977) in a more general context. In the Poisson set up, this algorithm is also denoted as Richardson- Lucy algorithm. In the above paper the problem is stated in a discrete formulation, i.e. the data $\{Y_i\}_{i=1\ldots n}$ is modeled as random variables with distribution $Po((A\lambda)_i)$, a Poisson distribution with intensity $(A\lambda)_i$ and $\lambda = (\lambda_1, \ldots, \lambda_m)$ being a vector. Here $(A\lambda)_i = \sum_{j=1}^m a_{ij}\lambda_j$ and the matrix A fulfills

$$\sum_{i=1}^n a_{ij} = 1$$
$$a_{ij} \geq 0, \forall i = 1, \ldots, n, j = 1, \ldots, m,$$

and we assume w.l.o.g. that

$$\sum_{j=1}^m \lambda_j = 1.$$

We are interested in estimating the discrete unknown intensity $\{\lambda_j\}_{j=1,\ldots,m}$ of this Poisson process which models the emission density in medical imaging techniques like PET. The emission space is seen as a grid with λ_j constant in each box j. The 'exact solution' for each discretization is going to be denoted with $\{\lambda_j\}_{j=1,\ldots,m}$ and the discretization of the matrix A with a_{ij}.The number of total emissions in box j has mean $\sum_{i=1}^n \lambda_j a_{ij} = \lambda_j$.
This is a typical estimation problem from incomplete data which can be solved with the help of the EM algorithm. The complete data is represented by the number of emissions in box j detected in tube i $\{N_{ij}\}_{i=1\ldots n, j=1\ldots m}$, which is unobservable. The likelihood function is

$$L_N(\lambda) = \prod_{i=1}^n \prod_{j=1}^m \exp(-\lambda_j a_{ij}) \frac{(\lambda_j a_{ij})^{N_{ij}}}{N_{ij}!}$$

and the log-likelihood function is

$$l_N(\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m} -\lambda_j a_{ij} + N_{ij}\log(\lambda_j a_{ij}) - \log(N_{ij}!)$$
$$= -\sum_{j=1}^{m}\lambda_j + \sum_{i=1}^{n}\sum_{j=1}^{m} N_{ij}\log(\lambda_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} N_{ij}\log(a_{ij}) - \log(N_{ij}!).$$

In the estimation step of the EM algorithm we start with an initial value f^{old} for λ and we compute the conditional expectation as

$$\mathbf{E}(N_{ij} \mid \{Y_i\}_{i=1,\dots,n}, f^{old}) \overset{(1)}{=} \mathbf{E}(N_{ij} \mid f^{old}, Y_i) \overset{(2)}{=} f_j^{old}\frac{Y_i a_{ij}}{\sum_k f_k^{old} a_{ik}}.$$

In (1) we used that $\{Y_i\}_{i=1,\dots,n}$ are independent random variables and in (2) that, since for any $i = 1,\dots,n$ $\{N_{ij}\}_{j=1,\dots,m}$ are independent Poisson distributed random variables with parameter $f_j^{old}a_{ij}$ and $\sum_{j=1}^m N_{ij} = Y_i$, the conditional expectation of N_{ij} given $\sum_{j=1}^m N_{ij}$ is a binomial random variable with parameters $\left(Y_i, \frac{f_j^{old}a_{ij}}{\sum_{j=1}^m f_j^{old}a_{ij}}\right)$ and $\mathbf{E}(N_{ij} \mid f^{old}, Y_i) = f_j^{old}\frac{Y_i a_{ij}}{\sum_{j=1}^m f_j^{old}a_{ij}}$. After estimating N_{ij} with $\mathbf{E}(N_{ij} \mid f^{old}, \{Y_i\}_{i=1,\dots,n})$, in the maximization step we choose N_{ij} as the maximum likelihood estimator in $l_{\mathbf{E}(N_{ij}|\{Y_i\}_{i=1,\dots,n},f^{old})}(\cdot)$, which is again equal to $\mathbf{E}(N_{ij} \mid f^{old}, Y_i) = f_j^{old}\frac{Y_i a_{ij}}{\sum_{j=1}^m f_j^{old}a_{ij}}$. Hence, we obtain

$$f_j^{new} = \mathbf{E}(\sum_{i=1}^{n} N_{ij} \mid f^{old}, \{Y_i\}_{i=1,\dots,n}) = \sum_{i=1}^{n}\mathbf{E}(N_{ij} \mid f^{old}, \{Y_i\}_{i=1,\dots,n})$$
$$= f_j^{old}\sum_{i=1}^{n}\frac{Y_i a_{ij}}{\sum_{k=1}^m f_k^{old}a_{ik}},\ j = 1,\dots,m. \tag{2}$$

This is a gradient-type algorithm which maximizes the log- likelihood of the data $\{Y_i\}_{i=1,\dots,n}$ with respect to $\{\lambda_j\}_{j=1,\dots,m}$. We remark that

$$\sum_{j=1}^{m} f_j^{new} = \sum_{j=1}^{m} f_j^{old}\sum_{i=1}^{n}\frac{a_{ij}Y_i}{\sum_{k=1}^m a_{ik}f_k^{old}}$$
$$= \sum_{i=1}^{n} Y_i\sum_{j=1}^{m}\frac{a_{ij}f_j^{old}}{\sum_{k=1}^m a_{ik}f_k^{old}} = \sum_{i=1}^{n} Y_i. \tag{3}$$

We also notice that, since $\{Y_i\}_i$ are independent, Poisson distributed random variables, $\sum_{i=1}^n Y_i$ is also Poisson distributed with parameter

$$\sum_{i=1}^{n}(A\lambda)_i = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\lambda_j = \sum_{j=1}^{m}\lambda_j\sum_{i=1}^{n} a_{ij} = \sum_{j=1}^{m}\lambda_j = 1.$$

This algorithm can be extended in an obvious manner to the model (1). In fact, the continuous version of the EM algorithm for positive linear integral equations was proposed for the first time in (Kondor 1983). In (Shepp & Vardi 1982) a semi-continuous model was proposed for emission tomography, where the intensity of the Poisson process λ is a density. In this case the random variables $\{Y_i\}_{i=1\ldots n}$ are distributed as $Po((A\lambda)_i)$, where $(A\lambda)_i = \int_\Omega a(t_{ni},s)\lambda(s)\,ds$, the kernel $a_i(s) = a(t_{ni},s)$ fulfills

$$\sum_{i=1}^{n} a_i(s) = 1\, a.e. \qquad (4)$$
$$a_i(s) \geq 0\, a.e.,$$

and it holds

$$\int_\Omega \lambda(s)\,ds = 1\, a.e. \qquad (5)$$

From Kuhn- Tucker conditions for maximizing the log- likelihood of the data we obtain in analogy to (2) the fixed- point semi-continuous EM algorithm

$$f_{k+1} = f_k \sum_{i=1}^{n} \frac{a_i Y_i}{\int_\Sigma a_i(s) f_k(s)\,ds}, \qquad (6)$$

where $f_0 \in L^1(\Omega)$ is a positive initial guess.
In (Mair, Rao & Anderson 1996) the authors remark that the emission intensity is not necessarily a density and the theoretical properties of the EM algorithm can be extended to the continuous case for the weak topology. Hence, a more general approach is studied in this paper. In this case we want to reconstruct the intensity λ of a Poisson process as a finite signed Borel measure belonging to the Banach space $(\mathrm{B}, \|\cdot\|_\mathrm{B})$ of Borel measures on Ω with the total variation norm $\|\lambda\|_\mathrm{B} = \sup\{\lambda(C) : C \subseteq \Omega \text{ Borel set}\}$. Now, we have independent Poisson distributed random variables $\{Y_i\}_{i=1,\ldots,n}$ with parameters $\int_\Omega a_i(s)\,d\lambda(s)$, with $a_i \geq 0$, $\sum_{i=1}^n a_i(s) = 1$ a.e. and $\int_\Omega d\lambda = 1$. It follows that $\sum_{i=1}^n Y_i$ is also a Poisson random variable with parameter $\sum_{i=1}^n \int_\Omega a_i(x)\,d\lambda(x) = \int_\Omega \sum_{i=1}^n a_i(x)\,d\lambda = \int_\Omega d\lambda = 1$. The EM algorithm is written as

$$f_{p+1}(s) = f_p(s) \sum_{i=1}^{n} \frac{Y_i a_i(s)}{(T f_p(s))_i} \qquad (7)$$

where $T : \mathrm{B} \to \mathbf{R}^n$, $(T\lambda)_i = \int_\Omega a_i(x)\,d\lambda(x)$. We remark that f_p is a random measure i.e. a measurable mapping from a probability space $(\Gamma, K, \mathbf{P})$ to $(\mathrm{B}, \|\cdot\|_\mathrm{B})$. We endow the space of random measures with the metric

$$\beta(\lambda,\nu) = \int_\Gamma \|\lambda(s) - \nu(s)\|_\mathrm{B}\, d\mathbf{P}(s),$$

see (Crauel 2002). Similar to the finite dimensional case we remark that $\mathbf{P}-$almost surely

$$\begin{aligned}\int_\Omega df_{p+1} &= \int_\Omega \sum_{i=1}^{n} \frac{Y_i a_i(s)}{(Tf_p)_i} df_p \\ &= \sum_{i=1}^{n} \frac{Y_i}{(Tf_p)_i} \int_\Omega a_i df_p = \sum_{i=1}^{n} Y_i. \end{aligned} \tag{8}$$

Note again, the analogy to the basic algorithm (2) and the equality (3).

1.2 Convergence Properties

The *numerical* convergence of the EM algorithm in the discrete model towards the maximum likelihood estimator was investigated in a series of papers see e.g. (Vardi, Shepp & Kaufman 1985), (Csiszár & Tusnády 1984), (Cover 1984) or (Iusem 1992) and is satisfactorily understood nowadays. However, the more subtle issue is the investigation of statistical properties of this estimator. This turns out to be a surprisingly difficult problem and we are far from a theoretical understanding. Since computing the maximum likelihood estimator in the continuous case is an ill-posed problem, letting the EM algorithm converge to the MLE will give an inconsistent estimator in general, when the dimension of the underlying parameter space is too large compared to the number of observations. This inconsistency of the algorithm was noticed in (Anderson, Mair & Rao 1996). This problem can be overcome by additional regularization of the maximum likelihood estimator e.g. by smoothing every step of the EM algorithm like in (Eggermont 1999), (Latham & Anderssen 1994), (Silverman, Jones, Nychka & Wilson 1990) or by introducing a penalization term into the EM algorithm like in (Eggermont& LaRiccia 1996), (Fessler 1994) and many others. Nonetheless, the (unpenalized) EM algorithm (and faster variants thereof) is used in practice (Natterer 2001), (Hudson & Larkin 1994), (Ahn & Fessler 2003) and typically it is terminated in an early stage of iteration, tending to surprisingly good results of reconstruction. The typical situation (observed and documented in numerous simulation studies and confirmed on many real life applications see e.g. (Shepp & Vardi 1982), (Bissantz , Mair & Munk 2008), (Vardi & Lee 1993), (Mair, Rao & Anderson 1996), (Szkutnik 2000) and the references given there) is the following. If the EM algorithm is initialized with a uniform image, the iterates initially improve, then after a certain point gradually deteriorate in appearance and accuracy. We can observe in Figure 1 and Figure 2 this behavior of the EM algorithm for both simulated and experimental Poisson data for a convolution and uniform initial guess.
The source for the experimental data is 4π- microscopy which aims for a reconstruction of a certain protein distribution in a cell (see e.g. (Lang, Müller, Engelhard & Hell 2007) and (Vicidomini, Hell & Schönle 2009)). In our experimental data, the distribution of the trans-Golgi protein taken from the VERO cells of a green African

monkey is studied. The aim of the research is to develop a vaccine against a broad range of viruses.

1.3 Self-Regularization

We would like to call the above described property *self-regularization* due to the fact that iteration itself has the effect of regularization up to a certain point. Subsequent iterations tend to undersmoothing until finally the ML estimator is achieved which, due to the ill posedness of the problem, lacks consistency, in general. A similar phenomenon for other iterative algorithms has been recently investigated by (Bissantz, Hohage, Munk & Ruymgaart 2007) and in the context of boosting algorithms by (Bühlmann & Yu 2003). For the linear algorithms considered there it is possible to compute the mean square error as a function of the iteration step k and to compute the minimizing $k = k(n, \lambda, \sigma)$, which typically depends on the sample size n, the smoothness of the true signal λ, the noise level σ and the regularization properties of the underlying algorithm. Early stopping of these algorithms introduces already a regularization step, which can be shown to be minimax-optimal in certain settings (see (Bissantz, Hohage, Munk & Ruymgaart 2007)).
For the EM algorithm in the present setting, a similar behavior can be observed numerically, however a theoretically understanding is very difficult due to the nonlinearity of the EM algorithm in the Poisson case, i.e. the algorithm can not be written as an iteration of a linear operator T acting on the data, $f^{n+1} = T^k f^n$.

1.4 Stopping Rules

This observation has initiated some research how to stop the EM properly. We will give a brief overview. One approach in choosing the stopping rule for the EM algorithm is based on statistical hypothesis testing. A first stopping criterion based on Pearson's criterion test was proposed in (Veklerov & Llacer 1987). In (Hebert, Leahy & Singh 1988) a new rule is proposed for approximating the maximum likelihood estimator by a stopped EM algorithm, using the Poisson distribution of the data which allows the use of Pearson's criterion for multinomial distribution. Herbert (1990) proposed another stopping rule using Wilcoxon's signed rank statistics. These methods suffer from lack of robustness, which was overcome to some extent through the multi- scale rule proposed in (Bissantz , Mair & Munk 2005). Roughly speaking this criterion tests simultaneously at all points whether the residuals are consistent with the distribution of the noise.
Cross- validation as a stopping rule for the EM algorithm was proposed in (Coakley 1991) and (Coakley & Llacer 1991) and implemented in (Llacer, Veklerov, Coakley, Hoffman & Nunez 1993). A single set of data is randomly split to produce two (or

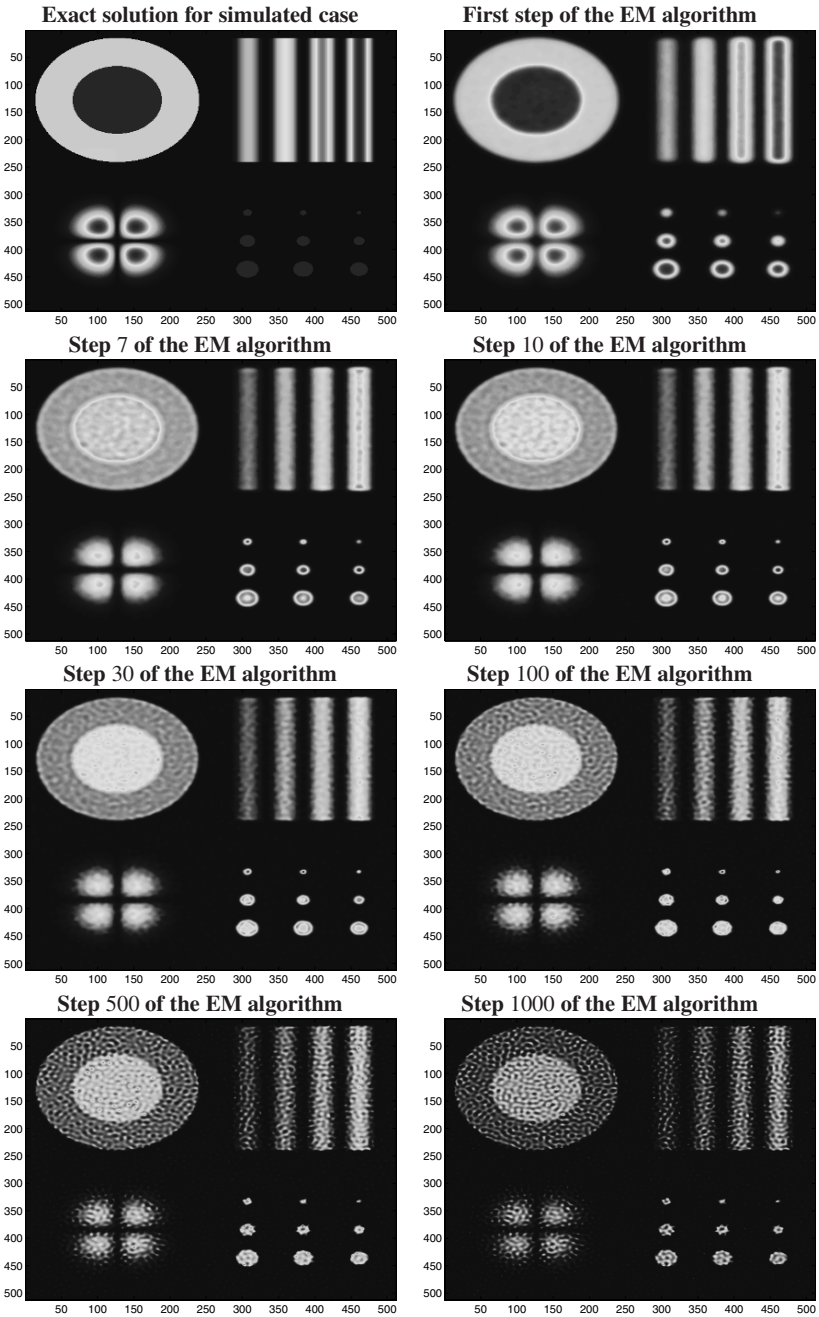


Fig. 1 Evolution of the EM algorithm for simulated data with convolution operator and uniform initial guess

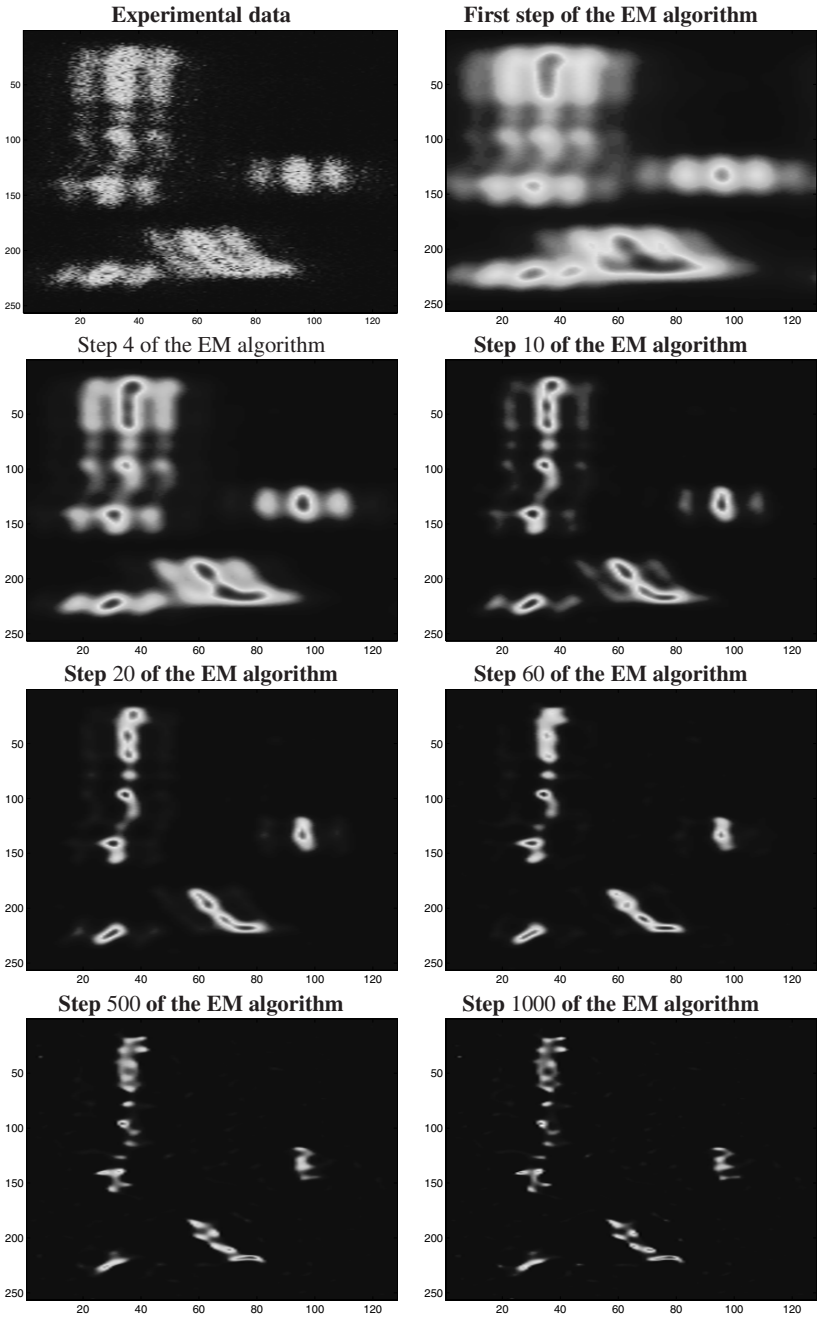


Fig. 2 Evolution of the EM algorithm for experimental data with convolution operator and uniform initial guess

more) data sets possessing similar statistical properties. The estimate corresponding to the full data is obtained by adding the image estimates for each data set. This stopping rule proves to be robust in practical applications but shows difficulties at high count levels. To overcome this, in (Johnson 1994) a jackknife criterion and in (Coakley 1996) a bootstrap method is proposed.
Another approach, using the limit properties of the EM algorithm was proposed in (Kontaxakis & Tzanakos 1993). The algorithm is stopped when a threshold value has been reached which corresponds to convergence to the maximum likelihood estimator. But, as we argued before, this does not seem to be a good idea since the problem we want to study is ill-posed and the maximum likelihood estimator will be very irregular and hence is far away from the exact solution.
Nonetheless, all these methods lack any theoretical foundation, in the sense, that there are no results available which guarantee statistical convergence of the properly stopped EM algorithm to the true intensity λ. To our knowledge the only work in this direction is due to (Resmerita, Engl, & Iusem 2007) in the context of a deterministic linear model. Their result, however, requires a supplementary condition on the boundedness of the iterations, which was not proven until now (see (Resmerita, Engl & Iusem 2008)).
The convergence analysis so far does not explain the self-regularization property of the EM algorithm. In our paper we take a first step in this direction and we explain this regularizing behavior in some specific cases. Moreover, we prove that for the right scaling this self- regularization improves asymptotically and we derive lower bounds for its total error.

2 Scaling properties of the EM algorithm

In this section we study the influence of the scaling of the kernel of the operator A on the total error of the EM algorithm. To this end we mention that property (8) of the EM algorithm is crucial for its convergence analysis. We will see that this property implies a rescaling of the kernel in order to be consistent. We consider the setting introduced in (Anderson, Mair & Rao 1996) (see also the Introduction). Under the normalizing condition (4) we can not obtain a consistent estimator by stopping this algorithm.

Lemma 1. *The total error $\|f_p - \lambda\|_\beta$ of the step p of the EM algorithm (7) is bounded from below by $2\exp(-1)$, for any $p \geq 1$ and $n \geq 1$.*

Proof. The following inequalities hold true:

$$\begin{aligned}
\|f_p-\lambda\|_\beta &= \int_\Gamma \|f_p(\gamma)-\lambda\|_{\mathrm{B}}\, d\mathbf{P}(\gamma)\\
&= \int_\Gamma \sup\left\{|f_p(\gamma)(C)-\lambda(C)| : C \text{ Borel set },C\subseteq\Omega\right\} d\mathbf{P}(\gamma)\\
&\geq \int_\Gamma |f_p(\gamma)(\Omega)-\lambda(\Omega)|\, d\mathbf{P}(\gamma)\\
&\overset{(7)}{=} \int_\Gamma |\sum_{i=1}^n Y_i(\gamma)-\int_\Omega d\lambda|\, d\mathbf{P}(\gamma)
\end{aligned}$$

$$\begin{aligned}
&= \mathbf{E}\left[|\sum_{i=1}^n Y_i-\int_\Omega d\lambda|\right]\\
&\overset{(4)}{=} \mathbf{E}\left[|\sum_{i=1}^n Y_i-1|\right]\\
&= 2\exp(-1),
\end{aligned}$$

since for any Poisson random variable Y with parameter θ it holds

$$\begin{aligned}
\mathbf{E}[|Y-\theta|] &= \sum_{k=0}^\infty |k-\theta|\exp(-\theta)\frac{\theta^k}{k\,!}\\
&= \exp(-\theta)\left\{\sum_{k=0}^{[\theta]}(\theta-k)\frac{\theta^k}{k\,!}+\sum_{k=[\theta]+1}^\infty (k-\theta)\frac{\theta^k}{k\,!}\right\}\\
&= \exp(-\theta)\left\{\sum_{k=0}^{[\theta]}\frac{\theta^{k+1}}{k\,!}-\sum_{k=0}^{[\theta]-1}\frac{\theta^{k+1}}{k\,!}+\sum_{k=[\theta]}^\infty\frac{\theta^{k+1}}{k\,!}-\sum_{k=[\theta]+1}^\infty\frac{\theta^{k+1}}{k\,!}\right\}\\
&= 2\exp(-\theta)\frac{\theta^{[\theta]+1}}{[\theta]\,!} = 2\theta\mathbf{P}\{Y=[\theta]\}.
\end{aligned}$$

Here $[\theta]$ is the integer part of θ. This suggests that the L^1 risk does not depend on n and it can not converge towards 0. □

Hence we notice that the normality property for the kernel i.e. $\sum_{i=1}^n a_i = 1$ implies that any stopping rule will not provide a consistent estimator. We remark the scaling properties of the EM algorithm: by multiplying the step f_k with a constant α, the next step f_{k+1} does not change, while if the data is scaled by a constant α, every iteration of the EM algorithm will be multiplied with α. In other iterative regularization methods like Landweber or iterated Tikhonov (see (Engl, Hanke, Neubauer 1996)) scaling the k-step or the data by a constant α implies scaling the $(k+1)$-iteration by α. Moreover, the scaling of the kernel $\sum_{i=1}^n a_i = \alpha$ implies that the sample size $\sum_{i=1}^n Y_i$ is Poisson distributed with parameter α. As we will see in the next lemma, a necessary condition for the lower bound to go to 0 is that α converges to infinity as n goes to infinity. Therefore we consider now the rescaled EM algorithm

$$f_{p+1}(s) = \frac{f_p(s)}{\alpha} \sum_{i=1}^{n} \frac{Y_i a_i(s)}{(T f_p(s))_i} \tag{9}$$

Lemma 2. *Let us assume that* $\sum_{i=1}^{n} a_i = \alpha$ *and consider the rescaled algorithm (9). Then* $\sum_{i=1}^{n} Y_i$ *is Poisson distributed with parameter* α *and the expected TV- risk has a lower bound equal to* $2\frac{\exp(-\alpha)}{[\alpha]!}\alpha^{[\alpha]}$.

Proof. The proof is similar to the proof of Lemma 1. We have

$$\begin{aligned}
\|f_p - \lambda\|_\beta &\geq \int_\Gamma |f_p(\gamma)(\Omega) - \lambda(\Omega)|\, d\mathbf{P}(\gamma) \\
&= \int_\Gamma |\frac{1}{\alpha}\sum_{i=1}^{n} Y_i(\gamma) - \int_\Omega d\lambda|\, d\mathbf{P}(\gamma) \\
&= \frac{1}{\alpha}\mathbf{E}\left[|\sum_{i=1}^{n} Y_i - \alpha\int_\Omega d\lambda|\right] \\
&= \frac{1}{\alpha}\mathbf{E}\left[|\sum_{i=1}^{n} Y_i - \alpha|\right] \\
&= 2\exp(-\alpha)\frac{\alpha^{[\alpha]}}{[\alpha]\,!}
\end{aligned} \tag{10}$$

□

Lemma 3. *Let us consider the rescaled algorithm with* $\alpha = \alpha_n$ *and* $\sum_{i=1}^{n} a_i = \alpha_n$. *If* $\alpha_n \to \infty$ *as* $n \to \infty$ *then*

$$\|f_p - \lambda\|_\beta \geq \frac{\sqrt{2\pi}}{e}\alpha_n^{-\frac{1}{2}} + \mathbf{o}(1). \tag{11}$$

Proof. We use the properties of Γ function to obtain a lower bound for the risk. It is well known that for integral, positive values $[\alpha_n]$ we have that $\Gamma([\alpha_n]) = ([\alpha_n]-1)\,!$. We abbreviate that $a_n \approx b_n$ if $\frac{a_n}{b_n} \overset{n\to\infty}{\to} 1$. We also have with Stirling's formula $\frac{\Gamma(\alpha_n)}{\alpha_n^{\alpha_n-\frac{1}{2}}\exp(-\alpha_n)} = \sqrt{2\pi} + \mathbf{o}(1)$. It follows that

$$\exp(-[\alpha_n])\,\frac{[\alpha_n]^{[\alpha_n]-\frac{1}{2}}}{([\alpha_n]-1)\,!} = \sqrt{2\pi} + \mathbf{o}(1).$$

Hence it holds

$$\exp(-[\alpha_n])\frac{[\alpha_n]^{[\alpha_n]}}{[\alpha_n]\,!} \approx \frac{\sqrt{2\pi}}{\sqrt{[\alpha_n]}}.$$

Since $1 \leq [\alpha_n] \leq \alpha_n < [\alpha_n]+1$ we have that $\exp(-[\alpha_n]-1) < \exp(-\alpha_n) \leq \exp(-[\alpha_n])$ and $[\alpha_n]^{[\alpha_n]} \leq \alpha_n^{[\alpha_n]}$. Putting these relations together we can write the

lower bound as

$$\frac{\sqrt{2\pi}}{e}\frac{1}{\sqrt{[\alpha_n]}} \approx \exp\left(-[\alpha_n]-1\right)\frac{[\alpha_n]^{[\alpha_n]}}{[\alpha_n]\,!} \leq \exp\left(-\alpha_n\right)\frac{\alpha_n^{[\alpha_n]}}{[\alpha_n]\,!}.$$

□

Remark 1. As a direct consequence of the Lemma 2 under the assumption (4) for the semi-continuous formulation (6) of the EM algorithm for $\lambda \in L^1(\Omega)$, the total error of the step k of the EM algorithm $\mathbf{E}\int_\Omega |f_k(s)-\lambda(s)|\,ds$ is bounded from below by $2\exp(-1)$, for any $k \geq 1$ and $n \geq 1$. Moreover, Lemma 2 and 3 also holds for the L^1- risk $\mathbf{E}\int_\Omega |f_k(s)-\lambda(s)|\,ds$.

From Stirling's formula it holds $\exp(-n)\frac{n^n}{n!} = \mathbf{o}\left(\frac{1}{\sqrt{n}}\right)$. Hence the right hand side of (10) $\exp(-\alpha)\frac{\alpha^{[\alpha]}}{[\alpha]\,!} \to 0$ for $\alpha = n$. This suggests us that instead of the data Y_i to use $\frac{1}{n}Y_i$ and to scale the kernel a_i proportionally, i.e. $\sum_{i=1}^n a_i = n$. Since $a_i(s) = a(t_{ni},s)$, this condition implies $\frac{1}{n}\sum_{i=1}^n a(t_{ni},s) \overset{n\to\infty}{\to} \int_\Sigma a(t,s)\,dt$. We will see that, while in the case of the normal scaling in (4) the numerical simulations do not suggest that the minimum of the L^1-risk goes to zero, for the $n-$scaled kernel and the EM algorithm

$$f_{k+1} = \frac{f_k}{n}\sum_{i=1}^n \frac{a_i Y_i}{\int_\Sigma a_i(s) f_k(s)\,ds}.$$

this minimum goes to zero and, from Stirling's formula, a lower bound for the L^1- risk is $\mathbf{o}\left(\frac{1}{\sqrt{n}}\right)$.

We illustrate in the following the scaling property of the EM algorithm by some numerical results. We fix the operator A as a convolution with a gaussian kernel restricted on the domain $[0,1]\times[0,1]$. First we choose the exact solution λ as the Heaviside function and we apply the EM algorithm scaled with 1, $\sqrt{n}$ and n, respectively. The number of simulations is always $N = 100$. We display in Figure 3 (a) the minimum L^1- risk as the sample size increases from 300 to 15,000. The lack of consistency becomes apparent in the first case and the improvement for a bigger scale. In Figure 3 (b) we consider a smooth function $\lambda(s) = 1+\cos(s)$ and the same scalings as before. The previous remark applies to this simulation, too. Moreover, the improvement in the rates for the minimum error of the $\sqrt{n}$, respectively n scaling is obvious. In Figure 4 we plotted the total error $\mathbf{E}\|f_k-\lambda\|_{L^1}$ for the first 100 iterations if the exact solution is the Heaviside function and the smooth function, respectively. For the $\sqrt{n}$ and n scaled algorithm we notice that the error decreases up to a minimum in the first iterations, then increases and tends to converge in the last iterations.

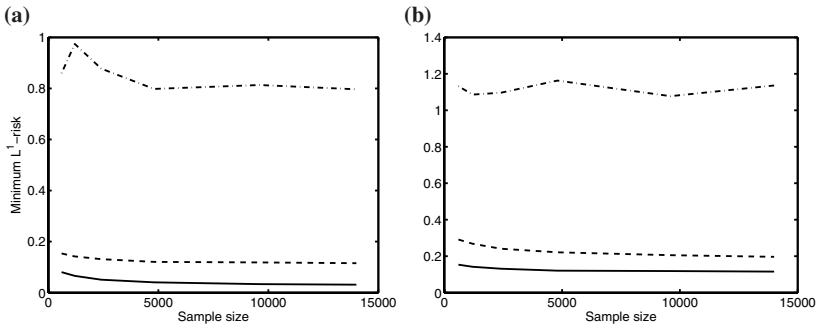


Fig. 3 Comparison of the minimum risk for 1 (dotted discontinuous line), $\sqrt{n}$ (discontinuous line) and n (continuous line) scaled EM algorithm with uniform initial guess for a smooth function as exact solution (a) and for a Heaviside function as exact solution (b)

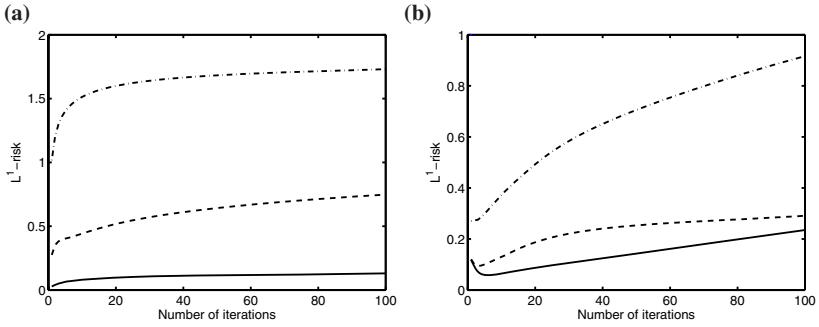


Fig. 4 Comparison of the L^1 risk for the first 100 iterations of the 1 (dotted discontinuous line), $\sqrt{n}$ (discontinuous line) and n (continuous line) scaled EM algorithm with uniform initial guess for smooth exact solution (a) and for Heaviside exact solution (c)

3 The effect of the initial guess

In this section we will give a heuristic explanation for the observation, that early stopping of the EM algorithm leads to the self- regularization of the true solution. For this purpose we consider a multiplicative error model with deterministic noise level $\delta > 0$. Our simplified data model will be

$$Y_i^\delta = \delta \int_\Omega a_i(s)\lambda(s)\,ds, \quad i = 1,\ldots,n,$$

where $\delta = 1$ corresponds to the noiseless case. We can write the iterations of the EM algorithm as

$$f_{k+1}^{\delta} = f_k^{\delta} A^T \left(\frac{\delta A\lambda}{A f_k^{\delta}} \right),$$

where

$$A^T : \mathbf{R}^n \to L^1(\Omega), A^T y = \sum_{i=1}^{n} a_i(s) y_i.$$

As it is custom in many applications (Shepp & Vardi 1982) we investigate the situation of a uniform initial solution. Let us assume

$$f_0^{\delta} = 1 \, a.e., , A f_0^{\delta} = c. \tag{12}$$

It follows

$$f_1^{\delta} = f_0^{\delta} \delta A^T \left(\frac{A\lambda}{A f_0^{\delta}} \right) = \frac{1}{c} \delta A^T A\lambda,$$

and

$$\begin{aligned} f_2^{\delta} &= f_0^{\delta} \delta A^T \left(\frac{A\lambda}{A f_0^{\delta}} \right) A^T \left(\frac{A\lambda}{A f_0^{\delta} A^T \left(\frac{A\lambda}{A f_0^{\delta}} \right)} \right) \\ &= \frac{\delta}{c} A^T A\lambda A^T \left(\frac{A\lambda}{A A^T \left(\frac{A\lambda}{c} \right)} \right) \\ &= \delta A^T A\lambda A^T \left(\frac{A\lambda}{A A^T A\lambda} \right). \end{aligned}$$

Finally, we get

$$f_3^{\delta} = \delta A^T A\lambda A^T \left(\frac{A\lambda}{A A^T A\lambda} \right) A^T \left(\frac{A\lambda}{A A^T A\lambda A^T \left(\frac{A\lambda}{A A^T A\lambda} \right)} \right).$$

Let us assume now that $c = 1$ to simplify notation further. Then we obtain

$$\begin{aligned} f_1^{\delta} &= f_0^{\delta} \delta A^T \frac{A\lambda}{A f_0^{\delta}} = \delta A^T A\lambda, \\ f_2^{\delta} &= \delta A^T A\lambda A^T \left(\frac{A\lambda}{A A^T A\lambda} \right), \\ f_3^{\delta} &= \delta A^T A\lambda A^T \left(\frac{A\lambda}{A A^T A\lambda} \right) A^T \left(\frac{A\lambda}{A A^T A\lambda A^T \left(\frac{A\lambda}{A A^T A\lambda} \right)} \right). \end{aligned}$$

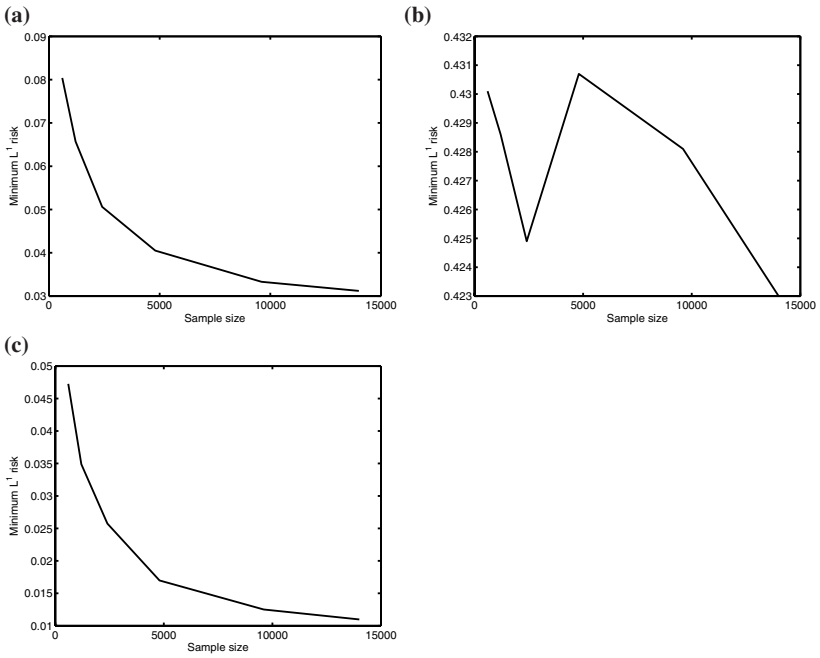


Fig. 5 Minimum L^1-risk of the n- scaled EM algorithm for smooth exact solution for uniform initial guess (a), for parabolic initial guess (b) and for exact solution as initial guess (c)

Let us assume in addition that the true solution λ is an eigen function of A^TA with corresponding eigenvalue θ (i.e. $A^TA\lambda = \theta\lambda$). Then

$$\begin{aligned} f_0^\delta &= 1, \\ f_1^\delta &= \delta\theta\lambda, \\ f_2^\delta &= f_3^\delta = \cdots = \delta\lambda. \end{aligned}$$

This provides an explanation for the self-regularization behavior of the EM algorithm: We obtain a good approximation of the solution already in the first step of the algorithm, and the improvement or the worsening of this approximation in the second iteration depends on the behavior of $\delta\theta$. As long as $|\delta\theta - 1| > |\delta - 1|$ the first step is improved by the second iteration for any δ and as $\delta \to 1$ this holds for any $\theta \neq 1$.
If the initial guess $f_0^\delta = \lambda$ then we have in the first step of the EM algorithm $f_1^\delta = \delta\lambda$ and hence

$$f_k^\delta = \delta\lambda, k \geq 1$$

under the assumptions (12).
In Figure 5 we compare the minimum L^1 risk for different initial guesses for the $n-$ scaled EM algorithm. Again, the same setting has been used as at the end of the section 2. As expected, starting with the exact solution presents the smallest total error. The uniform initial guess has also a small risk and the much bigger minimum L^1 error is obtained for a parabolic initial guess which is far away from the exact solution.

Acknowledgements M. Pricop and A. Munk acknowledge support of DFG SFB 755 and FOR 916. We are grateful to P. Marnitz, A. Esner, S. Mek and A. Schoenle for the 4π images used in the introduction.

References

Ahn, S., Fessler, J.A. (2003). Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms, *IEEE Transactions on Medical Imaging* **22**: 613-626.

Anderson, J. M. M., Mair, B. A. & Rao, M. (1996). Positron emission tomography: a review, *Control and Cybernetics* **25**: 1089–1111.

Antoniadis, A. & Bigot, J. (2006). Poisson inverse problems, *The Annals of Statistics* **34**: 2132–2158.

Bertero, M. & Boccacci, P. (1998). *Introduction to inverse problems in imaging,* Bristol.

Bertero, M., Boccacci, P., Desiderà G. & Vicidomini G. (2009). Image deblurring with Poisson data: from cells to galaxies, *Inverse Problems*, to appear.

Bissantz, N., Hohage, T., Munk, A. & Ruymgaart, F. (2007) Convergence rates of general regularization methods for statistical inverse problems and applications, *SIAM J. Numer. Anal.* **45**: 2610–2636.

Bissantz, N., Mair, B.A. & Munk, A. (2005). A multi-scale stopping criterion for MLEM reconstruction in PET, *IEEE Nuclear Science Symposium Conference Record* **6**: 3376–3379.

Bissantz, N., Mair, B.A. & Munk, A. (2008). A statistical stopping rule for MLEM reconstructions in PET, it IEEE Nucl. Sci. Symp. Conf. Rec. **8**: 4198-4200.

Bühlmann, P. & Yu, B. (2003). Boosting with the L_2 loss: regression and classification, *J. Amer. Statist. Assoc.* **98**: 324–339.

Cavalier, L. & Koo, J.-Y. (2002). Poisson intensity estimation for tomographic data using a wavelet shrinkage approach, *Transactions on Information Theory* **48**: 2794–2802.

Chang, I. S. & Hsiung, C. A. (1994). Asymptotic consistency of the maximum likelihood estimate in positron emission tomography and applications, *The Annals of Statistics* **22**: 1871–1883.

Coakley, K.J. (1991). A Cross-Validation Procedure for stopping the EM algorithm and deconvolution of neutron depth profiling spectra, *IEEE Transactions on Nuclear Science* **38**: 9–15.

Coakley, K.J. (1996). Bootstrap method for nonlinear filtering of EM-ML reconstructions of PET images, *International Journal of Imaging Systems and Technology* **7**: 54-61.

Coakley, K.J. & Llacer, J. (1991). The use of Cross- Validation as a stopping rule in emission computed tomography image reconstruction, *SPIE Proc. Image Phys: Med. Imaging V* **1443**: 226–233.

Cover, T. M. (1984). An algorithm for maximizing expected log investment return *Transactions on Information Theory* **30**: 369–373.

Crauel, Hans (2002). *Random probability measures on Polish spaces,* Taylor & Francis, London.

Csiszár, I. & Tusnády, G. (1984). Information geometry and alternating minimization procedures, *Statistics & Decisions* **1**: 205–237.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B* **39**: 1–38.

Eggermont, P. P. B. (1993). Maximum entropy regularization for Fredholm integral equations of the first kind, *SIAM Journal on Mathematical Analysis* **24**: 1557–1576.

Eggermont, P. P. B. (1999). Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind, *Applied Mathematics and Optimization* **39**: 75–91.

Eggermont, P. P. B. & LaRiccia, V. N. (1996). Maximum penalized likelihood estimation and smoothed EM algorithms for positive integral equations of the first kind, *Numerical Functional Analysis and Optimization* **17**: 737–754.

Eggermont, P. P. B. & LaRiccia, V. N. (2001). *Maximum penalized likelihood estimation. Vol. I,* Springer-Verlag, New York.

Engl, H. W. , Hanke, M. & Neubauer, A. (1996). *Regularization of inverse problems,* 375. Kluwer Academic Publishers Group, Dordrecht.

Fessler, J.A. (1994). Penalized weighted least-squares image-reconstruction for positron emission tomography, *IEEE Transactions on Medical Imaging* **13**: 290–300.

Hebert, T.J. (1990). Statistical stopping criteria for iterative maximum-likelihood reconstruction of emission images, *Physics in Medicine and Biology* **35**: 1221–1232.

Hebert, T., Leahy, R. & Singh, M. (1988). Fast MLE for SPECT using an intermediate polar representation and a stopping criterion, *IEEE Transactions on Nuclear Science* **35**: 615–619.

Hudson, M. & Larkin, R. (1994). Accelerated Image Reconstruction using Ordered Subsets of Projection Data, *IEEE Trans. Med. Imag* **13**: 601–609.

Iusem, Alfredo N. (1992). A short convergence proof of the EM algorithm for a specific Poisson model, *Rebrape* **6**: 57–67.

Johnson, V.E. (1994) A note on stopping rules in EM-ML reconstructions of ECT images, *IEEE Transactions on Medical Imaging* **13**, 569-571.

Johnstone, I. M. & Silverman, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems, *The Annals of Statistics* **18**: 251–280.

Koo, J.-Y. & Chung, H.-Y. (1998). Log-density estimation in linear inverse problems, *The Annals of Statistics* **26**: 335–362.

Kondor, A. (1983). Method of convergent weights — An iterative procedure for solving Fredholm's integral equations of the first kind, *Nuclear Instruments and Methods in Physics Research* **216**: 177-181.

Kontaxakis, G. & Tzanakos, G. (1993). Further study of a stopping rule for the EM algorithm, *Bioengineering Conference, 1993., Proceedings of the 1993 IEEE Nineteenth Annual Northeast Volume* , **18-19**: 52 - 53.

Korostelëv, A. P & Tsybakov, A. B. (1993). *Minimax theory of image reconstruction,* Springer-Verlag, New York.

Lang, M. & Müller, T. & Engelhard, J. & Hell, S. (2007). 4Pi microscopy of type A with 1-photon excitation in biological fluorescence imaging, *Opt. Express* **15**: 2459.

Latham, G. A.& Anderssen, R. S. (1994). On the stabilization inherent in the EMS algorithm, *Inverse Problems.* **10**, 161–183.

Llacer, J., Veklerov, E.& Coakley, K.J. (1993). Statistical analysis of maximum-likelihood estimator images of human brain FDG PET studies, *IEEE Transactions on Medical Imaging* **12**: 215-231.

Mair, B. A, Rao, M. & Anderson, J. M. M. (1996). Positron emission tomography, Borel measures and weak convergence, *Inverse Problems* **12**: 965–976.

Meinshausen, N., Rice, J. & Schücker, T. (2006). Testing for monotonicity in the Hubble diagram, *eprint arXiv:astro-ph/0612556* **12**.

Mülthei, H. N. & Schorr, B. (1989). On properties of the iterative maximum likelihood reconstruction method, *Mathematical Methods in the Applied Sciences* **11**: 331–342.

Natterer, F. (2001). *The mathematics of computerized tomography,* Philadelphia.

Resmerita, E., Engl, H. W. & Iusem, A. N. (2007). The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis, *Inverse Problems* **23**: 2575–2588.

Resmerita, E., Engl, H. W. & Iusem, A. N. (2008). Corrigendum: The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis, *Inverse Problems* **24**: 1p.

Shepp, L.A. & Vardi, Y. (1982). Maximum Likelihood Reconstruction for Emission Tomography, *IEEE Transactions on In Medical Imaging* **1**: 113–122.

Silverman, B. W., Jones, M. C., Nychka, D. W. & Wilson, J. D. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography, *Journal of the Royal Statistical Society. Series B.* **52**: 271–324.

Szkutnik, Z. (2000). Unfolding intensity function of a Poisson process in models with approximately specified folding operator, *Metrika* **52**: 1–26.

Vardi, Y. & Lee, D. (1993). From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems. With discussion. *J. Roy. Statist. Soc. Ser. B* **55**: 569–612.

Vardi, Y., Shepp, L. A. & Kaufman, L. (1985). A statistical model for positron emission tomography, *Journal of the American Statistical Association* **80**: 8–37.

Veklerov, E. & Llacer, J. (1987). Stopping rule for the MLE algorithm based on statistical hypothesis-testing, *IEEE Transactions on Medical Imaging* **6**: 313-319 .

Vicidomini, G., Hell, S. & Schönle, A. Automatic deconvolution of 4Pi-microscopy data with arbitrary phase, *Opt. Lett.*, Submitted.

Sequential Design of Computer Experiments for Constrained Optimization

Brian J. Williams, Thomas J. Santner, William I. Notz and Jeffrey S. Lehman

Abstract This paper proposes a sequential method of designing computer or physical experiments when the goal is to optimize one integrated signal function subject to constraints on the integral of a second response function. Such problems occur, for example, in industrial problems where the computed responses depend on two types of inputs: manufacturing variables and noise variables. In industrial settings, manufacturing variables are determined by the product designer; noise variables represent field conditions which are modeled by specifying a probability distribution for these variables. The update scheme of the proposed method selects the control portion of the next input site to maximize a posterior expected "improvement" and the environmental portion of this next input is selected to minimize the mean square prediction error of the objective function at the new control site. The method allows for dependence between the objective and constraint functions. The efficacy of the algorithm relative to the single-stage design and relative to a design assuming independent responses is illustrated. Implementation issues for the deterministic and measurement error cases are discussed as are some generalizations of the method.

Brian J. Williams
Los Alamos National Laboratory, P.O. Box 1663, MS F600, Los Alamos, NM 87545, e-mail: brianw@lanl.gov

Thomas J. Santner and William I. Notz
Department of Statistics, The Ohio State University, Cockins Hall, 1958 Neil Ave., Columbus, OH 43210, e-mail: tjs@stat.osu.edu, win@stat.osu.edu

Jeffrey S. Lehman
JPMorganChase, Home Finance Marketing Analytics, Columbus, OH 43240, e-mail: jeff_lehman@bankone.com

T. Kneib, G. Tutz (eds.), *Statistical Modelling and Regression Structures*,
DOI 10.1007/978-3-7908-2413-1_24,

1 Introduction

Computer models refer to settings in which a mathematical description of a physical system is implemented numerically via computer code so that system "responses" can be computed for any set of inputs. In a *computer experiment*, the inputs are manipulated to study their effect on the physical system that the computer code represents. For example, Bernardo et al. (1992) used computer-aided design simulators to model electrical current reference and voltage shifter circuits. Haylock & O'Hagan (1996) modeled the radiation dose received by body organs after ingesting radioactive iodine. Chang et al. (2001) modeled "proximal bone stress shielding" and "implant relative motion" for an in vivo hip prosthesis. Ong (2004) modeled the stability of acetabular hip components.

Computer experiments are attractive alternatives to physical experiments when the latter involve high-dimensional inputs, when running the corresponding physical experiment poses ethical issues, or when the conduct of the physical experiment would involve substantial time or other resources. Motivated by these concerns, numerous authors have developed statistical techniques both for prediction of the response at untried input sites based on a (small) training sample of computed responses and for selection of the inputs at which to compute the training sample, i.e., the analysis and design of computer experiments. Both frequentist (Sacks et al. 1989, Welch et al. 1992) and Bayesian (Currin et al. 1991, O'Hagan 1992) principles have been used to suggest methodology to predict responses at untried input sites for computer experiments (see also Santner et al. 2003). This paper follows the Bayesian approach; it regards the unknown function calculated by the computer code to be a realization of a random function whose properties embody the prior information about the code output.

We propose a sequential design for computer experiments when the goal is to optimize the mean of one computer code (the *objective* function) under constraints defined by the mean of a second computer code (the *constraint* function). To motivate this formulation, consider Chang et al. (2001), who study the design of a hip prosthesis. Their computer code calculated a pair of *competing* responses which depend on both *manufacturing* (control) input variables and *environmental* input variables. The computer outputs were the "proximal bone stress shielding" and "implant relative motion." Biomechanically, long-term shielding of the bone from stress can cause bone resorption and thus weaken the hip; conversely, too much motion of the implant within the femur can cause the implant to loosen. In this application, the control input variables describe the geometry of the implant while the environmental input variables account for variability in patient bone properties and the level of patient activity. The distribution of the environmental variables in human populations was inferred from studies in the orthopedic literature. Chang et al. (2001) determined the combination of geometry variables that minimized *mean* stress shielding subject to an upper limit on the mean implant relative motion where both responses were averaged over a discrete probability distribution for the environmental variables.

While the mathematical programming literature contains numerous algorithms to solve constrained and unconstrained optimization problems, essentially all such tech-

niques require too many function evaluations to be *directly* useful in many computer experiments because of the substantial length of time the codes require to calculate the response(s). As a result, the computer experiment literature has coupled the use of statistical predictors in place of exact computer code evaluations with traditional optimization algorithms. For example, Bernardo et al. (1992) implemented an algorithm for response minimization that sequentially focuses on the region of the input variable space where the optimum appears to be located. Jones et al. (1998) and Schonlau et al. (1998) introduced a criterion-based sequential strategy for response minimization. Williams et al. (2000b) extended the methodology of Schonlau et al. (1998) to situations in which there are both control and environmental input variables.

Schonlau et al. (1998) proposed extending their single-response expected improvement algorithm to accommodate multiple signals for constrained optimization problems under the assumptions that the objective and constraint signals be modeled as mutually independent *and* both were computable at any input. Their methods are not useable in the hip replacement problem because neither of these assumptions holds in this application. First, modeling the computer outputs as mutually independent is unreasonable because the computer outputs represent competing design objectives and thus the outputs are negatively correlated with large values for one tending to be associated with small values for the other. Second, neither the objective function nor the constraint function in the hip design application were directly computable because each was a mean over the environmental variables; in this case, because the enviromental distribution consisted of twelve points, *twelve runs* of the code were required to calculate a *single* value of either the objective or constraint functions at any desired control variable. With each run of the finite element code requiring five to ten hours of workstation time, a single value of either the objective or constraint function would require roughly five days to calculate.

The Bayesian model of Section 2 that is used in this paper allows for both computer experiments and physical (spatial) applications that contain measurement error. The details of the proposed *biVariate constraIned exPectEd impRovement* algorithm (called VIPER hereafter) are presented for this general model in Section 3. An example comparing the performance of this algorithm to a single-stage design is given in Section 4 and to a sequential algorithm that assumes the constraint and objective signals can be modeled as independent processes. Section 5 contains a discussion of several important issues regarding implementation and extensions of the algorithm.

2 Modeling

We base the development in Section 3 on the following Bayesian model. Our interest is in expectations of the "smooth" signals $z_1(\mathbf{x})$ and $z_2(\mathbf{x})$ defined on a common compact domain $\mathscr{X} \subset \mathbb{R}^p$. In computer experiments, each $z_i(\mathbf{x})$ is observed exactly while in spatial applications a corrupted version of each $z_i(\mathbf{x})$ is observed. Denote the *observed* responses by $y_1(\mathbf{x})$ and $y_2(\mathbf{x})$.

For $i = 1, 2$, the prior information about the signal and the observed responses are specified by regarding them as draws from the random functions

$$Z_i(\mathbf{x}) = \boldsymbol{\beta}_i^\top \mathbf{f}_i(\mathbf{x}) + W_i(\mathbf{x}) \quad \text{and} \quad Y_i(\mathbf{x}) = Z_i(\mathbf{x}) + \varepsilon_i(\mathbf{x}) \tag{1}$$

For $i = 1, 2$, in physical experiments, $\varepsilon_i(\cdot)$ is a zero-mean Gaussian white noise process with unknown variance $\sigma_i^2 > 0$ that represents measurement error; in computer experiments, $\sigma_i^2 \equiv 0$ so that $y_i(\mathbf{x}) = z_i(\mathbf{x})$. We refer to $Z_i(\cdot)$ as the *signal* process.

All signal processes are statistically independent of all noise processes and the two noise processes are independent. The term $\boldsymbol{\beta}_i^\top \mathbf{f}_i(\mathbf{x})$ of the signal process is the (nonstationary) mean of the $Z_i(\cdot)$ and $Y_i(\cdot)$ processes while the "residual" $W_i(\cdot)$ is a stationary Gaussian stochastic process having mean 0, correlation function $R_i(\cdot)$, and unknown process variance $\tau_i^2 > 0$. The vector of regression coefficients $\mathbf{f}_i(\cdot)$ is a $k_i \times 1$ set of known regression functions and $\boldsymbol{\beta}_i \in \mathbb{R}^{k_i}$ is a vector of unknown regression parameters. We let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ denote the vector of all $k = k_1 + k_2$ regression parameters. The model (1) is completed upon specification of a positive definite joint covariance structure for the signal processes. Stationarity implies that the covariance (correlation) between $Z_i(\mathbf{x}_1)$ and $Z_i(\mathbf{x}_2)$ depends only on the difference $\mathbf{x}_1 - \mathbf{x}_2$. In the general discussion of Section 3, we allow the arbitrary form for $\mathrm{Cov}[Z_1(\mathbf{x}_1), Z_2(\mathbf{x}_2)] = \tau_1 \tau_2 R_{12}(\mathbf{x}_1 - \mathbf{x}_2)$ (subject to positive definiteness of the full bivariate covariance matrix at any finite set of inputs).

Most correlation functions commonly used in practice are members of some parametric family. We assume that the correlation function $R_i(\cdot)$ depends on an unknown parameter vector $\boldsymbol{\xi}_i$, for $i = 1, 2$, and that the cross-correlation function $R_{12}(\cdot)$ depends on the unknown parameter vector $\boldsymbol{\xi}_{12}$. The joint correlation parameter vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_{12})$ is allowed to take any value for which the covariance structure of the joint process $\{\mathbf{Z}(\mathbf{x}) = (Z_1(\mathbf{x}), Z_2(\mathbf{x})), \mathbf{x} \in \mathscr{X}\}$ is positive definite. The means $\boldsymbol{\beta}_i^\top \mathbf{f}_i(\mathbf{x})$ and the correlation parameter $\boldsymbol{\xi}$ determine the permissible sample paths of this joint process. In Section 4, we introduce a specific spatial autoregressive model for the signal processes where we use the product power exponential correlation function. We complete specification of the prior by assuming the noninformative prior distribution,

$$[\boldsymbol{\beta}, \tau_1^2] \propto \frac{1}{\tau_1^2}$$

for the final stage of this hierarchical model.

In the formulas that follow, it is notationally convenient to adopt the following reparameterization of the model in terms of the variance τ_1^2 of $Z_1(\cdot)$ and the variance ratios: $\eta = \tau_2^2/\tau_1^2$, $\rho_1 = \sigma_1^2/\tau_1^2$, and $\rho_2 = \sigma_2^2/\tau_2^2 = \sigma_2^2/(\eta\tau_1^2)$. The quantity η is the ratio of the signal process variances and the ρ_i are the inverse of signal-to-noise ratios.

In this article, all posterior distributions assume that we are *given* the unknown vector of parameters $\boldsymbol{\gamma} = (\eta, \rho_1, \rho_2, \boldsymbol{\xi})$. It is possible to carry out a fully Bayesian analysis by integrating $\boldsymbol{\gamma}$ out of these posterior distributions. However, due to the

substantial additional computational complexity of this approach, we adopt the simpler strategy of setting $\boldsymbol{\gamma}$ equal to its posterior mode and proceed by substituting this mode for $\boldsymbol{\gamma}$ wherever required.

Let $\mathbf{x}_c$ and $\mathbf{x}_e$ represent the control and environmental variable vectors, and denote their corresponding domains by $\mathscr{X}_c$ and $\mathscr{X}_e$. We assume that the environmental variables have a joint probability distribution with finite support $\{\mathbf{x}_{e,j}\}_{j=1}^{n_e}$ and associated probabilities (weights) $\{w_j\}_{j=1}^{n_e}$; in practice, this assumption is sufficiently flexible that it can serve as an adequate approximation for many continuous environmental variable distributions. The functions $\mu_i(\cdot)$ are the mean of the signal processes over the environmental variables given by

$$\mu_i(\mathbf{x}_c) = \sum_{j=1}^{n_e} w_j \, z_i(\mathbf{x}_c, \mathbf{x}_{e,j}) \, .$$

Our goal is to identify the control variable settings $\mathbf{x}_c^*$ that minimize $\mu_1(\cdot)$ subject to a constraint on $\mu_2(\cdot)$, i.e.,

$$\mathbf{x}_c^* = \underset{\mathbf{x}_c \in \mathscr{X}_c}{\operatorname{argmin}} \; \mu_1(\mathbf{x}_c) \quad \text{subject to} \quad \mu_2(\mathbf{x}_c^*) \leq U \, . \tag{2}$$

A straightforward modification of the algorithm presented in this paper can be applied to the situation where the constraint function is bounded from below or within an interval. Prior uncertainty in $\mu_i(\cdot)$ is induced directly from the Z_i processes, i.e., the prior of $\mu_i(\cdot)$ is specified by the distribution of $M_i(\mathbf{x}_c) = \sum_{j=1}^{n_e} w_j \, Z_i(\mathbf{x}_c, \mathbf{x}_{e,j})$.

3 A Minimization Algorithm

The algorithm described in this section uses statistical predictors of the objective and constraint functions in conjunction with traditional optimization algorithms to solve (2). The formulas required to implement the algorithm are stated for a general bivariate $(Z_1(\mathbf{x}), Z_2(\mathbf{x}))$ process and, in Section 4, are specialized to the spatial autoregressive model of Kennedy & O'Hagan (2000). In brief, the algorithm

1. Computes both responses for the set of points in an initial (space-filling) design.
2. Uses the information from these runs to select the next point according to a bivariate expected improvement criterion.
3. Continues selecting points using the necessary information from all of the previous runs until a stopping criterion is met.

3.1 The VIPER Algorithm

The first stage of the VIPER algorithm observes $y_1(\mathbf{x})$ and $y_2(\mathbf{x})$ at every input site for an initial design $S_n = \{\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t\}$ where the notation $\mathbf{x}_j^t = (\mathbf{x}_{c,j}^t, \mathbf{x}_{e,j}^t)$ is used to emphasize that the input sites are *training data* at which the $y_1(\cdot)$ and $y_2(\cdot)$ are to be evaluated. An attractive choice of S_n is a space-filling design (see Chapter 5 of Santner et al. (2003) for several methods of generating space-filling designs). Let $\mathbf{Y}_i^n$ denote the vector of random responses associated with S_n for $i \in \{1,2\}$. Finally, let $S_n^c = \{\mathbf{x}_{c,1}^t, \ldots, \mathbf{x}_{c,n}^t\}$. denote the control variable portions of S_n.

For any potential new control variable site $\mathbf{x}_c$, define the *improvement* at $\mathbf{x}_c$ to be

$$i_n(\mathbf{x}_c) = \max\{0, \mu_1^{\min} - \mu_1(\mathbf{x}_c)\} \times \chi[\mu_2(\mathbf{x}_c) \le U],$$

where $\mu_1^{\min} = \min\{\mu_1(\mathbf{x}_{c,i}^t) : \mu_2(\mathbf{x}_{c,i}^t) \le U\}$ is the minimum of $\mu_1(\cdot)$ over feasible points in S_n^c and $\chi[A]$ is 1 if A occurs and is 0 otherwise. Thus $i_n(\mathbf{x}_c)$ measures the amount of improvement in the value of the objective function $\mu_1(\cdot)$ at the candidate site $\mathbf{x}_c$ compared with the minimum of $\mu_1(\cdot)$ over the current feasible training data points subject to $\mathbf{x}_c$ satisfying the $\mu_2(\cdot)$ constraint. We take the distribution of

$$I_n(\mathbf{x}_c, \boldsymbol{\gamma}) = \max\{0, M_1^{\min} - M_1(\mathbf{x}_c)\} \times \chi[M_2(\mathbf{x}_c) \le U]$$

as a prior for $i_n(\cdot)$ where $M_1^{\min}$ is defined to be

$$\min\{M_1(\mathbf{x}_{c,i}^t) : \mathrm{E}\{M_2(\mathbf{x}_{c,i}^t) \,|\, \mathbf{Y}_1^n, \mathbf{Y}_2^n, \boldsymbol{\gamma}\} - t_{2n-k,.95}\sqrt{\mathrm{Var}\{M_2(\mathbf{x}_{c,i}^t) \,|\, \mathbf{Y}_1^n, \mathbf{Y}_2^n, \boldsymbol{\gamma}\}} \le U\}$$

and $t_{\nu,.95}$ is the upper 95th percentile of the t-distribution with ν degrees of freedom and $k = k_1 + k_2$ is the total number of unknown parameters that describe the means of $Z_1(\cdot)$ and $Z_2(\cdot)$. The quantity $M_1^{\min}$ is an intuitive estimate of $\mu_1^{\min}$ because it is the minimum of $M_1(\cdot)$ at the control sites that *appear* to be in the $\mu_2(\mathbf{x}_c)$-based feasible region, as judged by a point-wise (posterior) 95% confidence band for $\mu_2(\mathbf{x}_c)$.

With the above notation, we present a more detailed description of the proposed algorithm and then provide implementation specifics in Section 3.2. Additional implementation details can be found in Williams et al. (2000a).

S0: Choose the initial set of design points $S_n = \{\mathbf{x}_1^t, \ldots, \mathbf{x}_n^t\}$.

S1: Estimate the covariance parameter vector $\boldsymbol{\gamma}$ by the mode of the posterior density of $\boldsymbol{\gamma}$ given $(\mathbf{Y}_1^n, \mathbf{Y}_2^n)$ (given by (6), below). Let $\widehat{\boldsymbol{\gamma}}_n$ denote this posterior mode.

S2: Select the $(n+1)$-st *control* variable site, $\mathbf{x}_{c,n+1}^t$, to maximize the *posterior expected improvement* given the current data, i.e.,

$$\mathbf{x}_{c,n+1}^t = \underset{\mathbf{x}_c \in \mathscr{X}_c}{\mathrm{argmax}}\ \mathrm{E}\{I_n(\mathbf{x}_c, \widehat{\boldsymbol{\gamma}}_n) \,|\, \mathbf{Y}_1^n, \mathbf{Y}_2^n, \widehat{\boldsymbol{\gamma}}_n\}, \tag{3}$$

where $\mathrm{E}\{\cdot \,|\, \mathbf{Y}_1^n, \mathbf{Y}_2^n, \boldsymbol{\gamma}\}$ denotes the posterior conditional mean given the observed data $(\mathbf{Y}_1^n, \mathbf{Y}_2^n)$ and the parameter vector $\boldsymbol{\gamma}$.

S3: Choose the *environmental* variable site $\mathbf{x}_{e,n+1}^t$ corresponding to $\mathbf{x}_{c,n+1}^t$ to minimize the *posterior mean square prediction error* given the current data, i.e.

$$\mathbf{x}_{e,n+1}^t = \underset{\mathbf{x}_e \in \mathscr{X}_e}{\operatorname{argmin}} \, \mathrm{E}\left\{ \left[\widehat{M}_1^{n+1}(\mathbf{x}_{c,n+1}^t) - M_1(\mathbf{x}_{c,n+1}^t)\right]^2 \,\middle|\, \mathbf{Y}_1^n, \mathbf{Y}_2^n, \widehat{\boldsymbol{\gamma}}_n \right\}, \quad (4)$$

where $\widehat{M}_1^{n+1}(\cdot)$ is the posterior mean of $M_1(\cdot)$, based on the data from the n-point design S_n *and* the $Y_1(\cdot)$ and $Y_2(\cdot)$ signals at the location $(\mathbf{x}_{c,n+1}^t, \mathbf{x}_e)$.

S4: If the stopping criterion is not met, set $S_{n+1} = S_n \cup \{(\mathbf{x}_{c,n+1}^t, \mathbf{x}_{e,n+1}^t)\}$, calculate $y_1(\mathbf{x}_{c,n+1}^t, \mathbf{x}_{e,n+1}^t)$ and $y_2(\mathbf{x}_{c,n+1}^t, \mathbf{x}_{e,n+1}^t)$, increment n to $(n+1)$, and go to *S1*. If the criterion is met, the global minimizer is estimated to be the minimizer of the empirical Best Linear Unbiased Predictor (EBLUP) of $M_1(\cdot)$ subject to the EBLUP of $M_2(\cdot)$ satisfying the upper bound constraint. Specific stopping criteria are discussed in the examples of Section 4.

The parameter estimation in Step *S1* can be very time consuming. The objective functions to be optimized in (3) and (4) can have numerous local optima. These optimizations are carried out using the simplex algorithm of Nelder & Mead (1965). Our code makes repeated attempts to find an optimal solution to avoid getting trapped in the local optima. A quasi-Newton algorithm is applied to the candidate solutions of the simplex algorithm. One point of each starting simplex in (3) is obtained by searching a Latin Hypercube design (LHD) for good candidates; the remaining points are determined randomly. Each starting simplex in (4) is determined randomly. In Section 4, we examine a modification to this algorithm where parameter estimates from *S1* are used to add groups of several points sequentially according to *S2–S4*.

3.2 Implementation Details

In the following, we let $\mathscr{T}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denote the q-variate t-distribution with location–shift $\boldsymbol{\mu}$, scale-matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom, i.e., the distribution with joint density function

$$f(\mathbf{w}) = \frac{\Gamma([\nu+q]/2)}{|\boldsymbol{\Sigma}|^{1/2}(\nu\pi)^{q/2}\Gamma(\nu/2)} \left(1 + \frac{1}{\nu}(\mathbf{w}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w}-\mu)\right)^{-(\nu+q)/2}, \quad \mathbf{w} \in \mathbb{R}^q. \quad (5)$$

3.2.1 Step S0: Initial Design

While not the only possible space-filling designs, we have found that maximin distance LHDs or, perhaps better, maximin distance orthogonal array LHDs (Tang 1993) work well.

3.2.2 Step S1: Maximizing the Posterior of $\boldsymbol{\gamma}$ given $\mathbf{Y}_1^n$ and $\mathbf{Y}_2^n$

The probability density function of the posterior distribution of $\boldsymbol{\gamma}$ given $\mathbf{Y}_1^n$ and $\mathbf{Y}_2^n$, is

$$p(\boldsymbol{\gamma}\,|\,\mathbf{Y}_1^n,\mathbf{Y}_2^n) \propto p(\boldsymbol{\gamma})\,|\mathbf{R}|^{-1/2}\,|\mathbf{F}^\top\mathbf{R}^{-1}\mathbf{F}|^{-1/2}\,[\widehat{\tau}_1^2]^{-(n-k/2)}\,, \tag{6}$$

where $p(\boldsymbol{\gamma})$ is the prior distribution of $\boldsymbol{\gamma}$ (see Handcock & Stein 1993), $\mathbf{F}$ and $\mathbf{R}$ are the regression and correlation matrices of the vector $(\mathbf{Y}_1^{n\top},\mathbf{Y}_2^{n\top})^\top$, and $\widehat{\tau}_1^2$ is the posterior estimate of τ_1^2 given $(\mathbf{Y}_1^n,\mathbf{Y}_2^n)$ and $\boldsymbol{\gamma}$.

3.2.3 Step S2: Selection of Control Variables

We obtain an expression for the posterior expected improvement (3). Let $\mathbf{Z}_c$ denote the column vector $(\mathbf{M}_1^{n\top},\,\mathbf{Y}_1^{n\top},\,\mathbf{Y}_2^{n\top})^\top$ of length $3n$ where $\mathbf{M}_1^n$ is the vector of $M_1(\cdot)$ values evaluated at S_n^c. Given $\mathbf{M}_1^n$, note that $I_n(\mathbf{x}_c)$ is a function only of $(M_1(\mathbf{x}_c),M_2(\mathbf{x}_c))$. Hence, we evaluate the posterior expected improvement using

$$\mathrm{E}\{\,I_n(\mathbf{x}_c)\,|\,\mathbf{Y}_1^n,\mathbf{Y}_2^n,\boldsymbol{\gamma}\} = \mathrm{E}_{\mathbf{M}_1^n\,|\,\mathbf{Y}_1^n,\mathbf{Y}_2^n,\boldsymbol{\gamma}}\left\{\,\mathrm{E}\{\,I_n(\mathbf{x}_c)\,|\,\mathbf{Z}_c,\boldsymbol{\gamma}\}\,\right\}. \tag{7}$$

To evaluate the inner expectation we require the posterior distribution of $(M_1(\mathbf{x}_c),$ $M_2(\mathbf{x}_c))$, given $\mathbf{Z}_c$ and $\boldsymbol{\gamma}$, which is a shifted, bivariate t:

$$[\,M_1(\mathbf{x}_c)\,,M_2(\mathbf{x}_c)\,|\,\mathbf{Z}_c,\boldsymbol{\gamma}] \;\sim\; \mathscr{T}_2\,(\,\mathbf{m}_c\,,\widehat{\tau_{1,c}^2}\,\mathbf{R}_c\,,3n-k\,), \tag{8}$$

where the posterior mean $\mathbf{m}_c$ is the vector of BLUPs of $M_1(\mathbf{x}_c)$ and $M_2(\mathbf{x}_c)$ given $\mathbf{Z}_c$, and $\widehat{\tau_{1,c}^2}\,\mathbf{R}_c$ is proportional to the posterior covariance matrix. Using (8), a formula for the inner conditional expectation (7) can be written in terms of the quantities: $\widehat{r} \equiv \mathbf{R}_{c,12}/\sqrt{\mathbf{R}_{c,11}\mathbf{R}_{c,22}}$, $U_1 \equiv (M_1^{\min}-\mathbf{m}_{c,1})/\sqrt{\widehat{\tau_{1,c}^2}\mathbf{R}_{c,11}}$, $U_2 \equiv (U-\mathbf{m}_{c,2})/\sqrt{\widehat{\tau_{1,c}^2}\mathbf{R}_{c,22}}$, $\zeta_{\widehat{r}}^2(z) = (1-\widehat{r}^2)(z^2+3n-k)/(3n-1-k)$, and $C(z) \equiv \sqrt{3n-k}$ $t_{3n-2-k}\left(z\sqrt{3n-2-k}/\sqrt{3n-k}\right)\,/\,\sqrt{3n-2-k}$ as

$$\begin{aligned}\mathrm{E}\{\,I_n(\mathbf{x}_c)\,|\,\mathbf{Z}_c,\boldsymbol{\gamma}\} = \sqrt{\widehat{\tau_{1,c}^2}\mathbf{R}_{c,11}}\times\Bigg[&U_1\,T_2\,(\,U_1\,,U_2\,,\mathbf{0}_2,\widehat{r},3n-k\,)\,+\\ &C(U_1)\,T_{3n-1-k}\left(\frac{U_2-\widehat{r}U_1}{\zeta_{\widehat{r}}(U_1)}\right)+\widehat{r}\,C(U_2)\,T_{3n-1-k}\left(\frac{U_1-\widehat{r}U_2}{\zeta_{\widehat{r}}(U_2)}\right)\Bigg],\end{aligned} \tag{9}$$

where $t_\nu(\cdot)$ ($T_\nu(\cdot)$) denotes the *probability density function* (*cumulative distribution function*) of the standard univariate t distribution with ν degrees of freedom, and $T_2\,(\,\cdot\,,\widehat{r},\nu\,)$ denotes the bivariate t *cumulative distribution function* with ν degrees of freedom, location-shift $(0,0)^\top$, and scale-matrix $\begin{pmatrix}1 & \widehat{r}\\ \widehat{r} & 1\end{pmatrix}$. Note that $\widehat{r}$ is the posterior correlation between $M_1(\mathbf{x}_c)$ and $M_2(\mathbf{x}_c)$ given $\mathbf{Z}_c$ and $\boldsymbol{\gamma}$.

The conditional posterior expected improvement in (9) has the following interpretation. It forces $\mathbf{x}_{c,n+1}^t$ to be chosen roughly in an area of the control variable space where $M_1(\cdot)$ is predicted to be small *or* there is high uncertainty in the prediction of $M_1(\cdot)$, and the constraint is satisfied with high probability. When $M_1(\cdot)$ and $M_2(\cdot)$ are positively correlated, high uncertainty in the prediction of $M_2(\cdot)$ can also contribute to favorable consideration of a candidate point.

We use Monte Carlo simulation to compute the (outer) unconditional posterior expected improvement in (7) by integrating (9). A random sample of size N_c is obtained from the posterior distribution of $\mathbf{M}_1^n$ given $\mathbf{Y}_1^n$, $\mathbf{Y}_2^n$ and $\boldsymbol{\gamma}$. For each sample, the minimum loss $M_1^{\min}$ is obtained and the expectation in (9) is computed. We estimate the posterior expected improvement as the average of these quantities over the N_c draws. One feature of using (7) is that the *same* Monte Carlo sample can be used to estimate the posterior expected improvement at *all* control sites $\mathbf{x}_c$ (this method will provide dependent estimates of the improvement across different control sites). This follows from the fact that the posterior distribution of $\mathbf{M}_1^n$ given $\mathbf{Y}_1^n$, $\mathbf{Y}_2^n$ and $\boldsymbol{\gamma}$ does not depend on $\mathbf{x}_c$.

3.2.4 Step S3: Selection of Environmental Variables

In words, we select the environmental portion of the input corresponding to $\mathbf{x}_{c,n+1}^t$ to minimize the prediction error of the predicted mean at $\mathbf{x}_{c,n+1}^t$. We evaluate the expectation in (4). Let $J_n(\mathbf{x}_e) = [\widehat{M}_1^{n+1}(\mathbf{x}_{c,n+1}^t) - M_1(\mathbf{x}_{c,n+1}^t)]^2$ be the squared prediction error at $\mathbf{x}_{c,n+1}^t$ as a function of $\mathbf{x}_e$. Set

$$\mathbf{Z}_e = (Z_1(\mathbf{x}_{c,n+1}^t, \mathbf{x}_e), Z_2(\mathbf{x}_{c,n+1}^t, \mathbf{x}_e), \mathbf{Y}_1^{n\top}, \mathbf{Y}_2^{n\top})^\top .$$

Recall that $\widehat{M}_1^{n+1}(\mathbf{x}_{c,n+1}^t)$ is the posterior mean of $M_1(\mathbf{x}_{c,n+1}^t)$ given $\mathbf{Z}_e$ and $\boldsymbol{\gamma}$. Hence $\mathrm{E}[J_n(\mathbf{x}_e) \,|\, \mathbf{Y}_1^n, \mathbf{Y}_2^n, \boldsymbol{\gamma}]$ can be evaluated as

$$\mathrm{E}_{Z_1(\mathbf{x}_{c,n+1}^t, \mathbf{x}_e), Z_2(\mathbf{x}_{c,n+1}^t, \mathbf{x}_e) \,|\, \mathbf{Y}_1^n, \mathbf{Y}_2^n, \boldsymbol{\gamma}} \{ \mathrm{E}[J_n(\mathbf{x}_e) \,|\, \mathbf{Z}_e, \boldsymbol{\gamma}] \} . \tag{10}$$

An analytic expression for the inner expectation can be obtained from the fact that the posterior distribution of $M_1(\mathbf{x}_{c,n+1}^t)$ given $\mathbf{Z}_e$ and $\boldsymbol{\gamma}$ is a scaled and shifted univariate t:

$$[M_1(\mathbf{x}_{c,n+1}^t) \,|\, \mathbf{Z}_e, \boldsymbol{\gamma}] \sim \mathscr{T}_1(m_1^e, \widehat{\tau_{1,e}^2} R_e, 2n+2-k) . \tag{11}$$

The posterior mean m_1^e is the BLUP of $M_1(\mathbf{x}_{c,n+1}^t)$ based on responses from the design S_n and the signals evaluated at $(\mathbf{x}_{c,n+1}^t, \mathbf{x}_e)$, and $\widehat{\tau_{1,e}^2} R_e$ is proportional to the posterior variance. Then,

$$\mathrm{E}\{J_n(\mathbf{x}_e) \,|\, \mathbf{Z}_e, \boldsymbol{\gamma}\} = \frac{2n+2-k}{2n-k} \widehat{\tau_{1,e}^2} R_e , \tag{12}$$

which is the variance of the posterior distribution in (11).

Finally, a closed form expression for the posterior mean square prediction error in (10) at $\mathbf{x}_e$ is obtained by computing the expectation of (12) with respect to the conditional distribution of $(Z_1(\mathbf{x}^t_{c,n+1},\mathbf{x}_e),Z_2(\mathbf{x}^t_{c,n+1},\mathbf{x}_e))$ given $\mathbf{Y}_1^n$, $\mathbf{Y}_2^n$ and $\boldsymbol{\gamma}$. The outer expectation in (10), the integral of (12), is

$$\mathrm{E}\left\{J_n(\mathbf{x}_e)\,|\,\mathbf{Y}_1^n,\mathbf{Y}_2^n,\boldsymbol{\gamma}\right\} = \frac{1}{2n-k}\left[\mathbf{M}_e^\top \mathbf{Q}_e \mathbf{M}_e + \frac{2n-k}{n-1-k/2}\,\widehat{\tau}_1^2\right] R_e\,,$$

where $\mathbf{M}_e^\top = (\mathbf{m}^\top, \mathbf{Y}_1^{n\top}, \mathbf{Y}_2^{n\top})$, $\mathbf{m}$ contains the posterior means of $Z_1(\mathbf{x}^t_{c,n+1},\mathbf{x}_e)$ and $Z_2(\mathbf{x}^t_{c,n+1},\mathbf{x}_e)$, given $\mathbf{Y}_1^n$, $\mathbf{Y}_2^n$ and $\boldsymbol{\gamma}$, and $\mathbf{Q}_e$ is the matrix of the quadratic form defining $\widehat{\tau^2_{1,e}}$.

4 An Autoregressive Model and Example

Below we present recommendations concerning the implementation choices for VIPER based on an autoregressive bivariate model; these recommendations are based on a variety of examples, of which that in Section 4.2 is typical. We also discuss the fundamental problem of quantifying the benefit of using the sequential algorithm compared with a one-stage design that takes the same total number of observations as the sequential algorithm in a space-filling manner and then uses the same constrained optimizer used by VIPER in its final stage.

4.1 A Bivariate Gaussian Stochastic Process Model

The following example illustrates the operation of the VIPER algorithm in the computer expements set-up (no measurement error). In this example, we specialize the general formulas of Section 3 to a variant of the nonisotropic spatial autoregressive model of Kennedy & O'Hagan (2000). For $\mathbf{x} \in \mathscr{X}$, the $Z_i(\mathbf{x})$ model of (1) is taken to have constant mean β_i, $i = 1,2$, with the $\{W_i(\cdot)\}_i$ processes built from independent covariance stationary Gaussian processes. $W_1(\cdot)$ is a mean zero, stationary Gaussian stochastic process with process variance τ_1^2, and correlation function $R_1(\cdot)$; $W_2(\cdot)$ is defined by

$$W_2(\mathbf{x}) = rW_1(\mathbf{x}) + W_\delta(\mathbf{x})\,, \tag{13}$$

where $W_\delta(\cdot)$ is a mean zero stationary Gaussian stochastic process independent of $W_1(\cdot)$ with process variance τ_δ^2, and covariance function $R_\delta(\cdot)$. Thus $\eta \equiv \tau_2^2/\tau_1^2 = r^2 + \tau_\delta^2/\tau_1^2$ and, upon specification of $R_1(\cdot)$ and $R_\delta(\cdot)$, it is straightforward to determine the form of the correlation function $R_2(\cdot)$ for the signal process $Z_2(\cdot)$, and the cross-correlation function $R_{12}(\cdot)$ between $Z_1(\cdot)$ and $Z_2(\cdot)$, as follows

$R_2(\mathbf{x}_1-\mathbf{x}_2) = \mathrm{Cor}[Z_2(\mathbf{x}_1),Z_2(\mathbf{x}_2)] = [r^2R_1(\mathbf{x}_1-\mathbf{x}_2) + (\eta-r^2)R_\delta(\mathbf{x}_1-\mathbf{x}_2)]/\eta$ and

$$R_{12}(\mathbf{x}_1-\mathbf{x}_2)=\mathrm{Cor}[Z_1(\mathbf{x}_1),Z_2(\mathbf{x}_2)]=rR_1(\mathbf{x}_1-\mathbf{x}_2)/\sqrt{\eta}.$$

The bivariate Gaussian process $(Z_1(\mathbf{x}_1),Z_2(\mathbf{x}_2))$ has positive definite covariance structure if and only if $r^2<\eta$. When $r=0$, $Z_1(\cdot)$ and $Z_2(\cdot)$ are independent Gaussian processes. The size and direction of the dependence between $Z_1(\mathbf{x}_1)$ and $Z_2(\mathbf{x}_2)$ is measured by the cross-correlation function $R_{12}(\cdot)$. The cross-correlation is strongest when $\mathbf{x}_1=\mathbf{x}_2$ and decreases as $\mathbf{x}_1$ and $\mathbf{x}_2$ move further apart. The correlation function $R_2(\cdot)$ is a weighted average of the correlation functions $R_1(\cdot)$ and $R_\delta(\cdot)$; the $W_\delta(\cdot)$ process plays a more prominent role in adjusting $W_1(\cdot)$ locally to obtain $W_2(\cdot)$ when $r^2\ll\eta$, because $R_2(\cdot)\approx R_\delta(\cdot)$ in this case.

The calculations for the example below are made using the power exponential correlation function. For $h\in\{1,\delta\}$, the power exponential correlation function is given by

$$R_h(\mathbf{x}_1-\mathbf{x}_2)=\prod_{i=1}^{p}\exp\left(-\theta_i^h\,|x_{1,i}-x_{2,i}|^{\alpha_i^h}\right), \tag{14}$$

where $\theta_i^h>0$ and $0<\alpha_i^h\le 2$, $i=1,2$. Smaller values of the range parameters θ_i^h indicate increased dependence between the responses at fixed input sites. If $\alpha_i^h=2$, the process is infinitely mean square differentiable and the sample paths are infinitely differentiable in the i-th direction; for all other allowable values of α_i^h, the process is mean square continuous but not differentiable in the i-th direction. The correlation parameter vector for the power exponential family is $\boldsymbol{\gamma}=\left(\eta,\{\rho_i\},\theta_1^1,\ldots,\theta_p^1,\alpha_1^1,\ldots,\alpha_p^1,\theta_1^\delta,\ldots,\theta_p^\delta,\alpha_1^\delta,\ldots,\alpha_p^\delta,r\right)$.

4.2 An Example with Six Input Variables

This is a six input example with $\mathbf{x}_c=(x_1,x_2)$ and $\mathbf{x}_e=(x_3,x_4,x_5,x_6)$. The objective function is the *mean* of $z_1(\mathbf{x}_c,\mathbf{x}_e)=n_1(\mathbf{x}_c)\times o_1(\mathbf{x}_e)$ with respect to the distribution of $\mathbf{X}_e=(x_3,x_4,x_5,x_6)$ specified below where

$$n_1(x_1,x_2)=\left[x_2-\frac{5.1x_1^2}{4\pi^2}+\frac{5x_1}{\pi}-6\right]^2+\;10\left(1-\frac{1}{8\pi}\right)\cos(x_1)+10,$$

$$o_1(x_3,x_4,x_5,x_6)=2\,(2x_3)^2+4.5\,(2x_4)^{1.5}+2x_4+14x_3x_5+2\sqrt{x_4x_6}$$

and $(x_1,x_2,x_3,x_4,x_5,x_6)\in[-5,10]\times[0,15]\times[0,1]^4$. The $\mathbf{X}_e=(X_3,X_4,X_5,X_6)$ environmental variable is taken to have the discrete uniform distribution over the $3^4=81$ points in four-fold cross product of $\{0.25,0.50,0.75\}$. The objective function is

$$\mu_1(x_1,x_2)=\mathrm{E}\{z_1(x_1,x_2,\mathbf{X}_e)\}=n_1(x_1,x_2)\frac{1}{81}\sum o_1(x_3,x_4,x_5,x_6) \tag{15}$$

where the sum is over the 81 points (x_3,x_4,x_5,x_6) in $\{0.25,0.50,0.75\}^4$. The objective function roughly varies between 5 and 2,000 (see Figure 1); $\mu_1(x_1,x_2)$ has

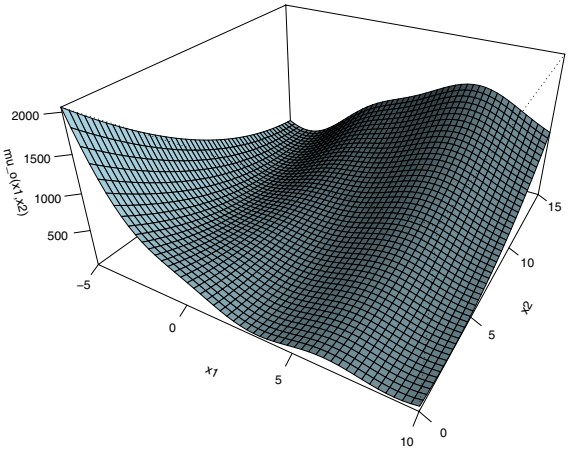


Fig. 1 The objective function (15) for the worked example.

three global minima—at $(\pi, 2.275)$, $(3\pi, 2.475)$, and $(-\pi, 12.275)$, with a (common) minimum objective function value of 1.526798.

The constraint function is the *mean* of $z_2(\mathbf{x}_c, \mathbf{x}_e) = n_2(\mathbf{x}_c) \times o_2(\mathbf{x}_e)$ with respect to the same 81 point discrete uniform distribution for $\mathbf{X}_e = (X_3, X_4, X_5, X_6)$ as $\mu_1(\cdot)$ where

$$\begin{aligned} n_2(x_1, x_2) = &- \sqrt{(10.5 - x_1)(x_1 + 5.5)(x_2 + 0.5)} \\ &- \frac{1}{30}\left(x_2 - \frac{5.1x_1^2}{4\pi^2} + \frac{5x_1}{\pi} - 6\right)^2 - \frac{1}{3}\left(1 - \frac{1}{8\pi}\right)\cos(x_1) - \frac{1}{3}, \\ o_2(x_3, x_4, x_5, x_6) = &\ 1.2\,(2x_4)\,(2x_3)^{1.3} + 4.5\,(2x_4)^3 + 2.0\,(2x_4)^{0.6} \\ &+ 3.5\,(4x_3x_5)^{1.7} + (4x_4x_6)^{0.7} \end{aligned}$$

and $(x_1, x_2, x_3, x_4, x_5, x_6) \in [-5, 10] \times [0, 15] \times [0, 1]^4$. The constraint function

$$\mu_2(x_1, x_2) = \mathrm{E}\{z_2(x_1, x_2, \mathbf{X}_e)\}$$

takes values -307.2373, -201.3822, -104.8448 at the three global $\mu_1(\cdot)$ minimizers $(-\pi, 12.275)$, $(\pi, 2.275)$, and $(3\pi, 2.475)$, respectively.

We use the constraint bound $\mu_2(x_1, x_2) \leq -250$ which gives the unique value $(-\pi, 12.275)$ as the true constrained global minimizer. This example has moderately correlated objective and constraint functions. If $\mathbf{X} = (X_1, \ldots, X_6)$ has independent components with each X_i uniformly distributed over $[0, 1]$, then $\mathrm{Cor}(z_1(\mathbf{X}), z_2(\mathbf{X})) \approx -0.51$. Thus, for this example, the bivariate predictor of the objective and constraint functions based on (13) may be more accurate than separate predictors based on individual data from each output.

In this example, we study the effect of several implementation choices on the performance of VIPER. These choices are discussed next and then the performance criteria for the resulting algorithm are defined.

- *Use of standardized or non-standardized responses:* We always center the set of objective (constraint) function values by subtracting the sample mean of the objective (constraint) values. In some runs we divided the centered objective and constraint function values by their sample standard deviation while in others we only centered the values. We denote this two-level factor by *Standardized* below.
- *Number of Monte Carlo samples used to estimate (9):* In various runs, the number of Monte Carlo samples was either 100 or 1,000. We denote this two-level factor by *MC runs* below.
- *Choice of predictor:* Schonlau et al. (1997) proposed predicting the constraint and objective functions in constrained optimization problems using independent Gaussian stochastic processes. We compare the results of using the VIPER algorithm based on independent predictors of the objective and constraint functions, which we denote by VIPER-IND, with VIPER based on the bivariate predictor obtained from (13)-(14), which we denote by VIPER-BIV. We denote this two-level factor by *Predictor* below.
- *Sequential versus one-stage design:* In addition to implementation choices, we also attempt to quantify the improvement in using the sequential algorithm compared with constrained optimization based on predictors that use the same total number of observations as the sequential procedure, but taken in a space-filling manner. We answered this question in the case of Example 4.2 by running VIPER for each combination of the three factors above, starting VIPER with 45 initial observations per output and stopping VIPER after 70 observations have been collected per output. We also ran the same constrained optimizer used by the final stage of VIPER, but based on a one-stage design that takes 70 observations per output. In the latter case, the 70 observations were taken according to a maximin distance LHD, while in the former the initial 45 observations are selected by the same criterion. We denote this two-level factor by *Design* below.

We evaluated the performance of VIPER using two criteria. The first criterion is the Euclidean distance

$$\|\mathbf{x}_{c,opt} - \widehat{\mathbf{x}}_{c,opt}\| \tag{16}$$

between the true constrained global minimizer $\mathbf{x}_{c,opt}$ and its estimator $\widehat{\mathbf{x}}_{c,opt}$ obtained as the solution to the constrained optimization problem defined by $\widehat{\mu}_1(\cdot)$ and $\widehat{\mu}_2(\cdot)$, which are the EBLUPs of $\mu_1(\cdot)$ and $\mu_2(\cdot)$ based on the total number of output evaluations. This is a bottom-line measure of performance. Our second criterion is the square root of the sum of the squared relative prediction errors of the objective and constraint functions at $\mathbf{x}_{c,opt}$

$$\left(\left[\frac{(\mu_1(\mathbf{x}_{c,opt}) - \widehat{\mu}_1(\widehat{\mathbf{x}}_{c,opt}))}{\mu_1(\mathbf{x}_{c,opt})} \right]^2 + \left[\frac{(\mu_2(\mathbf{x}_{c,opt}) - \widehat{\mu}_2(\widehat{\mathbf{x}}_{c,opt}))}{\mu_2(\mathbf{x}_{c,opt})} \right]^2 \right)^{\frac{1}{2}}. \tag{17}$$

Notice that our estimate of $\mu_1(\mathbf{x}_{c,opt})$ is the value of the estimated $\mu_1(\cdot)$ at the estimated $\mathbf{x}_{c,opt}$. Of course, both criteria are tied to the prediction accuracy of $z_1(\cdot)$ and $z_2(\cdot)$.

Table 1 ANOVA table for the main effects plus two-way interaction model for response (16).

Source	df	Sum of Squares	F ratio	P-value
Predictor	1	0.0502	1.0334	0.3560
Standardized	1	0.1632	3.3618	0.1262
MC runs	1	0.1364	2.8086	0.1546
Design	1	10.94	225.38	< 0.0001
Predictor $\times$ Standardized	1	0.0510	1.0506	0.3524
Predictor $\times$ MC runs	1	0.0284	0.5840	0.4793
Predictor $\times$ Design	1	0.1493	3.0740	0.1399
Standardized $\times$ MC runs	1	< 0.0001	0.0018	0.9676
Standardized $\times$ Design	1	0.0017	0.0351	0.8588
MC runs $\times$ Design	1	0.0751	1.5459	0.2689

Table 2 Estimates of the *Standardized* main effect for response (16).

Standardized	Mean	Standard Error
No	1.2984	0.0779
Yes	1.0963	0.0779

Table 3 Estimates of the *Design* main effect for response (16).

Design	Mean	Standard Error
Sequential	0.3703	0.0779
One-stage	2.0244	0.0779

Table 4 Estimates of the *Predictor* $\times$ *Design interaction* effects for response (16).

Interaction	Mean	Standard Error
Viper-BIV $\times$ Sequential	0.2177	0.1102
Viper-BIV $\times$ One-stage	2.0650	0.1102
Viper-IND $\times$ Sequential	0.5229	0.1102
Viper-IND $\times$ One-stage	1.9838	0.1102

4.3 Implementation Recommendations

Below we report the results of running VIPER according to the 2^4 factorial experiment described in Section 4.2, with the factors *Predictor*, *Standardized*, *MC runs*, *Design* where the responses are (16) and then (17).

4.3.1 Use of Standardization

Typical of other examples, Tables 1, 2, 5, and 6 clearly show that basing the prediction of the optimal $\mathbf{x}_c$ and the associated $\mu_1(\mathbf{x}_c)$ and $\mu_2(\mathbf{x}_c)$ on the standardized responses improves prediction accuracy. This is true whether VIPER uses the BIV or IND predictor. Only in the case where the range of the response is narrow, will there not be a noticable advantage for standardizing $y(\cdot)$. Therefore, we recommend always standardizing the response.

Table 5 ANOVA table for the main effects plus two-way interaction model for response (17).

Source	df	Sum of Squares	F ratio	P-value
Predictor	1	0.0107	2.8871	0.1500
Standardized	1	0.0361	9.7687	0.0261
MC runs	1	0.0122	3.3022	0.1289
Design	1	11.028	2984	< 0.0001
Predictor $\times$ Standardized	1	0.0003	0.0730	0.7978
Predictor $\times$ MC runs	1	0.0014	0.3816	0.5638
Predictor $\times$ Design	1	0.0459	12.423	0.0168
Standardized $\times$ MC runs	1	0.0016	0.4272	0.5422
Standardized $\times$ Design	1	0.0021	0.5708	0.4840
MC runs $\times$ Design	1	0.0147	3.9690	0.1030

Table 6 Estimates of the *Standardized* main effect for response (17).

Standardized	Mean	Standard Error
No	1.0340	0.0215
Yes	0.9390	0.0215

Table 7 Estimates of the *Design* main effect for response (17).

Design	Mean	Standard Error
45	0.1562	0.0215
70	1.8167	0.0215

Table 8 Estimates of the *Predictor* $\times$ *Design interaction* effects for response (17).

Interaction	Mean	Standard Error
Viper-BIV $\times$ Sequential	0.0768	0.0304
Viper-BIV $\times$ One-stage	1.8444	0.0304
Viper-IND $\times$ Sequential	0.2356	0.0304
Viper-IND $\times$ One-stage	1.7889	0.0304

4.3.2 Monte Carlo Sampling

While as few as 100 MC samples can be adequate to produce estimates of the posterior expected improvement (7), our experience is that VIPER produced more accurate and stable estimates when 1,000 MC samples were drawn. Although the main effects of MC were not significant for any of the three criteria, the mean of each criterion was smaller for 1,000 MC runs than for 100 MC runs. For (16), the mean for 100 MC runs is 1.2897 and for 1,000 MC runs is 1.1050. For (17), the mean for 100 MC runs is 1.0141 and for 1,000 MC runs is 0.9588.

For the examples we have investigated (problems of 6 and fewer input dimensions), using more than 1,000 MC samples did not improve the performance of VIPER relative to results based on 1,000 MC samples. However if the dimension of the input space grows, we conjecture that larger numbers of MC samples would be required for VIPER to operate most effectively.

4.3.3 Choice of Predictor

Although the main effect of *Predictor* is not significant in Table 1 or 5 for (16) or (17), the interaction between *Predictor* and *Design* is significant for (17). Tables 4 and 8 indicate that VIPER-BIV yields smaller mean errors for the sequential procedure but slightly larger mean errors for the one-stage procedure. Thus using the bivariate-model predictor appears to be superior to using independent predictors in the sequential VIPER procedure. Because the sequential procedure outperforms the one-stage procedure (see Section 4.4.1), the VIPER-BIV algorithm is recommended.

Other examples we have run have had both smaller and greater correlations between the objective and constraint functions. In general, VIPER-BIV is more effective with examples having large values of $\mathrm{Cor}(z_1(\mathbf{X}), z_2(\mathbf{X}))$ because of the ability of the bivariate predictor to glean information about each function from all $2n$ observations.

Thus it would seem that VIPER-BIV would always be preferred to VIPER-IND, at least when using a sequential strategy. However, there is one cautionary note to this recommendation. VIPER-BIV requires greater computational effort than VIPER-IND. This is because if n observations have been taken from both the objective and constraint functions, the BIV predictor of $z_1(\cdot)$ must invert a $2n \times 2n$ correlation matrix whereas the VIPER-IND predictors must invert two $n \times n$ correlation matrices. Thus if it is known that the objective and constraint functions are likely to be only mildly correlated, VIPER-IND will be quicker to run and will produce the same answer as VIPER-BIV.

4.3.4 Stopping Rules

This example stops VIPER after a fixed number of additional sites have been added to the initial input sites. In a more automatic mode, one requires an adaptive stopping rule to terminate VIPER. We recommend that the algorithm should *not* be terminated after a cycle in which there is a single small maximum expected improvement. The reasons for this are (1) there can be numerical uncertainties in the calculated value of maximum expected improvement caused by the fact that the expected improvement surface can contain many local optima and (2) the expected improvement surface can change somewhat dramatically from one iteration to the next as correlation parameters are re-estimated. We have investigated stopping the algorithm when a longer series of small expected improvements suggested that it be terminated. We recommend terminating the algorithm when both a moving *average* and a moving *range* of the expected improvements become "small." Formally, suppose that m iterations have been completed, then let $\widehat{I}_j$ denote the observed maximum expected improvement at stage j, let $\mathscr{A}_j = (\widehat{I}_j + \cdots + \widehat{I}_{j-g+1})/g$ denote a moving average of window length g terminating at stage j, and let $\mathscr{R}_j = \max\{\widehat{I}_j, \ldots, \widehat{I}_{j-g+1}\} - \min\{\widehat{I}_j, \ldots, \widehat{I}_{j-g+1}\}$ denote the range of the same g maximum expected improvements. Then we terminate the algorithm if $\mathscr{A}_m \leq 0.0005$ and $\mathscr{R}_m \leq 0.005$ although, in general, these tolerances are problem specific. The stopping criteria should be "small" relative to typical values of the moving averages and ranges observed at the start of the algorithm. It is

also recommended that the stopping criteria be met by two or more successive values of the moving average and range before stopping the algorithm.

4.4 Other Conclusions

Below we report the results of pursuing a sequential design strategy relative to adopting a fixed design. Finally, we compare accuracy in predicting values of the objective and constraint functions based on the final designs produced by VIPER-BIV and VIPER-IND.

4.4.1 Sequential versus One-stage Design

In every example we have run, the sequential design produced by VIPER outperforms a naive one-stage optimization based on the same number of code runs. This is clearly seen in Tables 1, 3, 5, and 7. The improvement in this example is on the order of 6-10 based on Tables 3 and 7. Intuitively, this is not surprising. The sequential procedure makes use of information at each stage about current estimates of the location of the optimum and where the constraint is satisfied, whereas the one-stage strategy can not.

4.4.2 Effect of Predictor on Overall Prediction Accuracy

While the goal of VIPER is to identify global optima, it is still of interest to assess the effect of *Predictor* on the overall accuracy of prediction. We assess this effect in two ways. First, we compute, for a space-filling set of $\mathbf{x}$ input values, the average empirical root mean square prediction error (ERMSPE) for $z_1(\mathbf{x})$ and $z_2(\mathbf{x})$ predictors based on the final designs produced by VIPER-BIV and VIPER-IND. The ERMSPE of the predictor $\widehat{z}(\mathbf{x})$ of a generic function $z(\mathbf{x})$ over the set of points $\{\mathbf{x}_i\}_{i=1}^m$ is defined to be

$$\left(\frac{1}{m}\sum_{i=1}^{m}\left(\widehat{z}(\mathbf{x}_i)-z(\mathbf{x}_i)\right)^2\right)^{1/2}.$$

For each of $z_1(\cdot)$ and $z_2(\cdot)$, we computed the ERMSPE in Example 4.2 for the VIPER-BIV and VIPER-IND predictors, each algorithm starting with 45 inputs (taken to be a maximin distance LHD in 6-dimensions) and stopping after 70 evaluations. We evaluated the $z_1(\mathbf{x})$ and $z_2(\mathbf{x})$ predictors at a set of 300 6-d inputs that were seletected to form a maximin distance LHD (50 points per dimension). These ERMSPE values are given in Table 9. The function $z_1(\cdot)$ ranges over a relatively narrow range compared with $z_2(\cdot)$. In this case, the IND predictor is slightly more accurate for the narrow-range response and greatly inferior for the wide-range re-

Table 9 ERMSPE values for the VIPER-BIV and VIPER-IND predictors.

	$z_1(\cdot)$	$z_2(\cdot)$
VIPER-BIV	0.7103	89.2995
VIPER-IND	0.6885	115.0035

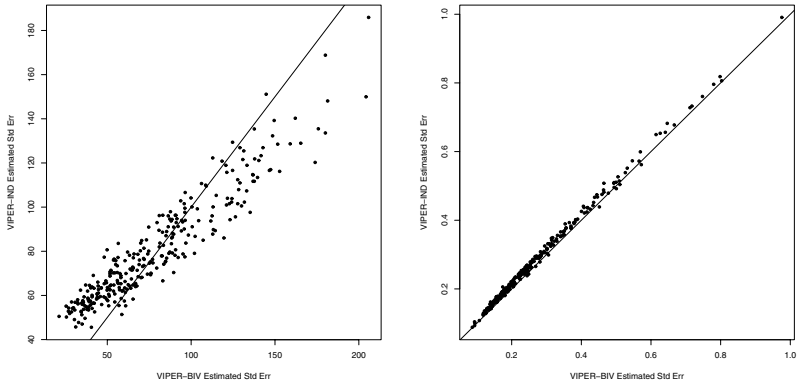


Fig. 2 Comparison of estimated standard errors using VIPER-BIV and VIPER-IND of $\{z_1(\mathbf{x}_i)\}$ and $\{z_2(\mathbf{x}_i)\}$ for the 300 space-filling points $\{\mathbf{x}_i\}$ at which these functions are evaluated in the worked example.

sponse. In most, but not all examples that we constructed, the bivariate predictor has smaller ERMSPE for both components.

Second, we study the estimated standard errors of the 300 space-filling points from the previous paragraph. Figure 2 plots the estimated standard errors for the VIPER-IND and VIPER-BIV predictors of $z_1(\cdot)$ and $z_2(\cdot)$. The left panel corresponds to $z_1(\cdot)$ and the scatter about the 45° line shows that VIPER-BIV produces smaller estimated standard errors for those points with "small" estimated standard errors and larger standard errors for those points with "large" estimated standard errors. The right panel corresponds to $z_2(\cdot)$ and the scatter about the 45° line shows that VIPER-BIV almost uniformly yields smaller estimated standard errors. In this case, VIPER-BIV outperforms VIPER-IND.

5 Discussion

Our implementation of VIPER uses traditional minimization methods, appropriate for rapidly computed functions, in three places. First, the correlation parameters of the empirical best linear unbiased predictors are estimated by restricted maximum likelihood (or other methods that also require numerical optimization); second, the control portion of the updated input site is selected to maximize the poster expected improvement; third, the environmental portion of the updated input site is chosen to minimize the posterior mean square prediction error. The maximization of the

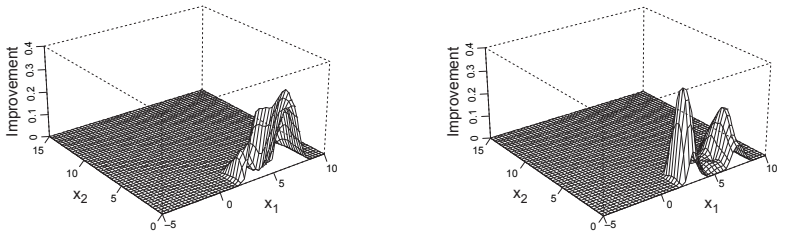


Fig. 3 Typical estimated posterior expected improvement surfaces.

posterior expected improvement is particularly difficult; typically, this surface is highly multi-modal because it is nearly zero throughout most of the control variable space, but has areas of large improvement concentrated in specific regions (see Figure 3). Even using multiple starting points, standard minimization algorithms such as the Nelder-Mead simplex algorithm, can become trapped in a local optimum of the expected improvement surface if no starting points are located near these regions (Nelder & Mead 1965).

The multiple-mode nature of all three of the surfaces described in the previous paragraph motivated the use of a space-filling set of sites to locate promising starting points. The starting points for each optimization were obtained by estimating the posterior expected improvement on a 200-point maximin distance LHD in the control variable space, and selecting the five input sites with the highest value for further improvement. The Nelder-Mead simplex algorithm was applied to improve each of these five input values with random generation of points. Finally a quasi-Newton algorithm was applied to the output of the Nelder-Mead stopping value in a further attempt to maximize the surface.

When the control variable space is of high dimension, the number of sites needed to fill the control space is large, extending the run time of the algorithm. This situation can potentially be remedied by a strategy that concentrates starting points near areas where local improvement has been demonstrated, and in boundary regions where global searches are often made because of large prediction uncertainty.

Correlation parameter estimation is the most time-consuming component of running the VIPER algorithm. One approach to reducing the time associated with this aspect of the algorithm is to update the correlation parameter estimates only after *groups* of input sites have been added, rather than after each input site has been added. While it is computationally simple to implement this strategy using a fixed group size, say update the correlation parameter estimates after every five inputs have been added, a more sophisticated approach to the grouping is based on the following observations. Correlation parameter estimates tend to fluctuate considerably at the outset of the algorithm, and then they stabilize as more input sites are added to the design. Thus a correlation parameter estimation approach that requires more frequent updates for smaller designs and fewer updates for larger designs has intuitive appeal. This would reduce run times because the time to perform correlation parameter estimation increases substantially with design size.

The spatial autoregressive model of Section 4 is a specific application of the general formulas from Section 3. In some settings, it may be more reasonable to assume that the strongest cross-correlation between $Z_1(\mathbf{x}_1)$ and $Z_2(\mathbf{x}_2)$ occurs when $\mathbf{x}_1$ and $\mathbf{x}_2$ differ by some unknown spatial shift parameter $\boldsymbol{\Delta}$. For example, suppose $Z_1(\cdot)$ and $Z_2(\cdot)$ represent the spatial distribution of rainfall over a fixed geographical area at two distinct time points. The spatial shift parameter would model the movement of any weather systems between the two times at which rainfall measurements are taken. Ver Hoef & Barry (1998) include spatial shift in their development of methods for fitting variogram and cross-variogram models to spatial data. Equation (13) is easily modified as $W_2(\mathbf{x}) = rW_1(\mathbf{x}-\boldsymbol{\Delta}) + W_\delta(\mathbf{x})$ to include a spatial shift parameter $\boldsymbol{\Delta}$. This modified model gives $R_{12}(\mathbf{x}_1-\mathbf{x}_2) = rR_1(\mathbf{x}_1-\mathbf{x}_2+\boldsymbol{\Delta})/\sqrt{\eta}$, so the dependence between $Z_1(\mathbf{x}_1)$ and $Z_2(\mathbf{x}_2)$ is strongest when $\mathbf{x}_2-\mathbf{x}_1 = \boldsymbol{\Delta}$.

An alternative to the spatial autoregressive model that treats the objective and constraint functions symmetrically was constructed by Ver Hoef & Barry (1998). Let $W_0(\cdot)$, $W_1(\cdot)$ and $W_2(\cdot)$ denote mutually independent, mean zero, Gaussian white noise processes. Set

$$Z_i(\mathbf{w}) = \sqrt{1-\rho_i^2}\, W_i(\mathbf{w}) + \rho_i W_0(\mathbf{w}-\boldsymbol{\Delta}_i)\,,$$

where $-1 \le \rho_i \le 1$ and the $\boldsymbol{\Delta}_i$ are spatial shift parameters. The models for the objective and constraint functions are the integrated Z_i processes with respect to a square integrable moving average function $g_i(\cdot|\boldsymbol{\theta}_i)$ plus a nonzero mean $\boldsymbol{\beta}^\top \mathbf{f}_i(\cdot)$:

$$Y_i(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{f}_i(\mathbf{x}) + \int g_i(\mathbf{w}-\mathbf{x}|\boldsymbol{\theta}_i)\, Z_i(\mathbf{w})\, d\mathbf{w}\,.$$

Many commonly used variogram models can be reproduced with this moving average construction.

VIPER can be applied to output that contains measurement errors. If, in (1), $\varepsilon_1(\mathbf{x})$ and $\varepsilon_2(\mathbf{x})$ are independent zero-mean Gaussian white noise processes, then $\mathbf{R}_1$ and $\mathbf{R}_2$ are modified by adding the variances of the associated Gaussian measurement errors $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$ to the diagonal elements of $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively. In particular, for the spatial autoregressive model,

$$\mathrm{Cov}(Z_1(\mathbf{x}_1), Z_1(\mathbf{x}_2)) = \tau_1^2 R_1(\mathbf{x}_1-\mathbf{x}_2) + \sigma_1^2 \delta_{[\mathbf{x}_1=\mathbf{x}_2]}$$

where $\delta_{[\mathbf{x}_1=\mathbf{x}_2]} = 1$ or 0 according as $\mathbf{x}_1 = \mathbf{x}_2$ or $\mathbf{x}_1 \neq \mathbf{x}_2$. Similarly

$$\mathrm{Cov}(Z_2(\mathbf{x}_1), Z_2(\mathbf{x}_2)) = \tau_1^2 \left[r^2 R_1(\mathbf{x}_1-\mathbf{x}_2) + (\eta - r^2) R_\delta(\mathbf{x}_1-\mathbf{x}_2)\right] + \sigma_2^2 \delta_{[\mathbf{x}_1=\mathbf{x}_2]}.$$

The matrix $\mathbf{R}_{12}$ is unchanged from the deterministic, non-error case.

Another case where VIPER applies is when the environmental variables have different distributions that define the objective and constraint functions. While this may not occur frequently, the general formulas of Section 3 apply with the modification that the $\mu_1(\cdot)$ and $\mu_2(\cdot)$ estimates must be summed over different support sets with their associated weights.

Section 4 presented one approach to identifying a stopping rule for the VIPER algorithm. Other approaches are certainly worth considering. For example, a stopping rule based on the sequence of predicted global optimum values of the empirical BLUP and the associated prediction uncertainties could be examined. The difficulty in identifying a "best" rule lies in the fact that the criterion used for stopping is not monotone as the number of input sites is increased and the actual values of the objective function being optimized are never directly observed. This is the one aspect of the algorithm that requires monitoring by the investigator, as choice of an appropriate stopping criterion is problem specific and requires tracking the expected improvement sequence in its entirety. Also, when the total number of observations is limited, there are significant issues involving allocation of effort between the initial design and the subsequent sequential search that will affect the stopping rule.

In applications where $y_1(\cdot)$ and $y_2(\cdot)$ require separate computer runs, VIPER can be made more efficient by modifying the algorithm to select only *one* of $y_1(\cdot)$ and $y_2(\cdot)$ at each stage. For example, it may be rather simple to determine the feasibility of a subregion of the domain in which the minimizer of the objective function appears to be located; in such a case, further evaluation of $y_2(\cdot)$ would be wasteful of resources. It is not difficult to define appropriate maximum expected improvements for each of $y_1(\cdot)$ and $y_2(\cdot)$ and, in each cycle, to select both the next function to be evaluated as well as the control portion of the next input at which to evaluate that function. The modified VIPER would select the function $y_i(\cdot)$ having greater maximum expected improvement and $\mathbf{x}_c$ as the point that produces this maximum for the selected function. Thus, for example, the algorithm might decide to evaluate only $y_1(\mathbf{x})$ from some point onward. See Lehman (2002) for more details about such a modification.

The full Bayesian approach to working with the correlation parameters is also readily available. The posterior expected improvement (9) and mean square prediction error (12) could be estimated based on a *sample* from the posterior distribution of $\boldsymbol{\gamma}$ given $\mathbf{Y}_1^n$ and $\mathbf{Y}_2^n$. This distribution is provided, up to the normalizing constant, in (6). Sampling from this distribution is non-trivial, requiring an MCMC approach. However, the time required to obtain such a sample for each iteration of the algorithm may turn out to be much less than the current approach of obtaining the posterior mode.

Finally we note that the VIPER algorithm can be extended to incorporate multiple (≥ 2) constraints. Identify the control variable settings $\mathbf{x}_c^*$ that minimize $\mu_1(\cdot)$ subject to constraints on $\mu_i(\cdot)$, $i = 2, \ldots, q$, i.e.,

$$\mathbf{x}_c^* = \underset{\mathbf{x}_c \in \mathscr{X}_c}{\operatorname{argmin}}\ \mu_1(\mathbf{x}_c) \quad \text{subject to} \quad \mu_i(\mathbf{x}_c^*) \leq u_i, \quad i = 2, \ldots, q\,.$$

The improvement function is extended to incorporate the additional constraints:

$$i_n(\mathbf{x}_c) = \max\{0, \mu_1^{\min} - \mu_1(\mathbf{x}_c)\} \times \chi[\mu_2(\mathbf{x}_c) \leq u_2] \times \cdots \times \chi[\mu_q(\mathbf{x}_c) \leq u_q]\,.$$

For simplicity of notation, we suppose $(M_1(\mathbf{x}_c), \ldots, M_q(\mathbf{x}_c))$ has a multivariate Gaussian distribution with mean vector $\mathbf{m}$ and covariance matrix $\tau_1^2\mathbf{R}$. The mean and

covariance are computed conditional on all quantities required to calculate the posterior expected improvement analogous to (3). We assume τ_1^2 is distributed as inverse chi-square with ν degrees of freedom and scale parameter $\nu\widehat{\tau}_1^2$. Define the following quantities: $U_1 = (M_1^{\min} - \mathbf{m}_1)/\sqrt{\widehat{\tau}_1^2 \mathbf{R}_{11}}$, $U_i = (u_i - \mathbf{m}_i)/\sqrt{\widehat{\tau}_1^2 \mathbf{R}_{ii}}$ for $i = 2, \ldots, q$, and $\mathbf{C} = \text{diag}(\mathbf{R})^{-1/2}\, \mathbf{R}\, \text{diag}(\mathbf{R})^{-1/2}$. Let $\begin{pmatrix} 1 & \mathbf{c}_i^\top \\ \mathbf{c}_i & \mathbf{C}_i \end{pmatrix}$ denote the matrix formed by permuting the i-th row and column of $\mathbf{C}$ with its first row and column. Define $\widetilde{\mathbf{C}}_i = \mathbf{C}_i - \mathbf{c}_i \mathbf{c}_i^\top$,

$$\widetilde{\mathbf{U}}_i = \left(\frac{U_1 - U_i\, \mathbf{c}_{i1}}{\sqrt{\widetilde{\mathbf{C}}_{i,11}}}, \ldots, \frac{U_{i-1} - U_i\, \mathbf{c}_{i,i-1}}{\sqrt{\widetilde{\mathbf{C}}_{i,(i-1,i-1)}}}, \frac{U_{i+1} - U_i\, \mathbf{c}_{i,i}}{\sqrt{\widetilde{\mathbf{C}}_{i,ii}}}, \ldots, \frac{U_q - U_i\, \mathbf{c}_{i,q-1}}{\sqrt{\widetilde{\mathbf{C}}_{i,(q-1,q-1)}}} \right),$$

and $\widetilde{\mathbf{R}}_i = \text{diag}(\widetilde{\mathbf{C}}_i)^{-1/2}\, \widetilde{\mathbf{C}}_i\, \text{diag}(\widetilde{\mathbf{C}}_i)^{-1/2}$.

The posterior expected improvement is given by

$$\mathrm{E}\{I_n(\mathbf{x}_c)\} = \sqrt{\widehat{\tau}_1^2 \mathbf{R}_{11}} \left[U_1\, T_{q,\mathbf{C}}(U_1, \ldots, U_q, \nu) + \sum_{i=1}^{q} \mathbf{C}_{1i}\, C_\nu(U_i)\, T_{q-1,\widetilde{\mathbf{R}}_i}\left(\frac{\widetilde{\mathbf{U}}_i}{\zeta_\nu(U_i)}, \nu - 1 \right) \right]$$

where $T_{p,\boldsymbol{\Sigma}}(\cdot, \nu)$ is the p-variate t cumulative distribution function with ν degrees of freedom, zero mean and scale matrix $\boldsymbol{\Sigma}$,

$$C_\nu(w) = \sqrt{\frac{\nu}{\nu - 2}}\, t_{\nu-2}\left(w\sqrt{\frac{\nu - 2}{\nu}} \right), \quad \zeta_\nu(w) = \sqrt{\frac{w^2 + \nu}{\nu - 1}},$$

and $t_\nu(\cdot)$ is the standard t density function with ν degrees of freedom. It is straightforward to extend this improvement criterion to accommodate lower bound or interval constraints.

A practical problem implementing such an algorithm with multiple constraints is that correlation parameter estimation will become substantially more difficult if product correlation structures are used to model each response. The additional computation of the multivariate t cumulative distribution function will slow estimation of posterior expected improvements. One approach to modeling in such a setting is to increase the complexity of the global trend (regression) part of the model at the expense of the local trend (correlation) part of the model. This would reduce the dimensionality of the correlation parameter vector, making sampling or estimation easier. More generally, empirical work would also need to be accomplished to determine if the algorithm proposed herein is as successful at locating the constrained optimum with simple correlation parameter models as it is with the more complicated product structures.

Acknowledgements This work was sponsored, in part, by grants DMS-0406026 and DMS-0806134 (The Ohio State University) from the National Science Foundation. The authors would like to thank Han Gang for computational help.

References

Bernardo, M. C., Buck, R., Liu, L., Nazaret, W. A., Sacks, J. & Welch, W. J. (1992). Integrated circuit design optimization using a sequential strategy, *IEEE Transactions on Computer-Aided Design* **11**: 361–372.

Chang, P. B., Williams, B. J., Bawa Bhalla, K. S., Belknap, T. W., Santner, T. J., Notz, W. I. & Bartel, D. L. (2001). Robust design and analysis of total joint replacements: Finite element model experiments with environmental variables, *Journal of Biomechanical Engineering* **123**: 239–246.

Currin, C., Mitchell, T. J., Morris, M. D. & Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments, *Journal of the American Statistical Association* **86**: 953–963.

Handcock, M. S. & Stein, M. L. (1993). A bayesian analysis of kriging, *Technometrics* **35**: 403–410.

Haylock, R. G. & O'Hagan, A. (1996). On inference for outputs of computationally expensive algorithms with uncertainty on the inputs, *in* J. Bernardo, J. Berger, A. Dawid & A. Smith (eds), *Bayesian Statistics*, Vol. 5, Oxford University Press, pp. 629–637.

Jones, D. R., Schonlau, M. & Welch, W. J. (1998). Efficient global optimization of expensive black–box functions, *Journal of Global Optimization* **13**: 455–492.

Kennedy, M. C. & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available, *Biometrika* **87**: 1–13.

Lehman, J. (2002). *Sequential Design of Computer Experiments for Robust Parameter Design*, PhD thesis, Department of Statistics, Ohio State University, Columbus, OH USA.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization, *Computer Journal* **7**: 308–313.

O'Hagan, A. (1992). Some bayesian numerical analysis, *in* J. Bernardo, J. Berger, A. Dawid & A. Smith (eds), *Bayesian Statistics*, Vol. 4, Oxford University Press, pp. 345–363.

Ong, L.-T. (2004). *Stability of Uncemented Acetabular Components: Design, Patient-Dependent, and Surgical Effects & Complications Due to Trapped Interfacial Fluid*, PhD thesis, Sibley School of Mechanical and Aerospace Engineering, Cornell University.

Sacks, J., Schiller, S. B. & Welch, W. J. (1989). Designs for computer experiments, *Technometrics* **31**: 41–47.

Santner, T. J., Williams, B. J. & Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*, Springer Verlag, New York.

Schonlau, M., Welch, W. J. & Jones, D. R. (1997). Global versus local search in constrained optimization of computer models, *Technical Report RR-97-11*, University of Waterloo and General Motors.

Schonlau, M., Welch, W. J. & Jones, D. R. (1998). Global versus local search in constrained optimization of computer models, *in* N. Flournoy, W. Rosenberger & W. Wong (eds), *New Developments and Applications in Experimental Design*, Vol. 34, Institute of Mathematical Statistics, pp. 11–25.

Tang, B. (1993). Orthogonal array-based latin hypercubes, *Journal of the American Statistical Association* **88**: 1392–1397.

Ver Hoef, J. M. & Barry, R. P. (1998). Constructing and fitting models for cokriging and multivariable spatial prediction, *Journal of Statistical Planning and Inference* **69**: 275–294.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J. & Morris, M. D. (1992). Screening, predicting, and computer experiments, *Technometrics* **34**: 15–25.

Williams, B. J., Santner, T. J. & Notz, W. I. (2000a). Sequential design of computer experiments for constrained optimization of integrated response functions, *Technical Report 658*, Department of Statistics, The Ohio State University.

Williams, B. J., Santner, T. J. & Notz, W. I. (2000b). Sequential design of computer experiments to minimize integrated response functions, *Statistica Sinica* **10**: 1133–1152.